UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA DE INGENIERIA DE SISTEMAS INFORMÁTICOS

MINERÍA DE DATOS PARA LA GENERACIÓN DE CONOCIMIENTO DE LA ATENCIÓN A RECLAMACIONES Y DE LOS SONDEOS DE PRECIOS PARA LA UNIDAD DE ANÁLISIS DE CONSUMO Y MERCADOS DE LA DEFENSORÍA DEL CONSUMIDOR

PRESENTADO POR:

MOISÉS DANIEL HERRERA CRIOLLO

WALTER OSWALDO LEMUS VÁSQUEZ

VLADIMIR ALBERTO SÁNCHEZ CASTANEDA

ÍTALO ALEXANDER UMAÑA RUBIO

PARA OPTAR AL TITULO DE:

INGENIERO DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, OCTUBRE DE 2020

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MAESTRO ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

MSc. FRANCISCO ANTONIO ALARCÓN SANDOVAL

FACULTAD DE INGENIERIA Y ARQUITECTURA

DECANO:

PhD. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

Ing. JULIO ALBERTO PORTILLO

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

Ing. RUDY WILFREDO CHICAS VILLEGAS

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA DE INGENIERIA DE SISTEMAS INFORMATICOS

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO DE SISTEMAS INFORMATICOS

Título:

MINERÍA DE DATOS PARA LA GENERACIÓN DE CONOCIMIENTO DE LA ATENCIÓN A RECLAMACIONES Y DE LOS SONDEOS DE PRECIOS PARA LA UNIDAD DE ANÁLISIS DE CONSUMO Y MERCADOS DE LA DEFENSORÍA DEL CONSUMIDOR

Presentado por:

MOISÉS DANIEL HERRERA CRIOLLO

WALTER OSWALDO LEMUS VÁSQUEZ

VLADIMIR ALBERTO SÁNCHEZ CASTANEDA

ÍTALO ALEXANDER UMAÑA RUBIO

Trabajo de Graduación Aprobado por:

Docente Asesor:

MSc. ELMER ARTURO CARBALLO RUIZ

SAN SALVADOR, OCTUBRE DE 2020

Trabajo de Graduación Aprobado por:

Docente Asesor:

MSc. ELMER ARTURO CARBALLO RUIZ

AGRADECIMIENTOS

Desde el día en que decidimos como equipo aventurarnos en la temática de nuestro trabajo de graduación supe que sería un gran reto el cual fue posible culminar gracias al apoyo de muchas personas que influyeron en cierta medida para que esto fuera posible.

Agradecer primeramente a mis padres Ángel Herrera y Ana Criollo quienes siempre me han apoyado y motivado en todas mis metas y objetivos en todo lo largo de mi vida, quienes han estado incondicionalmente en los momentos en los que más los he necesitado, y que para mí siempre han sido, son y serán un ejemplo a seguir.

Agradezco al gran equipo que me acompaño en el desarrollo del presente trabajo: Alberto, Walter e Ítalo, con quienes tuve la dicha de ser compañero en las aulas y en este proyecto y sobre quienes puedo decir con toda categoría son personas muy capaces, grandes profesionales y con los cuales logré tener una gran sinergia lo cual influyó en el resultado tan satisfactorio que hemos obtenido.

Así mismo agradecer al ingeniero Elmer Arturo Carballo quien fue nuestro docente y en está ocasión nuestro asesor profesional el cual nos brindó de sus conocimientos, nos supo orientar y responder a todas las inquietudes y desafíos con los cuales nos encontramos en el camino.

Agradezco al Licenciado Jorge Salvador Pocasangre gerente de Sistemas Informáticos y a la Licenciada Diana Carolina Castro Orellana jefa de la Unidad de Análisis de Consumo y Mercados de la Defensoría del Consumidor quienes nos abrieron las puertas en la institución y estuvieron al pendiente y brindando seguimiento debido al proyecto.

Gracias a todos los familiares, amigos y conocidos de mis compañeros o míos quienes apoyaron en su momento de manera directa o indirecta cuando se les necesito y que supieron responder con mucha amabilidad, poniendo también de su parte para que la tan anhelada meta fuera cumplida.

Moisés Herrera

AGRADECIMIENTOS

Agradecer a Dios, por brindarme sabiduría, paciencia, fuerza y salud, también, por guiarme en todo momento, nunca dejarme desamparado, y así, permitirme llegar hasta este punto y poder culminar este gran paso en mi vida.

A mi madre, ya que este logro se lo debo a ella, por su paciencia, amor, apoyo incondicional en todas las etapas de mi vida, por ser mi inspiración, por estar siempre a mi lado, alentándome en todo momento y siempre creyendo en mí.

A mi padre, que en paz descanse, por ser mi ejemplo a seguir, por brindarme su apoyo, por creer indudablemente que podía lograrlo, alentarme a nunca darme por vencido y a luchar siempre por lo que quiero.

A mi hermano y a mis primos Wendy Luna y Victor Luna, por brindarme consejos, ayuda, apoyo y alentarme a continuar siempre adelante y de la mejor manera posible.

A mis amigos, compañeros de grupo, Moisés, Alberto e Ítalo; doy gracias a Dios por ponerlos en mi camino, conocíamos el reto al cual nos enfrentábamos con este proyecto, pero en ningún momento dudamos en tomarlo, conozco y admiro el potencial que cada uno de ellos posee, y no dudo que este solo es el primer logro de muchos que tendrán en su vida.

A nuestro asesor, Ing. Elmer Arturo Carballo por orientarnos de la mejor manera posible en la realización de este proyecto, por compartirnos su conocimiento y poder culminarlo de manera exitosa.

A la Licda. Diana Carolina Castro y al Lic. Jorge Salvador Pocasangre, por abrirnos las puertas en la Defensoría del Consumidor, que sin su ayuda este proyecto no hubiese sido realidad.

A mi familia, compañeros, amigos y demás personas que directa o indirectamente me acompañaron y me apoyaron para lograr dar este paso, les agradezco de corazón por hacer que este camino fuese más fácil y por estar ahí siempre que los necesité.

Walter Lemus

AGRADECIMIENTOS

A mis padres, por creer en mí y apoyarme en cada paso de mi formación académica, por sus consejos, valores, motivación constante, inculcarme siempre la perseverancia y constancia que nos caracteriza como familia y ser mi inspiración para culminar con éxito una nueva etapa en mi vida, pero ante todo por su amor.

A mis hermanas porque siempre me brindan amor, ayuda, paciencia y apoyo incondicional.

A mis abuelos por cuidarme siempre y por tantos consejos que me han brindado.

A nuestro asesor ingeniero Elmer Arturo Carballo, por su profesionalidad y apoyo presentado en el desarrollo de este trabajo, pero principalmente por su calidad en la docencia.

A nuestro jurado Ingeniera Nelly Henríquez e Ingeniero Arnoldo Rivas por su guía y conocimiento y ser parte importante en la realización del proyecto, así todos los docentes de la Universidad de El Salvador que de alguna manera apoyaron y nos brindaron de su conocimiento a lo largo de nuestra formación académica.

A la Licenciada Diana Carolina Castro Orellana y al Licenciado Jorge Salvador Pocasangre de la Defensoría del Consumidor, por brindarnos su apoyo y herramientas necesarias para la elaboración exitosa de este logro.

A mis compañeros de estudio y amigos Moisés Herrera, Walter Lemus e Ítalo Umaña un agradecimiento muy especial por la dedicación y el esmero por el cual logramos realizar exitosamente este trabajo de graduación. A quienes les deseo muchos éxitos en su vida profesional y no dudo que lograran realizar todas sus metas.

Por último, a mis amigos y compañeros tanto de estudio como de trabajo por su apoyo y ayuda que nos han brindado desde el primero al último día.

Alberto Castaneda

AGRADECIMIENTOS

A mis padres por todo el apoyo brindado durante todo el transcurso de la carrera ya que sin ellos no hubiera podido culminar este logro.

A mis amigos y compañeros Alberto, Moisés y Walter por toda la dedicación y esfuerzo les deseo muchos éxitos en sus futuros proyectos.

Y por último a familia, amigos y compañeros por todo el apoyo brindado.

Italo Umaña

Contenido

1 Introducción

En la nueva era digital las empresas generan grandes volúmenes de datos y con mucha frecuencia estos son solamente almacenados, pocas veces son tratados y analizados, no fue hasta que empresas punteras del sector descubrieron la importancia que puede tener la exploración de los datos, y los resultados que se pueden obtener al disponer de información útil para la toma de decisiones. Por ello nace la minería de datos o exploración de datos. Básicamente, es un conjunto de herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información interesante, novedosa y potencialmente útil de grandes bases de datos o volúmenes de datos que puede ser utilizada como soporte para la toma de decisiones.

En este trabajo se trata de presentar al lector a partir los antecedentes de la minería de datos como tal, desde un ámbito global hasta un contexto de nuestro país, retomando las bases estadísticas y probabilísticas en las que se basa la Minería de Datos y como después de estas surgen otras para cubrir necesidades que surgen en las empresas que tienen grandes cantidades de información.

Según los objetivos del análisis de los datos, los algoritmos pueden clasificarse de dos tipos, “Algoritmos supervisados” y “Algoritmos no supervisados”. Los algoritmos supervisados generan un modelo predictivo basado en datos de entrada y salida. La palabra “supervisados” viene de la idea de tener un conjunto de datos previamente etiquetado y clasificado, llamados “datos de entrenamiento”, el algoritmo va “aprendiendo” a clasificar a partir de estos datos comparando con el resultado del modelo. Entre estos están las técnicas de predicción como las regresiones, y las técnicas de clasificación, como los árboles de decisión, redes neuronales, y las clasificaciones bayesianas que se derivan del teorema de Bayes. Como se mencionó anteriormente el segundo tipo de algoritmo es el no supervisado, estos al ser la contraparte del tipo anterior, no parten de un conjunto de datos etiquetados para extraer un patrón o dicho en otras palabras permite extraer conocimiento de un conjunto de datos donde a priori se desconoce. En este tipo de algoritmo se encuentran los algoritmos de “clustering” y “reglas de asociación”.

Entre las técnicas, algoritmos y modelos para la minería de datos tenemos, “Regresión lineal”, esta se puede definir como el modelo de relación entre dos variables que se representan con una ecuación lineal. Asimismo, existen las “Series Temporales”, que se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. También, se encuentran las Redes Neuronales que son un modelo computacional vagamente inspirado en el comportamiento del cerebro humano, estas consisten en un conjunto de unidades que son llamadas “neuronas artificiales”, dichas “neuronas” están conectadas entre sí para transmitirse señales. La información que entra atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida. Igualmente, están los modelos de “Clasificación”, estos son una forma de análisis de datos que extrae modelos que describen clases de datos importantes. Este modelo también llamado “clasificadores”, predicen etiquetas de clase categórica como “discretas” o “no ordenadas”. Entre los algoritmos de clasificación tenemos: “Algoritmo de Naive Bayes”, “Algoritmo de K vecinos cercanos” y “Árboles de decisión”. Por último, se encuentra el algoritmo de “Reglas de asociación”, que tiene como objetivo encontrar relaciones dentro un conjunto de transacciones. Y los algoritmos son “Apriori”, “FP Growth” y “Eclat”.

Existen distintas plataformas, herramientas o lenguajes de programación destinados para facilitarnos el análisis y exploración de los datos, entre ellas están: “Konstanz Information Miner” o simplemente conocido como “Knime”, el lenguaje de programación “R” o la suite de minería de datos del lenguaje de programación Python. Además, existen otras herramientas que hacen de complemento a las utilizadas en minería de datos como, por ejemplo, “Talend Open Studio” que es una herramienta de ETL (Extracción, Transformación y Carga) que nos permiten realizar el tratamiento de los datos y “Power BI” que es una aplicación que nos apoya en la creación de visualizaciones interactivas como líneas de tiempo, gráficos, jerarquías, entre otras.

El proyecto se realiza para dos grandes temas, el primero es acerca de los sondeos de precios de mercados y supermercados y el segundo las atenciones a reclamaciones. En la planificación del proyecto se utilizó el marco de trabajo ágil “Scrum”, utilizando sus artefactos: Product Backlog, Story Mapping, Historias de Usuario, entre otros. Además, para diagramar los procesos se hace uso de la notación gráfica estandarizada “Modelo y Notación de Procesos de Negocio” o por sus siglas en inglés (BPMN), esta notación permite el modelado de procesos de negocios, en formato de flujo de trabajo. Para la generación de los diagramas de los ETL’s y Almacenes de datos se hace uso de un método global basado en UML, y otros diagramas como los “Paquetes de información”.

Para los sondeos de precios de los productos, se divide en 3 Sprint, en el primero se trabaja con respecto a los datos de los sondeos de precios del MAG (Ministerio de Agricultura y Ganadería). Se describen las historias de usuario, se diseñan los diagramas de procesos BPMN, el paquete de información, y se añade la explicación de cómo se diseñó el ETL desde archivos históricos en formato Excel que se tienen ahorita a un Data Warehouse, y la aplicación de modelos de minería en los programas de Knime, los lenguajes de programación R y Python, se diseñan en Power BI los reportes de BI (inteligencia de negocios) y de minería. El segundo Sprint se trabaja con los datos de los sondeos de precios de mercados, estos sondeos son realizados por la Dirección de Vigilancia De Mercado (DVM). Se describe las historias de usuario, se diseñan los diagramas de procesos BPMN, el paquete de información y se añade la explicación de cómo se diseñó el ETL, además, de cómo realizar la carga desde un archivo Excel a un Data Warehouse. En el tercer Sprint se toman los datos del sistema SPRS (Sistema de Recolección de Precios de Supermercados), de igual manera que en los Sprint anteriores, se diseñaron los diagramas de procesos, el paquete de información y su explicación de cómo se realiza el ETL.

Para las atenciones a reclamaciones se divide igualmente en tres Sprint, el primero es sobre la extracción, transformación y carga de los datos, primero se analiza la situación actual, determinando los orígenes de datos, se diseña el “Staging Area” que es la zona intermedia donde se unifican las tablas de los orígenes de datos, para luego realizar las transformaciones necesarias y crear las dimensiones y las tablas de hecho del Data Warehouse. En el segundo Sprint se realiza la minería de datos, se toman los datos del Datawarehouse y se generan los flujos de trabajo donde se aplican los diferentes algoritmos de minería de datos y así, satisfacer cada una de las historias de usuario. Finalmente, en el sprint tres, se realizan los informes de inteligencia de negocios y los informes de minería de datos, estos informes se realizan con la plataforma de Power Bi.

2 Objetivos

2.1 Objetivo General

Elaborar una solución de Minería de datos para la generación de conocimiento de la atención a reclamaciones y de los sondeos de precios mediante la identificación de patrones de comportamiento en los datos por medio de exploración y analítica.

2.2 Objetivos Específicos

- Identificar orígenes de datos de los sondeos de precios realizados por la Defensoría y trasladarlos a un área de stage.
- Diseñar y desarrollar procesos de extracción, transformación y carga de los datos de sondeos de precios del área de stage hacia un modelo multidimensional.
- Aplicar técnicas de minería descriptiva y predictiva en los datos de los sondeos de precios mediante diferentes algoritmos que permitan explorar los datos en búsqueda de patrones de comportamiento.
- Proporcionar una solución de visualización producto de la explorativa y analítica de los datos de sondeos de precios mediante informes de inteligencia de negocios que contribuya a la toma de decisiones.
- Identificar orígenes de datos de las atenciones a reclamaciones brindadas por la Defensoría y trasladarlos a un área de stage.
- Diseñar y desarrollar procesos de extracción, transformación y carga de los datos de las atenciones a reclamaciones del área de stage hacia un modelo multidimensional.
- Aplicar técnicas de minería descriptiva y predictiva en los datos de las atenciones a reclamaciones mediante diferentes algoritmos que permitan explorar los datos en búsqueda de patrones de comportamiento.
- Proporcionar una solución de visualización producto de la explorativa y analítica de los datos de las atenciones a reclamaciones mediante informes de inteligencia de negocios que contribuya a la toma de decisiones.

3 Alcances

- La solución será de uso exclusivo de la Defensoría del Consumidor.

- La solución se realizará en conjunto con la Unidad de Análisis de Consumo y Mercado, además se coordinará soporte con la Gerencia de Sistemas Informáticos.

- El proceso se realizará de manera iterativa, realizándose en dos iteraciones atendiendo en la primera iteración los datos que cuenta la UACM sobre los sondeos de precios y tomando en la segunda iteración los datos sobre atenciones y reclamaciones.

- La iteración 1 será desarrollada para 3 diferentes sondeos de precios, los cuales son: sondeos de precios del informe diario de precios de productos del Ministerio de Agricultura y Ganadería, los sondeos realizados en Mercados por la unidad de “Dirección de Vigilancia De Mercado” de la Defensoría del Consumidor y de los sondeos en supermercados con el sistema SUPERMERCADOS.

- La iteración 2 será desarrollada con la información del Sistema de Atenciones a Reclamaciones (SARA).

- Esta fase contara con 3 Sprint que se dividen en Construcción del modelo multidimensional, Minería de datos y visualización.

- Las dos Iteraciones contaran con las siguientes iteraciones Construcción de un Modelo Multidimensional, Exploración de Minería de Datos, Desarrollo de informes de Inteligencia de Negocio.

4 Limitaciones

- El poco presupuesto asignado a la Gerencia de Sistemas Informáticos para la compra de nuevos equipos y adquisiciones de licencias, que potencien la exploración de los datos.

5 Justificación

La Unidad de Análisis de Consumo y Mercados de la Defensoría del Consumidor actualmente ejecuta un análisis para los sondeos de precios de diferentes productos, estos análisis se efectúan únicamente para la realización de reportes y/o boletines que se entregan periódicamente a la presidencia de la institución u otras instituciones que los requieran, sin embargo, no se realiza un análisis profundo sobre los datos históricos con los que se poseen, la Unidad no cuenta con las herramientas TIC's necesarias que le permita estar preparada y actuar de manera anticipada ante cualquier eventualidad, es por ello, que con la realización de la solución se pretende emplear técnicas de minería de datos para generar conocimiento, descubrir patrones ocultos, además, se pretende realizar un análisis predictivo en las bases de precios, siempre con el propósito de que la información apoye el actuar oportunamente en caso que se estén violentando los derechos de los consumidores.

Para el caso del análisis de los datos de la atención a reclamaciones, no se cuentan con herramientas que permitan identificar es de especial interés para la Unidad, conocer predicciones de ofertas/demandas, altas y bajas en las atenciones, tipos de clientes, motivos de reclamos, ubicaciones geográficas de empresas reincidentes para ciertos tipos de reclamos, entre otros. Siendo esta información de suma importancia y de utilidad para la planificación de recursos, toma de decisiones y mejora de procesos.

La minería de datos se ha vuelto una parte importante y vital en las empresas o instituciones donde se emplea actualmente ya que está permite a las empresas o instituciones ser más competitivos en el mercado tomando decisiones más acertadas en base al conocimiento generado de los mismos datos con los que cuenta la institución, ya que el uso de minería de datos genera un valor agregado a la toma de decisiones.

6 Importancia

Actualmente toda institución ya sea pública o privada debe reconocer la importancia que existe en el uso de las Tecnologías de Información y Comunicaciones (TIC), ya que estas permiten identificar mejoras o ventajas competitivas en los procesos, procedimientos o servicios que brindan como institución, así poder lograr el apoyo de manera que se alcancen las metas y objetivos como institución.

La minería de datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de datos, siendo el objetivo general el extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. La minería de datos es fundamental para realizar un análisis exhaustivo que arroje información de interés en las organizaciones para la toma de decisiones, Además, la aplicación de técnicas de minería de datos no se realiza en un solo sector determinado, esta puede ser aplicada en la mayoría de sectores productivos o económicos, para comprender que la minería de datos puede brindar información útil y valiosa en los diferentes sectores.

La Defensoría del Consumidor es la única entidad autorizada y designada por Ley, para velar por los derechos de los consumidores, dicha institución posee como objetivo proteger y promover efectivamente los derechos de las personas consumidoras, a fin de procurar el equilibrio, certeza y seguridad jurídica en sus relaciones con los proveedores. El uso de técnicas de Minería de Datos seria de una gran utilidad, ya que la aplicación de estas técnicas puede generar un valor añadido no solo en organizaciones privadas, sino, también en organizaciones públicas, autónomas, y semi-autónomas.

Asimismo, se debe de tomar en cuenta que a medida que se han ido automatizando los procesos y procedimientos que se realizan en las diferentes instituciones o empresas, la cantidad de información generada y almacenada por estos sistemas transaccionales en bases de datos o archivos (archivos de texto, archivos en Excel, entre otros) ha ido en aumento, donde al aplicar procesos de extracción, transformación y carga, se puedan obtener unos datos limpios, y al aplicar técnicas de minería de datos, se pueda obtener conocimiento que permita tomar decisiones más acertadas.

7 Antecedentes

7.1 Antecedentes Generales.

La minería de datos ha venido a revolucionar la forma de trabajar con los datos y la información que se puede obtener de ella, teniendo un fuerte impacto en la toma de decisiones. Actualmente se puede encontrar en todo el mundo, desde empresas nacionales, empresas multinacionales, entidades de gobiernos, etc. Pero este término no es reciente, su historia data desde hace algunos años y la extracción de patrones a partir de los datos y el análisis de estos se ha producido durante mucho tiempo atrás.

La estadística ha sido la primera ciencia que históricamente extrae información de los datos básicamente mediante metodologías procedentes de las matemáticas. En el año **1763** se publica un artículo póstumo de Thomas Bayes en relación con un teorema denominado "Teorema de Bayes" que es utilizado para calcular la probabilidad de un suceso, teniendo información de antemano sobre ese suceso. Este teorema es fundamental para la minería de datos y la probabilidad, ya que nos permite la comprensión de realidades complejas basadas en probabilidades estimadas.

En **1805**, Adrien-Marie Legendre y Carl Friedrich Gauss desarrollaron y aplicaron la primera forma de regresión, que fue denominada "Método de Mínimos Cuadrados", que es un procedimiento de análisis numérico en la que, dados un conjunto de datos, se intenta determinar la función continua que mejor se aproxime a esos datos. Esta técnica se utilizó para determinar las órbitas de los cuerpos sobre el Sol. El análisis de regresión tiene como objetivo el estimar las relaciones entre las variables, y el método específico. Esta es una de las herramientas clave en la minería de datos. Ya en **1845** se acuño la frase *"Business Intelligence"* (Inteligencia de negocios) o por sus siglas en ingles *"BI"*, en la *"Cyclopædia of Commercial and Business Anecdotes"* (Enciclopedia de las anécdotas comerciales y de negocios) y se utilizó para describir como un banquero se beneficiaba de la información, al reunirla y actuar sobre ella antes de la competencia.

El comienzo de la era de la computadora que hace posible la recopilación y el procesamiento de grandes cantidades de datos, comienza en **1936**, con el artículo "*On Computable Numbers, with an Application to the Entscheidungsproblem"* (En Números computables, con una aplicación a los problemas de decisión). Donde Alam Turing presentó la idea de una Máquina Universal que era capaz de realizar cálculos como en los computadores de hoy en día. En **1943** (McCulloch & Pitts, 1943) publicaron un artículo denominado "*A logical calculus of the ideas immanent in nervous activity*" (Un cálculo lógico de las ideas inmanentes a la actividad nerviosa) crearon un modelo conceptual de una red neuronal, donde cada una de estas neuronas puede hacer 3 cosas: recibir entradas, procesar entradas y generar salidas.

Arthur Samuel en **1959** acuñó el término "Aprendizaje automático", mientras trabajaba para la compañía estadounidense IBM. El aprendizaje automático surgió de la búsqueda de inteligencia artificial para que las máquinas aprendieran de los datos, abordando el problema con varios métodos simbólicos, a los que luego denominaron "redes neuronales". En este mismo año se pública el primer artículo relacionado con los árboles de decisión, una técnica muy importante en la minería de datos y se tituló "*Matching and Prediction on the Principle of Biological*

Classification" (*Emparejamiento y predicción sobre el principio de clasificación biológica*) por el investigador británico William Belson,

Ya en los años de **1970**, se podía encontrar sistemas de administración de bases de datos, en el cual era posible almacenar y consultar datos. Donde ya los almacenes de datos permitían a los usuarios pasar de una forma de pensar orientada a las transacciones a una forma más analítica de ver los datos. Sin embargo, era muy limitado poder extraer una información sofisticada de los almacenes de datos de modelos multidimensionales.

En la década de **1980,** HNC Software una empresa estadounidense, tenía la marca registrada de "D*atabase mining*" (Minería de base de datos). Esta, estaba destinada a proteger un producto llamado "DataBase Mining Workstation" (Estación de Trabajo para Minería de Bases de Datos). Que era una herramienta que permitía crear modelos de redes neuronales, actualmente ya no se encuentra disponible. En estadística, ese mismo año, Gordon V. Kass introdujo un algoritmo recursivo de clasificación no binario llamado *"Chi-square Automatic Interaction Detection"* (Detección de Interacciones Automáticas Chi-cuadrado) por sus siglas en inglés *"CHAID",* y fue un acontecimiento muy importante para que, en el año **1984**, los investigadores Leo Breiman, Jerome Friedman, Richard Olshen y Charles Stone introdujeron un algoritmo para la construcción de árboles, y este lo aplicaron a problemas de regresión y clasificación. Casi simultáneamente el proceso de inducción mediante árboles de decisión comenzó a ser usado por la comunidad del Aprendizaje automático. En **1989** se acuño el término de "*Knowledge Discovery in Databases"* (Descubrimiento del conocimiento en bases de datos) por sus siglas en inglés "*KDD*" es acuñado por Gregory Piatetsky-Shapiro. Según (Han, Kamber, & Pei, 2012) este término se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. Ese mismo año se desarrolló el taller original KDD-89 y fue el inicio de una serie de talleres de KDD que luego se convirtieron en conferencias.

Ya en la década de **1990**, Se empezó a escuchar el término "Minería de Datos" en la comunidad de bases de datos. Tanto las empresas minoristas y la comunidad financiera estaban utilizando la minería para analizar los datos y reconocer las tendencias para aumentar la base de clientes. También les permitía predecir las fluctuaciones en las tasas de interés, precios en acciones y la demanda que podían tener de sus clientes. En esta época se dieron a conocer las máquinas de vectores de soporte que son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T. Y en **1992**, se sugirió una mejora a la maquina original que permitía la creación de clasificadores no lineales.

Ya en el **siglo XXI**, la minería de datos se encuentra muy bien posicionada en los negocios, en las ingenierías, en la ciencia, medicina, entre otros. El termino de Minería de Datos y Big Data ya son más comunes. Y todo gracias a que los costos de recopilación de datos y los equipos o sistemas para tratarlos se vuelven más baratos. Y por ende las principales técnicas de la minería de datos como, arboles de decisiones, clustering, entre otras, se vuelven más fáciles de aplicarlas a grandes cantidades de datos.

7.2 Antecedentes en El Salvador

La Minería de Datos es un tema que ha venido a revolucionar el panorama tecnológico mundial, permitiendo que las pequeñas, medianas y grandes empresas, así como carteras de estado y organizaciones que cuenten con grandes volúmenes de datos almacenados, permitan tomar decisiones más acertadas u obtener información valiosa.

Actualmente la minería de datos en El Salvador no se encuentra en un estado tan avanzado, existen muchas empresas o entidades de gobierno que no conocen estas técnicas, ni mucho menos a lo que se refiere el término, especialmente las PYMES y carteras de estado.

Al momento son pocas las compañías que aplican estas técnicas, especialmente las grandes corporaciones multinacionales como es el caso de empresas telefónicas como Millicom Tigo y Claro, entidades financieras como Banco Agrícola y Banco DAVIVIENDA, y otras empresas como AVIANCA y CROWLEY, dichas entidades han observado su potencial en empresas de otros países donde el tema si está muy desarrollado, y al no aplicarlo parten de una posición no muy ventajosa con respecto a ellos.

En las carteras de estado, existen pocas o ninguna entidad que haga uso de estas técnicas, es por ello que se ha visto una importante pérdida en su credibilidad al momento de tomar decisiones importantes. Esto se puede deber en primer lugar por la forma arbitraria en que se toma, ya que en la mayoría de los casos no se cuenta con información útil, datos estadísticos o proyecciones obtenidas basándose en datos pasados con las que se pueda contrarrestar o justificar el porqué de la decisión tomada.

Además, actualmente el país no cuenta con la suficiente oferta de personas especializadas en esta área, ya sea por la falta de carreras, postgrados o maestrías en las universidades o el costo que tienen las pocas opciones que se encuentran en el mercado, es por ello que las grandes empresas tratan de acaparar a los especialistas, o especializar a sus trabajadores.

7.3 Línea de tiempo.

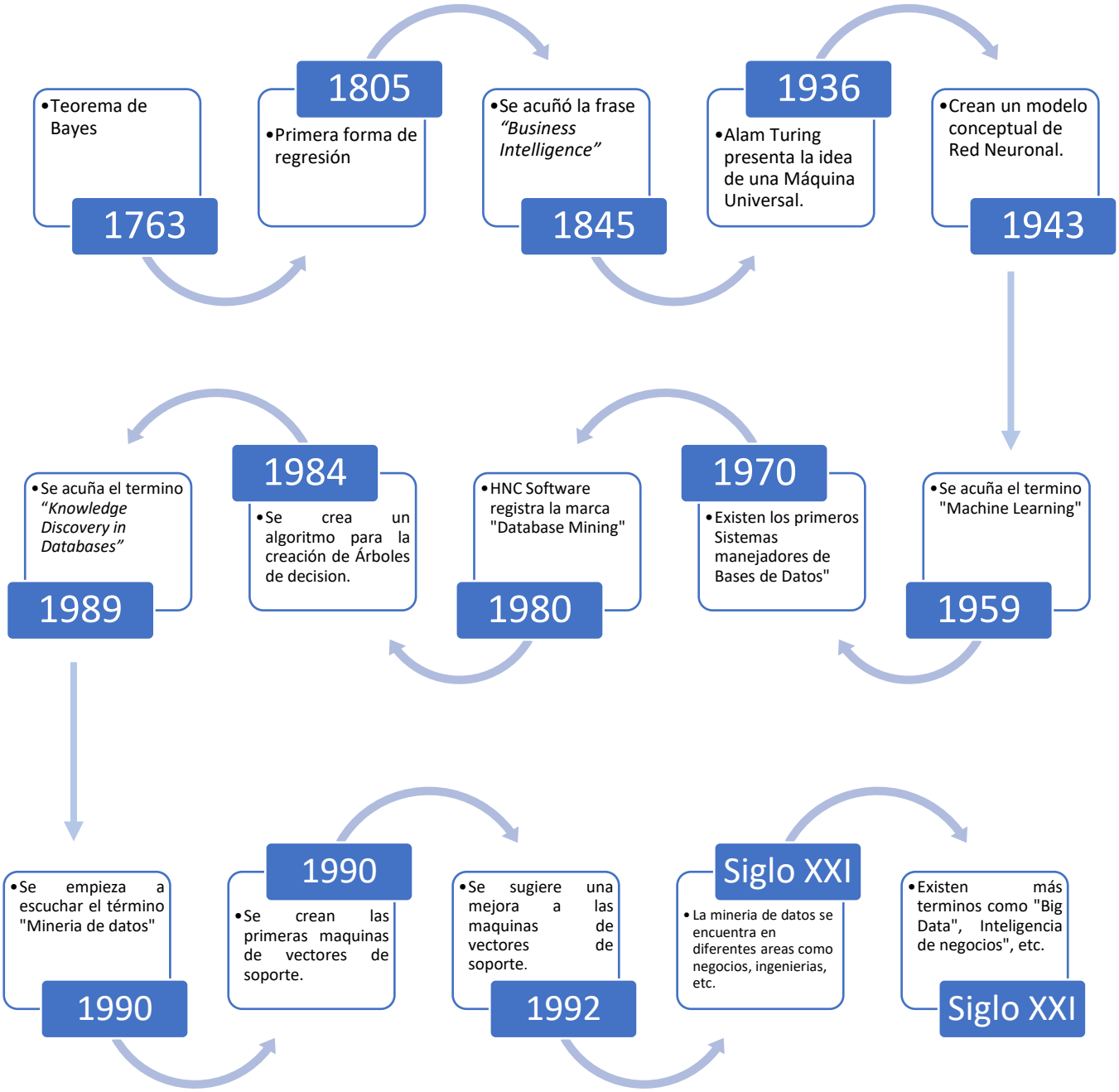


Figura 1 Línea de tiempo.

8 Marco Teórico

La minería de datos o a veces llamada también como descubrimiento de datos o de conocimiento, es un campo de la estadística y de la ciencia de la computación, la cual se refiere al proceso de preparar, sondear, explorar y analizar grandes volúmenes de datos desde diferentes perspectivas con el objetivo de encontrar información útil y/o patrones de comportamiento que se encuentren ocultos de manera automática o semi automática.

La minería de datos surge para intentar ayudar a comprender el contenido de grandes volúmenes de datos. De forma general, los datos son la materia prima y cuando se le atribuye algún significado se convierte en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.

8.1 Características y objetivos de la minería de datos

La minería de datos se caracteriza por:

- Explorar los datos que se encuentran en las profundidades de las bases de datos (por ejemplo, los Almacenes de Datos), que algunas veces contienen información almacenada durante varios años.
- En algunos casos, los datos se consolidan en un data warehouse o en data marts y se mantienen en servidores de Internet e Intranet.
- El entorno de la minería de datos suele tener una Arquitectura Cliente Servidor.
- Las herramientas de la minería de datos ayudan a extraer el mineral de la información registrado en archivos corporativos o en registros públicos, archivados.
- El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias, para efectuar preguntas ad-hoc y obtener rápidamente respuestas.
- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos. (EcuRed, s.f.)

8.2 Aplicaciones de la minería de datos

Entre las aplicaciones más populares de la minería de datos están:

I. **Análisis de datos financieros**: se emplea tanto en el sector bancario como en el de las finanzas. Se busca proveer datos que aseguren que es posible practicar análisis sistemáticos en condiciones avanzadas y con garantías de fiabilidad. Algunos ejemplos son:
    - Diseño y construcción de almacenes de datos para el análisis multidimensional de datos.
    - Predicción de pago de préstamos y análisis de políticas de crédito de cliente.
    - Clasificación y el agrupamiento de los clientes para la creación de ofertas personalizadas.

- Detección de blanqueamiento de dinero y otros delitos financieros.

II. **Industria minorista:** el sector de las ventas al por menor recoge grandes cantidades de datos provenientes de las ventas, el historial de compra de los clientes o el transporte de mercancías. La cantidad de datos recogidos continúa expandiéndose rápidamente debido al aumento de la facilidad, disponibilidad y popularidad de la web y las transacciones online. La minería de datos con sus aplicaciones para la industria minorista ayuda a identificar patrones de compra de los clientes y tendencias. De esta forma, las empresas están en condiciones de proporcionar una mejor calidad de servicio al cliente, aumentando su satisfacción y facilitando su retención. Entre estas aplicaciones destacan las que permiten:
    - El análisis multidimensional de las ventas, los clientes, los productos, el tiempo y la región.
    - Los análisis de la eficacia de las campañas de ventas.
    - La recomendación personalizada de productos.
    - Las referencias cruzadas de artículos.

III. **Industria de las telecomunicaciones:** en este sector, los datos son especialmente importantes para alcanzar una buena comprensión del negocio. La minería de datos y aplicaciones específicamente diseñadas para esta área, ayudan en la identificación de los patrones de telecomunicaciones, facilitan la detección de actividades fraudulentas y posibilitan el hacer un mejor uso de los recursos, mejorando la calidad del servicio. Entre las más ventajosas están:
    - Análisis multidimensional de datos de telecomunicaciones.
    - Análisis de patrones fraudulentos.
    - Identificación de patrones inusuales, hábitos y tendencias.
    - Asociación multidimensional y análisis de patrones secuenciales.

IV. **Análisis de datos biológicos**: el campo de la biología es uno de los más beneficiados por los avances de la tecnología. La genómica, la proteómica, la genómica funcional y la minería de datos aplicada a la investigación de los seres vivos son sólo algunos ejemplos, una lista donde no hay que olvidarse de la bioinformática. La minería de datos con sus aplicaciones aporta una contribución importante para el análisis de datos biológicos:
    - Integración semántica de las bases de datos genómicos y proteómicos heterogéneos distribuidos
    - Alineamiento, indexación, búsqueda de similitudes y análisis comparativo de múltiples secuencias de nucleótidos.
    - Descubrimiento de patrones y análisis de redes genéticas.
    - Identificación de patrones de proteínas estructurales.

(Logicalis, 2014)

8.3 Clasificación de los algoritmos de minería

Los algoritmos o las técnicas de minería se pueden clasificar en dos tipos según el objetivo del análisis de los datos, estos son:

8.3.1 Algoritmos o técnicas supervisadas

Los algoritmos supervisados también conocidos como algoritmos predictivos. Predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos

descriptivos). A partir de datos cuya etiqueta se conoce, se induce una relación entre dichas etiquetas y serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado.

Entre los algoritmos o técnicas que pertenecen a esta categoría tenemos:

- Regresión.
- Clasificación.
- Series temporales.

8.3.2 Algoritmos o técnicas no supervisadas

Los algoritmos no supervisados descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

- Clusterización.
- Asociación.

8.4 Técnicas de exploración de la información

8.4.1 Regresión lineal

La regresión lineal se puede definir como el modelo de la relación entre dos variables que se representa con una ecuación lineal. Donde las dos variables se consideran una explicativa y la otra dependiente. Se debe de determinar que las dos variables estén asociadas, donde se obtiene el coeficiente de correlación donde los valores van de -1 a 1 dependiendo de la fuerza de correlación.

$$Y = a + bX$$

Ecuación 1 Ecuación de la recta.

Como bien se conoce, la regresión lineal es representada por la Ecuación 1 donde *Y* es la variable dependiente y X es la variable explicativa o independiente. La pendiente de la recta es *b*, y *a* es la intersección.

8.4.1.1 Ventajas y Desventajas

Entre las ventajas tenemos:

- Distingue entre los roles de las diferentes variables. Los coeficientes pueden interpretarse en términos de los factores determinantes.
- Se trata de una herramienta útil para estudiar e identificar las posibles relaciones entre los cambios observados en dos conjuntos diferentes de variables.
- Suministra datos para confirmar hipótesis acerca de si dos variables están relacionadas.
- Proporciona un medio visual para probar la fuerza de una posible relación y agilizar la toma de decisiones.
- Desarrollo de proyectos para la búsqueda de mejoras de la calidad.

Existen diferentes tipos de métodos de regresión lineal aplicados a diferentes situaciones, pero entre los que más se destacan están la regresión lineal simple y la múltiple.

Regresión lineal simple

Es un modelo que debe de cumplir $y_i = a + bx_i + u_i$ donde:

- y_i representa el valor de la variable respuesta para la observación i-ésima.
- x_i representa el valor de la variable explicativa para la observación i-ésima.
- u_i representa el error para la observación i-ésima que se asume normal. $u_i \sim N(0, \sigma)$.

Y *a* y *b* ya son conocidas como la intersección y la pendiente respectivamente. El objetivo es calcular una recta que se ajuste lo mejor posible a los datos, la cual quedaría de la siguiente manera:

$$\hat{y} = \hat{a} + \hat{b}x$$

Ecuación 2 Ecuación de la recta de regresión.

El modelo de regresión lineal simple tiene un residuo llamado así porque es una diferencia que se calcula entre cada valor y_i de la variable respuesta y su estimación $\hat{y}_i$ es representado por e_i.

La regresión lineal simple tiene 5 hipótesis las cuales son:

Linealidad: Los datos deben ser razonablemente rectos, por lo que la relación de *x* e *y* debe ser una recta. (Cascos)

Homogeneidad: El valor promedio del error es cero. $E[u_i] = 0$. (Cascos)

Homocedasticidad: Los datos deben de ser constantes en su dispersión, a esto se le conoce como homocedásticos ver Figura 2, donde el valor promedio del error es cero, si esto no se cumple los datos son heterocedásticos.

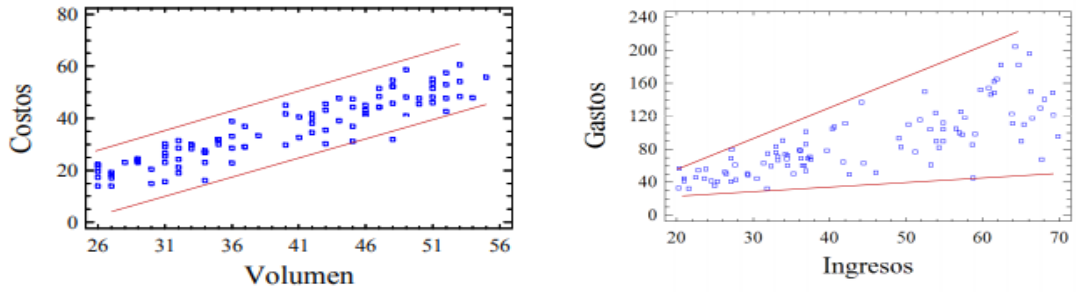


Figura 2 Grafico de costos vs volumen donde los datos son homocedásticos y heterocedásticos respectivamente.

(Cascos)

Independencia:

- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás.
- Habitualmente, se sabe por el tipo de datos si son adecuados o no para el análisis.
- En general, las series temporales no cumplen la hipótesis de independencia.

(Cascos)

Normalidad:

Los errores u_i se asumen que siguen una distribución normal.

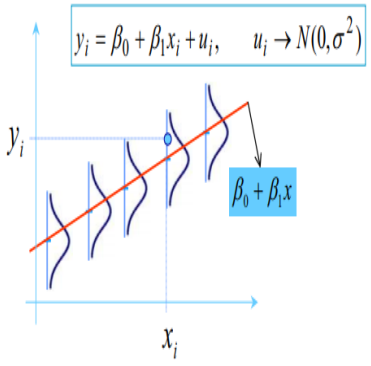


Figura 3 El error en los datos son normales a priori.

(Cascos)

Estimar y predecir:

Estos términos, aunque son similares y se obtienen de igual forma su precisión representada matemáticamente varia para lo cual tomaremos los siguientes conceptos.

- Estimar es obtener el valor medio de la variable Y para cierto valor X = x_0.
- Predecir es conocer el valor que tomará la variable Y para cierto valor X = x_0.

El intervalo de confianza para la respuesta promedio es:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\sqrt{s_R^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)}$$

Ecuación 3 Confianza para al estimar un $\hat{y}_0$.

El intervalo de confianza para la predicción de una nueva respuesta es:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\sqrt{s_R^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)}$$

Ecuación 4 Confianza para estimar un $\hat{y}_0$.

Cabe recalcar que la longitud de este intervalo es mayor que la del anterior (menos precisión) porque no corresponde a un valor medio sino a uno específico.

Regresión lineal múltiple

El Análisis de Regresión Lineal Múltiple permite establecer relaciones que se producen entre una variable dependiente *Y* e un conjunto de variables independientes (X_1, X_2, ... X_K). Este análisis se aproxima más a lo que sucede en la realidad fenómenos, hechos y procesos sociales, los cuales deben ser explicados por una serie de variables que afectan directa o indirectamente en su precisión.

En el análisis de regresión múltiple lo más frecuente es que tanto la variable dependiente como las independientes sean variables continúas medidas en escala de intervalo. La nota matemática de este modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Ecuación 5 Ecuación de regresión lineal múltiple.

En este modelo se asumen la mayoría de las hipótesis planteadas para el análisis de regresión lineal simple como son:

- Linealidad
- Homogeneidad
- Homocedasticidad
- Independencia
- Normalidad

Y se añaden:

- *n > k+1*. El número de datos debe de ser mayor a *k+1*.
- Colinealidad: Ninguna variable explicativa es combinación lineal de las demás (las x_i son linealmente independientes). Además, esta hipótesis maneja el evento de que si una x_i es exactamente igual a otra puede manejarse el modelo con menos variables. También hay que considerar si alguna de las x_i está fuertemente correlacionada con otras.

Ahora se verán las hipótesis de este modelo en su representación matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix}$$

Ecuación 6 Forma matricial del análisis de regresión lineal múltiple.

Debido al principio de colinealidad y al interpretar esta matriz como una nube de puntos en el espacio nos daremos cuenta que trabajaremos con *k+1* dimensiones, por lo tanto, si *k=2* será la representación de un plano en el espacio.

Se puede resumir la solución a la matriz para conocer β_n forma matricial:

$$\hat{\beta} = [x'x]^{-1}x'y$$

Ecuación 7 Resultado de forma matricial de MCO.

Para contrastar con la hipótesis si una variable x_i influye en la variable y_i se usará el estadístico teórico t – Student y la forma matricial $[x'x]^{-1} = q_{i+1,i+1}$. En la cual se establece la hipótesis nula $H_0: \beta_i = 0$ y la alternativa $H_1: \beta_i \neq 0$ donde el estadístico observado seria.

$$t = \frac{\hat{\beta}_i + \beta_i}{s_R\sqrt{q_{i+1,i+1}}} \; cuando\ es\ la\ H_0\ la\ ecuacion\ sera\ t = \frac{\hat{\beta}_i}{s_R\sqrt{q_{i+1,i+1}}}$$

Ecuación 8 Contraste de la hipótesis nula.

Confianza en estimar y predecir

Al igual que en la regresión lineal simple teníamos un nivel de confianza para las predicciones o estimaciones que se realizan a la variable dependiente, su cálculo se realiza mediante las siguientes expresiones:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},(n-k-1)}\, s_R \sqrt{(1\ x_{10}\ x_{20}\ \cdots\ x_{k0})[x'x]^{-1}\begin{bmatrix}1\\x_{10}\\x_{20}\\\vdots\\x_{k0}\end{bmatrix}}$$

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},(n-k-1)}\, s_R \sqrt{1+(1\ x_{10}\ x_{20}\ \cdots\ x_{k0})[x'x]^{-1}\begin{bmatrix}1\\x_{10}\\x_{20}\\\vdots\\x_{k0}\end{bmatrix}}$$

Ecuación 9 Confianza para la estimación y predicción respectivamente.

(de la Fuente Fernández, 2011)

Ejemplo

Utilizaremos el dataset de Iris flower o Fisher's Iris el cual es un conjunto de datos multivariados introducido por el estadístico y biólogo británico Ronald Fisher en su artículo de 1936, para poder realizar una regresión lineal en KNIME. Usaremos el simple regression tree para la regresión en la plataforma.

Se divide el conjunto de datos de iris en entrenamiento y pruebas. Se hará uso del algoritmo CART este algoritmo intenta encontrar la mejor dirección para los valores perdidos enviándolos en cada dirección y seleccionando el que produce el mejor resultado. Se coloca un nodo final para calcular ciertas estadísticas.

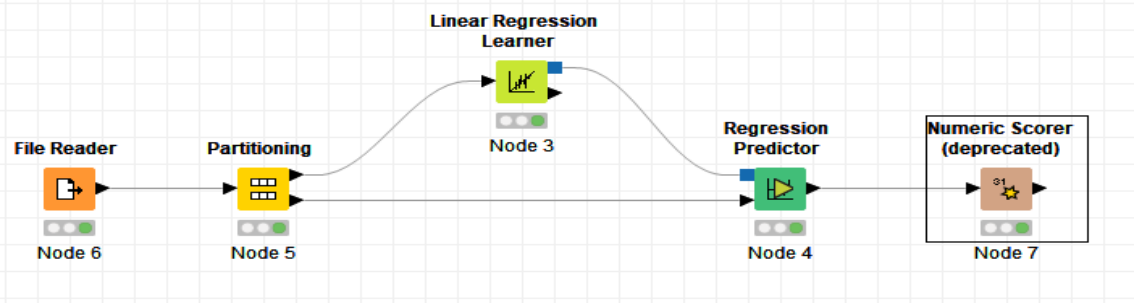
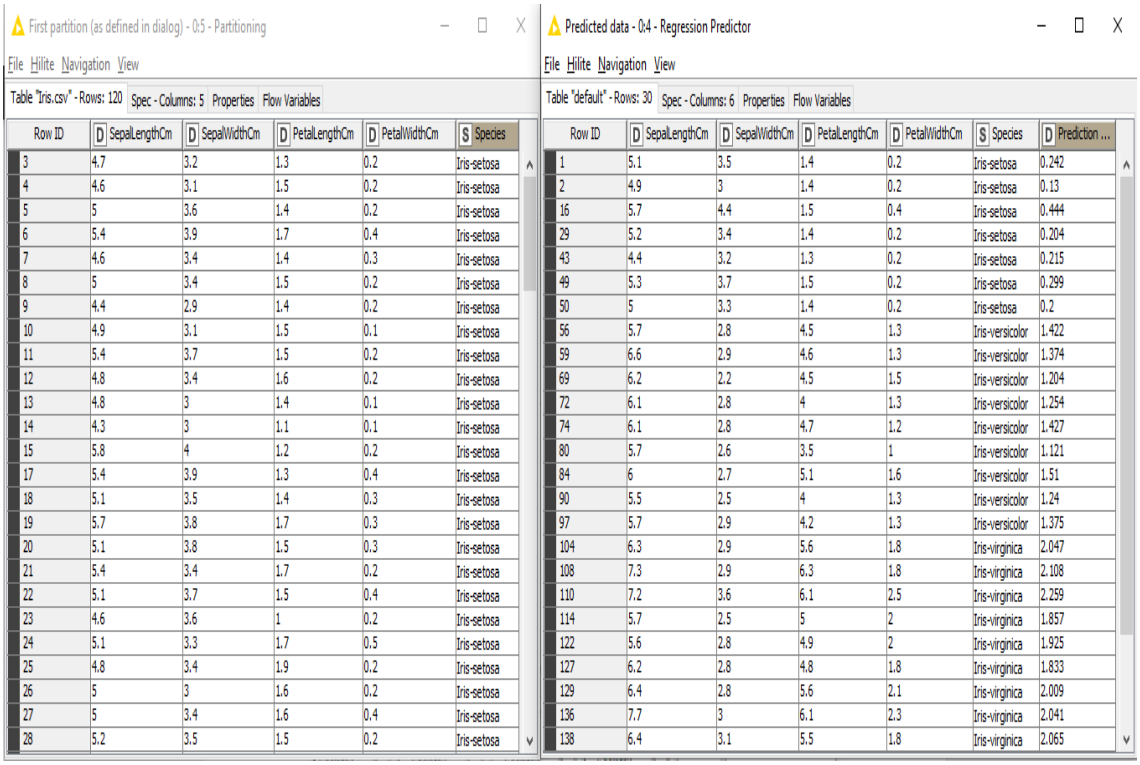


Figura 4 Nodo Numeric Scorer para obtener cálculos estadísticos.

Por último, podemos ver la predicción la cual es el cuadro del lado derecho de la Figura 5.



First partition (as defined in dialog) - 0:5 - Partitioning

File Hilite Navigation View

Table "Iris.csv" - Rows: 120 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa
25	4.8	3.4	1.9	0.2	Iris-setosa
26	5	3	1.6	0.2	Iris-setosa
27	5	3.4	1.6	0.4	Iris-setosa
28	5.2	3.5	1.5	0.2	Iris-setosa

Predicted data - 0:4 - Regression Predictor

File Hilite Navigation View

Table "default" - Rows: 30 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Prediction ...
1	5.1	3.5	1.4	0.2	Iris-setosa	0.242
2	4.9	3	1.4	0.2	Iris-setosa	0.13
16	5.7	4.4	1.5	0.4	Iris-setosa	0.444
29	5.2	3.4	1.4	0.2	Iris-setosa	0.204
43	4.4	3.2	1.3	0.2	Iris-setosa	0.215
49	5.3	3.7	1.5	0.2	Iris-setosa	0.299
50	5	3.3	1.4	0.2	Iris-setosa	0.2
56	5.7	2.8	4.5	1.3	Iris-versicolor	1.422
59	6.6	2.9	4.6	1.3	Iris-versicolor	1.374
69	6.2	2.2	4.5	1.5	Iris-versicolor	1.204
72	6.1	2.8	4	1.3	Iris-versicolor	1.254
74	6.1	2.8	4.7	1.2	Iris-versicolor	1.427
80	5.7	2.6	3.5	1	Iris-versicolor	1.121
84	6	2.7	5.1	1.6	Iris-versicolor	1.51
90	5.5	2.5	4	1.3	Iris-versicolor	1.24
97	5.7	2.9	4.2	1.3	Iris-versicolor	1.375
104	6.3	2.9	5.6	1.8	Iris-virginica	2.047
108	7.3	2.9	6.3	1.8	Iris-virginica	2.108
110	7.2	3.6	6.1	2.5	Iris-virginica	2.259
114	5.7	2.5	5	2	Iris-virginica	1.857
122	5.6	2.8	4.9	2	Iris-virginica	1.925
127	6.2	2.8	4.8	1.8	Iris-virginica	1.833
129	6.4	2.8	5.6	2.1	Iris-virginica	2.009
136	7.7	3	6.1	2.3	Iris-virginica	2.041
138	6.4	3.1	5.5	1.8	Iris-virginica	2.065

Figura 5 Predicciones lineal de la data set Iris.

8.4.2 Series Temporales

Se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados.

El estudio descriptivo de series temporales se basa en la idea de descomponer la variación de una serie en varias componentes básicas. Este enfoque no siempre resulta ser el más adecuado, pero es interesante cuando en la serie se observa cierta tendencia o cierta periodicidad.

Las componentes o fuentes de variación que se consideran habitualmente son las siguientes:

1. **Tendencia:** Se puede definir como un cambio a largo plazo que se produce en relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.
2. **Efecto Estacional:** Muchas series temporales presentan cierta periodicidad o, dicho de otro modo, variación de cierto periodo (anual, trimestral, mensual). Por ejemplo, el paro laboral aumenta en general en invierno y disminuye en verano. Estos tipos de efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos, desestacionalizando la serie original.
3. **Componente Aleatoria:** Una vez identificados los componentes anteriores y después de haberlos eliminado, persisten unos valores que son aleatorios. Se pretende estudiar qué tipo de comportamiento aleatorio presentan estos residuos, utilizando algún tipo de modelo probabilístico que los describa.

De las tres componentes reseñadas, las dos primeras son componentes determinísticas, mientras que la última es aleatoria. Así, se puede denotar que:

$$X_t = \mathrm{T}_t + E_t + I_t$$

Ecuación 10 Serie temporal.

Donde T_t es la tendencia, E_t es la componente estacional, que constituyen la señal o parte determinística, e I_t es el ruido o parte aleatoria. El aislamiento de la componente aleatoria se suele abordar de dos maneras.

1. **Enfoque descriptivo:** Se estima T_t y E_t y se obtiene I_t como:

$$I_t = \mathrm{X}_t - T_t - E_t$$

Ecuación 11 Expresión enfoque descriptivo.

2. **Enfoque de Box-Jenkins:** Se elimina de X_t la tendencia y la parte estacional (mediante transformaciones o filtros) y queda sólo la parte probabilística. A esta última parte se le ajustan modelos paramétricos.

Clasificación descriptiva

1. **Estacionarias**: Una serie es estacionaria cuando es estable, es decir, cuando la media y la variabilidad son constantes a lo largo del tiempo. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo. Es una serie básicamente estable a lo largo del tiempo.
2. **No Estacionarias:** Son series en las cuales la media y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante.

Estimación de la tendencia

Para estimar la tendencia supondremos que tenemos una serie no estacionaria sin componente estacional, es decir, que la serie se puede descomponer en:

$$X_t = \mathrm{T}_t + I_t$$

Ecuación 12 Tendencia para una serie de tiempo no estacionaria.

Tendencia determinista

En este caso supondremos que la tendencia es una función determinística. La función más sencilla posible es una recta.

Tendencia evolutiva (medias móviles)

Se supone que la tendencia es una función que evoluciona lentamente y que puede aproximarse en intervalos muy cortos (por ejemplo, de 3 o 5 datos) por una función simple del tiempo. En general se supone una recta, pero ahora sus coeficientes van cambiando suavemente en el tiempo.

Ejemplos

Antes de comenzar aclararemos el set de datos son unos valores de horas agrupados utilizados en ejemplos de KNIME, y tiene como objetivo seleccionar un cluster o agrupamiento.

Se divide en dos particiones la tabla una para entrenar y otra para probar. Una se evalua con la expresión x(t), x(t-1), x(t-2), ..., x(t-lag). La opción de retraso L en este nodo es útil para la predicción de series de tiempo. Se generan las Regresiones Lineal y Polinomial.

Por último, generaremos las gráficas de regresión lineal que son dos cargados uno para cada tipo de regresión.

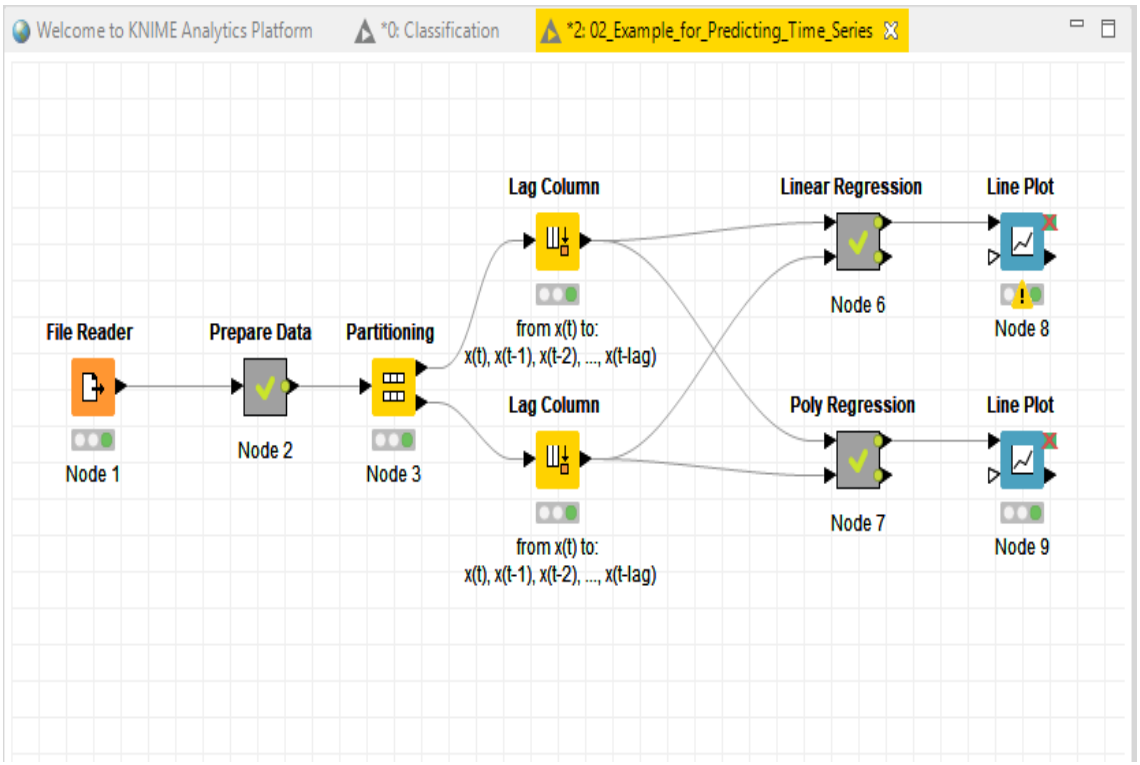


Figura 6 Nodos Line Plot.

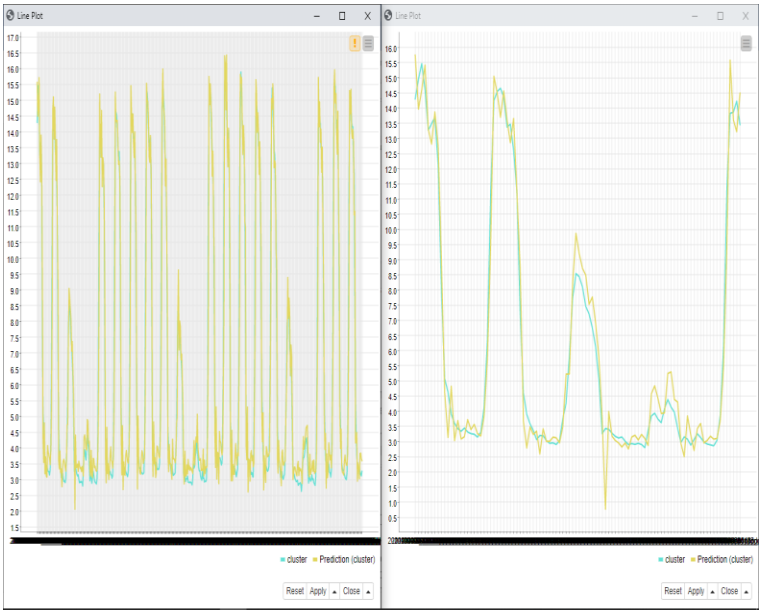


Figura 7 Resultado de las estimaciones para la Series temporales.

En la Figura 7 se puede observar el grafico derecho la regresión Lineal y el izquierdo la Polinomial.

8.4.3 Clasificación

La clasificación es una forma de análisis de datos que extrae modelos que describen clases de datos importantes. Dichos modelos, llamados clasificadores, predicen etiquetas de clase categóricas (discretas, no ordenadas).

La clasificación de datos es un proceso de dos pasos, que consiste en un paso de aprendizaje (donde se construye un modelo de clasificación) y un paso de clasificación (donde el modelo se utiliza para predecir las etiquetas de clase para los datos).

8.4.3.1 Algoritmo de Naive Bayes

Es un Algoritmo de Clasificación basado en el *Teorema de Bayes* con una suposición entre los predictores. Naive Bayes es fácil construir y es útil para una cantidad de datos muy grande.

El Algoritmo de clasificación de Naive Bayes asume que el efecto de una característica en particular en una clase es independiente de otras características. Por ejemplo, un solicitante de préstamo es deseable o no dependiendo de sus ingresos, historial de préstamos y transacciones anteriores, edad y ubicación. Incluso si estas características son interdependientes, estas características se consideran de forma independiente.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Ecuación 13 Teorema de Bayes.

Donde:

$P(h)$: Es la probabilidad de que la hipótesis h sea cierta (independientemente de los datos). Esto se conoce como la probabilidad previa de h.

$P(D)$: Probabilidad de los datos (independientemente de la hipótesis). Esto se conoce como probabilidad previa.

$P(h|D)$: Es la probabilidad de la hipótesis h dada los datos D. Esto se conoce como la probabilidad posterior.

$P(D|h)$: Es la probabilidad de los datos d dado que la hipótesis h era cierta. Esto se conoce como probabilidad posterior.

Ventajas:

- Es fácil y rápido predecir la clase de conjunto de datos de prueba. También funciona bien en la predicción multiclase.
- Cuando se mantiene la suposición de independencia, El algoritmo de Naive Bayes funciona mejor en comparación con otros modelos como la Regresión Logística y se necesitan menos datos de entrenamiento.
- Funciona bien en el caso de variables de entrada categóricas comparada con variables numéricas.

Desventajas:

- Si la variable categórica tiene una categoría en el conjunto de datos de prueba, que no se observó en el conjunto de datos de entrenamiento, el modelo asignará una probabilidad de 0 y no podrá hacer una predicción. Esto se conoce a menudo como frecuencia cero. Para resolver esto, podemos utilizar la técnica de alisamiento.
- Otra limitación de Naive Bayes es la asunción de predictores independientes. En la vida real, es casi imposible que obtengamos un conjunto de predictores que sean completamente independientes.

8.4.3.2 K-NN o Nearest Neighbours (Vecinos más cercanos)

El método K-NN emplea la clasificación supervisada estimando la distancia de cierto número de muestras (K vecinos) a la muestra que se pretende clasificar, determinando su pertenencia a la clase de la que encuentre más vecinos etiquetados, considerando el criterio de mínima distancia. Esta técnica es válida solo para datos numéricos, no para clasificadores de textos.

Como funciona:

1. Calcular la distancia entre el ítem a clasificar y el resto de ítems del dataset de entrenamiento.
2. Seleccionar los «k» elementos más cercanos (con menor distancia, según la función que se use).
3. Realizar una «votación de mayoría» entre los k puntos: los de una clase/etiqueta que <<dominen>> decidirán su clasificación final.

8.4.3.3 Árboles de decisiones

Los árboles de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas. Esta técnica permite que personas u organizaciones comparen posibles acciones. Estos árboles se utilizan como técnica de la minería de datos, donde nos permite preparar, sondear y se explorar datos para obtener información oculta de ellos. Esta técnica se aplica en forma de algoritmos a conjuntos de datos. Los árboles permiten tener soluciones a problemas de predicciones, clasificaciones o segmentaciones.

Esta técnica permite:

- Segmentación: establecer que grupos son importantes para clasificar un cierto ítem.
- Clasificación: asignar ítems a uno de los grupos en que está particionada una población.
- Predicción: establecer reglas para hacer predicciones de ciertos eventos.
- Reducción de la dimensión de los datos: Identificar que datos son los importantes para hacer modelos de un fenómeno.
- Identificación-interrelación: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
- Recodificación: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante

(Bouza, 2014)

La mayoría de los algoritmos utilizados para construir un árbol son variaciones de uno genérico llamado “Greedy algorithm” que básicamente va desde la raíz hacia abajo (Top-Down) buscando de manera recursiva los atributos que generan el mejor árbol hasta encontrar el óptimo global con una estructura de árbol lo más simple posible. Los algoritmos más conocidos son:

- **ID3:**

Este algoritmo fue propuesto por J Ross Quinlan 1975 en su libro “Machine learning” vol. 1. Básicamente ID3 construye un árbol de decisión (DT) desde un set fijo de “ejemplos”, el DT generado se usa para clasificar futuros ejemplos. Cada ejemplo tiene varios atributos que pertenecen a una clase (como los valores sí o no). Los nodos de “hoja” del árbol (leaf nodes) contienen el nombre de la clase, mientras que los nodos “no-hoja” son los nodos de decisión donde cada uno de ellos (cada rama) corresponde un posible valor del atributo. Cada nodo de decisión es una prueba del atributo con otro árbol que comienza a partir de él.

- **C4.5:**

Este algoritmo fue desarrollado por Ross Quinlan en 1993 [71] y básicamente es una versión avanzada del algoritmo ID3.

Este algoritmo se incluyen las siguientes capacidades o ventajas:

a) Manejo de valores continuos y discretos.
b) Tiene la capacidad de manejar valores de atributos faltantes.
c) Es capaz de generar un set de reglas que son mucho más fáciles de interpretar para cualquier tipo de árbol.
d) Este algoritmo construye un gran árbol y lo concluye con una “poda” de las ramas de manera de simplificarlo de manera de generar resultados más fáciles de entender y haciéndolo menos dependiente de la data de prueba.

- **C5.0:**

Al igual que sus predecesores, este algoritmo construye los árboles en base a un conjunto de datos de entrenamiento optimizado bajo el criterio de ganancia de información y corresponde a una evolución de su versión anterior, el algoritmo C4.5. Las mayores ventajas de esta versión tienen que ver con la eficiencia en el tiempo de construcción de árbol, el uso de memoria y la obtención de árboles considerablemente más pequeños que en el C4.5 con la misma capacidad predictiva.

- **CART (Classification And Regression Trees):**

Los árboles de clasificación y regresión (CART) fueron desarrollados en los años ochenta por Breiman, Freidman, Olshen y Stone en el libro Classification and regression trees (Breiman et al. 1984). La metodología CART utiliza datos históricos para construir árboles de clasificación o de regresión, los cuales son usados para clasificar o predecir nuevos datos. Estos árboles CART pueden manipular fácilmente como variable respuesta variables numéricas y categóricas.

(Berrios, 2014)

Ejemplo

Para este ejemplo se utilizará como data set los datos meteorológicos diarios. Manejaremos los valores faltantes encontrados en las celdas de la tabla de entrada. Definiremos columnas y para

cada una de ellas se definen intervalos, conocidos como bins. Cada uno de estos contenedores tiene un nombre único, un rango definido y bordes de intervalo abiertos o cerrados. Cada columna se reemplaza por la columna de tipo cadena agrupada o se agrega una nueva columna de tipo cadena agrupada.

Se obtendrán valores estadísticos como mínimo, máximo, media, desviación estándar, varianza, mediana, suma total, número de valores faltantes y recuento de filas en todas las columnas numéricas, y cuenta todos los valores nominales junto con sus ocurrencias.

Particionamos la tabla de entrada y se divide en dos particiones entrenar y probar datos. Se induce un árbol de decisión de clasificación en la memoria principal. El atributo de destino debe ser nominal. Los otros atributos utilizados para la toma de decisiones pueden ser nominales o numéricos. Por último, se predice el valor de la clase para nuevos patrones.

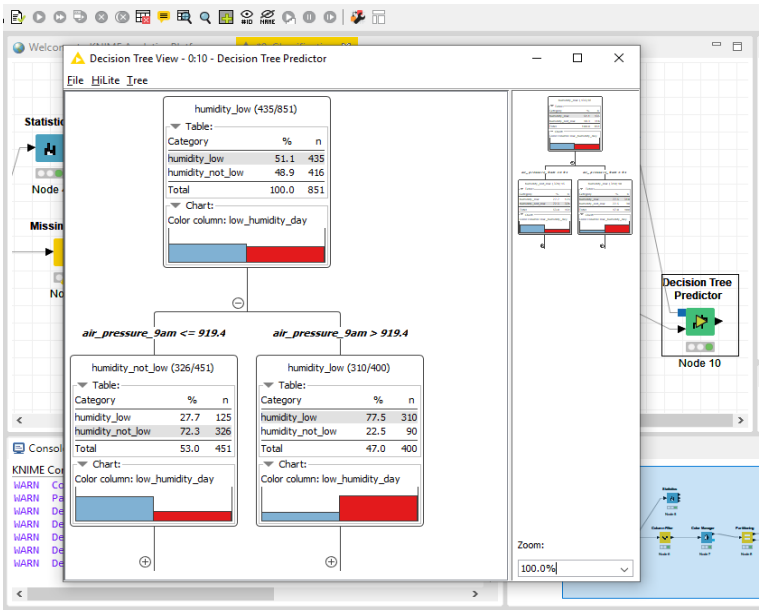


Figura 8 Nodo Decision Tree Predictor.

8.4.4 Reglas de Asociación

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta. En este contexto, el término transacción hace referencia a cada grupo de eventos que están asociados de alguna forma, por ejemplo:

- La cesta de la compra en un supermercado.
- Los libros que compra un cliente en una librería.
- Las páginas web visitadas por un usuario.
- Las características que aparecen de forma conjunta.

(Amat Rodrigo, 2018)

A cada uno de los eventos o elementos que forman parte de una transacción se le conoce como ítem y a un conjunto de ellos itemsets. Una transacción puede estar formada por uno o varios ítems, en el caso de ser varios, cada posible subconjunto de ellos es un itemsets distinto. Por

ejemplo, la transacción T = {A, B, C} está formada por 3 ítems (A, B y C) y sus posibles itemsets son: {A, B, C}, {A,B}, {B,C}, {A,C}, {A}, {B} y {C}.

Esta técnica de minería de datos hace uso de tres conceptos muy importantes para el desarrollo de la misma, así como el de poder realizar el análisis correspondiente a los resultados de la misma.

Soporte

El valor de soporte de X con respecto al conjunto de transacciones T está dado por el radio del número de transacciones que contienen el conjunto de elementos de X. Una regla con bajo soporte nos indica que esta pudo haber aparecido por casualidad.

$$soporte(X) = \frac{(NumTransacciones \subseteq X)}{(NumTotalTransacciones)}$$

Ecuación 14 Soporte.

Confianza

El valor de confianza está definido por la proporción de transacciones que contienen "∪" con respecto al número de transacciones que contienen X. Definida como la probabilidad de exista una relación entre antecedente y consecuente.

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

Ecuación 15 Confianza.

Lift

Se define como el radio del soporte observado a lo esperado si X y Y fuesen independientes. En caso de tener un valor de 1 quiere decir que los subconjuntos X y Y son independientes, mientras que un valor mayor indica que están positivamente correlacionados, casó contrario negativamente correlacionados.

$$lift(X \Rightarrow Y) = \frac{confianza(X \rightarrow Y)}{soporte(Y)}$$

Ecuación 16 Lift.

Existen varios algoritmos diseñados para identificar itemsets frecuentes y reglas de asociación. A continuación, se describen algunos de los más utilizados. (Monteserin, 2018)

Algoritmo Apriori

Fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, tiene dos etapas:

- Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite (itemsets frecuentes).
- Convertir esos itemsets frecuentes en reglas de asociación.

El concepto clave del algoritmo Apriori es su medida antimonotónica de soporte. Apriori supone que todos los subconjuntos de un conjunto de elementos frecuentes deben ser frecuentes (propiedad de Apriori). Si un conjunto de elementos es infrecuente, todos sus superconjuntos serán infrecuentes. (Amat Rodrigo, 2018)

FP Growth

Los investigadores de Han et al. propusieron en el 2000 un nuevo algoritmo llamado FP-Growth Basado en una mejora del algoritmo A priori, define una primera fase que consiste en descubrir los elementos frecuentes, y en una segunda fase en la que se generan las reglas de asociación de los elementos frecuentes encontrados, basado a parámetros de soporte y confianza mínimos. La principal diferencia con el algoritmo Apriori es la implementación usada en FP Growth, la cual es más eficiente al hacer uso de un árbol de elementos frecuentes que puede ser procesado más rápidamente que la estructura de datos usada en Apriori. (Amat Rodrigo, 2018)

Eclat

En el 2000, Zaki propuso un nuevo algoritmo para encontrar patrones frecuentes (itemsets frecuentes) llamado Equivalence Class Transformation (Eclat). La principal diferencia entre este algoritmo y Apriori es la forma en que se escanean y analizan los datos. El algoritmo Apriori emplea transacciones almacenadas de forma horizontal, es decir, todos los elementos que forman una misma transacción están en la misma línea. El algoritmo Eclat, sin embargo, analiza las transacciones en formato vertical, donde cada línea contiene un ítem y las transacciones en las que aparece ese ítem.

Cabe destacar que, el algoritmo Eclat, permite la identificación de itemsets frecuentes, pero no genera reglas de asociación. A pesar de ello, una vez identificados los itemsets frecuentes, se puede aplicar la segunda parte del algoritmo Apriori para obtenerlas. (Amat Rodrigo, 2018)

Ejemplo

Para este ejemplo se utilizará como data set los datos meteorológicos diarios. Se usa el algoritmo Apriori. Los atributos utilizados para asociación son la columna donde obtendrá los itemsets, el elemento mínimo que puede contener un itemset, soporte máximo y la confianza. Luego se hace la división de la columna Antecedente que contiene una colección de estos, arrojado por el nodo de la asociación que implementa Apriori.

Por último, mostramos la tabla donde se muestra el consecuente, el antecedente, la confianza y el lift entre otros.

Table View - 2:18 - Data to Report (association rules)

File Hilite Navigation View Output

Row ID	S Conseq...	S Antecedants	I ItemSe...	D Relativ...	D RuleCo...	D Absolut...	D Relativ...	D RuleLift	D RuleLift%	D Absolut...	D Relativ...	D rule qu...
Row111	Rihanna	The Pussycat Dolls	215	11.364	92.3	233	12.3	3.607	360.71	484	25.581	19,844.5
Row113	Britney Spears	The Pussycat Dolls	213	11.258	91.4	233	12.3	3.313	331.34	522	27.59	19,468.2
Row115	Lady Gaga	The Pussycat Dolls	201	10.624	86.3	233	12.3	2.671	267.13	611	32.294	17,346.3
Row104	Beyoncé	The Pussycat Dolls	195	10.307	83.7	233	12.3	3.989	398.85	397	20.983	16,321.5
Row106	Christina Ag...	The Pussycat Dolls	193	10.201	82.8	233	12.3	3.851	385.06	407	21.512	15,980.4
Row110	Katy Perry	The Pussycat Dolls	178	9.408	76.4	233	12.3	3.056	305.58	473	25	13,599.2
Row109	Madonna	The Pussycat Dolls	166	8.774	71.2	233	12.3	3.142	314.21	429	22.674	11,819.2
Row101	Shakira	The Pussycat Dolls	150	7.928	64.4	233	12.3	3.818	381.83	319	16.86	9,660
Row108	Avril Lavigne	The Pussycat Dolls	149	7.875	63.9	233	12.3	2.901	290.15	417	22.04	9,521.1
Row103	Ke$ha	The Pussycat Dolls	143	7.558	61.4	233	12.3	3.208	320.77	362	19.133	8,780.2
Row97	Black Eyed P...	The Pussycat Dolls	141	7.452	60.5	233	12.3	3.766	376.63	304	16.068	8,530.5

Figura 9 Nodo de Data to Report *para mostrar los resultados del análisis.*

8.5 Herramientas o lenguajes utilizados para minería de datos

8.5.1 Knime

Su nombre viene de Konstanz Information Miner y se pronuncia */naim/.* Es una plataforma de análisis de datos, informes e integración, de código abierto, integra varios componentes para machine learning y data mining. Está construido sobre la plataforma Eclipse y es extremadamente flexible y potente. (colaboradores de Wikipedia, KNIME, 2019)

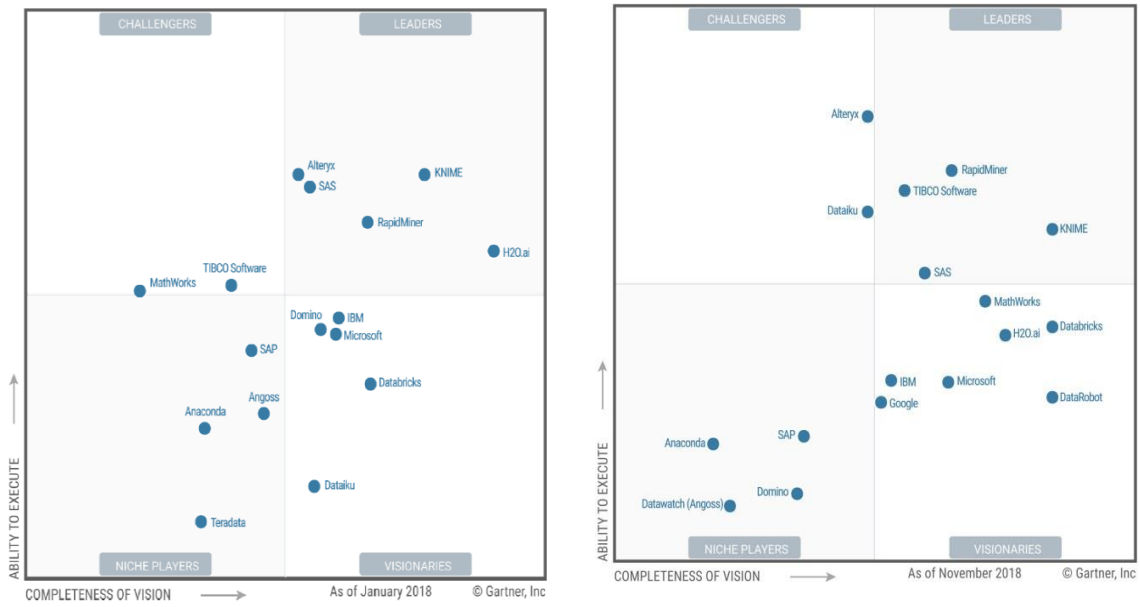


Figura 10 Cuadrante mágico de Gartner de enero de 2018 vs noviembre de 2018.

Para finales del 2017 y principios del 2018, Knime era la mejor punteada según el grupo Gartner para Data Science and Machine Learning Platforms (Plataformas de ciencia de datos y aprendizaje automático) colocándola en la parte superior del cuadrante debido a su capacidad de ejecución y visión del mercado altamente cambiada, como lo muestra la Figura 10 pero esta presenta el versus con los datos obtenidos para finales del 2018 la cual aunque ha bajado en puntuación sigue en una buena postura, siendo una de las plataformas mejor colocadas gracias a sus servicios.

Los estudios que se realizaron para poder seleccionar KNIME como software para la ejecución de minería de datos se especifican en el

Anexo 2: Definición, Evaluación y Selección de la Herramienta de Minería de Datos, y se concluye que KNIME presenta una robustez y maduración bastante alta comparado con sus competidores, incluyendo bibliotecas o librerías de terceros. Otro de los motivos es sobre la base de código abierto una gran comunidad que presenta. Y su principal ventaja es el costo nulo que presenta su uso para la Defensoría del Consumidor.

Está concebido como una herramienta gráfica y dispone de una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva.

Los nodos implementan distintos tipos de acciones que pueden ejecutarse sobre una tabla de datos:

- Manipulación de filas, columnas, entre otros., como muestreos, transformaciones, agrupaciones, entre otros.
- Visualización (histogramas, graficas de barra, matriz de puntos, gráfico lineal, entre otros.).
- Creación de modelos estadísticos y de minería de datos, como árboles de decisión, máquinas de vector soporte, regresiones, entre otros.
- Validación de modelos, como curvas ROC, generar matrices de clasificación, gráficos de dispersión, entre otros.
- Scoring o aplicación de dichos modelos sobre conjuntos a nuevos de datos.
- Creación de informes a medida gracias a su integración con BIRT (Business Intelligence and Reporting Tools, Inteligencia de negocio y herramientas de informes). (colaboradores de Wikipedia, KNIME, 2019)

KNIME permite la integración con otras herramientas de análisis como son Weka, Python, R e incluso crear clases propias con Java. Así como, los complementos adicionales permiten la integración de métodos para minería de texto, minería de imágenes, así como análisis de series temporales. Otros de los puntos a destacar es que KNIME son las tecnologías que emplea para su desarrollo como son Eclipse, Java, Python entre otros. (colaboradores de Wikipedia, KNIME, 2019)

8.5.2 Lenguaje de programación “R”

R es un entorno y un lenguaje de programación con un enfoque al análisis estadístico. R nació como una implementación de software libre del lenguaje “S” en 1993 por Robert Gentleman y Ross Iharka del departamento de Estadística de la Universidad de Auckland. R es uno de los lenguajes de programación más utilizados en el ámbito de la investigación científica siendo, además, popular en los campos de aprendizaje automático (machine learning), minería de datos, bioinformática y matemática financiera.

R es parte del sistema GNU y se distribuye bajo la licencia GNU PL. se encuentra disponible para los sistemas Windows, Unix, Mac y GNU/Linux. (colaboradores de Wikipedia, R (lenguaje de programación), 2020)

8.5.2.1 Características

- R proporciona un amplio abanico de herramientas estadísticas como por ejemplo modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, gráficas, etc.
- R hereda de S su orientación a objetos.
- R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.
- R permite generar gráficos con alta calidad.
- Puede utilizarse también como herramienta para el cálculo numérico.

8.6 Herramienta para la Extracción, Transformación y carga de los datos

8.6.1 Talend Open Studio

Talend Open Studio es una solución de Business Intelligence de código abierto y de fácil uso para la integración de los datos. Esta solución permite el desarrollo rápido y reduce costes de implementación por medio de un desarrollo gráfico con conectores prediseñados conectando a todos los sistemas de datos fuente y de destino.

Talend provee diferentes softwares de integración de código abierto, como son:

- Data Integration.
- Big Data.
- Data preparation.
- Enterprise Service Bus.
- Data Quality.
- Stitch Data Loader.

(Talend Inc., 2020)

Para nuestro proyecto ocuparemos el software “Data Integration” ya que este nos provee los diferentes nodos para la extracción, transformación y carga mejor conocido como ETL de diferentes repositorios o motores de bases de datos.

8.7 Herramienta para la visualización de datos

8.7.1 Power BI

Power BI es un servicio para el análisis empresarial desarrollado por Microsoft, esta aplicación nos proporciona visualizaciones interactivas como pueden ser líneas de tiempo, gráfico de burbujas, segmentado por jerarquías, entre muchos otros que se encuentran entre las aplicaciones para Power BI y capacidades de inteligencia empresarial con una interfaz simple para que usuarios finales tengan la capacidad de crear sus propios informes y paneles. (colaboradores de Wikipedia, Power BI, 2019)

Componentes claves del ecosistema de Power BI:

- **Power BI Desktop:** La aplicación basada en escritorio de Windows para equipos y escritorios, principalmente para diseñar y publicar informes en el servicio.

- **Power BI Service**: El servicio en línea basado en SaaS (software como servicio) (anteriormente conocido como Power BI para Office 365, ahora denominado PowerBI.com o simplemente Power BI).
- **Power BI Mobile Apps**: Las aplicaciones de Power BI Mobile para dispositivos Android y iOS, así como para teléfonos y tabletas Windows.
- **Power BI Gateway:** Puertas de enlace que se usan para sincronizar datos externos dentro y fuera de Power BI. En el modo empresarial, también pueden usar los flujos y PowerApps en Office 365.
- **Power BI Embedded:** La REST API de Power BI se puede usar para crear paneles e informes en las aplicaciones personalizadas que sirven a los usuarios de Power BI, así como a los usuarios que no son de Power BI.
- **Power BI Report Server:** Una solución de informes de Power BI local para empresas que no almacenan o no los datos en el servicio Power BI basado en la nube. Power BI Visuals Marketplace
- **Power BI Visuals Marketplace:** Un mercado de objetos visuales personalizados y objetos visuales con tecnología R. (colaboradores de Wikipedia, Power BI, 2019)

9 Formulación Del Problema

9.1.1 Definición del problema

Actualmente la Defensoría del Consumidor atiende reclamaciones por parte de la población mediante los diferentes mecanismos a su disposición, estos pueden ser de forma presencial en cualquiera de las 14 ventanillas centralizadas departamentales, en ventanillas de atención en cualquiera de las 6 sedes de Ciudad Mujer, además de las diferentes defensorías móviles, también, la defensoría pone a disposición de los consumidores diferentes medios electrónicos, como es el chat en su sitio web, chat por WhatsApp, correo electrónico, y llamadas por call center o llamadas por línea directa que se encuentran disponibles en distintas alcaldías en todo el país. Todas estas atenciones se envían a uno de los 4 Centros de Solución de Controversias y es ahí donde se ingresa dicha información en el Sistema de Atención a Reclamos y Atenciones (SARA), luego de ello se le asigna un técnico, este da seguimiento durante todos los pasos por los cuales un caso es tratado desde que se toma la solicitud hasta el cierre del caso, si se acordaron montos debido a una conciliación, si no se llegó a conciliar, si se archiva el caso, si posteriormente se reapertura, entre otras variantes que estos pueden tener, son diversos procesos los que se ven inmersos en la atención de personas usuarias, para los cuales la Defensoría del Consumidor está en búsqueda de una mejora continua, que le permita generar conocimiento sobre las atenciones brindadas, según el caso, región, sector, entre otros. Además, se cuenta con información sobre casos cerrados y los montos recuperados, de manera que permita brindar un servicio de calidad la población en general a quienes se debe.

Además la Defensoría del Consumidor, mediante la unidad de Análisis de Consumo y Mercados procesa sondeos de precios de distintas fuentes, para el caso de la dirección de Vigilancia de Mercado (DVM), los sondeos de precios se orientan a harinas, productos en supermercados, mercados, establecimientos mayoristas, artículos de la canasta básica y de temporada, además se cuenta con información de precios de otras entidades como es el Ministerio de Agricultura y Ganadería (MAG) con lo cual se busca brindar a la población información fidedigna sobre las fluctuaciones, rangos válidos de precios, en síntesis cuanto es lo que deben estar dispuestos a pagar por X o Y producto y que garantías esperar si en dado caso se les estén violentando sus derechos como consumidores.

9.2 Diagnóstico del problema

Dentro del diagnóstico se determinan las causas de la problemática actual u oportunidad de mejora, con este fin, se utilizan las siguientes técnicas: Entrevista, lluvia de ideas, y la matriz FODA.

9.2.1 Entrevista

9.2.1.1 Gerencia de Sistemas Informáticos

Se sostuvo una entrevista con el gerente de Sistemas Informáticos para tener la perspectiva del lado de tecnología sobre las necesidades de la institución para apoyar en la mejora continua y brindar así un servicio eficiente y de calidad a la población definiendo como tema de principal interés las atenciones, en términos generales se desea solucionar brechas en el conjunto de procesos que se ven inmersos en la atención del cliente, identificar inconsistencias en la información que se genera, mejorar los tiempos de respuesta en la atención a las personas usuarias, se tiene interés por analizar además si el personal con el que cuenta la institución

satisface las demanda de la población para hacer una mejor planificación y orientación efectiva de los recursos, información o patrones de empresas que se ven implicadas en casos en varias ocasiones, que el servicio que se está brindando este a la altura del servicio que se requiere por la población.

9.2.1.2 Unidad de Análisis de Consumo y Mercados

Se sostuvo así mismo una entrevista con la Jefa de la Unidad de Análisis de Consumo y Mercados (UACM), en la cual se hizo énfasis en el tiempo en el que se invierte en limpiar los datos proporcionados por el Sistema de Atención a Reclamos y Atenciones (SARA) ya que según un estudio realizado por el técnico encargado de limpiar dichos datos, la base de datos actual del sistema SARA puede llegar a contener hasta un 70% de errores o inconsistencias en los datos, por lo tanto la unidad emplea demasiado tiempo limpiándolos y dedicando poco tiempo al análisis, es por ello que no se tienen estimaciones sobre los tiempos de respuesta a las reclamaciones, patrones de comportamiento de las reclamaciones, demandas, aumento y disminuciones en las atenciones brindadas, los motivos de reclamos, los tipos de clientes quienes realizan las reclamaciones y cuáles son las empresas más reincidentes, también, se pone a disposición la mayor parte del recurso a la recepción de los datos provenientes de distintas fuentes y unidades para sacar reportes específicos solicitados por la presidencia de la Defensoría del Consumidor y el área de Comunicaciones, así mismo es el caso del análisis que se realiza al sondeo de precios, en algunos casos no es posible visualizar que X o Y producto va al alza o se han disparado repentinamente los precios ya que los recursos satisfacen únicamente las necesidades diarias, máxime mensuales más no proyecciones en base a datos históricos, por ejemplo los datos para los precios de mercados y de temporada, los inspectores se dirigen a diferentes mercados para recolectar datos en el sitio, además, se cuenta con los datos que el Ministerio de Agricultura y Ganadería (MAG) envía periódicamente, sobre informes de precios de la canasta básica, en el caso de supermercados envían sus paquetes de precios semanalmente, por consiguiente se tienen datos suficientes históricamente para realizar un análisis a conciencia siendo este de mucho provecho para la institución.

9.2.2 Lluvia de ideas

Situación: La defensoría del consumidor en relación a su proceso de mejora continua desea brindar un servicio de calidad a la población para las atenciones a reclamos y asesorías, así como brindar información oportuna referente a los precios de los productos.

- Algunas personas prefieren archivar sus casos, contrario a pasar por todas las actividades que deben realizar para confrontar a las partes solicitadas.
- En algunos casos los tiempos de respuesta a la población son demasiado largos debido a la burocracia que se pueda ver inmersa en los procesos.
- Se produce inconsistencia en la información generado en el transcurso del proceso de atención a casos.
- Puede que algunas áreas geográficas o en algunos métodos por los cuales se reciben solicitudes no se tengan los recursos necesarios para satisfacer la demanda de la población.
- Dentro de la cantidad nutrida de procesos que se llevan a cabo en torno a la atención de las personas usuarias posiblemente existan algunos en los cuales sea necesario realizar optimizaciones o solucionar brechas.

- Invertir la mayor cantidad de los recursos de la Unidad de Análisis de Consumo y Mercados (UACM) en recibir y procesar los datos de las diversas fuentes les proporcionan impide se realice un análisis profundo de las variables inmersas en el sondeo de precios.
- El análisis de los datos que maneja la UACM se realiza de forma manual.
- No se logra apreciar en algunos casos alzas en los precios de X o Y productos, o que pueda provocar que se disparen repentinamente.

9.2.3 Matriz FODA

Se plantea un análisis de la matriz FODA para las áreas de interés que abordará el proyecto, en el contexto de la generación de nuevo conocimiento a partir de la exploración de los datos, para la mejora del servicio a las personas usuarias y la mejora continua:

FORTALEZAS	OPORTUNIDADES
- La institución cuenta con una gerencia informática que permite brindar soluciones de tecnologías de la información a los diversos procesos de atención a las personas usuarias. - Aparte de los inspectores que recolectan datos acerca del sondeo de precios, cadenas de supermercados y otras entidades proporcionan datos respecto de los precios que manejan. - La Unidad de Análisis de Consumo y Mercado (UACM) está especializada en brindar información de valor a la presidencia de la Defensoría del Consumidor, así como insumos a comunicaciones para la elaboración de boletines informativos para la población. - El hecho de ser la UACM la unidad con la que se trabajará este proyecto es una fortaleza ya que se especializan en el tema del análisis de los datos.	- Solucionar brechas y/u optimizar los procesos inmersos en la atención de las personas usuarias permitiría brindar un servicio de mejor calidad. - Reducir los tiempos de atención a los casos influirá en una percepción positiva de la población. - Conocer una aproximación de la demanda de la población en los distintos medios de solicitud permitirá planificar y/o reorientar los recursos para satisfacer los requerimientos de las personas usuarias.
DEBILIDADES	AMENAZAS
- Por ser una institución de carácter público cuenta con un presupuesto poco flexible y a su vez limitado, lo cual disminuye su poder de inversión en investigación. - El empleo de la mayor parte del recurso humano de la UACM en recibir y procesar datos de las diferentes fuentes que maneja deja poco tiempo para realizar el análisis respectivo de las métricas inmersas en el sondeo de precios y atención a reclamos y asesorías en base a datos históricos.	- Una planificación sin los criterios correctos podría conducir a la saturación de algunos de los medios de solicitud. - Tiempos de respuesta demasiado largos concluyen en mayor número de casos archivados. - No estar en un proceso de mejora continua y acortamiento de brechas conduciría a un decaimiento del servicio a la población afectando negativamente en su opinión.

	- No brindar información fidedigna y oportuna respecto a los precios de los productos a la población generará desconocimiento de cuando hacer efectivos los derechos de los consumidores en caso se estén violentando.

Tabla 1 matriz FODA.

9.3 Problema general

La UACM recibe datos de las atenciones a reclamaciones del Sistema de Atención a Reclamaciones y Asesorías (SARA) y del sondeo de los precios de diversos orígenes (DVM, MAG, supermercados). Para todos estos datos se realiza un procesamiento de forma manual, invirtiéndose gran parte del tiempo en limpiar y preparar datos de las atenciones a casos, y en recibir y consolidar datos para el caso del sondeo de precios, impidiendo realizar un análisis de datos históricos.

9.4 Problemas específicos

9.4.1 Sondeos de precios

- Complejidad en el análisis de datos históricos de los sondeos de precios.
- No se visualizan en algunos casos alzas en los precios de los productos.
- Con el análisis actual se dificulta anticiparse a los hechos sobre fluctuaciones en los precios que permita alertar a la población y hacer efectivas las medidas pertinentes.
- Calculo engorroso debido al trabajo que conlleva el análisis de los sondeos de los precios de forma manual.
- Aplicar modelos matemáticos y/o estadísticos manualmente resulta complicado y demora mucho.

9.4.2 Atenciones a reclamaciones

- No se logra determinar demanda a futuro del servicio de atención a reclamaciones que permita realizar una planificación anticipada o reorientación de los recursos.
- No se logra medir el nivel de cumplimiento o satisfacción de la población en cuanto a los montos en los casos cerrados de atención a reclamaciones.
- Solucionar errores o inconsistencias de llenado de los datos de las atenciones a reclamaciones es una tarea que demora mucho tiempo.
- No hay control o seguimiento de causas atendidas que den paso a tomar medidas en posteriores casos.

10 Propuesta De Solución

10.1 Descripción

La Unidad de Análisis de Consumo y Mercados (UACM), desea brindar información de valor para la toma de dicha información obtenida de conocimiento a partir de patrones de comportamiento y predicciones basadas en los datos históricos de atención a reclamaciones y los sondeos de precios con los que cuenta la unidad, para llegar a esa situación deseada se presenta la solución propuesta mediante un diagrama de enfoque de sistemas.

Además, para una mejor comprensión de la solución propuesta, se representa de forma gráfica, el proceso de extracción de conocimiento a partir de modelos de minería de datos basados en los datos históricos de atención a reclamaciones y los sondeos de precios (en adelante MIDAS) en ella se describe el flujo de trabajo entre los distintos actores implicados en los procesos que conlleva el proyecto.

10.1.1 Enfoque de sistemas

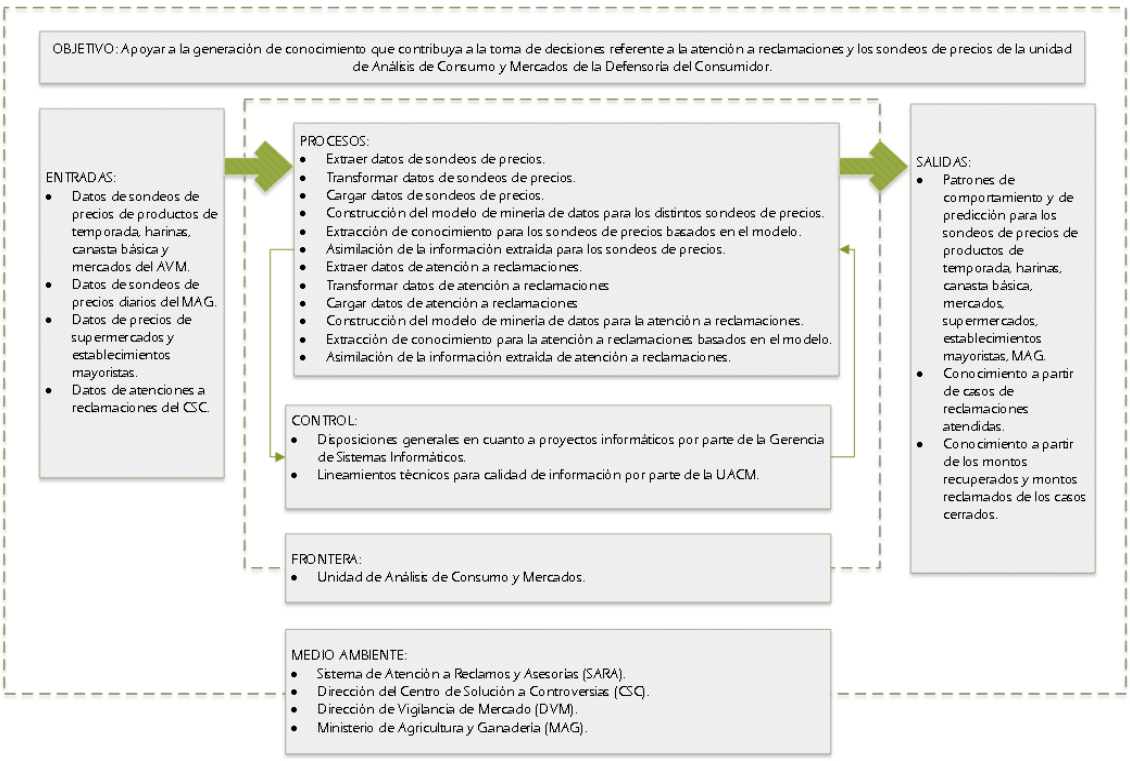


Figura 11 Enfoque de Sistemas de la solución propuesta por el proyecto MIDAS.

10.1.2 Proceso de extracción de conocimiento

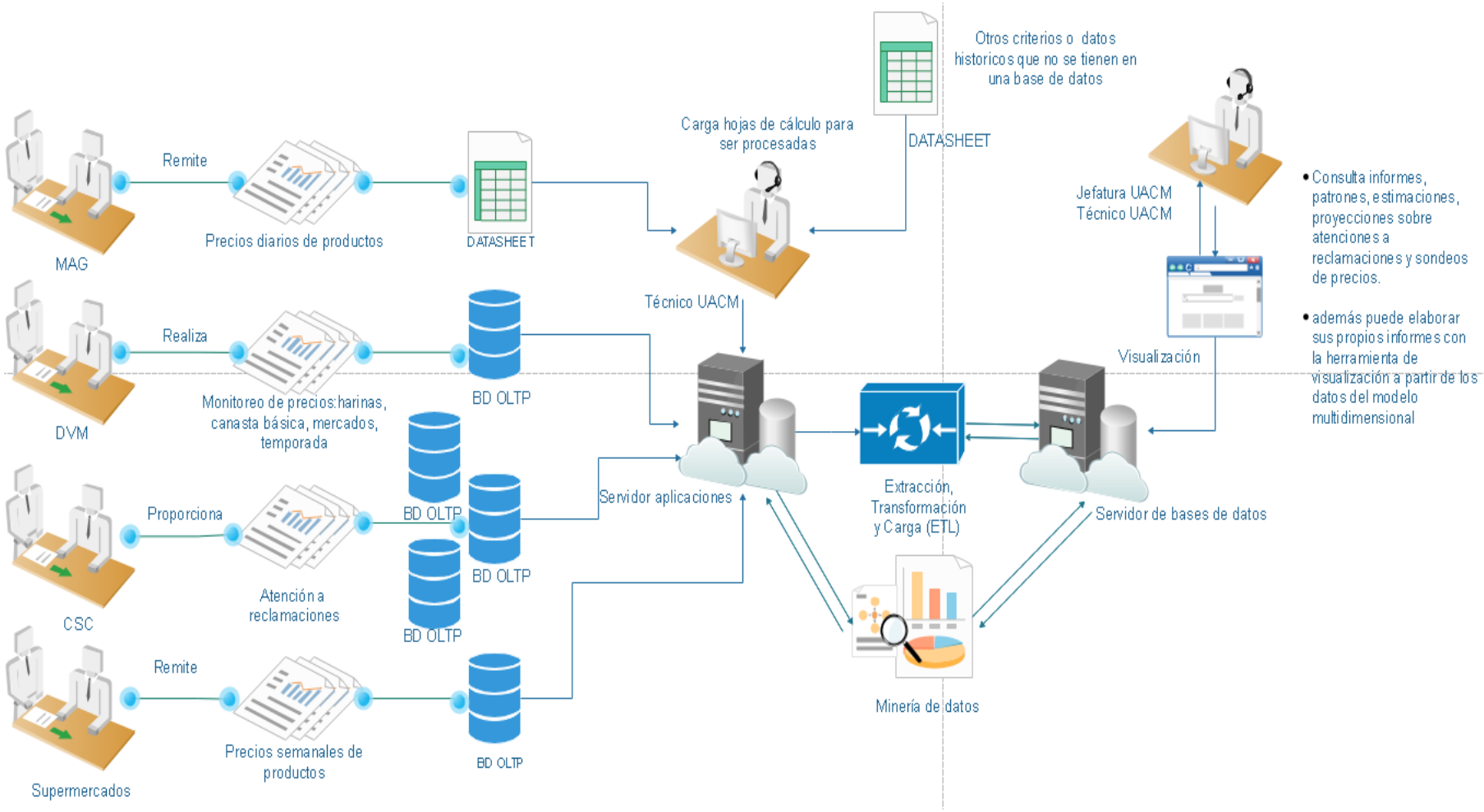


Figura 12 Propuesta de proceso de extracción de conocimiento para el proyecto MIDAS

La Figura 12 muestra el flujo de trabajo de la solución propuesta en el cual se tienen como orígenes de datos los sondeos de precios de Mercados, supermercados, los remitidos por el Ministerio de Agricultura y Ganadería (MAG), para el caso de atenciones se tienen los datos del "Sistema de Atención a Reclamaciones y Asesorías (SARA)", además los casos recibidos mediante "Defensoría en Línea" y su versión anterior como origen histórico.

La UACM deberá proporcionar al proceso los datos provenientes del MAG ya que solo se cuentan con hojas de cálculos como orígenes de datos, la solución se encargará del proceso de ETL (Extracción, transformación y carga).

Así mismo deberá brindar al proceso el conjunto de criterios (campos equivalentes, campos resumidos, agrupados u otros) dichos datos no se tienen almacenados en una estructura relacional si no en hojas de cálculo y son dinámicos por lo tanto el proceso únicamente será el encargado de procesarlos en conjunto con los orígenes de bases de datos.

Una vez se tienen los datos dentro de las estructuras del staging área mediante procesos ETL son trasladados al modelo multidimensional correspondiente, posteriormente se aplican técnicas de minería de datos siendo los resultados almacenados en una base de datos para finalmente mediante una herramienta de visualización que sea capaz de mostrar los resultados e inclusive diseñar otro tipo de informes a partir del data warehouse permitiendo esto apoyar a la toma de decisiones.

10.2 Componentes de la solución

10.2.1 Almacenes de datos: Sondeos de precios y atención a reclamaciones

- **Orígenes de datos:** Correspondiente a bases de datos transaccionales y otros conjuntos de datos como hojas de cálculo.
- **Procesos ETL:** Traslado de los datos de los orígenes al área de pruebas y posterior a los modelos multidimensionales realizando tareas de extracción, transformación y carga.
- **Staging área:** Base de datos intermedia que facilita el trabajo de los procesos ETL previó a su traslado a los modelos multidimensionales.
- **Data marts:** Bases de datos con estructura multidimensional que facilitan el análisis de los datos.

10.2.2 Minería de datos

- **Transformación de los datos:** Se transforman los sets de datos de acuerdo a la necesidad de cada técnica de minería.
- **Algoritmos:** Se aplican algoritmos en base a la técnica seleccionada al set de datos para cada caso.
- **Resultados:** Se transforman los resultados de los algoritmos para ser almacenados en una base de datos.

10.2.3 Visualización

- **Informes de inteligencia de negocios:** Teniendo como origen los almacenes de datos se presentan informes de alto impacto mediante una herramienta de visualización.
- **Informes de inteligencia de negocios minería de datos:** Se proporcionan informes de con los resultados de la búsqueda de conocimiento realizada con minería de datos.

11 Metodología

11.1 Actividades

Se hace una combinación de las metodologías ágiles con la metodología CRISP-DM teniendo una forma de trabajo iterativa e incremental para la fase 1 y únicamente incremental para la fase 2 retomando algunas de las tareas propias de las metodologías para abordar proyectos de minería de datos como lo es CRISP-DM.

N°	Actividad
	Almacén de datos
1	Refinamiento del requerimiento de información
2	Diseño del staging area
3	Diseño de procesos ETL Staging área
4	Desarrollo de procesos ETL staging área
5	Pruebas staging área
6	Diseño del staging area
7	Diseño de procesos ETL Staging área
8	Desarrollo de procesos ETL staging área
9	Pruebas modelo multidimensional
10	Documentación
	Minería de datos
1	Exploración de los datos
2	Técnica de asociación
2.1	Comprensión del negocio
2.2	Comprensión de los datos
2.3	Preparación de los datos
2.4	Modelar
2.5	Evaluar resultados
2.6	Preparar resultados
3	Técnica de clasificación
4	Técnica de pronóstico
5	Técnica de agrupamiento
	Visualización
1	Elaboración de informes de inteligencia de negocios
2	Visualización resultados minería de datos

Tabla 2 Actividades de desarrollo de las iteraciones

La Tabla 2 muestra las actividades principales llevadas a cabo en las fases, dividiéndose estas en 3 grandes grupos (Almacén de datos, minería de datos, visualización), para el caso de los numerales del 2.1 al 2.6 se repiten para cada una de las técnicas utilizadas (Asociación, clasificación, pronostico y agrupamiento).

CRISP-DM	En el presente proyecto
Modelar	Técnica
Modelar	Algoritmo
Comprensión de los datos	Población
Comprensión de los datos	Variables
Comprensión del negocio	Hipótesis
Modelar	Procedimiento
Evaluar y preparar resultados	Resultados
Evaluar y preparar resultados	Interpretación de resultados

Tabla 3 Tabla equivalencias CRISP-DM respecto proyecto MIDAS

La Tabla 3 muestra el termino correspondiente utilizado en el presente trabajo para cada una las tareas comprendidas dentro de la metodología CRISP-DM.

11.2 Estándares

11.2.1 Estándares de documentación

Para la documentación se establecen los siguientes estándares:

1) Tipo y tamaño de la letra:
   a) Párrafos: Tipo de letra: Arial, Tamaño de letra: 11.
   b) Titulo 1: Tipo de letra: Arial, Tamaño de letra: 14.
   c) Titulo 2: Tipo de letra: Arial, Tamaño de letra: 13.
   d) Titulo 3: Tipo de letra: Arial, Tamaño de letra: 12.
   e) Titulo 4: Tipo de letra: Arial, Tamaño de letra: 11.
   f) Titulo 5: Tipo de letra: Arial, Tamaño de letra: 11.
   g) Titulo 6: Tipo de letra: Arial, Tamaño de letra: 11.
2) Tamaño de las imágenes:
   a) 1 x 1: utilizado para nodos en los flujos de trabajo
   b) 5 x 5: utilizado para las imágenes pequeñas.
   c) 8 x 10: utilizado para las imágenes medianas.
   d) 8 x 15: utilizado para los reportes de visualización.
   e) 10 x 15: utilizado para las imágenes media grandes.
   f) 15 x 15: utilizado para los flujos de trabajo en Knime.
   g) 20 x 17: utilizado para algunos diagramas.
3) Título de imágenes, ecuaciones y tablas:
   a) Formato para imágenes con el título: “Figura ##:” Tamaño de letra: 9 y tipo cursiva.
   b) Formato para ecuación con el título: “Ecuación ##:” Tamaño de letra: 9 y tipo cursiva.
   c) Formato para tablas con el título: “Tabla ##:” Tamaño de letra: 9 y tipo cursiva.
4) Funciones y Nodos:
   a) Al nombrar funciones o nodos estos estarán entre comillas, negrita y cursivas.

11.2.2 Estándares de base de datos

Para las bases de datos de los modelos multidimensionales y minería de datos se aplica la nomenclatura siguiente:

5) Para el nombre de las bases de datos la especificación es:
   a) nombre_descriptivo_sondeos.

i) Ejemplo: mining_sondeos y dw_sondeos.

6) Para la especificación de los nombres de las tablas:
   a) jerarquía_nombre:
      i) Ejemplo: dim_tiempo, stg_tiempo, fact_sondeos_mag.
7) Para el nombre de los campos:
   a) clave_subrogada_nombre_tabla
      i) Ejemplo: sk_tiempo, sk_producto, sk_categoria.
   b) Id_nombre_tabla
      i) Ejemplo: id_tiempo, id_producto, id_categoria.

11.3 Paquetes de información

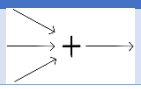
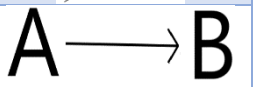
Un paquete de información define las relaciones entre el tema y las medidas clave de rendimiento. El diagrama del paquete de información tiene un objetivo altamente específico, que proporciona un alcance específico para los requisitos del usuario. Debido a que los diagramas de paquetes de información apuntan a lo que desean los usuarios, son efectivos para facilitar la comunicación entre el personal técnico y los usuarios, lo que indica cualquier inconsistencia entre los requisitos y lo que entregará el almacén de datos. (Balanced Insight, Inc., 2000) En la Tabla 4 Se visualiza un paquete de información.

Tema:	**Análisis de venta**					
JERARQUIAS	Tiempo	Productos	Marca	Establecimiento	Categoría	Genero
	Año	Producto	Marca	País	Categoría	Genero
	Mes			Área / Región		
	Semana			Departamento		
	Día			Municipio		
				Tienda		
Hechos Medidos:	**Forcast de venta, presupuesto de venta.**					

Tabla 4 Ejemplo de paquete de información.

11.4 Nomenclatura de diagramas multidimensionales (Modelado conceptual)

Para obtener un modelo conceptual de un proceso ETL no hay una nomenclatura estándar, pero se pueden adoptar algunas propuestas como es la extensión UML para este ámbito, la cual ha sido propuesta por Sergio Luján-Mora y Juan Trujillo, en la cual se ha tenido un conjunto reducido pero potente de mecanismos ETL, para poder reducir la complejidad de la propuesta. La cual se puede ver en la Tabla 5.

Mecanismo ETL (Estereotipo)	Descripción	Icono
Aggregation	Agrega los datos (SUM, AVG, MAX/MIN, COUNT, etc.) en base a algún criterio.	
Conversion	Cambia los tipos de datos, el formato o calcula nuevos datos (atributos derivados) a partir de los datos existentes.	

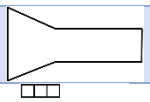
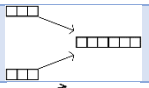
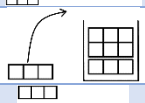
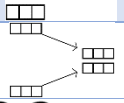
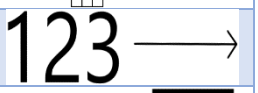
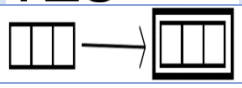
Filter	Filtra los datos no deseados y verifica la calidad de los datos en base a restricciones.	
Incorrect	Redirige los registros incorrectos o descartados a un destino separado para su posterior verificación; sólo se puede usar con Filter, Loader y Wrapper.	
Join	Une dos fuentes de datos relacionadas entre sí a través de uno o varios atributos.	
Loader	Carga los datos en el destino de un proceso ETL (en un hecho o dimensión del DW).	
Log	Controla y registra la actividad de otro mecanismo ETL, con el fin de auditar las transformaciones realizadas.	
Merge	Integra los datos provenientes de dos o más fuentes de datos con atributos compatibles.	
Surrogate	Genera una clave substituta única, que se emplea para reemplazar la clave empleada en las fuentes de datos.	
Wrapper	Transforma una fuente de datos nativa en una fuente de datos basada en registros.	

Tabla 5 Mecanismos ETL y su representación en UML.

(Luján-Mora & Trujillo, 2003)

Además, se cuenta con una nomenclatura basada en UML para la creación de estructuras multidimensionales tal y como se muestra en la Tabla 6.

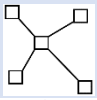
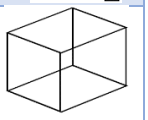
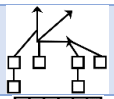
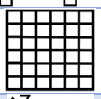
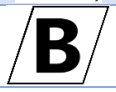
Concepto MD (Estereotipo)	Descripción	Icono
StarPackage	Paquetes de este estereotipo representan esquemas estrellas, compuestos de hechos y dimensiones; se emplea en el nivel 1.	
FactPackage	Paquetes de este estereotipo representan hechos, compuestos de medidas y relacionados con las dimensiones; se emplea en el nivel 2	
DimensionPackage	Paquetes de este estereotipo representan dimensiones, compuestas de jerarquías; se emplea en el nivel 2	
Fact	Clases de este estereotipo representan hechos, compuestos de medidas; se emplea en el nivel 3	
Dimension	Clases de este estereotipo representan dimensiones, compuestas de jerarquías; se emplea en el nivel 3	
Base	Clases de este estereotipo representan niveles de jerarquía en una dimensión; se emplea en el nivel 3	

Tabla 6 Conceptos multidimensionales y su representación en UML.

(Luján-Mora & Trujillo, 2003)

11.5 Estándares minería de datos

N° de caso: C-<Abreviatura de la técnica>- correlativo

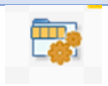
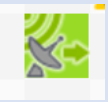
Técnica	Que técnica se utilizó.
Algoritmos	Que algoritmo se utilizó
Población	Qué conjunto de datos conforman la población.
Variables	Que variables son inmersas en el caso
Hipótesis	Cuál es el objetivo de minería de datos o hipótesis para el caso.
Procedimiento	Cuál es el tipo de solución que se trabajó.
Resultados	Los resultados esperados.
Interpretación de resultados	Que indicadores o criterios se evalúan para concluir el resultado.
Herramienta de software	Con que tecnología se trabajó.

Tabla 7 Plantilla para ficha de contenido del caso de minería de datos.

La Tabla 7 representa la plantilla para la ficha del contenido del caso de minería de datos siendo la contextualización en cada una de las historias de usuario que se abordaron referente a la temática de minería de datos.

11.6 Nodos Talend utilizados en la solución

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
tCreateTable	Este nodo permite crear una tabla en una base de datos especificando un esquema.	Creación de la tabla producto	
tDBClose	Cierra la transacción confirmada en una base de datos conectada.	Cerrar la conexión a la base de datos.	
tDBCommit	Sirve para enviar los cambios a la base de datos a partir de una conexión mediante transacción.	Enviar los registros que fueron insertados en una tabla a la base de datos.	
tDBConnection	Se utiliza para establecer la conexión con una base de datos	Establecer la conexión con PostgreSQL	
tDBInput	Obtener un flujo de datos a partir de una consulta de base de datos.	Consultar regitros de dos tablas de una base de datos y convertirlos al flujo de datos.	
tDBOutput	Permite escribir datos en una tabla de la BD, realizando acciones sobre la tabla o los registros.	Truncar una tabla y luego insertar registros provenientes del flujo de datos	
tDBRollback	Cancela la confirmación de transacción en una base de datos conectada para evitar confirmar parte de una transacción involuntariamente.	Si al ejecutar un job se produce un error, devuelve a la base de datos al estado previo.	

tDBRow	Permite realizar una consulta de tipo DML o DDL a una base de datos	Deshabilitar integridad referencial, truncar una tabla mediante SQL	
tDBSCDETL	Sincroniza una tabla de origen a una dimensión SCD con campos tipo SCD 1 y/o 2	Sincronizar la tabla producto a la dimensión producto.	
tELTPostgresqlMap	Permite transformar una o varias tablas de una base de datos postgres en una o muchas salidas.	Conectar sondeos_mag y plaza mediante JOIN o implicit JOIN y luego sacar obtener una única salida.	
tETLPostgresqlInput	Leer el contenido de una dimensión o tabla de la base de datos.	Leer los datos de la dimensión tiempo	
tETLPostgresqlOutput	Sirve para escribir en una tabla el flujo de datos proveniente de un tELTPostgresqlMap	Mayormente se utiliza para alimentar la tabla de hechos.	
tFileInputDelimited	Permite leer un archivo delimitado y convertirlo en un flujo de datos	Leer un archivo delimitado por comas (.CSV)	
tFilterRow	Permite filtrar registros estableciendo una o varias condiciones	Filtrar registros que en una columna tenga valores diferentes de vacío	
tFlowMeter	Cuenta el número de filas procesadas en el flujo definido, por lo que el componente tFlowMeterCatcher puede detectar este número para fines de registro	Reportar los datos que no se cargan a la tabla.	
tJavaRow	Permite introducir código personalizado que puede integrar en un programa Talend	Crear la estructura y notificación del correo.	
tLogCatcher	Funciona como una función de registro activada por uno de los tres: excepción Java, tDie o tWarn, para recopilar y transferir datos de registro.	Capturar el error del job ejecutado.	
tLogRow	Imprime en pantalla un flujo de datos en forma tabular especificando el esquema.	Imprimir en pantalla los registros del flujo de datos antes de hacer commit.	
tMap	Sirve para transformar uno o varios orígenes de datos en una o varias salidas de flujo de datos	Leer 3 flujos de datos y obtener una única salida	

tRedirectOutput	Permite escribir en un archivo de texto plano en el ordenador lo proveniente de un tLogRow	En lugar de imprimir en pantalla, escribir en un archivo de texto plano el log de registros insertados.	
tReplace	Permite reemplazar el contenido de una o varias columnas del flujo de datos.	Reemplazar “$” por “” en un flujo de datos leído desde una hoja de calculo.	
tRunJob	Este nodo permite invocar la ejecución de un Job.	Invocar la ejecución del job de normalización	
tSentMail	Envía correos electrónicos y archivos adjuntos a destinatarios definidos	Enviar el correo con el error del job.	
tUniqRow	Permite ingresar un flujo de datos con registros repetidos y obtener una salida sin registros repetidos	Obtener solo registros únicos para plazas.	

Tabla 8 Nodos de Talend Open Studio utilizados en los flujos de trabajo.

La Tabla 8 muestra describe cada uno de los nodos utilizados en la herramienta Talend para abordar cada uno de los casos en los flujos de trabajo extracción, transformación y carga.

11.7 Nodos Knime utilizados en la solución

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
DB Query Reader	Permite realizar una consulta a una base de datos	Obtener precios de los productos por cada plaza	
DB SQL Executor	Permite ejecutar una consulta DDL	Crear la extensión tablefunc en postgres para realizar pivot	
Create Collection Column	Convertir varias columnas en una sola columna de tipo collection	Crear una columna de colección de datos para productos sondeados en un día	
Number To String	Convertir una columna de tipo numerico a cadena de caracteres	Convertir un id a String para realizar una unión	
Joiner	Realizar uniones entre tablas	Realizar una unión de dos flujos de datos mediante claves	
GroupBy	Permite agrupar un flujo de datos mediante algún método de agregación	Agrupar precios por producto mostrando el promedio por cada producto.	
Double configuration	Permite Crear una variable de tipo double a	Establecer rago para el soporte mínimo en reglas de asociación	

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
	la cual se establecen minimos y máximos		
Create Bit Vector	Convierte varias columnas en una sola de tipo vector de bit's	Crear un vector de bits para introducirse al algoritmo Apriori	
Association Rule Learner (Borgelt)	Permite generar reglas de asociación mediante el algoritmo A priori	Encontrar Reglas de asociación entre los productos	
RowID	Asigna un identificador de registro a cada fila en un flujo de trabajo	Si se necesita un identificador extra para la fila.	
Partitioning	Divide la población en dos particiones para entrenar y probar datos.	Divide la población en datos de entrenamiento y prueba.	
Decision Tree Learner	Permite inducir un árbol de decisión que será almacenado en memoria principal.	Genera un modelo que se usara para clasificar los datos de prueba.	
Decision Tree Predictor	Utiliza el modelo para predecir nuevas clases para nuevos patrones.	Crea una clasificación para los datos de prueba.	
Database Writer (legacy)	Permite escribir un set de datos en una tabla de la base de datos.	Almacenar los resultados de un flujo de trabajo de minería de datos.	
Decision Tree To Image	Crea una imagen de la vista del árbol de decisión, en formato PNG.	Crea una imagen de la vista del árbol de decisión.	
Image Writer (Port)	Escribe una imagen en la URL especificada.	Crea una imagen PNG en la URL especifica.	
Scorer	Compara dos columnas por sus pares de valor de atributo y muestra la matriz de confusión.	Genera una matriz de confusión para interpretar el nivel resultados.	
Pivoting	Permite realizar la operación de pivot a un set de datos, aplicando funciones de agregación.	Cuando necesitamos cambiar la disposición de ciertas filas a columnas agrupando y aplicando funciones de agregación.	
RowFilter	Permite filtrar filas en un set de datos.	Si se desea filtrar los datos para el resto del flujo de trabajo.	
Rule Based RowFilter	Permite filtrar filas de un data set en base a reglas	Cuando se quiere hacer filtros más avanzados	
Generates date&time	Sirve para generar valores de fecha y tiempo	Generar valores futuros de fechas para las proyecciones	

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
Arima Learner	Sirve para entrenar las predicciones de series temporales mediante el algoritmo ARIMA	Entrenar predicciones para precios de productos de sondeos	
Arima Predictor	Sirve para realizar las predicciones de series temporales mediante el algoritmo ARIMA	Predecir precios de los productos para las próximas semanas	
Line Plot	Permite generar un gráfico de líneas	Cuando se quiere poner en perspectiva 2 o más variables	
Metanode	Permite encapsular un flujo de trabajo en un solo nodo existiendo la posibilidad de parámetros de entrada y salida.	Cuando un flujo de trabajo es demasiado extenso	
Counter	Permite crear una columna de cuenta de los registros.	Para utilizarse como columna de unión con otro set de datos	
Column Appender	Permite anexar una columna a un set de datos sin la operación de unión	Cuando se desea anexar una columna generado en el flujo de trabajo	
Missing Value	Permite identificar valores faltantes y definir que trato se les dará	Cuando deseamos colocar un valor en específico a valores faltantes	
Missing Value (Apply)	Aplicar los criterios definidos en el nodo Missin Value al set de datos.	Es de uso obligatorio cuando se usa el nodo Missing Value	
Column Resorter	Permite cambiar el orden en el que están dispuestas las columnas de un set de datos.	Cuando vamos a escribir en la BD y queremos que las columnas se creen en un orden específico.	
Concatenate	Permite anexar un dataset seguido de otro dataset	Para anexar datos de una predicción seguidos de los datos de prueba	
Column Rename	Permite renombrar las columnas de un set de datos.	Cuando necesitamos que una columna tenga un nombre en específico.	
Column List Loop Start	Inicia un bucle sobre un listado de columnas	Iniciar bucle y realizar iteración por cada producto	
Empty Table Switch	Sirve para saltar la ejecución en un bucle a la siguiente iteración cuando cumpla el criterio de tabla vacía	Para prevenir errores en ejecuciones de bucles.	

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
Loop End	Marca el final del bucle		
Chunk Loop Start	Iniciar bucle especificando cantidad de registros a abordar en cada iteración	Se usa cuando los datos que se recorren en cada iteración son modelos	
Loop End (Column Append)	Finalizar bucle y agregar a una tabla el resultado de cada iteración	Cuando se quiere acumular los resultados en cada iteración	
Cell to Model	Convierte una celda de tabla a un modelo	Cuando se quiere utilizar el modelo que viene en cada celda para predecir	
Arima Parameter Extractor	Extrae las variables contenidas en el modelo de ARIMA	Cuando se desea extraer algún dato del modelo dentro de variables	
Table Column to Variable	Convierte una columna de tabla del flujo de datos en variable	Cuando se necesita utilizar como filtro en otro nodo	
Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
Reset on Load	Reinicia todos los nodos en un flujo de trabajo	Cuando se tienen salidas a hojas de cálculo, texto plano o base de datos y se requiere sobrescribirlas, esto ejecutando nuevamente el flujo de trabajo.	
Table Creator	Permite crear una tabla de flujo de datos especificando nombres de columnas y sus valores	Si se necesita una serie de variables un insumo factible es obtenerlo de una tabla.	
Table Row To Variable	Convierte cada columna de una tabla en una variable	Cuando se necesita una serie de variables se pueden crear en una tabla y posteriormente ser transformadas mediante este nodo.	
PostgreSQL Connector (legacy)	Permite crear una conexión con una base de datos Postgres SQL	Cuando no se requiere especificar una cadena de conexión JDBC y se desea conectar específicamente con una base de datos Postgres SQL.	
String Manipulation	Permite realizar operaciones mediante funciones a los campos string.	Cuando necesitamos realizar transformaciones en los campos string como (lower, upper, replace, substr)	

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
Ungroup	Permite Desagrupar una columna de tipo colección de datos.	Cuando necesitamos obtener filas como elementos hay en un campo de tipo colección de datos.	
Cell Splitter	Permite dividir una columna en varias columnas definiendo un separador.	Cuando necesitamos obtener columnas como numero elementos separados hay en un string.	
Column Filter	Permite seleccionar las columnas que se desea mantener en el flujo de trabajo.	Cuando algunas de las columnas dejan de ser necesarias.	
Column Aggregator	Permite aplicar funciones de agregación sin necesidad de agrupar los datos.	Por ejemplo, si deseamos obtener en una nueva columna la concatenación de 2 columnas.	
Rank	Permite enumerar los datos en base a un agrupador y un ordenador.	Cuando deseamos obtener el primer elemento por cada grupo de un set de grupos de datos.	
Numeric Outliers	Detecta y trata los valores atípicos para cada una de las columnas seleccionadas individualmente.	Genera el modelo para eliminar de la población los registros con montos reclamados igual a cero.	
Numeric Outliers (Apply)	Trata los valores atípicos en los datos de entrada de acuerdo con los parámetros de la entrada del modelo.	Elimina de la población los registros con montos reclamados igual a cero.	
Cell Replacer	Reemplaza las celdas de una columna por un diccionario de datos en tabla.	Cambia la columna de counter por los nombres de las clases.	
Sorter	Permite ordenar las filas de un set de datos	Para posteriormente particionar los datos	
Unpivoting	Permite transformar múltiples columnas enfilas.	Cuando se necesita cambiar la representación de los datos de columnas a filas	
Constant Value Column	Permite crear una columna con un valor constante.	Cuando se necesita clasificar las filas de un data set.	
Math Formula	Permite trabajar con fórmulas y expresiones matemáticas.	Cuando se necesita aproximar un resultado.	
String to Date&Time	Permite convertir de una cadena de caracteres a formato fecha	Cuando se necesita los datos en formato fecha	

Nombre	Descripción	Utilidad o ejemplo de uso.	Representación
Extract Date&Time Fields	Permite extraer los campos derivados de una fecha	Cuando se necesita (el nombre del mes, año o trimestre) de una fecha.	
End IF	Aplicar el criterio al evaluar una condición	Aplicar el criterio del nodo Empty Table	
Reference Column Filter	Permite filtrar los datos en base a un nombre de columna de referencia	Cuando se quiere filtrar de forma dinámica dentro de un loop	
K-Means	Permite generar los centros de clúster para un número predefinido de clústeres (sin número dinámico de clústeres)	Asigna el mejor clúster a cada dato, según los centros de clústeres más cercanos.	
Color manager	Permite asignar un color según los valores de una columna.	Asigna un color diferente a cada clúster.	
Scatter Plot (Local)	Crea una gráfica de dispersión de dos atributos seleccionables	Utilizado para generar un gráfico de dispersión entre las variables.	
Cluster Assigner	Asigna el mejor clúster según el dato.	Utilizado para asignar el mejor clúster a los datos nuevos.	
Entropy scorer	Genera indicadores de rendimiento para los resultados de los algoritmos de agrupamiento.	Utilizado para calcular las estadísticas sobre el rendimiento de los algoritmos de agrupamiento.	

Tabla 9 Nodos de Knime utilizados en los flujos de trabajo.

La Tabla 9 muestra describe cada uno de los nodos utilizados en la herramienta Knime para abordar cada uno de los casos en los flujos de trabajo de minería de datos.

12 Planificación

12.1 Product Backlog

(ID) Sprint	(ID) Historia	Requisitos	Origen	Esfuerzo (Días)	Esfuerzo (Horas)	Prioridad
Sprint 1: Sondeo de precios MAG						
1	RSPP01	Transformación de los datos en Excel a una base de datos relacional.	Jefa UACM	3	24	1
1	RSPP02	Migración de datos de una base de datos relacional a un modelo multidimensional.	Jefa UACM	3	24	1
1	RSPP03	Informe de las fluctuaciones y el comportamiento de los precios de granos básicos	Jefa UACM	2	16	1
1	RSPP04	Gráfico de variación semanal de precios de granos básicos	Jefa UACM	2	16	1
1	RSPP05	Patrones de comportamiento y/o predicciones de los precios de los granos básicos.	Jefa UACM	3	24	1
1	RSPP06	Normalidad de los precios de granos básicos	Jefa UACM	3	24	1
				16	128	
Sprint 2: Sondeo de precios de Mercados						
2	RSPH01	Transformación de los datos en Excel a una base de datos relacional.	Jefa UACM	2	16	2
2	RSPH02	Migración de datos de una base de datos relacional a un modelo multidimensional.	Jefa UACM	2	16	2
2	RSPH03	Informe de las fluctuaciones y el comportamiento de los precios de granos básicos	Jefa UACM	3	24	2
2	RSPH04	Patrones de comportamiento y/o predicciones de los precios de los granos básicos.	Jefa UACM	2	16	2
2	RSPH05	Normalidad de los precios de granos básicos	Jefa UACM	2	16	2
				11	88	
Sprint 3: Sondeo de precios de Supermercados						

(ID) Sprint	(ID) Historia	Requisitos	Origen	Esfuerzo (Días)	Esfuerzo (Horas)	Prioridad
3	RSPS01	Migración de datos de una base de datos relacional a un modelo multidimensional	Jefa UACM	2	16	3
3	RSPS02	Informe de las fluctuaciones y el comportamiento de los precios de granos básicos	Jefa UACM	2	16	3
3	RSPS03	Informe de las fluctuaciones y el comportamiento de los precios de granos básicos desde distintas variables	Jefa UACM	2	16	3
3	RSPS04	Mapa de precios de los productos por municipio	Jefa UACM	2	16	3
3	RSPS05	Mapa de precios de los productos por región	Jefa UACM	2	16	3
3	RSPS06	Mapa de precios de los productos por departamento, región y establecimiento.	Jefa UACM	2	16	3
				12	80	
Sprint 4: Construcción del modelo multidimensional para Atenciones a Reclamaciones						
4	RA101	Identificar orígenes de datos del modelo.	Jefa UACM	0.5	4	4
4	RA102	Identificar tablas dentro de los orígenes de datos.	Jefa UACM	2.5	20	4
4	RA103	Identificar campos a migrar.	Jefa UACM	1.5	12	4
4	RA104	Migrar tablas al staging area.	Jefa UACM	5	40	4
4	RA105	Diseñar los paquetes de información.	Jefa UACM	1	8	4
4	RA106	Diseño del modelo multidimensional	Jefa UACM	2	16	4
4	RA107	Diseño de Diagramas UML de los ETL.	Jefa UACM	2.5	20	4
4	RA108	Desarrollo de los ETL.	Jefa UACM	22.5	180	4
4	RA109	Pruebas.	Jefa UACM	2	16	4
				39.5	316	
Sprint 5: Minería de datos para Atenciones a Reclamaciones						

(ID) Sprint	(ID) Historia	Requisitos	Origen	Esfuerzo (Días)	Esfuerzo (Horas)	Prioridad
5	RA201	Determinar la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año.	Jefa UACM	3.8	30	5
5	RA202	Segmentar consumidores en base a los motivos en los cuales han solicitado atención a la DC.	Jefa UACM	2	16	5
5	RA203	Clasificar la influencia que ha tenido la DC en las atenciones en base a la solución y montos recuperados.	Jefa UACM	3.8	30	5
5	RA204	Pronosticar casos a recibir en fechas futuras.	Jefa UACM	2	16	5
5	RA205	Identificar qué solución tendrán los casos recibidos en base a la edad de los consumidores.	Jefa UACM	3	24	5
5	RA206	Clasificar a los proveedores en base las soluciones que se han tenido.	Jefa UACM	3	24	5
5	RA207	Pronosticar casos solucionados en fechas futuras.	Jefa UACM	2	16	5
5	RA208	Identificar grupos de meses que reciben más casos.	Jefa UACM	3	24	5
				22.6	180	
Sprint 6: Visualización para Atenciones a Reclamaciones						
6	RA301	Mostrar en un mapa los proveedores que en más reiteradas ocasiones han sido denunciados por consumidores.	Jefa UACM	2	16	5
6	RA302	Visualizar las atenciones según el motivo.	Jefa UACM	2	16	5
6	RA303	Visualizar Atenciones brindadas.	Jefa UACM	1	8	5
6	RA304	Visualizar Atenciones brindadas, según oficina	Jefa UACM	2	16	5
6	RA305	Visualizar Atenciones brindadas, según región	Jefa UACM	1	8	5
6	RA306	Visualizar Atenciones brindadas, según sector	Jefa UACM	1	8	5
6	RA307	Atenciones en medios descentralizados.	Jefa UACM	2	16	5
6	RA308	Atenciones en ventanillas descentralizadas.	Jefa UACM	2	16	5
6	RA309	Casos cerrados y montos recuperados.	Jefa UACM	2	16	5

(ID) Sprint	(ID) Historia	Requisitos	Origen	Esfuerzo (Días)	Esfuerzo (Horas)	Prioridad
6	RA310	Mostrar número de casos según departamentos o municipios y filtrar si es caso aperturado o cerrado e indicar en qué departamento se ha dado la mayor cantidad de casos.	Jefa UACM	1.5	12	5
6	RA311	Mostrar cantidad de atenciones según forma de recepción por municipio, departamento, ventanilla u oficina e indicar por qué medio se ha dado la mayor cantidad de casos.	Jefa UACM	2	16	5
6	RA312	Visualizar la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año (referencia a RA201).	Jefa UACM	3	24	6
6	RA313	Visualizar la segmentación de consumidores en base a los motivos en los cuales han solicitado atención a la DC (referencia a RA202).	Jefa UACM	2	16	6
6	RA314	Visualizar las segmentaciones de meses donde se reciben más atenciones. (Referencia RA208).	Jefa UACM	2	16	6
6	RA315	Visualizar la solución de casos recibidos en base a la edad de los consumidores (Referencia RA205).	Jefa UACM	2	16	6
6	RA316	Visualizar la influencia que ha tenido la DC en las atenciones (Referencia a RA203).	Jefa UACM	2	16	6
6	RA317	Visualizar la clasificación de los proveedores en base las soluciones que se han tenido (Referencia a RA206).	Jefa UACM	1	8	6
6	RA318	Visualizar el pronóstico de las cantidades de casos nuevos para fechas posteriores. (referencia a RA204).	Jefa UACM	1	8	6
6	RA319	Visualizar el pronóstico casos solucionados en fechas futuras (referencia a RA207).	Jefa UACM	1	8	6
6	RA320	Mostrar en un mapa las atenciones que han sido brindadas a los consumidores.	Jefa UACM	2	16	5
				34.5	276	

12.2 Arquitectura de servidores

12.2.1 Diagrama de despliegue.

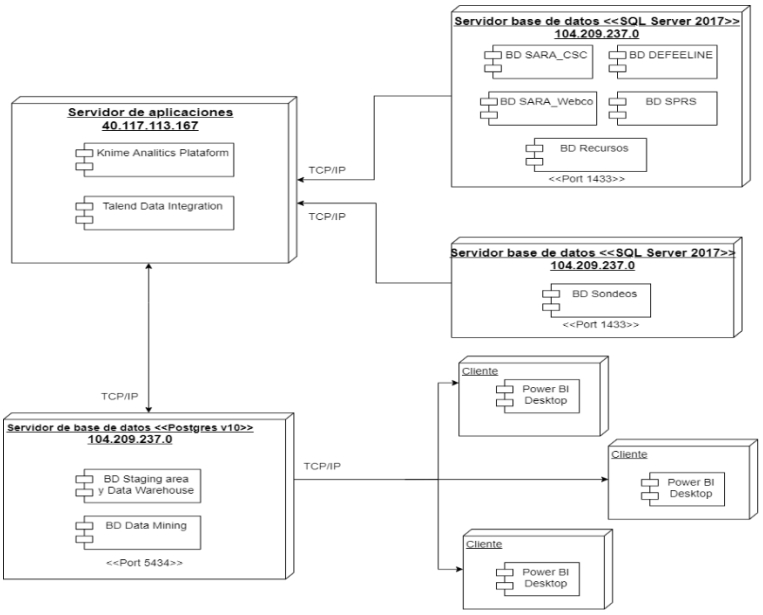


Figura 13 Diagrama de Despliegue.

12.2.1.1 Diagrama de Componentes Dominio.

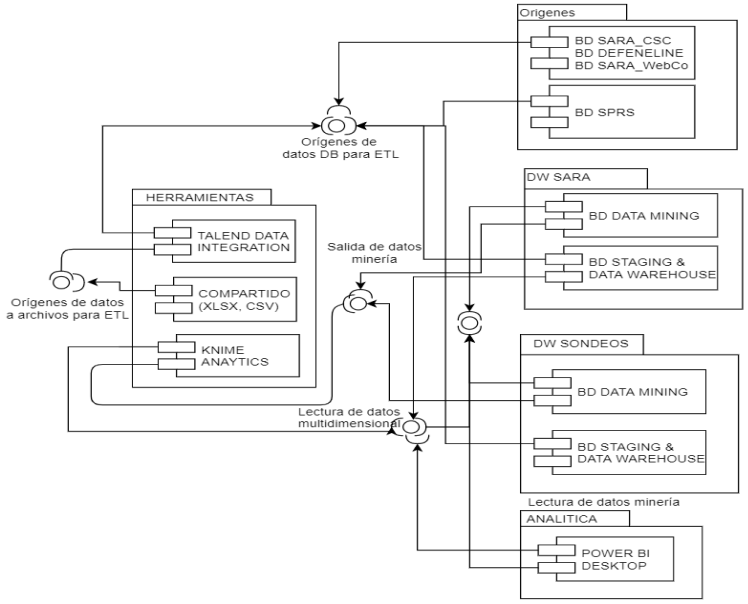


Figura 14 Diagrama de componentes del dominio.

13 Sprint 1

13.1 Descripción historias de usuario

Código	RSPP01
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea transformar los archivos que se tienen en formato Excel en bases de datos.
Razón	Para poder obtener los datos más rápidos y mejor resguardados.
Criterios de aceptación	Las estructuras de las tablas deberán estar normalizadas.
	Se deberán crear las tablas necesarias para normalizar de mejor manera posible, respetando la integridad de los datos.
	Las estructuras de la base de datos deberán respetar la consistencia de los datos.
Validación	Se comprobará que la base de datos este correctamente normalizada.
	Se comprobará que la normalización respeta la integridad de los datos.
	Se comprobará que la base de datos respeta la consistencia de los datos.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 10 Historia de Usuario RSPP01

Código	RSPP02
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea guardar la información de las bases de datos en estructuras de datos más robustas.
Razón	Para poder obtener de mejor manera los datos y poder generar información más rápido.
Criterios de aceptación	Se deberá crear un almacén de datos.
	Se deberá crear las dimensiones necesarias.
	Se deberá tener al menos una tabla de hechos.
Validación	Comprobar que el almacén de datos sea respetando un esquema como estrella, constelación o copo de nieve.
	Comprobar que el almacén de datos deberá tener las dimensiones necesarias como la dimensión tiempo, precios o productos.
	Comprobar que el almacén de datos deberá contener una tabla de hechos.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 11 Historia de Usuario RSPP02.

Código	RSPP03
Rol	Como técnico(a) UACM/ Jefatura
Funcionalidad	Se desea conocer un informe de las fluctuaciones y el comportamiento de los precios de granos básicos.
Razón	Para informar oportunamente a presidencia de la Defensoría y a la población.

Criterios de aceptación	Se debe considerar los distintos tipos de los granos arroz, frijol maíz y sorgo en sus variaciones.
	Se debe presentar su precio actual, variación diaria, semanal y mensual respecto al día seleccionado.
	La variación se debe expresar en términos monetarios como en forma porcentual.
Validación	Comprobar que se muestre información para los productos especificados.
	Comprobar que se muestre información en los lapsos de tiempo especificados.
	Comprobar que las variaciones se muestren en las unidades de medida especificadas.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 12 Historia de Usuario RSPP03.

Código	**RSPP04**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer un gráfico de variación semanal de precios de granos básicos.
Razón	Para informar oportunamente a presidencia de la Defensoría y a la población.
Criterios de aceptación	Debe considerar los distintos tipos de los granos arroz, frijol maíz y sorgo en sus variaciones.
	Deberá representar los días correspondientes a una semana mostrando el precio que cada producto poseía por cada día.
	Deberá identificar con distinto color los productos en el gráfico mostrando (x=fecha vs y=precio).
Validación	Comprobar que se muestre información para los productos especificados.
	Comprobar que se muestren los precios a lo largo de una semana.
	Comprobar que se diferencien los productos en el gráfico y que involucra las variables.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 13 Historia de Usuario RSPP04.

Código	**RSPP05**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Deseo conocer patrones de comportamiento y/o predicciones de los precios de los granos básicos.
Razón	Para que se identifique alzas inminentes en los precios de ese grupo de productos.
Criterios de aceptación	Se deberá mostrar la salida con los filtros iniciales para este grupo de precios.
	Se deberá presentar gráficos y filtros que representen la información descubierta.

	Se podrá manipular los parámetros respecto a las variables que se relacionan en la salida.
Validación	Comprobar que las salidas muestren la información sin filtrar.
	Comprobar que la información descubierta se represente mediante un informe de manera gráfica.
	Comprobar que se pueda filtrar la información mediante los parámetros que tiene a disposición.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 14 Historia de Usuario RSPP05.

Código	**RSPP06**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer la normalidad de los precios de productos que se proporcionan diariamente por el MAG (granos básicos).
Razón	Para poder asignar inspectores a la investigación en los sitios en específico para corroborar las situaciones.
Criterios de aceptación	Se deberá mostrar la salida con los filtros iniciales para este grupo de precios.
	Se deberá considerar los productos comprendidos en los granos básicos y plazas o sitios visitados.
	Se deberá presentar un indicador que calificará el comportamiento de los precios.
Validación	Comprobar que las salidas muestren la información sin filtrar.
	Comprobar que se representen las variables involucradas en el análisis.
	Validar que la calificación que se le da como comportamiento sea congruente con la información presentada.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 15 Historia de Usuario RSPP06.

13.2 Refinamiento del requerimiento de información

13.2.1 Proceso BPMN

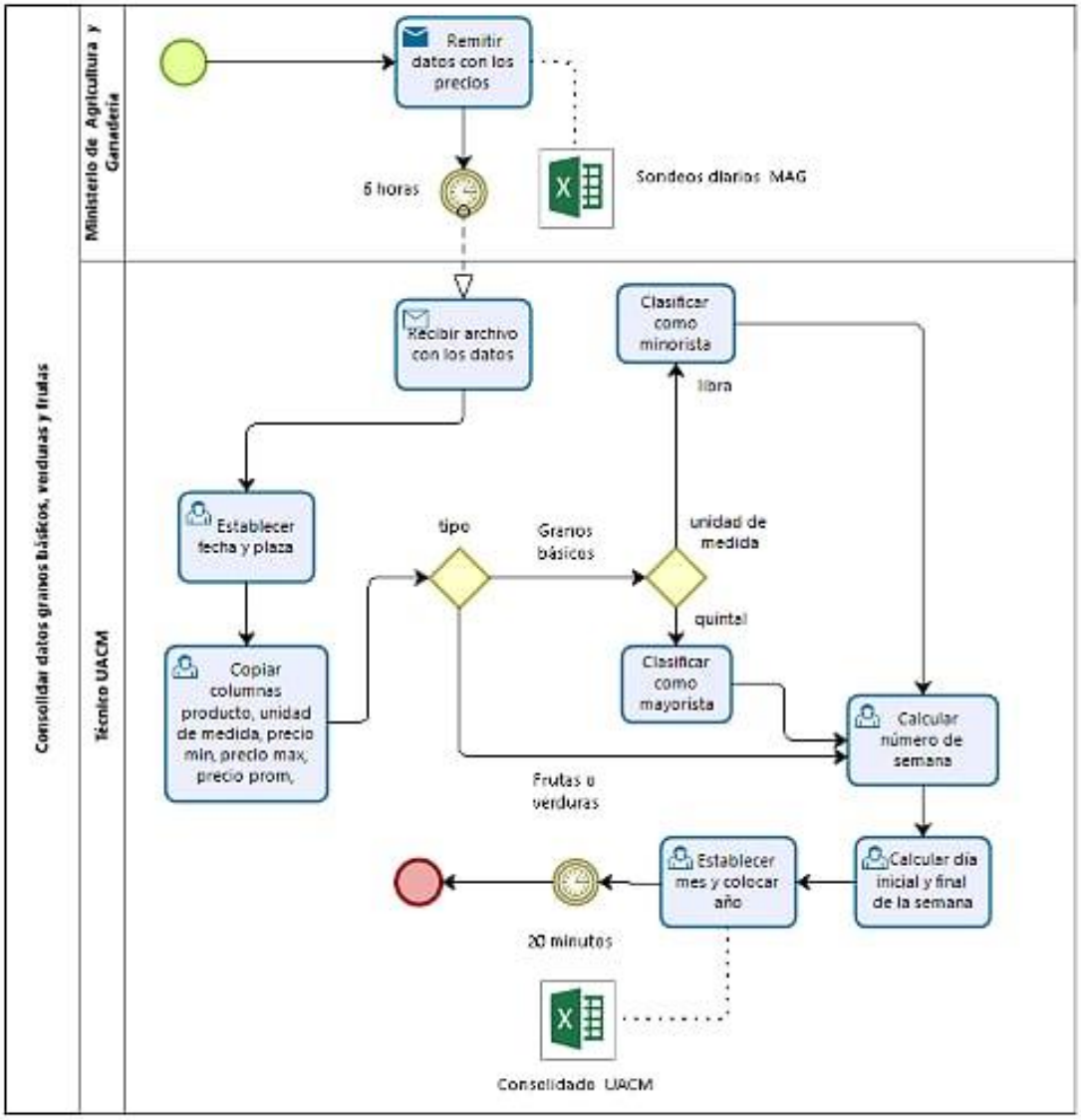


Figura 15 Proceso datos consolidados.

La Figura 15 muestra el diagrama BPMN del proceso actual que realiza la UACM para llevar a cabo el consolidado de precios enviados por el MAG.

13.2.2 Paquete de Información

Tema:		Variación diaría y semanal de granos básicos			
JERARQUIAS	Tiempo	Precios	Productos	Plaza	Unidad Medida
	Año	Precio minimo	Nombre del producto	Nombre de la plaza	Nombre de la unidad de medida
	Mes	Precio máximo			
	Semana	Precio Promedio			
	Dia				
Hechos Medidos:		Variación diaria (Medida calculada), Variación Semanal(Medida calculada)			

Figura 16 Paquete de información Variación diaria y semanal de granos básicos.

13.2.3 Casos de uso

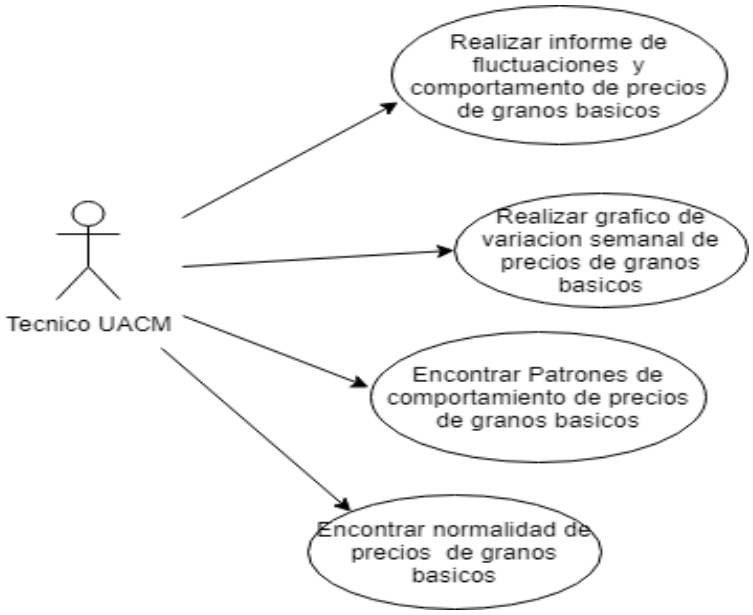


Figura 17 Diagrama de Casos de Uso Sondeo de Precios MAG.

13.3 Desarrollo de la iteración

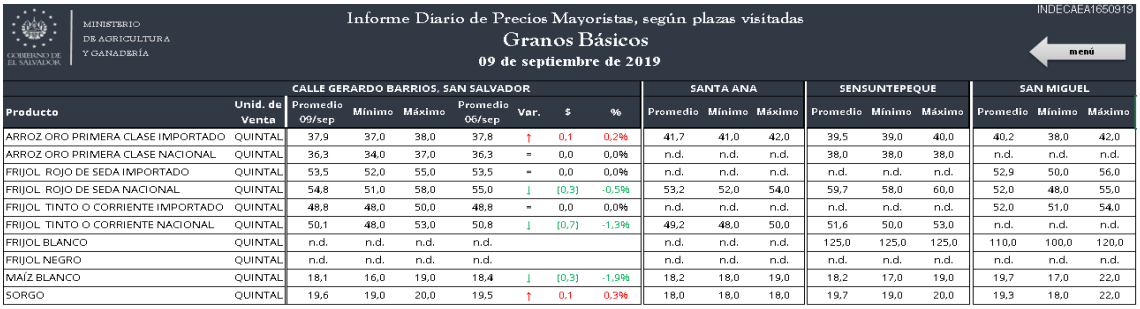
13.3.1 Integración de los datos

13.3.1.1 Extracción de los datos

Se tienen dos orígenes de datos para este sprint:

1- El informe diario de precios de productos agropecuarios el cual está disponible al público en el siguiente enlace: "http://www.mag.gob.sv/direccion-general-de-economia-agropecuaria/estadisticas-agropecuarias/informe-diario-de-precios-de-productos-agropecuarios/",

Y que además es remitido vía correo electrónico por el Ministerio de Ganadería y Agricultura a la Unidad de Análisis de Consumo y Mercados (UACM) diariamente.



MINISTERIO DE AGRICULTURA Y GANADERÍA — GOBIERNO DE EL SALVADOR

Informe Diario de Precios Mayoristas, según plazas visitadas

Granos Básicos

09 de septiembre de 2019

INDECAEA1650919

menú

		CALLE GERARDO BARRIOS, SAN SALVADOR							SANTA ANA			SENSUNTEPEQUE			SAN MIGUEL		
Producto	Unid. de Venta	Promedio 09/sep	Mínimo	Máximo	Promedio 06/sep	Var.	$	%	Promedio	Mínimo	Máximo	Promedio	Mínimo	Máximo	Promedio	Mínimo	Máximo
ARROZ ORO PRIMERA CLASE IMPORTADO	QUINTAL	37,9	37,0	38,0	37,8	↑	0,1	0,2%	41,7	41,0	42,0	39,5	39,0	40,0	40,2	38,0	42,0
ARROZ ORO PRIMERA CLASE NACIONAL	QUINTAL	36,3	34,0	37,0	36,3	=	0,0	0,0%	n.d.	n.d.	n.d.	38,0	38,0	38,0	n.d.	n.d.	n.d.
FRIJOL ROJO DE SEDA IMPORTADO	QUINTAL	53,5	52,0	55,0	53,5	=	0,0	0,0%	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	52,9	50,0	56,0
FRIJOL ROJO DE SEDA NACIONAL	QUINTAL	54,8	51,0	58,0	55,0	↓	(0,3)	-0,5%	53,2	52,0	54,0	59,7	58,0	60,0	52,0	48,0	55,0
FRIJOL TINTO O CORRIENTE IMPORTADO	QUINTAL	48,8	48,0	50,0	48,8	=	0,0	0,0%	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	52,0	51,0	54,0
FRIJOL TINTO O CORRIENTE NACIONAL	QUINTAL	50,1	48,0	53,0	50,8	↓	(0,7)	-1,3%	49,2	48,0	50,0	51,6	50,0	53,0	n.d.	n.d.	n.d.
FRIJOL BLANCO	QUINTAL	n.d.	n.d.	n.d.	n.d.				n.d.	n.d.	n.d.	125,0	125,0	125,0	110,0	100,0	120,0
FRIJOL NEGRO	QUINTAL	n.d.	n.d.	n.d.	n.d.				n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
MAÍZ BLANCO	QUINTAL	18,1	16,0	19,0	18,4	↓	(0,3)	-1,9%	18,2	18,0	19,0	18,2	17,0	19,0	19,7	17,0	22,0
SORGO	QUINTAL	19,6	19,0	20,0	19,5	↑	0,1	0,3%	18,0	18,0	18,0	19,7	19,0	20,0	19,3	18,0	22,0

Figura 18 Sección de precios mayoristas de los sondeos diarios del MAG.

MINISTERIO DE AGRICULTURA Y GANADERÍA — GOBIERNO DE EL SALVADOR

Informe Diario de Precios Minoristas, según plazas visitadas

Granos Básicos

09 de septiembre de 2019

INDECAEA1650919

menú

		CALLE GERARDO BARRIOS, SAN SALVADOR							SANTA ANA			SENSUNTEPEQUE			SAN MIGUEL		
Producto	Unid. de Venta	Promedio 09/sep	Mínimo	Máximo	Promedio 06/sep	Var.	$	%	Promedio	Mínimo	Máximo	Promedio	Mínimo	Máximo	Promedio	Mínimo	Máximo
ARROZ ORO PRIMERA CLASE IMPORTADO	LIBRA	0,45	0,45	0,45	0,45	=	0,00	0,00	0,50	0,50	0,50	0,49	0,40	0,50	0,49	0,40	0,50
ARROZ ORO PRIMERA CLASE NACIONAL	LIBRA	0,45	0,45	0,45	0,45	=	0,00	0,00	n.d.	n.d.	n.d.	0,50	0,50	0,50	n.d.	n.d.	n.d.
FRIJOL ROJO DE SEDA IMPORTADO	LIBRA	0,68	0,65	0,70	0,68	=	0,00	0,00	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,61	0,60	0,65
FRIJOL ROJO DE SEDA NACIONAL	LIBRA	0,68	0,65	0,75	0,68	↓	0,00	-0,01	0,65	0,65	0,65	0,83	0,75	0,90	0,63	0,60	0,65
FRIJOL TINTO O CORRIENTE IMPORTADO	LIBRA	0,58	0,55	0,60	0,58	=	0,00	0,00	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,62	0,60	0,65
FRIJOL TINTO O CORRIENTE NACIONAL	LIBRA	0,58	0,55	0,60	0,58	↓	-0,01	-0,01	0,61	0,60	0,65	0,71	0,70	0,75	0,65	0,65	0,65
FRIJOL BLANCO	LIBRA	1,00	1,00	1,00	1,00	=	0,00	0,00	n.d.	n.d.	n.d.	1,50	1,50	1,50	1,18	1,00	1,45
FRIJOL NEGRO	LIBRA	0,50	0,50	0,50	0,50	=	0,00	0,00	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,50	0,50	0,50
MAÍZ BLANCO	LIBRA	0,25	0,25	0,25	0,25	=	0,00	0,00	0,25	0,25	0,25	0,24	0,20	0,25	0,25	0,25	0,25
SORGO	LIBRA	0,25	0,25	0,25	0,25	=	0,00	0,00	0,25	0,25	0,25	0,26	0,25	0,30	0,25	0,25	0,25

Figura 19 Sección de precios minoristas de los sondeos diarios del MAG.

2- Documento de hoja de cálculo en el cual la UACM consolida los sondeos que día a día son remitidos por el MAG:

A	B	C	D	E	F	G
FECHA	PRODUCTO	PLAZA	TIPO	PRECIO PROMEDIO	PRECIO MÍNIMO	PRECIO MÁXIMO
03/06/2019	ARROZ ORO 1a. CLASE IMPORTADO	GERARDO BARRIOS	MAYORISTA	$28,78	$28,12	$28,88
03/06/2019	ARROZ ORO 1a. CLASE NACIONAL	GERARDO BARRIOS	MAYORISTA	$27,55	$25,84	$28,12
03/06/2019	FRÍJOL ROJO DE SEDA IMPORTADO	GERARDO BARRIOS	MAYORISTA	$37,62	$36,48	$38,00
03/06/2019	FRÍJOL ROJO DE SEDA NACIONAL	GERARDO BARRIOS	MAYORISTA	$40,66	$36,48	$41,80
03/06/2019	FRIJOL TINTO IMPORTADO	GERARDO BARRIOS	MAYORISTA	$35,53	$34,20	$36,48
03/06/2019	FRIJOL TINTO NACIONAL	GERARDO BARRIOS	MAYORISTA	$37,05	$33,44	$38,00
03/06/2019	MAÍZ BLANCO NACIONAL	GERARDO BARRIOS	MAYORISTA	$14,09	$13,68	$15,20
03/06/2019	MAICILLO	GERARDO BARRIOS	MAYORISTA	$12,35	$12,16	$12,92
03/06/2019	ARROZ ORO 1a. CLASE IMPORTADO	GERARDO BARRIOS	MINORISTA	$0,34	$0,34	$0,34
03/06/2019	ARROZ ORO 1a. CLASE NACIONAL	GERARDO BARRIOS	MINORISTA	$0,34	$0,34	$0,34
03/06/2019	FRÍJOL ROJO DE SEDA IMPORTADO	GERARDO BARRIOS	MINORISTA	$0,51	$0,49	$0,53
03/06/2019	FRÍJOL ROJO DE SEDA NACIONAL	GERARDO BARRIOS	MINORISTA	$0,53	$0,49	$0,57
03/06/2019	FRIJOL TINTO IMPORTADO	GERARDO BARRIOS	MINORISTA	$0,44	$0,42	$0,46
03/06/2019	FRIJOL TINTO NACIONAL	GERARDO BARRIOS	MINORISTA	$0,44	$0,42	$0,46
03/06/2019	MAÍZ BLANCO NACIONAL	GERARDO BARRIOS	MINORISTA	$0,19	$0,19	$0,19
03/06/2019	MAICILLO	GERARDO BARRIOS	MINORISTA	$0,19	$0,19	$0,19
03/06/2019	ARROZ ORO 1a. CLASE IMPORTADO	SENSUNTEPEQUE	MAYORISTA	$28,63	$28,12	$29,64
03/06/2019	FRÍJOL ROJO DE SEDA NACIONAL	SENSUNTEPEQUE	MAYORISTA	$41,33	$39,52	$41,80
03/06/2019	FRIJOL TINTO NACIONAL	SENSUNTEPEQUE	MAYORISTA	$37,43	$35,72	$38,00
03/06/2019	MAÍZ BLANCO NACIONAL	SENSUNTEPEQUE	MAYORISTA	$16,26	$15,20	$17,48
03/06/2019	MAICILLO	SENSUNTEPEQUE	MAYORISTA	$13,14	$12,16	$14,44
03/06/2019	ARROZ ORO 1a. CLASE IMPORTADO	SENSUNTEPEQUE	MINORISTA	$0,38	$0,38	$0,38
03/06/2019	FRÍJOL ROJO DE SEDA NACIONAL	SENSUNTEPEQUE	MINORISTA	$0,57	$0,57	$0,57
03/06/2019	FRIJOL TINTO NACIONAL	SENSUNTEPEQUE	MINORISTA	$0,53	$0,53	$0,53

Figura 20 Archivo de Excel consolidado de la UACM.

La extracción de los datos en la primera fuente de datos toma muchas consideraciones entre las cuales se encuentran:

1- Los datos son leídos como caracteres y se hacen las conversiones necesarias en el integrador de datos para evitar trabajo manual.
2- Los datos no son homogéneos por lo tanto dicha homogenización se realiza desde el integrador de datos.
3- Las plazas visitadas se muestran horizontalmente formando una matriz con los precios de los sondeos de los productos por lo tanto es necesario realizar un proceso “unpivot” y asignar cada grupo de precios su plaza correspondiente.
4- Es necesario convertir la fecha al formato que la base de datos la puede capturar.
5- La variación no se almacenará si no que se calculará posteriormente en los informes.
6- Existen sondeos de precios que no se realizan por lo tanto aparecen en cero, estos no son tomados en cuenta.
7- Algunos nombres de productos no coinciden con los que se manejan en el consolidado por lo tanto es necesario homogenizar.
8- Es necesario validar en la base de datos si ya existe el producto, plaza o unidad de medida para no repetir en los catálogos.

13.3.1.2 Staging área

13.3.1.2.1 Diseño del modelo staging área

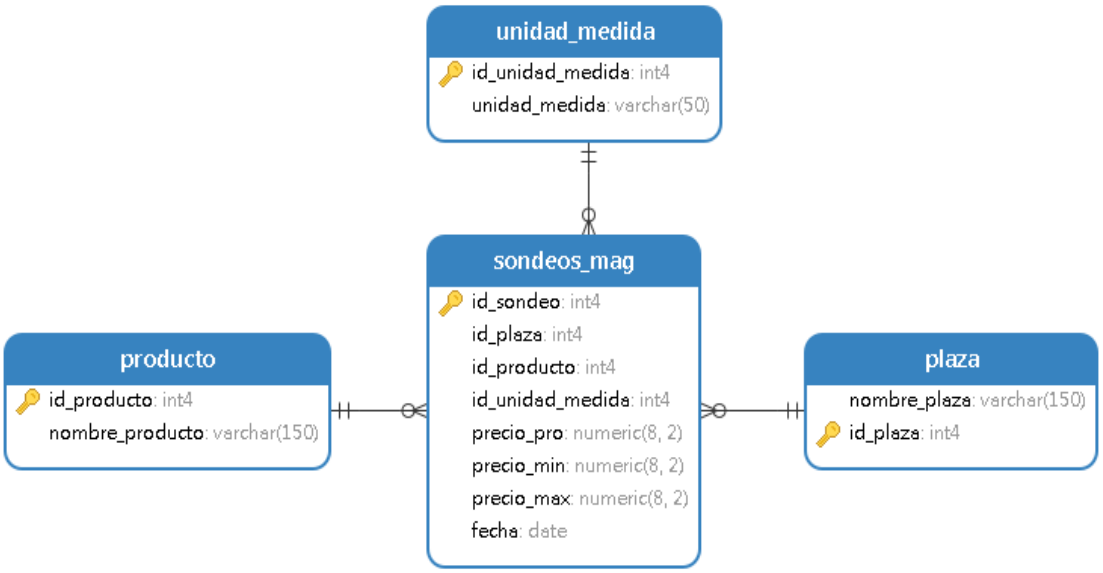
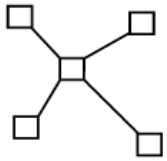


Figura 21 Diseño de modelo relacional (staging área).

13.3.1.3 Modelo multidimensional

13.3.1.3.1 Diseño Conceptual del Data mart (UML)

Especialización: Analizar los sondeos de precios del MAG.

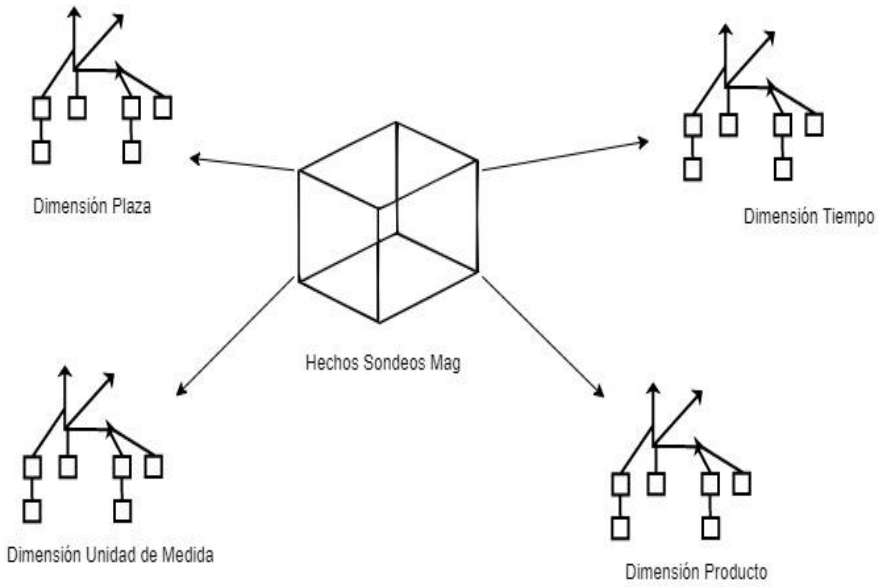


Nivel 1: Definición del modelo

Paquete Hechos:
Sondeos MAG.

Paquete Dimensiones:

1. Tiempo
2. Plaza
3. Unidad de medida
4. Producto



Nivel 2: Definición del esquema estrella.

Figura 22 Niveles 1 y 2 del diseño conceptual

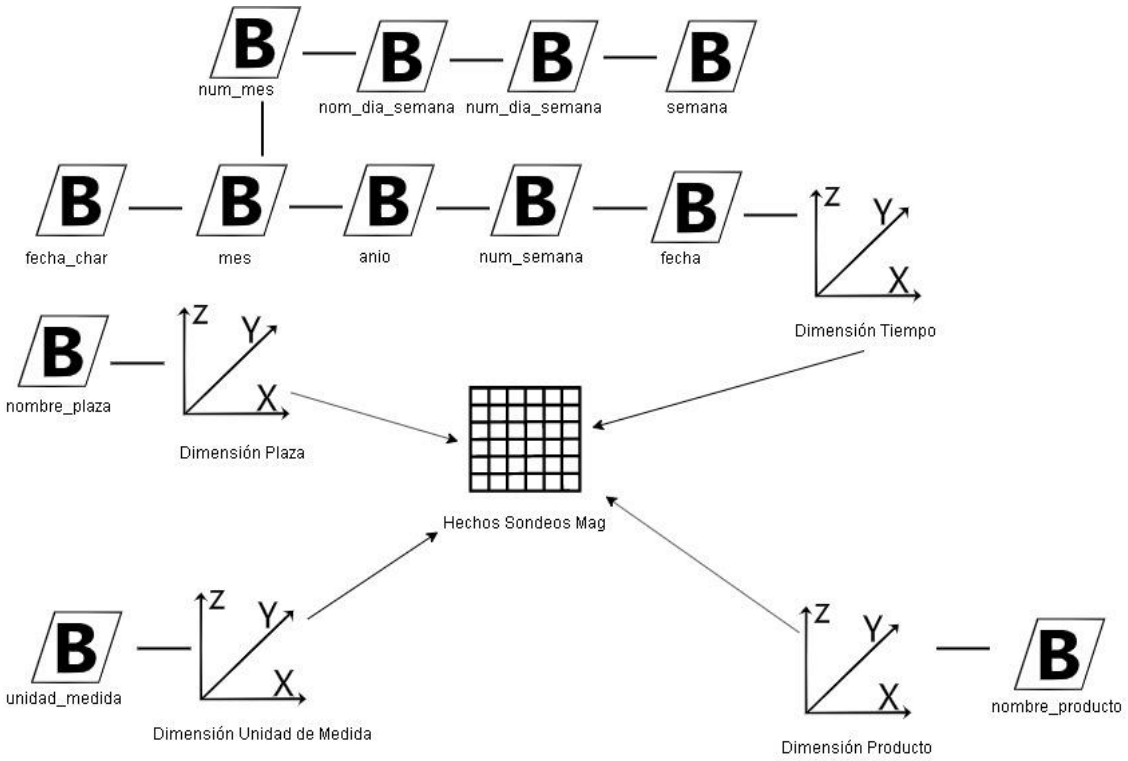


Figura 23 Nivel 3: Dimensiones/Hechos.

13.3.1.3.2 Diseño Físico Data Mart

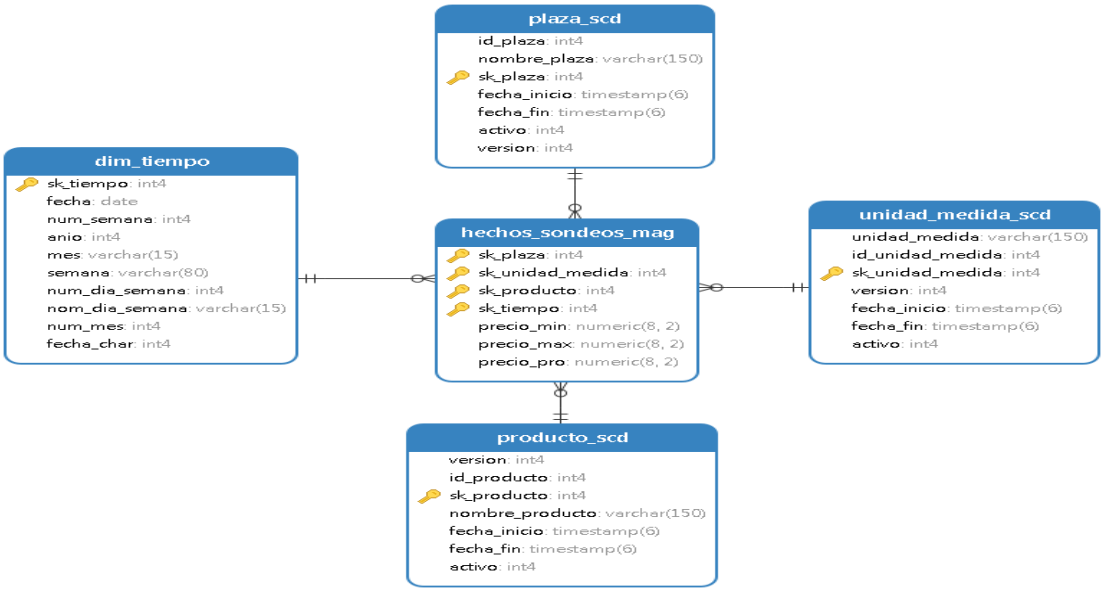


Figura 24 Modelo multidimensional data mart MAG.

13.3.1.3.3 Diseño de procesos ETL

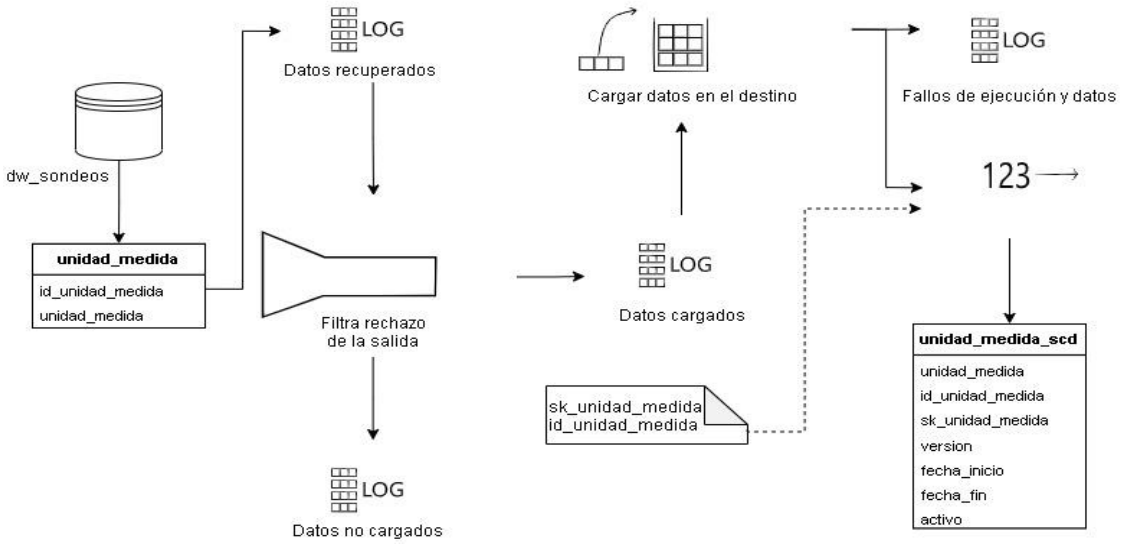


Figura 25 Dimensión unidad de medida.

La Figura 25 muestra uno de los diseños de los procesos ETL, el cual sigue los siguientes pasos:

1. Conectar a la base de datos de origen en este caso el staging_area
2. Unir las tablas que conformarán la dimensión unidad de medida.
3. Reportar al Log los datos recuperados, Filtrar y reportar datos cargados y no cargados
4. Calcular clave sustituta y cargar datos a la tabla de destino unidad_medida_scd.

13.3.1.3.4 Desarrollo de procesos ETL

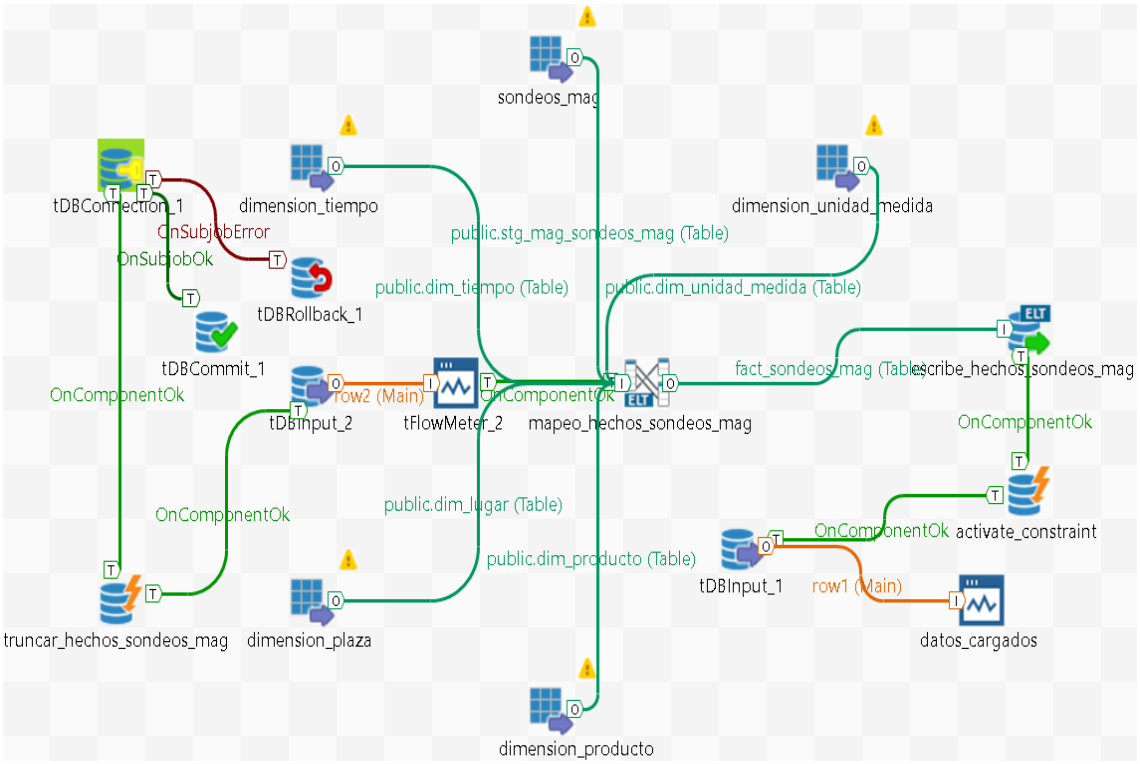


Figura 26 Job hechos_mag

La Figura 26 muestra el desarrollo de uno de los flujos de trabajo en este caso para el hecho a medir en el presente sprint en el cual se sigue el proceso enumerado a continuación:

1. Se establece la conexión con la base de datos.
2. Se truncan los datos de la tabla de hechos y se desactiva la integridad referencial.
3. Se reportan al log los datos recuperados de la tabla de normalización del staging área.
4. Se unen los datos de las dimensiones con la tabla de normalización del staging área para determinar las llaves sustitutas en la tabla de hechos.
5. Se almacenan los datos en la tabla de hechos y se activa nuevamente la integridad referencial.
6. Se reportan al log los datos almacenados en la tabla de hechos.

13.3.1.3.5 Pruebas

```
.-----------------+-------------+----------------+-------------------------------------------------------------------------.
|                                                  MAG_STAGING                                                            |
|=----------------+-------------+----------------+-------------------------------------------------------------------------=|
|datos_recuperados|datos_cargados|datos_no_cargados|criterios                                                               |
|=----------------+-------------+----------------+-------------------------------------------------------------------------=|
|52483            |51749        |734             |precio max, min y pro <> 0 y productos diferentes de frijol blanco y negro|
'-----------------+-------------+----------------+-------------------------------------------------------------------------'

13320 milliseconds
[statistics] disconnected
2020-01-26 23:42:31
.-----------------+-------------+----------------+-------------------------------------------------------------------------.
|                                                  MAG_STAGING                                                            |
|=----------------+-------------+----------------+-------------------------------------------------------------------------=|
|datos_recuperados|datos_cargados|datos_no_cargados|criterios                                                               |
|=----------------+-------------+----------------+-------------------------------------------------------------------------=|
|0                |0            |0               |precio max, min y pro <> 0 y productos diferentes de frijol blanco y negro|
'-----------------+-------------+----------------+-------------------------------------------------------------------------'

7669 milliseconds
[statistics] disconnected
```

Figura 27 Pruebas de ejecución al job de Staging MAG.

La Figura 27 presenta los resultados para dos ejecuciones del Job, en el cual se constatan la cantidad de registros recuperados del origen, cargados y no cargados en el destino, y en la segunda ejecución se valida el correcto funcionamiento del Job ya que no carga datos repetidos, así como el correcto funcionamiento del registro de Log ya que presenta cantidades correctas de registros cargados.

```
[statistics] disconnected
2020-02-02 10:14:28
.-----------------+--------------+-----------------+-----------------------------------------------------
|                                                                    DW_SONDEO_MAG
|=----------------+--------------+-----------------+-----------------------------------------------------
|datos_recuperados|datos_cargados|datos_no_cargados|criterios
|=----------------+--------------+-----------------+-----------------------------------------------------
|51800            |51783         |17               |Para los sondeos de productos en los cuales se cumple
'-----------------+--------------+-----------------+-----------------------------------------------------

1266 milliseconds
[statistics] disconnected
2020-02-02 12:11:15
.-----------------+--------------+-----------------+-----------------------------------------------------
|                                                                    DW_SONDEO_MAG
|=----------------+--------------+-----------------+-----------------------------------------------------
|datos_recuperados|datos_cargados|datos_no_cargados|criterios
|=----------------+--------------+-----------------+-----------------------------------------------------
|51800            |51783         |17               |Para los sondeos de productos en los cuales se cumple
'-----------------+--------------+-----------------+-----------------------------------------------------

1053 milliseconds
[statistics] disconnected
```

Figura 28 Pruebas de ejecución al job de fact_sondeos_mag

La Figura 28 muestra el log generado después de la ejecución del Job en el cual se muestra la cantidad de datos que fueron recuperados desde el Staging área, los que fueron cargados al

modelo multidimensional y los que no fueron cargados, cabe destacar que en la figura se muestra para dos ejecuciones en las cuales las cantidades son las mismas esto debido a que el tipo de carga se realiza es total, truncando la tabla de hechos y luego insertando los datos correspondientes.

sondeos_mag

	id_sondeo	id_plaza	id_producto	id_unidad_medida	precio_pro	precio_min	precio_max	fecha	fecha_carga
1	1	355	308	13	80	80	80	2015-08-19	2020-04-13 00:27:07
2	2	355	302	13	76.5	75	78	2015-08-19	2020-04-13 00:27:07
3	3	355	308	13	83.5	82	85	2015-08-26	2020-04-13 00:27:07
4	4	355	308	13	80	80	80	2015-09-02	2020-04-13 00:27:07
5	5	355	302	13	74	73	75	2015-09-02	2020-04-13 00:27:07
6	6	355	301	13	20	20	20	2015-09-02	2020-04-13 00:27:07
7	7	355	308	13	55.25	55	56	2017-01-18	2020-04-13 00:27:07
8	8	355	302	13	47	45	48	2017-01-18	2020-04-13 00:27:07
9	9	355	301	13	12.25	12	13	2017-01-18	2020-04-13 00:27:07
10	10	355	300	13	14.38	14	15	2017-01-18	2020-04-13 00:27:07
11	11	355	308	14	0.65	0.65	0.65	2017-01-18	2020-04-13 00:27:07
12	12	355	302	14	0.6	0.6	0.6	2017-01-18	2020-04-13 00:27:07
13	13	355	301	14	0.16	0.15	0.17	2017-01-18	2020-04-13 00:27:07
14	14	355	300	14	0.19	0.17	0.2	2017-01-18	2020-04-13 00:27:07

hechos_sondeos_mag

	precio_max	precio_min	precio_pro	sk_plaza	sk_unidad_medida	sk_producto	sk_tiempo
1	24	17	19.11	11,612	477	6,062	111
2	20	16	18.27	11,612	477	6,062	112
3	20	16	17.06	11,612	477	6,062	162
4	18	17	17.13	11,612	477	6,062	226
5	23	19	19.72	11,612	477	6,062	252
6	17	15	16	11,612	477	6,062	260
7	19	17.5	18	11,612	477	6,062	279
8	24	23	23.2	11,612	477	6,062	300
9	18	15	15.75	11,612	477	6,062	302
10	20	17	17.67	11,612	477	6,062	303
11	19	16.5	17.17	11,612	477	6,062	305
12	20	17	17.61	11,612	477	6,062	308
13	24	21.5	22.61	11,612	477	6,062	311
14	18	15	16.5	11,612	477	6,062	313

Figura 29 Muestra de datos para 2 tablas

13.3.2 Minería de datos

13.3.2.1 Exploración de los datos

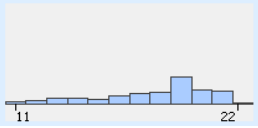
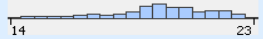
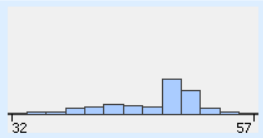
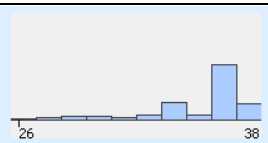
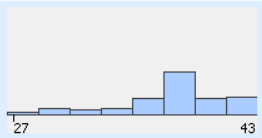
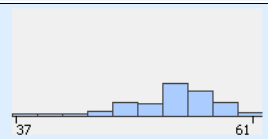
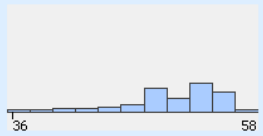
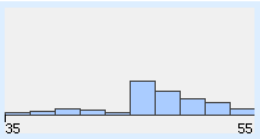
PRODUCTO	Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
MAICILLO	precio	11,48	18,3396	?	22,25	2,495	-0,8166	-0,1947	0	0	0	
MAIZ B. N	precio	14,09	19,6649	?	23,27	1,8927	-0,7343	0,3393	0	0	0	
FRIJOL TINTO N	precio	31,92	46,7176	?	57	4,5842	-0,8684	0,3535	0	0	0	
ARROZ ORO 1a. CLASE N	precio	25,93	34,86	?	38	2,5501	-1,6181	2,3325	0	0	0	
ARROZ ORO 1a. CLASE I	precio	27,42	37,4975	?	43	3,1652	-0,8889	0,8197	0	0	0	
FRIJOL ROJO DE SEDA N	precio	36,94	52,1838	?	60,57	4,5747	-0,956	0,8735	0	0	0	
FRIJOL ROJO DE SEDA I	precio	36,48	50,37	?	58	4,1109	-1,1228	1,1977	0	0	0	
FRIJOL TINTO I	precio	35,02	47,1095	?	55	3,9717	-0,8917	0,8376	0	0	0	

Tabla 16 Tabla comparativa de estadísticas por producto.

En la Tabla 16 se presenta un comparativo de estadísticas por producto salida generada desde la herramienta Knime Analytics se muestran estadísticos como mínimo, máximo, media, desviación estándar, varianza, mediana, suma total, número de valores faltantes y recuento de filas en todas las columnas numéricas, y el recuento todos los valores nominales junto con sus ocurrencias, tomando como base precios de los productos mayorista y promedio nacional (Sin desagregar por plaza).

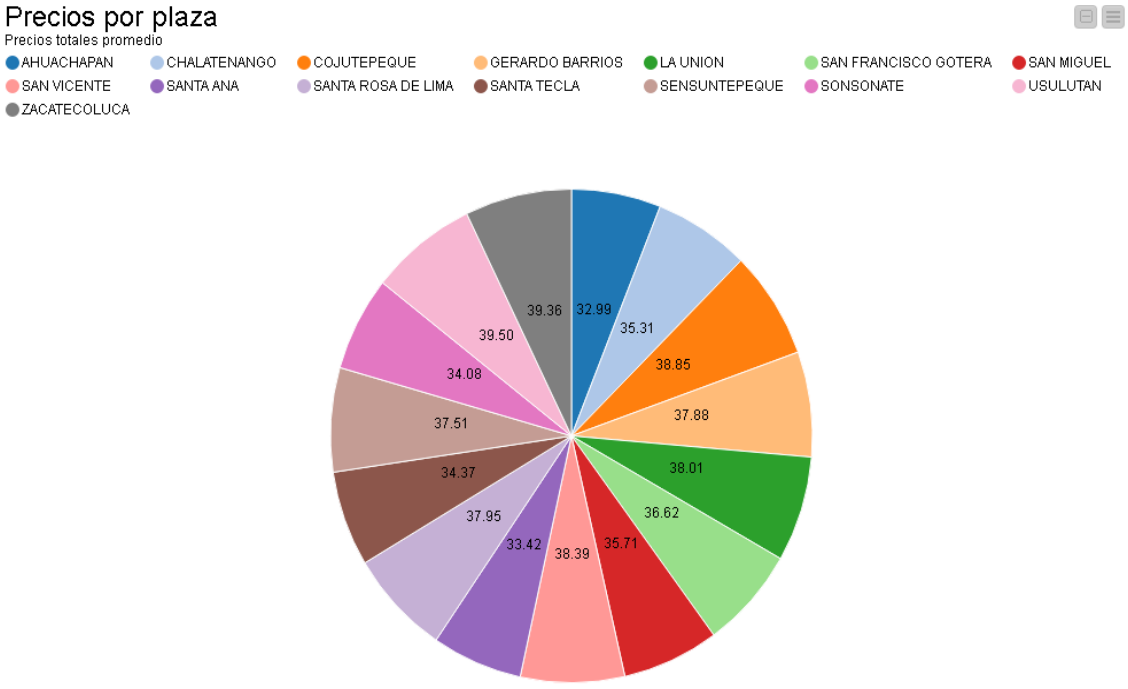


Figura 30 Gráfico de precio promedio por plaza.

Se calcula el precio promedio de los productos de granos básicos agrupados por las 15 plazas visitadas por el Ministerio de Agricultura y Ganadería, obteniéndose que en general los precios son más altos en Usulután.

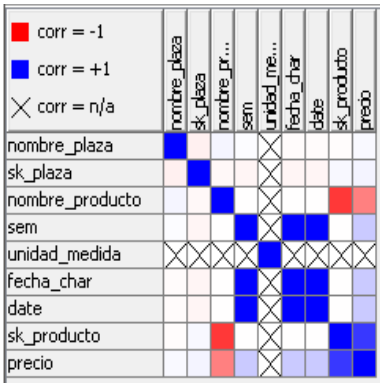


Figura 31 Correlación existente entre las variables.

La Figura 31 muestra la correlación existente entre variables para las cuales el color azul destaca alta correlación el rojo menor correlación y la “X” una correlación totalmente inexistente.

13.3.2.2 Técnica de asociación

13.3.2.2.1 Descubrir la relación entre los precios de los productos

13.3.2.2.1.1 Contenido del caso.

N° de caso: C-ASO-01	
Técnica	Asociación.
Algoritmos	A priori, FP Growth, Association Rule Learner.
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019.
Variables	Se analiza el precio promedio por producto, tomando cada día como una transacción.
Hipótesis	Descubrir la relación entre los precios de los productos.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Reglas de asociación.
Interpretación de resultados	Evaluar en base a indicadores de rendimiento (Support, Confidence, Lift).
Herramienta de software	Knime Analytics Platform.

Tabla 17 Contenido del caso N° C-ASO-01.

13.3.2.2.1.2 Población

SQL Statement

```
select * from crosstab(
  ' SELECT fecha, sk_producto,
      CASE
    WHEN aux2.precio>(aux2.promedio+aux2.promedio*0.05) THEN aux2.sk_producto
    ELSE NULL
  END
  AS esmayor
FROM
    (SELECT
    (SELECT AVG(p.precio_max)
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
    JOIN producto_scd pr ON pr.sk_producto = h.sk_producto
    WHERE pr.sk_producto=aux.sk_producto
    AND sk_unidad_medida=19
    GROUP BY pr.sk_producto
    )promedio, fecha,aux.sk_producto, precio
    FROM
    (SELECT t.fecha, pr.sk_producto,AVG(precio_max) precio
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
    JOIN producto_scd pr ON pr.sk_producto = h.sk_producto
    JOIN dim_tiempo t ON t.sk_tiempo = h.sk_tiempo
    WHERE sk_unidad_medida=19
    GROUP BY 1,2
    ORDER BY 1,2) aux)aux2'
) as (
  fecha DATE,
  "91" INT,
  "92" INT,
```

Preview results: Evaluate

Row ID	fecha	91	92	93	94	95	96	97	98
Row0	2018-09-03	91	92	93	?	?	?	?	?
Row1	2018-09-04	91	92	93	?	95	?	?	98
Row2	2018-09-05	91	92	?	?	?	?	?	?
Row3	2018-09-06	91	92	93	?	?	?	?	98
Row4	2018-09-07	91	92	93	?	?	?	97	?
Row5	2018-09-10	91	92	93	?	?	?	?	?
Row6	2018-09-11	91	92	?	?	?	?	97	?
Row7	2018-09-12	91	92	?	?	?	?	?	?
Row8	2018-09-13	91	92	93	?	?	?	?	98
Row9	2018-09-14	91	92	93	?	?	?	97	?

Figura 32 Set de datos para el caso C-ASO-01.

La Figura 32 muestra la población de datos utilizada, es proveniente del modelo multidimensional y corresponde a una consulta de tipo PIVOTE para los productos cuyo precio promedio es mayor que el promedio de los precios de todo el universo de datos para el mismo producto más el 5% distribuido por fecha, en pocas palabras se simula una transacción de super mercado en la tenemos un conjunto de productos, con un precio.

13.3.2.2.1.3 Variables

Las variables que se ven involucradas en la exploración son el producto, el precio del producto y la fecha, para el caso del precio se utiliza para validar si hubo un aumento respecto al promedio del precio del mismo producto.

13.3.2.2.1.4 Hipótesis

El objetivo de minería de datos que se quiere llevar a cabo o la hipótesis que se desea comprobar es: “Explorar la existencia de una relación en el alza de los precios de los productos de granos básicos, y descubrir cuáles de ellos son los que marcan tendencia”.

13.3.2.2.1.5 Procedimiento

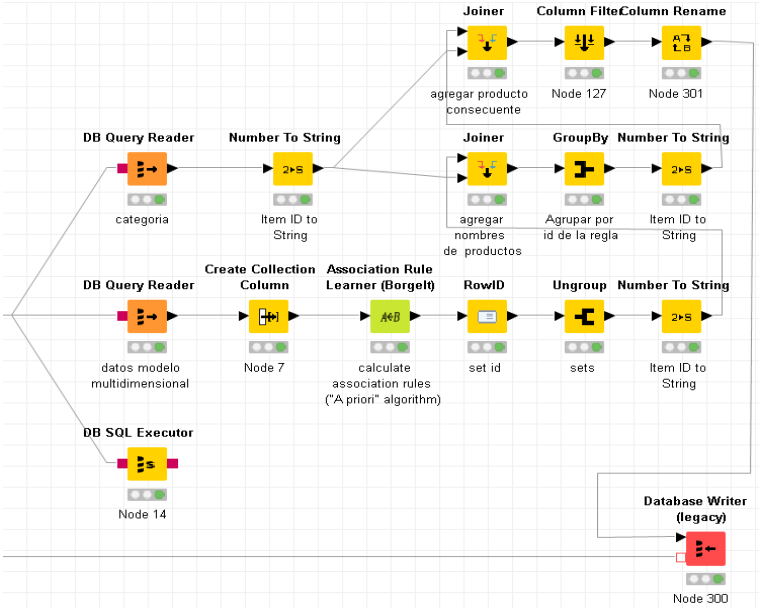


Figura 33 Flujo de trabajo C-ASO-01.

La Figura 33 muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-ASO-01 se resumen a continuación las acciones que se llevan a cabo:

- **DB Connector**: Establece la conexión con la base de datos Postgres para acceder a los datos del modelo multidimensional.
- **DB SQL Executor**: Ejecuta el script para crear la extensión tablefunc que permite utilizar funciones de tabla y el caso se necesita utilizar crosstab para pivotear los datos que se leen del modelo multidimensional.
- **DB Query Reader (datos del modelo multidimensional):** Obtiene los datos de las transacciones a ser exploradas con el algoritmo de minería de datos.
- **DB Query Reader (categoria):** Obtiene el catálogo con la categorización de las transacciones en este caso a que nombre de producto corresponde cada id.
- **Create collection column:** Lee un set de datos y crea una columna de colección de varias columnas que se le especifiquen en la configuración, teniendo la posibilidad de obviar valores faltantes.
- **Double configuration:** Se utiliza este nodo para definir el valor para el soporte mínimo del ítem set frecuentes de la minería.

- **Association Rule Learner (Borgelt):** Aplica el algoritmo A priori en busca de reglas de asociación sobre la columna colección que se creó anteriormente, definiendo en la configuración de este el valor mínimo que se desea para la confianza de las reglas, como salidas de interés a parte de las reglas de asociación devuelve el Soporte del ítem set, la confidencia de cada regla y su respectivo Lift (La probabilidad de que se dé el consecuente dado que se dio el antecedente).
- **RowID:** Asigna un identificador de registro a cada regla encontrada.
- **Ungroup:** Desagrupa la columna de Antecedente, obteniendo como resultado un registro por cada ítem contenido en el ítem set de antecedente de cada registro, Esto con la finalidad de obtener la categoría (Nombre del producto) asociado al identificador del ítem.
- **Number To String:** Es necesario convertir las columnas de antecedente (Set de datos de transacciones) y sk_producto (Set de datos de categoria) a Cadena de caracteres previó a realizar la unión que devuelva el nombre de la categoría.
- **Joiner:** Realiza la unión entre antecedente/consecuente con la categoría para obtener su nombre.
- **GroupBy:** Se agrupan los datos del flujo de trabajo agregándose por la columna de Identificador de la regla de modo que las reglas de asociación con sus respectivos antecedentes y consecuentes volverán a la forma de colección ahora presentando los nombres de las categorías en lugar de los identificadores para cada ítem del ítem set.

13.3.2.2.1.6 Resultados

Soporte	Confianza	Lift	Antecedente	Consecuente
108	90.8	5.9082	Frijol Tinto Nacional, Frijol Tinto Importado, Frijol Rojo De Seda Importado	Frijol Rojo De Seda Nacional
108	90	5.8589	Maíz Blanco Nacional, Frijol Tinto Nacional, Frijol Tinto Importado	Frijol Rojo De Seda Nacional
113	85.6	3.8779	Maíz Blanco Nacional, Frijol Tinto Nacional, Frijol Rojo De Seda Nacional	Frijol Rojo De Seda Importado
113	88.3	5.7471	Maíz Blanco Nacional, Frijol Tinto Nacional, Frijol Rojo De Seda Importado	Frijol Rojo De Seda Nacional
116	97.5	1.7582	Frijol Tinto Nacional, Frijol Tinto Importado, Frijol Rojo De Seda Importado	Maíz Blanco Nacional
116	96.7	4.379	Maíz Blanco Nacional, Frijol Tinto Nacional, Frijol Tinto Importado	Frijol Rojo De Seda Importado
116	90.6	4.4992	Maíz Blanco Nacional, Frijol Tinto Nacional, Frijol Rojo De Seda Importado	Frijol Tinto Importado

Figura 34 Resultados obtenidos para el caso C-ASO-01.

La Figura 34 muestra los resultados obtenidos mediante el flujo de trabajo detallado anteriormente.

13.3.2.2.1.7 Interpretación de resultados

Cada regla de asociación se interpreta de la siguiente manera, si ocurrieron los ítems del ítem set antecedente sucederá el consecuente con una de confianza de "valor de confianza", el soporte es la frecuencia relativa del ítem set contiene al ítem set antecedente conjunto con el ítem consecuente, para el caso del Lift es la probabilidad condicional de que se dé el consecuente dado que se dio el antecedente: aplicándolo a una de las reglas tenemos: Si los productos: MAIZ BLANCO NACIONAL, FRIJOL TINTO NACIONAL, FRIJOL ROJO DE SEDA NACIONAL. Aumentan de precio, también aumentará de precio el producto FRIJOL ROJO DE SEDA IMPORTADO con una confianza del 85.6%, el ítem set que contiene a los productos en cuestión (antecedente, consecuente en conjunto) aparece 113 veces en la población de datos, además el FRIJOL ROJO DE SEDA IMPORTADO aumenta su probabilidad de aumentar de precio en un 388% si los productos del antecedente aumentaron de precio, por consiguiente, FRIJOL ROJO DE SEDA IMPORTADO NACIONAL tiene menor probabilidad de aumentar de precio por sí solo.

13.3.2.2.2 Descubrir la relación entre los precios por plaza

13.3.2.2.2.1 Contenido del caso

N° de caso: C-ASO-02	
Técnica	Asociación
Algoritmos	A priori
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019
Variables	Se analiza el precio promedio por plaza, tomando cada día como una transacción
Hipótesis	Descubrir la relación entre los precios por plaza.
Procedimiento	Flujo de trabajo en Knime
Resultados	Reglas de asociación
Interpretación de resultados	Evaluar en base a indicadores de rendimiento (Support, Confidence, Lift)
Herramienta de software	Knime Analytics Platform

Figura 35 Contenido del caso N° C-ASO-02.

La Figura 35 muestra el contenido que se aborda para para el caso C-ASO-02.

13.3.2.2.2.2 Población

SQL Statement

```
select * from crosstab(
  ' SELECT semana, sk_plaza,
       CASE
    WHEN aux2.precio>(aux2.promedio+aux2.promedio*0.05) THEN aux2.sk_plaza
    ELSE NULL
  END
  AS esmayor
FROM
    (SELECT
    (SELECT AVG(p.precio_max)
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
```

Preview results: Evaluate

Row ID	semana	189	190	191	192
Row0	Semana del 2015-08-17 al 2015-08-21	189	191	193	195
Row1	Semana del 2015-08-24 al 2015-08-28	189	191	192	193
Row2	Semana del 2015-08-31 al 2015-09-04	189	191	192	193
Row3	Semana del 2015-09-07 al 2015-09-11	189	190	191	192
Row4	Semana del 2015-09-14 al 2015-09-18	189	191	192	193
Row5	Semana del 2015-09-21 al 2015-09-25	189	192	193	194
Row6	Semana del 2015-09-28 al 2015-10-02	189	191	193	194
Row7	Semana del 2015-10-05 al 2015-10-09	189	190	191	193
Row8	Semana del 2015-10-12 al 2015-10-16	189	191	192	?
Row9	Semana del 2015-10-19 al 2015-10-23	189	193	194	195

Figura 36 Selección de datos extraídos del modelo multidimensional desde Knime.

La población de datos utilizada, es proveniente del modelo multidimensional y corresponde a una consulta de tipo PIVOTE para los productos cuyo precio promedio es mayor que el promedio de los precios de todo el universo de datos para el mismo producto más el 5% distribuido por semana agrupado por plaza.

13.3.2.2.2.3 Variables

Las variables que se ven involucradas en la exploración son Las plazas, El precio del producto, La fecha, en el caso del precio se utiliza para validar si hubo un aumento respecto al promedio del precio de la misma plaza esto se hace para cada semana.

13.3.2.2.2.4 Hipótesis

El objetivo de minería de datos que se quiere llevar a cabo o la hipótesis que se desea comprobar es: “Explorar la existencia de una relación en el alza de los precios de los productos de granos básicos en las diferentes plazas, y descubrir cuáles de ellas son las que marcan tendencia”.

13.3.2.2.2.5 Procedimiento

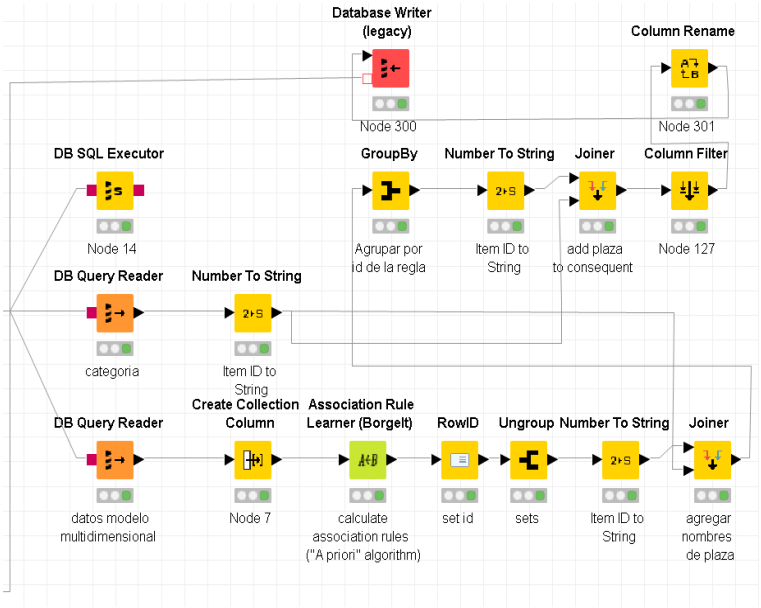


Figura 37 Flujo de trabajo C-ASO-02.

En la Figura 37 se muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-ASO-02 se resumen a continuación las acciones que se llevan a cabo:

- **DB Connector**: Establece la conexión con la base de datos Postgres para acceder a los datos del modelo multidimensional.
- **DB SQL Executor**: Ejecuta el script para crear la extensión tablefunc que permite utilizar funciones de tabla y el caso se necesita utilizar crosstab para pivotear los datos que se leen del modelo multidimensional.
- **DB Query Reader (datos del modelo multidimensional):** Obtiene los datos de las transacciones a ser exploradas con el algoritmo de minería de datos.
- **DB Query Reader (categoria):** Obtiene el catálogo con la categorización de las transacciones en este caso a que nombre de plaza corresponde cada id.
- **Create collection column:** Lee un set de datos y crea una columna de colección de varias columnas que se le especifiquen en la configuración, teniendo la posibilidad de obviar valores faltantes.
- **Double configuration:** Se utiliza este nodo para definir el valor para el soporte mínimo del ítem set frecuentes de la minería.
- **Association Rule Learner (Borgelt):** Aplica el algoritmo A priori en busca de reglas de asociación sobre la columna colección que se creó anteriormente, definiendo en la configuración de este el valor mínimo que se desea para la confianza de las reglas, como salidas de interés a parte de las reglas de asociación devuelve el Soporte del ítem set, la confidencia de cada regla y su respectivo Lift (La probabilidad de que se dé el consecuente dado que se dio el antecedente).
- **RowID:** Asigna un identificador de registro a cada regla encontrada.
- **Ungroup:** Desagrupa la columna de Antecedente, obteniendo como resultado un registro por cada ítem contenido en el ítem set de antecedente de cada registro, Esto

con la finalidad de obtener la categoría (Nombre de la plaza) asociado al identificador del ítem.

- **Number To String:** Es necesario convertir las columnas de antecedente (Set de datos de transacciones) y sk_plaza (Set de datos de categoria) a Cadena de caracteres previó a realizar la unión que devuelva el nombre de la categoría.
- **Joiner:** Realiza la unión entre antecedente/consecuente con la categoría para obtener su nombre.
- **GroupBy:** Se agrupan los datos del flujo de trabajo agregándose por la columna de Identificador de la regla de modo que las reglas de asociación con sus respectivos antecedentes y consecuentes volverán a la forma de colección ahora presentando los nombres de las categorías en lugar de los identificadores para cada ítem del ítem set.

13.3.2.2.2.6 Resultados

Soporte	Confianza	Lift	Antecedente	Consecuente
27	96.4	2.6173	San Francisco Gotera, Santa Rosa De Lima, Zacatecoluca, San Miguel	Gerardo Barrios
27	96.4	3.2506	Santa Rosa De Lima, Zacatecoluca, Gerardo Barrios, San Miguel	San Francisco Gotera
27	96.4	2.6173	San Francisco Gotera, Zacatecoluca, La Union, San Miguel	Gerardo Barrios
27	96.4	2.6173	San Francisco Gotera, Zacatecoluca, San Miguel, Usulutan	Gerardo Barrios
31	96.9	2.6295	San Francisco Gotera, Santa Rosa De Lima, La Unión, San Miguel	Gerardo Barrios
31	96.9	2.6295	San Francisco Gotera, Santa Rosa De Lima, San Miguel, Usulutan	Gerardo Barrios
31	96.9	3.2656	Santa Rosa De Lima, Gerardo Barrios, San Miguel, Usulutan	San Francisco Gotera
28	96.6	2.6207	San Francisco Gotera, La Union, San Miguel, Usulutan	Gerardo Barrios
28	96.6	2.5224	San Francisco Gotera, La Union, Gerardo Barrios, Usulutan	San Miguel

Figura 38 Resultados del algoritmo Apriori para el caso C-ASO-02

El algoritmo Apriori devuelve las reglas de asociación presentadas como Antecedente/Consecuente, así como sus respectivos indicadores de rendimiento obteniéndose la cantidad de 9 reglas de asociación.

13.3.2.2.2.7 Interpretación de resultados

Cuando los precios de los productos de granos básicos aumentan en las plazas SAN FRANCISCO GOTERA, LA UNION, SAN MIGUEL, USULUTAN, también lo hace en GERARDO BARRIOS además, la probabilidad de que los precios aumenten en dicha plaza dado que aumentaron en las anteriormente mencionadas es mayor a la probabilidad de que solo en esa plaza aumenten de precio los productos esto ya que el RuleLift es mayor que uno, además vale la pena destacar que los precios en la plaza San Miguel en aumento se encuentran en 8 de los antecedentes de las 8 reglas por lo tanto esta es una plaza que marca tendencia.

13.3.2.3 Técnica de clasificación

13.3.2.3.1 Clasificar el comportamiento de los precios en aumento, disminución o normal por producto.

13.3.2.3.1.1 Contenido del caso

N° de caso: C-CLA-01	
Técnica	Clasificación
Algoritmos	Arboles de decisión
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019
Variables	Se analiza el precio promedio por producto, tomando cada día como una transacción
Hipótesis	Clasificar los datos, en base al precio por producto por día sondeado discriminando por aumento, disminución o si se mantiene en términos normales.
Procedimiento	Flujo de trabajo en Knime
Resultados	Datos clasificados, reglas de asociación,
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión.
Herramienta de software	Knime Analytics Platform

Tabla 18 Contenido del caso N° C-CLA-01.

13.3.2.3.1.2 Población

SQL Statement

```
SELECT sk_tiempo,b.sk_producto, b.nombre_producto,precio,avg,
        CASE
    WHEN precio>avg+(avg*0.05) THEN 'MAYOR'
        WHEN precio BETWEEN avg-(avg*0.05) AND avg+(avg*0.05) THEN 'NORMAL'
        WHEN precio < avg-(avg*0.05) THEN 'MENOR'
    ELSE NULL
  END
  AS esmayor
FROM
    (SELECT pr.sk_producto,AVG(p.precio_max)
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
    JOIN producto_scd pr ON pr.sk_producto = h.sk_producto
    AND sk_unidad_medida=1
    GROUP BY pr.sk_producto) a
    JOIN
    (SELECT t.sk_tiempo,pr.nombre_producto, pr.sk_producto,AVG(precio_max) precio
    FROM hechos_sondeos_mag h
```

Preview results: Evaluate

Row ID	sk_tiempo	sk_pro...	nombre_producto	precio	avg	esmayor
Row0	1	8	ARROZ ORO 1a. CLASE IMPORTADO	41.667	39.481	MAYOR
Row1	1	4	ARROZ ORO 1a. CLASE NACIONAL	38	36.942	NORMAL
Row2	1	10	FRIJOL ROJO DE SEDA IMPORTADO	85.5	58.397	MAYOR
Row3	1	9	FRIJOL ROJO DE SEDA NACIONAL	89.4	59.902	MAYOR
Row4	1	7	FRIJOL TINTO IMPORTADO	85.333	54.392	MAYOR
Row5	1	3	FRIJOL TINTO NACIONAL	82.25	53.241	MAYOR
Row6	1	1	MAICILLO	24	19.943	MAYOR
Row7	1	2	MAIZ BLANCO NACIONAL	23.3	17.864	MAYOR
Row8	2	8	ARROZ ORO 1a. CLASE IMPORTADO	41.5	39.481	MAYOR
Row9	2	4	ARROZ ORO 1a. CLASE NACIONAL	39.5	36.942	MAYOR

Figura 39 Set de datos para el caso C-CLA-01

La población de datos utilizada consiste en una comparativa de los precios agrupado por productos para cada fecha sondeada, versus el promedio correspondiente, para definir si hubo un aumento, una disminución o se mantuvo dentro de los rangos esperados surgiendo las clases (MENOR, MAYOR y NORMAL).

13.3.2.3.1.3 Variables

Las variables que se ven involucradas en la exploración son la fecha del sondeo, el precio promedio por fecha y el precio promedio de la población de datos por producto.

13.3.2.3.1.4 Hipótesis

El objetivo de minería de datos que se quiere ejecutar, o la hipótesis que se quiere comprobar es la siguiente: “Realizar una clasificación de los sondeos de producto por fecha para determinar si existió un aumento, una disminución o los precios se mantuvieron de forma normal, para posteriormente obtener algunas reglas de asociación de los aumentos en los precios de los productos”.

13.3.2.3.1.5 Procedimiento

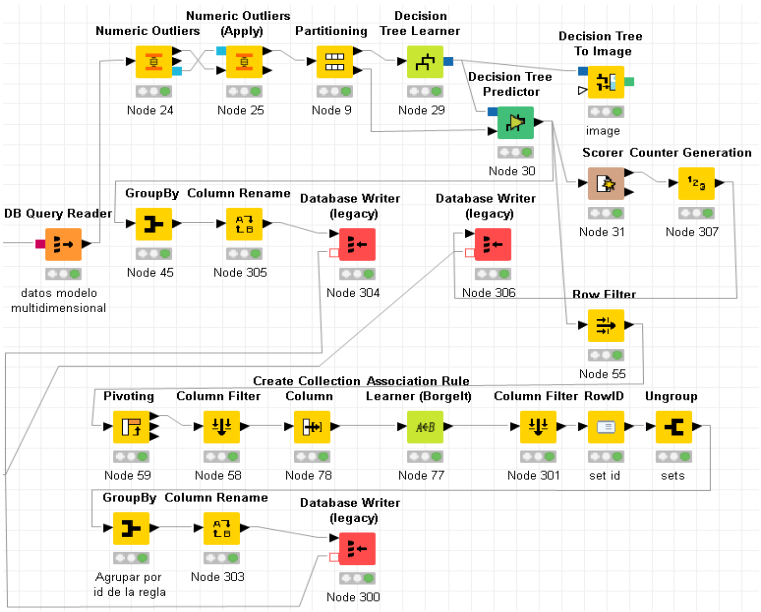


Figura 40 Flujo de trabajo C-CLA-01.

En la Figura 40 se muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-CLA-01 se resumen a continuación las acciones que se llevan a cabo:

Se aplica el algoritmo de árboles de decisión particionado los datos de manera absoluta, como variante, además de clasificar los datos, se filtran para las filas únicamente aquellas en las que el algoritmo clasificó como un aumento, posteriormente se pivotean los datos y se filtra las columnas para solo aquellas que contienen los precios por producto por transacción, se crea una columna de colección y se aplica Rule Learner para encontrar reglas de asociación.

13.3.2.3.1.6 Resultados

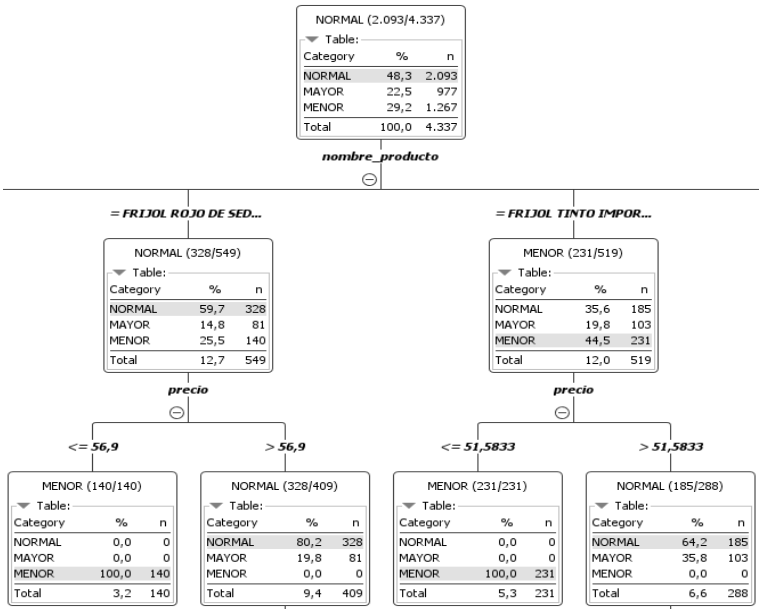


Figura 41 fragmento del árbol de decisión C-CLA-01.

La Figura 41 muestra un fragmento del árbol de decisión generado por el aprendizaje del algoritmo, en el cual se observa que se discrimina en primer momento porque producto es y posterior en base al precio que este tuvo lo clasifica por MAYOR, MENOR o NORMAL.

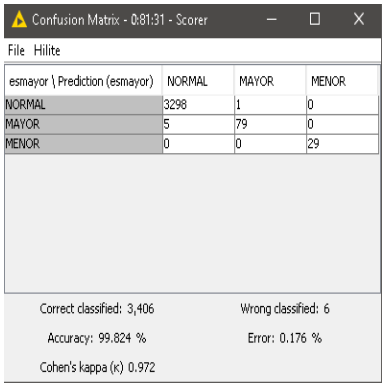


Figura 42 Indicadores de rendimiento C-CLA-01

La Figura 42 muestra los indicadores de rendimiento para el árbol de decisión generado los cuales corresponden a:

- **Matriz de confusión:** Donde se contabilizan las veces que el algoritmo asignó cada una de las clases vs la clase que se proporcionó en los datos de prueba.
- **Correctamente clasificados:** La cantidad de datos que se clasificaron de forma correcta.
- **Accuracy:** La exactitud que logró el algoritmo al clasificar.
- **Wrong clasified:** La cantidad de datos que se clasificó de forma errónea.
- **Error:** El porcentaje de error que tuvo el algoritmo.
- **Cohen's Kappa(k):** La estadística que toma en cuenta la probabilidad de que el acuerdo ocurra por casualidad, entre más se acerca a 1 los resultados del algoritmo son mejores.

nombre_producto	NORMAL...	MAYOR...	MENOR...
ARROZ ORO 1a. CLASE IMPORTADO	433	?	?
ARROZ ORO 1a. CLASE NACIONAL	433	?	?
FRIJOL ROJO DE SEDA IMPORTADO	388	17	?
FRIJOL ROJO DE SEDA NACIONAL	425	9	?
FRIJOL TINTO IMPORTADO	387	18	?
FRIJOL TINTO NACIONAL	420	14	?
MAICILLO	421	13	?
MAIZ BLANCO NACIONAL	392	13	29

Figura 43 Contador de clase por producto C-CLA-01

Se observa en la Figura 43 un conteo por cada una de las clases por producto mostrando "?" para los productos que obtuvieron 0 registros clasificados en dicha clase.

13.3.2.3.1.7 Interpretación de resultados

Se puede observar que el frijol tinto importado en el lapso de 2018 a octubre 2019 se ha mantenido normal respecto al promedio de su precio 387 veces versus las 18 veces que fue

normal y las 0 que fue menor, esto en base a la técnica de clasificación aplicando el algoritmo de árboles de decisión con una precisión aproximada del 100%, además como reglas de asociación generadas se tiene que el FRIJOL ROJO DE SEDA IMPORTADO, FRIJOL TINTO IMPORTADO y FRIJOL TINTO NACIONAL aumenta de precio cuando así lo hacen el resto de tipos de frijol.

13.3.2.3.2 Clasificar el comportamiento de los precios en aumento, disminución o normal por plaza.

13.3.2.3.2.1 Contenido del caso

N° de caso: C-CLA-02	
Técnica	Clasificación
Algoritmos	Arboles de decisión
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019.
Variables	Se analiza el precio promedio de los productos de granos básicos agrupados por plaza, tomando cada semana como una transacción.
Hipótesis	Clasificar los datos, en base al precio por plaza por semana discriminando por aumento, disminución o si se mantiene en términos normales.
Procedimiento	Flujo de trabajo en Knime
Resultados	Datos clasificados, reglas de asociación.
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión.
Herramienta de software	Knime Analytics Platform

Tabla 19 Contenido del caso N° C-CLA-02.

13.3.2.3.2.2 Población

SQL Statement

```
SELECT semana,b.sk_plaza, b.nombre_plaza,precio,avg,
        CASE
    WHEN precio>avg+(avg*0.05) THEN 'MAYOR'
        WHEN precio BETWEEN avg-(avg*0.05) AND avg+(avg*0.05) THEN 'NORMAL'
        WHEN precio < avg-(avg*0.05) THEN 'MENOR'
    ELSE NULL
  END
  AS esmayor
FROM
    (SELECT pr.sk_plaza,AVG(p.precio_max)
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
    JOIN plaza_scd pr ON pr.sk_plaza = h.sk_plaza
    AND sk_unidad_medida=25
    GROUP BY pr.sk_plaza) a
    JOIN
    (SELECT t.semana,pr.nombre_plaza, pr.sk_plaza,AVG(precio_max) precio
```

Preview results: Evaluate

Row ID	semana	sk_plaza	nombre_plaza	precio
Row0	Semana del 2015-08-17 al 2015-08-21	206	AHUACHAPAN	59.333
Row1	Semana del 2015-08-17 al 2015-08-21	196	CHALCHUAPA	60
Row2	Semana del 2015-08-17 al 2015-08-21	199	GERARDO BARRIOS	58.562
Row3	Semana del 2015-08-17 al 2015-08-21	205	LA UNION	67.286
Row4	Semana del 2015-08-17 al 2015-08-21	202	SAN FRANCISCO GOTERA	64.714
Row5	Semana del 2015-08-17 al 2015-08-21	197	SAN VICENTE	58.375
Row6	Semana del 2015-08-17 al 2015-08-21	189	SANTA ANA	62.333
Row7	Semana del 2015-08-17 al 2015-08-21	193	SANTA ROSA DE LIMA	62.571
Row8	Semana del 2015-08-17 al 2015-08-21	200	SANTA TECLA	60.429
Row9	Semana del 2015-08-17 al 2015-08-21	191	SENSUNTEPEQUE	59.625

Figura 44 Población de datos para caso C-CLA-02

La Figura 44 muestra la población de datos que en este caso a diferencia del caso C-CLA-01 se toma por semana analizando los precios por plaza.

La población de datos utilizada consiste en una comparativa de los precios agrupado por plazas para cada fecha sondeada, versus el promedio correspondiente, para definir si hubo un aumento, una disminución o se mantuvo dentro de los rangos esperados surgiendo las clases (MENOR, MAYOR y NORMAL).

13.3.2.3.2.3 Variables

Las variables que se ven involucradas en la exploración son la fecha del sondeo, el precio promedio por fecha y el precio promedio de la población de datos por plaza.

13.3.2.3.2.4 Hipótesis

El objetivo de minería de datos que se quiere ejecutar, o la hipótesis que se quiere comprobar es la siguiente: “Realizar una clasificación de los sondeos de plaza por semana para determinar si existió un aumento, una disminución o los precios se mantuvieron de forma normal, para posteriormente obtener algunas reglas de asociación de los aumentos en los precios de los productos por plaza”.

13.3.2.3.2.5 Procedimiento

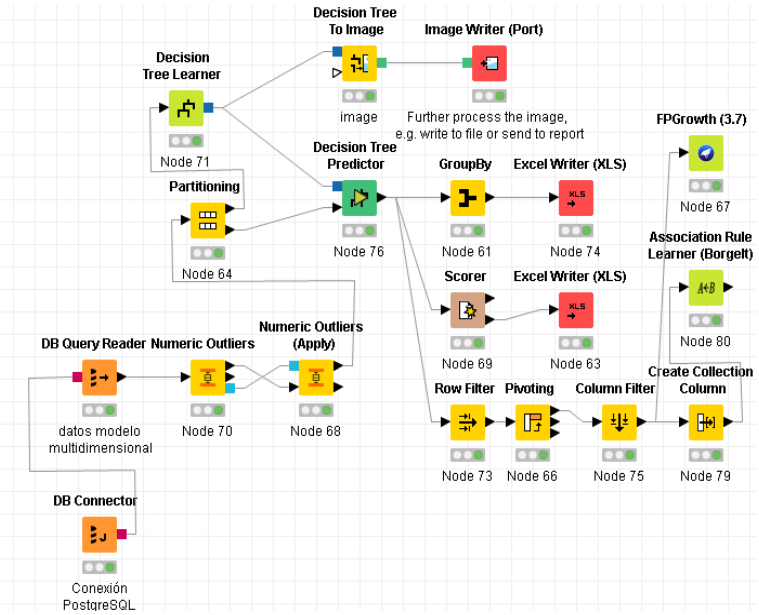


Figura 45 Flujo de trabajo para caso C-CLA-02.

La Figura 45 muestra el flujo de trabajo para el caso C-CLA-02 al igual a C-CLA-01 se utiliza arboles de decisión para clasificar, posteriormente se sacan los contadores por clase por plaza y luego se generan reglas de asociación con FP Growth Y Rule Learner.

13.3.2.3.2.6 Resultados

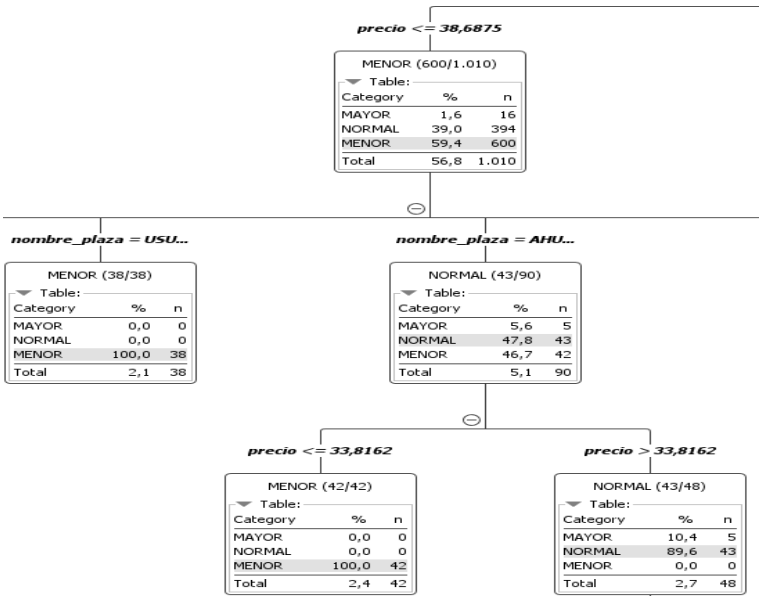


Figura 46 Fragmento del árbol de decisión C-CLA-02.

La Figura 46 muestra un fragmento del árbol de decisión generado por el aprendizaje del algoritmo el cual discrimina por plaza y por el precio promedio en cada de una de ellas para tomar la decisión de la clase a la que pertenecerá.

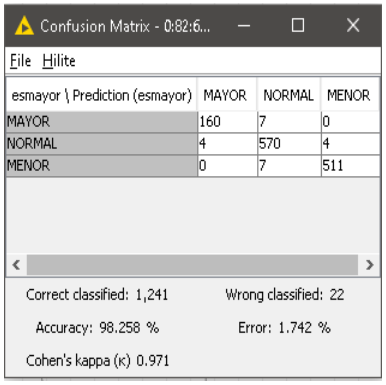


Figura 47 Indicadores de rendimiento C-CLA-02

La Figura 47 muestra los indicadores de rendimiento (precisión, clasificados correctamente e incorrectamente, error porcentual, Cohen’s Kappa y matriz de confusión) para el caso C-CLA-02.

nombre_plaza	MAYOR...	MENOR...	NORMAL...
AHUACHAPAN	12	39	27
ARMENIA	7	10	24
CHALATENANGO	19	19	40
CHALCHUAPA	10	12	19
COJUTEPEQUE	15	14	48
GERARDO BARRIOS	6	21	58
LA UNION	2	58	19
SAN FRANCISCO GOTERA	5	44	31
SAN MIGUEL	7	34	44
SAN VICENTE	12	26	41
SANTA ANA	15	46	24
SANTA ROSA DE LIMA	?	51	22
SANTA TECLA	4	45	27
SENSUNTEPEQUE	12	23	40
SONSONATE	21	37	22
USULUTAN	15	21	42
ZACATECOLUCA	5	18	50

Figura 48 Contador de clase por plaza C-CLA-02

La Figura 48 muestra los resultados para el algoritmo de árboles de decisión sobre la población especificada el conteo de cada de una de las clases por plaza.

13.3.2.3.2.7 Interpretación de resultados

Se puede observar que en las semanas de 2018 hacia adelante en la plaza La Unión los precios en general de los granos básicos se ha mantenido bajo así lo demuestran las 58 semanas en las cuales el algoritmo clasifico como menor versus las 19 veces que se mantuvo dentro de lo normal y las 2 que estuvo al alza, esto con una precisión del 98.26%. Para el caso de las reglas de asociación se tiene a Usulután y Sonsonate como plazas que generan tendencia ya que se ven implicadas como antecedentes en varias ocasiones por ejemplo cuando los precios de los granos básicos aumentan en las plazas de Usulután y Santa Ana, también lo hace en Ahuachapán esto con una confianza del 100% cabe mencionar también que la probabilidad que aumente el precio en dicha plaza aumenta con un Lift del 250% dado que aumento en las plazas antecedentes.

13.3.2.4 Técnica de pronóstico

13.3.2.4.1 Pronosticar precios de productos en las próximas semanas

13.3.2.4.1.1 Contenido del caso

N° de caso: C-FOR-01	
Técnica	Pronóstico
Algoritmos	ARIMA
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019
Variables	Se analiza el precio promedio de los productos de granos básicos, agrupados por semana
Hipótesis	Realizar predicciones de los precios para los productos de granos básicos para las próximas semanas.
Procedimiento	Flujo de trabajo en Knime

Resultados	Predicciones de los precios de los productos para las próximas semanas.
Interpretación de resultados	Mediante indicadores de rendimiento estadísticos como R^2, Mean abosolute error, Mean squered error, Mean absolute percentage error.
Herramienta de software	Knime Analytics Platform

Tabla 20 Contenido para el caso C-FOR-01.

13.3.2.4.1.2 Población

La población de datos utilizada es proveniente del modelo multidimensional, y consiste en precios para los productos de granos básicos de 2015-2019, agrupados por semana, para realizar predicciones para las próximas semanas.

SQL Statement

```
SELECT semana,fecha,anio,pr.sk_producto, nombre_producto, AVG(p.precio_pro)
    FROM hechos_sondeos_mag h
    JOIN precios_scd p ON p.sk_precio = h.sk_precio
    JOIN producto_scd pr ON pr.sk_producto = h.sk_producto
    JOIN dim_tiempo t ON t.sk_tiempo= h.sk_tiempo
    AND sk_unidad_medida=1
    GROUP BY 1,2,3,4
    ORDER BY anio ASC
```

Preview results: Evaluate

Row ID	ana	fecha	...	sk_pro...	nombre...	promedio
Row0	del 2015-08-17 al 2015-08-21	2015-08-19	...	1	MAICILLO	23.047
Row1	del 2015-08-17 al 2015-08-21	2015-08-19	...	2	MAIZ BLANC...	21.938
Row2	del 2015-08-17 al 2015-08-21	2015-08-19	...	3	FRIJOL TIN...	80.562
Row3	del 2015-08-17 al 2015-08-21	2015-08-19	...	4	ARROZ OR...	37.835
Row4	del 2015-08-17 al 2015-08-21	2015-08-19	...	7	FRIJOL TIN...	83
Row5	del 2015-08-17 al 2015-08-21	2015-08-19	...	8	ARROZ OR...	40.467
Row6	del 2015-08-17 al 2015-08-21	2015-08-19	...	9	FRIJOL ROJ...	86.85
Row7	del 2015-08-17 al 2015-08-21	2015-08-19	...	10	FRIJOL ROJ...	83.75
Row8	del 2015-08-17 al 2015-08-21	2015-08-20	...	1	MAICILLO	21.883
Row9	del 2015-08-17 al 2015-08-21	2015-08-20	...	2	MAIZ BLANC...	21.878

Figura 49 Set de datos para el caso C-FOR-01.

La Figura 49 muestra la consulta utilizada para obtener la población de datos a ser utilizada en la exploración de series temporales para lo cual precios de los productos sondeados están representados por fecha y por semana.

13.3.2.4.1.3 Variables

Las variables que se ven incluidas en el análisis son el tiempo, utilizando la periodicidad de la semana, el producto, representando para cada semana cada producto sondeado y finalmente precio, el precio que se observó en cada semana para cada producto sondeado.

13.3.2.4.1.4 Hipótesis

Por ser una técnica de previsión (Forecasting) el objetivo de la minería de datos en este caso o la hipótesis que se quiere llevar a cabo es realizar predicciones de los precios de los productos de granos básicos para semanas posteriores a las brindadas en la población de datos.

13.3.2.4.1.5 Procedimiento

Se presenta a continuación el flujo de trabajo en Knime que busca responder a la hipótesis planteada.

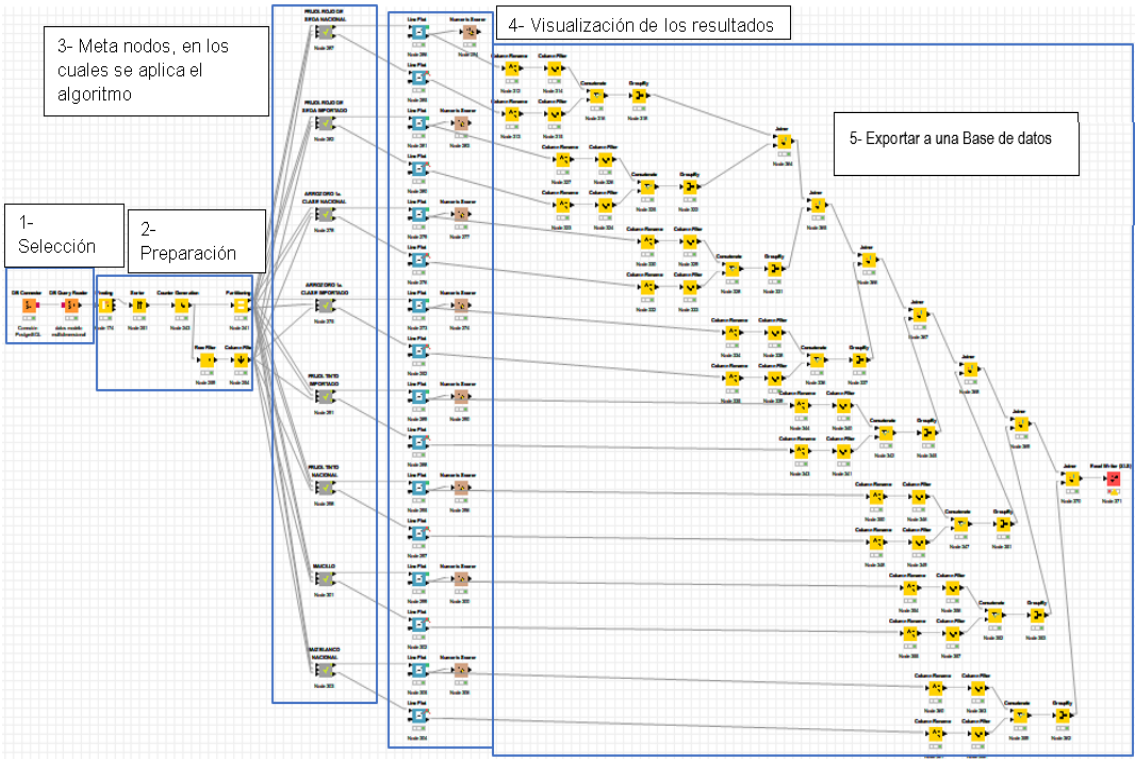


Figura 50 Señalización en el flujo de trabajo C-FOR-01

1- **Selección de los datos**: Se establece la conexión con la base de datos y se seleccionan los datos del modelo multidimensional mediante una consulta SQL.
2- **Preparación de los datos:** Se realiza un pinvoting en el cual la columna de agrupación es la semana colocando cada producto como una columna y como sus valores, la primera fecha para esa semana en los datos y el precio promedio del producto, se separan los datos de entrenamiento y de pruebas siendo desde la semana del 2015-08-17 al 2015-08-21 hasta la semana del 2017-09-18 al 2017-09-22 para entrenamiento y las restantes para pruebas.
3- **Meta nodos, en los cuales se aplica el algoritmo:** Se crea un metanodo por cada producto dentro de los cuales se encuentra el flujo de trabajo que utiliza el algoritmo de series de tiempo.

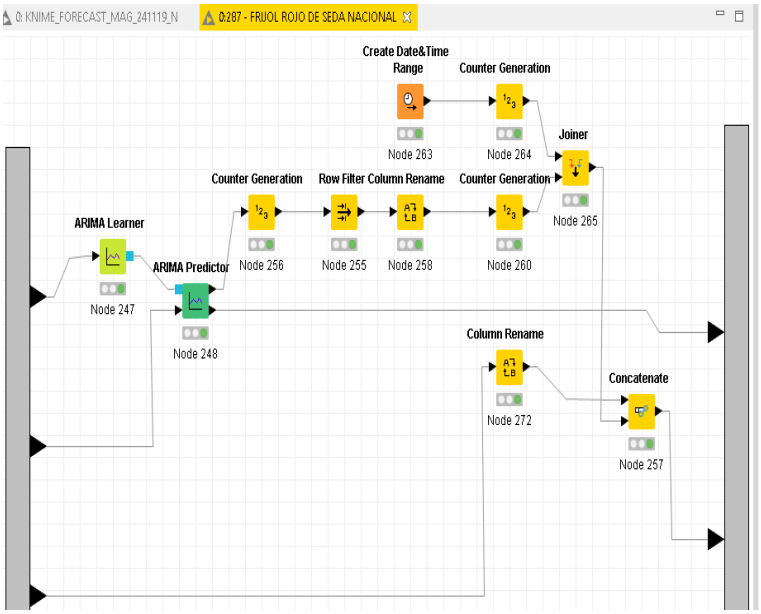


Figura 51 Flujo de trabajo metanodo FRIJOL DE SEDA NACIONAL caso C-FOR-01

La Figura 51 muestra el flujo de trabajo contenido en el meta nodo para “FRIJOL ROJO DE SEDA NACIONAL”, se recibe como entrada los datos de entrenamiento y de pruebas posteriormente se genera un contador para los datos que han sido pronosticados para filtrar los registros que comprenden la predicción, se coloca un segundo contador con una unidad de escala de 7 (1,7,14, etc), se genera una serie de fechas en un intervalo dado y se coloca un contador para dichos datos, posteriormente se realiza una unión de estos dos últimos contadores para determinar la fecha estimada a la cual se realizaron las predicciones, además de las dos entradas mencionadas anteriormente se tiene una tercera que son los datos sin procesar, se renombran las columnas de estos datos para estandarizarse respecto a los que se manejaron en la predicción, para finalmente ser concatenados , como salidas al metanodo se tienen los datos pronosticados para fechas futuras más los datos originales, y en la otra salida los datos pronosticados para fechas ya pasadas (como entrenó el algoritmo). Se realiza un proceso similar o equivalente en cada uno de los metanodos.

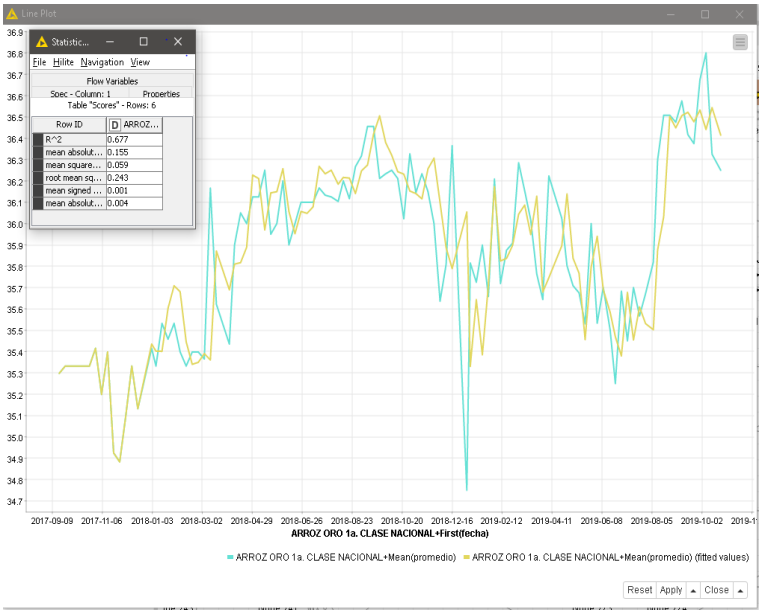


Figura 52 Gráfico comparativo de datos originales vs pronosticados para Arroz Oro 1a Clase Nacional.

4- En la Figura 52 se observan las salidas generadas para uno de los metanodos, las cuales consisten en el comparativo de datos originales vs datos pronosticados, así como sus indicadores de rendimiento, se observa en la tendencia como en los indicadores que son aceptables, se generan estas salidas para cada uno de los metanodos.

5- En la Figura 50 número 5 se concatenan los datos generados por cada metanodo, (datos originales, pronosticados del pasado y los que se predijeron para el futuro) para generar una salida a una hoja de cálculo para que pueda ser visualizada desde Power BI.

13.3.2.4.1.6 Resultados

Para manejar predicciones para multiples productos fue necesario la utilización de metanodos, para dar una solución modular, aplicando siempre algoritmo de series de tiempo en específico ARIMA, además generando salidas gráficas como de datos para ser procesados posteriormente.

13.3.2.4.1.7 Interpretación de resultados

Se evalúa el grado de precisión con el que entrenó el algoritmo y realizó las predicciones mediante los indicadores de rendimiento: El coeficiente de determinación (R cuadrado) nos dice en el caso de la Figura 52 que el modelo explica en un 68% lo que ocurre en la vida real, en el caso del error cuadrático medio que es de 0.059 nos indica un valor bajo por lo cual el ajuste realizado por el modelo tiene una precisión muy buena, así como también el error absoluto medio de 0.155.

13.3.2.5 Técnica de agrupamiento

13.3.2.5.1 Agrupar los meses en los cuales se identifican aumentos o bajas en los precios

13.3.2.5.1.1 Contenido del caso

N° de caso: C-AGR-01	
Técnica	Agrupamiento
Algoritmos	K-Means
Población	Datos provenientes del modelo multidimensional de precios de los granos básicos sondeados por el MAG, datos desde 2015 a 2019
Variables	Se analiza el precio de los productos
Hipótesis	Realizar Agrupamiento de precios promedio de productos por mes para identificar en que meses se concentran los precios más bajos o más altos.
Procedimiento	Programación en Python
Resultados	Grupos de Precios
Interpretación de resultados	Evaluar en base a los diferentes grupos de precios obtenidos por los algoritmos
Herramienta de software	Python

Tabla 21: Contenido del caso C-AGR-01.

13.3.2.5.1.2 Población

```
select extract(month from fecha),avg(precio_pro)precio_pro
from hechos_sondeos_mag
join producto_scd using (sk_producto)
join precios_scd using (sk_precio)
join dim_tiempo using (sk_tiempo)
join unidad_medida_scd using(sk_unidad_medida)
where sk_producto=82 and sk_unidad_medida=18 group by extract(month from fecha)
```

Figura 53 Población de datos caso C-AGR-01

La población de datos utilizada, es proveniente del modelo multidimensional y corresponde a una consulta donde un producto se agrupa por mes y se calcula el precio promedio para cada mes del año.

13.3.2.5.1.3 Variables

Las variables que se ven involucradas en la exploración el producto, el precio y el mes.

13.3.2.5.1.4 Hipótesis

El objetivo de minería de datos que se quiere llevar a cabo o la hipótesis que se desea comprobar es: Agrupar por mes el precio promedio del Producto para posteriormente poder identificar en cuales meses del año están concentrados los precios más bajos o en cuales se encuentran los precios más altos.

13.3.2.5.1.5 Procedimiento

1. Se importaron las siguientes librerías de Python

- Numpy: librería de Python, que le agrega mayor soporte para vectores y matrices
- Pandas: librería que nos sirve para manipular tablas de datos.
- Matplotlib.pyplot: libreria que contiene una colección de funciones para que se trabajen similar a Matlab.
- Seaborn: librería de visualización basada en matplotlib.
- Kmeans: librería de sklearn que implementa el algoritmo kmeans.
- Psycopg2: librería para PostgreSQL.

Librerias

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.cluster import KMeans
import psycopg2
```

Figura 54 Librerías de Python.

2. Se carga la población de datos desde el modelo multidimensional por medio de una conexión con postgreSQL usando la librería psycopg2 y el resultado se almacena en un Dataframe de pandas.

```
conn = psycopg2.connect(database = "mag", user = "postgres", password = "admin", host="localhost", port="5433" )
data = pd.read_sql('select extract(month from fecha),avg(precio_pro)precio_pro from hechos_sondeos_mag join producto
```

```
data.head()
```

	date_part	precio_pro
0	1.0	18.854750
1	2.0	18.716912
2	6.0	14.824599
3	10.0	16.049850
4	9.0	15.267389

Figura 55 Carga Población de Datos

3. Se Graficó los datos cargados para poder visualizar de mejor manera

```
plt.scatter(data['date_part'],data['precio_pro'])
plt.xlabel('Mes')
plt.ylabel('Precio Promedio')
plt.show()
```

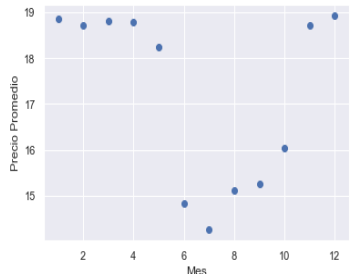


Tabla 22 Resultados de Kmeans.

4. Luego se implementa el Algoritmo de Kmeans de la librería sklearn. clúster indicándole cuantos clústeres queremos obtener en el resultado como se muestra en la figura 169 se ingresó 3.

```
x=data.copy()
```

```
kmeans=KMeans(3)
kmeans.fit(x)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

Figura 56 Implementación Kmeans

13.3.2.5.1.6 Resultados

Los resultados obtenidos son tres grupos de precios para los diferentes meses del año

- Grupo de precios número uno que son los que están conformados entre los meses de enero a mayo.
- Grupo de Precios numero dos que son los que están conformados entre los meses de junio a octubre.
- Grupo de precios número tres que son lo que están conformados entre noviembre y diciembre.

Clustering Result

```
clusters=x.copy()
clusters['cluster_pred']=kmeans.fit_predict(x)
```

Plot

```
plt.scatter(clusters['date_part'],clusters['precio_pro'],c=clusters['cluster_pred'],cmap='rainbow')
plt.xlabel('Mes')
plt.ylabel('Precio Promedio')
plt.show()
```

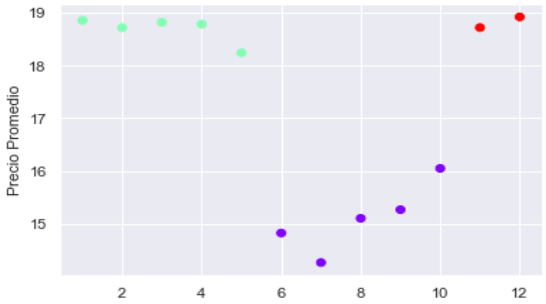


Figura 57 Resultados caso C-AGR-01.

13.3.3 Visualización

13.3.3.1 Informes de inteligencia de negocios

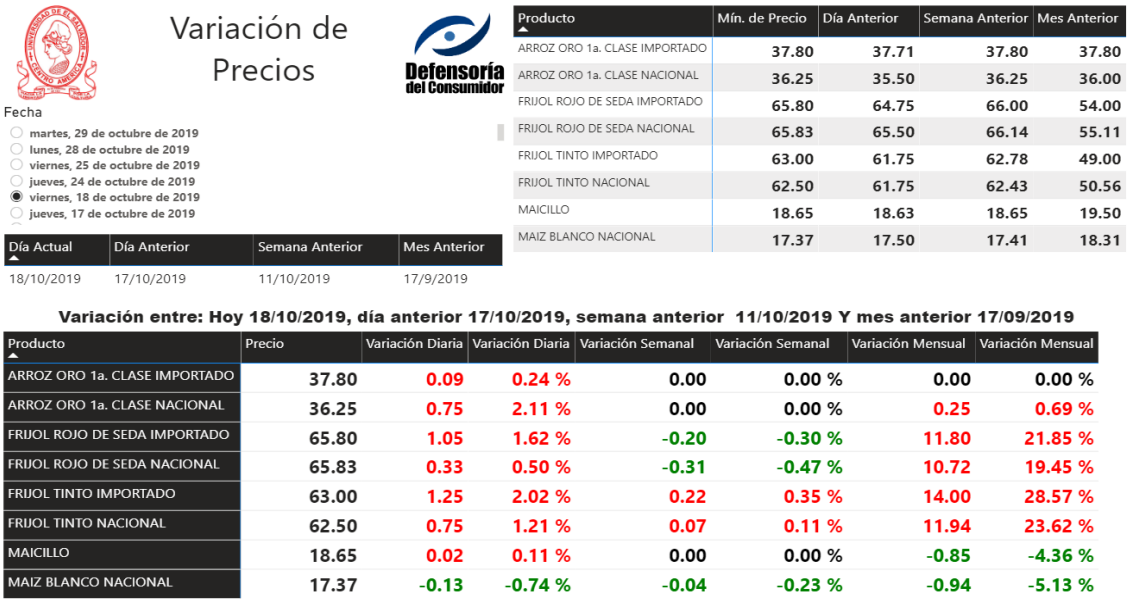


Figura 58 Informe de variación de precios de productos.

La Figura 58 presenta el informe de variación en el cual se puede visualizar la diferencia en el precio de los productos en términos porcentuales y monetarios desde una fecha concreta en comparación al día anterior, la semana anterior y el mes anterior.

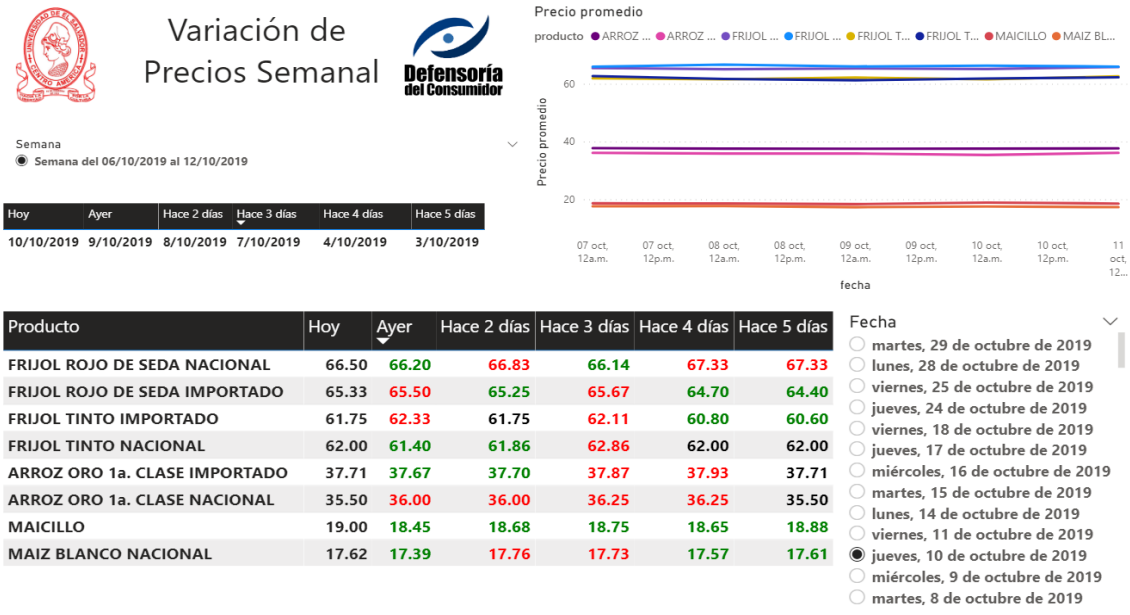


Figura 59 Diseño de la interfaz gráfica del informe de variación semanal.

La Figura 59 muestra el informe de variación semanal en el cual se visualiza a nivel gráfico y tabular el precio de los productos desde una fecha específica hasta una semana atrás día a día.

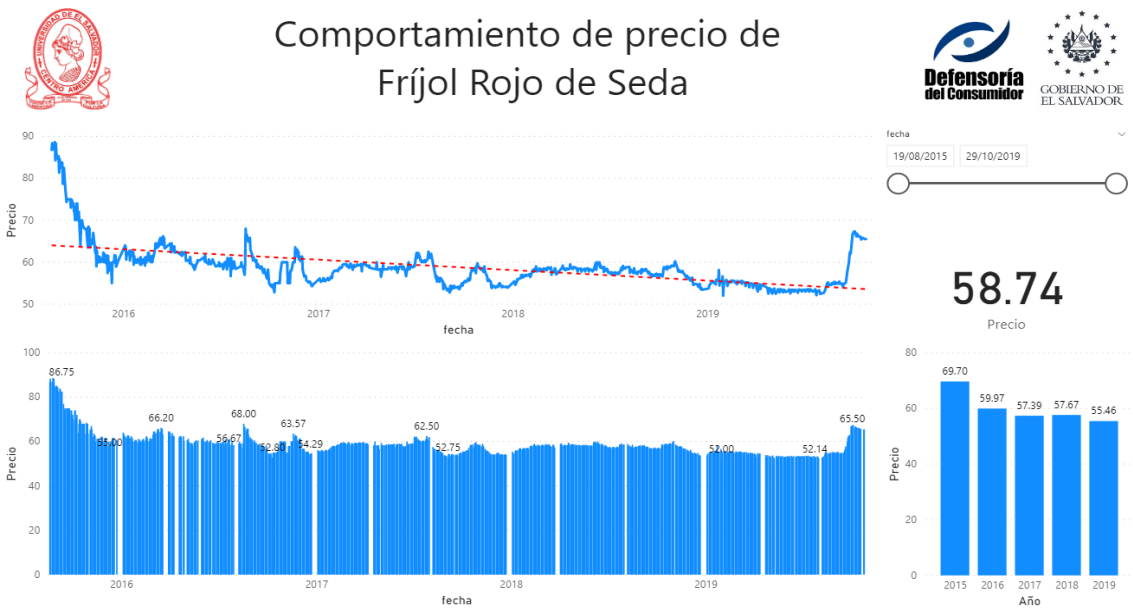


Figura 60 Informe sobre el comportamiento de precio del Frijol Rojo de Seda Nacional.

La Figura 60 muestra el informe sobre el comportamiento del precio del Frijol Rojo de Seda Nacional, este permite filtrar por fecha y muestra en diferentes gráficos como ha sido el comportamiento del precio de dicho producto, se tiene disponible este informe para los productos Frijol Rojo Tinto Nacional y Arroz Blanco Nacional adicionalmente.

13.3.3.2 Informes de resultados minería de datos

Antecedente	Consecuente	Soporte	Confianza	Lift
FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	FRIJOL TINTO NACIONAL	155	95.70	2.47
FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	MAIZ BLANCO NACIONAL	157	96.90	1.75
FRIJOL TINTO NACIONAL, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	FRIJOL TINTO IMPORTADO	155	87.10	4.14
FRIJOL TINTO NACIONAL, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	MAIZ BLANCO NACIONAL	172	96.60	1.75
FRIJOL TINTO NACIONAL, FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA IMPORTADO	FRIJOL ROJO DE SEDA NACIONAL	155	97.50	2.59
FRIJOL TINTO NACIONAL, FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA IMPORTADO	MAIZ BLANCO NACIONAL	154	96.90	1.75
FRIJOL TINTO NACIONAL, FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA NACIONAL	FRIJOL ROJO DE SEDA IMPORTADO	155	94.50	4.30
FRIJOL TINTO NACIONAL, FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA NACIONAL	MAIZ BLANCO NACIONAL	158	96.30	1.74
FRIJOL TINTO NACIONAL, FRIJOL TINTO IMPORTADO, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	MAIZ BLANCO NACIONAL	150	96.80	1.75
MAICILLO, MAIZ BLANCO NACIONAL, FRIJOL ROJO DE SEDA NACIONAL	FRIJOL TINTO NACIONAL	113	87.60	2.26
MAICILLO, MAIZ BLANCO NACIONAL, FRIJOL TINTO NACIONAL	FRIJOL ROJO DE SEDA NACIONAL	113	86.30	2.29
MAIZ BLANCO NACIONAL, FRIJOL ROJO DE SEDA NACIONAL, FRIJOL ROJO DE SEDA IMPORTADO	FRIJOL TINTO IMPORTADO	157	86.30	4.10

Figura 61 Visualizar la relación existente entre los precios de los productos (C-ASO-01).

Para el reporte del algoritmo a priori en la Figura 61, los datos se proyectan en 3 partes. La primera es una tabla donde se pueden observar los itemsets frecuentes que el algoritmo ha

descubierto junto con su consecuente. Además, se muestra la información correspondiente del Soporte, Confianza y Lift para cada uno de ellos.

Asimismo, en el lado derecho se proyecta el número de itemsets frecuentes que se ha descubierto el algoritmo y un recuento de los productos que han sido descubiertos como consecuentes.

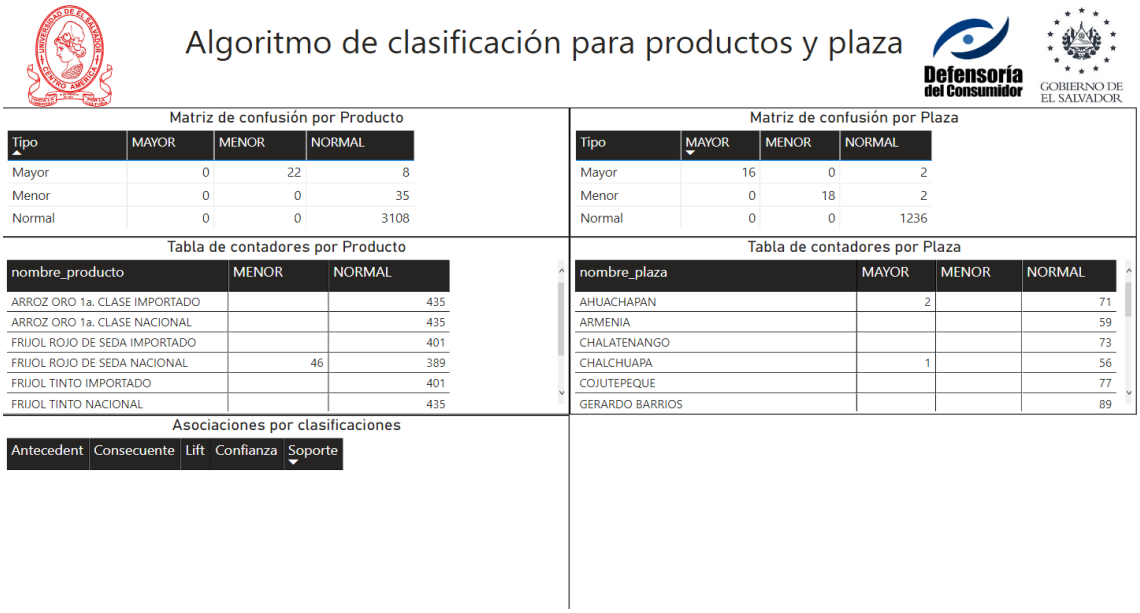


Figura 62 Visualizar la clasificación del comportamiento de los precios por producto y plaza.

La Figura 62 presenta la visualización para los resultados de la clasificación del comportamiento de los precios por producto y plaza, mostrando la matriz de confusión, contadores por clase y relaciones existentes.

Figura 63 Visualizar el pronóstico de los precios de los productos para las próximas semanas para Arroz Oro Importado y Nacional.

La Figura 63 muestra el gráfico de línea para dos de los productos Arroz oro importado y arroz oro nacional poniendo en perspectiva los datos originales en la línea celeste y los datos pronosticados en la línea azul.

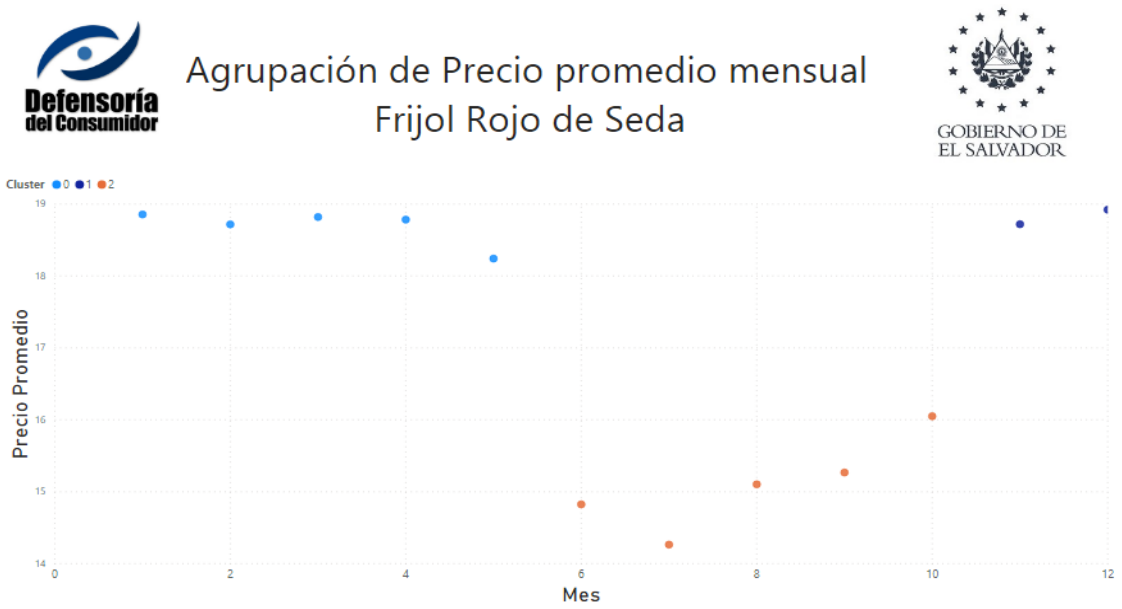


Figura 64 Visualizar el agrupamiento de los meses en los cuales se identifican bajas o altas en los precios

En la Figura 64 muestra el gráfico de dispersión para el producto frijol rojo de seda donde se representan los grupos para los diferentes meses del año representados con valores de 1 a 12 , el primero grupo de color naranja representa los meses donde se tuvo en precio promedio para el producto mientras los grupos de color azul y celeste representa los meses donde se tuvo precios arriba del promedio.

14 Sprint 2

14.1 Descripción historias de usuario

Código	RSPM01
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea transformar los archivos históricos que se tienen actualmente en formato Excel de los sondeos realizados en mercados a bases de datos.
Razón	Para poder obtener los datos más rápidos y mejor resguardados.
Criterios de aceptación	La estructura de las tablas deberá estar normalizada.
	Se deberán crear las tablas necesarias para normalizar de mejor manera posible, respetando la integridad de los datos.
	La estructura de la base de datos deberá respetar la consistencia de los datos.
Validación	Comprobar que la base de datos este correctamente normalizada.
	Comprobar que la normalización respeta la integridad de los datos.
	Comprobar que la base de datos respeta la consistencia de los datos.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 23 Primera historia de usuario para mercado.

Código	RSPM02
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea guardar la información de las bases de datos en estructuras de datos más robustas.
Razón	Para poder obtener de mejor manera los datos y poder generar información más rápido.
Criterios de aceptación	Se deberá crear un almacén de datos.
	Se deberán crear las dimensiones necesarias.
	Se deberá tener al menos una tabla de hechos.
Validación	Comprobar que el almacén de datos sea respetando un esquema como estrella, constelación o copo de nieve.
	Comprobar que el almacén de datos deberá tener las dimensiones necesarias como la dimensión tiempo, precios o productos.
	Comprobar que el almacén de datos deberá contener una tabla de hechos.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 24 Segunda historia de usuario para mercado.

Código	RSPM03
Rol	Como técnico(a) UACM/ Jefatura
Funcionalidad	Se desea conocer un informe de las fluctuaciones y el comportamiento de los precios de frutas, verduras, granos básicos, lácteos, aceites y margarinas.

Razón	Para informar oportunamente a presidencia de la Defensoría, a la población y boletines informativos a la población.
Criterios de aceptación	Se debe considerar los distintos tipos frutas, verduras, granos básicos, lácteos, aceites y margarinas.
	Se debe presentar su precio actual, variación semanal, mensual y anual respecto al día seleccionado.
	Las variaciones de los precios se deben expresar tanto en términos monetarios como en forma porcentual.
Validación	Se comprobará que se muestre la información para los productos especificados
	Se comprobará que se muestre información en los lapsos de tiempo especificados
	Se comprobará que las variaciones se muestren en las unidades de medida especificadas
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 25 Tercera historia de usuario para mercado.

Código	**RSPP04**
Rol	Como técnico(a) UACM/ Jefatura
Funcionalidad	Se desea conocer patrones de comportamiento y/o predicciones de frutas, verduras, granos básicos, lácteos, aceites y margarinas.
Razón	Para poder informar oportunamente a presidencia de la Defensoría del consumidor y a la población en general.
Criterios de aceptación	Deberá mostrar la salida con los filtros iniciales para este grupo de precios.
	Deberá presentar gráficos y filtros que representen la información descubierta.
	Podrá manipular los parámetros respecto a las variables que se relacionan en la salida
Validación	Se comprobará que las salidas muestren la información sin filtrar
	Se comprobará que la información descubierta se represente mediante un informe de manera gráfica
	Se comprobará que se pueda filtrar la información mediante los parámetros que tiene a disposición
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 26 Cuarta historia de usuario para mercado.

Código	**RSPM05**
Rol	Como técnico(a) UACM/ Jefatura
Funcionalidad	Se desea conocer la normalidad de los precios de productos que se proporcionan.
Razón	Para poder asignar inspectores a la investigación en los sitios en específico para corroborar las situaciones

Criterios de aceptación	Se deberá mostrar la salida con los filtros iniciales para este grupo de precios
	Se deberá considerar los productos comprendidos.
	Se deberá presentar un indicador que calificará el comportamiento de los precios
Validación	Se comprobará que las salidas muestren la información sin filtrar
	Se comprobará que se representen las variables involucradas en el análisis
	Se validará que la calificación que se le da como comportamiento sea congruente con la información presentada
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 27 Quinta historia de usuario para mercado.

14.2 Refinamiento del requerimiento de información

14.2.1 Proceso BPMN

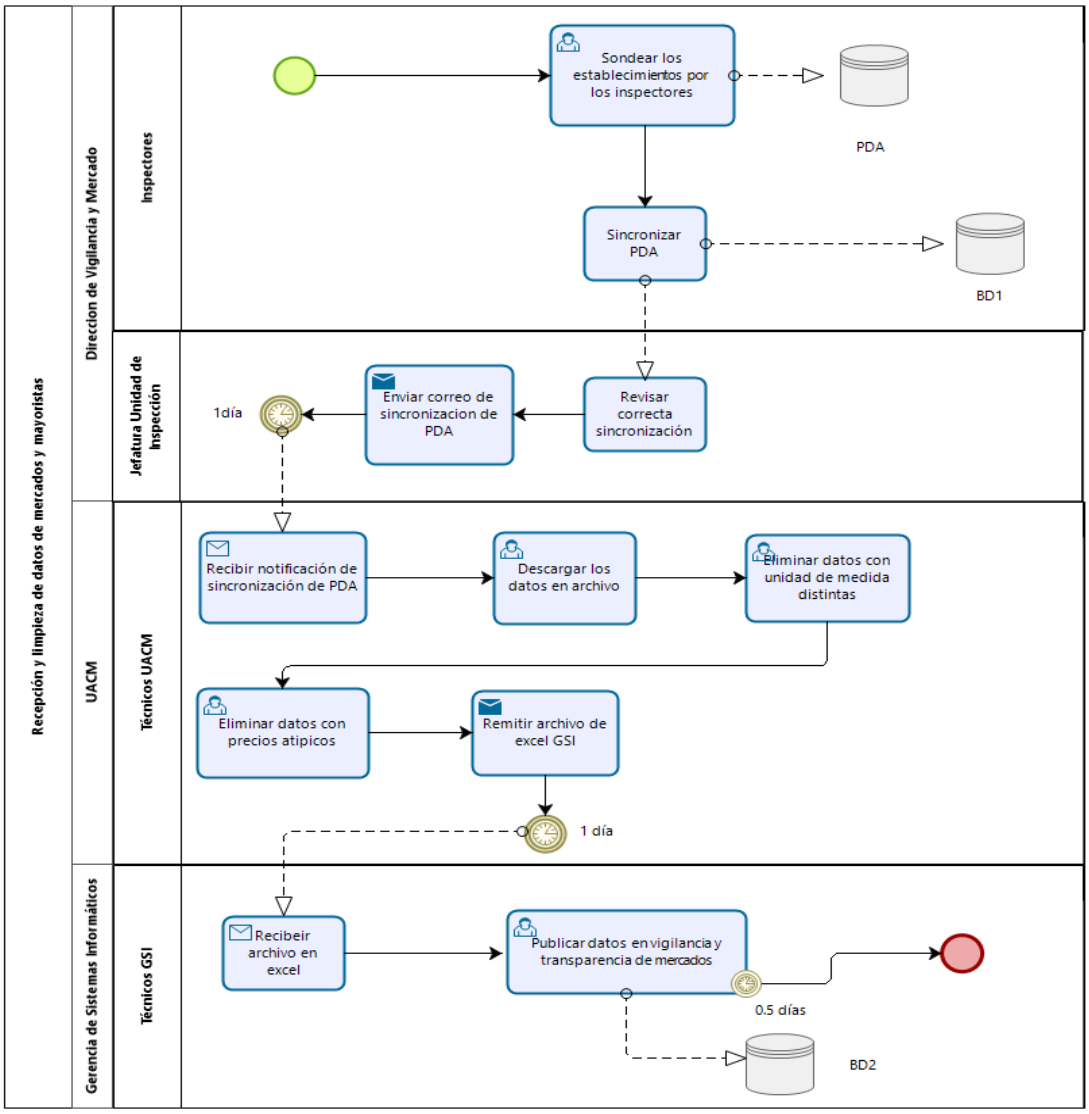


Figura 65 Proceso de consolidación de datos.

La Figura 65 muestra el diagrama BPMN del proceso actual que realiza la UACM para llevar a cabo el consolidado de precios de mercado y mayorista.

14.2.2 Paquete de Información

Tema:	**Variaciones de precios de frutas, verduras, granos básicos, lácteos, aceites y margarinas**					
JERARQUIAS	Tiempo	Productos	Marca	Establecimiento	Categoría	Unidad Medida
	Año	Nombre del producto	Nombre de la marca	Nombre del negocio	Nombre de la categoría	Nombre de la unidad de medida
	Mes			Nombre del establecimiento		

Tema:	Variaciones de precios de frutas, verduras, granos básicos, lácteos, aceites y margarinas					
	Semana			Nombre del departamento		
	Dia			Nombre del municipio		
				Nombre de la región		
Hechos Medidos:	Comportamiento de precios de frutas, verduras, granos básicos, lácteos, aceites y margarinas (Medida calculada)					

Tabla 28 Paquete de información de mercados y mayorista.

14.2.3 Casos de uso

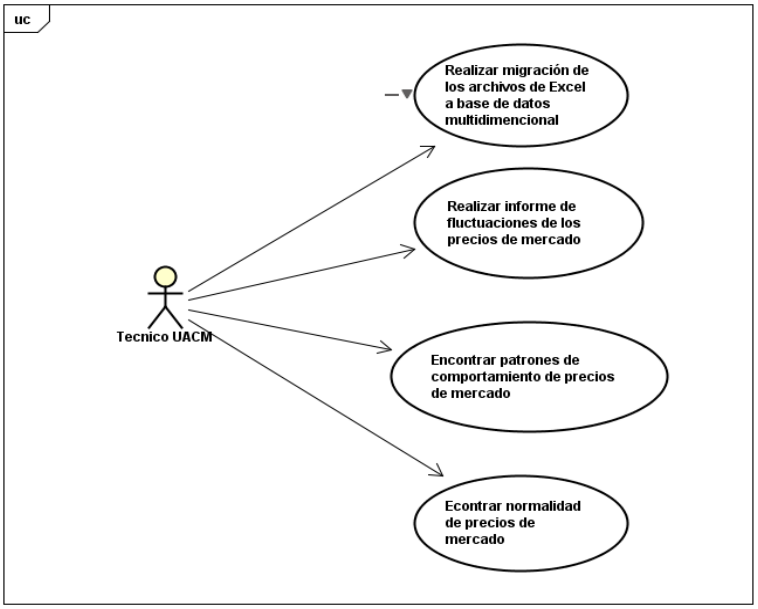


Figura 66 Diagrama de Casos de Uso Sondeo de Precios de mercados.

14.3 Desarrollo de la iteración

14.3.1 Integración de los datos

14.3.1.1 Extracción de los datos

Se cuenta con un documento de hoja de cálculo en Excel, en el cual la UACM consolida los sondeos semanales son remitidos por los inspectores. Y otro documento separado por comas (CSV) que presenta el consolidado histórico de los sondeos.

	ID sondeo	Nombre sondeo	Categoría	Producto	Principio	Marca	Cantidad	Unidad	Precio
2	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
3	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
4	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
5	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
6	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
7	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
8	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
9	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
10	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
11	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
12	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	7	Unidades	1
13	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
14	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
15	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
16	14	Sondeo de Mercados	Frutas	Guineo, Banano (unidad)		Sin Marca	8	Unidades	1
17	14	Sondeo de Mercados	Frutas	Guineo, Indio (unidad)		Sin Marca	10	Unidades	1

Figura 67 Archivo de sondeos de mercado consolidado en Excel por UACM.

Las consideraciones a tomar en cuenta para las fuentes de datos son:

1- Los datos son leídos como caracteres y se hacen las conversiones necesarias en el integrador de datos para evitar trabajo manual.
2- Los datos no son homogéneos por lo tanto dicha homogenización se realiza desde el integrador de datos.
3- Es necesario convertir la fecha al formato que la base de datos la puede capturar.
4- La variación no se almacenará si no que se calculará posteriormente en los informes.
5- Algunos nombres de productos no coinciden con los que se manejan en el consolidado por lo tanto es necesario homogenizar.
6- Existen sondeos de precios que no se asigna, estos no son tomados en cuenta.

Es necesario validar en la base de datos si ya existe el producto, plaza o unidad de medida para no repetir en los catálogos.

14.3.1.2 Staging área

14.3.1.2.1 Diseño del modelo staging área

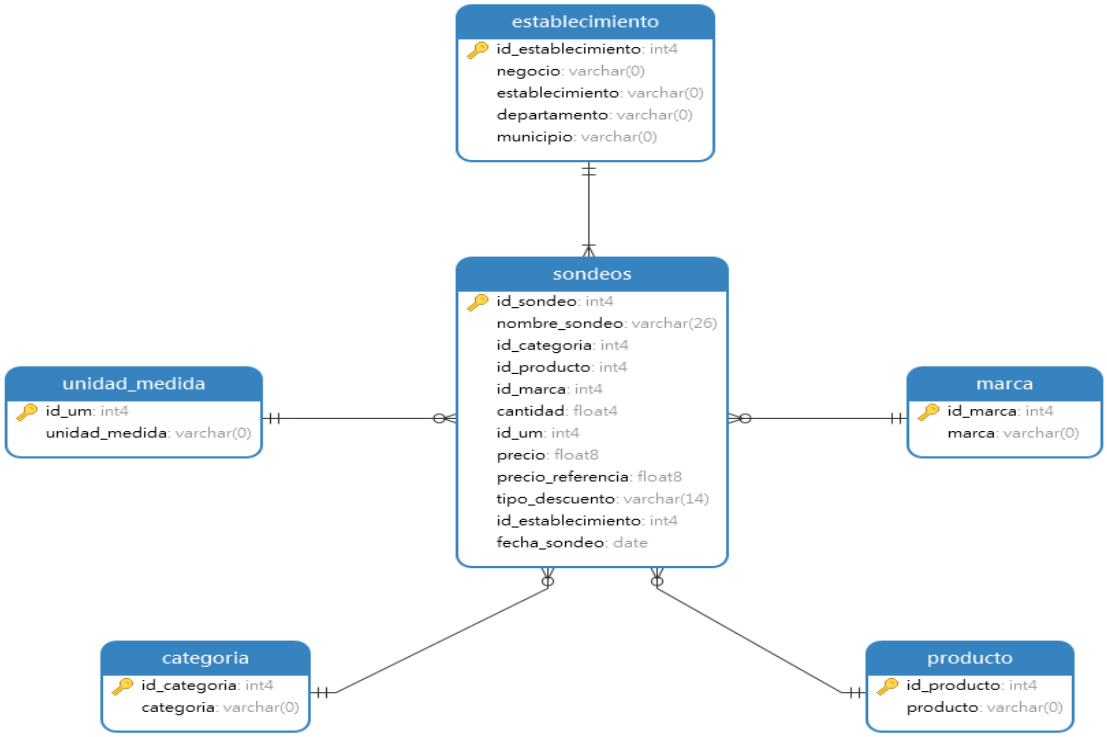
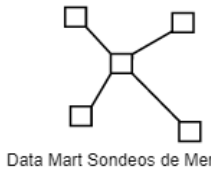


Figura 68 Diseño del modelo relacional.

14.3.1.3 Modelo multidimensional

14.3.1.3.1 Diseño Conceptual del Data mart (UML)



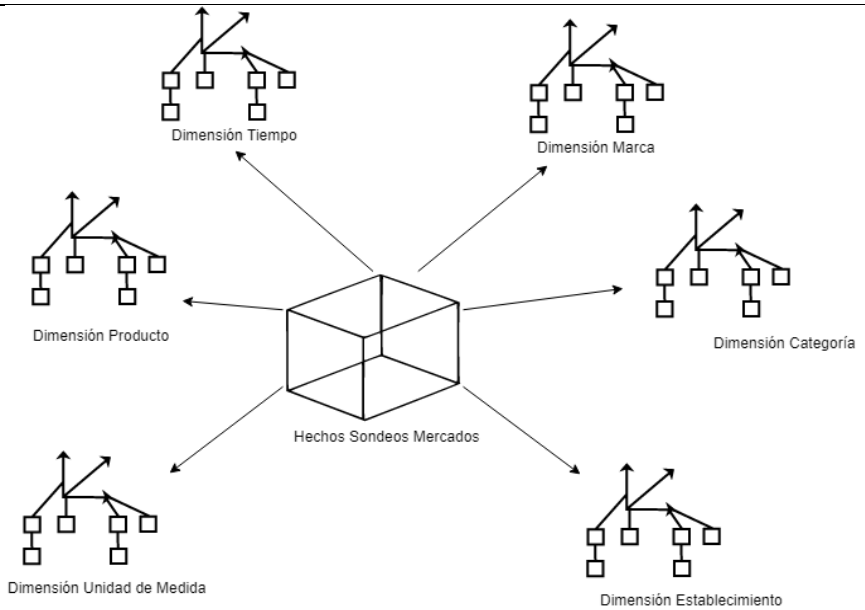
Nivel 1: Definición del modelo

Paquete Hechos:

Sondeo Mercados.

Paquete Dimensiones:

1. Tiempo
2. Marca
3. Categoría
4. Establecimiento
5. Unidad de Medida
6. Producto



Nivel 2: Definición del esquema estrella.

Figura 69 Niveles 1 y 2 del diseño conceptual sondeos Mercados.

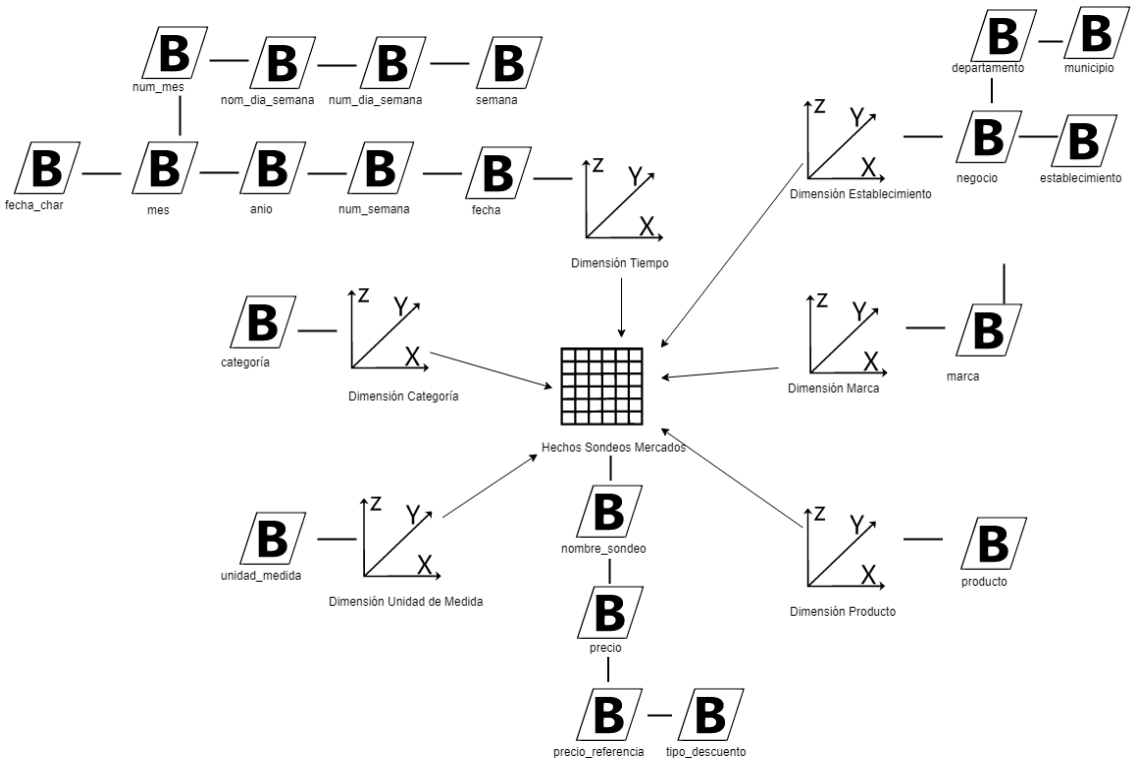


Figura 70 Nivel 3: Dimensiones/Hechos sondeos Mercados.

14.3.1.3.2 Diseño Físico Data Mart

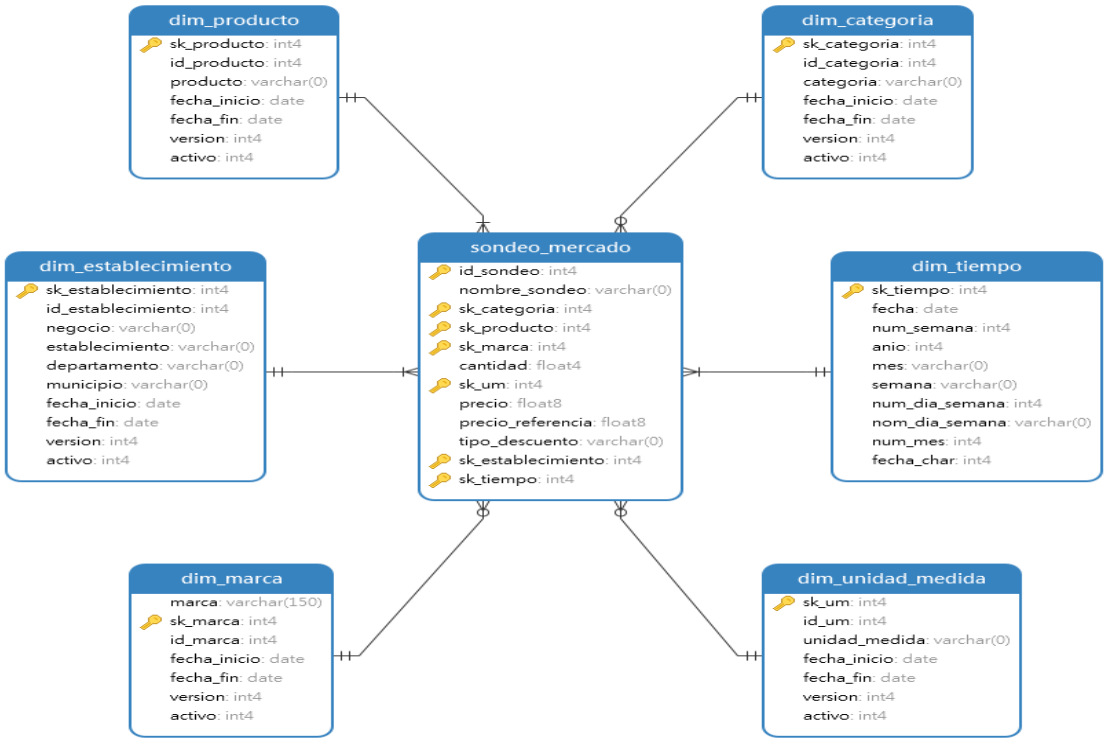


Figura 71 Diseño de Data Mart de sondeo de mercado.

14.3.1.3.3 Diseño de procesos ETL

Se presenta a continuación el diseño del proceso ETL de una de las dimensiones.

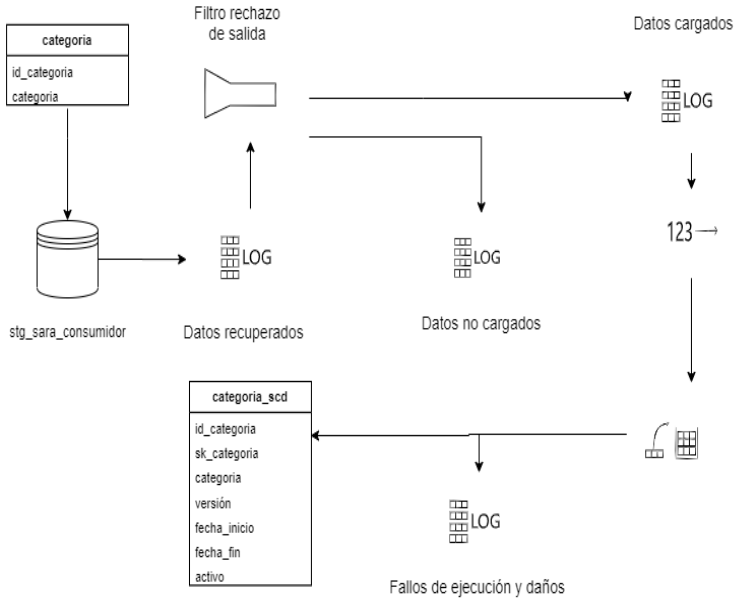


Figura 72 Dimensión categoría.

La Figura 72 muestra uno de los diseños UML de los procesos ETL este corresponde la dimensión categoría en el cual el proceso que se sigue se lista a continuación:

1. Conecta a la base de datos de origen en este caso el staging_area.
2. Une las tablas que conformarán la dimensión categoría, en este caso una única tabla categoría.
3. Reporta al Log los datos recuperados.
4. Filtra para datos cargados y no cargados
5. Reporta al Log datos cargados y datos no cargados.
6. Calcula clave sustituta.
7. Carga datos a la tabla de destino categoria_scd.

14.3.1.3.4 Desarrollo de procesos ETL

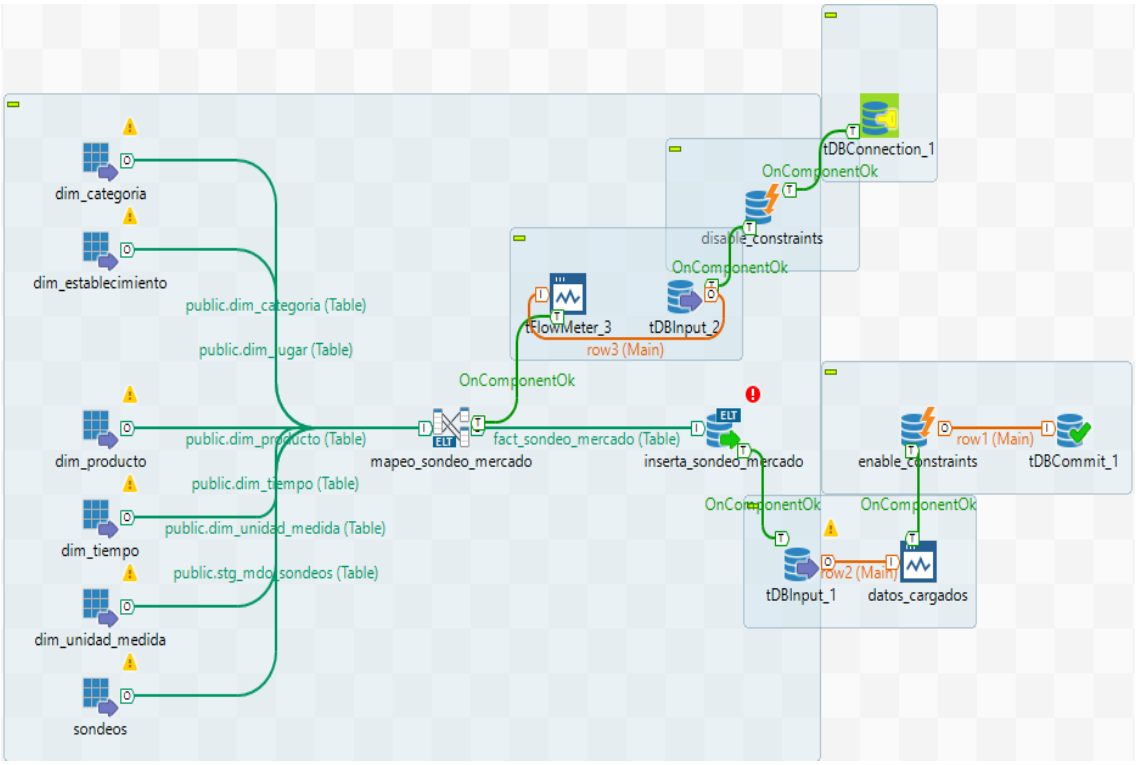


Figura 73 Job fact_mercados.

La Figura 73 muestra el desarrollo de uno de los flujos de trabajo en este caso para el hecho a medir en el presente sprint en el cual se sigue el proceso enumerado a continuación:

1. Se establece la conexión con la base de datos.
2. Se truncan los datos de la tabla de hechos y se desactiva la integridad referencial.
3. Se reportan al log los datos recuperados de la tabla de normalización del staging área.
4. Se unen los datos de las dimensiones con la tabla de normalización del staging área para determinar las llaves sustitutas en la tabla de hechos.
5. Se almacenan los datos en la tabla de hechos y se activa nuevamente la integridad referencial.
6. Se reportan al log los datos almacenados en la tabla de hechos.

14.3.1.3.5 Pruebas

```
.-----------------+------+---------+--------+----------+-------+-----------------------+----------------------+-----------+-------+------------
+-----------------+------+---------+----------.
|                                                                                                 STAGING_MERCADO
                    I     |
|=----------------+------+---------+--------+----------+-------+-----------------------+----------------------+-----------+-------+------------
+-----------------+------+---------+---------=|
|moment           |pid   |father_pid|root_pid|system_pid|project|job                    |job_repository_id     |job_version|context|origin       |label
|count |reference|thresholds|
|=----------------+------+---------+--------+----------+-------+-----------------------+----------------------+-----------+-------+------------
+-----------------+------+---------+---------=|
|2020-02-11 22:26:01|UL32bH|93Y0N9    |93Y0N9  |1836      |SONDEOS|configuraciones_iniciales|_Xqc-YDTCEeqZiYBjnqy9oA|0.2        |Default|tFlowMeter_38|Datos
recuperados|973081|null     |          |
|2020-02-11 22:26:01|UL32bH|93Y0N9    |93Y0N9  |1836      |SONDEOS|configuraciones_iniciales|_Xqc-YDTCEeqZiYBjnqy9oA|0.2        |Default|tFlowMeter_35|Datos cargados
|0     |null     |          |
|2020-02-11 22:35:27|UL32bH|93Y0N9    |93Y0N9  |1836      |SONDEOS|configuraciones_iniciales|_Xqc-YDTCEeqZiYBjnqy9oA|0.2        |Default|tFlowMeter_43|Datos no
cargados|3694  |null     |          |
'-----------------+------+---------+--------+----------+-------+-----------------------+----------------------+-----------+-------+------------
+-----------------+------+---------+---------'
```

Figura 74 Ejecución Job configuraciones_iniciales.

La Figura 74 presenta los resultados para la ejecución del Job, en el cual se crearon las siguientes tablas con los datos históricos de precios del mercado y los datos del sondeo Mercado, así como el correcto funcionamiento del registro de Log.

La Figura 75 muestra los datos que fueron recuperados desde el Staging área y contrastando los que fueron cargados al modelo multidimensional y los que no fueron cargados.

stg_mdo_sondeos

	id_sondeo	nombre_sondeo	id_categoria	id_producto	cantidad	id_um	precio	precio_referencia
1	1	Sondeo Mayorista de Granos	1	1	1	1	38	1.000
2	2	Sondeo Mayorista de Granos	1	2	1	1	40	1.000
3	3	Sondeo Mayorista de Granos	1	3	1	1	30	1.000
4	4	Sondeo Mayorista de Granos	1	4	1	1	14	1.000
5	5	Sondeo Mayorista de Granos	1	5	1	1	43	1.000
6	6	Sondeo Mayorista de Granos	1	5	1	2	0,4499999881	1.000
7	7	Sondeo Mayorista de Granos	1	6	1	1	46	1.000
8	8	Sondeo Mayorista de Granos	1	6	1	2	0,5	1.000
9	9	Sondeo Mayorista de Granos	1	1	1	1	38	1.000
10	10	Sondeo Mayorista de Granos	1	1	1	2	0,400000006	1.000
11	11	Sondeo Mayorista de Granos	1	2	1	1	33	1.000
12	12	Sondeo Mayorista de Granos	1	2	1	2	0,400000006	1.000
13	13	Sondeo Mayorista de Granos	1	3	1	1	31	1.000
14	14	Sondeo Mayorista de Granos	1	3	1	2	0,400000006	1.000

fact_sondeo_mercado

	id_sondeo	nombre_sondeo	sk_categoria	sk_producto	cantidad	sk_um	precio	precio_referencia
1	1	Sondeo Mayorista de Granos	1	1	1	1	38	1.000
2	2	Sondeo Mayorista de Granos	1	2	1	1	40	1.000
3	3	Sondeo Mayorista de Granos	1	3	1	1	30	1.000
4	4	Sondeo Mayorista de Granos	1	4	1	1	14	1.000
5	5	Sondeo Mayorista de Granos	1	5	1	1	43	1.000
6	6	Sondeo Mayorista de Granos	1	5	1	2	0,4499999881	1.000
7	7	Sondeo Mayorista de Granos	1	6	1	1	46	1.000
8	8	Sondeo Mayorista de Granos	1	6	1	2	0,5	1.000
9	9	Sondeo Mayorista de Granos	1	1	1	1	38	1.000
10	10	Sondeo Mayorista de Granos	1	1	1	2	0,400000006	1.000
11	11	Sondeo Mayorista de Granos	1	2	1	1	33	1.000
12	12	Sondeo Mayorista de Granos	1	2	1	2	0,400000006	1.000
13	13	Sondeo Mayorista de Granos	1	3	1	1	31	1.000
14	14	Sondeo Mayorista de Granos	1	3	1	2	0,400000006	1.000

Figura 75 Muestra los datos de las dos tablas.

14.3.2 Minería de datos

14.3.2.1 Exploración de los datos

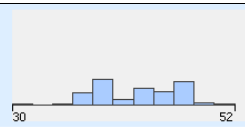
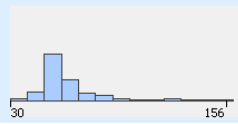
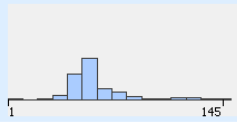
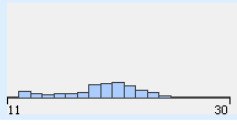
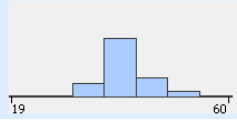
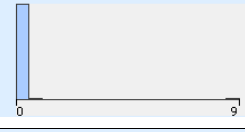
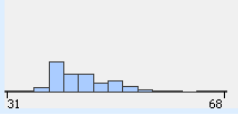
PRODUCTO	Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
AZUCAR	precio	30	42.0518	42	52	3.623	0.0054	-1.5348	0	0	0	
FRIJOL ROJO DE SEDA	precio	30	61.746	57	156.26	16.7867	2.3203	6.515	0	0	0	
FRIJOL TINTO	precio	0.55	57.1312	52.68	145	16.943	2.2103	5.8429	0	0	0	
MAÍZ	precio	11	18.8088	19	30	3.2296	-0.4262	-0.0967	0	0	0	
ARROZ BLANCO	precio	19	40.5895	40	60	4.1486	0.9072	0.6339	0	0	0	
ARROZ BLANCO (MINORISTA)	precio	0.25	0.4929	0.5	8.75	0.0646	76.8787	9,805.8127	0	0	0	
ARROZ PRECOCIDO (MAYORISTA)	precio	31	43.4167	42	67.65	4.5504	0.8032	0.2906	0	0	0	

Tabla 29 Comparativa de estadísticas por representativos en productos de mercado.

14.3.2.2 Series temporales

14.3.2.2.1 Descubrir la relación entre los precios de los productos

14.3.2.2.1.1 Contenido del caso.

N° de caso: C-FOR-01	
Técnica	Forecast
Algoritmos	Series Temporales
Población	Datos provenientes del modelo multidimensional de precios de los sondeos de mercados, datos desde 2013 a 2019
Variables	Se analiza el precio promedio del producto en base a la fecha de sondeo realizada.
Hipótesis	Predecir el comportamiento del precio del producto en el futuro.
Procedimiento	Flujo de trabajo en Knime, Programación en Python y R.
Resultados	Gráfico temporal, tendencia, estacionalidad.
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión.
Herramienta de software	Knime Analytics Platform, R, Python

Tabla 30 Contenido del caso C-FOR-01.

14.3.2.2.1.2 Población

La población de datos utilizada, es proveniente del modelo multidimensional y corresponde a una consulta donde se obtienen todos los precios promedios por mes de un producto ordenado por año y mes.

	precio	year	month
1	0.9051118	2013	10
2	0.9365591	2013	11
3	0.8901163	2013	12
4	0.7641304	2014	1
5	1.0599541	2014	2
6	0.9475369	2014	3
7	0.9173810	2014	4
8	0.9159341	2014	5
9	0.8502620	2014	6

Figura 76 Datos extraídos del modelo multidimensional.

14.3.2.2.1.3 Variables

Las variables que se ven involucradas en la exploración son: Precio, Año y Mes.

14.3.2.2.1.4 Hipótesis

El objetivo de minería de datos que queremos llevar a cabo o la hipótesis que deseamos comprobar es: “Determinar la estacionalidad de los productos de las categorías frutas y verduras”.

14.3.2.2.1.5 Procedimiento

1. Selección de los datos:

Se establece la conexión con la base de datos y se seleccionan los datos del modelo multidimensional mediante una consulta SQL.

2. Generar un modelo por cada producto:

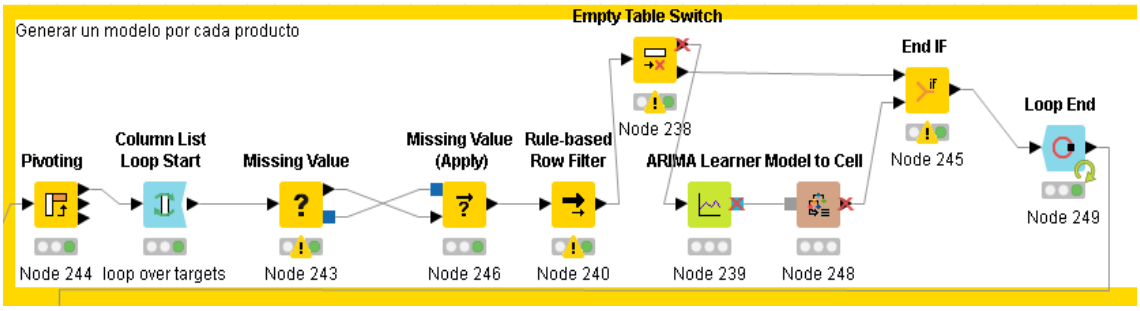


Figura 77 Bloque 2 del flujo de trabajo C-FOR-01.

La Figura 77 muestra el bloque del flujo de trabajo en el cual se genera un modelo por cada producto, se inicia pivoteando los datos presentando cada producto como una columna siendo las filas los precios por cada fecha sondeada, luego se inicia un bucle manejando los valores faltantes y realizando una condición en la cual si el producto de la iteración actual devuelve una tabla vacía pasa a la siguiente iteración sino calcula el modelo de series temporales con el algoritmo ARIMA, se convierte dicho modelo en una celda y pasa a la siguiente iteración hasta finalizar todos los productos.

3. Renombrar los modelos por el nombre correspondiente de cada producto:

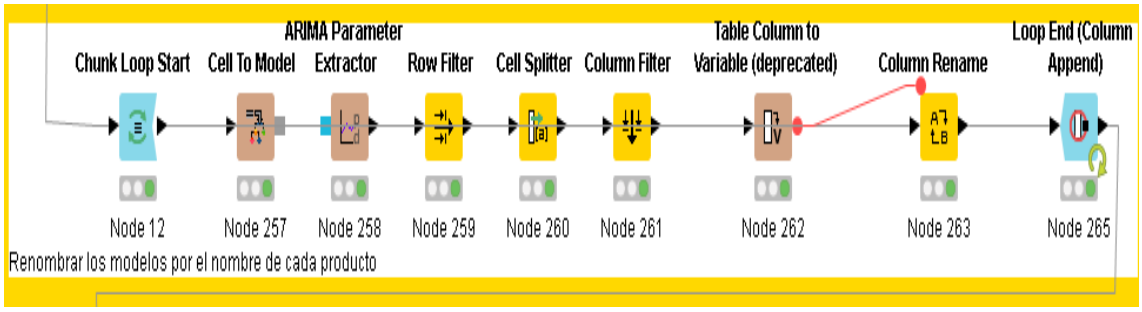


Figura 78 Bloque 3 del flujo de trabajo C-FOR-01.

La Figura 78 muestra el bloque del flujo de trabajo en el cual se renombran los modelos por el nombre de cada producto, se inicia el bucle de tipo “chunk” que aborda los datos provenientes del flujo de datos registro a registro, convirtiendo posteriormente cada celda a modelo, se extraen los parámetros del modelo de ARIMA, se filtra el que contiene el nombre del producto, se genera una variable a partir de ese valor y luego se renombra la columna que contiene el modelo por el nombre del producto, luego pasa a la siguiente iteración.

4. Generar predicciones por cada producto:

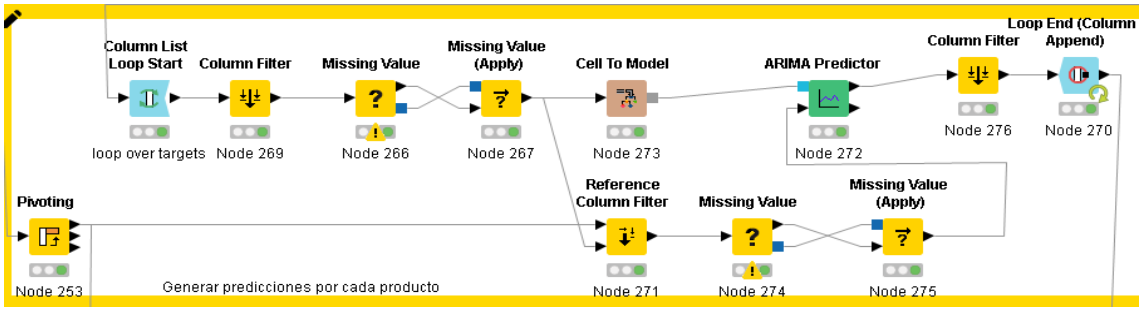


Figura 79 Bloque 4 del flujo de trabajo C-FOR-01.

La Figura 79 muestra el bloque del flujo de trabajo en el cual se generan las predicciones para cada producto se inicia el bucle, se filtra la columna que contiene el modelo, se manejan los valores faltantes, se convierte la celda a modelo, se filtran los datos a ser sujeto de la predicción para el producto de la iteración actual y entran al predictor de ARIMA posterior a ello se filtra la columna que contiene el forecast (predicción) y pasa a la siguiente iteración.

5. Asignar las fechas a los datos originales y predicciones:

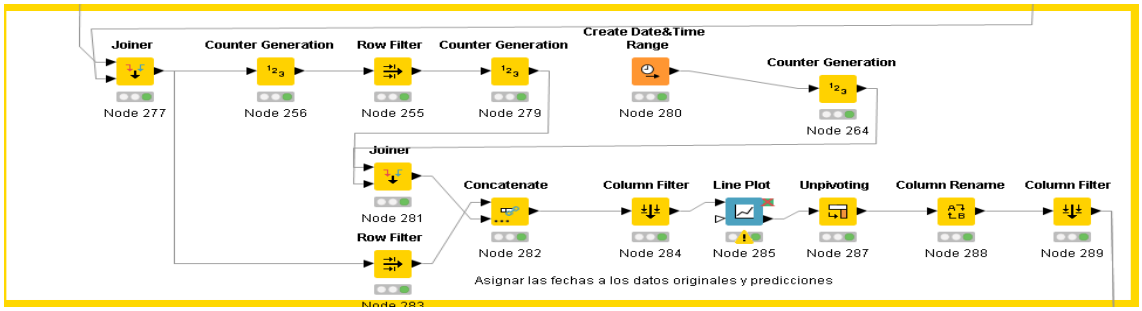


Figura 80 Bloque 5 del flujo de trabajo C-FOR-01.

La Figura 80 muestra el bloque del flujo de trabajo en el cual se asignan las fechas a los datos originales y las predicciones, primero se concatenan los datos del forecast provenientes del flujo de datos con los datos originales, se genera un contador y se filtran solo los valores predichos, se crea un rango de fechas y se genera igualmente un contador, se unen ambos conjuntos de datos mediante su contador y luego se concatenan con el resto de datos originales, se genera como primera salida un gráfico de línea, y luego se hace un pivotaje inverso en el cual los datos vuelven a la forma de registros.

6. Salidas a la base de datos:

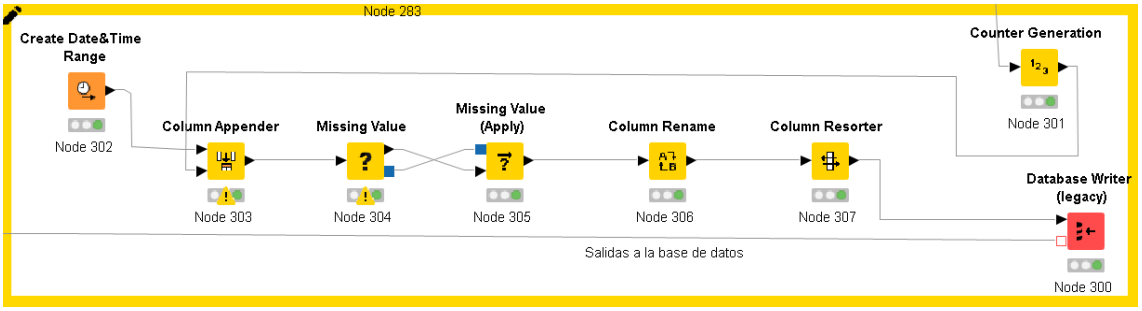


Figura 81 Bloque 6 del flujo de trabajo C-FOR-01.

En la Figura 81 se muestra el bloque del flujo de trabajo en el cual se generan las salidas a la cual se escriben los datos provenientes del flujo de datos a la base de datos, se genera un counter para los identificadores de los registros y se obtiene la fecha actual del sistema, y se procede a insertar los datos.

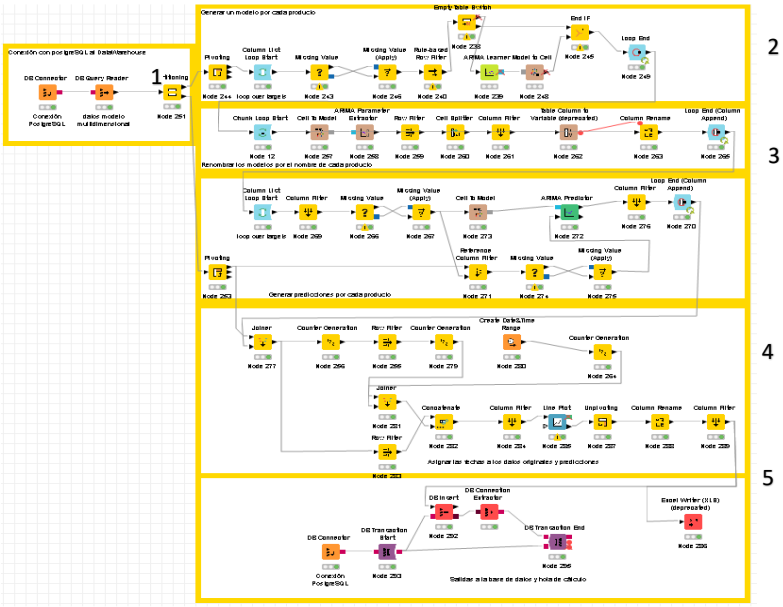


Figura 82 Señalización de las partes del flujo de trabajo para el caso C-FOR-01.

Para la estacionalidad se hace uso de la herramienta R debido a que KNIME no cuenta con nodos especiales para esto.

Se agrega las librerías siguientes:

1. *lubridate*: Permite trabajar con funciones para realizar análisis rápido y fácil, extracción y actualización de datos de fecha y hora, así como años, meses, días, horas, minutos y segundos.
2. *tidyverse:* Es una colección obvia de paquetes R diseñados para la ciencia de datos. Todos los paquetes comparten una filosofía de diseño, gramática y estructuras de datos subyacentes.
3. *tsibble*: Proporciona una clase 'tbl_ts' para datos temporales en un formato orientado a datos y modelos.

4. *feasts:* Ofrece una gran variedad de herramientas para analizar datos temporales ordenados en el formato tsibble.
5. *fable:* Proporciona métodos y herramientas para mostrar y analizar pronósticos de series de tiempo.

```
dataset$fecha <- as.Date(dataset$fecha , "%Y-%m-%d")

precio_fecha <- dataset

precio <- tsibble(
  Producto = precio_fecha$producto,
  Fecha = precio_fecha$fecha,
  Precio = precio_fecha$avg_precio,
  key = Producto,
  index = Fecha
)
```

Figura 83 Creación del objeto tsibble.

Se transforma la fecha del formato 'Y-m-dThh:ii:ss' al 'Y-m-d' para poder ser reconocible para la clase 'tbl_ts' en donde se crea un diccionario seteando las variables key e index, la primera define el único índice de tiempo y la segunda especifica la variable de índice del tiempo.

```
precio <- precio %>%
  group_by(Producto) %>%
  index_by(year_month = ~ yearmonth(.)) %>%
  summarise(
    Precio = mean(Precio, na.rm = TRUE)
  )
```

Figura 84 Realización del cálculo del promedio del precio de cada producto agrupado en base al mes y año.

Se agrupa en base al mes y año generando una nueva variable de tipo yearmonth, y en base a la agrupación de Producto se realiza un promedio del precio de cada producto. Lo siguiente será graficar la estacionalidad de cada producto. Con la función:

$$gg\_subseries(precio)$$

Ecuación 17 Implementación de la gráfica de estacionalidad para un producto.

14.3.2.2.1.6 Resultados

Se logró generar predicciones para múltiples productos de forma dinámica auxiliándose del algoritmo de series temporales ARIMA y el uso de bucles.

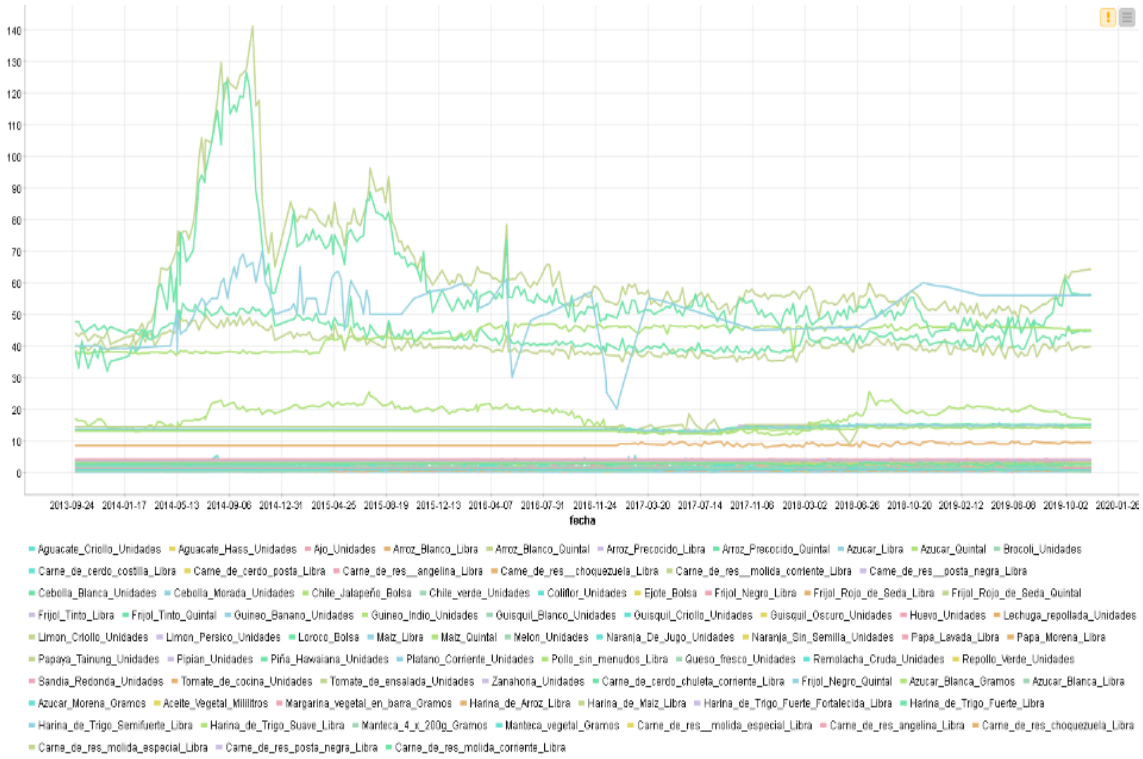


Figura 85 Gráfico de línea de las predicciones generadas por productos de mercados.

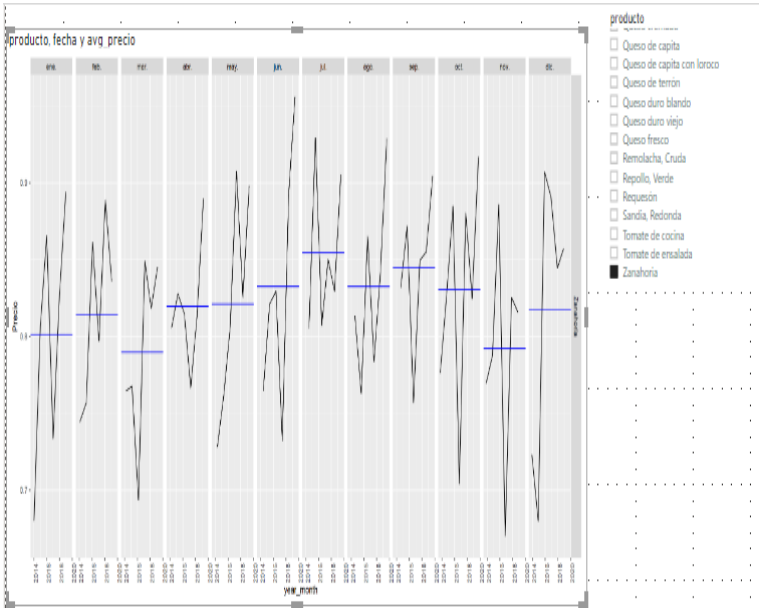


Figura 86 Estacionalidad de un producto (Zanahoria).

14.3.2.2.1.7 Interpretación de resultados

Es posible predecir para múltiples productos si se trabaja de forma iterativa, es de tener muy en cuenta que no todos los productos llevan el mismo comportamiento para lo cual si se quisiera una precisión más exacta se debería trabajar uno a uno, sin embrago la opción propuesta arroja resultados aceptables.

La estacionalidad del producto en este caso Zanahoria, podemos visualizar que las líneas horizontales representan las medias para cada mes, se permite ver claramente el patrón estacional subyacente y muestra los cambios en la estacionalidad a lo largo del tiempo el cual varía entre los 0.7 y los 0.9 y la media ubicada en los 0.8. En este se puede evidenciar un cambio estacional positivo en el mes de junio y uno negativo en noviembre y marzo.

14.3.2.3 Técnica de agrupamiento

14.3.2.3.1 Agrupar los meses en los cuales se identifican aumentos o bajas en los precios

14.3.2.3.1.1 Contenido del caso

N° de caso: C-AGR-01	
Técnica	Agrupamiento.
Algoritmos	K-means Clustering
Población	Datos provenientes del modelo multidimensional de precios de los sondeos de Mercados.
Variables	El mes, precio y producto.
Hipótesis	Realizar Agrupamiento de precios promedio de productos por mes para identificar en que meses se concentran los precios más bajos o más altos.
Procedimiento	Programación en Python
Resultados	Grupos de Precios
Interpretación de resultados	Evaluar en base a los diferentes grupos de precios obtenidos por los algoritmos
Herramienta de software	Python

Tabla 31 Contenido del caso C-AGR-01.

14.3.2.3.1.2 Población

La población de datos utilizada, es proveniente del modelo multidimensional y corresponde a una consulta donde se obtienen todos los precios promedios por año agrupados por mes

mes	precio_promedio
1	0,7271768993
2	0,7144816577
3	0,7297134816
4	0,7454738196
5	0,7807935277
6	0,7933202263
7	0,8631882777
8	0,9161750857
9	0,9459049114
10	0,8143189004
11	0,79084987
12	0,676512346

Figura 87 Población de Datos

14.3.2.3.1.3 Variables

Las variables involucradas en el análisis son:

- El Precio
- Los meses
- El producto

14.3.2.3.1.4 Hipótesis

El objetivo de minería de datos que se quiere llevar a cabo o la hipótesis que se desea comprobar es: Agrupar por mes el precio promedio del Producto para posteriormente poder identificar en cuales meses del año están concentrados los precios más bajos o en cuales se encuentran los precios más altos.

14.3.2.3.1.5 Procedimiento

1. Se importaron las siguientes librerías de Python
    - Numpy: librería de Python, que le agrega mayor soporte para vectores y matrices
    - Pandas: librería que nos sirve para manipular tablas de datos.
    - Matplotlib.pyplot: librería que contiene una colección de funciones para que se trabajen similar a Matlab.
    - Seaborn: librería de visualización basada en matplotlib.
    - Kmeans: librería de sklearn que implementa el algoritmo kmeans.
    - Psycopg2: librería para PostgreSQL.

Librerias

```
[ ]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     sns.set()
     from sklearn.cluster import KMeans
     import psycopg2
```

Figura 88 Librerías de Python

2. Se carga la población de datos desde el modelo multidimensional por medio de una conexión con postgreSQL usando la librería psycopg2 y el resultado se almacena en un Dataframe de pandas.

```
In [2]: conn = psycopg2.connect(database = "mercado", user = "postgres", password = "admin", host="localhost", port="5433" )
        data = pd.read_sql('select dt.num_mes,avg(sm.precio) precio from sondeo_mercado sm join dim_producto dp on dp.sk_producto=sm.sk_

In [3]: data.head()

Out[3]:
```

	num_mes	precio
0	1	0.727177
1	2	0.714482
2	3	0.729713
3	4	0.745474
4	5	0.780794

Figura 89 Población de Datos

3. Se Graficó los datos cargados para poder visualizar de mejor manera

```
plt.scatter(data['num_mes'],data['precio'])
plt.xlabel('Mes')
plt.ylabel('Precio Promedio')
plt.show()
```

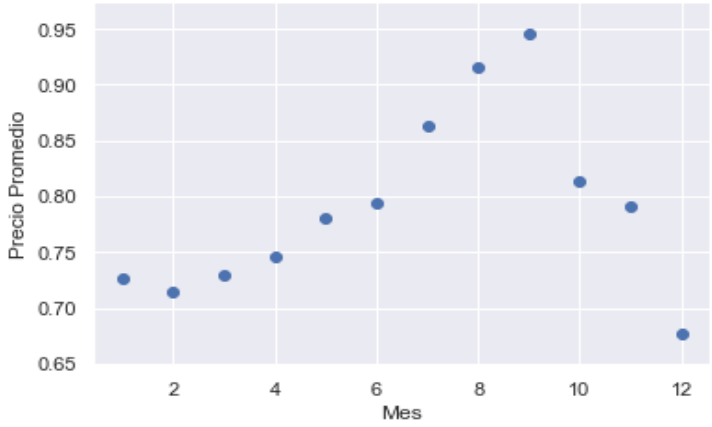


Figura 90 Grafico de Población de datos

4. Luego se implementa el Algoritmo de Kmeans de la librería sklearn. clúster indicándole cuantos clústeres queremos obtener en el resultado como se muestra en la Figura 91 se ingresó 3.

```
x=data.copy()
```

```
kmeans=KMeans(3)
kmeans.fit(x)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

Figura 91 Implementación de Kmeans

14.3.2.3.1.6 Resultados

Los resultados obtenidos son tres grupos de precios para los diferentes meses del año

- Grupo de precios número uno que son los que están conformados entre los meses de enero y abril.
- Grupo de Precios numero dos que son los que están conformados entre los meses de mayo y agosto.
- Grupo de precios número tres que son lo que están conformados entre septiembre y diciembre.

Clustering Result

```
clusters=x.copy()
clusters['cluster_pred']=kmeans.fit_predict(x)
```

Plot

```
plt.scatter(clusters['num_mes'],clusters['precio'],c=clusters['cluster_pred'],cmap='rainbow')
plt.xlabel('Mes')
plt.ylabel('Precio Promedio')
plt.show()
```

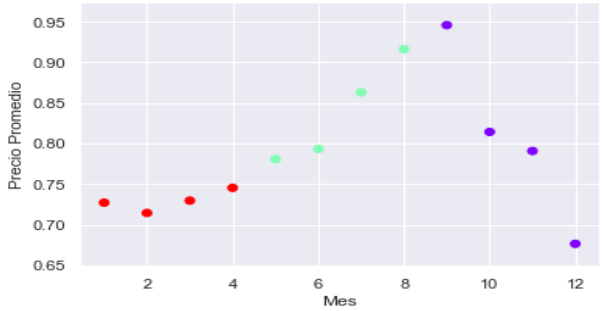


Figura 92 Resultados caso C-AGR-01

14.3.2.3.1.7 Interpretación de resultados

En la Figura 92 se observa los diferentes grupos para los diferentes meses, el primer grupo de color rojo representa los meses donde el producto tuvo precios que estuvieron en el promedio mientras que en el grupo de color verde y morado representa los meses donde se tuvieron variación de precios con respecto al promedio.

14.3.3 Visualización

14.3.3.1 Informes de inteligencia de negocios

La Figura 93 se visualizar la diferencia en el precio de los productos en términos porcentuales y monetarios con un filtro por fecha comparándolo al día, semana y mes anterior. Para las categorías Aceite y Lácteos, Frutas, Huevos y Carne, y Verduras.

Categoria	Producto	Precio	Semana Anterior	Mes Anterior	Año Anterior	Diferencia Semanal	Diferencia Semanal %	Diferencia Mensual	Diferencia Mensual %	Diferencia Anual	Diferencia Anual %
Aceites	Aceite Vegetal	1.33	1.37	1.34	1.38	-0.04	-2.84	-0.01	-0.82	-0.05	-3.41
Lacteos	Queso fresco	2.04	2.20	2.02	2.11	-0.16	-7.92	0.01	0.63	-0.08	-3.82
Margarinas y mantecas	Manteca vegetal	0.84	0.94	0.87	0.85	-0.10	-11.75	-0.03	-3.28	-0.01	-1.05
Margarinas y mantecas	Margarina vegetal en barra	1.28	1.27	1.26	1.29	0.01	0.63	0.02	1.36	-0.01	-0.89

Figura 93 Diseño de la interfaz gráfica informe de variación.

14.3.3.2 Informes de resultados minería de datos

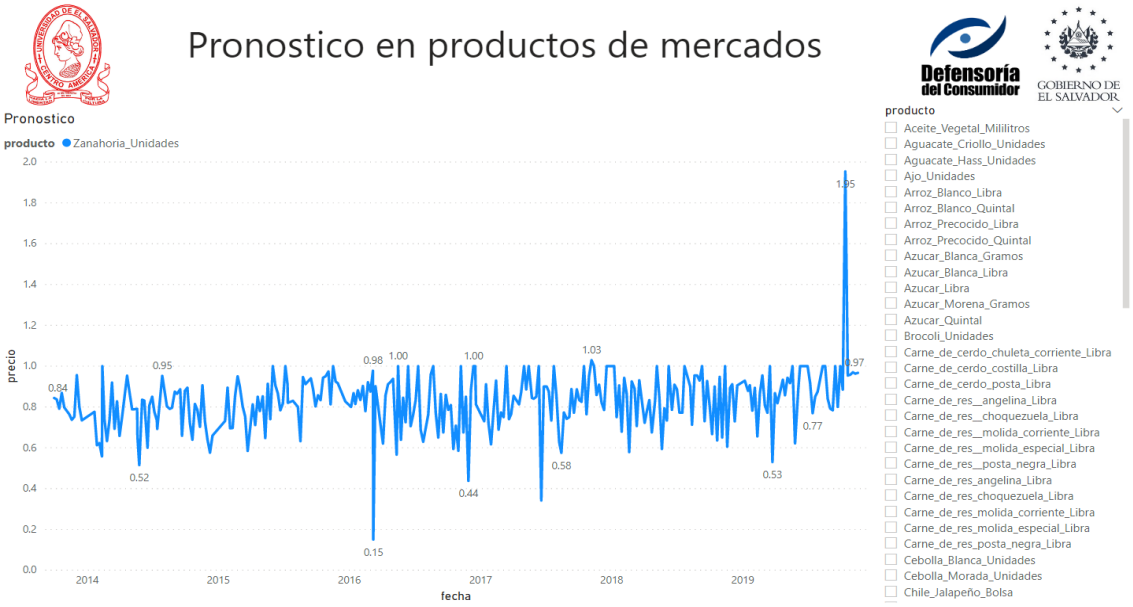


Figura 94 Reporte de minería para C-FOR-01.

Para el reporte de forecast en la Figura 94, los datos se proyectan en una gráfica de serie temporal que muestra los distintos valores del precio del producto en los años anteriores y una proyección de 4 semanas. Asimismo, en el lado izquierdo se proyecta el filtro que se puede realizar a la proyección.

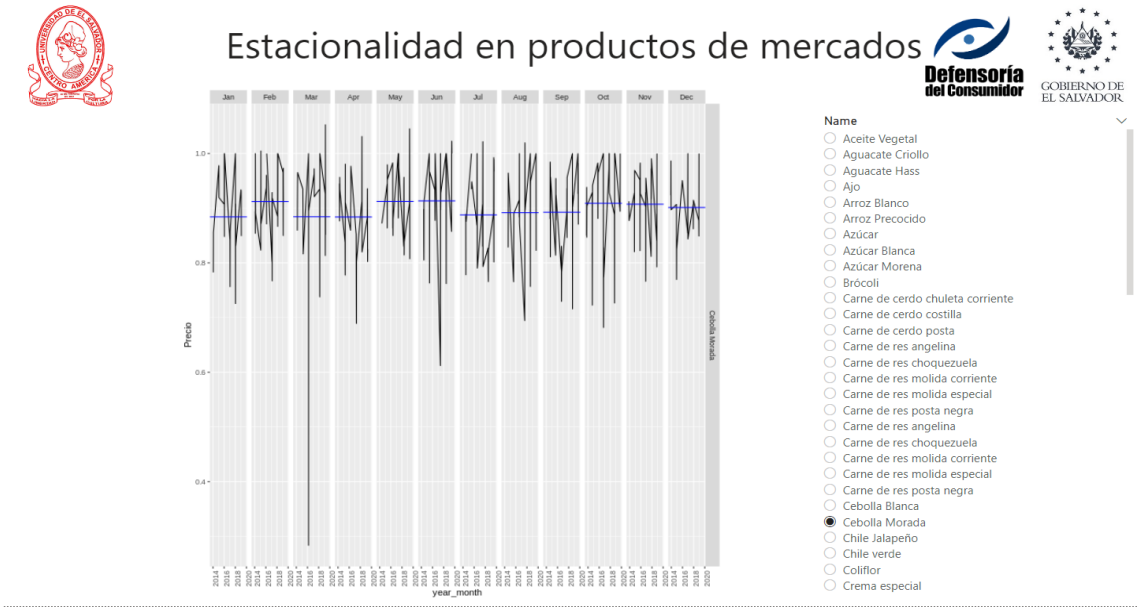


Figura 95 Reporte de minería para la estacionalidad C-FOR-01.

Para el reporte de estacionalidad en la Figura 95, los datos se proyectan en una gráfica que muestra la estacionalidad que tiene un producto determinado de cada mes por los distintos años a evaluar, se puede observar el precio medio del producto en los 12 meses del año. Asimismo, en el lado izquierdo se proyecta el filtro que se puede realizar a la proyección.

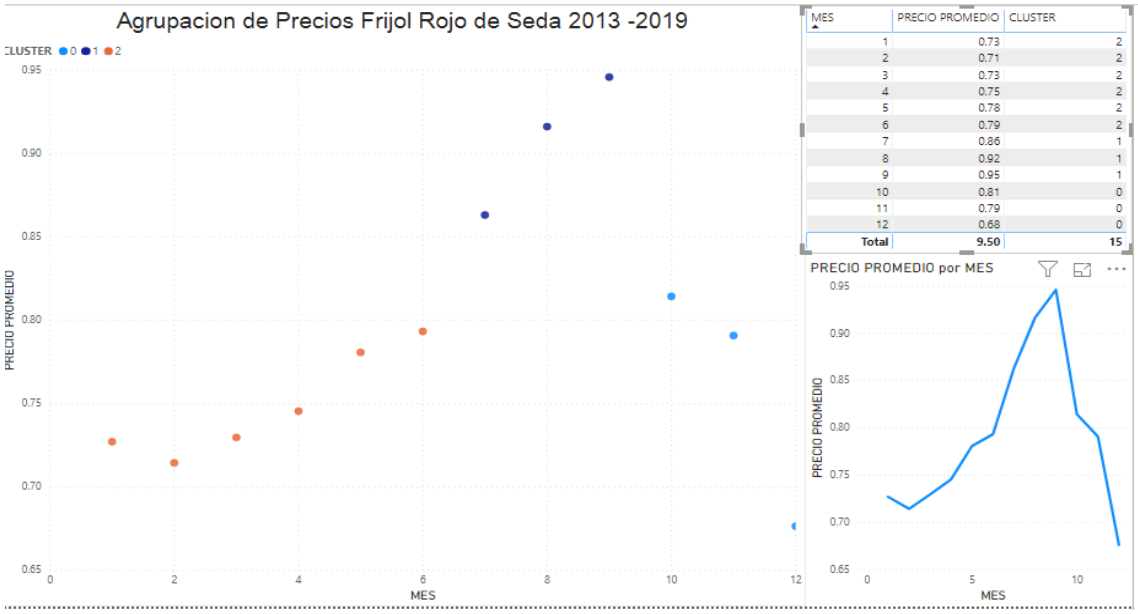


Figura 96 Informe Agrupación para Frijol Rojo de Seda

El informe está compuesto por dos gráficos y una tabla, la primera grafica muestra cómo se agrupan los precios en los diferentes meses del año representados en el eje X, el precio promedio representado en el eje Y, el segundo grafico muestra la tendencia de los precios promedio por mes y la última parte una tabla que muestra el detalle de los meses y el precio promedio.

15 Sprint 3

15.1 Descripción Historias de Usuario

Código	RSPS01
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea guardar la información de las bases de datos en estructuras de datos más robustas.
Razón	Para poder obtener de mejor manera los datos y poder generar información más rápido.
Criterios de aceptación	Se deberá crear un almacén de datos.
	Se deberán crear las dimensiones necesarias.
	Se deberá tener al menos una tabla de hechos.
Validación	Se comprobará que el almacén de datos sea respetando un esquema como estrella, constelación o copo de nieve.
	Se comprobará que el almacén de datos deberá tener las dimensiones necesarias como la dimensión tiempo, precios o productos.
	Se comprobará que el almacén de datos deberá contener una tabla de hechos.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 32 Historia de Usuario RSPS01.

Código	RSPS02
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar mediante un gráfico los precios de los productos de supermercados.
Razón	Para tener una visión general de sus fluctuaciones y así, detectar anomalías.
Criterios de aceptación	Se debe considerar cualquier producto de supermercado
	Se debe considerar un rango de fechas, este será variable, y deberá ser elegido por el usuario.
	Se deberá permitir además filtros como la cadena y sucursal.
Validación	Se comprobará que se muestre la información correcta para el producto seleccionado.
	Se comprobará que el rango de fechas sea válido
	Se comprobará que, al realizar un cambio en los filtros, este se realice con la información correcta.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 33 Historia de Usuario RSPS02.

Código	RSPS03
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar mediante un gráfico los precios de los productos de supermercados desde distintas variables.

Razón	Para tener una visión general de sus fluctuaciones y así, detectar anomalías.
Criterios de aceptación	Debe considerar cualquier categoría de producto de supermercado.
	Debe considerar si la categoría contiene una sub categoría
	Deberá permitir diferentes filtros como la cadena, sucursal y las fechas en las que se hará el análisis.
Validación	Comprobar que se muestre la información para la categoría seleccionada.
	Comprobar que se muestre la información correcta para la sub categoría.
	Comprobar que, al realizar un cambio en los filtros, este se realice con la información correcta.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 34 Historia de Usuario RSPS03.

Código	RSPS04
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Mostrar los precios de los productos por municipio.
Razón	Para tener una perspectiva más detallada de los precios por municipio.
Criterios de aceptación	Se debe considerar cualquier producto de supermercado y mercado.
	Se debe mostrar el precio de los productos en un mapa.
	Se deberá permitir diferentes filtros como el departamento, municipio, producto, unidad de medida u otros.
Validación	Se comprobará que se muestre la información correcta para el producto seleccionado.
	Que los municipios sean claramente identificables dentro del mapa.
	Que al colocar el puntero sobre el municipio se visualice la información correspondiente.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 35 Historia de Usuario RSPS04

Código	RSPS05
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Mostrar los precios de los productos por región.
Razón	Para tener una perspectiva más detallada de los precios por región.
Criterios de aceptación	Se debe considerar cualquier producto de supermercado y mercado.
	Se debe mostrar el precio de los productos en un mapa según su región.
	Se deberá permitir diferentes filtros como la región, municipio, producto, unidad de medida u otros.

Validación	Se comprobará que se muestre la información correcta para el producto seleccionado.
	Que las regiones sean claramente identificables dentro del mapa.
	Que al colocar el puntero sobre el municipio se visualice la información correspondiente.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 36 Historia de Usuario RSPS05.

Código	**RSPS05**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Mostrar los precios de los productos por departamento, región y establecimiento.
Razón	Para tener una perspectiva más detallada de los precios por departamento, región y establecimiento.
Criterios de aceptación	Se debe considerar cualquier producto de supermercado y mercado.
	Se debe mostrar el precio de los productos en un mapa
	Se deberá permitir diferentes filtros como el departamento, establecimiento, producto, unidad de medida u otros.
Validación	Se comprobará que se muestre la información correcta para el producto seleccionado.
	Que los establecimientos sean claramente identificables dentro del mapa.
	Que al colocar el puntero sobre el municipio se visualice la información correspondiente.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 37 Historia de Usuario RSPS06.

15.2 Refinamiento del requerimiento de información

15.2.1 Proceso BPMN

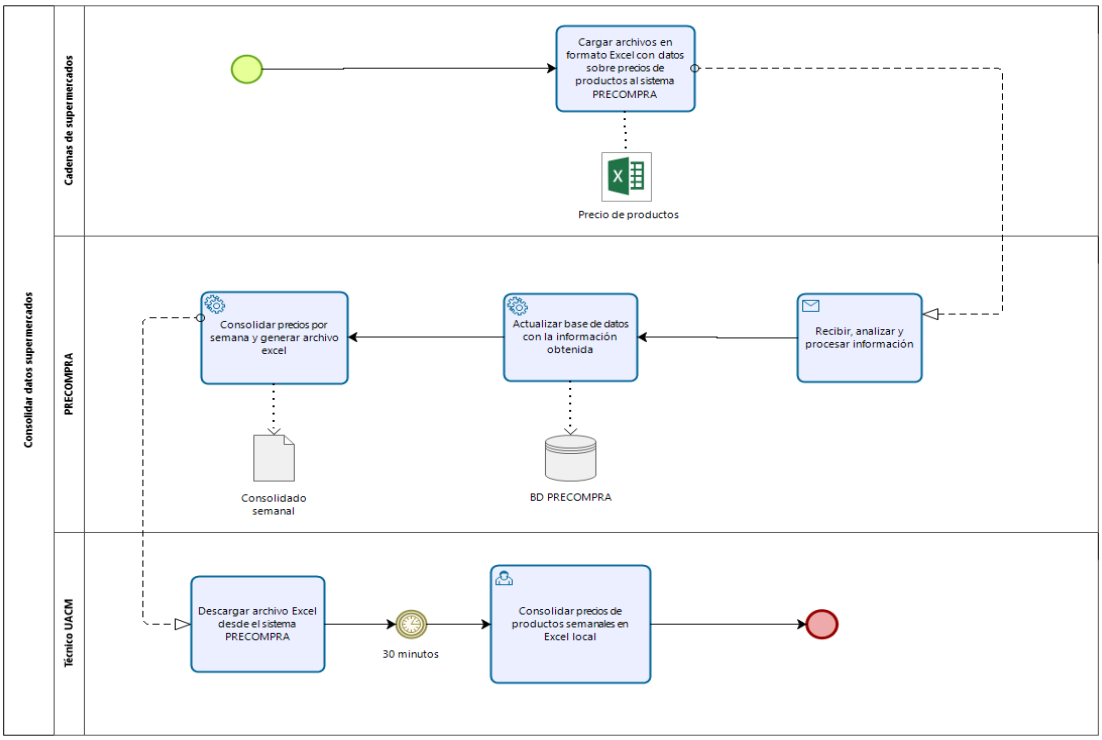


Figura 97 Proceso de consolidación de datos.

15.2.2 Paquetes de Información

Tema:	Variaciones de precios de productos de supermercados				
JERARQUIAS	**Tiempo**	**Sucursal**	**Producto**	**Categoría**	**Subcategoría**
	Año	Cadena	Producto	Categoría	Subcategoría
	Mes	Sucursal	Contenido neto		
	Día		Marca		
	Semana				
Hechos Medidos:	**Comportamiento de precios de supermercados (Medida calculada)**				

Tabla 38 paquete de información de Supermercados.

15.2.3 Casos de uso

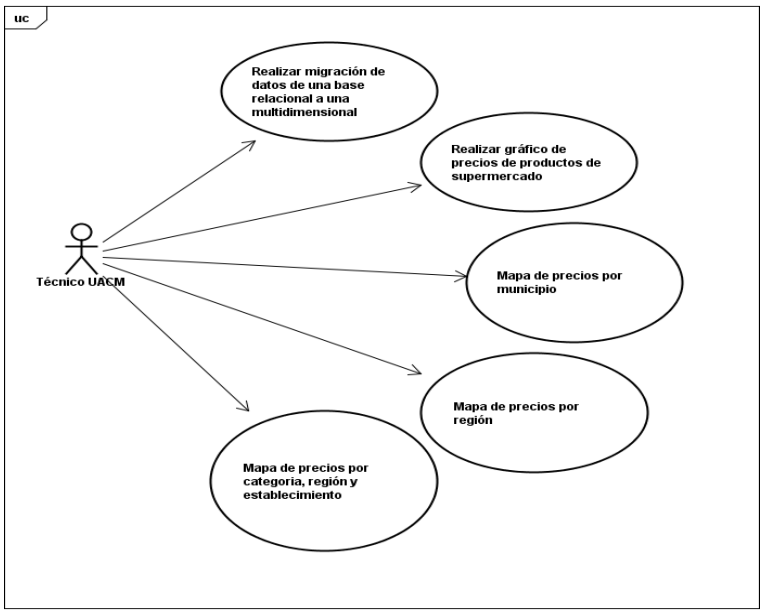


Figura 98 Diagrama de Casos de Uso Sondeo de Precios de Supermercados.

15.3 Desarrollo de la iteración

15.3.1 Integración de los datos

15.3.1.1 Extracción de los datos

Para la extracción de los datos para el ETL de supermercados, se hace desde la base de datos transaccional, dicha base de datos está alojada en un servidor virtual y en el motor de base de datos SQL Server 2016.

Las consideraciones a tomar en cuenta para las fuentes de datos son:

1- Los datos son leídos como caracteres y se hacen las conversiones necesarias en el integrador de datos para evitar trabajo manual.
2- Los datos no son homogéneos por lo tanto dicha homogenización se realiza desde el integrador de datos.
3- Es necesario convertir la fecha al formato que la base de datos la puede capturar.
4- La variación no se almacenará si no que se calculará posteriormente en los informes.
5- Algunos nombres de productos no coinciden con los que se manejan en el consolidado por lo tanto es necesario homogenizar.
6- Existen sondeos de precios que no se asigna, estos no son tomados en cuenta.
7- Es necesario validar en la base de datos si ya existe el producto, plaza o unidad de medida para no repetir en los catálogos.

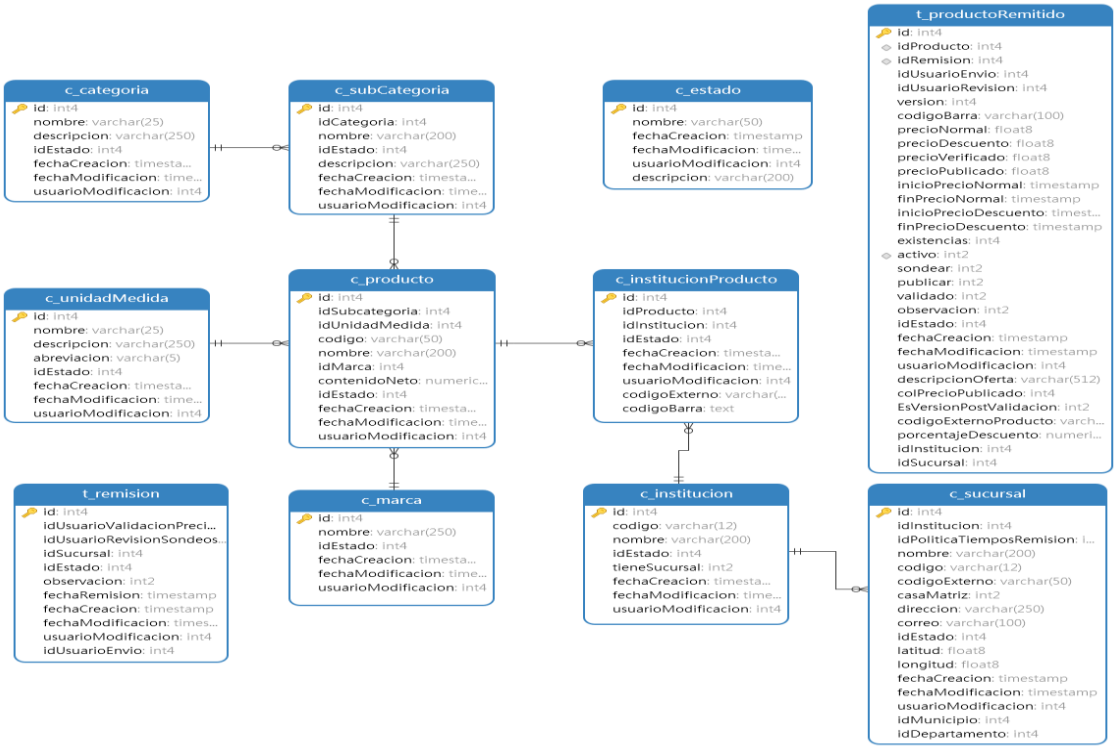


Figura 99 Modelo relacional de SPRS.

15.3.1.2 Staging área

15.3.1.2.1 Diseño del modelo staging área

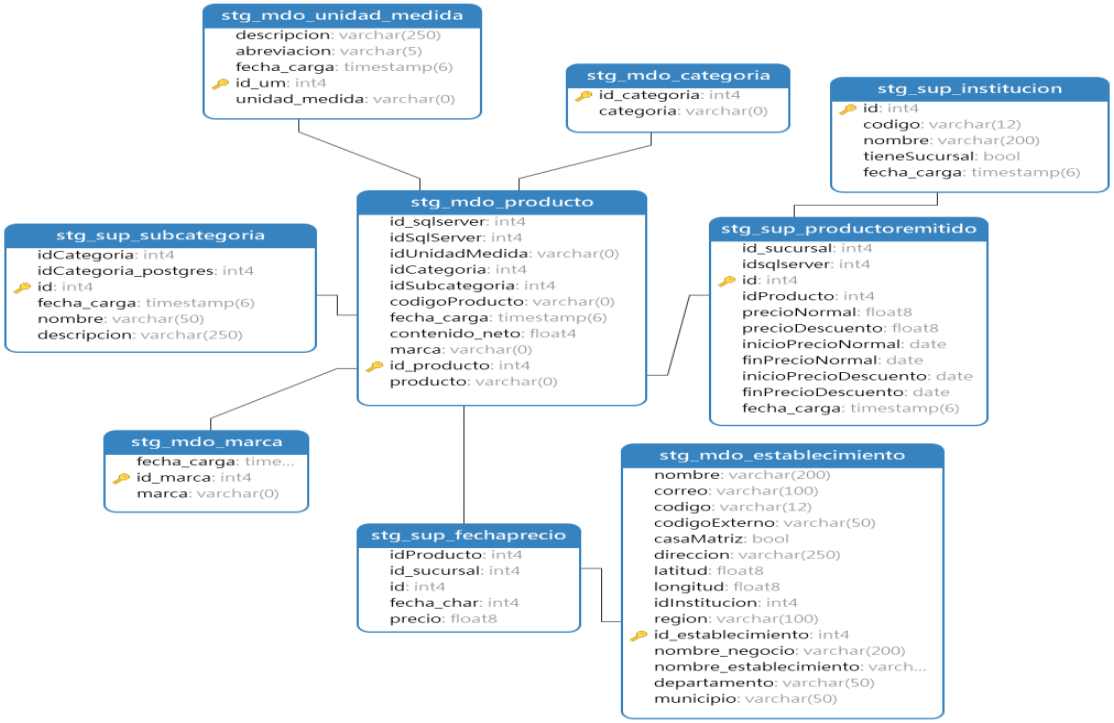
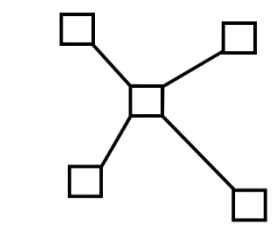


Figura 100 Diseño de modelo relacional (staging área).

15.3.1.3 Modelo multidimensional

15.3.1.3.1 Diseño Conceptual del Data mart (UML)

Especialización: Analizar los sondeos de precios de supermercados.



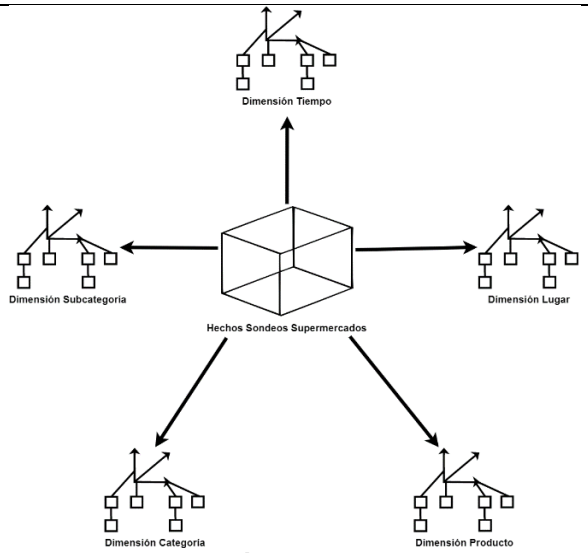
Nivel 1: Definición del modelo

Paquete Hechos:

Sondeo Supermercados.

Paquete Dimensiones:

1. Tiempo
2. Lugar
3. Producto
4. Categoría
5. Subcategoría



Nivel 2: Definición del esquema estrella.

Figura 101 Niveles 1 y 2 del diseño conceptual sondeos de Supermercados.

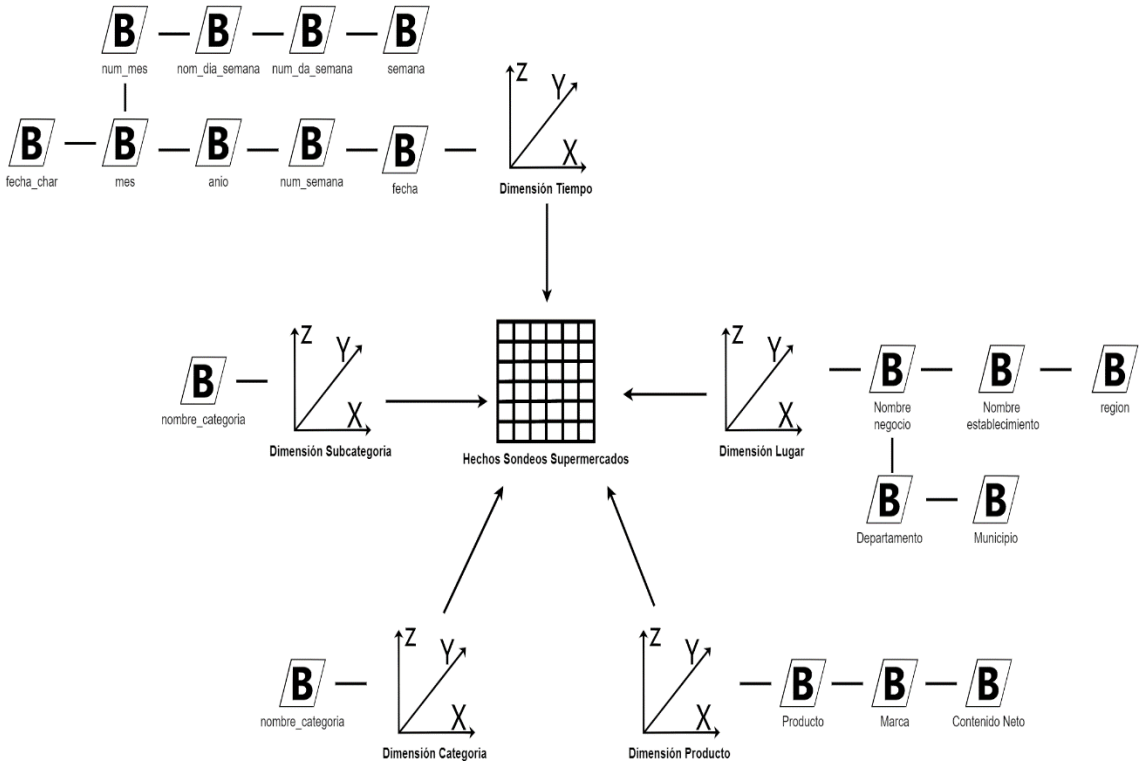


Figura 102 Nivel 3: Dimensiones/Hechos.

15.3.1.3.2 Diseño Físico Data Mart

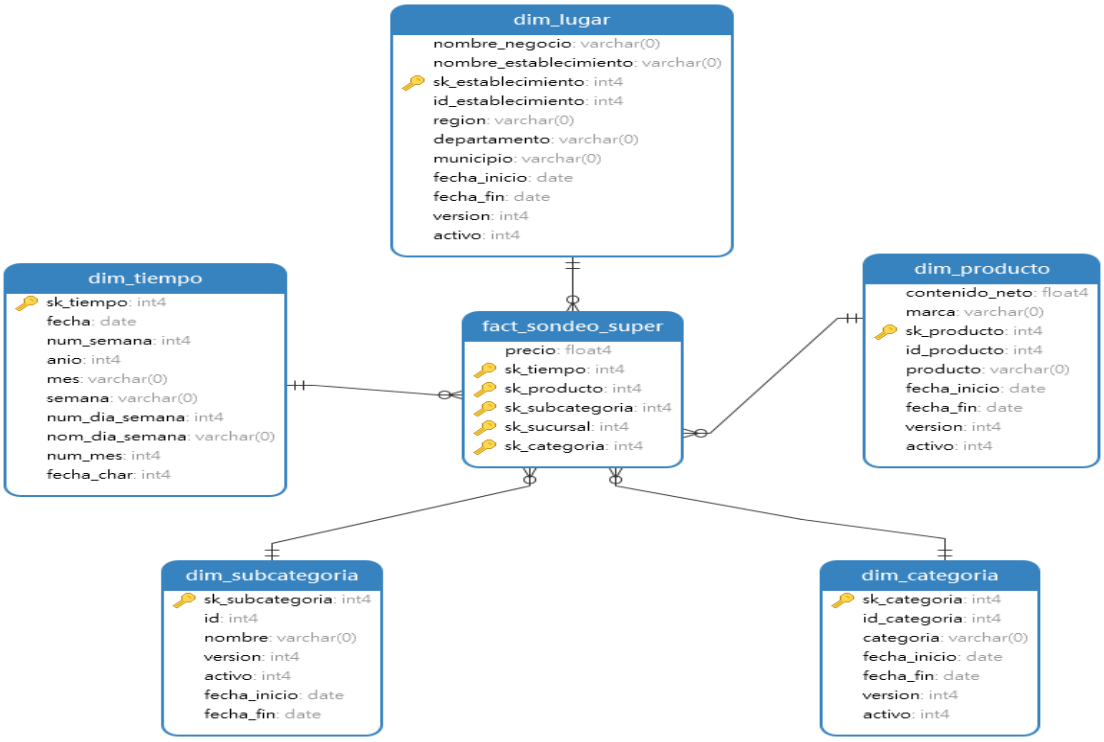


Figura 103 Modelo multidimensional Supermercados.

15.3.1.3.3 Diseño de procesos ETL

Se presenta a continuación el diseño del proceso ETL de una de las dimensiones.

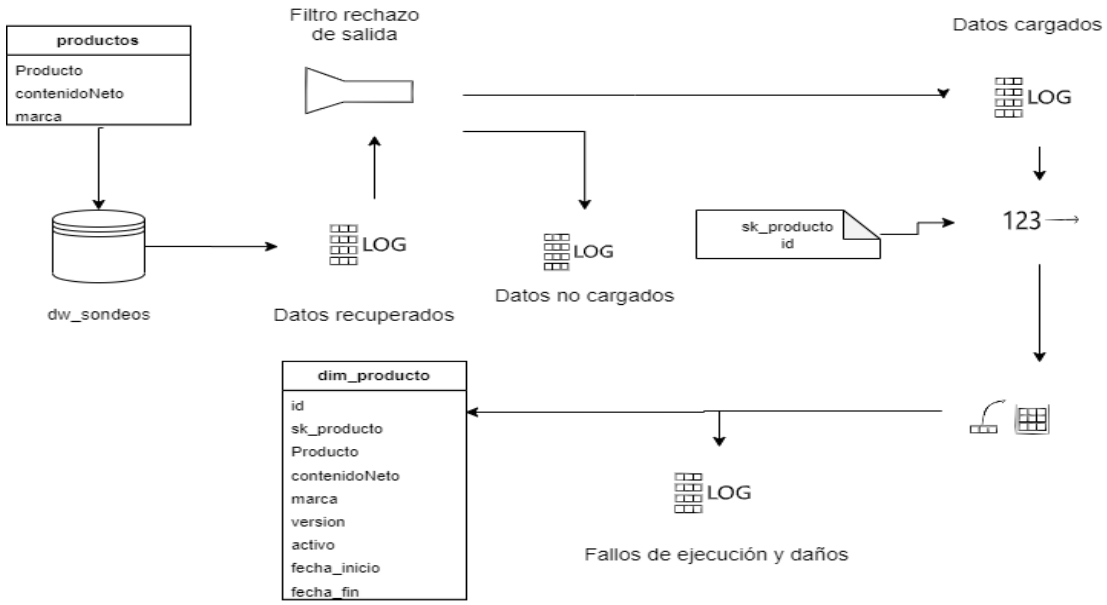


Figura 104 Dimensión producto.

La Figura 104 muestra uno de los diseños UML de los procesos ETL este corresponde la dimensión producto en el cual el proceso que se sigue se lista a continuación:

1. Conectar a la base de datos de origen en este caso el staging_area
2. Unir las tablas que conformarán la dimensión producto
3. Reportar al Log los datos recuperados.
4. Filtrar para datos cargados y no cargados
5. Reportar al Log datos cargados y datos no cargados.
6. Calcular clave sustituta.
7. Cargar datos a la tabla de destino dim_producto.

15.3.1.3.4 Desarrollo de procesos ETL

Se realizan procesos ETL para la parte del “Staging Area”, donde se obtienen los datos de la base de datos relacional y se preparan para cargarlos al modelo multidimensional.

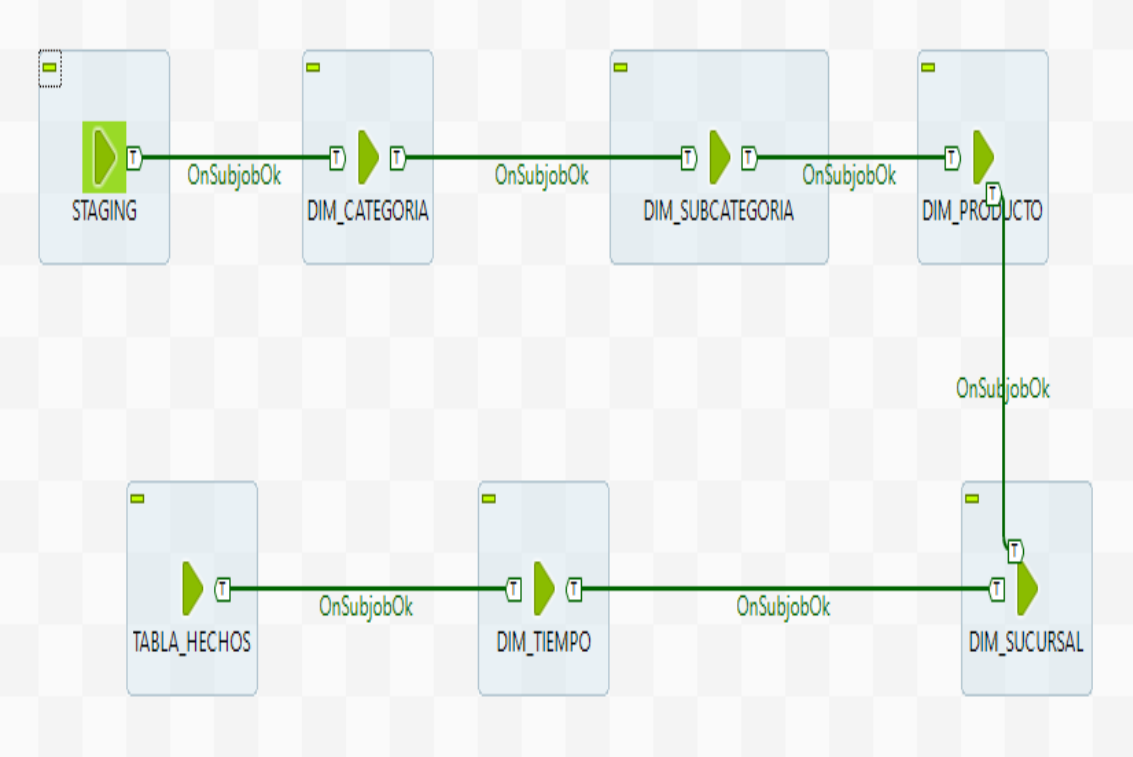


Figura 105 Job de orquestación de ejecución.

La Figura 105 muestra el macro Job, desde el cual se ordena la ejecución del conjunto de Jobs que tienen inmersas todas las transformaciones y cargas necesarias para los datos.

1. STAGING: Este Job ejecuta todas las transformaciones, extracciones, conversiones y el tratamiento necesario de los datos para que queden listos para ser cargados a las dimensiones y a la tabla de hechos.
2. DIM_CATEGORIA: Este Job se encarga de la creación de la dimensión categoría y realizar el ETL del Staging Área a la respectiva dimensión.
3. DIM_SUBCATEGORIA: Este Job se encarga de la creación de la dimensión subcategoría y realizar el ETL del Staging Área a la respectiva dimensión.

4. DIM_PRODUCTO: Este Job se encarga de la creación de la dimensión producto y realizar el ETL del Staging Área a la respectiva dimensión.
5. DIM_SUCURSAL: Este Job se encarga de la creación de la dimensión sucursal y realizar el ETL del Staging Área a la respectiva dimensión.
6. DIM_TIEMPO: Este Job se encarga de la creación de la dimensión tiempo y realizar el ETL del Staging Área a la respectiva dimensión.
7. TABLA_HECHOS: Este Job se encarga de la creación de la tabla de hechos, cargando los datos desde las respectivas dimensiones.

15.3.1.3.5 Pruebas

Para las pruebas técnicas en sondeos de precios de supermercados se ingresaron datos a la base de datos transaccional que se encuentra en el servidor de base de datos sql server, luego se corre el job “data_warehouse_sondeos” en Talend y nos ingresa los datos en las tablas Staging, luego se cargan a las dimensiones y por último se ingresa en la tabla de hechos “fact_sondeo_super”, para esta prueba se ingresaron un total de 1074478 registros (Figura 106) de los cuales se cargaron 70 a la tabla de hechos de supermercados (Ultima tabla de Figura 107).

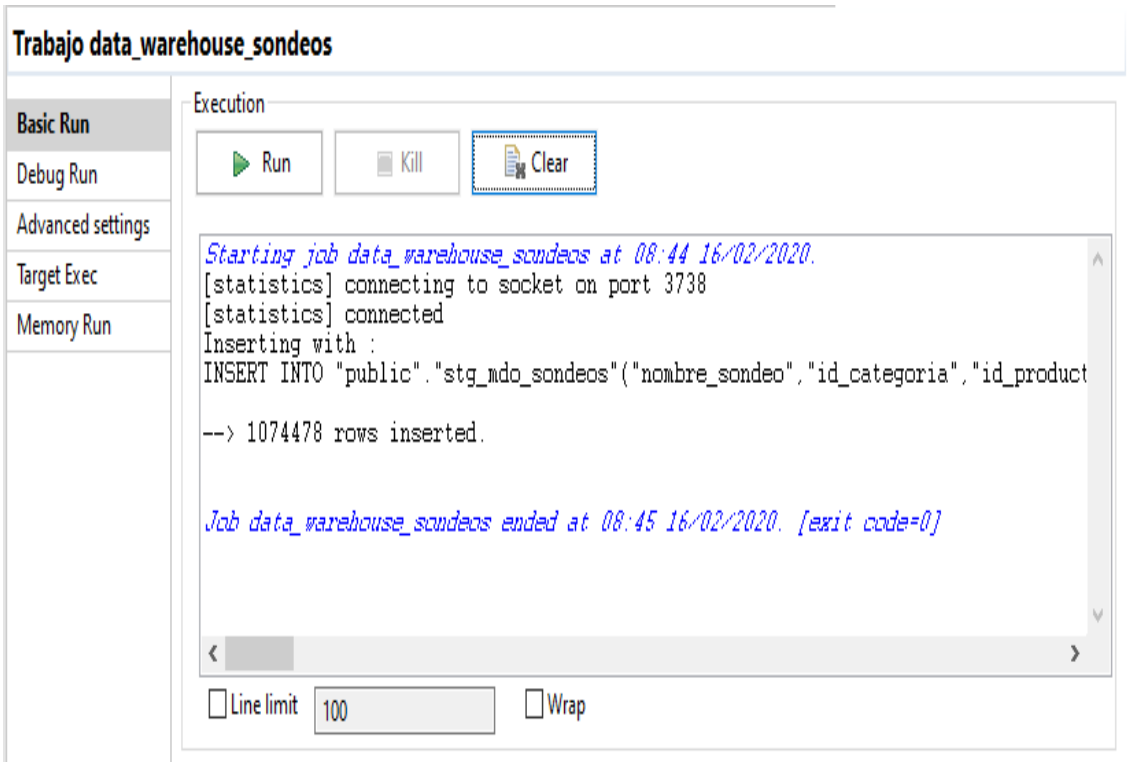


Figura 106 Ejecuciones de pruebas en Job ETL_supermercados.

```
2020-02-16 08:44:52
+-------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------+
|                                                                                    STAGING_SUPERMERCADOS                                                                                    |
|=------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------=|
|moment             |pid   |father_pid|root_pid|system_pid|project|job      |job_repository_id  |job_version|context|origin             |label            |count|reference|thresholds|
|=------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------=|
|2020-02-16 08:44:53|TrJFsQ|E4pZwu     |E4pZwu  |1936      |MIDAS  |staging_area|_EN724DVoEeqmtsGIB1Vlpw|0.1        |Default|tFlowMeter_71|Datos recuperados|70   |null     |          |
|2020-02-16 08:44:53|TrJFsQ|E4pZwu     |E4pZwu  |1936      |MIDAS  |staging_area|_EN724DVoEeqmtsGIB1Vlpw|0.1        |Default|tFlowMeter_66|Datos cargados   |42   |null     |          |
+-------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------+

Inserting with :
INSERT INTO "public"."fact_sondeo_mercado"("id_sondeo","nombre_sondeo","cantidad","precio","precio_referencia","tipo_descuento","sk_categoria","sk_producto","sk_um","sk_establecimiento","sk_tiempo")

--> 1074478 rows inserted.

2020-02-16 08:45:07
+-------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------+
|                                                                                    DW_SUPERMERCADOS                                                                                         |
|=------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------=|
|moment             |pid   |father_pid|root_pid|system_pid|project|job      |job_repository_id  |job_version|context|origin             |label            |count|reference|thresholds|
|=------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------=|
|2020-02-16 08:45:07|Pgaoor|E4pZwu     |E4pZwu  |1936      |MIDAS  |tabla_hechos|_jUVSMPmTEembt7pngYK81w|0.1        |Default|tFlowMeter_1|Datos recuperados|70   |null     |          |
|2020-02-16 08:45:07|Pgaoor|E4pZwu     |E4pZwu  |1936      |MIDAS  |tabla_hechos|_jUVSMPmTEembt7pngYK81w|0.1        |Default|tFlowMeter_4|Datos cargados   |70   |null     |          |
+-------------------+------+-----------+--------+----------+-------+---------+-------------------+-----------+-------+-------------------+-----------------+-----+---------+----------+
```

Figura 107 Log generado después de la ejecución.

15.3.2 Visualización

15.3.2.1 Informes de inteligencia de negocios

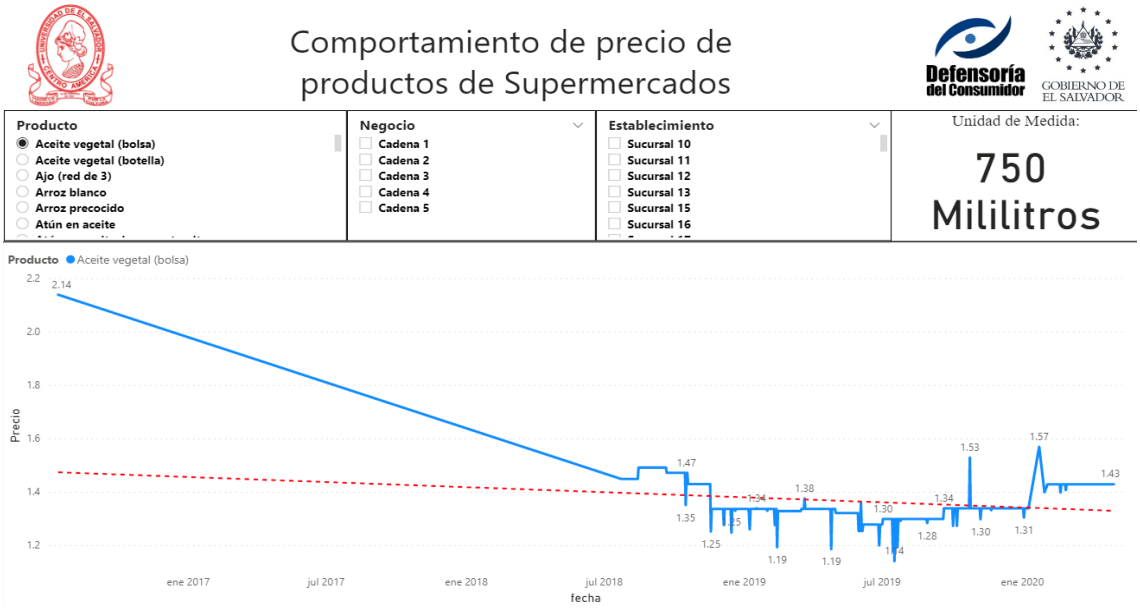


Figura 108 Reporte de precios de productos de supermercados.

La Figura 108 muestra un histórico de precios de los productos de supermercados, este reporte permite que se seleccionen diferentes parámetros como el producto, el negocio y el establecimiento. Además de mostrarnos la unidad de medida del producto seleccionado. Los filtros nos permiten seleccionar uno o más negocios y establecimientos para un producto.

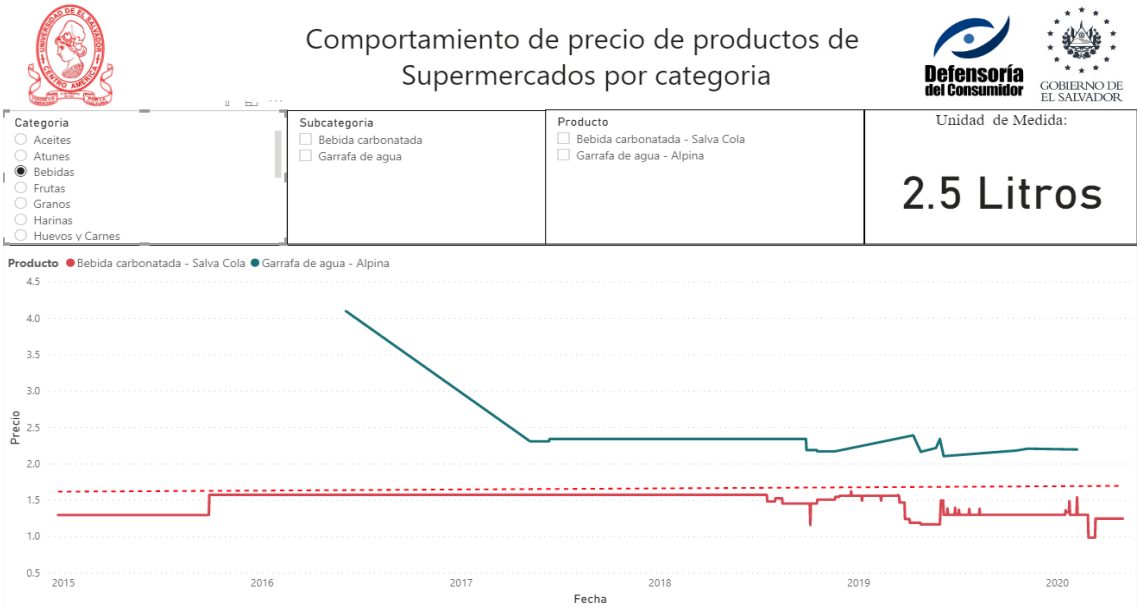


Figura 109 Reporte de precios de productos de supermercados por categoria.

La Figura 109 muestra un gráfico del precio de los productos de supermercados, el reporte nos permite seleccionar diferentes parámetros como la categoría, subcategoría y el producto. Además de mostrarnos la unidad de medida de los productos a la categoría seleccionada. Los filtros nos permiten seleccionar uno o más productos o subcategorías.

Figura 110 Mapa de precios de productos coloreado por municipio.

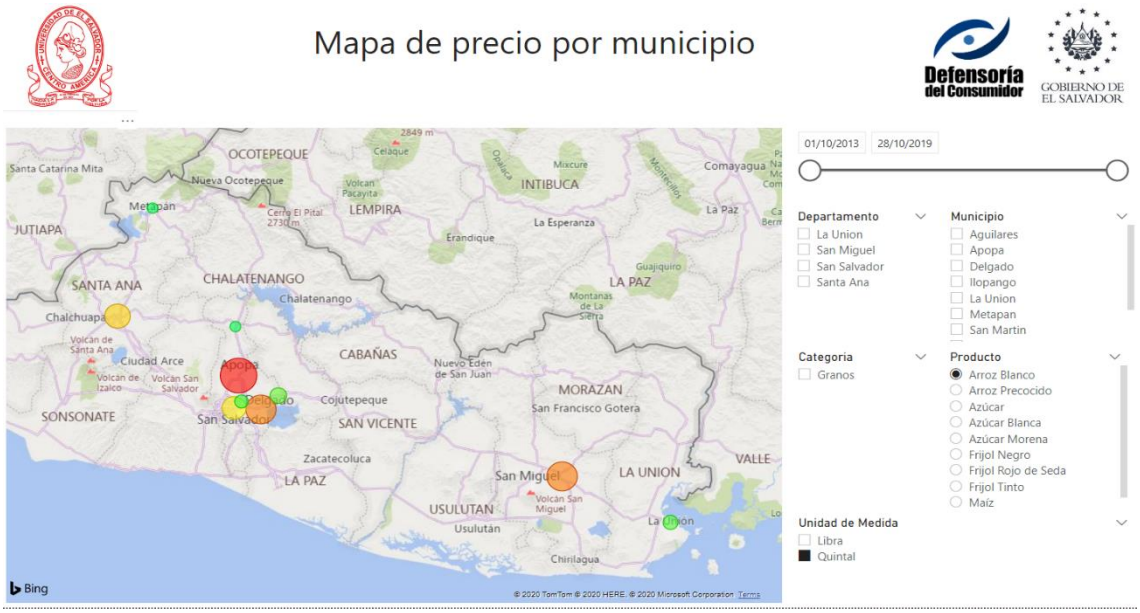


Figura 111 Mapa de precios de productos con burbujas según municipio.

La Figura 110 y Figura 111 muestra un mapa con los precios de los productos según el municipio seleccionado como parámetro. La representación del precio varía según el reporte, este puede ser mostrado coloreando el municipio al que pertenece o mostrando una burbuja sobre él. Los parámetros para este reporte son: departamento, municipio, producto, unidad de medida, categoría, y la fecha.

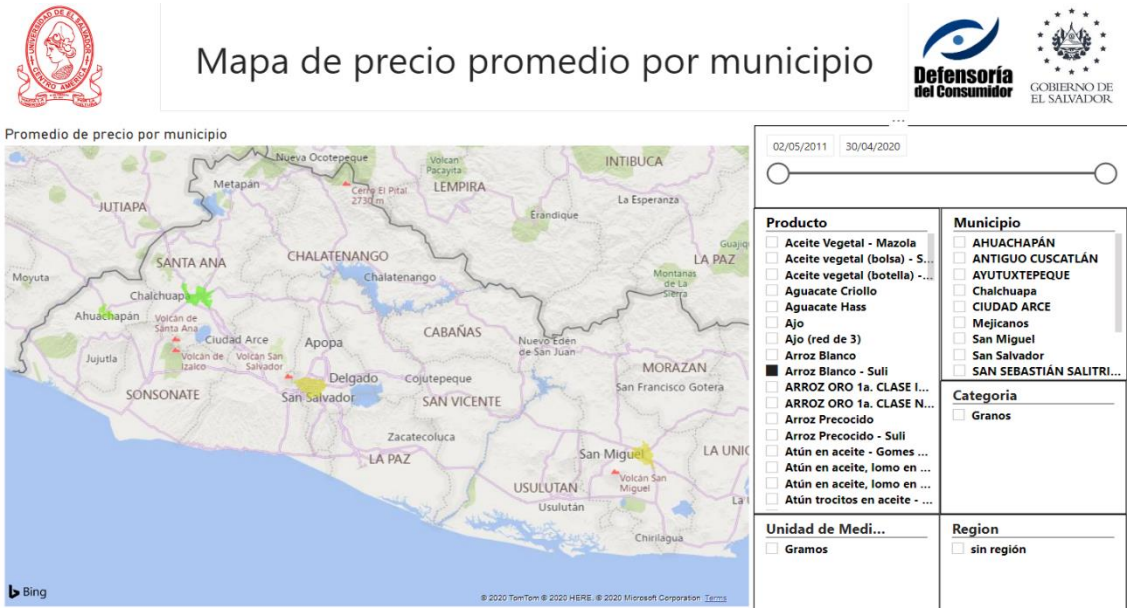


Figura 112 Mapa de precios de productos por región.

La Figura 112 muestra el reporte de precios de productos por región en un mapa, mostrando los precios de los productos en los municipios a la región a la que pertenece. En el reporte, el mapa muestra los precios según la región seleccionada por el filtro. Otros filtros a utilizar son: el producto, la categoría, la unidad de medida y la fecha.

15.3.2.1.1 Mapa de precios de productos por departamento, región y establecimiento

Figura 113 Mapa de precios de productos por departamento, región y establecimiento.

La Figura 113 muestra el reporte de precios de productos por departamento, región y establecimiento en un mapa, mostrando los precios de los productos por establecimiento. En el reporte, el mapa muestra los precios según la región, departamento, producto, la unidad de medida y la fecha.

15.3.3 Proceso de integración Data Warehouse

Posterior al desarrollo de las tres iteraciones se realizó un proceso de integración debido a que cada sprint se desarrolló bajo un data mart diferente, en dicho proceso de unificación se estandarizaron nombres de tablas del Staging área así como las tablas en las cuales se escribe en la base datos si son equivalentes, para que desde los ETL de los distintos data mart escriban a las mismas tablas así mismo se realizó para el modelo multidimensional, teniendo al inicio modelos multidimensionales o data mart bajo esquema estrella posteriormente pasaría a ser un solo Data Warehouse bajo el esquema constelación múltiples tablas de hechos conectadas o no con dimensiones en común con otras tablas de hechos.

Grupo sondeos	Tabla original	Campo	Tabla integrada	Campo	Modelo	BD	BD integrada
MAG	producto	id_producto	stg_mdo_producto	id_producto	S A	mag	dw_sondeos
MAG	producto	nombre_producto	stg_mdo_producto	producto	S A	mag	dw_sondeos
MAG	unidad_medida	id_unidad_medida	stg_mdo_unidad_medida	id_um	S A	mag	dw_sondeos
MAG	unidad_medida	unidad_medida	stg_mdo_unidad_medida	unidad_medida	S A	mag	dw_sondeos
MAG	plaza	id_plaza	stg_mdo_establecimiento	id_establecimiento	S A	mag	dw_sondeos
MAG	plaza	nombre_plaza	stg_mdo_establecimiento	nombre_establecimiento	S A	mag	dw_sondeos
MAG	sondeos_mag	fecha_carga	stg_mag_sondeos_mag	fecha_carga	S A	mag	dw_sondeos
MAG	sondeos_mag	id_sondeo	stg_mag_sondeos_mag	id_sondeo	S A	mag	dw_sondeos
MERCADOS	unidad_medida	id_um	stg_mdo_unidad_medida	id_um	S A	mercado	dw_sondeos
MERCADOS	unidad_medida	unidad_medida	stg_mdo_unidad_medida	unidad_medida	S A	mercado	dw_sondeos
MERCADOS	categoria	id_categoria	stg_mdo_categoria	id_categoria	S A	mercado	dw_sondeos
MERCADOS	categoria	categoria	stg_mdo_categoria	categoria	S A	mercado	dw_sondeos
MERCADOS	establecimiento	region	stg_mdo_establecimiento	region	S A	mercado	dw_sondeos
MERCADOS	establecimiento	id_establecimiento	stg_mdo_establecimiento	id_establecimiento	S A	mercado	dw_sondeos
SUPERMERCADOS	institucion	id	stg_sup_institucion	id	S A	supermercados	dw_sondeos
SUPERMERCADOS	institucion	codigo	stg_sup_institucion	codigo	S A	supermercados	dw_sondeos
SUPERMERCADOS	institucion	nombre	stg_sup_institucion	nombre	S A	supermercados	dw_sondeos
SUPERMERCADOS	institucion	tieneSucursal	stg_sup_institucion	tieneSucursal	S A	supermercados	dw_sondeos
SUPERMERCADOS	institucion	fecha_carga	stg_sup_institucion	fecha_carga	S A	supermercados	dw_sondeos
SUPERMERCADOS	sucursal	nombre_institucion	stg_mdo_establecimiento	nombre_establecimiento	S A	supermercados	dw_sondeos

Tabla 39 Ejemplo del conjunto de cambios en el modelo de staging área.

La Tabla 39 presenta una tabla de ejemplo con algunas de las modificaciones hechas a nivel de estructura de las bases de datos para la integración de los modelos en un mismo staging área. Se muestra el conjunto de cambios realizados en el modelo de staging área para sondeos en el cual:

- **Grupo sondeos:** representa el tópico abordado (Mag, Mercados o Supermercados) los cuales fueron desarrollados uno por cada iteración.
- **Tabla Original**: representa el nombre que originalmente se le proporcionó a la tabla en la respectiva iteración en la cual se trató ese tópico.
- **Campo:** Representa el nombre de los campos de la tabla.
- **Tabla integrada**: Posterior a la integración cual es el nombre que se le dio a la tabla.
- **Campo**: El nombre que se le dio al campo dentro de la tabla integrada.
- **Modelo**: Se refiere a modelo de staging area (S A) o Data Warehouse (DW).
- **BD**: representa el nombre de la base de datos en la cual estaba alojada la tabla en la iteración en la que se realizó.
- **BD Integrada**: Posterior a la integración cual es nombre de la base de datos donde está alojada la tabla.

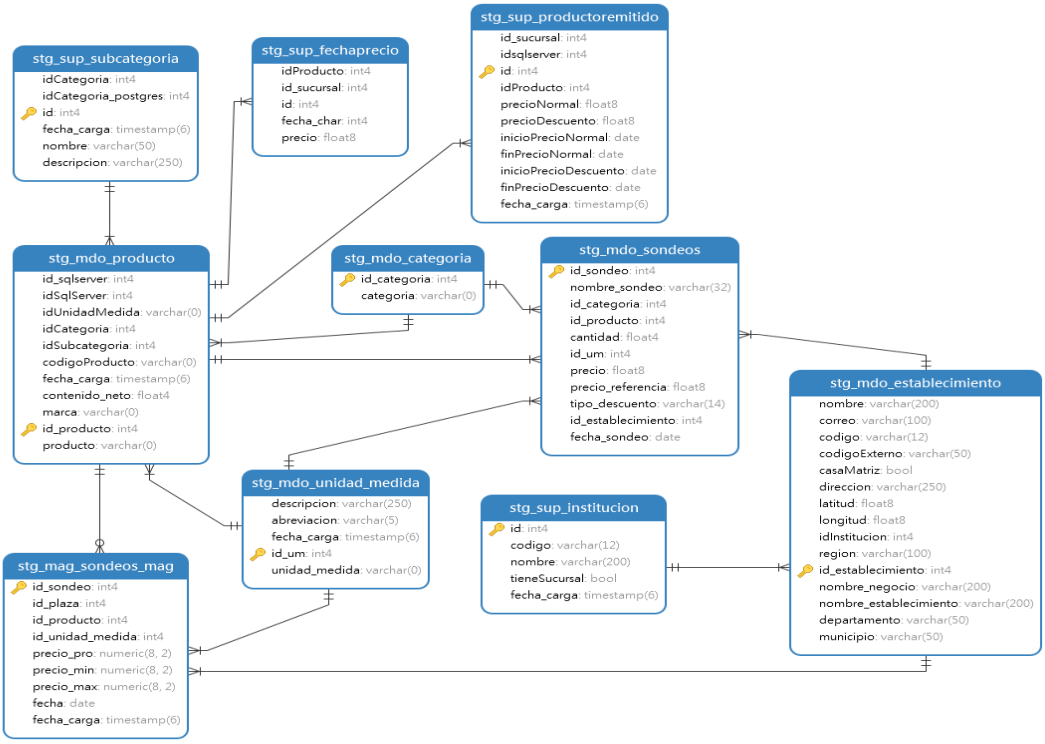


Figura 114 Modelo relacional para el staging área sondeos

La Figura 114 muestra el modelo relacional resultante posterior a la integración de los datos abordados en cada una de las iteraciones, el cual sirve para el staging área.

Grupo sondeos	Tabla original	Campo	Tabla integrada	Campo	Modelo	BD	BD integrada
MAG	dim_tiempo	sk_tiempo	dim_tiempo	sk_tiempo	DW	mag	dw_sondeos
MAG	dim_tiempo	fecha	dim_tiempo	fecha	DW	mag	dw_sondeos
MAG	plaza_scd	id_plaza	dim_lugar	id_establecimiento	DW	mag	dw_sondeos
MAG	plaza_scd	nombre_plaza	dim_lugar	nombre_establecimiento	DW	mag	dw_sondeos
MAG	unidad_medida_scd	unidad_medida	dim_unidad_medida	dim_unidad_medida	DW	mag	dw_sondeos
MAG	unidad_medida_scd	id_unidad_medida	dim_unidad_medida	id_um	DW	mag	dw_sondeos
MAG	producto_scd	fecha_fin	dim_producto	fecha_fin	DW	mag	dw_sondeos
MAG	producto_scd	activo	dim_producto	activo	DW	mag	dw_sondeos
MERCADOS	sondeo_mercado	nombre_sondeo	fact_sondeo_mercado	nombre_sondeo	DW	mercado	dw_sondeos
MERCADOS	sondeo_mercado	sk_categoria	fact_sondeo_mercado	sk_categoria	DW	mercado	dw_sondeos
MERCADOS	dim_marca	marca			DW	mercado	dw_sondeos
MERCADOS	dim_marca	sk_marca			DW	mercado	dw_sondeos
MERCADOS	dim_producto	sk_producto	dim_producto	sk_producto	DW	mercado	dw_sondeos
MERCADOS	dim_producto	id_producto	dim_producto	id_producto	DW	mercado	dw_sondeos
MERCADOS	dim_categoria	sk_categoria	dim_categoria	sk_categoria	DW	mercado	dw_sondeos
MERCADOS	dim_categoria	id_categoria	dim_categoria	id_categoria	DW	mercado	dw_sondeos
SUPERMERCADO	tabla_hechos	sk_tiempo	fact_sondeo_super	sk_tiempo	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_tiempo	sk_tiempo	dim_tiempo	sk_tiempo	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_tiempo	fecha	dim_tiempo	fecha	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_sucursal	sk_sucursal	dim_lugar	sk_establecimiento	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_sucursal	fecha_inicio	dim_lugar	fecha_inicio	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_subcategoria	sk_subcategoria	dim_subcategoria	sk_subcategoria	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_subcategoria	id	dim_subcategoria	id	DW	supermercados	dw_sondeos
SUPERMERCADO	dim_subcategoria	nombre	dim_subcategoria	nombre	DW	supermercados	dw_sondeos

Tabla 40 Ejemplo del conjunto de cambios en el modelo del Data Warehouse

En Tabla 40 se muestra el conjunto de cambios realizados en el modelo multidimensional para sondeos en el cual:

- **Grupo sondeos:** representa el tópico abordado (MAG, Mercados o Supermercados) los cuales fueron desarrollados uno por cada iteración.
- **Tabla Original**: representa el nombre que originalmente se le proporcionó a la tabla en la respectiva iteración en la cual se trató ese tópico.
- **Campo:** Representa el nombre de los campos de la tabla.
- **Tabla integrada**: Posterior a la integración cual es el nombre que se le dio a la tabla.
- **Campo**: El nombre que se le dio al campo dentro de la tabla integrada.
- **Modelo**: Se refiere a modelo de staging area (S A) o Data Warehouse (DW).
- **BD**: representa el nombre de la base de datos en la cual estaba alojada la tabla en la iteración en la que se realizó.
- **BD Integrada**: Posterior a la integración cual es nombre de la base de datos donde está alojada la tabla.

Para el modelo multidimensional al unir los 3 Data Mart se obtiene un modelo de tipo constelación en la cual existen 3 tablas de hecho las cuales están conectadas a dimensiones en algunos casos en común con otras tablas de hechos.

En el caso de los ETL tanto para el staging area como para el modelo multidimensional se realizaron cambios en los destinos de los datos tal como se especificó anteriormente en las tablas de conjunto de los cambios.

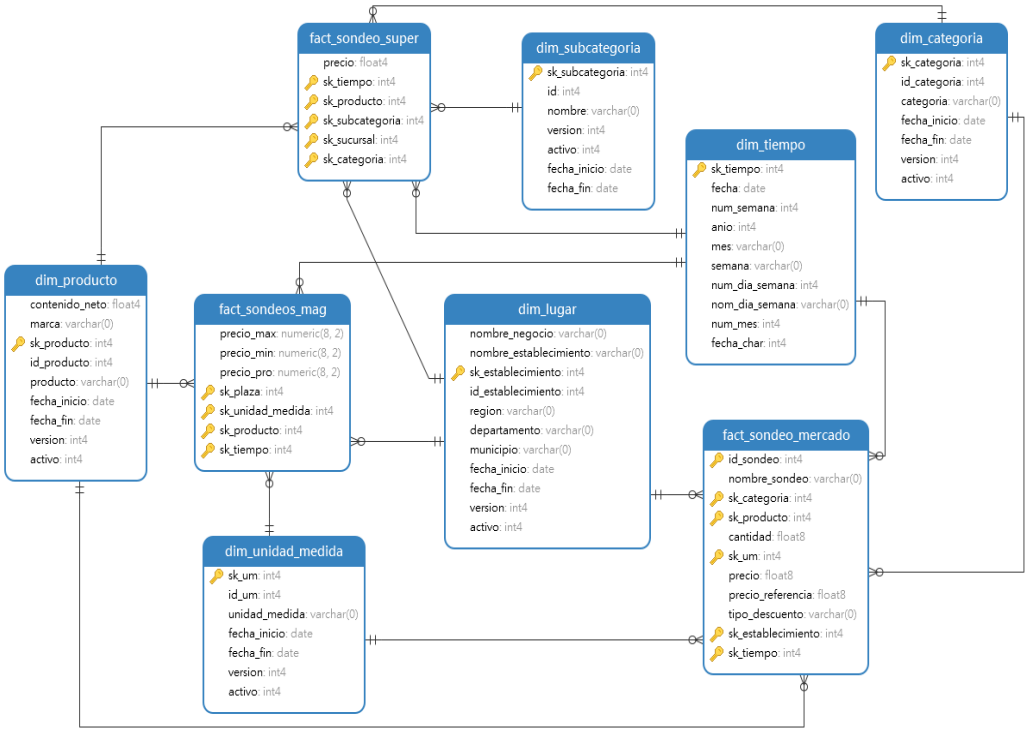


Figura 115 Modelo para el: Data Warehouse de sondeos

La Figura 115 muestra el modelo multidimensional, ya integrado comprendiendo las tablas de hechos para cada tópico, así como las respectivas dimensiones a las que acceden.

Respecto a los job's en los que se llevan a cabo los ETL se unificó en un único JOB la creación de las tablas de la base de datos, así como las secuencias necesarias para las dimensiones SCD como se observa en la Figura 116, únicamente quedaron en cada uno de sus job's correspondientes la creación de las tablas que alojan las dimensiones SCD.

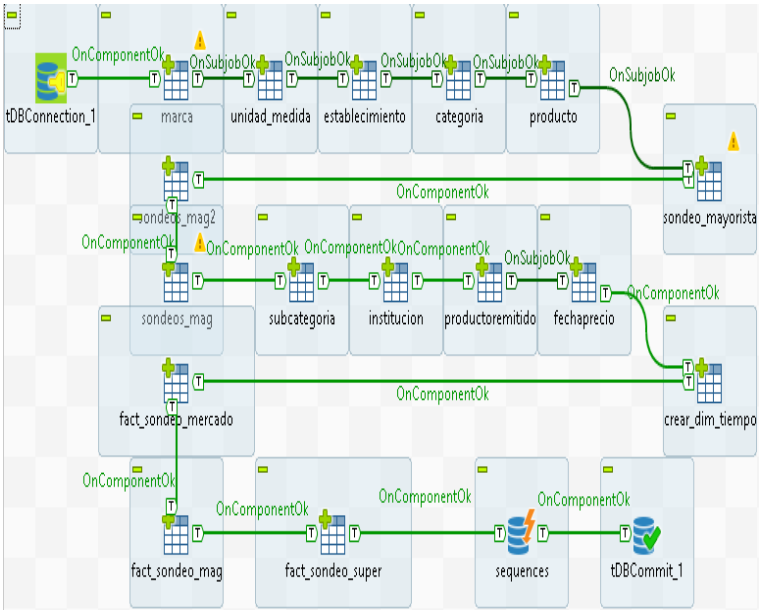


Figura 116 Creación de las estructuras de las tablas para la base de datos dw_sondeos.

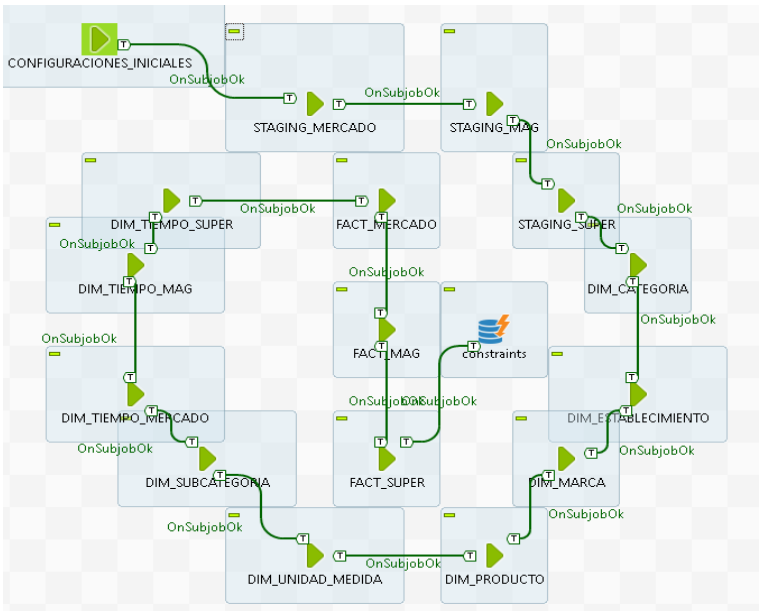


Figura 117 Job de orquestación de los ETL del data warehouse de sondeos

La Figura 117 muestra el job de orquestación de cada uno de los job que se utilizan para el Data Warehouse de sondeos, ya integrados de modo que el destino de los datos sea conforme como se detalló en la Tabla 39 para staging área y Data Warehouse respectivamente Tabla 40.

El orden de ejecución de los JOB's se enumera a continuación:

1. CONFIGURACIONES_INICIALES: Ejecuta el job de creación de las tablas como se expresó en la Figura 116.
2. STAGING_MERCADO: Ejecuta el job para el staging área del origen de datos para mercados.
3. STAGING_MAG: Ejecuta el job para el staging área del origen de datos para el Ministerio de Agricultura y Ganadería (MAG).
4. STAGING_SUPER: Ejecuta el job para el staging área del origen de datos para supermercados.
5. DIM_CATEGORIA: Ejecuta el job para la dimensión categoría en ella se cargan datos del staging área a la dimensión.
6. DIM_ESTABLECIMIENTO: Ejecuta el job para la dimensión establecimiento en ella se cargan datos del staging área a la dimensión.
7. DIM_MARCA: Ejecuta el job para la dimensión marca, en ella se cargan datos del staging área a la dimensión.
8. DIM_PRODUCTO: Ejecuta el job para la dimensión producto, en ella se cargan datos del staging área a la dimensión.
9. DIM_UNIDAD_MEDIDA: Ejecuta el job para la dimensión unidad de medida, en ella se cargan datos del staging área a la dimensión.
10. DIM_SUBCATEGORIA: Ejecuta el job para la dimensión subcategoria, en ella se cargan datos del staging área a la dimensión.
11. DIM_TIEMPO_MERCADO: Ejecuta el job para la dimensión tiempo, en ella se cargan datos del staging área de mercados a la dimensión tiempo.
12. DIM_TIEMPO_MAG: Ejecuta el job para la dimensión tiempo, en ella se cargan datos del staging área del MAG a la dimensión tiempo.
13. DIM_TIEMPO_SUPER: Ejecuta el job para la dimensión tiempo, en ella se cargan datos del staging área de supermercados a la dimensión tiempo.
14. FACT_MERCADO: Ejecuta el job para la tabla de hechos de mercado, en ella se cargan datos del staging área de mercados a la tabla de hechos de mercados.
15. FACT_MAG: Ejecuta el job para la tabla de hechos de mag, en ella se cargan datos del staging área del MAG a la tabla de hechos del MAG.
16. FACT_SUPERMERCADO: Ejecuta el job para la tabla de hechos de super mercados, en ella se cargan datos del staging área de supermercados a la tabla de hechos de supermercados.

15.3.4 Proceso de integración Workflow Knime

Inicialmente se construyeron flujos de trabajo (workflow's) en Knime Analytics Platform que implementan los modelos de minería de datos necesarios para satisfacer las hipótesis que se plantearon, posteriormente se realizó un proceso de unificación en el cual se afinaron los siguientes aspectos:

1- Se unifican todos los flujos de trabajo en uno solo mediante la utilización de metanodos.

2- Las salidas de datos inicialmente eran hacía hojas de cálculo (.xlsx) se modificó para que se escriban en tablas de una base de datos, para ser insumo de la visualización de datos en Power BI.
3- Inicialmente se realizaba una conexión a la base de datos del modelo multidimensional para la selección de los datos para los algoritmos de minería de datos por cada workflow, posteriormente se modificó para ser configurada en un único nodo el cual distribuye la configuración en el resto que la necesiten.
4- Se agrega un nodo al workflow que permita reiniciar los nodos previo a cada ejecución, esto pensando en la implementación mediante tareas programadas.

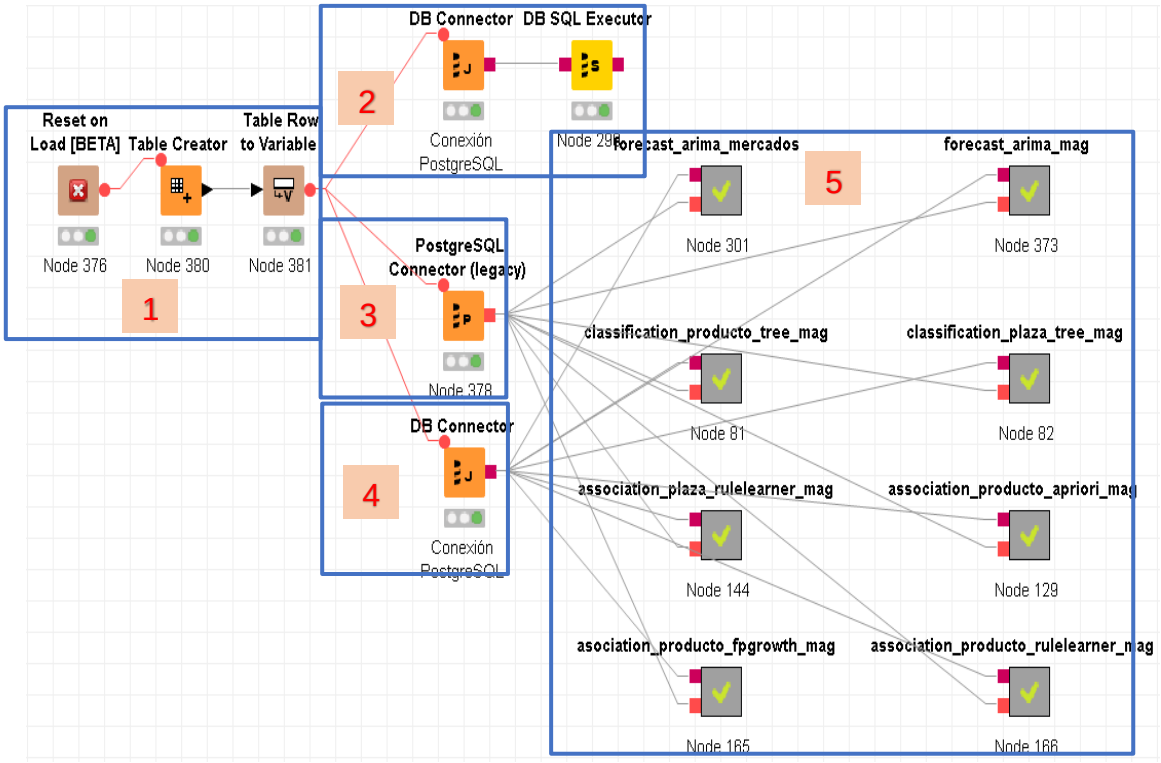


Figura 118 Flujo de trabajo integrado de minería de datos sondeos Knime

La Figura 118 muestra el flujo de trabajo integrado en el cual se lleva a cabo la ejecución de la minería de datos de sondeos mediante la herramienta Knime en general se realizan las siguientes actividades:

1. Se reinician los nodos en el flujo de trabajo, se crea una tabla con los parámetros de la conexión al Data Warehouse, así como a la base de datos en el que se escribirán las tablas con las salidas de la minería de datos, posteriormente se generan variables con cada uno de estos parámetros.
2. Se crea una sesión a la base de datos “mining_MIDAS” y se eliminan las tablas de las salidas si es que existen.
3. Se establece una conexión a la base de datos “mining_MIDAS”, está conexión se pasa como entrada a todos los metanodos para escribir en las tablas, las salidas de minería de datos.

4. Se crea una sesión de base de datos a “dw_sondeos” está sesión se pasa como entrada a cada uno de los metanodos y se utiliza para seleccionar los datos para los algoritmos de minería de datos en cada workflow.
5. Cada uno de los workflow que se desarrollaron se convierte en un metanodo para un solo flujo de trabajo quedando como se describe a continuación
   - **5.1. forecast_arima_mercados:** Técnica de Predicción mediante el algoritmo de series temporales ARIMA para sondeos de mercados.
   - **5.2. forecast_ arima_mag:** Técnica de Predicción mediante el algoritmo de series temporales ARIMA para sondeos del MAG.
   - **5.3. classification_producto_tree_mag:** Técnica de clasificación, algoritmo de árboles de decisión para productos en los sondeos del MAG.
   - **5.4. classification_plaza_tree_mag:** Técnica de clasificación, algoritmo de árboles de decisión para plazas en los sondeos del MAG.
   - **5.5. association_plaza_rulelearner_mag:** Técnica de asociación, algoritmo Rule Learner para plazas de sondeos del MAG.
   - **5.6. association_producto_apriori_mag:** Técnica de asociación, algoritmo Apriori para productos de sondeos del MAG.
   - **5.7. association_producto_fpgrowth_mag:** Técnica de asociación, algoritmo FPGrowth para productos de sondeos del MAG.
   - **5.8. association_producto_rulelearner_mag:** Técnica de asociación, algoritmo Rule Learner para productos de sondeos del MAG.

16 Sprint 4

16.1 Descripción Historias de Usuario

Código	RA101
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea identificar los orígenes de datos del modelo.
Razón	Para poder construir un modelo multidimensional a partir de ellos.
Criterios de aceptación	Se considerarán orígenes de datos que satisfagan las necesidades actuales de la UACM.
	Si son orígenes que corresponden a sistemas que no se encuentra en uso actualmente se incluirán igualmente como histórico.
	Los orígenes de datos serán correspondientes a las atenciones brindadas por la DC mediante SARA.
Validación	Se comprobará que los orígenes de datos satisfagan el requerimiento actual de datos que se hace por parte de la UACM.
	Se comprobará que no se excluya ningún origen que pueda afectar en las cantidades de información reportadas actualmente.
	Se comprobarán que los orígenes identificados correspondan a los orígenes tomados por los objetos de base de datos que producen el insumo actual de la UACM.
Valor del negocio	100
Puntos de historia	1
ROI	100

Tabla 41 Historia de Usuario RA101

Código	RA102
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea identificar tablas dentro de los orígenes de datos
Razón	Para incluir únicamente tablas donde se encuentren los campos que se necesitan.
Criterios de aceptación	Las tablas son definidas por los objetos de los orígenes de datos del requerimiento que la UACM realiza actualmente.
	Únicamente se incluirán tablas contenidas en las bases de datos identificadas como orígenes.
	Las tablas corresponden a estructuras que almacenan las atenciones que realiza la DC, así como sus catálogos asociados.
Validación	Se comprobará que las tablas provengan únicamente de los orígenes identificados.
	Se comprobará que no se dejen fuera tablas que contengan campos que son utilizados actualmente en los informes de la UACM.
	Se comprobará que todas las tablas que se encuentran en operaciones de la base de datos que originan al requerimiento actual sean incluidas.
Valor del negocio	200
Puntos de historia	7

ROI	29

Tabla 42 Historia de Usuario RA102

Código	**RA103**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea identificar los campos a migrar al staging area
Razón	Para poder construir un proceso ETL
Criterios de aceptación	Se considerarán solo los campos que satisfagan las necesidades actuales de la UACM.
	Los orígenes de datos serán correspondientes a las atenciones brindadas por la DC mediante SARA.
Validación	Se comprobará que los campos satisfagan el requerimiento actual de datos que se hace por parte de la UACM.
	Se comprobará que no se excluya ningún campo que pueda afectar la información.
	Se comprobarán que los campos identificados correspondan a los orígenes tomados por los objetos de base de datos que producen el insumo actual de la UACM.
Valor del negocio	300
Puntos de historia	4
ROI	75

Tabla 43 Historia de Usuario RA103

Código	**RA104**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea migrar tablas al staging área.
Razón	Para poder construir un modelo multidimensional a partir de ellos.
Criterios de aceptación	El staging área será implementado en una base de datos.
	Los orígenes de datos serán consolidados temporalmente en el staging área.
	El staging área mantendrá registros históricos de las cargas.
	El proceso deberá guardar los tiempos de ejecución.
	Se notificará por correo al surgir un error en la ejecución del staging area.
Validación	Se comprobará que el staging área se carguen todos los datos provenientes de los orígenes de datos.
	Se comprobará que no se excluya ningún origen identificado previamente.
	Se comprobará que el staging área cuente con datos que puedan solucionar problemas técnicos en la operación de ETL.
Valor del negocio	900
Puntos de historia	13
ROI	68

Tabla 44 Historia de Usuario RA104

Código	**RA105**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea diseñar los paquetes de información.

Razón	Para poder construir un modelo multidimensional a partir de del hecho que se desea medir.
Criterios de aceptación	El paquete de información deberá de contener un tema.
	El paquete de información contará con un objetivo claro que es lo que se desea medir para la construcción del modelo multidimensional.
	El paquete de información deberá mostrar las diferentes dimensiones o jerarquías.
Validación	Se comprobará que cuenten con las dimensiones para la correcta construcción del modelo multidimensional.
	Se comprobará que no se excluya ningún campo de la dimensión que pueda perjudicar la construcción del modelo multidimensional.
	Se comprobará que las dimensiones tengan concordancia con el objetivo que se desea medir para la construcción del modelo multidimensional.
Valor del negocio	500
Puntos de historia	3
ROI	167

Tabla 45 Historia de Usuario RA105

Código	**RA107**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea diseñar diagramas UML para los procesos de ETL
Razón	Para poder desarrollar los procesos de ETL
Criterios de aceptación	Se considerará los procesos correspondientes a la carga staging area.
	Se considerará los procesos para la carga del modelo multidimensional
	Se considera los procesos correspondientes a archivos externos Excel entre otros.
Validación	Se comprobará que diagramas cumplan la nomenclatura UML extendida propuesta por Sergio Luján-Mora y Juan Trujillo
	Se comprobará que los diagramas tengan los elementos necesarios para una correcta carga del staging area
	Se comprobará que los diagramas incluyan los elementos necesarios para una correcta transformación de datos
Valor del negocio	300
Puntos de historia	1
ROI	100

Tabla 46 Historia de Usuario RA107

Código	**RA108**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea crear procesos de extracción, transformación y carga para los datos del sistema de atenciones a los consumidores hacia un modelo multidimensional.
Razón	Para poder tener los datos correctos a la hora de realizar informes.
	Obtener todos los datos de las atenciones a los consumidores.

Criterios de aceptación	Realizar todas las transformaciones que sean necesarias en concordancia con la UACM.
	Realizar la limpieza a los datos que realiza la UACM en la actualidad.
Validación	Se comprobará que los datos cargados sean congruentes en relación a una base de muestra previamente trabajada por UACM.
	Se comprobará que las transformaciones se hayan realizado correctamente.
	Se comprobará la limpieza realizada a la información para que sea congruente con la que realiza la UACM.
Valor del negocio	1200
Puntos de historia	15
ROI	80

Tabla 47 Historia de Usuario RA108

Código	**RA109**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea realizar pruebas dentro del modelo multidimensional.
Razón	Para corroborar el buen funcionamiento del mismo.
Criterios de aceptación	Que los ETL's no contenga errores en la ejecución.
	Que se realicen las transformaciones necesarias a los datos.
	Que se realice una buena limpieza de los datos.
Validación	Se comprobará que el ETL no tenga errores de ejecución u otro tipo de error.
	Se comprobará que se realicen todas las transformaciones necesarias.
	Se comprobará que se realice una buena limpieza en los datos.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 48 Historia de Usuario RA109

16.2 Refinamiento del requerimiento de información

16.2.1 Procesos BPMN

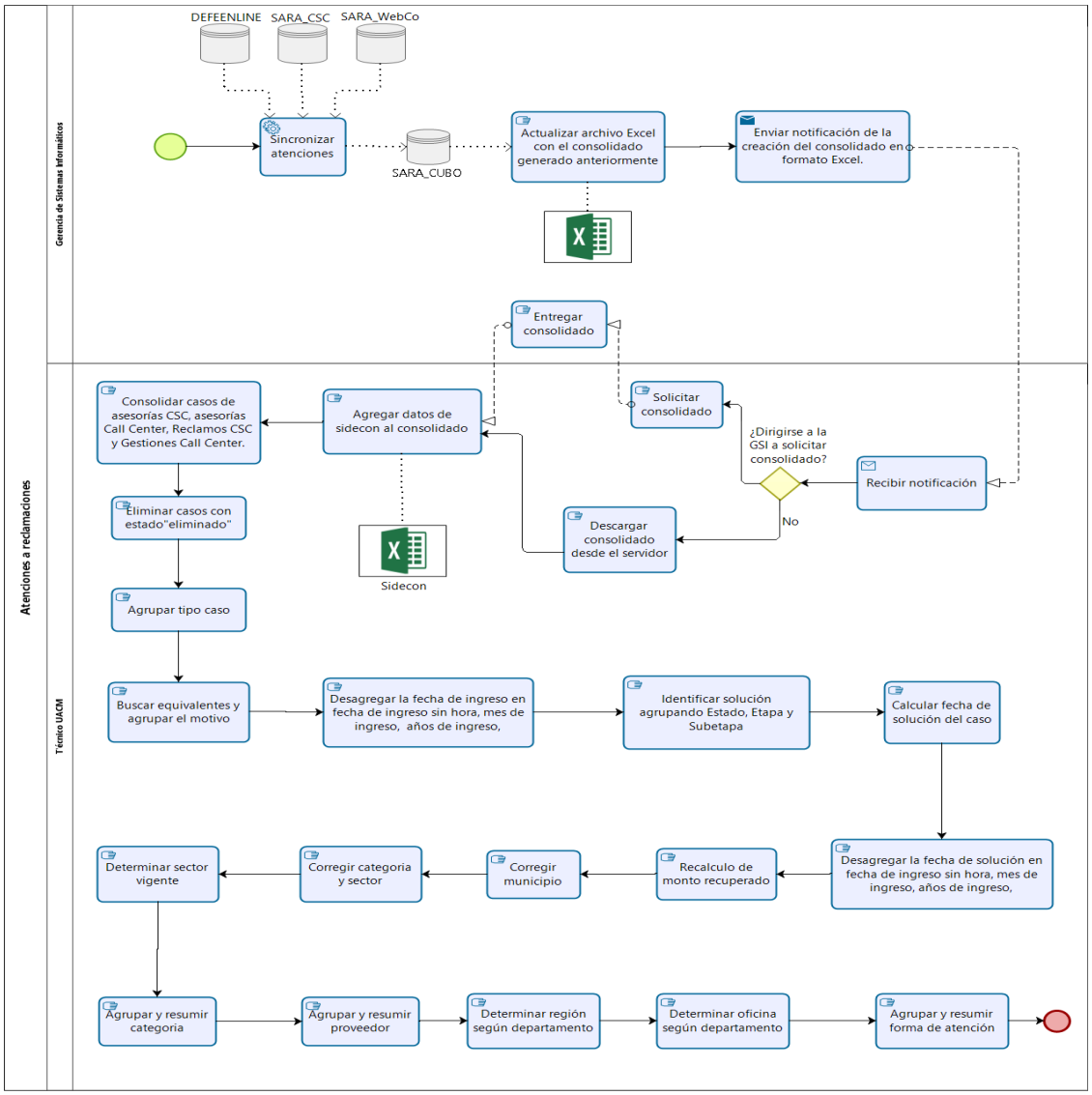


Figura 119 Diagrama BPMN Situación Actual SARA

La Figura 119 muestra el diagrama BPMN de la situación actual para el consolidado a las atenciones a reclamaciones. El primer paso que se realiza para la creación del consolidado es sincronizar las atenciones, esto se hace obteniendo los datos de tres bases de datos distintas, DEFEENLINE, SARA_CSC, SARA_WebCo. Los nuevos datos se cargan a la base de datos SARA_CUBO, luego, la gerencia de sistemas actualiza el archivo Excel y notifica a la GSI que el consolidado se encuentra listo.

El técnico de la UACM dispone de dos vías para obtener el consolidado, una es dirigiéndose a la GSI y obtener el consolidado personalmente, y la segunda, descargándolo desde un servidor. Luego de tener el consolidado se añaden los datos históricos con los que se cuentan, en este caso SIDECON dado que fue el primer sistema con el que se contó

en la Defensoría del Consumidor. Una vez realizado el paso anterior, se consolidan los casos de asesorías y reclamos del CSC (Centro de Solución de Controversia) y las asesorías y gestiones del Call Center. También, se eliminan los casos con estado eliminado, ya que estos no deben de ser tomados en cuenta para las estadísticas.

El siguiente paso es agrupar el tipo de caso, los tipos de caso originales se deben cambiar por su respectiva equivalencia y asignarle el nombre de “Tipo de caso agrupado”. Para el motivo, en algunas ocasiones se encuentran vacíos o con valores nuevos que no existían cuando se estableció el catálogo original, por lo tanto, se debe de buscar y asignar la equivalencia a los casos con campos vacíos y realizar una actualización para las equivalencias de los nuevos valores. Se debe de desagregar la fecha de ingreso en fecha de ingreso sin hora, mes de ingreso y año de ingreso, con la finalidad de que se puedan aplicar filtros de manera más fácil y rápida.

Teniendo los datos de estado, etapa y la subetapa del caso se determina la solución del caso, luego, se calcula la fecha de solución y se desagrega en fecha de ingreso sin hora, mes de ingreso y año de ingreso. Además, se hace un recalculo del monto recuperado, se corrige y estandariza el nombre del municipio, categoría y sector. Dado que la DC ha utilizado distintos listados de sectores y es necesario adecuarlos para que estos sean presentados, para ello la clasificación debe de ser uniforme y se deben determinar equivalencias en la nueva clasificación de sectores utilizando los campos de Categoría y Sector. Además, se deben de agrupar y resumir las categorías y proveedores. Asimismo, se determina la región y oficina según el departamento. Y, por último, se agrupa y resume la forma de recepción.

16.2.2 Paquetes de Información

Tema:	Cantidad de atenciones brindadas en la Defensoría del Consumidor										
JERARQUIAS	Tiempo	Consumidor	Motivo	Proveedor	Lugar	Oficina	Sector Categoría	Tipo caso	Atención	Forma de recepción	Técnico
	Fecha	Nombre del Consumidor	Nombre del motivo	Nombre del proveedor	Nombre del municipio	Nombre de la oficina	Nombre del sector	Nombre del tipo caso	Numero de caso	Nombre de la forma de recepción	Nombre del técnico
	Fecha concatenada	Nombre del genero	Nombre del motivo resumido	Nombre del proveedor resumido	Nombre del departamento	Nombre de la oficina original	Nombre de categoría	Nombre del tipo caso agrupado	Nombre de estado	Nombre de la forma de recepción agrupada	
	Año	Fecha de nacimiento del consumidor	Nombre del motivo financiero		Nombre de la región		Nombre de sector combinado	Sistema	Nombre de la etapa		
	Mes								Nombre de la subetapa		
	Semana								Nombre de la solución		
	Dia								Numero único		
									Nombre de la institución derivación		
									Nombre de la institución externa		
									Nombre de la ventanilla		
									Nombre del lugar móvil		
									Monto reclamado		
									Monto recuperado		
									Monto recuperado corregido		
Hechos Medidos:	Comportamiento de las atenciones brindadas										

Tabla 49 Paquete de información para cálculos de atenciones.

Tema:	Montos recuperados por atenciones					
JERARQUIAS	Tiempo	Atención	Motivo	Proveedor	Lugar	Consumidor
	Fecha	Numero de caso	Nombre del motivo	Nombre del proveedor	Nombre del municipio	Nombre del Consumidor
	Fecha concatenada	Nombre de estado	Nombre del motivo resumido	Nombre del proveedor resumido	Nombre del departamento	Nombre del genero
	Año	Nombre de la etapa	Nombre del motivo financiero		Nombre de la región	Fecha de nacimiento del consumidor
	Mes	Nombre de la subetapa				
	Semana	Nombre de la solución				
	Dia	Numero único				
		Nombre de la institución derivación				
		Nombre de la institución externa				
		Nombre de la ventanilla				
		Nombre del lugar móvil				
		Nombre del técnico receptor				
Hechos Medidos:	Monto recuperados y reclamados por atenciones					

Tabla 50 Paquete de información para montos recuperados y reclamos.

En base a los diferentes orígenes de datos que han sido proporcionados por la Defensoría del Consumidor se han determinado dos diferentes paquetes de información los cuales son Cantidad de atenciones brindadas en la Defensoría del Consumidor y Montos recuperados por atenciones en base al análisis de estos se ha determinado que para el primer paquete se han obtenido once jerarquías las cuales son: Tiempo, Consumidor, Motivo, Proveedor, Lugar, Oficina, Sector, Categoría, Tipo caso, Atención, Forma de recepción y Técnico y para el segundo paquete: Tiempo, Atención, Motivo, Proveedor, Lugar y Consumidor, asumiendo que la jerarquía de tiempo solo estará involucrada por la fecha a la cual será tratada para obtener las divisiones pertinentes.

16.2.3 Casos de uso

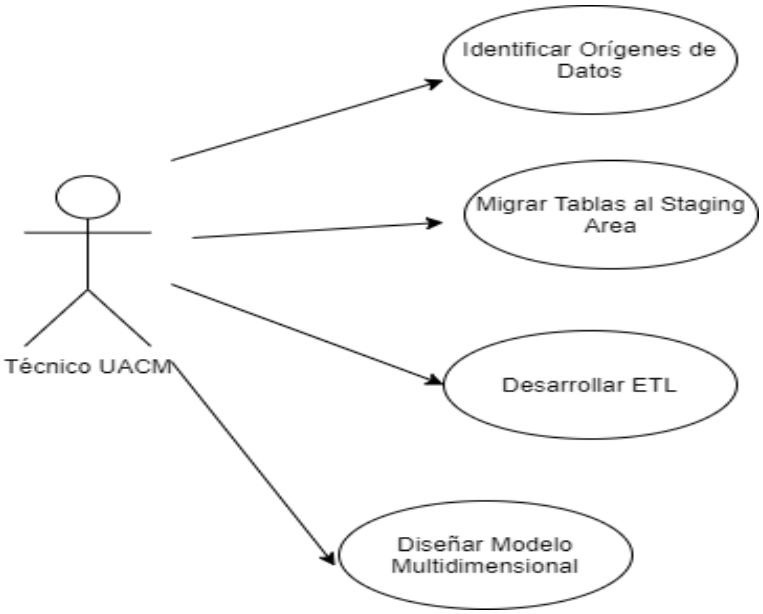


Figura 120 Diagrama de Casos de Uso

16.3 Actividades de desarrollo de la iteración

16.3.1 Integración de datos

16.3.1.1 Extracción de los datos

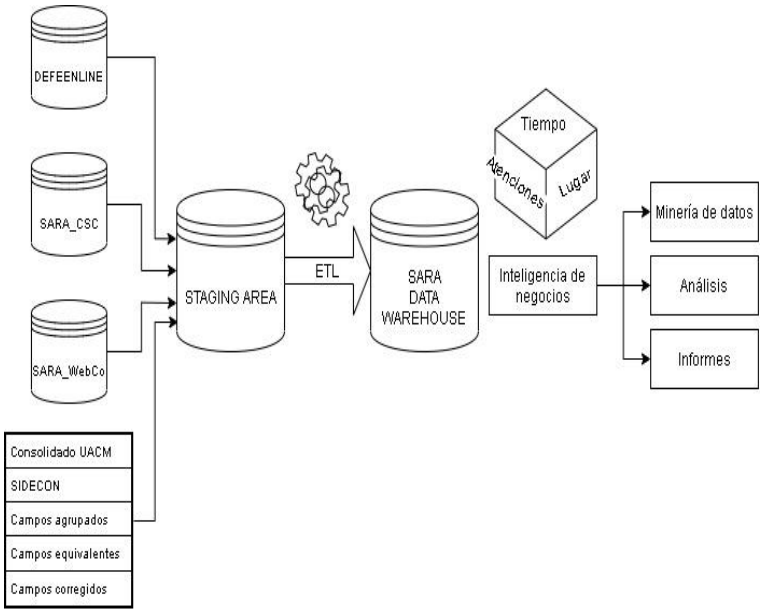


Figura 121 Flujo de extracción de los datos para la obtención de información.

La Figura 121 muestra el flujo de extracción de los datos en el cual se identifican los orígenes a ser utilizados en el presente trabajo:

1. DEFEENLINE: Es la base de datos en la que se almacenan atenciones en línea.
2. SARA_WebCo: Es la base de datos que tiene el histórico de atenciones en línea, es la base de datos que fue utilizada por el sistema de denuncias en línea, antecesor de DEFEENLINE.
3. SARA_CSC: Consiste en la base de datos central (Core institucional) de la de Defensoría del Consumidor en la cual se almacenan los diversos tipos de atenciones que la institución brinda (Avisos de infracción, Asesorías, Reclamos y anteriormente Gestiones).
4. Consolidado UACM: Actualmente la Unidad de Análisis de Consumo y Mercados utiliza como insumo Una hoja de cálculos separada por pestañas por los diversos tipos de atención al que le realizan procesos de revisión, limpieza y transformación de datos hasta lograr tener un consolidado que utilizan para la elaboración de informes, en dicho consolidado anexan las atenciones brindadas a nivel nacional mediante el “Sistema de Denuncias de Consumo” (SIDECON) ya que este histórico no proviene de los orígenes de base de datos antes especificados, además como para de la transformación de datos, para ciertas columnas se calcula su equivalente, agrupado o corregido.

Se enumera a continuación el proceso seguido para la extracción de los datos:

1. Por ser el tema de Atenciones brindadas por la Defensoría del Consumidor lo bastante complejo a nivel de tamaño y estructura se procedió a identificar en base al insumo que utiliza la UACM en primer momento los orígenes de datos, los cuales se detallaron y enumeraron anteriormente.
2. Una vez identificados los orígenes se procedió a identificar las tablas de esos orígenes que se utilizan en el insumo actual de la UACM.
3. Como las tablas contienen un número considerable de campos se discriminó por los campos utilizados únicamente en el insumo de la UACM; teniendo mapeadas las bases de datos, las tablas y los campos a utilizar se procede con el diseño y desarrollo de los ETL para migrar a una base de datos intermedia como se explica en la siguiente sección.

16.3.1.2 Staging área

16.3.1.2.1 Diseño de la base de datos

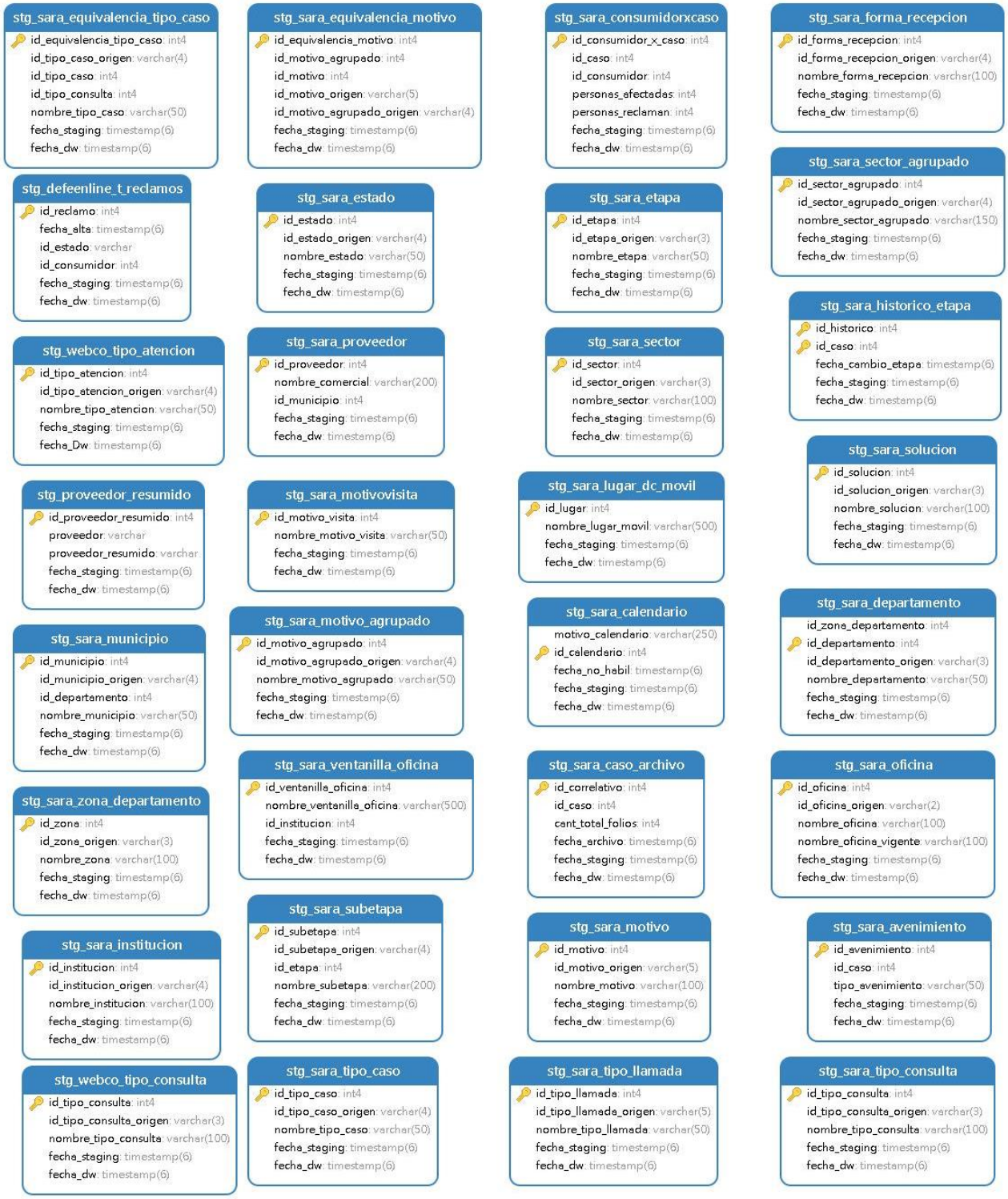


Figura 122 Diseño de modelo relacional (staging area).

16.3.1.2.2 Diseño de procesos ETL

16.3.1.2.2.1 Tabla staging área stg_sara_consultas

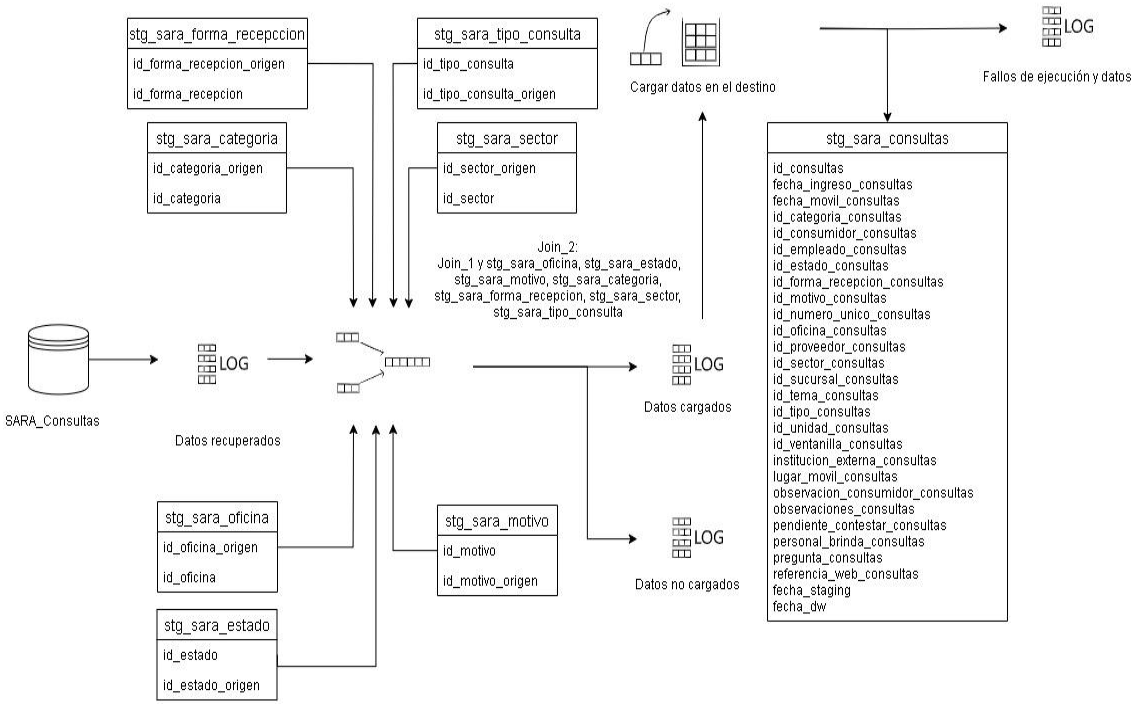


Figura 123 Diseño UML Job stg_sara_consultas.

La Figura 123 muestra el diseño UML para el Job stg_sara_consultas en el cual el proceso que se sigue se lista a continuación:

1. Conectar a la base de datos de origen y extraer los registros de la tabla SARA_Consultas.
2. Reportar al Log los datos recuperados.
3. Obtener el equivalente en el staging área de los identificadores de sus tablas relacionadas.
4. Reportar al Log datos cargados y datos no cargados.
5. Cargar datos a la tabla de destino stg_sara_consultas.
6. Reportar al Log fallos de ejecución y/o datos.

16.3.1.2.3 Desarrollo de procesos ETL

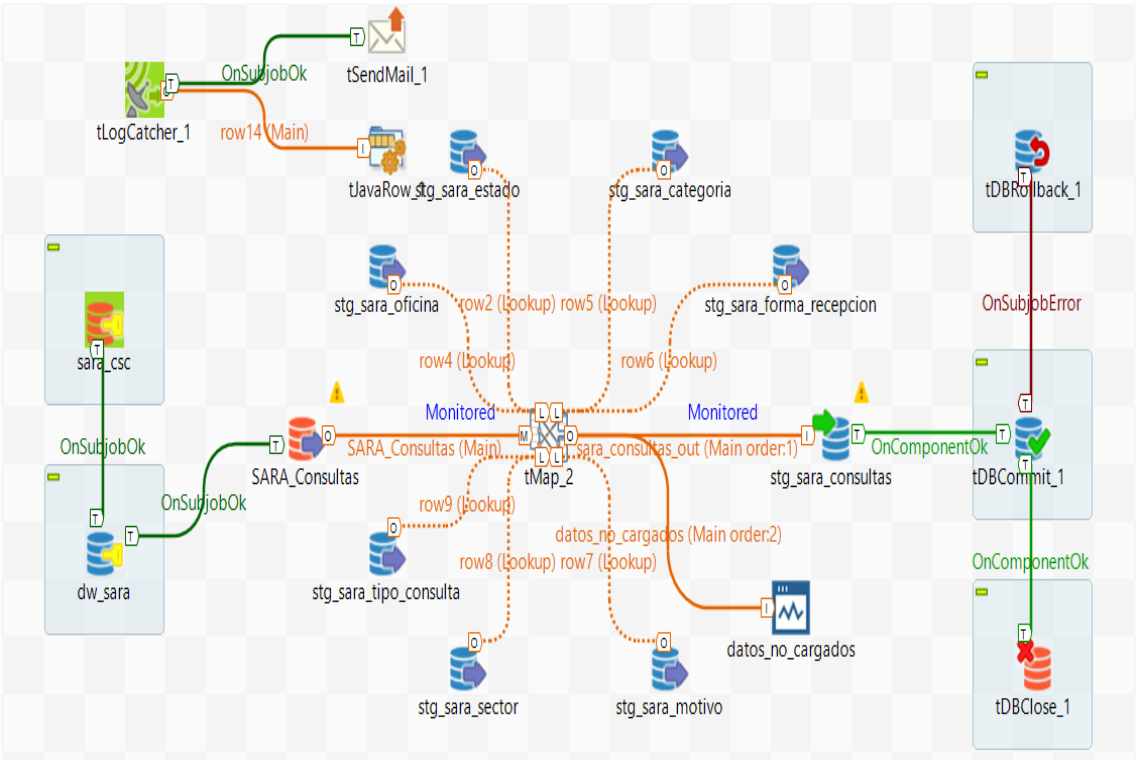


Figura 124 Job de stg_sara_consultas.

La Figura 124 muestra el Job stg_sara_consultas en el cual el proceso que se sigue se lista a continuación:

1. Se crea la conexión a la base de datos transaccional.
2. Se crea la conexión a la base de datos del staging área.
3. Se obtienen los registros de la tabla “SARA_Consultas”.
4. Se realizan las uniones a las tablas: stg_sara_oficina, stg_sara_estado, stg_sara_categoria, stg_sara_forma_recepción, stg_sara_tipo_consultas, stg_sara_sector y stg_sara_motivo para poder obtener los identificadores de cada tabla del staging.
5. Si algún dato no se carga correctamente se registrán en el nodo de datos_no_cargados.
6. Se cargan los datos a la tabla stg_sara_consultas y se cierran las conexiones.
7. Si existe un error en la ejecución, se captura y se envía por correo electrónico y se realiza un rollback.

16.3.1.2.4 Pruebas

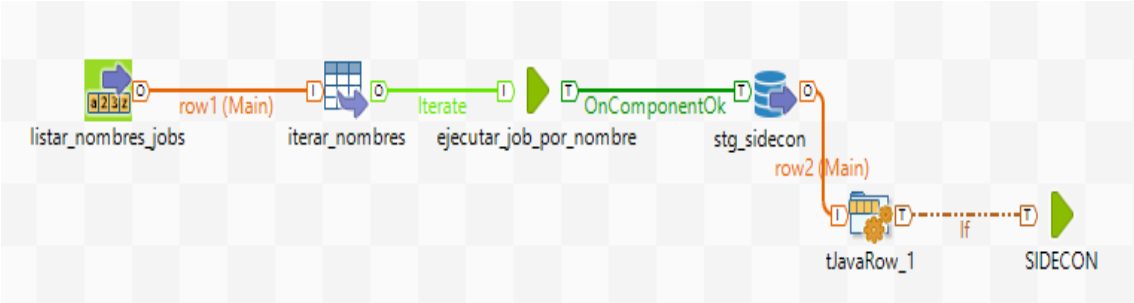


Figura 125 Job Ejecución staging area.

En la Figura 125 se observa el job que realiza una llamada para la ejecución de todos los job creados para el staging Area

- configuraciones_dw_sara_mherrera
- stg_sara_estado
- stg_sara_etapa
- stg_sara_subetapa
- stg_sara_calendario
- stg_sara_sector
- stg_sara_categoria
- stg_sara_zona_departamento
- stg_sara_departamento
- stg_sara_municipio
- stg_sara_forma_recepcion
- stg_sara_institucion
- stg_sara_lugar_dc_movil
- stg_sara_motivo
- stg_sara_motivos_por_sector
- stg_sara_motivovisita
- stg_sara_sector_agrupado
- stg_sara_motivo_agrupado
- stg_sara_solucion
- stg_sara_solucion_equivalencia
- stg_sara_tipo_caso
- stg_sara_tipo_consulta
- stg_sara_tipo_llamada
- stg_sara_ventanilla_oficina
- stg_webco_tipo_consulta
- stg_sara_equivalencia_motivo
- stg_sara_equivalencia_sector
- stg_sara_equivalencia_tipo_caso
- stg_sara_oficina
- stg_sara_empleado
- stg_sara_proveedor
- stg_sara_consumidor
- stg_sara_visita
- stg_sara_consultas
- stg_sara_caso
- stg_sara_caso_inf
- stg_sara_pd_formulario_generico
- stg_sara_historico_etapas
- stg_sara_avenimiento
- stg_sara_caso_archivo
- stg_sara_consumidorxcaso
- stg_sara_denuncia_no_personal
- stg_defeenline_asesoria
- stg_defeenline_t_reclamo
- stg_sara_remision
- stg_webco_consultas
- stg_sara_proveedorcc
- stg_proveedor_resumido
- stg_webco_tipo_atencion
- stg_sara_llamada

Se presentan los resultados para dos ejecuciones del Job, en el cual se constatan la cantidad de registros recuperados del origen, cargados y no cargados en el destino.

JOB	DATOS RECUPERADOS	DATOS CARGADOS	DATOS NO CARGADOS	RESULTADO	DURACION (seg)
stg_defeenline_asesoria	14420	14420	0	success	6.06
stg_defeenline_t_reclamo	12845	12845	0	success	1.07
stg_proveedor_resumido	22459	22459	0	success	2.87
stg_sara_avenimiento	384732	384732	0	success	5.98
stg_sara_calendario	1665	1665	0	success	0.67
stg_sara_caso	195406	195406	0	success	20.54
stg_sara_caso_archivo	164808	164808	0	success	3.89
stg_sara_caso_inf	23799	23799	0	success	1.90
stg_sara_categoria	312	312	0	success	0.73
stg_sara_consultas	703069	703069	0	success	42.35
stg_sara_consumidor	781581	781581	0	success	232.36
stg_sara_consumidorxcaso	4059	4059	0	success	0.66
stg_sara_denuncia_no_personal	14694	14694	0	success	2.05
stg_sara_departamento	16	15	0	success	0.56
stg_sara_empleado	1023	1023	0	success	0.80
stg_sara_equivalencia_motivo	303	303	0	success	0.72
stg_sara_equivalencia_sector	490	490	0	success	0.71
stg_sara_equivalencia_tipo_caso	5	5	0	success	0.80
stg_sara_estado	24	24	0	success	0.88
stg_sara_etapa	14	14	0	success	0.71
stg_sara_forma_recepcion	41	41	0	success	0.67
stg_sara_historico_etapas	1285604	1285604	0	success	24.38
stg_sara_institucion	45	44	1	success	0.59
stg_sara_llamada	639083	517587	121496	success	11.81
stg_sara_lugar_dc_movil	1820	1820	0	success	0.69
stg_sara_motivo	303	303	0	success	0.69
stg_sara_motivo_agrupado	15	15	0	success	0.62
stg_sara_motivos_por_sector	1106	1106	0	success	0.68
stg_sara_motivovisita	8	8	0	success	0.61
stg_sara_municipio	266	266	0	success	0.70
stg_sara_oficina	10	10	0	success	0.60

stg_sara_pd_formulario_generico	7499	7499	0	success	2.01
stg_sara_proveedor	28948	28948	0	success	4.16
stg_sara_proveedorcc	3193	3193	0	success	0.71
stg_sara_remision	8903	8903	0	success	0.92
stg_sara_sector	55	55	0	success	0.56
stg_sara_sector_agrupado	37	37	0	success	0.66
stg_sara_solucion	9	9	0	success	0.59
stg_sara_solucion_equivalencia	76	76	0	success	0.62
stg_sara_subetapa	96	96	0	success	0.65
stg_sara_tipo_caso	5	5	0	success	0.63
stg_sara_tipo_consulta	4	4	0	success	0.63
stg_sara_tipo_llamada	16	16	0	success	0.60
stg_sara_ventanilla_oficina	95	95	0	success	0.63
stg_sara_visita	285984	285984	0	success	10.64
stg_sara_zona_departamento	3	3	0	success	0.63
stg_webco_consultas	8371	8371	0	success	1.09
stg_webco_tipo_consulta	4	4	0	success	0.76

Tabla 51 Cantidad de registros recuperados del origen, cargados y no cargados en el destino.

Resultado obtenido en la base de datos:
Al ejecutar el job de staging area de ejecución se crearon las siguientes tablas con los datos de las bases de SARA

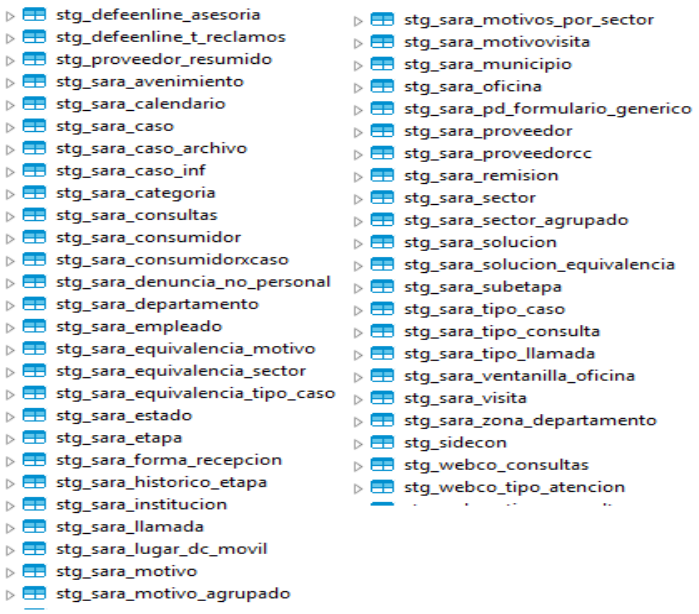


Figura 126 Resultado de Ejecución staging area.

Contenido tabla stg_sara_categoria

	id_sector	id_categoria	id_categoria_origen	nombre_categoria	fecha_staging	fecha_dw
1	1	1	C000	N/A	2020-05-25 18:23:16	[NULL]
2	2	2	C001	Suministro de agua	2020-05-25 18:23:16	[NULL]
3	29	3	C179	Red de distribución	2020-05-25 18:23:16	[NULL]
4	29	4	C180	Suministro de energía eléctrica	2020-05-25 18:23:16	[NULL]
5	30	5	C181	Aceites, grasas y margarinas	2020-05-25 18:23:16	[NULL]
6	30	6	C182	Agua embotellada	2020-05-25 18:23:16	[NULL]
7	30	7	C183	Aves	2020-05-25 18:23:16	[NULL]
8	30	8	C184	Azúcar	2020-05-25 18:23:16	[NULL]
9	30	9	C185	Bebidas carbonatadas	2020-05-25 18:23:16	[NULL]
10	30	10	C186	Bebidas energizantes	2020-05-25 18:23:16	[NULL]

Figura 127 Tabla stg_sara_categoria.

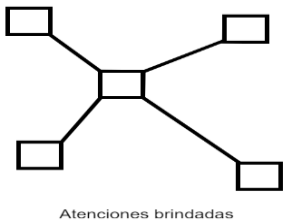
Contenido tabla sgt_sara_oficina

	id_oficina	id_oficina_origen	nombre_oficina	nombre_oficina_vigente
1	1	O1	Oficina Central	Oficina San Salvador
2	2	O2	Oficina Regional San Miguel	Oficina Regional San Miguel
3	3	O3	Oficina Regional Santa Ana	Oficina Regional Santa Ana
4	4	O4	Gerencia del Centro de Solución de Controversias de Sei	Gerencia del Centro de Solución de Controversias de Sei
5	5	O5	Oficina Call Center Plan de La Laguna	Oficina Call Center Plan de La Laguna
6	6	O6	Oficina Plan de La Laguna	Oficina Plan de La Laguna
7	7	O7	Oficina Educacion	Oficina Educacion
8	8	O8	Oficina Movil	Oficina Movil
9	9	O9	Oficina Ciudadania y Consumo	Oficina Ciudadania y Consumo
10	10	OX	Oficina Descentralizada	Oficina Descentralizada

Figura 128 Tabla stg_sara_oficina.

16.3.1.3 Modelo multidimensional

16.3.1.3.1 Diseño Conceptual del Data mart atenciones brindadas (UML)

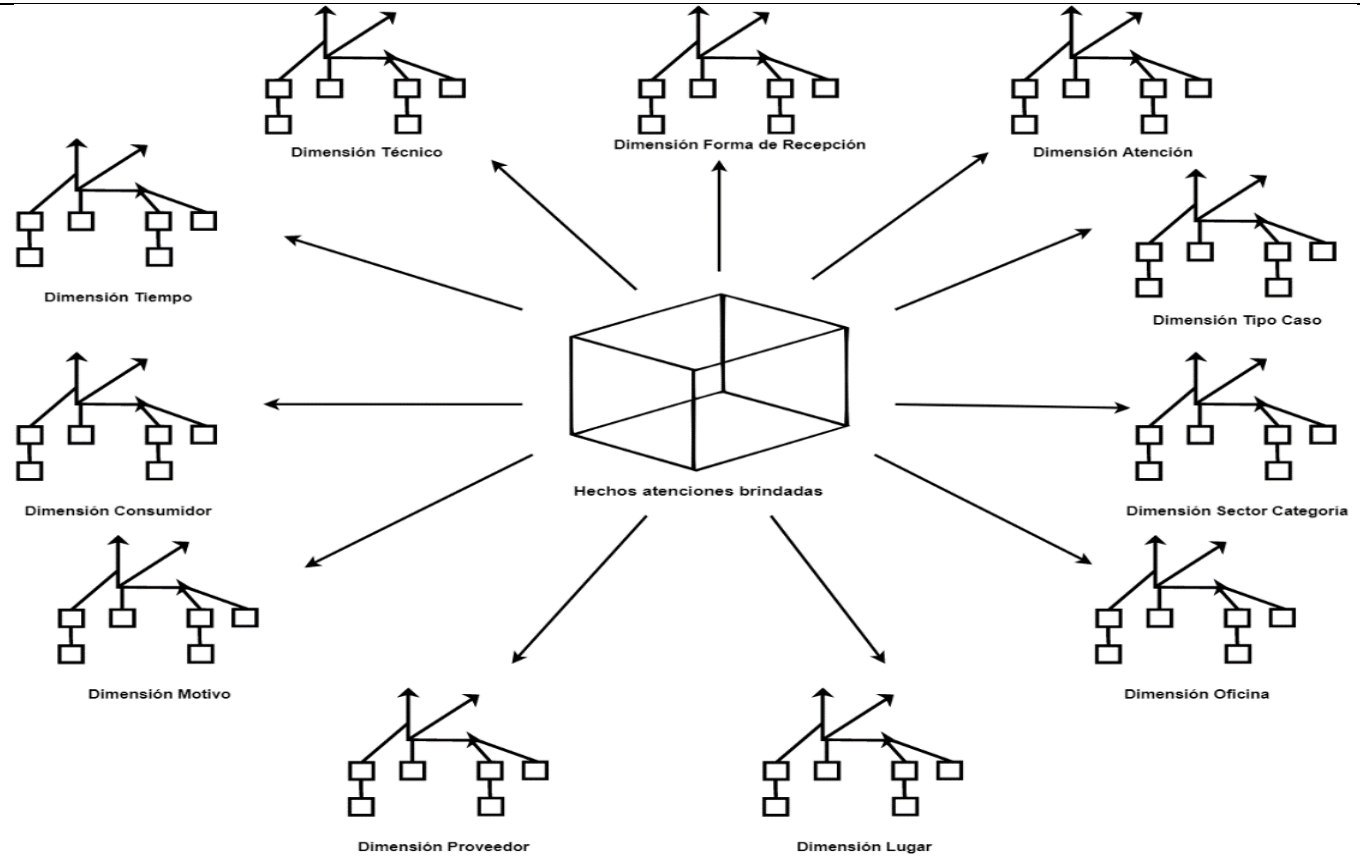


Nivel 1: Definición del modelo

Paquete Hechos:
Atenciones Brindadas.

Paquete Dimensiones:

1. Tiempo
2. Consumidor
3. Motivo
4. Proveedor
5. Lugar
6. Oficina
7. Categoría
8. Caso
9. Atención
10. Forma Recepción
11. Técnico



Nivel 2: Definición del esquema estrella.

Figura 129 Niveles 1 y 2 del diseño conceptual

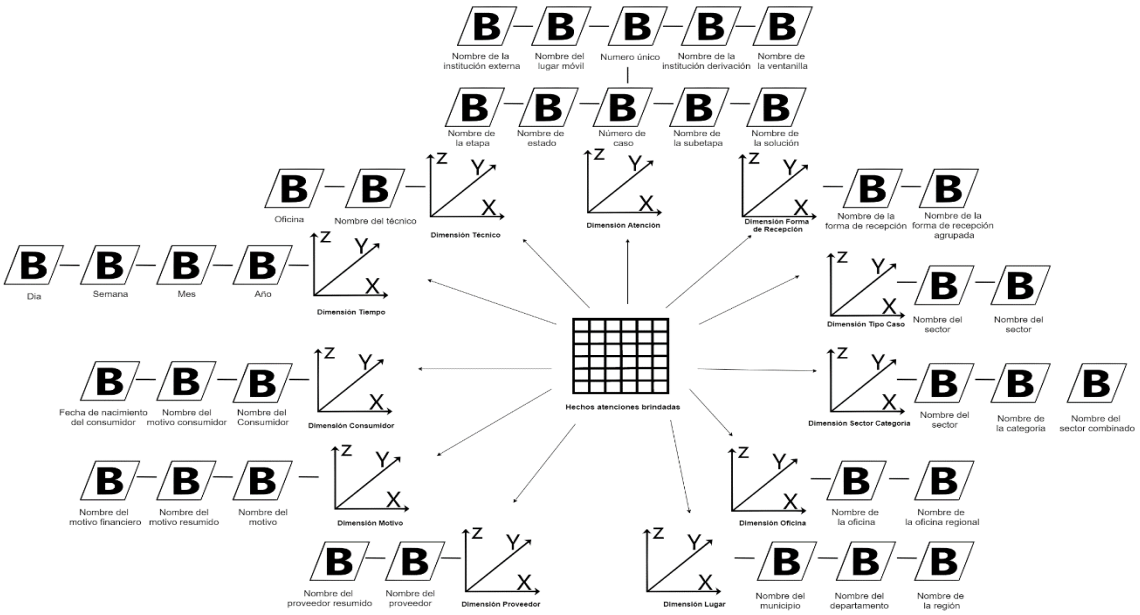
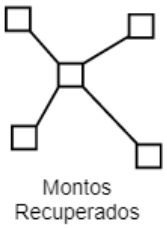
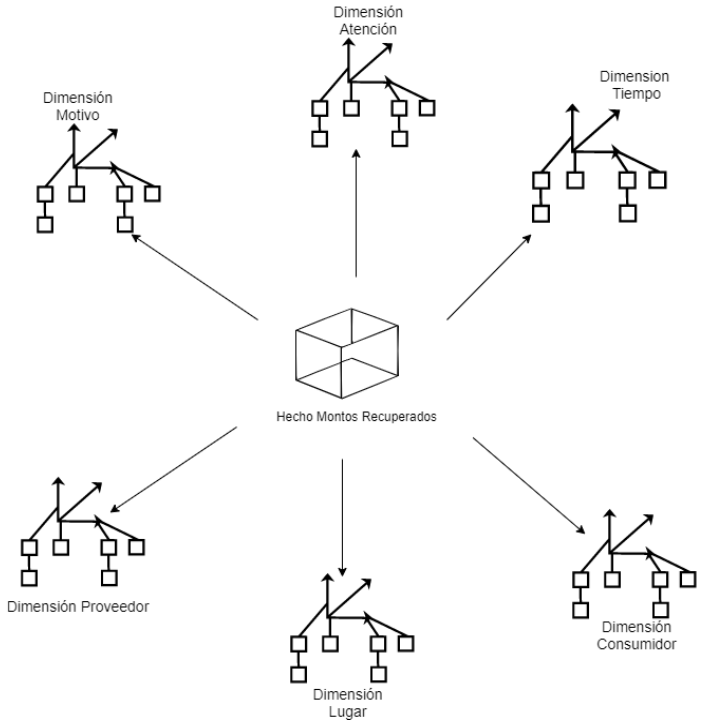


Figura 130 Nivel 3 - Diagrama Conceptual Data Mart Atenciones Brindadas.

16.3.1.3.2 Diseño conceptual del Data mart montos reclamados y recuperados (UML).



Nivel 1: Definición del modelo



Nivel 2: Definición del esquema estrella.

Paquete Hechos:
Montos Recuperados.

Paquete Dimensiones:

1. Tiempo
2. Proveedor
3. Lugar
4. Consumidor
5. Atención
6. Motivo

Figura 131 Niveles 1 y 2 del diseño conceptual

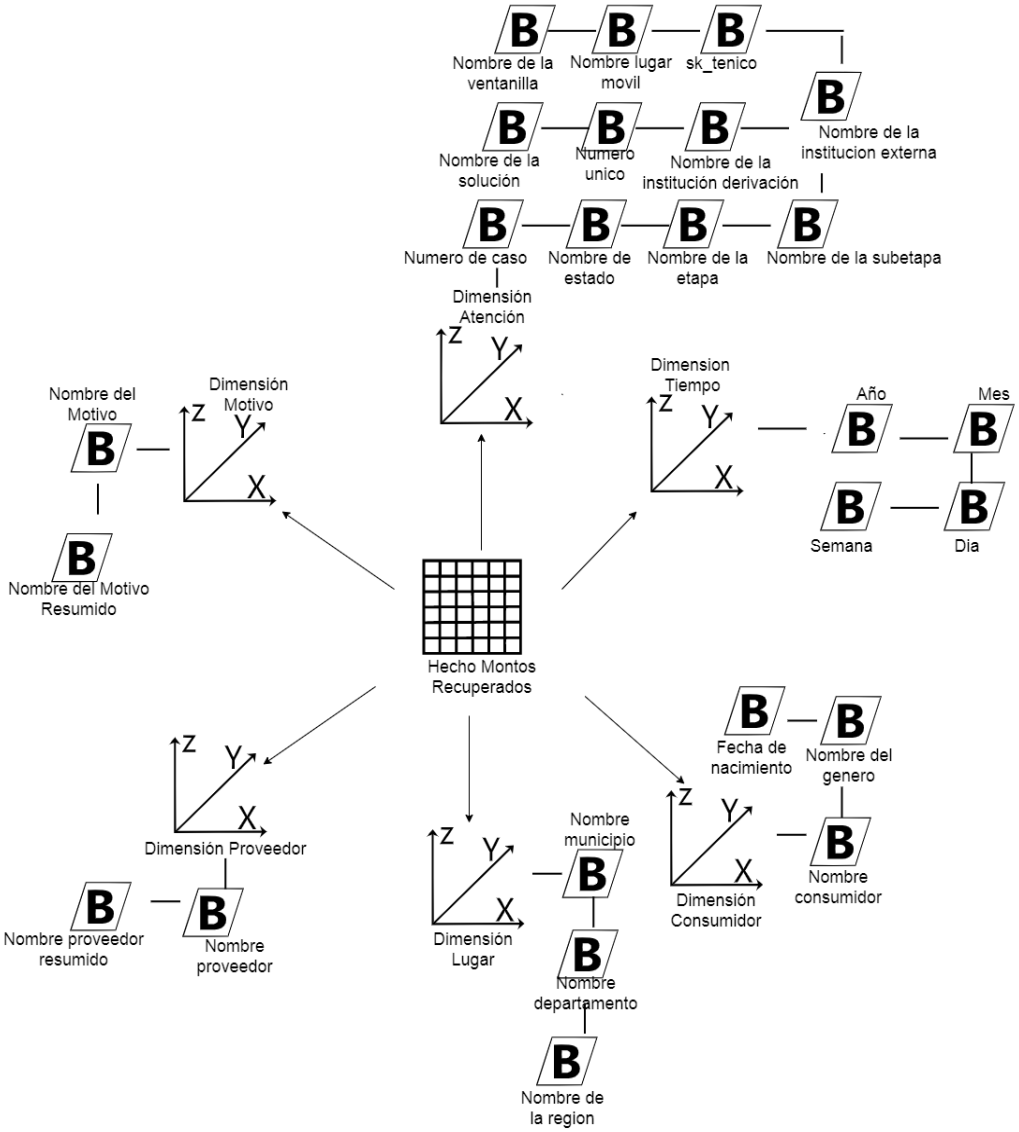


Figura 132 Nivel 3 Diagrama Conceptual Data Mart Montos Recuperados

16.3.1.3.3 Diseño de la base de datos

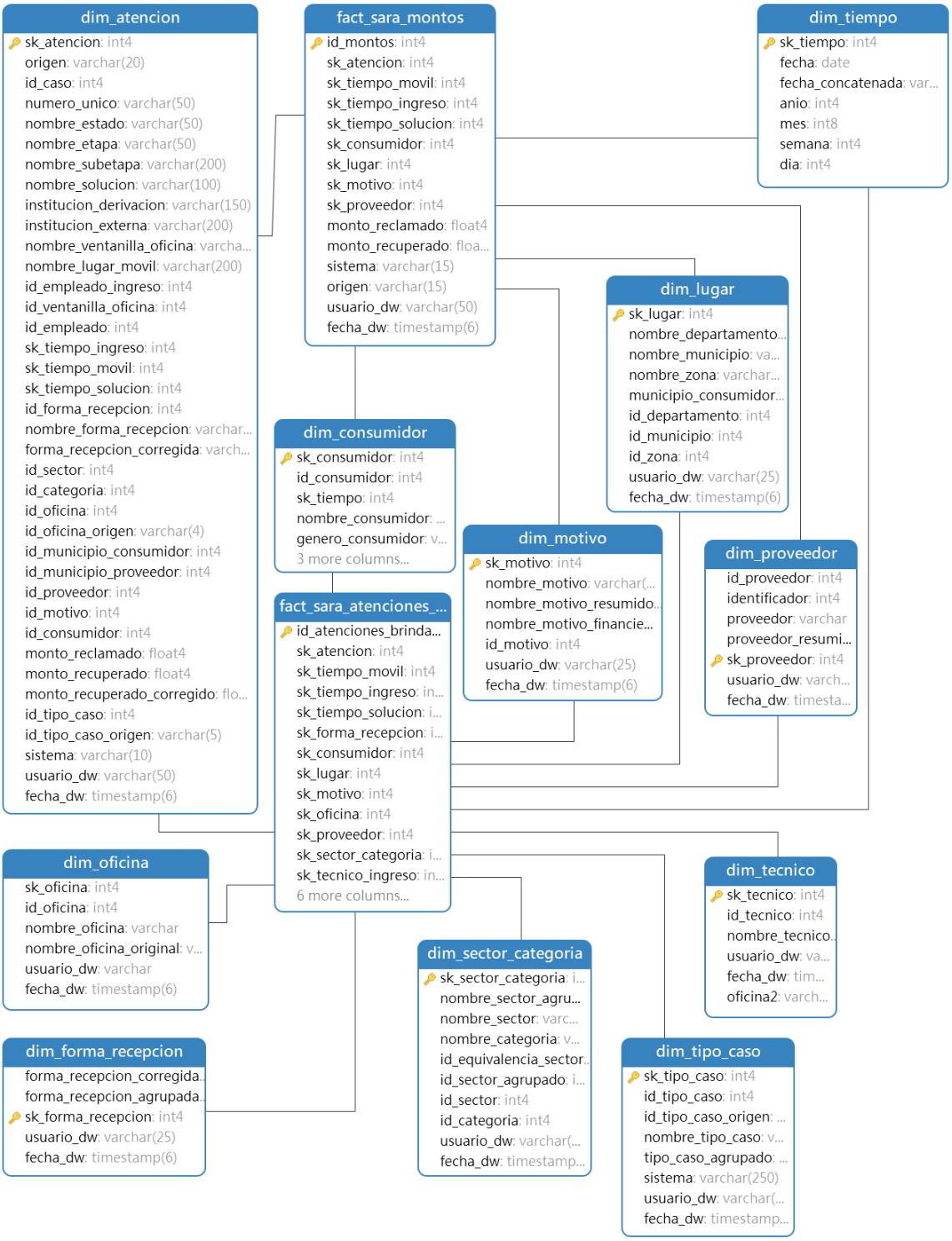


Figura 133 Diseño de modelo multidimensional.

16.3.1.3.4 Diseño de procesos de ETL

A continuación, se muestra el diseño de una Dimensión y una Tabla de Hechos

16.3.1.3.4.1 Dimensión Oficina

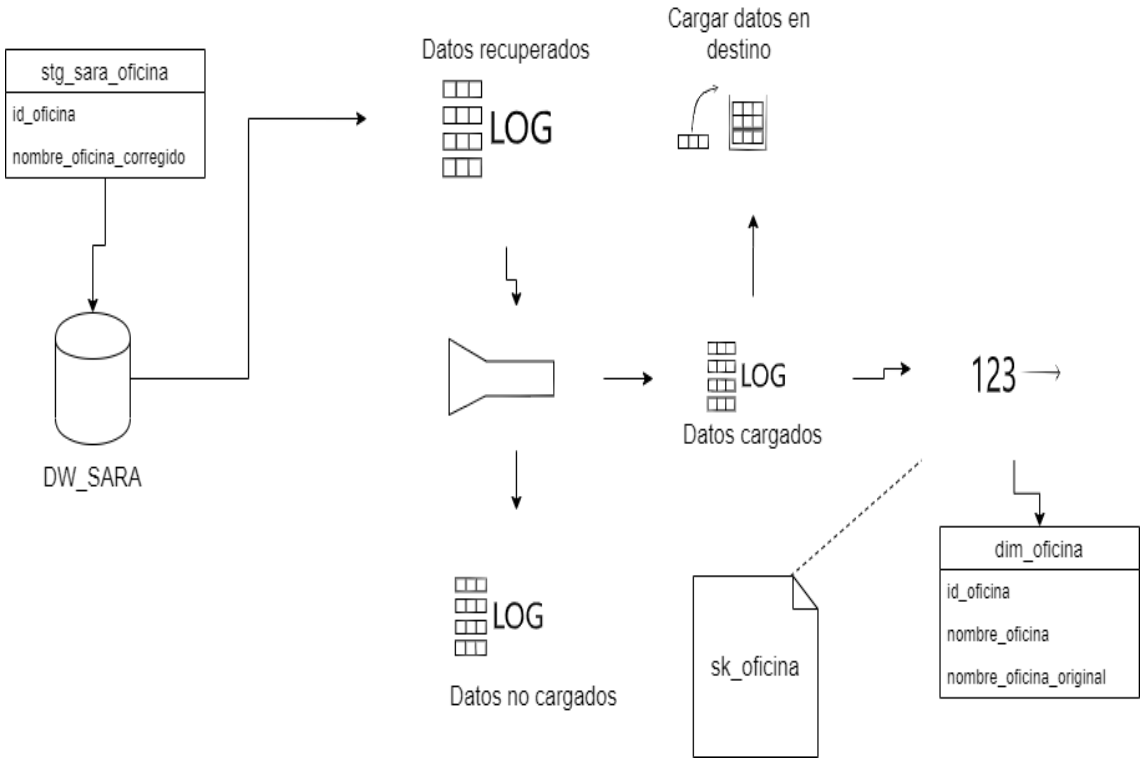


Figura 134 Diseño UML job dim_oficina.

La Figura 134 muestra el diseño UML para el Job dim_oficina en el cual el proceso que se sigue se lista a continuación:

1. Conectar a la base de datos de origen en este caso el staging_area
2. Unir las tablas que conformarán la dimensión lugar, en este caso stg_sara_oficina
3. Reportar al Log los datos recuperados.
4. Filtrar para datos cargados y no cargados
5. Reportar al Log datos cargados y datos no cargados.
6. Calcular clave sustituta.
7. Cargar datos a la tabla de destino dim_oficina.

16.3.1.3.4.2 Hecho montos reclamados/recuperados

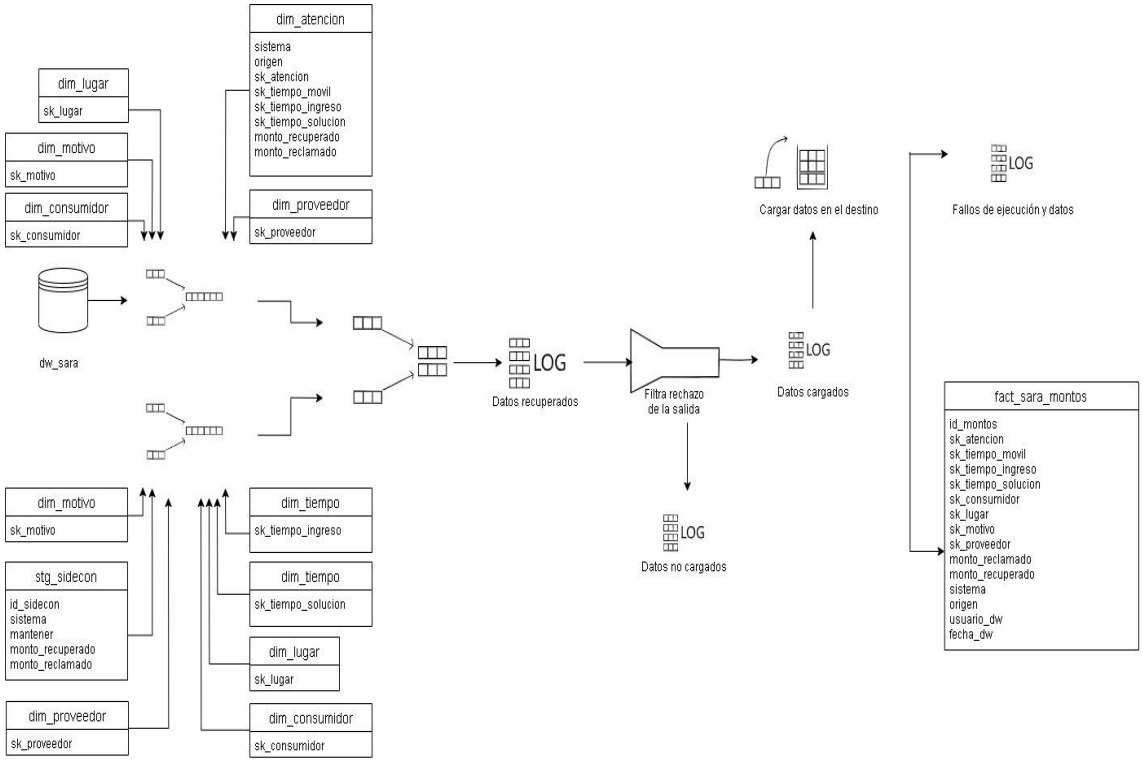


Figura 135 Diseño UML Job fact_sara_montos

La Figura 135 muestra el diseño UML para el Job fact_sara_montos en el cual el proceso que se sigue se lista a continuación:

1. Conectar a la base de datos de origen en este caso el dw_sara
2. Unir las tablas que conformarán el hecho:
   Atenciones SARA:
   a. dim_atencion
   b. dim_proveedor
   c. dim_consumidor
   d. dim_lugar
   e. dim_oficina
   f. dim_forma_recepcion
   g. dim_motivo
   h. dim_tecnico
   i. dim_sector_categoria
   j. dim_tipo_caso

   Atenciones SIDECON:
   a. stg_sidecon
   b. dim_motivo
   c. dim_proveedor
   d. dim_oficina
   e. dim_consumidor
   f. dim_tiempo
   g. dim_lugar
   h. dim_tipo_caso
   i. dim_sector_categoria
3. Hacer merge con datos de las dos uniones del numeral anterior.
4. Reportar al Log los datos recuperados.
5. Filtrar para datos cargados y no cargados
6. Reportar al Log datos cargados y datos no cargados.
7. Cargar datos a la tabla de destino fact_sara_montos.
8. Reportar al Log fallos de ejecución y/o datos.

16.3.1.3.4.3 Desarrollo de los procesos ETL

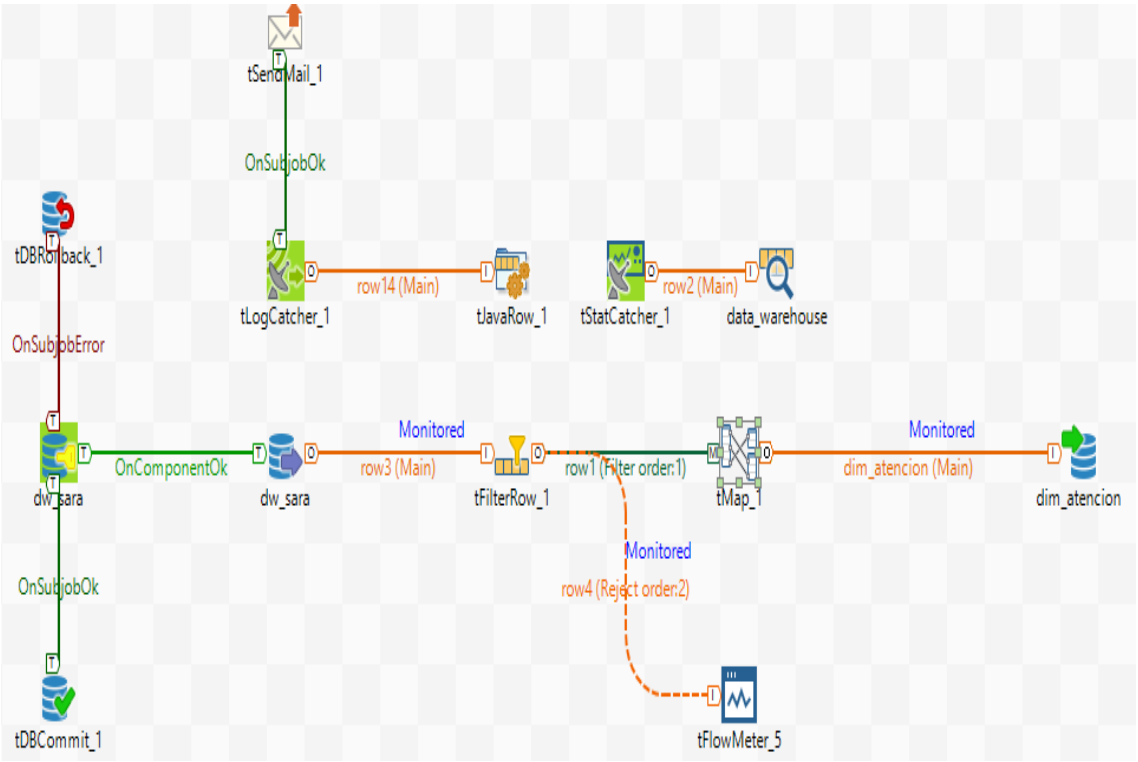


Figura 136 Job dim_atencion.

La Figura 136 muestra el Job dim_atencion en el cual el proceso que se sigue se lista a continuación:

1. Se crea la conexión a la base de datos del staging área.
2. Se obtienen los registros para cargar la dimensión atención.
3. Se excluyen los registros con tipo caso nulos
4. Si algún dato no se carga correctamente se registran en el nodo de datos_no_cargados.
5. Se cargan los datos a la tabla dim_atencion y se cierran las conexiones.
6. Si existe un error en la ejecución, se captura y se envía por correo electrónico y se realiza un rollback.

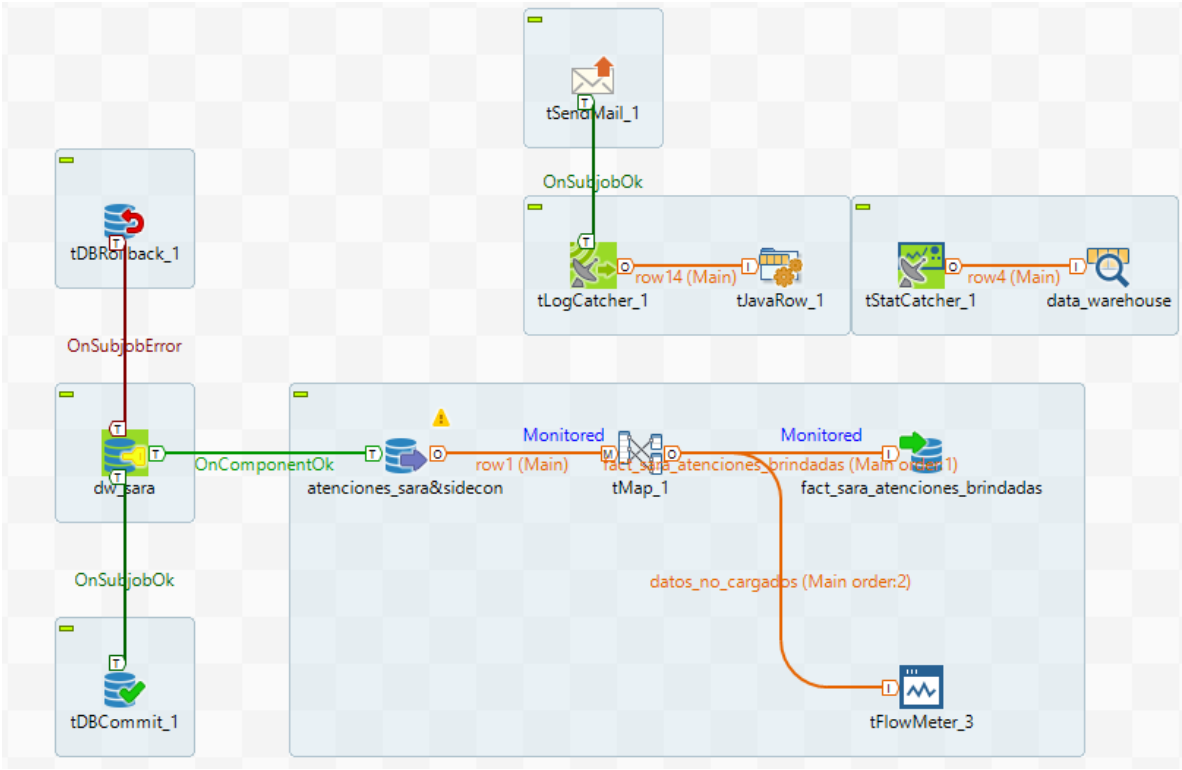


Figura 137 Job de fact_sara_atenciones_brindadas.

La Figura 137 muestra el Job fact_sara_atenciones_brindadas en el cual el proceso que se sigue se lista a continuación:

1. Se crea la conexión a la base de datos transaccional.
2. Se crea la conexión a la base de datos del modelo multimensional.
3. Se realizan las uniones a todas las dimensiones.
4. Si algún dato no se carga correctamente se registrán en el nodo de datos_no_cargados.
5. Se cargan los datos a la tabla fact_sara_atenciones_brindadas y se cierran las conexiones.
6. Si existe un error en la ejecución, se captura y se envía por correo electrónico y se realiza un rollback.

16.3.1.3.5 Pruebas

16.3.1.3.5.1 Pruebas de ejecución

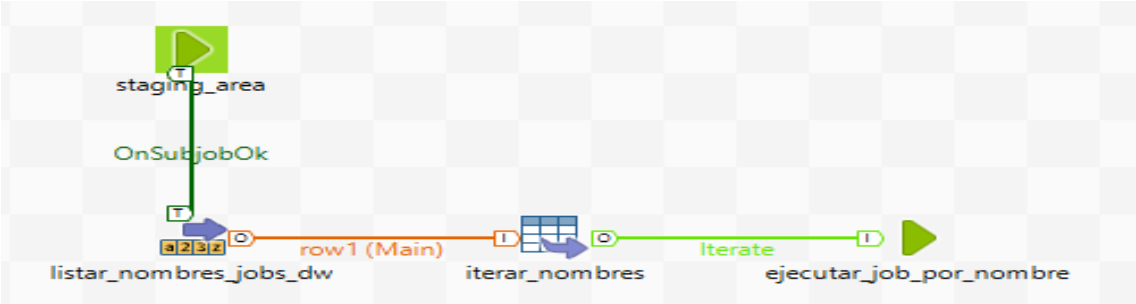


Figura 138 Job data_warehouse_ejecucion.

En la Figura 138 se observa el job que realiza una llamada al job ejecución staging area y luego realiza una llamada a todos los job del modelo multidimensional

- dim_tiempo
- dim_lugar
- dim_forma_recepcion
- dim_oficina
- dim_sector_categoria
- dim_consumidor
- dim_motivo
- dim_proveedor
- dim_tecnico
- dim_tipo_caso
- dim_atencion
- fact_sara_atenciones_brindadas
- fact_sara_montos

Se presentan los resultados para dos ejecuciones del Job, en el cual se constatan la cantidad de registros recuperados del origen, cargados y no cargados en el destino.

JOB	DATOS RECUPERADOS	DATOS CARGADOS	DATOS NO CARGADOS	RESULTADO	DURACION (seg)
dim_atencion	850520	850472	48	success	244.31
dim_consumidor	894978	894978	0	success	24.17
dim_forma_recepcion	14	14	0	success	1.00
dim_lugar	521	521	0	success	0.62
dim_motivo	685	392	293	success	1.32
dim_oficina	14	14	0	success	0.42
dim_proveedor	38139	38139	0	success	1.51
dim_sector_categoria	959	959	0	success	1.76
dim_tecnico	1063	1063	0	success	0.40
dim_tipo_caso	29	29	0	success	0.42
dw_generate_excel	967440	967440	0	success	1695.38

fact_sara_atenciones_brindadas	967440	967440	0	success	39.27
fact_sara_montos	174794	174794	0	success	9.19

Tabla 52 Resultados obtenidos de la ejecuación.

Resultado obtenido en la base de datos:
Al ejecutar el job de staging area de ejecución se crearon las siguientes tablas con los datos de las bases de SARA.

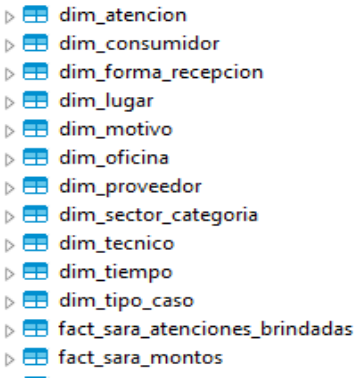


Figura 139 Tablas Modelo Multidimensional.

Contenido tabla dim_oficina

	123 sk_oficina	123 id_oficina	ABC nombre_oficina	ABC nombre_oficina_original
1	1	1	San Salvador	Oficina San Salvador
2	2	2	San Miguel	Oficina Regional San Miguel
3	3	3	Santa Ana	Oficina Regional Santa Ana
4	4	4	San Salvador	Gerencia del Centro de Solución de Controversias de Ser
5	5	5	Call Center	Oficina Call Center Plan de La Laguna
6	6	6	Plan de la Laguna	Oficina Plan de La Laguna
7	7	7	Oficina Educacion	Oficina Educacion
8	8	8	Oficina Movil	Oficina Movil
9	9	9	Plan de la Laguna	Oficina Ciudadania y Consumo
10	10	10	Oficina Descentralizada	Oficina Descentralizada

Figura 140 Tabla dim_oficina.

Contenido fact_sara_atenciones_brindadas

	123 id_atenciones_brindadas	123 sk_atencion	123 sk_tiempo_movil	123 sk_tiempo_ingreso	123 sk_tiempo_solucion
1	1	408.776	[NULL]	34.368	[NULL]
2	2	508.743	[NULL]	35.701	[NULL]
3	3	525.064	[NULL]	36.042	[NULL]
4	4	491.429	[NULL]	35.477	[NULL]
5	5	494.693	[NULL]	35.514	[NULL]
6	6	512.870	[NULL]	35.762	[NULL]
7	7	526.429	[NULL]	36.130	[NULL]
8	8	524.736	[NULL]	36.021	[NULL]
9	9	524.738	[NULL]	36.021	[NULL]
10	10	524.989	[NULL]	36.021	[NULL]

Figura 141 Tabla fact_sara_atenciones_brindadas.

16.3.1.3.5.2 Pruebas de sumarización

Se aplicaron los siguientes criterios en la transformación de los datos en base a requerimiento para asegurar que las sumarizaciones cuadren respecto con las salidas que se obtienen actualmente:

1. Se deben descartar las atenciones en estado Eliminado o Inactivo.
2. Se deben tomar como atenciones brindadas únicamente Asesorías del CSC, Asesorías Call Center, Reclamos del CSC y Gestiones del Call Center.
3. Unicamente se deben incluir los tipos de atención: Asesoría, Asesoría Web, Orientación de consumo, Denuncia personal, Denuncia colectiva, Derivación y Gestión y estas deben de cambiarse por su respectiva equivalencia:

Tipo de caso	Tipo de caso agrupado
Denuncia Personal	Denuncia
Gestión	Gestión
Denuncia Colectiva	Denuncia
Derivación	Derivación
Asesoría	Asesoría
Orientación de Consumo	Asesoría
Asesoría Web	Asesoría

Tabla 53 Equivalencia de tipo de caso.

4. Se debe derivar dos campos del motivo de la atención los cuales son motivo resumido y motivo financiero, estos se proporcionaron en una hoja de cálculos la cual es procesada dentro del ETL generando los dos campos adicionales.
5. Se debe derivar dos campos del Sector de la atención los cuales son Sector 2017 y Sector resumido, estos se proporcionaron en una hoja de cálculos la cual es procesada dentro del ETL generando los dos campos adicionales.
6. Se debe calcular la solución que la atención tiene en el momento actual la cual se base en la etapa, subetapa y estado igualmente fue proporcionada cual es la solución correspondiente para una atención para las distintas combinaciones de los parámetros mencionados.
7. Se debe calcular la fecha correspondiente a la solución la cual sigue la siguiente regla:
    a. Para las gestiones de Call Center esta debe ser igual a la fecha de ingreso.
    b. Para las denuncias y gestiones del CSC está solo existe para los casos que no están abiertos y debe ser igual a la fecha máxima de entre: fecha de solución, fecha de cierre, fecha de desistimiento y fecha de envío al Tribunal Sancionador, si no se tiene dato para ninguna de esas fechas debe utilizarse la fecha de archivo.
8. Para el campo del monto recuperado se debe mostrar valor únicamente a las atenciones cerradas en audiencia de gestión, avenimiento y conciliación.
9. Para el caso del nombre del proveedor ya que se puede ingresar de varias formas dependiendo como el consumidor lo conozca se vuelve necesario calcular un nombre de proveedor resumido el cual también es procesado desde un archivo de hoja de cálculo y procesado por el ETL para sacar el su nombre de proveedor resumido correspondiente.

10. En el caso del campo Oficina cuando este sea Oficina Móvil u Oficina Descentralizada se debe calcular una oficina sustituta que corresponde a la oficina a la que el técnico está asignado, ya que un mismo técnico dentro del sistema SARA puede tener varios perfiles y cada uno de ellos asignado a distinta oficina, el valor que se desea colocar si se cumple la condición fue proporcionado en una hoja de cálculo también para ser procesado dentro del ETL.
11. Se debe derivar dos campos de la forma de recepción de la atención los cuales son forma de recepción corregida y forma de recepción agrupada, estos se proporcionaron en una hoja de cálculos la cual es procesada dentro del ETL generando los dos campos adicionales.

17 Sprint 5

17.1 Descripción historias de usuario

Código	RA201
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se determine la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año.
Razón	Para monitorear en que meses algunos proveedores realizan prácticas abusivas y proceder con las respectivas inspecciones.
Criterios de aceptación	El catálogo de los proveedores debe estar depurados en la medida de lo posible.
	Los proveedores en conjunto de las cantidades de atenciones en las que se vieron implicados constituirán las transacciones.
	La relación a encontrar es si hubo o no aumento en las atenciones a proveedores mes a mes.
Validación	Validar que los resultados correspondan a relaciones encontradas por la minería de datos en conjunto con sus indicadores de rendimiento.
	Que se muestre únicamente proveedores para los cuales el aumento en las atenciones fue significativo.
	Se comprobará que los indicadores de rendimiento sean aceptables para las relaciones.
Valor del negocio	900
Puntos de historia	10
ROI	90

Tabla 54 Historia de Usuario RA201.

Código	RA202
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se segmenten los consumidores en base a los motivos en los cuales han solicitado atención a la DC.
Razón	Para brindar programas de educación en consumo a esos grupos de personas.
Criterios de aceptación	El catálogo de los consumidores debe estar depurado en la medida de lo posible.
	Se debe determinar una cantidad mínima de casos en los motivos para que estos sean sujetos del análisis.
	Se debe etiquetar cada consumidor al segmento al que se ha asignado.
Validación	Validar que la cantidad de grupos satisfaga el mínimo correspondiente a los programas que se deseen recomendar.
	Que se muestren únicamente consumidores a los cuales se les ha brindado atenciones de reclamos.
	Debe crear las agrupaciones únicamente para los motivos en los que más se hayan atendido consumidores.
Valor del negocio	900
Puntos de historia	5
ROI	180

Tabla 55 Historia de Usuario RA202.

Código	RA203
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se clasifique la influencia que ha tenido la DC en las atenciones en base a la solución y montos recuperados.
Razón	Para destacar la labor realizada en materia de los derechos de los consumidores.
Criterios de aceptación	La solución a tomar como base debe ser determinada por la combinación del estado, etapa y subetapa de los casos.
	Se deben tomar montos recuperados únicamente de casos cerrados en audiencia de gestión, avenimiento y conciliación.
	Se deben tomar 3 clasificaciones alta influencia, mediana influencia, baja influencia.
Validación	Se comprobará que únicamente se hayan utilizado montos recuperados con las soluciones especificadas anteriormente.
	Se validará que a los casos sujetos al análisis se les haya determinado una solución.
	Se validará que los resultados arrojados por el indicador de rendimiento sean aceptables.
Valor del negocio	900
Puntos de historia	10
ROI	90

Tabla 56 Historia de Usuario RA203.

Código	RA204
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se pronostiquen casos a recibir en fechas futuras.
Razón	Para planificar recursos en base a la demanda.
Criterios de aceptación	Se tomarán en cuenta casos recibidos sin importar la solución a la que se haya llegado.
	Se deben descartar los casos que hayan sido eliminados lógicamente.
	Se debe pronosticar por lo menos un mes.
Validación	Se comprobará que los indicadores de rendimiento sean aceptables para el requerimiento.
	Se validará que los resultados sean congruentes respecto a los de años anteriores.
	Se validará que las cantidades que se utilicen en el análisis sean correspondientes a las que la UACM reporta actualmente.
Valor del negocio	800
Puntos de historia	5
ROI	160

Tabla 57 Historia de Usuario RA204.

Código	RA205
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea identificar qué solución tendrán los casos recibidos en base a la edad y otros parámetros que se consideren relevantes de los consumidores.

Razón	Para tener una noción del resultado y poder tomar las acciones necesarias.
Criterios de aceptación	La solución a tomar como base debe ser determinada por la combinación del estado, etapa y subetapa de los casos.
	Se debe de depurar las fechas de nacimiento de los consumidores.
	Las clasificaciones serán las diferentes soluciones que pueda presentar un caso.
Validación	Comprobar que se haya utilizado únicamente fechas de nacimiento validas de consumidores.
	Validar que se tomen rangos de edades aceptables y bien definidas.
	Se validará que los resultados arrojados por el indicador de rendimiento sean aceptables.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 58 Historia de usuario RA205.

Código	**RA206**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea clasificar a el comportamiento de los proveedores en base los montos reclamados, montos recuperados y solución.
Razón	Para poner en perspectiva el actuar de los proveedores cuando son sujeto de denuncias interpuestas en su contra.
Criterios de aceptación	La solución a tomar como base debe ser determinada por la combinación del estado, etapa y subetapa de los casos.
	Las atenciones a tomar para el análisis deben tener una solución ya que se necesita un monto recuperado.
	Las clasificaciones serán los distintos comportamientos del proveedor.
Validación	Se validará que a los casos sujetos al análisis se les haya determinado una solución.
	Se validará que se incluyan únicamente proveedores a los cuales se les haya interpuesto denuncias en un número considerable de veces.
	Se validará que los resultados arrojados por el indicador de rendimiento sean aceptables.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 59 Historia de usuario RA206.

Código	RA207
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se pronostiquen casos solucionados en fechas futuras.
Razón	Para establecer un estimado de las metas y así poder medir el rendimiento.
Criterios de aceptación	Se tomarán en cuenta casos solucionados sin importar el tipo de solución.
	Se deben descartar los casos que hayan sido eliminados lógicamente.
	Se debe pronosticar por lo menos un mes.
Validación	Se comprobará que los indicadores de rendimiento sean aceptables para el requerimiento.
	Se validará que los resultados sean congruentes respecto a los de años anteriores.
	Se validará que las cantidades que se utilicen en el análisis sean correspondientes a las que la UACM reporta actualmente.
Valor del negocio	800
Puntos de historia	5
ROI	160

Tabla 60 Historia de usuario RA207.

Código	RA208
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer en qué meses se dan la mayor cantidad de casos.
Razón	Para poder reorientar el recurso entre las distintas unidades en base a la demanda.
Criterios de aceptación	Se deberá realizar utilizando un análisis de minería de datos.
	Se debe identificar cuales meses reciben mayor demanda de casos.
	Se debe mostrar al menos dos grupos o segmentaciones.
Validación	Se comprobará que se calcule la segmentación de casos por mes para los diferentes años.
	Se comprobará que los resultados obtenidos se guarden en una base datos.
	Se comprobará que se utilice toda la historia de datos.
Valor del negocio	400
Puntos de historia	8
ROI	50

Tabla 61 Historia de usuario RA208.

17.2 Caso de uso

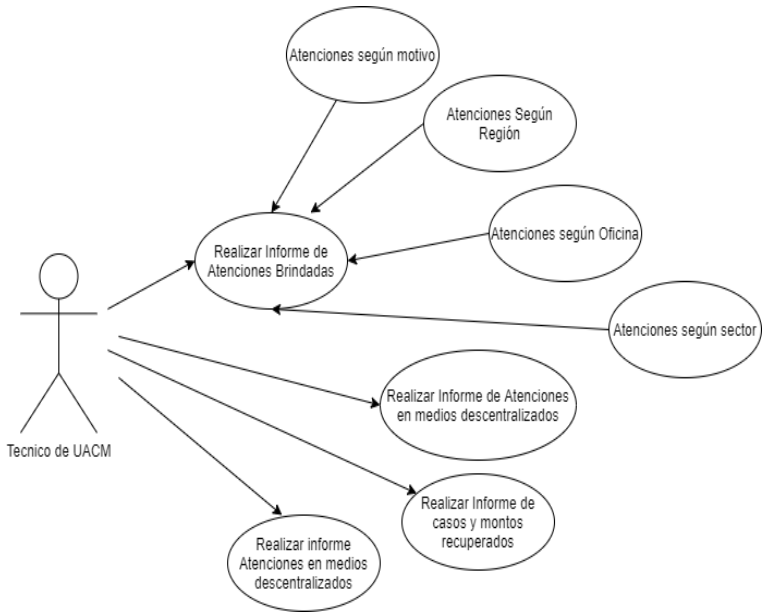


Figura 142 Diagrama Casos de Uso Sprint 2.

17.3 Exploración de los datos

Nombre ingles	**Nombre español**	**Descripción**
Column	Columna	Nombre de la variable hacer analizada.
Min	Mínimo	Valor mínimo que presenta la variable.
Mean	Media	Valor la media que presenta la variable.
Median	Mediana	Valor la mediana que presenta la variable.
Max	Máximo	Valor máximo que presenta la variable.
Std. Dev.	Desviación estándar	Cantidad de variación que presenta los valores de la variable.
Skewness	Asimetría estadística	Valor que permite saber el grado de simetría o asimetría que presenta la variable.
No. Missing	Valores ausentes	Ayuda a predecir la precisión de los estadísticos de la variable.
Histogram	Histograma	Distribución de frecuencias de la variable.

Tabla 62 Descripción de nombre de columnas en exploración de datos.

Nombre Solución	Estadísticos

Desistimiento

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	No. NaNs	No....	No...	Median	Row count	Histogram
monto_reclamado	0	4,458,082	1,602.298	49,474.865	2,447,762,264.562	83.041	7,386.618	14,327,745.576	0	0	0	0	?	8942	0 4.458.082
monto_recuperado...	?	?	NaN	NaN	NaN	NaN	NaN	0	8942	0	0	0	?	8942	?

Falta de Ratificación y Prevención

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall ...	No. missings	N...	N...	No...	Median	Row co...	Histogram
monto_reclamado	0	106,424.398	383.419	2,310.39	5,337,899.743	25.835	927.133	2,242,232.07	0	0	0	0	?	5848	0 106.424
monto_recuperado	?	?	NaN	NaN	NaN	NaN	NaN	0	5848	0	0	0	?	5848	?

Tribunal Sancionador

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	N...	N...	N...	Median	Row co...	Histogram
monto_reclamado	0	3,420,000	3,278.698	49,338.144	2,434,252,479.005	53.849	3,351.685	28,541,070.337	0	0	0	0	?	8705	0 3.420.000
monto_recuper...	?	?	NaN	NaN	NaN	NaN	NaN	0	8705	0	0	0	?	8705	?

Conciliación

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	N...	N...	N...	N...	M...	Row co...	Histogram
monto_reclamado	0	999,992,128	41,825.311	6,125,393.721	37,520,448,241,040.086	161.798	26,390.144	1,125,853,727.066	0	0	0	0	?	26918	0E0 1E9
monto_recuperado	0.01	2,438,789.75	1,255.992	26,487.745	701,600,641.687	62.063	4,601.782	27,216,100.525	5249	0	0	0	?	26918	0 2.438.790

Nombre Solución	Estadísticos
Avenimiento	

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. mis...	...	...	...	Median	Row co...	Histogram
monto_reclamado	0	999,999,168	9,451.301	2,847,853.695	8,110,270,667,499.46	347.945	122,067.613	1,177,225,671.987	2524	0	0	0	?	127081	0E0 1E9
monto_recuperado	0.01	300,000	297.566	2,624.714	6,889,121.008	63.799	5,487.283	23,101,217.306	49447	0	0	0	?	127081	0 300.000

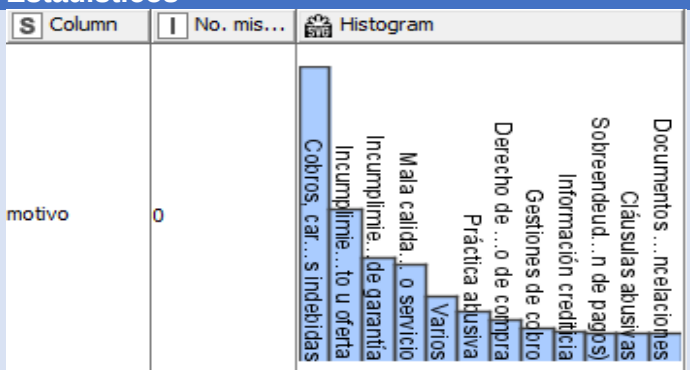
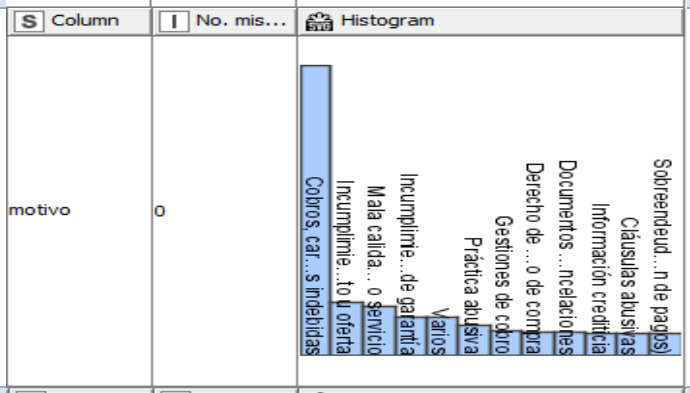
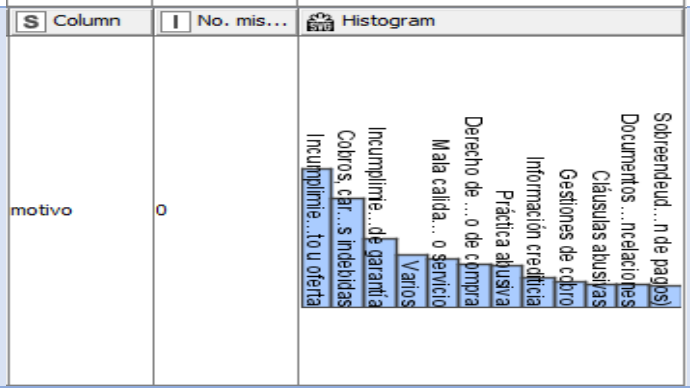
Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	...	...	...	Median	Row co...	Histogram
monto_reclamado	0	60,000	1,572.33	4,932.587	24,330,409.779	7.096	66.162	727,988.649	0	0	0	0	?	463	0 60,000
monto_recuperado	45	6,795.37	1,803.113	2,407.905	5,798,007.355	1.579	1.901	14,424.9	455	0	0	0	?	463	45 6.795

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	...	...	...	Median	Row count	Histogram
monto_reclamado	0	4,780	340.433	868.47	754,239.982	3.447	12.395	24,851.62	0	0	0	0	?	73	0 4.780
monto_recuperado	?	?	NaN	NaN	NaN	NaN	NaN	0	73	0	0	0	?	73	?

Cerrado por razones de oficio

Tabla 63 Tabla comparativa de estadísticas por solución

En la Tabla 63 se presenta un comparativo de estadísticas de las distintas soluciones que presenta una atención, esta es generada desde la herramienta Knime Analytics en la cual se muestran estadísticos como mínimo, máximo, media, desviación estándar, varianza, mediana, suma total, número de valores faltantes y recuento de filas en todas las columnas numéricas, y el recuento todos los valores nominales junto con sus ocurrencias, basándose en los montos reclamados y recuperado.

Nombre de solución	Estadísticos
Desistimiento	
Falta de Ratificación y Prevención	
Tribunal Sancionador	

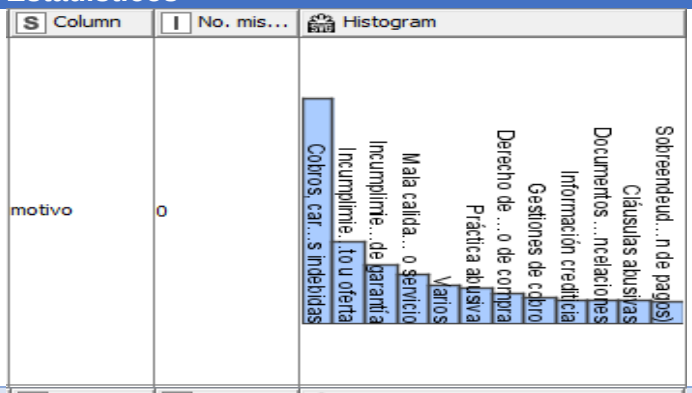
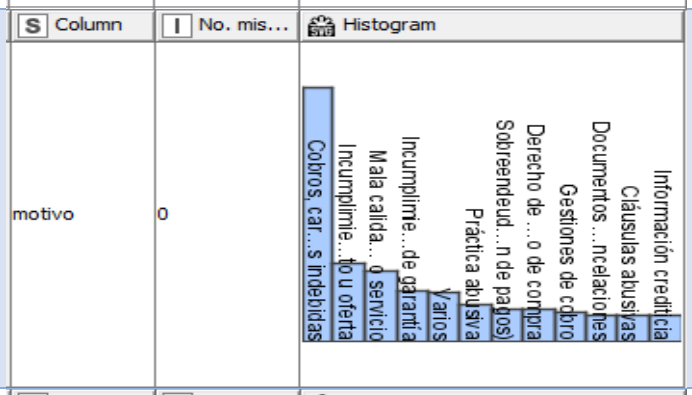
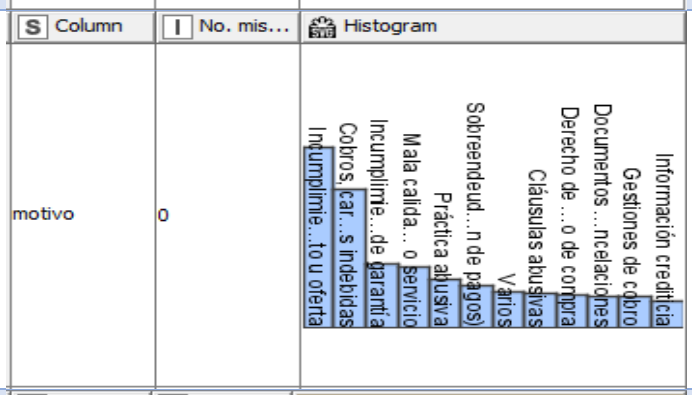
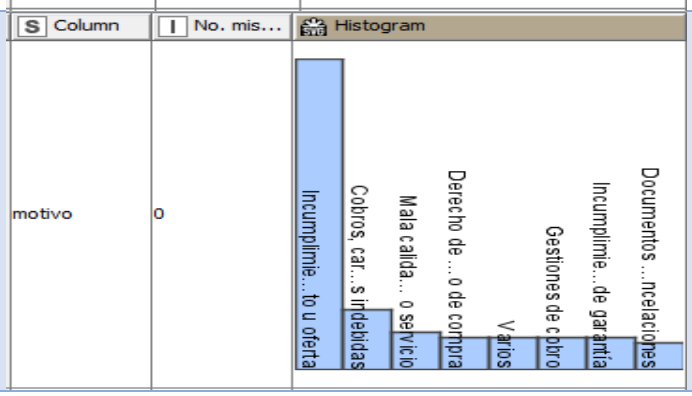
Nombre de solución	Estadísticos
Conciliación	
Avenimiento	
Audiencia de Gestión	
Cerrado por razones de oficio	

Tabla 64 Tabla comparativa de estadísticas por motivo.

En la Tabla 64 se presenta un histograma de las distintas soluciones que presenta una atención, esta es generada desde la herramienta Knime Analytics tomando como base los motivos.

17.4 Técnica de asociación

17.4.1 Determinar la relación existente entre el aumento de las denuncias hacia proveedores

17.4.1.1 Contenido del caso

N° de caso: C-ASO-01	
Técnica	Asociación.
Algoritmo	A priori
Población	Datos provenientes del modelo multidimensional SARA desde el año 2005 hasta el año 2019.
Variables	Se analizan proveedores, los 12 meses del año para los años desde los cuales se tienen datos disponibles.
Hipótesis	Determinar la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Reglas de asociación.
Interpretación de resultados	Evaluar en base a indicadores de rendimiento (Support, Confidence, Lift).
Herramienta de software	Knime Analytics Platform

Tabla 65 Contenido del caso C-ASO-01.

17.4.1.2 Población

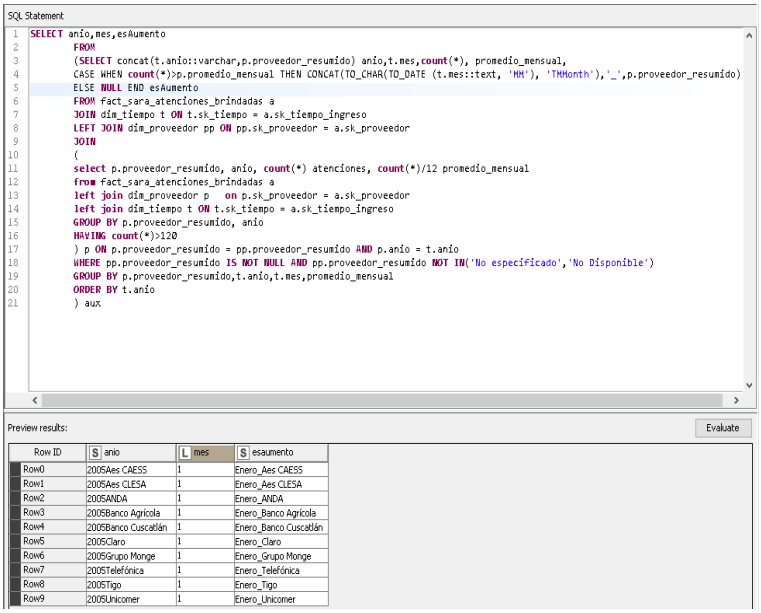


Figura 143 Set de datos para el caso C-ASO-01.

La Figura 143 muestra la consulta para obtener el set de datos el cual corresponde al aumento en la cantidad de reclamos para los proveedores del mes respecto al promedio mensual según el año, para los proveedores que poseen más de 120 reclamos anuales, se utiliza el campo derivado proveedor_resumido.

17.4.1.3 Variables

Las variables involucradas en la exploración de los datos son el proveedor resumido, y de la dimensión tiempo los años y meses, se disponen los datos por todos los meses en los cuales se han interpuesto reclamos en contra de los proveedores en cuestión para los años en los cuales se tienen disponibles datos.

17.4.1.4 Hipótesis

El objetivo de minería de datos o hipótesis a llevar a cabo en esta historia de usuario es: "Determinar la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año".

17.4.1.5 Procedimiento

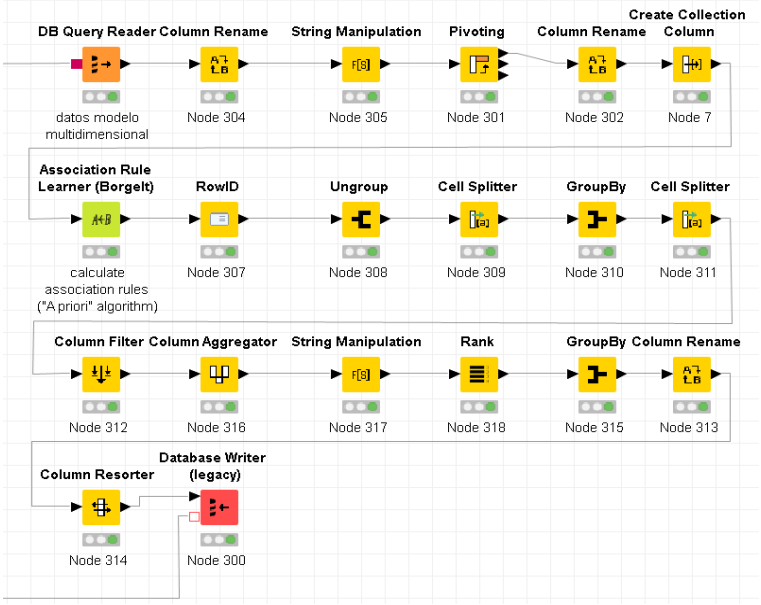


Figura 144 Flujo de trabajo aplicando A priori para el caso C-ASO-01.

La Figura 144 muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-ASO-01 se detalla a continuación las acciones que se llevan a cabo:

- **DB Query Reader:** Se ejecuta la consulta para seleccionar el set de datos.
- **Column Rename (Node 304):** Se renombra la columna "esaumento" por "esaumentooriginal".
- **String Manipulation (Node 305):** Para el campo "esaumentooriginal" se remueven los acentos y se convierte todo a mayúsculas, el resultado se guarda en el campo "esaumento".
- **Pivoting (Node 301):** Se define como variable de agrupamiento el año, como columna pivote el mes y como función de agregación First para el campo "esaumento".
- **Column Rename (Node 302):** Se renombran los números correspondientes a los 12 meses del año (1-12) por enero-diciembre como nombres de columnas.

- **Create Collection Column (Node 7):** Se crea una columna de colección con las 12 columnas correspondientes a los 12 meses del año ignorando valores faltantes para cada mes.
- **Association Rule Learner (Borgelt):** Se calculan las reglas de asociación con la columna de colección con un tamaño mínimo de set 2 y soporte mínimo de 7.
- **RowId (Node 307):** Se crea una columna id para cada regla generada por el algoritmo.
- **Ungroup (Node 308):** Se desagrupa la columna del antecedente, obteniendo n filas por cada item en la colección de cada regla.
- **Cell Splitter (Node 309):** Se separa el contenido de la columna Antecedente por "_" obteniendo 2 columnas correspondientes al mes y el nombre del proveedor.
- **GroupBy (Node 310):** Se agrupan los datos por la columna id utilizando como función de agregación "Concatenate" para la columna Antecedente (Únicamente mes) y para el resto de campos "First".
- **Cell Splitter (Node 311):** Se separa el contenido de la columna Consecuente por "_" obteniendo 2 columnas correspondientes al mes y el nombre del proveedor.
- **Column Filter (Node 312):** Se filtra las columnas manteniendo únicamente Concatenate(Antecedent (#1)) (String), Mean(ItemSetSupport) (Number (double)), Mean(RuleConfidence%) (Number (double)), Mean(RuleLift%) (Number (double)), First(Consequent)_Arr[0] (String), First(Consequent)_Arr[1] (String):
- **Column Aggregator (Node 316):** Se utiliza la función de agregación "Concatenate" para las columnas Concatenate (Antecedent (#1)) y First (Consequent)_Arr [0] que son únicamente nombres de los meses correspondientes al antecedente y consecuente.
- **String Manipulation (Node 317):** Se calcula la longitud de la columna Concatenada y se almacena en el campo length.
- **Rank (Node 318):** Se enumera el set de datos utilizando como atributo de agrupamiento el nombre del proveedor y como atributo de rank la longitud del campo concatenado ordenado descendentemente.
- **GroupBy (Node 315):** Se agrupa el set de datos por proveedor utilizando como función de agregación First así obteniendo el de mayor longitud en cada grupo.
- **Column Rename (Node 313):** Se renombran las columnas en base al sentido que dan los valores que contienen, antes de almacenar en la base de datos.
- **Column Resorter (Node 314):** Se ordenan las columnas de forma lógica antes de almacenar en la base de datos.
- **Database Writer (Node 300):** Se almacenan en una tabla llamada "sara_associationproveedormes" los resultados de la minería de datos del flujo de trabajo actual, el nodo realiza las tareas de crear la tabla con los campos que satisfagan el set de datos que se almacenará en ellos.

17.4.1.6 Resultados

Row ID	proveedor_resumido	...	..	r...	associationRule
Row0	AES CAESS	85.7	6	5,304.8	MAYO, NOVIEMBRE, JUNIO
Row1	AES CLESA	100	8	6,820.4	MAYO, OCTUBRE, NOVIEMBRE
Row2	AES EEO	100	8	5,967.9	AGOSTO, MAYO, SEPTIEMBRE
Row3	ANDA	100	10	4,774.3	ENERO, MARZO, FEBRERO
Row4	BANCO ABANK	100	8	5,570	OCTUBRE, NOVIEMBRE
Row5	BANCO AGRICOLA	85.7	6	5,304.8	JUNIO, JULIO
Row6	BANCO CUSCATLAN	100	7	6,820.4	ENERO, FEBRERO, MAYO
Row7	BANCO DAVIVIENDA SALVADORE...	100	7	5,570	MARZO, ENERO
Row8	BANCO DE AMERICA CENTRAL	87.5	7	4,873.8	JUNIO, JULIO
Row9	BANCO PROMERICA	85.7	6	5,967.9	FEBRERO, SEPTIEMBRE
Row10	CLARO	100	9	5,570	JULIO, OCTUBRE, ENERO
Row11	DIGICEL	88.9	8	5,304.8	NOVIEMBRE, OCTUBRE
Row12	DISTRIBUIDORA DE ELECTRICID...	85.7	6	4,774.3	FEBRERO, MAYO
Row13	GRUPO MONGE	100	8	3,713.3	DICIEMBRE, ENERO
Row14	GRUPO Q	100	7	6,962.5	NOVIEMBRE, OCTUBRE
Row15	OMNISPORT	100	10	7,957.1	DICIEMBRE, JULIO, NOVIEMBRE, OCTUBRE, SEPTIEMBRE
Row16	SIMAN	85.7	6	5,304.8	OCTUBRE, SEPTIEMBRE
Row17	TELEFONICA	100	7	3,978.6	NOVIEMBRE, ENERO
Row18	TIGO	85.7	6	7,957.1	NOVIEMBRE, DICIEMBRE
Row19	UNICOMER	100	8	6,188.9	DICIEMBRE, ENERO, NOVIEMBRE

Figura 145 Resultados obtenidos para el caso C-ASO-01.

La Figura 145 muestra los resultados obtenidos mediante el flujo de trabajo detallado anteriormente las columnas que se observan de izquierda a derecha se explican a continuación:

- **Proveedor_resumido:** Corresponde al nombre del proveedor para los cuales se descubrieron reglas de asociación.
- **Ruleconfidence:** Representa la confianza con las que se obtuvo la regla de asociación valores cercanos a 100 representan mejores resultados.
- **ItemSetSupport:** Representa el soporte del set de ítems en otras palabras lo tan frecuentes que fueron los ítems del antecedente dentro del set de datos.
- **RuleLift:** Representa el aumento de la probabilidad de que ocurra el consecuente dado que ocurrió en el antecedente por lo tanto la probabilidad de que ocurra el consecuente por si solo es menor que la probabilidad antes de mencionada, los valores de lift con resultados aceptables deben ser mayores al 100%
- **AssociationRule:** Representa la regla de asociación para la cual el ultimo item del item set que conforma la regla corresponde al consecuente.

17.4.1.7 Interpretación de resultados

Se tiene una regla de asociación para cada proveedor de entre los cuales se generó por lo menos una regla de asociación, está corresponde a la de mayor tamaño del set de ítems, la confianza en todos los casos es mayor del 85% el soporte varía de 6-10, un ejemplo de regla de asociación es BANCO CUSCATLAN [ENERO, FEBRERO] -> [MAYO], para lo cual del 2005-2019 se repite 7 veces el item set antecedente (soporte) y para todo ellos siempre fue el mismo consecuente.

17.5 Técnica de clasificación

17.5.1 Clasificar la influencia que ha tenido la DC en las atenciones en base a la solución y montos recuperados

17.5.1.1 Contenido del caso.

N° de caso: C-CLA-01	
Técnica	Clasificación.
Algoritmos	Bayes Ingenuo, Arboles de decisión, Regresión logística, Redes neuronales
Población	Datos del modelo multidimensional SARA 2005-2019.
Variables	Se analiza los montos reclamados y recuperados, tomando únicamente casos cerrados en audiencia de gestión, avenimiento y conciliación.
Hipótesis	Clasificar los datos, en base a los montos reclamados y recuperados separado por influencia baja, mediana o alta.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos clasificados.
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión.
Herramienta de software	Knime Analytics Platform.

Tabla 66 Contenido del caso C-CLA-01.

17.5.1.2 Población

Consiste en las atenciones en casos cerrados en audiencia de gestión, avenimiento y conciliación, para poder determinar la influencia de la DC en cada una de las atenciones surgiendo las clases: baja, mediana y alta.

SQL Statement

```
SELECT a.monto_reclamado, a.monto_recuperado, b.nombre_solucion,
    CASE
    WHEN a.monto_recuperado > a.monto_reclamado THEN 'alta'
    WHEN a.monto_reclamado > a.monto_recuperado THEN 'baja'
    WHEN a.monto_reclamado = a.monto_recuperado THEN 'mediana'
    ELSE
        'mediana'
    END AS influencia
FROM fact_sara_montos a
JOIN dim_atencion b ON a.sk_atencion = b.sk_atencion
WHERE b.nombre_solucion IN
    ('Audiencia de Gestión', 'Avenimiento', 'Conciliación');
```

Preview results: Evaluate

Row ID	monto_...	monto_...	nombre_...	influencia
Row0	41	41	Avenimiento	mediana
Row1	1.5	0	Avenimiento	baja
Row2	65	61	Avenimiento	baja
Row3	3	3	Avenimiento	mediana
Row4	9	9	Avenimiento	mediana
Row5	31.65	0	Avenimiento	baja
Row6	9.43	0	Avenimiento	baja
Row7	9.43	9.43	Avenimiento	mediana
Row8	12	12	Avenimiento	mediana
Row9	9.43	9.43	Avenimiento	mediana

Figura 146 Set de datos para el caso C-CLA-01.

17.5.1.3 Variables

Las variables que se ven involucradas en la exploración son la solución de la atención, el monto reclamado y el monto recuperado.

17.5.1.4 Hipótesis

El objetivo de minería de datos que queremos ejecutar, o la hipótesis que queremos comprobar es la siguiente: “Realizar una clasificación de las atenciones por solución y determinar qué tipo de influencia baja, mediana o alta ejerce la Defensoría del Consumidor en los montos reclamados y recuperados”.

17.5.1.5 Procedimiento

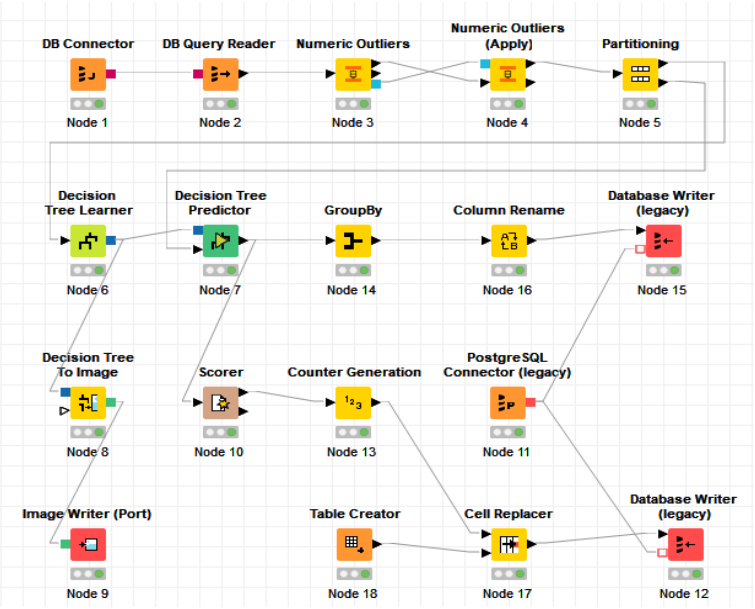


Figura 147 Flujo de trabajo aplicando arboles de decisión para el caso C-CLA-01.

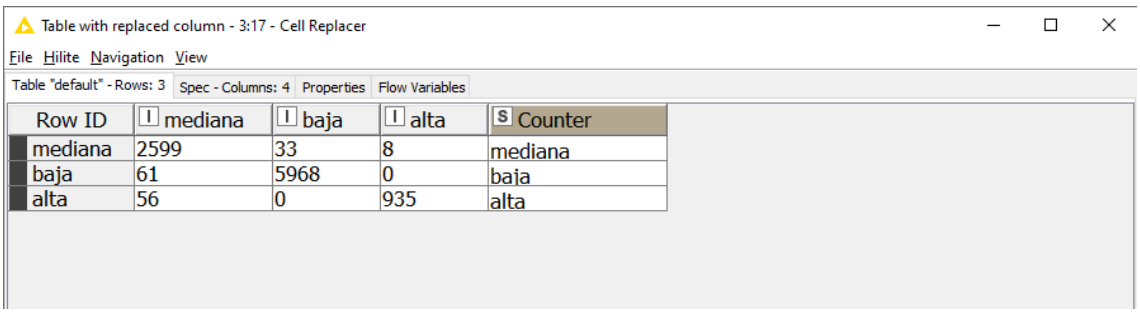
Descripción del procedimiento:

- **DB Connector**: Establece la conexión con la base de datos Postgres para acceder a los datos del modelo multidimensional.
- **DB Query Reader:** Obtiene los datos de los montos de las atenciones para ser exploradas con el algoritmo de minería de datos.
- **Numeric Outliers:** Lee el set de datos que se generó en la consulta y crea el modelo para eliminar de la población los registros con montos reclamados igual a cero**.**
- **Numeric Outliers (Apply):** Obtiene un set de datos y un modelo para poder eliminar para ese set de datos los montos igual a cero en los montos reclamados.
- **Partitioning:** Divide el set de datos en particiones para entrenamiento que tendrá un tamaño relativo del 91% del set, el 9% restante es para pruebas, estas se tomaran desde la parte superior no será aleatorio.
- **Decision Tree Learner:** Con el primer set de datos de la partición que corresponde a los datos de entrenamientos, se hace un modelo del árbol de decisión.

- **Decision Tree To Image:** Crea una imagen de la vista del árbol de decisión según al modelo en formato PNG.
- **Image Writer (Port):** Exporta una imagen en formato PNG de la vista del árbol de decisión.
- **Decision Tree Predictor:** Recibe el set de datos de prueba y haciendo uso del modelo creado a partir de los datos de entrenamiento, clasifica el set de pruebas en las diferentes clases como son influencia baja, mediana y alta.
- **Scorer:** Recibe el set de datos de prueba ya clasificado y compara las columnas para obtener una matriz de confusión para interpretar los resultados ya que en base a ello se puede obtener la presión del modelo.
- **Counter Generation:** Se agrega una nueva columna como contador a la matriz de confusión, esta será usada para representar las clases en la matriz de confusión.
- **Table Creator:** Crea un diccionario conteniendo las clases de la clasificación esta será utilizada para reemplazar los en la matriz de confusión.
- **Cell Replacer:** Reemplaza el contador por los datos obtenidos a partir del diccionario de datos y mostrar los nombres de las clases en la matriz de confusión.
- **PostgreSQL Connector (legacy):** Establece la conexión con la base de datos Postgres para poder escribir en la base de datos de minería.
- **Database Writer (legacy):** Crea y escribe una tabla para guardar los resultados obtenidos en la matriz de confusión.
- **GroupBy:** Se agrupan los datos del flujo de trabajo agregándose por la columna de solución sus respectivas clases y contando los resultados para mostrar una nueva columna con estos y así poder contabilizar los resultados, según la solución y la clase.}
- **Column Rename:** Cambia los nombres de las columnas prediction(influencia) a prediction y count(influencia) a count, para posteriormente ser guardadas en una base de datos.
- **Database Writer (legacy):** Crea y escribe una tabla para guardar los resultados de la clasificación de los datos de pruebas.

17.5.1.6 Resultados

Para partición de datos de entrenamiento y prueba del 90% y utilizando arboles de decisión se obtuvieron los resultados siguientes:



Row ID	mediana	baja	alta	Counter
mediana	2599	33	8	mediana
baja	61	5968	0	baja
alta	56	0	935	alta

Figura 148 Matriz de confusión para clasificación para el caso C-CLA-01.

Correct classified: 9,502	Wrong classified: 158
Accuracy: 98.364 %	Error: 1.636 %
Cohen's kappa (κ) 0.969	

Figura 149 Precisión de la clasificación para el caso C-CLA-01.

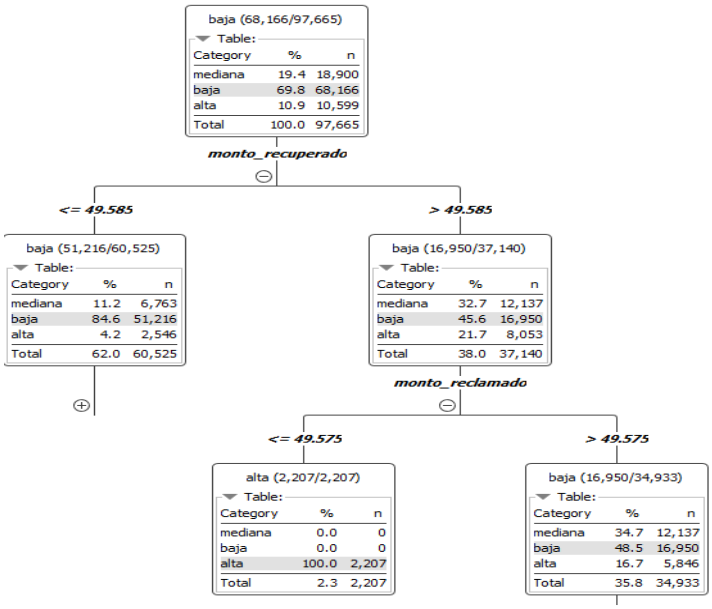


Figura 150 Árbol de clasificación para el caso C-CLA-01.

Renamed/Retyped table - 3:16 - Column Rename

File Hilite Navigation View

Table "default" - Rows: 7 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	nombre_solucion	prediction	count
Row0	Audiencia de Gestión	baja	45
Row1	Avenimiento	alta	684
Row2	Avenimiento	baja	5025
Row3	Avenimiento	mediana	2505
Row4	Conciliación	alta	259
Row5	Conciliación	baja	931
Row6	Conciliación	mediana	211

Figura 151 Resultados de la clasificación para el caso C-CLA-01.

17.5.1.7 Interpretación de resultados

De los algoritmos que se probaron el que obtuvo mejores resultados fue Arboles de Decisión con una precisión del 98.364%. Al analizar la matriz de confusión generada por el algoritmo de árboles de decisión se puede ver que presenta una clasificación distribuida que se logra visualizar entre las distintas clases que la componen siendo la mediana y bajas las más pobladas, pero logrando identificar algunas reglas distintas para cada una de las clases.

17.5.2 Identificar qué solución tendrán los casos recibidos en base a la edad y otros parámetros que se consideren relevantes de los consumidores

17.5.2.1 Contenido del caso

N° de caso: C-CLA-02	
Técnica	Clasificación.
Algoritmos	Bayes Ingenuo, Arboles de decisión, Regresión logística.
Población	Datos del modelo multidimensional SARA 2005-2019.
Variables	Se analiza el sexo, municipio, edad, motivo, y sector, tomando únicamente casos cerrados que no estén en Abiertos.
Hipótesis	Clasificar los datos, en base al sexo, municipio, edad, motivo, y sector discriminando por las diferentes soluciones que presenta un caso.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos clasificados.
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión y Coeficiente kappa.
Herramienta de software	Knime Analytics Platform y WEKA.

Tabla 67 Contenido del caso C-CLA-02.

17.5.2.2 Población

Consiste en las atenciones en donde estos sean casos cerrados, para poder determinar la solución en que cada una de las atenciones tendrá al finalizar el proceso.

SQL Statement

```
1  ECT b.nombre_solucion, c.genero_consumidor, b.id_municipio_consumidor,
2  ombre_motivo_resumido, EXTRACT(YEAR FROM age(e.fecha, d.fecha)) AS age,
3  ector_2017 AS sector
4  M fact_sara_atenciones_brindadas a
5  N dim_atencion b ON b.sk_atencion = a.sk_atencion
6  N dim_consumidor c ON c.sk_consumidor = a.sk_consumidor
7  N dim_tiempo d ON c.sk_tiempo = d.sk_tiempo
8  N dim_tiempo e ON a.sk_tiempo_ingreso = e.sk_tiempo
9  N dim_motivo f ON a.sk_motivo = f.sk_motivo
10 N dim_sector_categoria g ON a.sk_sector_categoria = g.sk_sector_categoria
11 N dim_tipo_caso h ON a.sk_tipo_caso = h.sk_tipo_caso
12 RE b.nombre_solucion <> 'Abierto' AND EXTRACT(YEAR FROM age(d.fecha)) > 18;
```

Figura 152 Set de datos para el caso C-CLA-02.

17.5.2.3 Variables

Las variables que se ven involucradas en la exploración son el sexo, municipio, edad, motivo, y sector.

17.5.2.4 Hipótesis

El objetivo de minería de datos que queremos ejecutar, o la hipótesis que queremos comprobar es la siguiente: “Realizar una clasificación de las atenciones por el sexo, municipio, edad, motivo, y sector y determinar qué tipo de solución que tendrá la atención al finalizar el proceso por parte de la Defensoría del Consumidor”.

17.5.2.5 Procedimiento configuración

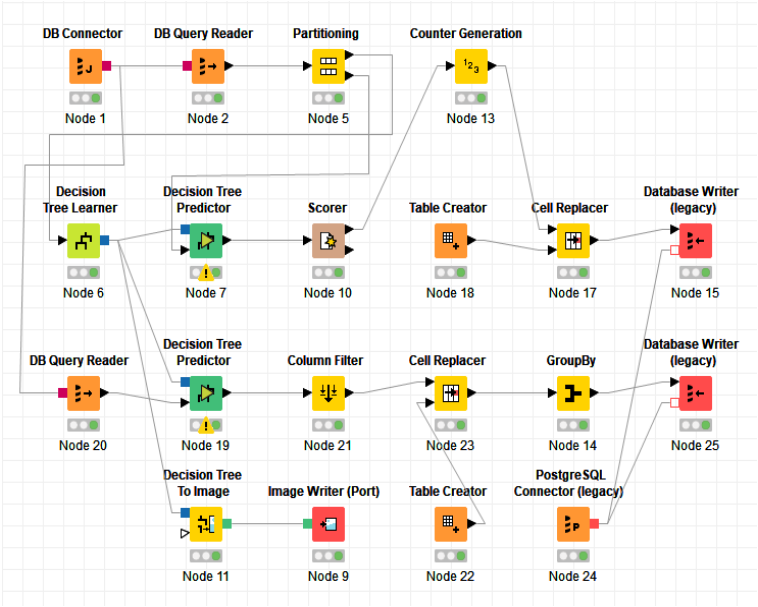


Figura 153 Flujo de trabajo aplicando arboles de decisión para el caso C-CLA-02.

La Figura 153 muestra el flujo de trabajo para el caso C-CLA-02 al igual a C-CLA-01 se utiliza arboles de decisión para clasificar el agregado en esta parte es que además de utilizar los datos de prueba, se introducen los datos que tienen como estado iniciado en el nodo Decision Tree Predictor.

17.5.2.6 Resultados

Confusion matrix - 4:10 - Scorer

File Hilite Navigation View

Table "spec_name" - Rows: 8 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Avenimiento	Conciliación	Otra	Tribunal Sancionador	Falta de Ratificación ...	Desistimiento	Audiencia de Gestión	Cerrado por razones de oficio
Avenimiento	10930	71	20	25	0	0	0	0
Conciliación	2105	35	1	19	0	0	0	0
Otra	80	0	1818	0	0	0	0	0
Tribunal Sancionador	786	45	1	30	0	0	0	0
Falta de Ratificación y Pre...	511	1	1	2	0	0	0	0
Desistimiento	630	8	0	5	0	0	0	0
Audiencia de Gestión	73	4	2	2	0	0	0	0
Cerrado por razones de ofi...	0	0	0	0	0	0	0	0

Figura 154 Matriz de confusión de Arboles de decisión para el caso C-CLA-02.

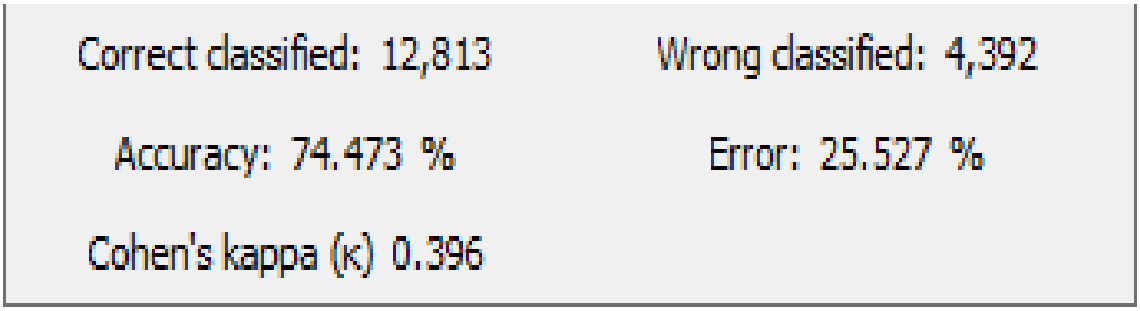


Figura 155 Indicadores de rendimiento de Arboles de decisión para el caso C-CLA-02.

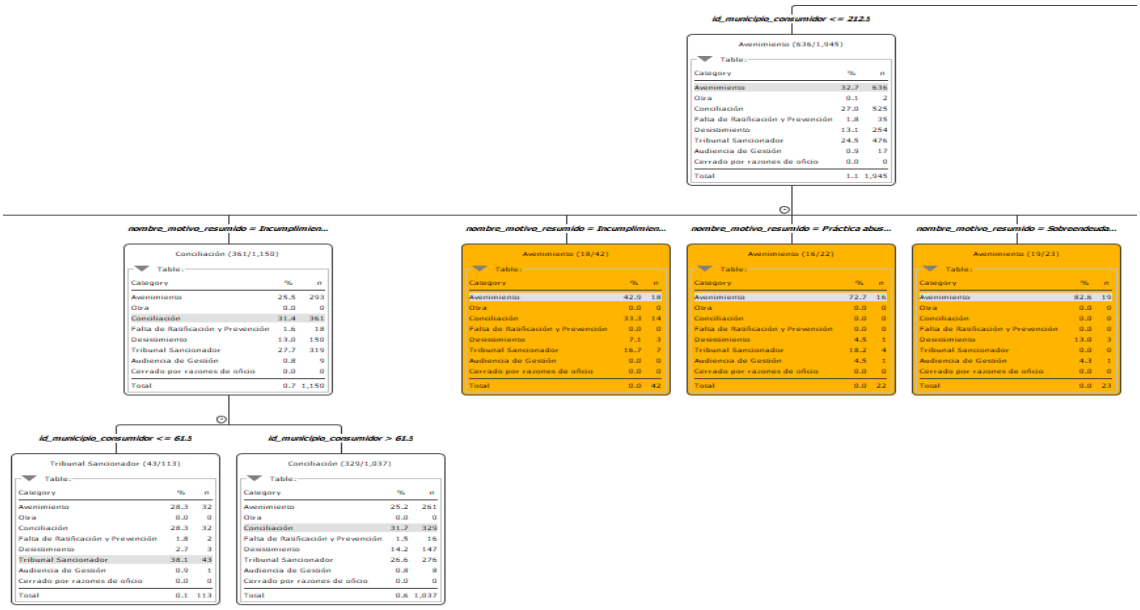


Figura 156 Árbol de decisión para el caso C-CLA-02.

table - 4:14 - GroupBy

e Navigation View

fault" - Rows: 93057 Spec - Columns: 6 Properties Flow Variables

ID	genero_consumidor	id_municipio_consumidor	nombre_motivo_resumido	age	sector	Prediction (no
0	0	Apaneca	Cobros, cargos y comisiones indebidas	74	Agua potable	Avenimiento
1	0	Apaneca	Incumplimiento de garantía	47	Vehículos	Avenimiento
2	0	Armenia	Cobros, cargos y comisiones indebidas	56	Agua potable	Avenimiento
3	0	California	Mala calidad del producto o servicio	10	Equipo informático	Avenimiento
4	0	Caluco	Cobros, cargos y comisiones indebidas	14	Telecomunicaciones	Avenimiento
5	0	Chiltiupán	Cobros, cargos y comisiones indebidas	13	Telecomunicaciones	Avenimiento
6	0	Chiltiupán	Cobros, cargos y comisiones indebidas	14	Tributos y servicios	Avenimiento
7	0	Chiltiupán	Cobros, cargos y comisiones indebidas	15	Agua potable	Avenimiento
8	0	Chiltiupán	Cobros, cargos y comisiones indebidas	18	Agua potable	Avenimiento
9	0	Chiltiupán	Cobros, cargos y comisiones indebidas	18	Vehículos	Avenimiento
10	0	Chiltiupán	Cobros, cargos y comisiones indebidas	20	Agua potable	Avenimiento
11	0	Chiltiupán	Cobros, cargos y comisiones indebidas	31	Agua potable	Avenimiento
12	0	Chiltiupán	Cobros, cargos y comisiones indebidas	33	Servicios	Avenimiento
13	0	Chiltiupán	Cobros, cargos y comisiones indebidas	40	Agua potable	Avenimiento
14	0	Chiltiupán	Cobros, cargos y comisiones indebidas	41	Agua potable	Avenimiento
15	0	Chiltiupán	Cobros, cargos y comisiones indebidas	44	Telecomunicaciones	Avenimiento
16	0	Chiltiupán	Cobros, cargos y comisiones indebidas	47	Créditos	Avenimiento
17	0	Chiltiupán	Cobros, cargos y comisiones indebidas	57	Telecomunicaciones	Avenimiento
18	0	Chiltiupán	Incumplimiento de contrato u oferta	16	Inmuebles	Avenimiento
19	0	Chiltiupán	Incumplimiento de contrato u oferta	16	Servicios	Avenimiento
20	0	Chiltiupán	Incumplimiento de contrato u oferta	16	Telecomunicaciones	Avenimiento
21	0	Chiltiupán	Incumplimiento de contrato u oferta	17	Inmuebles	Avenimiento
22	0	Chiltiupán	Incumplimiento de garantía	15	Equipo informático	Avenimiento
23	0	Chiltiupán	Incumplimiento de garantía	30	Equipo informático	Avenimiento

Figura 157 Resultados de la clasificación para el caso C-CLA-02.

17.5.2.7 Interpretación de resultados

De los algoritmos que se probaron el que obtuvo mejores resultados fue Arboles de Decisión con una precisión del 74.473% los otros se descartaron por un coeficiente menor a 39.6% que fue el de Arboles de Decisión esto nos indica la coincidencia de la predicción con la clase real y se determina que la hipótesis no generará ningún valor al negocio por la baja coincidencia que tendrá con los datos reales. Al analizar la matriz de confusión se puede ver que la clasificación que presentan es debido a la probabilidad y no una distribución, se aprecia visualizando las clases avenimiento y conciliación, en estas se concentra la clasificación, pero se obvia clases que en los datos reales presentan una población menor.

17.5.3 Clasificar el comportamiento de los proveedores en base los montos reclamados, montos recuperados y solución

17.5.3.1 Contenido del caso

N° de caso: C-CLA-03	
Técnica	Clasificación.
Algoritmos	Arboles de decisión
Población	Datos del modelo multidimensional SARA 2005-2019.
Variables	Se analizan los proveedores, los montos reclamados y recuperados, tomando únicamente los casos con solución audiencia de gestión, avenimiento y conciliación.
Hipótesis	Clasificar el comportamiento de los proveedores en base los montos reclamados, montos recuperados y solución.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos clasificados.
Interpretación de resultados	Mediante indicadores de rendimiento precisión y matriz de confusión.
Herramienta de software	Knime Analytics Platform.

Tabla 68 Contenido del caso C-CLA-03.

17.5.3.2 Población

SQL Statement

```
SELECT * FROM
(SELECT
t.anio,
f.origen,
f.sk_atencion,
CASE WHEN p.proveedor_resumido = 'No Disponible' THEN p.proveedor ELSE p.proveedor_resumido END proveedor_resumido,
COALESCE(f.monto_reclamado,0)/COALESCE(f.monto_reclamado,0) monto_reclamadoP,
COALESCE(a.monto_recuperado_corregido,0)/COALESCE(f.monto_reclamado,0) monto_recuperadoP,
a.nombre_solucion,
CASE
    WHEN nombre_solucion IN('Avenimiento','Audiencia de Gestión') AND COALESCE(monto_recuperado_corregido,0) BETWEEN 1 AND COALESCE(f.m
    WHEN  nombre_solucion IN('Avenimiento','Audiencia de Gestión') AND COALESCE(monto_recuperado_corregido,0) BETWEEN COALESCE(f.monto_
    WHEN  nombre_solucion IN('Avenimiento','Audiencia de Gestión') AND COALESCE(monto_recuperado_corregido,0) > COALESCE(f.monto_reclam

    WHEN nombre_solucion IN('Conciliación') AND COALESCE(monto_recuperado_corregido,0) BETWEEN 1 AND COALESCE(f.monto_reclamado,0)*0.33
    WHEN  nombre_solucion IN('Conciliación') AND COALESCE(monto_recuperado_corregido,0) BETWEEN COALESCE(f.monto_reclamado,0)*0.33 AND
    WHEN  nombre_solucion IN('Conciliación') AND COALESCE(monto_recuperado_corregido,0) > COALESCE(f.monto_reclamado,0)*0.66 THEN 'Menc
    ELSE 'No sujeto'
    END actuar_proveedor
FROM fact_sara_montos f
INNER JOIN dim_atencion a ON a.sk_atencion = f.sk_atencion
INNER JOIN dim_proveedor p ON p.sk_proveedor = f.sk_proveedor
INNER JOIN(

SELECT DISTINCT proveedor_resumido FROM(
select CASE WHEN p.proveedor_resumido = 'No Disponible' THEN p.proveedor ELSE p.proveedor_resumido END proveedor_resumido,
count(*) atenciones,t.anio
        from fact_sara_montos a
        left join dim_proveedor p   on p.sk_proveedor = a.sk_proveedor
        left join dim_tiempo t ON t.sk_tiempo = a.sk_tiempo_ingreso
```

Preview results: Evaluate

Row ID	anio	origen	sk_atencion	proveedor_resumido	monto_reclamadop	monto_recuperadop	nombre_solucion	actuar_proveedor
Row0	2007	Re of	655503	Banco Agrícola	1	1	Avenimiento	Ágil y muy dispuesto
Row1	2007	Re of	655504	Banco Cuscatlán	1	1	Avenimiento	Ágil y muy dispuesto
Row2	2007	Re of	655516	ANDA	1	0.461	Avenimiento	Ágil y dispuesto
Row3	2007	Re of	655528	ANDA	1	0.562	Avenimiento	Ágil y dispuesto
Row4	2007	Re of	655497	Unicomer	1	0.251	Avenimiento	Ágil y poco dispuesto
Row5	2007	Re of	655501	Banco Agrícola	1	1	Avenimiento	Ágil y muy dispuesto
Row6	2007	Re of	655515	Claro	1	1	Avenimiento	Ágil y muy dispuesto
Row7	2007	Re of	655517	ANDA	1	0.751	Avenimiento	Ágil y muy dispuesto
Row8	2007	Re of	655519	ANDA	1	0.239	Avenimiento	Ágil y poco dispuesto
Row9	2007	Re of	655507	Claro	1	0.193	Avenimiento	Ágil y poco dispuesto

Figura 158 Set de datos para el caso C-CLA-03.

La Figura 158 muestra la consulta para obtener el set de datos el cual corresponde al actuar de los proveedores en relación a la solución y los montos recuperados y reclamados de las denuncias interpuestas en su contra, para los proveedores con más de 70 denuncias anuales, se utiliza el campo derivado proveedor_resumido.

17.5.3.3 Variables

Las variables que se ven involucradas en la exploración son Solución, Monto reclamado, Monto recuperado y Proveedor resumido. Para el monto reclamado este análisis se toma como el 100%. Y para monto recuperado si se encuentra entre 0 y 33% el actuar del proveedor fue poco dispuesto, si se encuentra entre el 33% y el 66% el actuar del proveedor fue dispuesto y si sobrepasa el 66% fue muy dispuesto.

17.5.3.4 Hipótesis

El objetivo de minería de datos o hipótesis a llevar a cabo en esta historia de usuario es: “Clasificar a el comportamiento de los proveedores en base los montos reclamados, montos recuperados y solución”.

17.5.3.5 Procedimiento

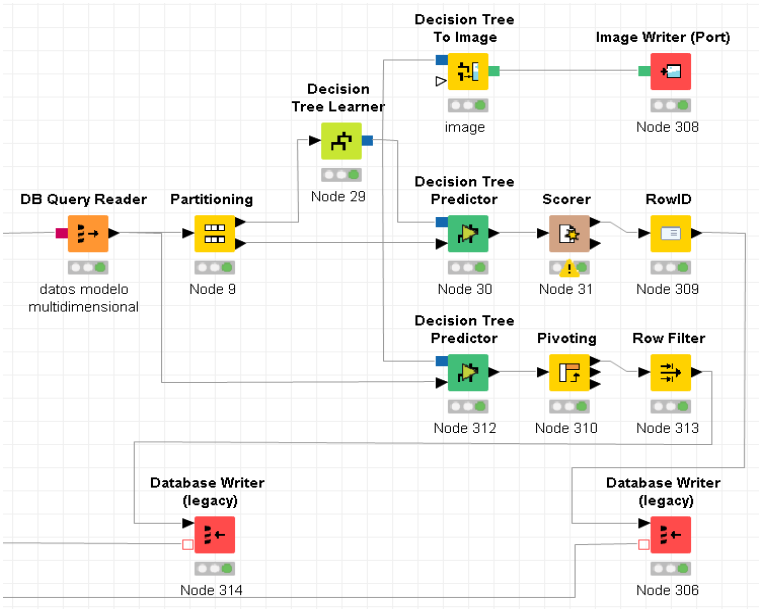


Figura 159 Flujo de trabajo aplicando Árboles de decisión para el caso C-CLA-03.

La Figura 159 muestra el flujo de trabajo para el caso C-CLA-03 al igual a C-CLA-01 se utiliza arboles de decisión para clasificar utilizando el mismo procedimiento para generar el modelo y obtener la información.

17.5.3.6 Resultados

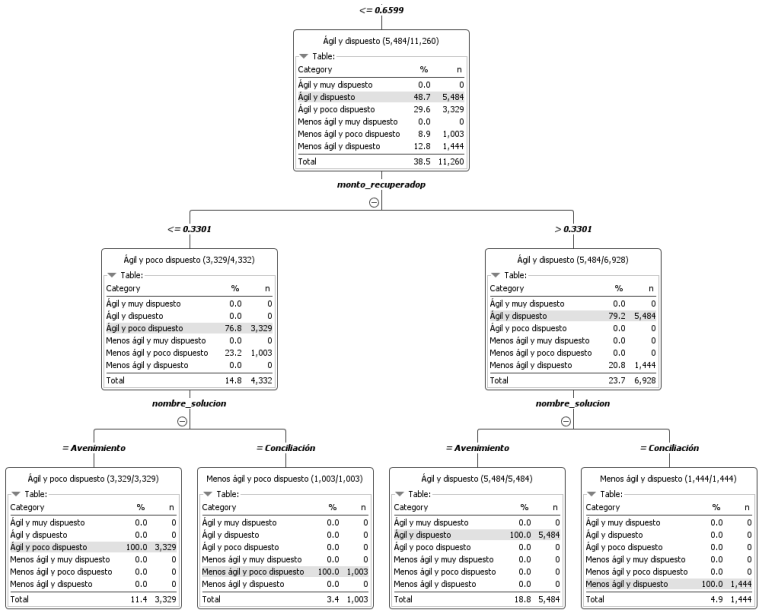
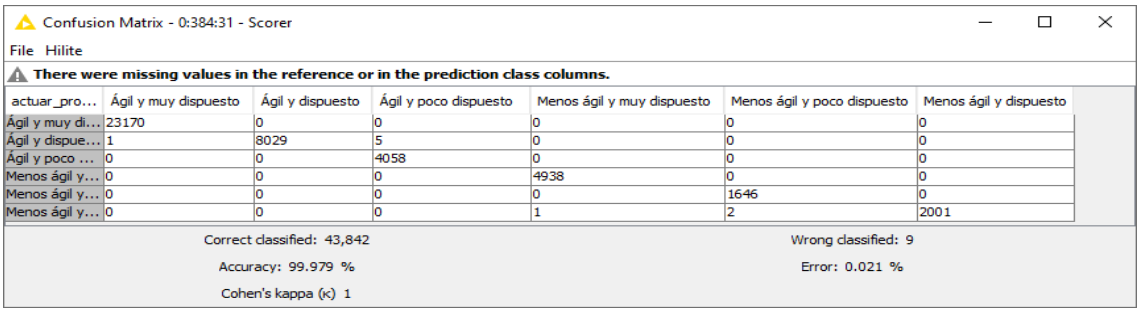


Figura 160 Fragmento del árbol de decisión para el caso C-CLA-3.

La Figura 160 muestra un fragmento del árbol de decisión generado el primer nodo corresponde al nodo raíz de esa rama en el cual se observa los posibles caminos y la

clasificación correspondiente en base a los dos criterios que el algoritmo determino como significativos los cuales son la solución y el monto recuperado en términos porcentuales.


Confusion Matrix - 0:384:31 - Scorer

File Hilite

There were missing values in the reference or in the prediction class columns.

actuar_pro...	Ágil y muy dispuesto	Ágil y dispuesto	Ágil y poco dispuesto	Menos ágil y muy dispuesto	Menos ágil y poco dispuesto	Menos ágil y dispuesto
Ágil y muy di...	23170	0	0	0	0	0
Ágil y dispue...	1	8029	5	0	0	0
Ágil y poco ...	0	0	4058	0	0	0
Menos ágil y...	0	0	0	4938	0	0
Menos ágil y...	0	0	0	0	1646	0
Menos ágil y...	0	0	0	1	2	2001

Correct classified: 43,842 Wrong classified: 9

Accuracy: 99.979 % Error: 0.021 %

Cohen's kappa (κ) 1

Figura 161 Indicadores de rendimiento para el caso C-CLA-03.

proveedor_resumido	Menos ...	Menos ...	Menos ...	Ágil y di...	Ágil y m...	Ágil y p...
ANDA	2138	4232	1603	11204	20178	5205
Aes CLESA	?	1	?	2	75	3
Banco Abank	44	157	57	38	284	51
Banco Agrícola	55	147	17	89	909	51
Banco Cuscatlán	40	261	47	72	873	81
Banco Davivienda Salvadoreño	45	76	24	98	421	65
Banco De América Central	18	80	15	27	502	28
Banco Promerica	10	23	5	11	194	17
Cedecel	7	36	2	7	131	8
Claro	416	1084	217	701	2491	439
Club De Playas Salinitas	56	86	14	33	312	23
Compramerica, S.A. De C.V.	?	?	?	1	86	?
Digicel	40	132	44	44	523	45
Editorial Oceano De El Salvador	1	7	5	22	148	11
Grupo Monge	64	541	126	134	2056	333
Jinsal S.A. de C.V.	1	8	8	2	134	16
Natural Learning Corporation...	6	15	1	10	75	6
Omnisport	59	346	129	41	643	101
Operadora Del Sur	2	14	1	10	415	12
Salazar R. S. A. De C. V.	62	114	42	61	178	30
Scotiabank El Salvador	6	26	8	12	214	13
Siman	6	41	6	14	388	21
Sky	9	27	15	40	93	43
Sovipe Comercial	9	86	27	10	330	22
Telefónica	79	221	55	75	610	76
Tigo	164	711	91	638	3492	526
Todoticket	1	2	?	?	61	1
Unicomer	27	132	55	74	1182	137

Figura 162 Contadores de clase por proveedor para caso C-CLA-03.

La Figura 162 muestra los contadores de cada una de las clases para cada proveedor, mostrando únicamente resultados para los proveedores que tienen un número considerable de reclamos en su contra.

17.5.3.7 Interpretación de resultados

Para el caso de los indicadores de rendimiento la exactitud del 99% arrojada por el algoritmo da mucha certeza sobre las clases asignadas para el actuar del proveedor:

- **Ágil y poco dispuesto:** Fue de forma ágil porque se brindó respuesta favorable al consumidor en etapas tempranas del proceso y fue poco dispuesto pues la pretensión del consumidor se cumplió en un porcentaje menor.
- **Ágil y dispuesto:** Fue de forma ágil porque brindó respuesta favorable al consumidor en etapas tempranas del proceso y fue dispuesto pues la pretensión del consumidor se cumplió en un porcentaje aceptable.
- **Ágil y muy dispuesto:** Fue de forma ágil porque brindó respuesta favorable al consumidor en etapas tempranas del proceso y fue muy dispuesto pues la pretensión del consumidor se cumplió en gran medida un porcentaje cercano o mayor a su pretensión.
- **Menos ágil y poco dispuesto:** Fue de forma menos ágil porque brindó respuesta favorable al consumidor en la etapa de conciliación y fue poco dispuesto pues la pretensión del consumidor se cumplió en un porcentaje menor.
- **Menos ágil y dispuesto:** Fue de forma menos ágil porque brindó respuesta favorable al consumidor en la etapa de conciliación y fue dispuesto pues la pretensión del consumidor se cumplió en un porcentaje aceptable.
- **Menos ágil y muy dispuesto:** Fue de forma menos ágil porque brindó respuesta favorable al consumidor en la etapa de conciliación y fue muy dispuesto pues la pretensión del consumidor se cumplió en gran medida un porcentaje cercano o mayor a su pretensión.

En el caso de los datos de prueba solo en 9 ocasiones el algoritmo asigno una clase que no le correspondía, teniendo un error del 0.021%.

En términos generales la cantidad de atenciones clasificadas en actuar del proveedor “Ágil y muy dispuesto” es mayor que las del resto de clasificaciones, sin embargo, hay algunos proveedores en los cuales la respuesta hacia el consumidor ha sido “Ágil y poco dispuesto” como lo es el caso de ANDA.

17.6 Técnica de pronóstico (forecast)

17.6.1 Pronosticar casos a recibir en fechas futuras

17.6.1.1 Contenido del caso.

N° de caso: C-FOR-01	
Técnica	Pronostico.
Algoritmos	Series temporales (ARIMA)
Población	Datos del modelo multidimensional SARA 2005-2019.

Variables	Se analizan los casos recibidos por tipo de caso a lo largo del tiempo.
Hipótesis	Pronosticar atenciones a recibir por tipo caso.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos pronosticados.
Interpretación de resultados	Mediante indicadores de rendimiento.
Herramienta de software	Knime Analytics Platform.

Tabla 69 Contenido del caso C-FOR-01.

17.6.1.2 Población

```
SELECT
concat(t.anio::varchar, t.semana::varchar) semana,
t.fecha,
t.anio,
tc.sk_tipo_caso,
CASE WHEN tipo_caso_agrupado = 'Asesoria' THEN 'Asesoría'
ELSE tipo_caso_agrupado END tipo_caso_agrupado,
o.nombre_oficina,
count(*) atenciones
FROM fact_sara_atenciones_brindadas a
LEFT Join dim_tipo_caso tc ON tc.sk_tipo_caso = a.sk_tipo_caso
LEFT JOIN dim_tiempo t ON t.sk_tiempo = a.sk_tiempo_ingreso
LEFT JOIN dim_oficina o ON o.sk_oficina = a.sk_oficina
GROUP BY
    concat(t.anio::varchar, t.semana::varchar),
    t.fecha,
    t.anio,
    tc.sk_tipo_caso,
    CASE WHEN tipo_caso_agrupado = 'Asesoria' THEN 'Asesoría' ELSE tipo_caso_agrupado END,
    o.nboombre_oficina
    ORDER BY anio ASC
```

ew results: Evaluat

Row ID	semana	fecha	anio	sk_tipo...	tipo_ca...	nombre...	atencio...
ow0	20053	2005-01-10	2005	3	Derivación	San Miguel	6
ow1	20053	2005-01-10	2005	3	Derivación	San Salvador	9
ow2	20053	2005-01-10	2005	3	Derivación	Santa Ana	34
ow3	20053	2005-01-10	2005	5	Gestión	San Miguel	1
ow4	20053	2005-01-10	2005	5	Gestión	San Salvador	42
ow5	20053	2005-01-10	2005	14	Asesoría	San Miguel	100
ow6	20053	2005-01-10	2005	14	Asesoría	San Salvador	282
ow7	20053	2005-01-10	2005	14	Asesoría	Santa Ana	125
ow8	20053	2005-01-10	2005	18	Denuncia	San Miguel	93
ow9	20053	2005-01-10	2005	18	Denuncia	San Salvador	349

Figura 163 Set de datos para el caso C-FOR-01.

La Figura 163 muestra la consulta para obtener el set de datos a utilizar la cual corresponde a las atenciones agrupadas por tipo de caso, oficina, semana fecha y año.

17.6.1.3 Variables

Las variables a utilizar se detallan a continuación:

- Semana: Como agrupador de cuantas atenciones se han brindado por semana.
- Fecha: Como agrupador de cuantas atenciones se han brindado en la fecha.
- Año: Como agrupador de cuantas atenciones se han brindado en el año.
- Tipo de caso: Corresponde a la categorización de las atenciones que los casos que se atienden los cuales corresponden a Denuncia, Asesoría, Gestión y Derivación.
- Oficina: Corresponde a los lugares desde los cuales se puede brindar una atención.
- Atenciones: Es el contador de las atenciones brindadas por los agrupadores.

17.6.1.4 Hipótesis

Pronosticar casos a recibir en fechas futuras, para poder reorientar recursos si fuese necesario.

17.6.1.5 Procedimiento

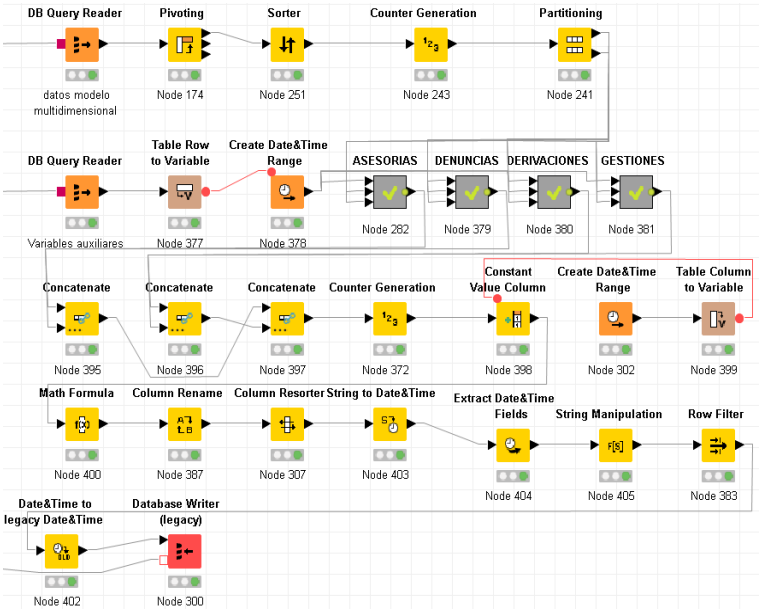


Figura 164 Flujo de trabajo macro para el caso C-FOR-01.

La Figura 164 muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-FOR-01 se detalla a continuación las acciones que se llevan a cabo:

- **DB Query Reader (datos modelo multidimensional):** Se ejecuta la consulta de selección de los datos.
- **Pivoting (Node 174):** Se toma como columna de agrupador la semana, como columna pivote el tipo de caso y como funciones de agregación First (Fecha y Año) y Sum (Atenciones) respectivamente.
- **Sorter (Node 251):** Se ordenan los datos por fecha ascendente.
- **Counter Generation (Node 243):** Se enumeran los datos del 1 al n.
- **Partitioning (Node 241):** Se particionan los datos en 75% entrenamiento y 25% pruebas.
- **DB Query Reader (Variables auxiliares):** Se ejecuta la consulta para obtener los siguientes valores: dia_inicio, dia_fin, cantidad_dias. anio_antes y anio_actual.
- **Table Row to Variable (Node 377):** Convierte los valores obtenidos en el nodo anterior en variables.
- **Create Date&Time Range (Node 378):** Crea un rango de fechas futuras tomando como día de inicio la variable dia_inicio, fecha final la variable dia_fin y cantidad de registros la variable cantidad_dias.
- **ASESORIAS (Node 282):** Metanodo que pronostica las atenciones de tipo Asesoría a recibir, recibe 3 parámetros La partición de datos de entrenamiento, la partición de datos de pruebas y el rango de fechas creado.
- **DENUNCIAS (Node 379):** Metanodo que pronostica las atenciones de tipo Denuncias a recibir, recibe 3 parámetros La partición de datos de entrenamiento, la partición de datos de pruebas y el rango de fechas creado.

- **DERIVACIONES (Node 380):** Metanodo que pronostica las atenciones de tipo Derivación a recibir, recibe 3 parámetros La partición de datos de entrenamiento, la partición de datos de pruebas y el rango de fechas creado.
- **GESTIONES (Node 381):** Metanodo que pronostica las atenciones de tipo Gestión a recibir, recibe 3 parámetros La partición de datos de entrenamiento, la partición de datos de pruebas y el rango de fechas creado.
- **Concatenate (Node 395):** Realiza la concatenación entre los dataset resultantes de los metanodos ASESORIAS y DENUNCIAS.
- **Concatenate (Node 396):** Realiza la concatenación entre los dataset resultantes de los metanodos DERIVACIONES y GESTIONES.
- **Concatenate (Node 397):** Realiza la concatenación de los dataset resultantes de los nodos 395 y 396.
- **Counter Generation (Node 372):** Enumera los datos del dataset del 1 al n.
- **Create Date&Time Range (Node 302):** Extrae la fecha de ejecución.
- **Table Column to Variable (Node 399):** Convierte la fecha de ejecución en una variable.
  Constant Value Column (Node 398): Para la variable que contiene la fecha de ejecución crea en el dataset una columna con ese valor constante.
- **Math Formula (Node 400):** Aproxima los datos pronosticados a valores enteros.
- **Column Rename (Node 306):** Se renombra la columna contadora para utilizarse como Id.
- **Column Resorter (Node 307):** Se ordenan las columnas del dataset antes de escribir en la base de datos.
- **String to Date&Time (Node 403):** Convierte la columna de la fecha de creación a Date&Time.
- **Extract Date&Time Fields (Node 404):** Extrae el año y el nombre del mes a partir del campo fecha.
- **String Manipulation (Node 405):** Concatena el nombre del mes con el año y lo coloca en el campo del nombre del mes.
- **Row Filter (Node 383):** Filtra los registros colocando como inicio y fin las variables anio_antes y anio_actual para el campo Year extraído anteriormente.
- **Date&Time to legacy Date&Time:** Convierte el campo fecha de Local Date a Date antes de almacenar en la base de datos.
- **DataBase Writer (Node 300):** Se escribe en la base de datos mining_midas en la tabla sara_forecastrecepcionadas los resultados de la minería de datos resultantes de este flujo de trabajo.

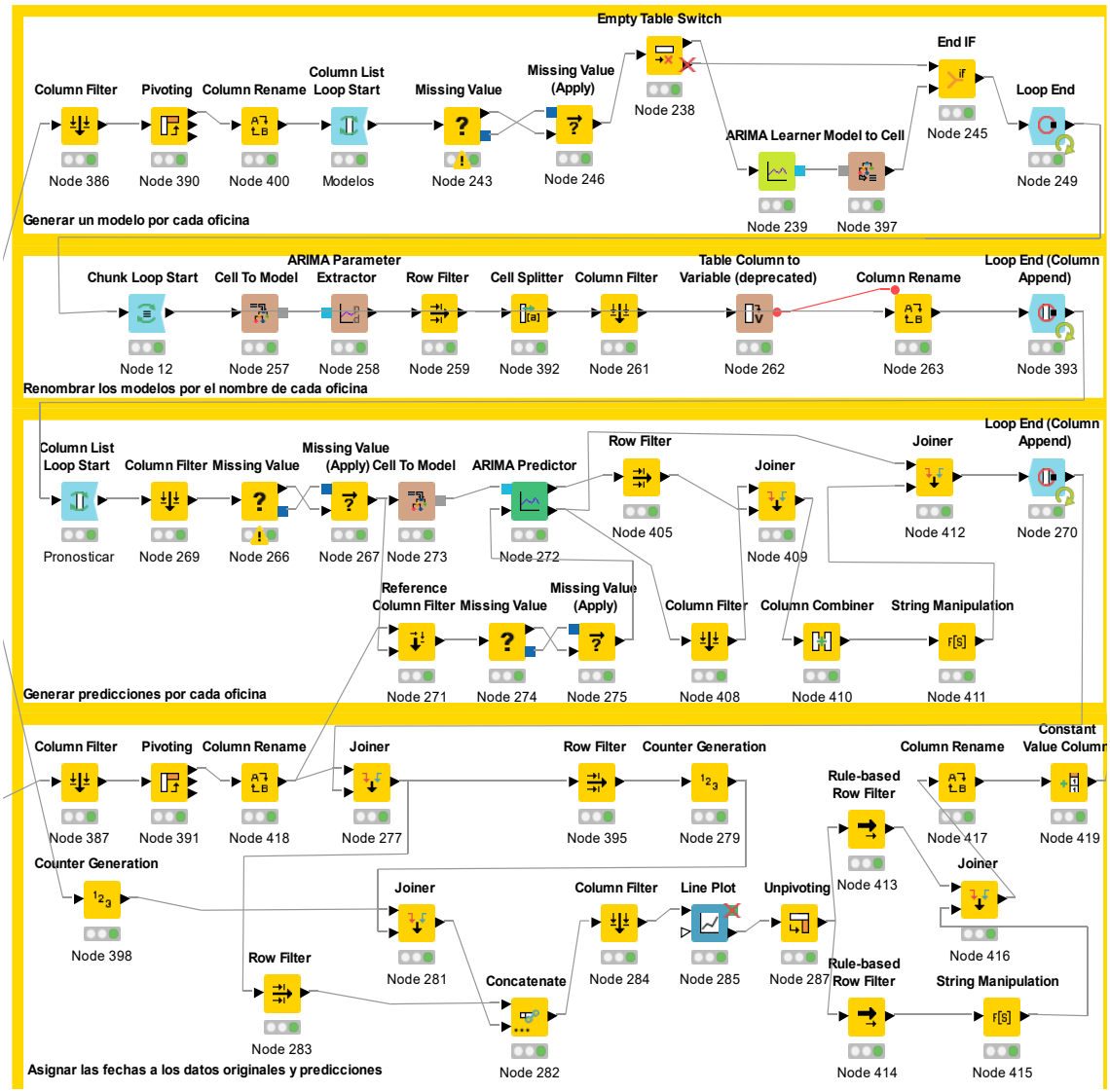


Figura 165 Flujo de trabajo metanodo DENUNCIAS aplicando ARIMA para el caso C-FOR-01.

La Figura 165 muestra el flujo de trabajo desarrollado para el metanodo de DENUNCIAS se detalla a continuación las acciones que se llevan a cabo:

- **Column Filter (Node 386):** Se filtran las columnas fecha, oficina y cantidad de denuncias.
- **Pivoting (Node 390):** Se toma como columna pivote la oficina, como grupo la fecha y como función de agregación “Sum” para la cantidad de Denuncias.
- **Column Rename (Node 400):** Se renombran las columnas correspondientes a las 5 oficinas retirando complementos generados por el pivoting.
- **Column List Loop Start (Modelos):** Se inicia un loop tomando como iteraciones cada una de las columnas de las oficinas.
- **Missing Value (Node 243):** Se toma el criterio de retirar del dataset los registros que tengan valores faltantes en los tipos de dato “Long”.

- **Missing Value Apply (Node 246):** Se aplica el criterio de valores faltantes del nodo anterior.
- **Empty Table Switch (Node 238):** Inicia decisión para el criterio de tabla vacía en la iteración actual.
- **ARIMA Learner (Node 239):** Genera el modelo de arima en la iteración actual con los datos de entrenamiento.
- **Modelo to Cell (Node 397):** Convierte el modelo generado en una celda de la tabla.
- **End IF (Node 245):** Finaliza decisión para el criterio de tabla vacía si es verdadero pasa al siguiente nodo.
- **Loop End (Node 249):** Pasa a la siguiente iteración hasta finalizar el Loop.
- **Chunk Loop Start (Node 12):** Inicia nuevo Loop donde las iteraciones corresponden a los modelos generados anteriormente.
- **Cell to Model (Node 257):** Convierte la celda de la iteración actual en modelo.
- **ARIMA Parameter Extractor (Node 258):** Extrae los parámetros del modelo ARIMA.
- **Row Filter (Node 259):** Filtra la fila que contiene el nombre de la columna con la que generó el modelo.
- **Cell Splitter (Node 392):** Realiza un corte en la cadena de caracteres para separar el nombre de la columna que generó el modelo.
- **Column Filter (Node 261):** Filtra la columna que contiene únicamente el nombre de la columna que generó el modelo.
- **Table Column to Variable (Node 262):** Convierte el nombre de la columna que generó el modelo en una variable.
- **Column Rename (Node 263):** Renombra el modelo con el nombre de la columna que lo originó.
- **Loop End (Column Append) (Node 393):** Se pasa a la siguiente iteración hasta terminar el bucle, se colectan los resultados.
- **Column List Loop Start (Pronosticar):** Se inicia un bucle donde los modelos identificados por los nombres de cada una de las oficinas corresponden a las iteraciones.
- **Column Filter (Node 269):** Se filtra la columna para la cual en la iteración actual contiene el modelo.
- **Missing Value (Node 266):** Se define el criterio de retirar las filas que contengan vacío en el modelo.
- **Missing Value Apply (Node 267):** Se aplica el criterio de valores faltantes.
- **Cell to Model (Node 273):** Se convierte la celda en modelo.
- **Column Filter (Node 387):** Se filtra para los datos de pruebas las columnas fecha, oficina y cantidad de denuncias.
- **Pivoting (Node 391):** Se toma como columna pivote la oficina, como grupo la fecha y como función de agregación “Sum” para la cantidad de Denuncias.
- **Column Rename (Node 418):** Se renombran las columnas correspondientes a las 5 oficinas retirando complementos generados por el pivoting.
- **Reference Column Filter (Node 271):** Se filtran los datos de entrenamiento para la oficina de la iteración actual pasando como referencia el nombre del modelo.
- **Missing Value (Node 274):** Se define el criterio de colocar cero en la cantidad de atenciones para las filas que carezcan de valor en dicha columna.

- **Missing Value Apply (Node 275):** Se aplica el criterio de valores faltantes.
- **ARIMA Predictor (Node 272):** Se pronostican las denuncias a recibir en 8 periodos (El algoritmo asumo que los datos de entrenamiento y pruebas se encuentran en períodos igualmente espaciados).
- **Row Filter (Node 405):** Se filtran únicamente las filas correspondientes a los 8 periodos pronosticados.
- **Column Filter (Node 408):** Se filtran las filas que contienen el pronóstico de los datos de prueba para la cantidad de denuncias recibidas.
- **Joiner (Node 409):** Se unen los datos resultantes de los nodos 405 y 408 obteniéndose 2 columnas.
- **Column Combiner (Node 410):** Se combinan las 2 columnas del nodo anterior definiendo como separador el carácter coma.
- **String Manipulation (Node 411):** Se retiran los caracteres "?" y "," el primero correspondiente a valores faltantes y el segundo al separador para la columna resultante del nodo 410.
- **Joiner (Node 412):** Se unen las columnas correspondientes a los datos originales de pruebas y los datos pronosticados.
- **Loop End (Column Append) (Node 270):** Se pasa a la siguiente iteración o se finaliza el bucle, se recolectan los resultados.
- **Joiner (Node 277):** Se obtiene la fecha para los datos originales de prueba procedentes del nodo 418.
- **Row Filter (Node 283):** Se obtienen las filas únicamente de los datos de prueba originales y pronosticados.
- **Row Filter (Node 395):** Se filtran las filas para los 8 periodos pronosticados en las 5 oficinas.
- **Counter Generation (Node 279):** Se enumeran los registros con espaciamiento de 7 correspondiente a los 7 días de una semana.
- **Counter Generation (Node 398):** Se enumeran las fechas de la serie de tiempo que el metanodo recibe de parámetro del macronodo.
- **Joiner (Node 281):** Se unen los datos resultantes de los nodos 398 y 279 teniendo ya una fecha en concreto para los 8 periodos pronosticados.
- **Concatenate (Node 282):** Se concatenan los data set resultantes de los nodos 281 y 283.
- **Column Filter (Node 284):** Se filtran las columnas correspondientes a los datos originales y pronosticados para las 5 oficinas y la fecha a la que corresponden.
- **Line Plot (Node 285):** Se previsualizan en un gráfico de líneas los resultados.
- **Unpivoting (Node 287):** Se transforman los datos de las columnas de las 5 oficinas y sus predicciones a filas.
- **Rule-based Row Filter (Node 413):** Se filtran los registros únicamente para los datos originales correspondientes a las oficinas.
- **Rule-based Row Filter (Node 414):** Se filtran los registros únicamente para los datos pronosticados correspondientes a las oficinas
- **String Manipulation (Node 415):** Se retira la cadena de caracteres del nombre de la oficina el complemento "(#1)" que denota a un dato pronosticado en este caso.
- **Joiner (Node 416):** Se unen los dataset resultantes de los nodos 413 y 415.

- **Column Rename (Node 417):** Se renombran las columnas a “Oficina”, “Original”, “Pronóstico” y “Fecha”.
- **Constant Value Column (Node 419):** Se crea una nueva columna con el valor constante “Denuncia” la cual denota que los resultados de este flujo de trabajo corresponden al tipo de caso Denuncia.

17.6.1.6 Resultados

Se presentan los resultados obtenidos para el pronóstico de los casos a recibir por tipo de caso.

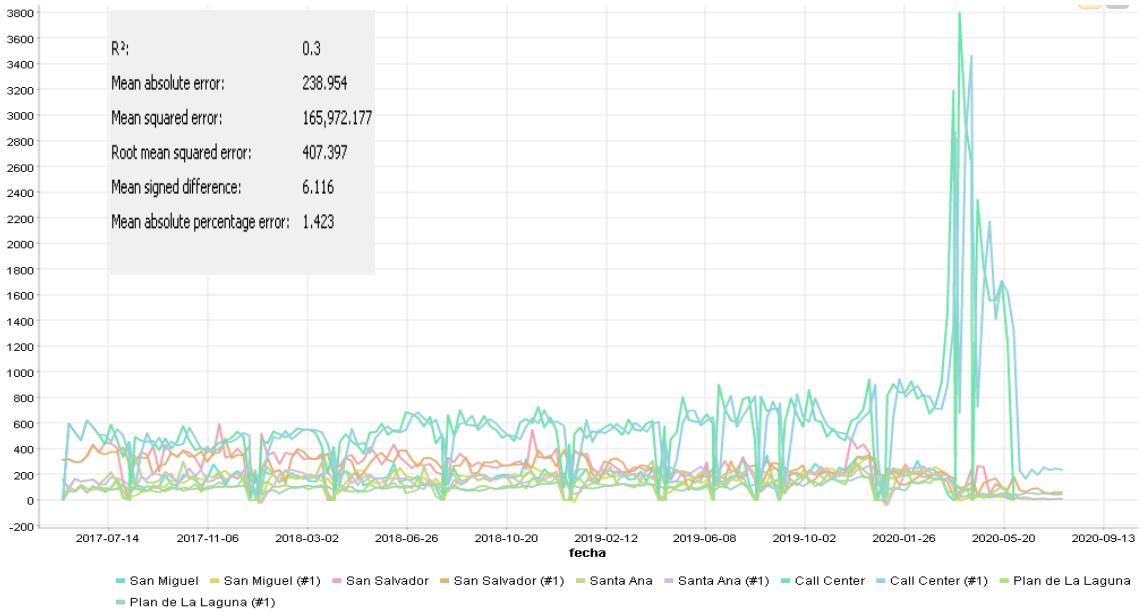


Figura 166 Gráfico de líneas Asesorías vs predicción Asesorías.

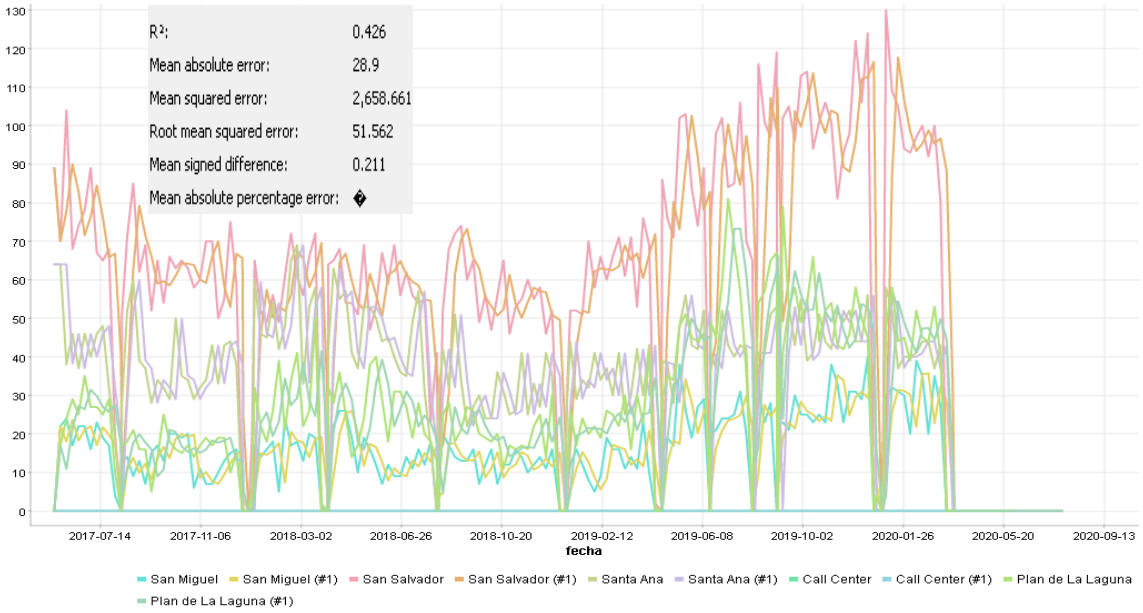


Figura 167 Gráfico de líneas Denuncias vs predicción Denuncias.

17.6.1.7 Interpretación de resultados

Para el caso del $R^2 = 0.3$ en las asesorías analizando el significado directo de este indicador muestra en qué medida el modelo se ajusta a la variable que se intenta explicar varía de 0 a 1 entre mayor es mejor es el resultado. El error absoluto medio viene dado en la escala en el que se muestran los datos por lo cual es un indicador de precisión y consiste en la sumatoria de los errores absolutos de cada observación por lo tanto existe un 238.954 de error medio en las predicciones, Por otro lado el error medio cuadrado castiga aquellos periodos donde la diferencia fue más alta en comparación de otros por eso se observa un valor de 165972.177 y culla raíz cuadrada que corresponde al error cuadrático medio es más comparable con 407.397 además se encuentra la diferencia media firmada de 6.116 la cual corresponde a la media de las diferencias del dato observado versus el dato que se pronosticó y finalmente se tiene el error porcentual absoluto medio el cual es una medida con mayor interpretación debido a que muestra en términos porcentuales sin importar la escala para este ejercicio se obtuvo un 1.423 lo cual es bastante aceptable, entonces concluyendo en este campo en el que no tenemos precedente de cuándo puede decidir un ser humano solicitar asesoría a la Defensoría sea tenga una problemática a nivel de consumo o no, a pesar del valor bajo de R^2 el predictor es estadísticamente significativo cuando de pronosticar cuantas asesorías se brindaran en las próximas semanas.

Respecto a las Denuncias, se tiene un $R^2 = 0.426$ un error absoluto medio 28.9, un error medio cuadrado 2658.661, un error cuadrático medio de 51.562, una diferencia media firmada de 0.211 y en el caso del error porcentual absoluto medio, que no se pudo calcular debido a que existen observaciones de los datos originales con valor 0, se puede concluir que aun siendo el error R^2 mayor que en el caso de las asesorías aún es bajo, pero se observa una diferencia firmada mucho más baja lo cual es positivo por lo tanto la variable a explicar sigue teniendo valor para los datos que se han pronosticado, además podemos observar al final del gráfico una caída de la recepción de denuncias tanto en los datos originales como en los pronosticados esto debido al COVID-19.

Respecto a las gestiones y derivaciones se tiene una caída de la recepción de estos tipos de caso debido a la puesta en vigencia de la Ley de procesos administrativos (LPA) la Defensoría ya no está facultada para brindar este tipo de atenciones.

17.6.2 Pronosticar casos solucionados en fechas futuras.

17.6.2.1 Contenido del caso.

N° de caso: C-FOR-02	
Técnica	Pronostico.
Algoritmos	Series temporales (ARIMA)
Población	Datos del modelo multidimensional SARA 2005-2019.
Variables	Se analizan los casos solucionados de tipo de caso gestión a lo largo del tiempo.
Hipótesis	Pronosticar atenciones a solucionar por oficina
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos pronosticados.
Interpretación de resultados	Mediante indicadores de rendimiento.

Herramienta de software	Knime Analytics Platform.

Tabla 70 Contenido de Caso C-FOR-02.

17.6.2.2 Población

SQL Statement

```
select dt.anio,last_day(dt.fecha)fecha,count(*)cant from fact_sara_atenciones_brindadas fs
left join dim_tipo_caso dtc
on (dtc.sk_tipo_caso = fs.sk_tipo_caso)
left join dim_atencion da
on(da.sk_atencion =fs.sk_atencion)
left join dim_tiempo dt
on (dt.sk_tiempo =fs.sk_tiempo_solucion)
left join dim_oficina dof on(dof.sk_oficina =fs.sk_oficina)
where dof.nombre_oficina ='San Salvador'
and dtc.tipo_caso_agrupado ='Gestión'
and da.nombre_solucion in(
'Avenimiento',
'Cerrado por razones de oficio',
'Conciliación',
'Falta de Ratificación y Prevención',
'Tribunal Sancionador'
)
and fs.sk_tiempo_solucion  is not null
group by dt.anio,dt.fecha
order by dt.anio,dt.fecha )t
group by t.anio,t.fecha
```

Preview results:

Row ID	anio	fecha	count
Row0	2008	2008-02-29	5
Row1	2008	2008-03-31	53
Row2	2008	2008-04-30	287
Row3	2008	2008-05-31	256
Row4	2008	2008-06-30	245
Row5	2008	2008-07-31	265
Row6	2008	2008-08-31	237
Row7	2008	2008-09-30	244
Row8	2008	2008-10-31	228
Row9	2008	2008-11-30	153

Figura 168 Set de datos para Caso FOR-02.

En la Figura 168 muestra la consulta para obtener el set de datos a utilizar la cual corresponde a la suma atenciones cerradas a final de cada mes por año y filtrado por las diferentes oficinas.

17.6.2.3 Variables

Las variables a utilizar se detallan a continuación: año, última fecha del mes, sumatoria a atenciones solucionadas a final de mes y oficina.

17.6.2.4 Hipótesis

Pronosticar casos a solucionar en fechas futuras, para poder realizar diferentes evaluaciones de desempeño.

17.6.2.5 Procedimiento

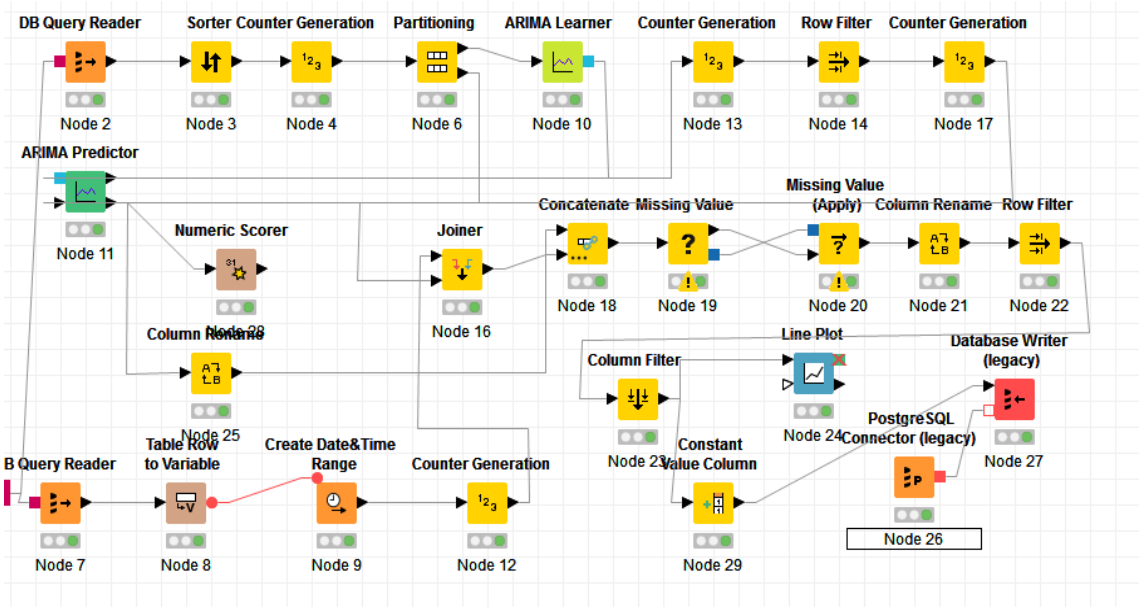


Figura 169 Flujo de Trabajo Caso C-FOR-02.

La Figura 169 muestra el flujo de trabajo desarrollado para cumplir con el objetivo de minería de datos para el caso C-FOR-02 se detalla a continuación las acciones que se llevan a cabo:

- **DB Query Reader (Node 2):** Se ejecuta la consulta al modelo multidimensional
- **Sorter (Node 3):** Se ordenan los datos por fecha ascendente
- **Counter Generation (Node 4):** Se enumeran los datos del 1 al n.
- **Partitioning (Node 6):** Se particionan los datos en 75% entrenamiento y 25% pruebas.
- **DB Query Reader (Node 7):** Se ejecuta la consulta para obtener los siguientes valores:
    - dia_inicio: La última fecha en la BD para la que se tienen atenciones +7.
    - dia_fin: 5 años hacia el futuro.
    - cantidad_dias: días entre el día inicio y fin.
    - anio_antes: Año anterior al año actual.
    - anio_actual: Año en curso.
- **Table Row to Variable (Node 8):** Convierte los valores obtenidos en el nodo anterior en variables.
- **Create Date&Time Range (Node 9):** Crea un rango de fechas futuras tomando como día de inicio la variable dia_inicio, fecha final la variable dia_fin y cantidad de registros la variable cantidad_dias.
- **Arima Learner (Node 10):** Se entrena el algoritmo de series temporales ARIMA con el set de datos de entrenamiento, obteniendo como resultado el modelo
- **Arima Predictor (Node 11):** Se entrena el algoritmo de series temporales ARIMA con el set de datos de entrenamiento, obteniendo como resultado el modelo.
- **Counter Generation (Node 12):** Se enumeran los datos provenientes del Nodo 9
- **Counter Generation (Node 13):** Se enumeran los datos provenientes del Nodo 11

- **Row Filter (Node 14):** Se filtran las filas provenientes del Nodo 13 dejando únicamente las de los nuevos periodos pronosticados.
- **Counter Generation (Node 17):** Se enumeran los datos provenientes del Nodo 14
- **Joiner (Node 16):** Se realiza un join con los datos provenientes del nodo 12 y nodo 17
- **Concatenate (Node 18):** Realiza la concatenación de los datos resultantes entre el nodo 25 y Nodo 16
- **Mising Values (Node 19):** definen los criterios a aplicar a los valores faltantes en el caso de la cantidad de denuncias solucionadas se coloca 0.
- **Missing Values Apply (Node 20):** Se aplican los criterios a los valores faltantes.
- **Column Rename (Node 21):** Se renombran las columnas que tienen los casos cerrados y la predicción de casos cerrados
- **Row Filter (Node 22):** Se filtran los datos por medio del año
- **Column Filter (Node 23):** Se filtran las columnas para solo dejar la columna de fecha y la cantidad de casos cerrados y la predicción de casos cerrados.
- **Line Plot (Node 24):** Se grafican los resultados.
- **Numeric Scorer (Node 28):** Se obtienen los indicadores de rendimiento.

17.6.2.6 Resultados

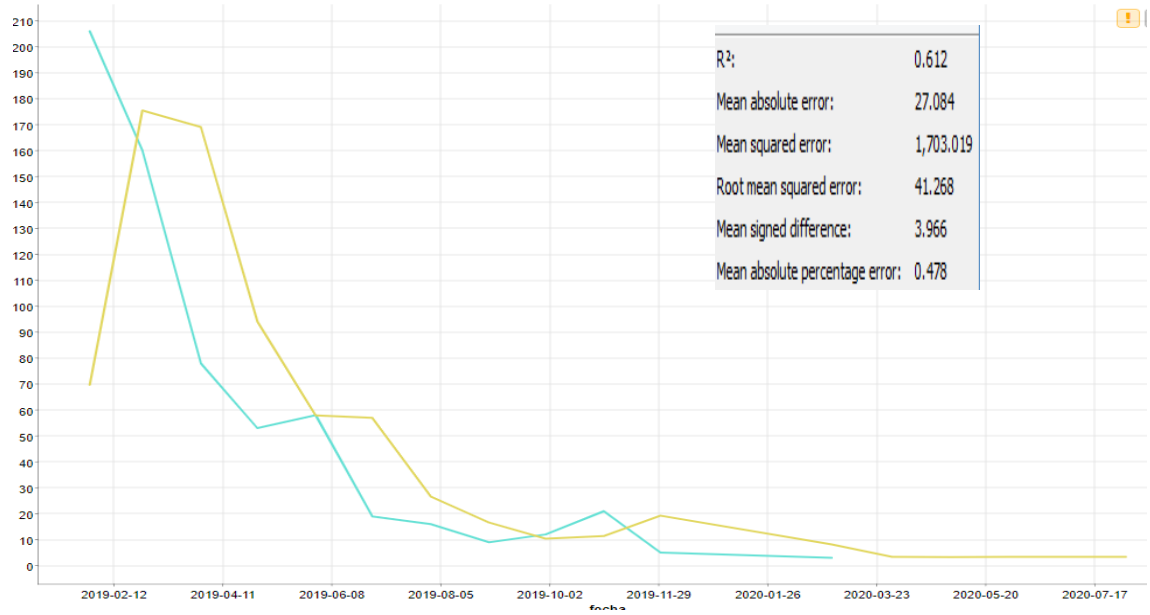


Figura 170 Predicción de Casos solucionados - Oficina San Salvador

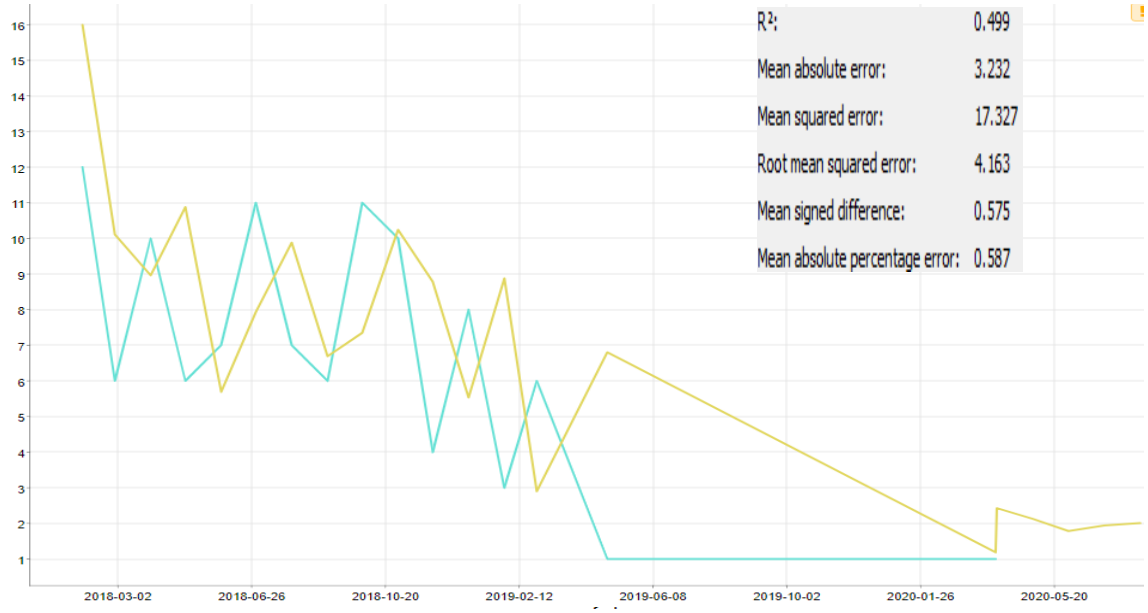


Figura 171 Predicción Casos Solucionados - Oficina Call Center

17.6.2.7 Interpretación de resultados

Se presenta el análisis para dos de las oficinas, la primera “San Salvador” Para el caso del $R^2 = 0.612$ el significado directo de este indicador muestra en qué medida el modelo se ajusta a la variable que se intenta explicar varía de 0 a 1 entre mayor es mejor es el resultado sin embargo un modelo no se puede descartar ni aceptar con está única medida, El error absoluto medio viene dado en la escala en el que se muestran los datos por lo cual es un indicador de precisión y consiste en la sumatoria de los errores absolutos de cada observación por lo tanto existe un 27.084 de error medio en las predicciones, Por otro lado el error medio cuadrado castiga aquellos periodos donde la diferencia fue más alta en comparación de otros por eso se observa un valor de 1703.019 y culla raíz cuadrada que corresponde al error cuadrático medio es más comparable con 41.268 además se encuentra la diferencia media firmada de 3.966 la cual corresponde a la media de las diferencias del dato observado versus el dato que se pronosticó y finalmente se tiene el error porcentual absoluto medio el cual es una medida con mayor interpretación debido a que muestra en términos porcentuales sin importar la escala para este ejercicio se obtuvo un 0.478 lo cual es bastante aceptable, entonces concluyendo en este campo en el que no tenemos precedente de cuando se pueda dar por solucionada una atención se tiene un estimado de en base a las atenciones pasadas cuantas atenciones se podrían solucionar para la oficina de San Salvador

Para el análisis para la oficina Call center se obtuvo un $R^2 = 0.612$, un El error absoluto medio de 3.322, un error medio cuadrado de 17.327, la raíz cuadrada del error medio cuadrado de 4.163 y una diferencia media firmada de 0.211 y el error porcentual absoluto medio de 0.587 en base a los indicadores de rendimiento se puede decir que se tiene unos resultados aceptables para las predicciones pero la falta de datos en los diferentes últimos meses de atenciones solucionas afecta en parte la predicción y en base a los últimos datos registradas se espera que la atenciones a solucionar continúen similar a los últimos meses.

17.7 Técnica de agrupamiento

17.7.1 Segmentar consumidores en base a los motivos en los cuales han solicitado atención a la DC

17.7.1.1 Contenido del caso

N° de caso: C-AGR-01	
Técnica	Agrupamiento.
Algoritmos	K-means Clustering
Población	Datos provenientes del modelo multidimensional SARA desde el año 2005 hasta el año 2020.
Variables	Se analiza el sector, motivo, categoría, forma de recepción, tipo de caso y el municipio del consumidor.
Hipótesis	Agrupar a los consumidores mediante sus variables por los motivos en los cuales solicitaron atenciones a la DC.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos agrupados
Interpretación de resultados	-
Herramienta de software	Knime Analytics Platform.

Tabla 71 Contenido del caso C-AGR-01.

17.7.1.2 Población

La población de datos utilizada consiste en determinar las variables que se tienen en común en la atención brindada al consumidor, entre estas tenemos los identificadores del sector, motivo, categoría, forma de recepción, tipo de caso, el género del consumidor y el municipio, tal y como se muestra en la Figura 172.

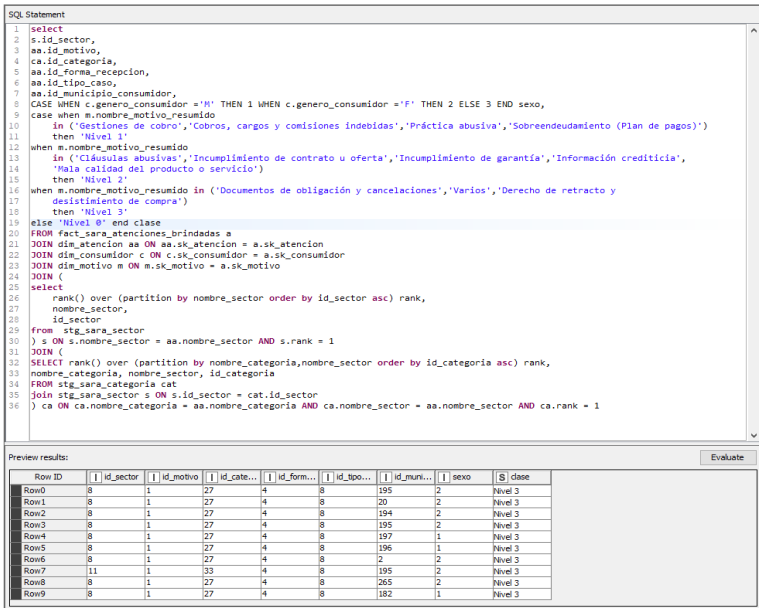


Figura 172 Contenido del caso C-AGR-01.

17.7.1.3 Variables

Las variables que se ven involucradas en la exploración son los identificadores de sector, motivo, categoría, forma de recepción, tipo de caso, el género del consumidor y el municipio.

17.7.1.4 Hipótesis

El objetivo de minería de datos que queremos ejecutar, o la hipótesis que queremos comprobar es la siguiente: “Realizar una segmentación a los consumidores en base a los motivos en los cuales se solicitan atenciones a la Defensoría del Consumidor”.

17.7.1.5 Procedimiento configuración

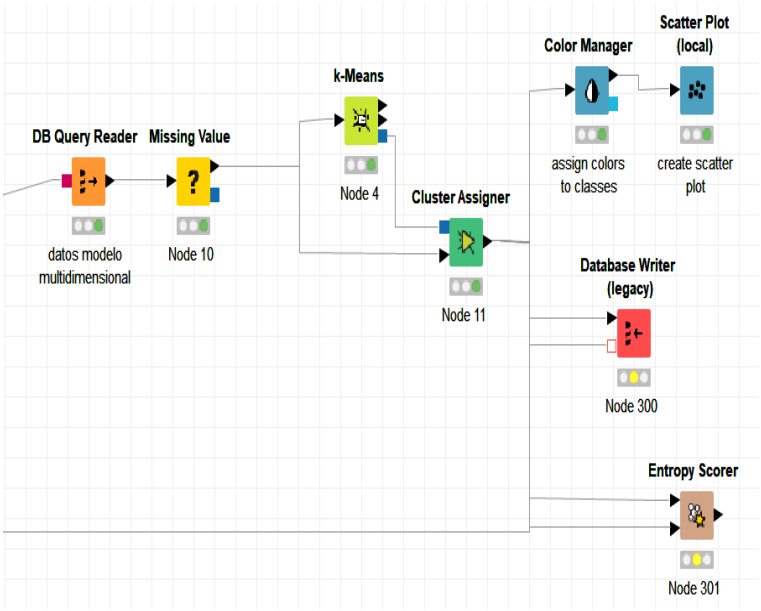


Figura 173: Flujo de trabajo aplicando k-means para el caso C-AGR-01.

Descripción del procedimiento:

- **DB Connector**: Establece la conexión con la base de datos Postgres para acceder a los datos del modelo multidimensional.
- **DB Query Reader:** Obtiene los datos de las atenciones para ser exploradas con el algoritmo de minería de datos.
- **Missing Value:** Calcula un valor mediante los valores más frecuentes y se lo asigna a los datos faltantes.
- **K-Means:** Aplica el algoritmo K-Means para generar el número de clústeres configurado en las propiedades del nodo (en nuestro caso 3 clúster) y así, asignar el mejor clúster a los datos.
- **Color Maganer:** Asigna una serie de colores dependiendo de la variable asignada, en este caso asignará un color diferente a cada clúster encontrado por el algoritmo.
- **Scatter Ploter:** Utilizado para generar un gráfico de puntos de dispersión de manera rápida.
- **Cluster Asigner:** Este nodo se utiliza para asignar el clúster a los datos pasados como parámetros.

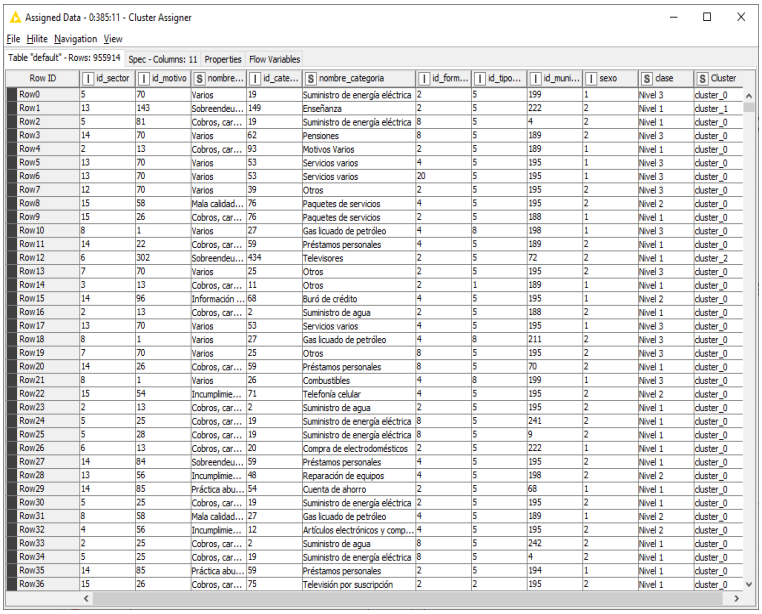
- **Database Writer (legacy):** Utilizado para almacenar los datos en la base de datos.
- **Entropy Scorer:** Genera una serie de indicadores para evaluar los datos obtenidos con el algoritmo de agrupamiento k means.

17.7.1.6 Resultados

Para realizar el agrupamiento de los datos, se ha hecho uso del algoritmo de K-means, este algoritmo es de tipo no supervisado y consiste en agrupar los datos en "k" grupos basándose en sus características.

Primero asigna un centro de clúster o centroide por cada clúster (en nuestro caso serán 3 clústeres), el algoritmo calcula la distancia euclidiana entre el dato y los centroides, y se asigna el clúster que tenga menor distancia, se actualiza el centroide de cada grupo tomando como nuevo centroide la posición promedio de los datos pertenecientes a dicho grupo, se vuelve a calcular la distancia euclidiana y se asigna el nuevo clúster a cada dato, estos últimos dos pasos se repiten hasta que los centroides no se mueven.

Para nuestro caso las variables a tomar en cuenta en el algoritmo son: id_sector, id_motivo, id_categoria, id_forma_recepcion, id_tipo_caso, id_municipio_consumidor, sexo. Y da como resultado:



Row ID	id_sector	id_motivo	nombre...	id_cate...	nombre_categoria	id_form...	id_tipo...	id_muni...	sexo	clase	Cluster
Row0	5	70	Varios	19	Suministro de energía eléctrica	2	5	199	1	Nivel 3	cluster_0
Row1	13	143	Sobreendeu...	149	Enseñanza	2	5	222	2	Nivel 1	cluster_1
Row2	5	81	Cobros, car...	19	Suministro de energía eléctrica	8	5	4	2	Nivel 1	cluster_0
Row3	14	70	Varios	62	Pensiones	8	5	189	2	Nivel 3	cluster_0
Row4	2	13	Cobros, car...	93	Motivos Varios	2	5	189	1	Nivel 1	cluster_0
Row5	13	70	Varios	53	Servicios varios	4	5	195	1	Nivel 3	cluster_0
Row6	13	70	Varios	53	Servicios varios	20	5	195	1	Nivel 3	cluster_0
Row7	12	70	Varios	39	Otros	2	5	195	2	Nivel 3	cluster_0
Row8	15	58	Mala calidad...	76	Paquetes de servicios	4	5	195	2	Nivel 2	cluster_0
Row9	15	26	Cobros, car...	76	Paquetes de servicios	2	5	188	1	Nivel 1	cluster_0
Row10	8	1	Varios	27	Gas licuado de petróleo	4	8	198	1	Nivel 3	cluster_0
Row11	14	22	Cobros, car...	59	Préstamos personales	4	5	189	2	Nivel 1	cluster_0
Row12	6	302	Sobreendeu...	434	Televisores	2	5	72	2	Nivel 1	cluster_2
Row13	7	70	Varios	25	Otros	2	5	195	2	Nivel 3	cluster_0
Row14	3	13	Cobros, car...	11	Otros	2	1	189	1	Nivel 1	cluster_0
Row15	14	96	Información ...	68	Buró de crédito	4	5	195	1	Nivel 2	cluster_0
Row16	2	13	Cobros, car...	2	Suministro de agua	2	5	188	2	Nivel 1	cluster_0
Row17	13	70	Varios	53	Servicios varios	4	5	195	1	Nivel 3	cluster_0
Row18	8	1	Varios	27	Gas licuado de petróleo	4	8	211	2	Nivel 3	cluster_0
Row19	7	70	Varios	25	Otros	8	5	195	2	Nivel 3	cluster_0
Row20	14	26	Cobros, car...	59	Préstamos personales	8	5	70	2	Nivel 1	cluster_0
Row21	8	1	Varios	26	Combustibles	4	8	199	1	Nivel 3	cluster_0
Row22	15	54	Incumplimie...	71	Telefonía celular	4	5	195	2	Nivel 2	cluster_0
Row23	2	13	Cobros, car...	2	Suministro de agua	2	5	195	2	Nivel 1	cluster_0
Row24	5	25	Cobros, car...	19	Suministro de energía eléctrica	8	5	241	2	Nivel 1	cluster_0
Row25	5	28	Cobros, car...	19	Suministro de energía eléctrica	8	5	9	2	Nivel 1	cluster_0
Row26	6	13	Cobros, car...	20	Compra de electrodomésticos	2	5	222	1	Nivel 1	cluster_0
Row27	14	84	Sobreendeu...	59	Préstamos personales	4	5	195	2	Nivel 1	cluster_0
Row28	13	56	Incumplimie...	48	Reparación de equipos	4	5	198	2	Nivel 2	cluster_0
Row29	14	85	Práctica abu...	54	Cuenta de ahorro	2	5	68	1	Nivel 1	cluster_0
Row30	5	25	Cobros, car...	19	Suministro de energía eléctrica	2	5	195	2	Nivel 1	cluster_0
Row31	8	58	Mala calidad...	27	Gas licuado de petróleo	4	5	189	1	Nivel 2	cluster_0
Row32	4	56	Incumplimie...	12	Artículos electrónicos y comp...	4	5	195	2	Nivel 2	cluster_0
Row33	2	25	Cobros, car...	2	Suministro de agua	8	5	242	2	Nivel 1	cluster_0
Row34	5	25	Cobros, car...	19	Suministro de energía eléctrica	8	5	4	2	Nivel 1	cluster_0
Row35	14	85	Práctica abu...	59	Préstamos personales	2	5	194	1	Nivel 1	cluster_0
Row36	15	26	Cobros, car...	75	Televisión por suscripción	2	2	195	2	Nivel 1	cluster_0

Figura 174 Clústeres generados por el algoritmo k-means.

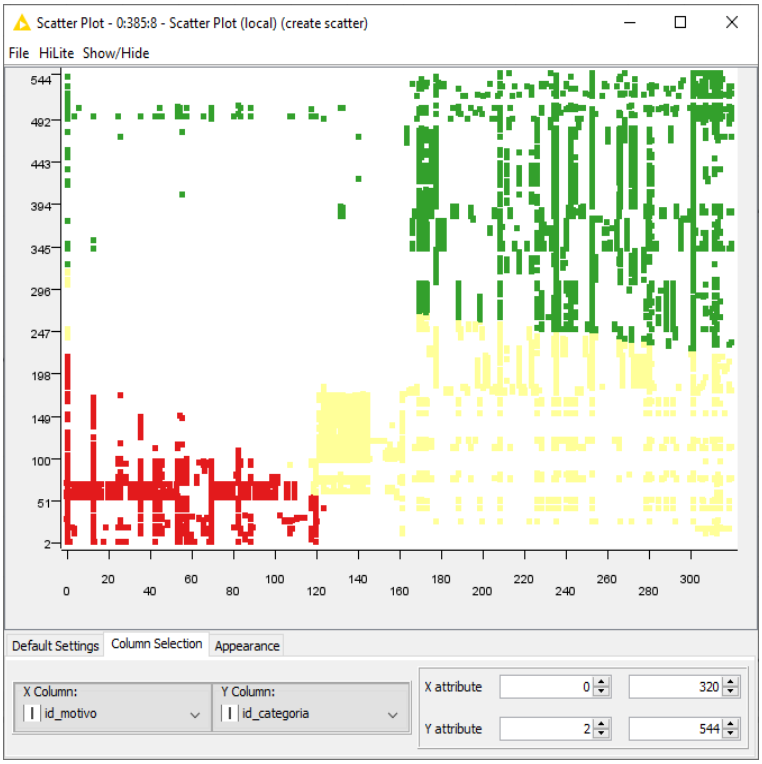


Figura 175 Clústeres generados por el algoritmo k-means de manera gráfica.

Como se puede observar en la Figura 174, al generar la gráfica de puntos de dispersión y tomando como eje Y el id de la categoría, como eje X el id del motivo, se pueden observar los 3 clústeres encontrados, cada clúster con un color diferente.

17.7.1.7 Interpretación de resultados

El primer paso que se realizó es determinar una clase a la cual pertenece cada dato, para esto se tomó como parámetro el motivo resumido, tal y como se muestra a continuación:

Clase	Motivo resumido
Nivel 1	Gestiones de cobros, cobros, cargos y comisiones indebidas, práctica abusiva, sobreendeudamiento (Plan de pagos)
Nivel 2	Cláusulas abusivas, incumplimiento de contrato u oferta, incumplimiento de garantía, información crediticia, mala calidad del producto o servicio.
Nivel 3	Documentos de obligación y cancelaciones, varios, derecho de retracto y desistimiento de compra

Tabla 72 Parámetros tomados para determinar el campo "Clase".

Luego de esto, para generar los clústeres mediante el algoritmo de k means, se incluyeron las variables que tuvieran más relación a la atención brindada al consumidor, como lo es el sector, motivo, categoría, forma de recepción, tipo de caso, el municipio, así como el género y el motivo resumido.

Realizar la segmentación por solo una variable, como el motivo resumido sesgaría mucho los resultados, ya que solo se realizaría mediante solo una variable de todas las que intervienen en la atención al consumidor, es por ello que se agregan las demás variables, y realizar esta segmentación manualmente, tomando todas las variables sería un proceso

muy engorroso de realizar, es por ello que se opta por utilizar minería de datos, ya que los algoritmos facilitan esta tarea de manera rápida y eficaz.

Tomando como referencia la columna “clase” y realizando un conteo de cuantos datos pertenecen a cada uno de los niveles tenemos:

Clase	Recuento
Nivel 1	510,916
Nivel 2	285,913
Nivel 3	159,085

Tabla 73 Recuento de los datos por clase.

Clase	Recuento
Cluster 0	567,295
Cluster 1	254,972
Cluster 2	133,647

Tabla 74 Recuento de datos por clúster.

Interpretando los resultados podemos ver que entre el campo clase y el clúster hay una cierta similitud en los datos:

Clase	Recuento	Cluster	Recuento	Variación	Variación %
Nivel 1	510,916.00	**Cluster 0**	567,295.00	- 56,379.00	-10%
Nivel 2	285,913.00	**Cluster 1**	254,972.00	30,941.00	12%
Nivel 3	159,085.00	**Cluster 2**	133,647.00	25,438.00	19%

Tabla 75 Comparativa entre recuento de clases y clúster.

Tomando como referencia el recuento entre las clases y los clústeres, podemos notar que el algoritmo segmentó los grupos con una diferencia máxima del 19%. Para verificar los resultados de agrupación, Knime nos proporciona un nodo llamada “Entropy Scorer”. La entropía mide cuánto coinciden las etiquetas de los clústeres con unos datos etiquetados previamente. En este caso la etiqueta “Clase” y “Cluster”.

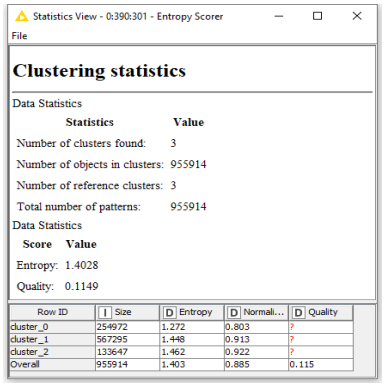


Figura 176 Estadísticas obtenidas por el nodo Entropy Scorer.

Estas estadísticas nos dicen que comparando las columnas “Clase” contra “Cluster”, obtenemos una entropía total de 1.40 (Menos es mejor) y una calidad de 0.11 (Más es mejor, valor máximo 1), con estos resultados podemos interpretar que la primera etiqueta que se le dio a los datos (columna clase) no se realizó de manera correcta (solo nos otorga una calidad del 11%), ya que solo se etiqueta mediante una sola variable (motivo de la

atención), en cambio, la realizada por el algoritmo k means toma muchas más variables que interviene en la atención brindada al consumidor

17.7.2 Identificar grupos de meses que reciben más casos.

17.7.2.1 Contenido del caso.

N° de caso: C-AGR-02	
Técnica	Agrupamiento.
Algoritmos	K-means Clustering
Población	Datos provenientes del modelo multidimensional SARA desde el año 2005 hasta el año 2020.
Variables	Se analizan la cantidad de atenciones recibidas por mes.
Hipótesis	Agrupar los meses que se reciben más atenciones.
Procedimiento	Flujo de trabajo en Knime.
Resultados	Datos agrupados
Interpretación de resultados	--
Herramienta de software	Knime Analytics Platform.

Tabla 76 Contenido del caso C-AGR-02.

17.7.2.2 Población

SQL Statement

```
1  select
2      dt.anio ,
3      dt.mes,
4      count(*)
5      from fact_sara_atenciones_brindadas fsab
6  join dim_tiempo dt on
7      (dt.sk_tiempo = fsab.sk_tiempo_ingreso)
8  group by
9      dt.mes,
10     dt.anio
```

Preview results: Evaluate

Row ID	anio	mes	count
Row0	2005	1	6820
Row1	2006	1	44343
Row2	2007	1	52522
Row3	2008	1	17153
Row4	2009	1	5156
Row5	2010	1	7426
Row6	2011	1	5232
Row7	2012	1	5980
Row8	2013	1	5980
Row9	2014	1	6294

Figura 177 Set de datos para el caso C-AGR-02

17.7.2.3 Variables

Las variables involucradas son: año, Mes y Cantidad de atenciones

17.7.2.4 Hipótesis

Identificar los grupos de meses en los diferentes años donde se reciben más atenciones.

17.7.2.5 Procedimiento configuración

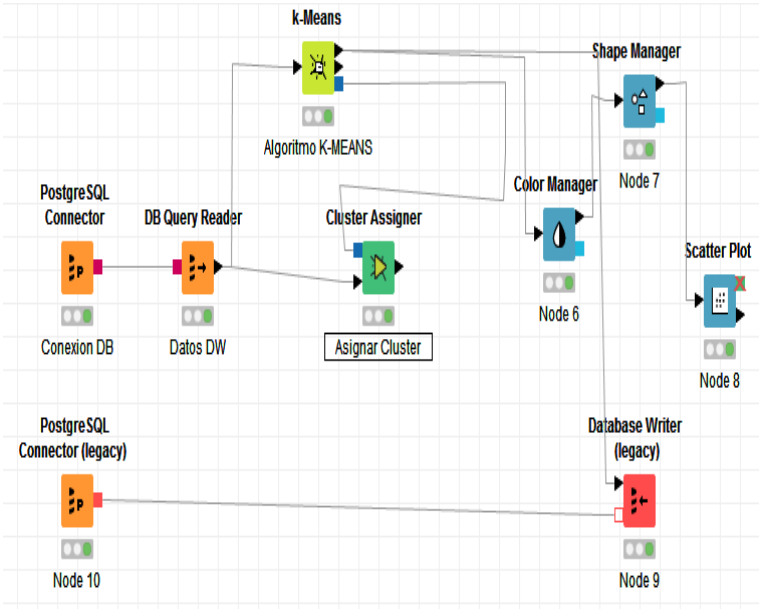


Figura 178 Flujo de trabajo aplicando el algoritmo K-means.

La Figura 178 muestra el flujo de trabajo para el caso C-AGR-02 al igual a C- AGR -01 se utiliza el algoritmo de K-means para agrupar utilizando el mismo procedimiento que el caso anterior para generar el modelo y obtener la información.

17.7.2.6 Resultados

Row ID	I anio	L mes	L count	S Cluster
Row0	2005	1	6820	cluster_1
Row1	2006	1	44343	cluster_0
Row2	2007	1	52522	cluster_0
Row3	2008	1	17153	cluster_1
Row4	2009	1	5156	cluster_1
Row5	2010	1	7426	cluster_1
Row6	2011	1	5232	cluster_1
Row7	2012	1	5980	cluster_1
Row8	2013	1	5980	cluster_1
Row9	2014	1	6294	cluster_1
Row10	2015	1	5183	cluster_1
Row11	2016	1	7173	cluster_1
Row12	2017	1	6707	cluster_1
Row13	2018	1	6441	cluster_1
Row14	2019	1	6904	cluster_1
Row15	2020	1	8037	cluster_1
Row16	2008	2	5667	cluster_1
Row17	2009	2	5307	cluster_1
Row18	2010	2	6788	cluster_1
Row19	2011	2	4754	cluster_1
Row20	2012	2	5449	cluster_1
Row21	2013	2	5173	cluster_1
Row22	2014	2	5441	cluster_1
Row23	2015	2	4699	cluster_1
Row24	2016	2	6358	cluster_1

Figura 179 Resultados de la asignación de los clústeres

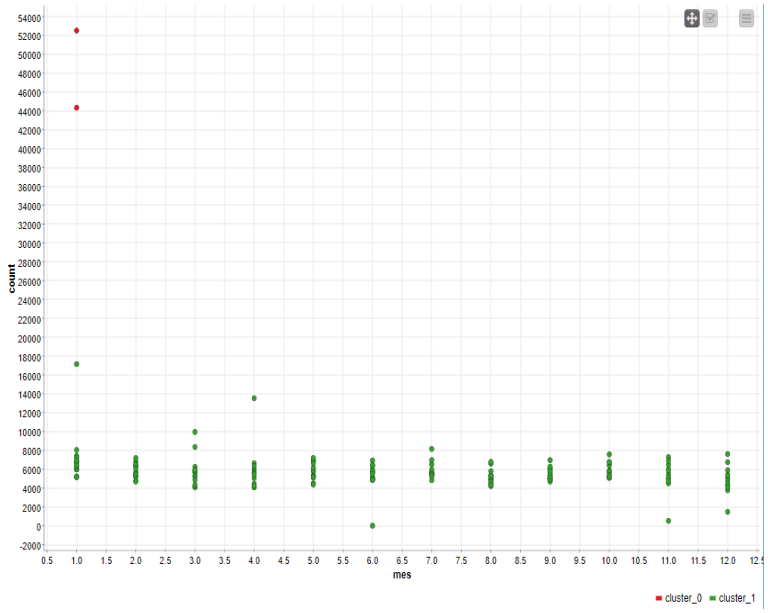


Figura 180 Representación gráfica de los clústeres generados

17.7.2.7 Interpretación de resultados

Al aplicar el algoritmo K-MEANS se generan dos grupos para el total de atenciones que se reciben por mes como se observa en la Figura 180 se tiene los datos representados con el color verde pertenecientes al cluster_1 y cluster_0.

18 Sprint 6

18.1 Descripción Historias de Usuario

Código	RA301
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar en un mapa los proveedores que en más reiteradas ocasiones han sido denunciados por consumidores.
Razón	Para tener una perspectiva más detallada de la ubicación de proveedores reincidentes.
Criterios de aceptación	Se debe identificar el municipio al que pertenecen.
	Los proveedores deben de estar depurados en la medida de lo posible.
	Se debe mostrar la cantidad de denuncias interpuestas hacia los proveedores en cada municipio.
Validación	Que al colocar el puntero sobre el municipio se visualice los proveedores y cantidad de denuncias correspondientes.
	Que se muestre una cantidad top de proveedores por municipio.
	Que los municipios sean claramente identificables dentro del mapa.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 77 Historia de Usuario RA301.

Código	RA302
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar las atenciones según el motivo.
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Se debe de filtrar por casos cerrados y recibidos.
	Se debe obtener un acumulado de todo el año.
	Se debe identificar en un gráfico de barra las cantidades de atenciones por motivo.
Validación	Verificar que se filtre de forma correcta los casos según los tipos mencionados.
	Que se muestre una tabla con las descripciones y motivos diversos.
	Que estén ordenadas de manera descendente.
Valor del negocio	600
Puntos de historia	3
ROI	200

Tabla 78 Historia de usuaria RA302.

Código	RA303
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer un informe de las atenciones brindadas
Razón	Para tener una representación gráfica de las atenciones por tipo.
Criterios de aceptación	Se debe mostrar información para los tipos de caso asesoría, denuncia, derivación y gestión

	Se debe mostrar un cuadro comparativo de las atenciones brindadas por tipo de caso.
	Se debe mostrar un gráfico de barra con la totalidad de casos para el año en curso y el año anterior
Validación	Se comprobará que muestre la diferencia de casos en cantidad y porcentaje según el mes actual y el mes anterior.
	Se comprobará que se muestre información en los lapsos de tiempo especificados.
	Se validará que la información mostrada concuerde con los informes que genera la UACM
Valor del negocio	400
Puntos de historia	3
ROI	133

Tabla 79 Historia de usuario RA303.

Código	**RA304**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer un informe de las atenciones brindadas según región.
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Se debe mostrar que muestre información para todas las regiones que se tengan disponibles.
	Se debe mostrar un cuadro comparativo por región para el mes en curso y para el acumulado.
	Se debe mostrar un gráfico de pastel para las atenciones brindadas por región.
Validación	Se comprobará que muestre información para los tipos de caso asesoría, denuncia, derivación y gestión.
	Se comprobará que muestre la cantidad de casos por región tanto en cantidad como en porcentaje.
	Se comprobará que se muestre información en los lapsos de tiempo especificados.
	Se validará que la información mostrada concuerde con los informes que genera la UACM.
Valor del negocio	400
Puntos de historia	5
ROI	80

Tabla 80 Historia de usuario RA304.

Código	**RA305**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer un informe de las atenciones brindadas según oficina
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Se comprobará que muestre información para todas las oficinas.
	Se debe mostrar un cuadro comparativo por tipo de oficina para el mes en curso y para la totalidad del año.
	Se debe mostrar un gráfico de pastel para las atenciones brindadas por tipo de oficina.

Validación	Se comprobará que muestre información para los tipos de caso asesoría, denuncia, derivación y gestión
	Se comprobará que muestre la cantidad de casos por oficina tanto en cantidad como en porcentaje.
	Se comprobará que se muestre información en los lapsos de tiempo especificados.
	Se validará que la información mostrada concuerde con los informes que genera la UACM
Valor del negocio	400
Puntos de historia	3
ROI	133

Tabla 81 Historia de usuario RA305.

Código	**RA306**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea conocer un informe de las atenciones brindadas por sector
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Se comprobará que muestre información para todos los sectores
	Se debe mostrar un gráfico de barra con las totalidades de atenciones por sector para el mes en curso
	Se debe mostrar un gráfico de barra con las totalidades de atenciones por sector para lo transcurrido del año en curso
Validación	Se comprobará que muestre la cantidad de casos por sector tanto en cantidad como en porcentaje.
	Se comprobará que se muestre información en los lapsos de tiempo especificados.
	Se validará que la información mostrada concuerde con los informes que genera la UACM
Valor del negocio	400
Puntos de historia	3
ROI	133

Tabla 82 Historia de usuario RA306.

Código	**RA307**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se genere un reporte para las atenciones dadas en medios descentralizados.
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Cuadro comparativo para las atenciones brindadas por medios descentralizados para el mes en curso y el mes anterior.
	Cuadro comparativo para las atenciones brindadas por medios descentralizados acumuladas en el año en curso y el mismo tiempo del año anterior.
	En cada cuadro comparativo es necesario que se muestre la variación entres las fechas especificadas.
Validación	Se validará que la información mostrada concuerde con los informes que genera la UACM

	Que se muestre la variación tanto en cantidad como en términos porcentuales.
	Se validará que se contengan totales al final de cada tabla.
Valor del negocio	500
Puntos de historia	5
ROI	188

Tabla 83 Historia de Usuario RA307.

Código	**RA308**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea la creación de un cuadro comparativo para las atenciones en ventanillas descentralizadas.
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Cuadro comparativo para las atenciones brindadas por ventanillas descentralizadas para el mes en curso y el mes anterior.
	Cuadro comparativo para las atenciones brindadas por ventanillas descentralizadas acumuladas en el año en curso y el mismo tiempo del año anterior.
	En cada cuadro comparativo es necesario que se muestre la variación entre las fechas especificadas.
Validación	Se validará que la información mostrada concuerde con los informes que genera la UACM
	Que se muestre la variación tanto en cantidad como en términos porcentuales.
	Se validará que se contengan totales al final de cada tabla.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 84 Historia de Usuario RA308.

Código	**RA309**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea que se genere un reporte de los casos cerrados y los montos recuperados.
Razón	Para tener una representación gráfica de las atenciones por este tópico.
Criterios de aceptación	Se desea generar un cuadro comparativo para las denuncias y gestiones cerradas para lo que va del año en curso y al mismo tiempo, pero del año anterior.
	Se desea que en el cuadro comparativo para las denuncias y gestiones cerradas exista una columna sobre la variación que se ha tenido.
	Se desea generar un cuadro comparativo donde se muestren los casos cerrados, los casos con devoluciones y los montos que se recuperaron de todo el año hasta el mes en curso.
Validación	Se validará que la información mostrada concuerde con los informes que genera la UACM

	Que se muestre la variación tanto en cantidad como en términos porcentuales.
	Se validará que se contengan totales al final de cada tabla.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 85 Historia de Usuario RA309.

Código	**RA310**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea generar un mapa de El Salvador interactivo sobre los casos recibidos y los casos cerrados.
Razón	Para visualizar la información de forma rápida y clara sobre las atenciones en los distintos departamentos del país.
Criterios de aceptación	Que se muestre la información de todos los departamentos del país.
	Que la información se pueda filtrar entre rangos de fechas.
	Que exista la posibilidad de filtrar los casos entre recibidos, casos cerrados o ambos.
Validación	Que al colocar el puntero sobre el departamento se visualice se muestren los casos cerrados y recibidos.
	Se validará que la información mostrada corresponda a los filtros seleccionados.
	Se corroborará que se pueda filtrar entre casos recibidos, casos cerrados o ambos.
Valor del negocio	500
Puntos de historia	4
ROI	125

Tabla 86 Historia de Usuario RA310.

Código	**RA311**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea generar un informe sobre las atenciones dadas según su forma de recepción por diferentes filtros.
Razón	Para poder obtener la información de las atenciones brindadas por las diferentes formas de recepción de forma rápida y clara.
Criterios de aceptación	Se desea que en el informe se muestre a través de un mapa interactivo para poder visualizar de forma rápida el número de atenciones dadas según su forma de recepción.
	El informe deberá contener diferentes filtros como forma de recepción, rango de fechas.
	El informe deberá mostrar por qué forma de recepción se ha dado el mayor número de casos.
Validación	Que al colocar el puntero sobre el departamento se muestre la cantidad de atenciones según su forma de recepción.
	Se validará que la información mostrada corresponda a los filtros seleccionados.

	Se corroborará que el indicador que muestra la forma de recepción con mayor numero sea correcto.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 87 Historia de Usuario RA311.

Código	**RA312**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar la relación existente entre el aumento de las denuncias hacia proveedores respecto a los meses del año.
Razón	Para observar cuando algunos proveedores realicen prácticas abusivas.
Criterios de aceptación	Se debe mostrar los meses del año.
	Debe mostrar que proveedores se ven denunciados por cada mes.
	Se debe mostrar solo proveedores con cantidades significativas de denuncias.
Validación	Que se valide que la confianza como criterio de aceptación sea >=85%.
	Que se muestren únicamente conjunto de proveedores frecuentes con cantidad mayor o igual a 3.
	Que se muestren resultados de forma tabular y gráfica.
Valor del negocio	900
Puntos de historia	8
ROI	113

Tabla 88 Historia de Usuario RA312.

Código	**RA313**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar la segmentación de consumidores en base a los motivos en los cuales han solicitado atención a la DC (referencia a RA202).
Razón	Para brindar programas de educación en consumo a esos grupos de personas.
Criterios de aceptación	Se deberá generar un informe que muestre los datos generados a partir de la minería.
	El informe deberá contar con un gráfico de puntos.
	El informe deberá contar con una tabla que muestre las cantidades por cada segmento.
Validación	Se corroborará que la información sea correcta y concisa.
	Validar que en el gráfico de puntos se muestren los distintos segmentos identificado por color.
	Se validará que en el gráfico se permita seleccionar una o varias segmentaciones para ser mostradas.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 89 Historia de usuario RA313.

Código	RA314
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar de en qué meses se brindan más atenciones a los consumidores.
Razón	Para poder reorientar el recurso entre las distintas unidades en base a la demanda.
Criterios de aceptación	Se deberá generar un informe que muestre los datos generados a partir de la minería.
	El informe deberá contar con un gráfico de puntos.
	El informe deberá contar con una tabla que muestre las cantidades por cada segmento.
Validación	Se corroborará que la información sea correcta y concisa.
	Validar que en el gráfico de puntos se muestren los distintos segmentos identificado por color.
	Se validará que en el gráfico se permita seleccionar una o varias segmentaciones para ser mostradas.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 90 Historia de usuario RA314.

Código	RA315
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar qué solución tendrán los casos recibidos en base a la edad y otros parámetros que se consideren relevantes de los consumidores.
Razón	Para tener una noción del resultado y poder tomar las acciones necesarias.
Criterios de aceptación	Se deberá generar un informe que muestre los datos generados a partir de la minería.
	Debe mostrar el indicador de rendimiento con su resultado.
	Debe contabilizar las cantidades por cada una de las clasificaciones
Validación	Que los resultados sean claros y concisos.
	Se validará que se muestren los resultados de forma tabular y/o gráfica.
	Que se identifique con colores cada una de las distintas clasificaciones.
Valor del negocio	500
Puntos de historia	5
ROI	100

Tabla 91 Historia de Usuario RA315.

Código	RA316
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar la influencia que ha tenido la DC en las atenciones.
Razón	Para destacar la labor realizada en materia de los derechos de los consumidores.
	Debe mostrar el indicador de rendimiento con su resultado.

Criterios de aceptación	Debe contabilizar las cantidades por cada una de las clasificaciones
	Se debe mostrar para los últimos años únicamente, año con año.
Validación	Que los resultados sean claros y concisos.
	Se validará que se muestren los resultados de forma tabular y/o gráfica.
	Que se identifique con colores cada una de las distintas clasificaciones.
Valor del negocio	900
Puntos de historia	5
ROI	180

Tabla 92 Historia de Usuario RA316

Código	**RA317**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar el comportamiento de los proveedores en base a los montos reclamados, montos recuperados y soluciones que se han tenido
Razón	Para poner en perspectiva el actuar de los proveedores cuando son sujeto de denuncias interpuestas en su contra.
Criterios de aceptación	Se deberá generar un informe que muestre los datos generados a partir de la minería.
	Debe mostrar el indicador de rendimiento con su resultado.
	Debe contabilizar las cantidades por cada una de las clasificaciones
Validación	Debe mostrar el indicador de rendimiento con su resultado.
	Debe contabilizar las cantidades por cada una de las clasificaciones
	Que se identifique con colores cada una de las distintas clasificaciones.
Valor del negocio	500
Puntos de historia	3
ROI	167

Tabla 93 Historia de usuario RA317.

Código	**RA318**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar el pronóstico de los casos a recibir en fechas futuras.
Razón	Para planificar recursos en base a la demanda.
Criterios de aceptación	Esta salida se deberá elaborar con datos obtenidos del resultado de la explorativa mediante minería de datos.
	Se debe mostrar mediante un gráfico de líneas.
	El gráfico de líneas debe mostrar por separado los datos reales vs los estimados.
Validación	Se debe validar que las líneas correspondientes a los datos reales y estimados se visualicen de diferente color.
	Se validará que en el eje horizontal se muestre la variable tiempo.
	El segmento de datos pronosticados debe ser identificables dentro de la gráfica y debe ser de al menos un mes.
Valor del negocio	500

Puntos de historia	3
ROI	167

Tabla 94 Historia de Usuario RA318.

Código	**RA319**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar el pronóstico de los casos solucionados en fechas futuras.
Razón	Para establecer un estimado de las metas y así poder medir el rendimiento.
Criterios de aceptación	Se debe mostrar mediante un gráfico de líneas.
	Se debe incluir un filtro que permita visualizar una o varias soluciones dentro del gráfico de líneas.
	El gráfico de líneas debe mostrar por separado los datos reales vs los estimados.
Validación	Se debe validar que las líneas correspondientes a los datos reales y estimados se visualicen de diferente color.
	Se debe validar que los datos presentados en la gráfica correspondan a los seleccionados en el filtro.
	El segmento de datos pronosticados debe ser identificables dentro de la gráfica y debe ser de al menos un mes.
Valor del negocio	500
Puntos de historia	3
ROI	167

Tabla 95 Historia de usuario RA319.

Código	**RA320**
Rol	Como técnico(a) UACM/ Jefatura.
Funcionalidad	Se desea visualizar en un mapa las atenciones que han sido brindadas a los consumidores.
Razón	Para tener una perspectiva más detallada de la ubicación de las atenciones brindadas.
Criterios de aceptación	Se debe identificar el municipio al que pertenecen.
	Se debe mostrar la cantidad de las atenciones brindadas.
	Se debe mostrar la cantidad detallada de las atenciones brindadas según el tipo de caso.
Validación	Que al colocar el puntero sobre el municipio se visualice las atenciones brindadas.
	Que se muestre las cantidades de las atenciones brindadas según su tipo de caso por municipio.
	Que los municipios sean claramente identificables dentro del mapa.
Valor del negocio	600
Puntos de historia	5
ROI	120

Tabla 96 Historia de usuario RA320.

18.2 Caso de uso

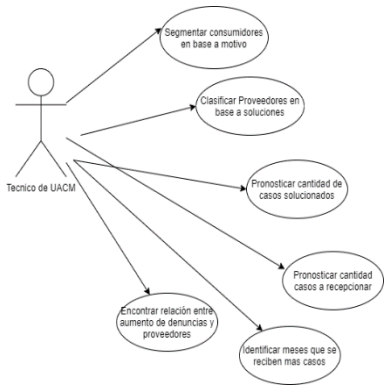


Figura 181 Diagrama de Casos de Uso Sprint 6.

18.3 Reportes de inteligencia de negocios

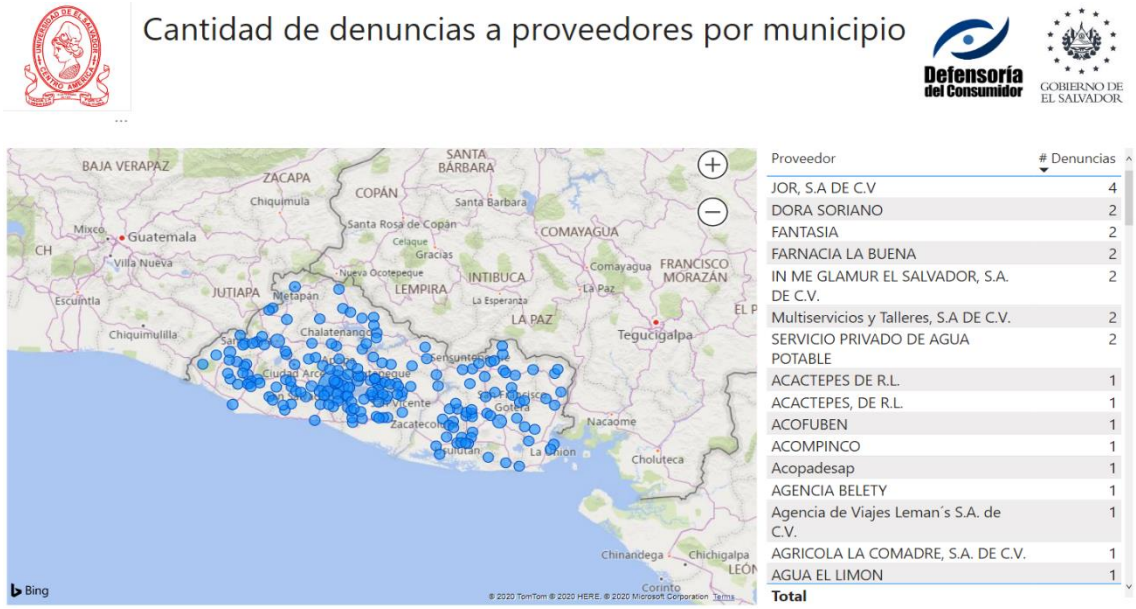


Figura 182 Reporte de denuncias a proveedores por municipio.

La Figura 182 presenta el reporte de las denuncias hechas a los proveedores por municipio. Estas denuncias se muestran a través de un mapa, una lista que muestra el nombre del proveedor y la cantidad de denuncias interpuestas ante él. Por cada municipio, se despliega un punto de color azul que representa la cantidad de denuncias, el tamaño del punto varía según la cantidad de ellas. Si se selecciona un punto, este actualiza los datos de la lista de proveedores con los datos del municipio seleccionado.

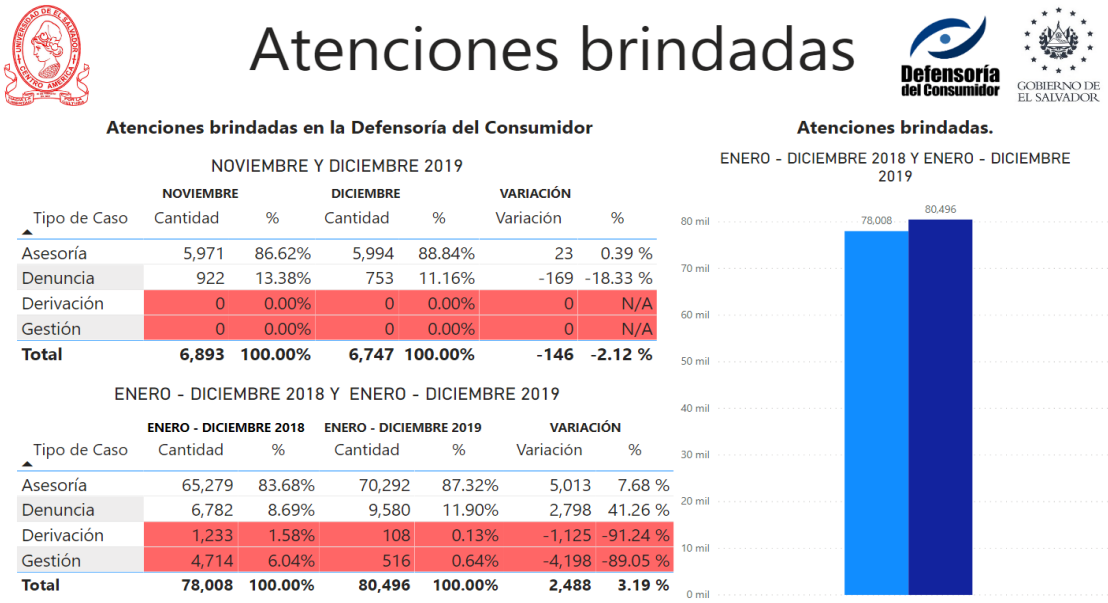


Figura 183 Diseño reporte atenciones brindadas.

La Figura 183 presenta el reporte de las atenciones brindadas. La primera tabla muestra la variación de atenciones entre el mes seleccionado y el mes anterior. La segunda muestra las atenciones brindadas acumuladas en el año hasta el mes seleccionado y la variación que se ha tenido en ambas tablas. Asimismo, se muestra un gráfico de barras de la cantidad acumulada de atenciones brindadas entre el año seleccionado en el filtro y el año anterior.

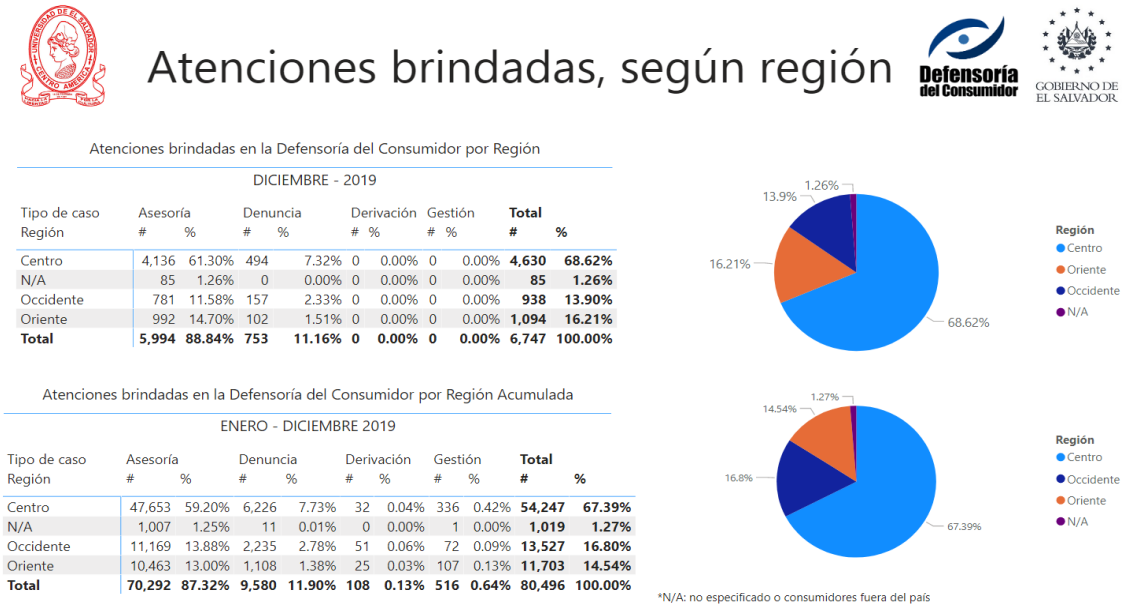


Figura 184 Diseño reporte atenciones brindadas, según región

La Figura 184 muestra el reporte de atenciones brindadas según región. El reporte consiste en dos tablas y dos gráficos de pastel que muestran la cantidad de atenciones por tipo de caso en las distintas regiones del país. La primera tabla muestra la cantidad de atenciones brindadas en el mes, este se selecciona a través de un filtro. La segunda tabla muestra el

número de atenciones acumuladas en el año hasta el mes seleccionado en el filtro. De igual manera, los gráficos de pastel muestran los valores totales de las atenciones según región.

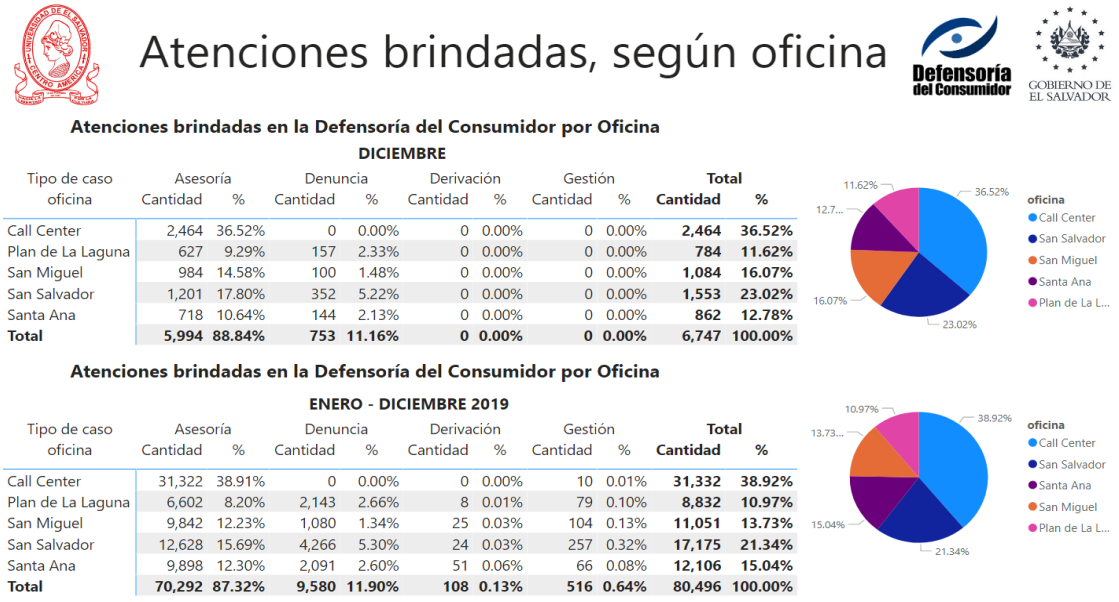


Figura 185 Diseño reporte atenciones brindadas, según oficina.

La Figura 185 muestra el reporte de atenciones brindadas según oficina. El reporte consiste en dos tablas y dos gráficos de pastel que muestran la cantidad de atenciones por tipo de caso las distintas oficinas del país. La primera tabla muestra la cantidad de atenciones brindadas en el mes, este se selecciona a través de un filtro. La segunda tabla muestra el número de atenciones acumuladas en el año hasta el mes seleccionado en el filtro. De igual manera, los gráficos de pastel muestran los valores totales de las atenciones según oficina.

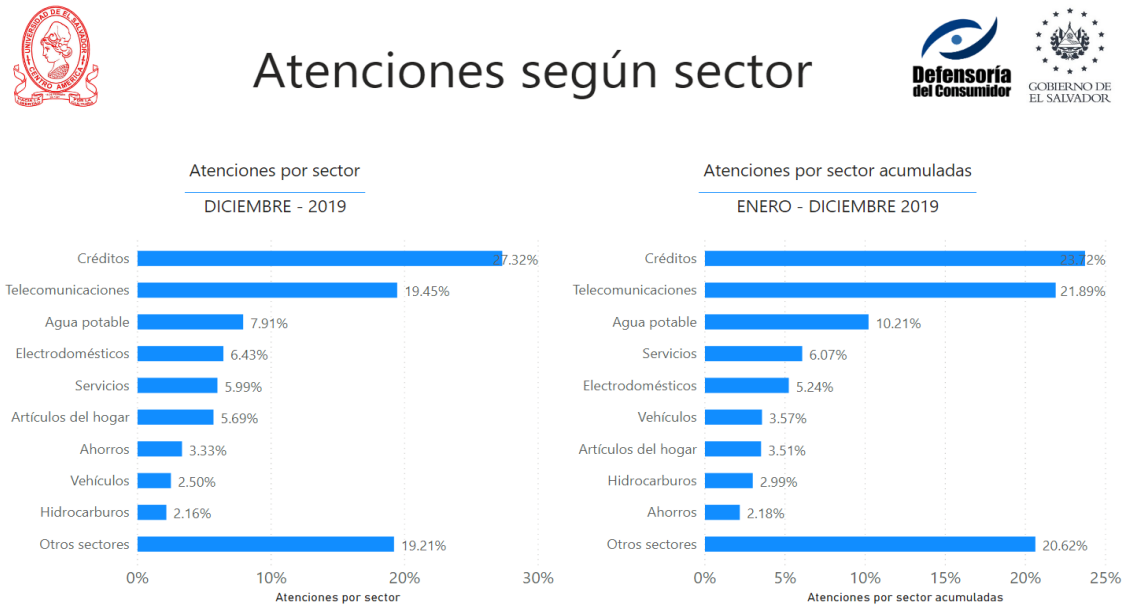


Figura 186 Diseño Reporte Atenciones, según sector.

La Figura 186 presenta el reporte de Atenciones según sector. Este reporte está compuesto por dos gráficos de barras horizontales que muestran las atenciones brindadas en términos porcentuales por la Defensoría del Consumidor en los diferentes sectores.

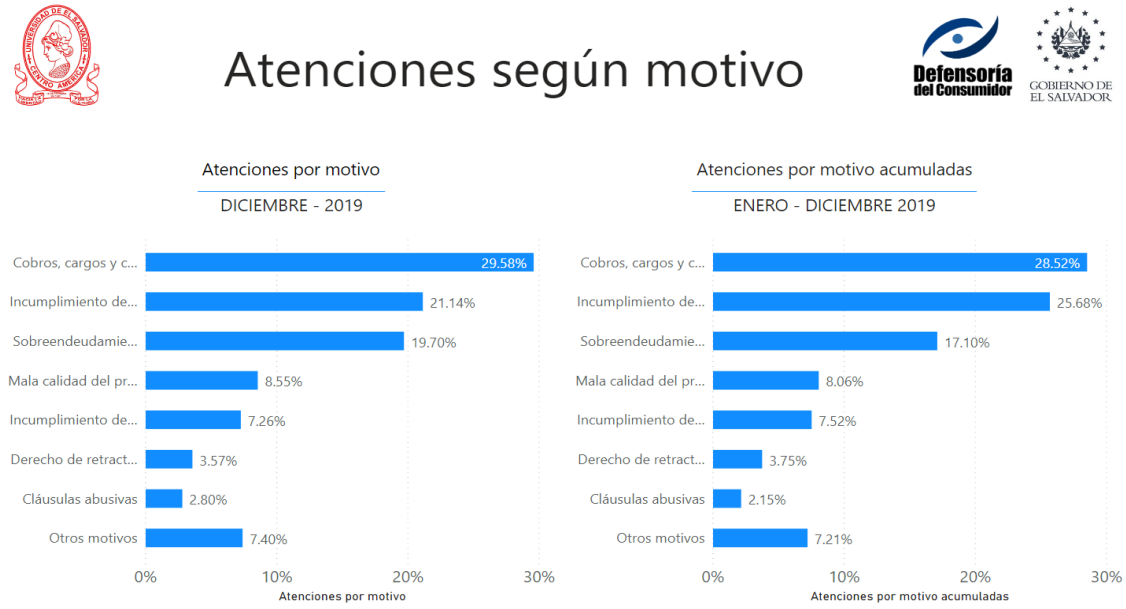
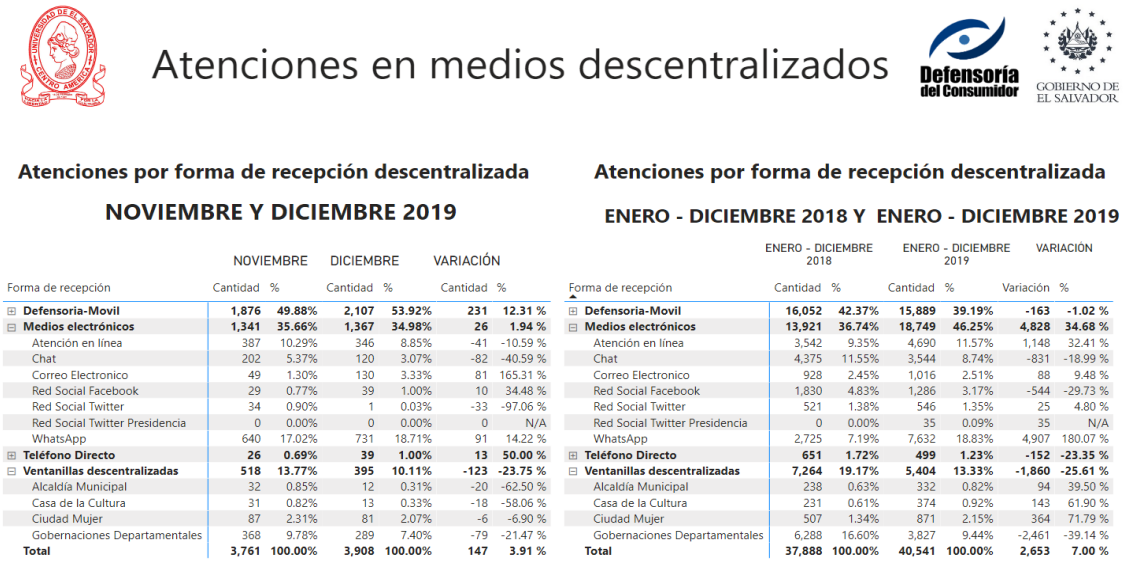


Figura 187 Diseño Reporte Atenciones, según motivo.

La Figura 187 presenta el reporte de Atenciones según motivo. Este reporte está compuesto por dos gráficos de barras horizontales que muestran las atenciones brindadas en términos porcentuales por la Defensoría del Consumidor por los diferentes motivos, el primer gráfico muestra las atenciones del mes seleccionado en un filtro y el segundo gráfico muestra el acumulado de los meses del año hasta el mes seleccionado en el filtro.

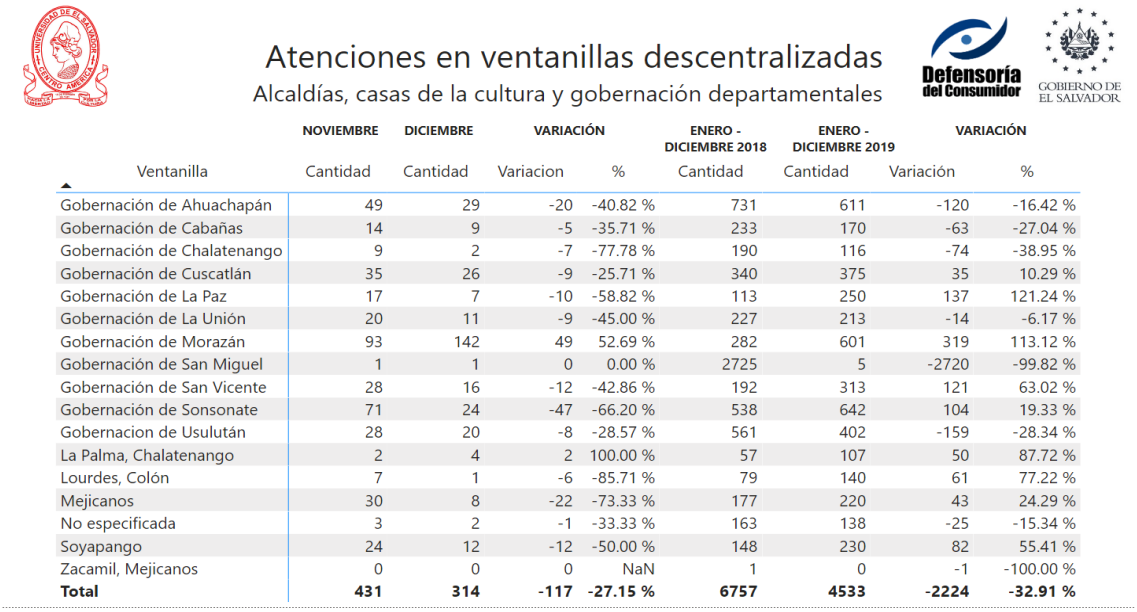


	NOVIEMBRE		DICIEMBRE		VARIACIÓN	
Forma de recepción	Cantidad	%	Cantidad	%	Cantidad	%
Defensoria-Movil	**1,876**	**49.88%**	**2,107**	**53.92%**	**231**	**12.31 %**
Medios electrónicos	**1,341**	**35.66%**	**1,367**	**34.98%**	**26**	**1.94 %**
Atención en línea	387	10.29%	346	8.85%	-41	-10.59 %
Chat	202	5.37%	120	3.07%	-82	-40.59 %
Correo Electronico	49	1.30%	130	3.33%	81	165.31 %
Red Social Facebook	29	0.77%	39	1.00%	10	34.48 %
Red Social Twitter	34	0.90%	1	0.03%	-33	-97.06 %
Red Social Twitter Presidencia	0	0.00%	0	0.00%	0	N/A
WhatsApp	640	17.02%	731	18.71%	91	14.22 %
Teléfono Directo	**26**	**0.69%**	**39**	**1.00%**	**13**	**50.00 %**
Ventanillas descentralizadas	**518**	**13.77%**	**395**	**10.11%**	**-123**	**-23.75 %**
Alcaldía Municipal	32	0.85%	12	0.31%	-20	-62.50 %
Casa de la Cultura	31	0.82%	13	0.33%	-18	-58.06 %
Ciudad Mujer	87	2.31%	81	2.07%	-6	-6.90 %
Gobernaciones Departamentales	368	9.78%	289	7.40%	-79	-21.47 %
Total	**3,761**	**100.00%**	**3,908**	**100.00%**	**147**	**3.91 %**

	ENERO - DICIEMBRE 2018		ENERO - DICIEMBRE 2019		VARIACIÓN	
Forma de recepción	Cantidad	%	Cantidad	%	Variación	%
Defensoria-Movil	**16,052**	**42.37%**	**15,889**	**39.19%**	**-163**	**-1.02 %**
Medios electrónicos	**13,921**	**36.74%**	**18,749**	**46.25%**	**4,828**	**34.68 %**
Atención en línea	3,542	9.35%	4,690	11.57%	1,148	32.41 %
Chat	4,375	11.55%	3,544	8.74%	-831	-18.99 %
Correo Electronico	928	2.45%	1,016	2.51%	88	9.48 %
Red Social Facebook	1,830	4.83%	1,286	3.17%	-544	-29.73 %
Red Social Twitter	521	1.38%	546	1.35%	25	4.80 %
Red Social Twitter Presidencia	0	0.00%	35	0.09%	35	N/A
WhatsApp	2,725	7.19%	7,632	18.83%	4,907	180.07 %
Teléfono Directo	**651**	**1.72%**	**499**	**1.23%**	**-152**	**-23.35 %**
Ventanillas descentralizadas	**7,264**	**19.17%**	**5,404**	**13.33%**	**-1,860**	**-25.61 %**
Alcaldía Municipal	238	0.63%	332	0.82%	94	39.50 %
Casa de la Cultura	231	0.61%	374	0.92%	143	61.90 %
Ciudad Mujer	507	1.34%	871	2.15%	364	71.79 %
Gobernaciones Departamentales	6,288	16.60%	3,827	9.44%	-2,461	-39.14 %
Total	**37,888**	**100.00%**	**40,541**	**100.00%**	**2,653**	**7.00 %**

Figura 188 Diseño Reporte Atenciones en medios descentralizados.

El Figura 188 presenta el reporte de las atenciones por medios descentralizados. La primera tabla muestra la cantidad de atenciones y la cantidad porcentual para el mes seleccionado en un filtro, así como el mes anterior y su variación. La segunda tabla muestra las cantidades acumuladas en el año hasta el mes seleccionado como parámetro, el acumulado del mes anterior y su variación.



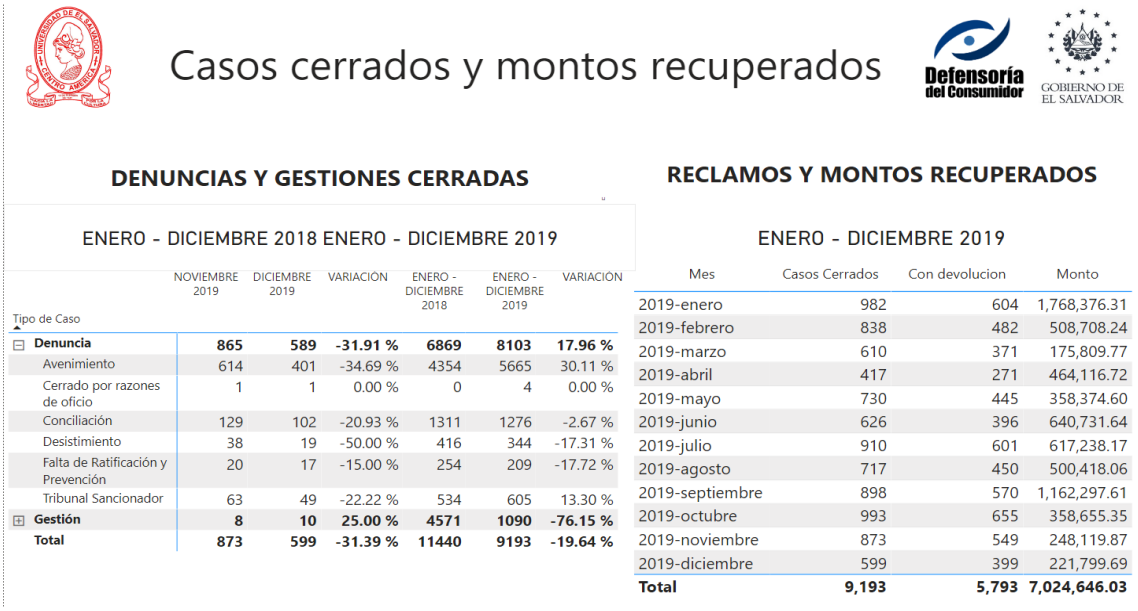
Atenciones en ventanillas descentralizadas

Alcaldías, casas de la cultura y gobernación departamentales

	NOVIEMBRE	DICIEMBRE	VARIACIÓN		ENERO - DICIEMBRE 2018	ENERO - DICIEMBRE 2019	VARIACIÓN	
Ventanilla	Cantidad	Cantidad	Variacion	%	Cantidad	Cantidad	Variación	%
Gobernación de Ahuachapán	49	29	-20	-40.82 %	731	611	-120	-16.42 %
Gobernación de Cabañas	14	9	-5	-35.71 %	233	170	-63	-27.04 %
Gobernación de Chalatenango	9	2	-7	-77.78 %	190	116	-74	-38.95 %
Gobernación de Cuscatlán	35	26	-9	-25.71 %	340	375	35	10.29 %
Gobernación de La Paz	17	7	-10	-58.82 %	113	250	137	121.24 %
Gobernación de La Unión	20	11	-9	-45.00 %	227	213	-14	-6.17 %
Gobernación de Morazán	93	142	49	52.69 %	282	601	319	113.12 %
Gobernación de San Miguel	1	1	0	0.00 %	2725	5	-2720	-99.82 %
Gobernación de San Vicente	28	16	-12	-42.86 %	192	313	121	63.02 %
Gobernación de Sonsonate	71	24	-47	-66.20 %	538	642	104	19.33 %
Gobernacion de Usulután	28	20	-8	-28.57 %	561	402	-159	-28.34 %
La Palma, Chalatenango	2	4	2	100.00 %	57	107	50	87.72 %
Lourdes, Colón	7	1	-6	-85.71 %	79	140	61	77.22 %
Mejicanos	30	8	-22	-73.33 %	177	220	43	24.29 %
No especificada	3	2	-1	-33.33 %	163	138	-25	-15.34 %
Soyapango	24	12	-12	-50.00 %	148	230	82	55.41 %
Zacamil, Mejicanos	0	0	0	NaN	1	0	-1	-100.00 %
Total	**431**	**314**	**-117**	**-27.15 %**	**6757**	**4533**	**-2224**	**-32.91 %**

Figura 189 Diseño Reporte Atenciones en ventanillas descentralizados.

La Figura 189 muestra el reporte de las atenciones en ventanillas descentralizadas. Este consta de una tabla donde se muestra la cantidad de atenciones brindadas, se muestra el mes, el mes anterior, la cantidad acumulada del año hasta el mes seleccionado y el acumulado del año anterior, así como su variación en cantidad y en porcentaje.



Casos cerrados y montos recuperados

DENUNCIAS Y GESTIONES CERRADAS

ENERO - DICIEMBRE 2018 ENERO - DICIEMBRE 2019

Tipo de Caso	NOVIEMBRE 2019	DICIEMBRE 2019	VARIACIÓN	ENERO - DICIEMBRE 2018	ENERO - DICIEMBRE 2019	VARIACIÓN
Denuncia	**865**	**589**	**-31.91 %**	**6869**	**8103**	**17.96 %**
Avenimiento	614	401	-34.69 %	4354	5665	30.11 %
Cerrado por razones de oficio	1	1	0.00 %	0	4	0.00 %
Conciliación	129	102	-20.93 %	1311	1276	-2.67 %
Desistimiento	38	19	-50.00 %	416	344	-17.31 %
Falta de Ratificación y Prevención	20	17	-15.00 %	254	209	-17.72 %
Tribunal Sancionador	63	49	-22.22 %	534	605	13.30 %
Gestión	**8**	**10**	**25.00 %**	**4571**	**1090**	**-76.15 %**
Total	**873**	**599**	**-31.39 %**	**11440**	**9193**	**-19.64 %**

RECLAMOS Y MONTOS RECUPERADOS

ENERO - DICIEMBRE 2019

Mes	Casos Cerrados	Con devolucion	Monto
2019-enero	982	604	1,768,376.31
2019-febrero	838	482	508,708.24
2019-marzo	610	371	175,809.77
2019-abril	417	271	464,116.72
2019-mayo	730	445	358,374.60
2019-junio	626	396	640,731.64
2019-julio	910	601	617,238.17
2019-agosto	717	450	500,418.06
2019-septiembre	898	570	1,162,297.61
2019-octubre	993	655	358,655.35
2019-noviembre	873	549	248,119.87
2019-diciembre	599	399	221,799.69
Total	**9,193**	**5,793**	**7,024,646.03**

Figura 190 Diseño Reporte casos cerrados y montos recuperados

muestra el reporte de casos cerrados y montos recuperados. La primera tabla muestra la cantidad de atenciones por tipo de caso que han sido cerrados en el mes seleccionado y el mes anterior y el acumulado del año hasta el mes seleccionado, el año anterior, y la variación. La segunda tabla muestra la cantidad de casos que han sido cerrados en el año hasta el mes seleccionado, se muestra la cantidad de casos con devolución, así como el monto total de estas.

Figura 191 Diseño Reporte de número de casos según departamento.

La Figura 191 presenta el reporte de casos según departamento. Este reporte muestra en un mapa el número de casos que se han recibidos o se han cerrados, según el valor del filtro. El número de casos se muestran mediante círculos de color azul, el tamaño del circulo es proporcional al valor o sea al número de casos.

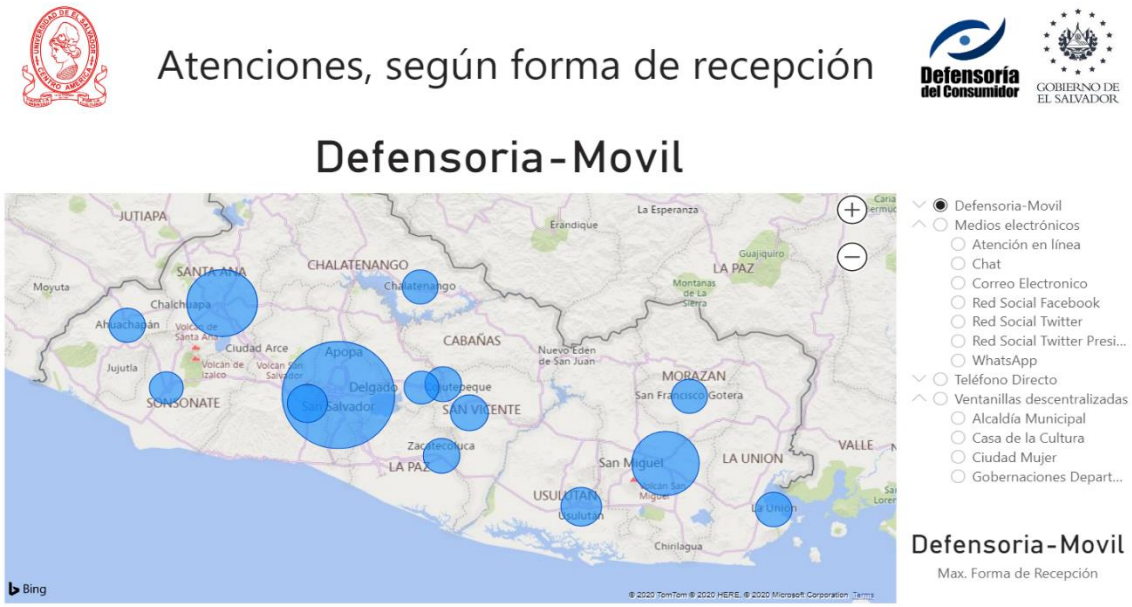


Figura 192 Diseño Reporte de atenciones, según forma de recepción.

La Figura 192 presenta el reporte de atenciones según forma de recepción. Este reporte muestra en un mapa el número de atenciones brindadas. El número de atenciones se muestra mediante círculos de color azul, el tamaño del circulo es proporcional al valor o sea al número de casos. De igual manera, se muestra un filtro de la forma de recepción y un indicador que nos muestra la forma de recepción con más atenciones dadas en el mes seleccionado como parámetro.

Figura 193 Diseño del reporte de atenciones brindadas por municipio.

La Figura 193 presenta el reporte de atenciones brindadas por municipio. Este reporte muestra en un mapa el número de atenciones brindadas. El número de atenciones se muestra mediante círculos de color azul, el tamaño del circulo depende del número de atenciones, entre más grande es el valor, más grande es el circulo.

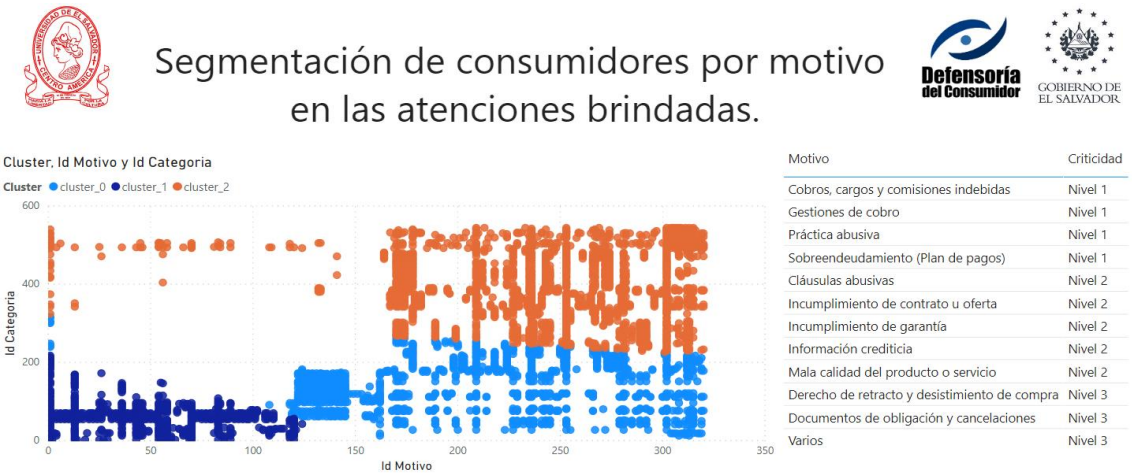
18.4 Reportes de minería de datos

Proveedor	Regla de asociación	Soporte	Confianza (%)	Lift
AES CAESS	MAY, NOVEMBER, JUNE	6.00	85.70	5,305
AES CLESA	MAY, OCTOBER, NOVEMBER	8.00	100.00	6,820
AES EEO	AUGUST, MAY, SEPTEMBER	8.00	100.00	5,968
ANDA	JANUARY, MARCH, FEBRUARY	10.00	100.00	4,774
BANCO ABANK	OCTOBER, NOVEMBER	8.00	100.00	5,570
BANCO AGRICOLA	JUNE, JULY	6.00	85.70	5,305
BANCO CUSCATLAN	FEBRUARY, JANUARY, MAY	7.00	100.00	6,820
BANCO DAVIVIENDA SALVADORENO	MARCH, JANUARY	7.00	100.00	5,570
BANCO DE AMERICA CENTRAL	JUNE, JULY	7.00	87.50	4,874
BANCO PROMERICA	FEBRUARY, SEPTEMBER	6.00	85.70	5,968
CLARO	JULY, OCTOBER, JANUARY	9.00	100.00	5,570
DIGICEL	NOVEMBER, OCTOBER	8.00	88.90	5,305
DISTRIBUIDORA DE ELECTRICIDAD DEL SUR	FEBRUARY, MAY	6.00	85.70	4,774
GRUPO MONGE	DECEMBER, JANUARY	8.00	100.00	3,713
GRUPO Q	NOVEMBER, OCTOBER	7.00	100.00	6,963
OMNISPORT	DECEMBER, JULY, NOVEMBER, OCTOBER, SEPTEMBER	10.00	100.00	7,957
SIMAN	OCTOBER, SEPTEMBER	6.00	85.70	5,305
TELEFONICA	FEBRUARY, JANUARY	7.00	100.00	3,979
TIGO	NOVEMBER, DECEMBER	6.00	85.70	7,957
UNICOMER	DECEMBER, JANUARY, NOVEMBER	8.00	100.00	6,189

Tomando como base proveedores que superan las 120 atenciones anuales se compara su promedio mensual por año con cada uno de los meses obteniendo los meses en los que historicamente se a observado un aumento en las denuncias, se presenta la columna asociación en la cual el ultimo mes para cada proveedor es producto de los aumentos de los meses anteriores.

Figura 194 Diseño Reporte de relación entre el aumento de las reclamaciones hacia proveedores respecto a los meses del año.

La Figura 194 muestra el reporte de la relación entre el aumento de las reclamaciones hacia proveedores respecto a los meses del año, este reporte de minería muestra las reglas de asociación obtenidas por medio de algoritmos de minería de datos en una tabla, así como los indicadores de rendimiento como el soporte, confianza y lift. Además, se agrega una descripción del reporte.



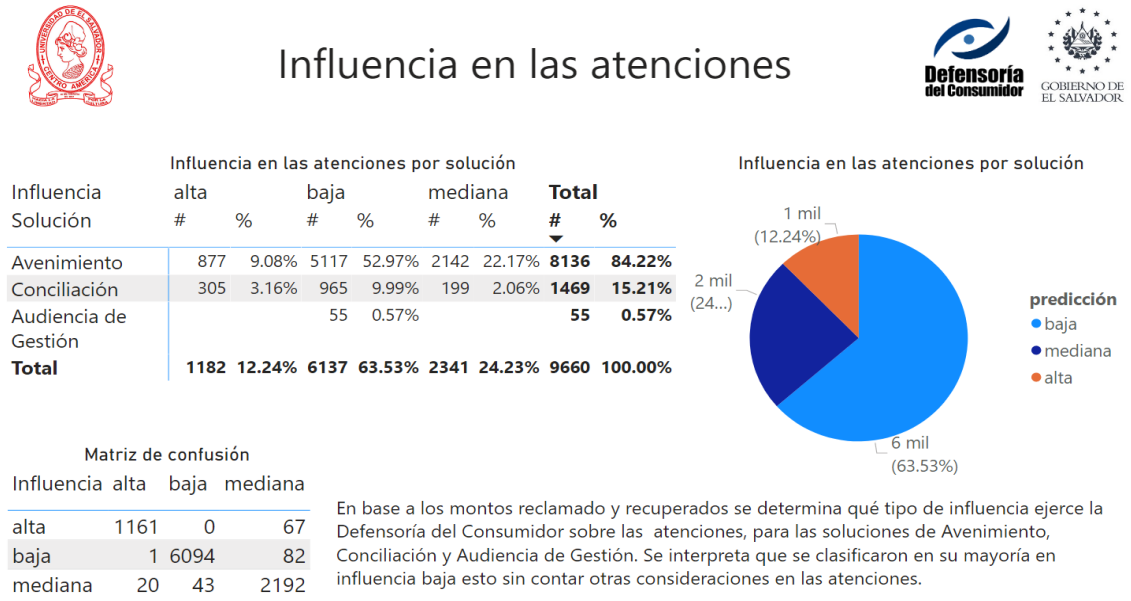
Motivo	Criticidad
Cobros, cargos y comisiones indebidas	Nivel 1
Gestiones de cobro	Nivel 1
Práctica abusiva	Nivel 1
Sobreendeudamiento (Plan de pagos)	Nivel 1
Cláusulas abusivas	Nivel 2
Incumplimiento de contrato u oferta	Nivel 2
Incumplimiento de garantía	Nivel 2
Información crediticia	Nivel 2
Mala calidad del producto o servicio	Nivel 2
Derecho de retracto y desistimiento de compra	Nivel 3
Documentos de obligación y cancelaciones	Nivel 3
Varios	Nivel 3

Se agrupan las atenciones brindadas a los consumidores en base al motivo, sector, categoría, forma de recepción, tipo de caso, municipio, genero de tal manera que se le puedan brindar programas de educación en consumo a estos grupos de personas.

Figura 195 Diseño del reporte segmentación de consumidores en base a los motivos.

La Figura 195 presenta el reporte de segmentación de consumidores en base a los motivos. Este reporte muestra los grupos o cluster que se generan a partir del algoritmo de

clusterización por medio de un gráfico de dispersión y coloreado por cluster. También, se muestra una tabla con una lista de motivos que en un primer momento se tomó como parámetro para agrupar las atenciones brindadas.



Influencia	alta		baja		mediana		Total	
Solución	#	%	#	%	#	%	#	%
Avenimiento	877	9.08%	5117	52.97%	2142	22.17%	**8136**	**84.22%**
Conciliación	305	3.16%	965	9.99%	199	2.06%	**1469**	**15.21%**
Audiencia de Gestión			55	0.57%			**55**	**0.57%**
Total	**1182**	**12.24%**	**6137**	**63.53%**	**2341**	**24.23%**	**9660**	**100.00%**

Influencia	alta	baja	mediana
alta	1161	0	67
baja	1	6094	82
mediana	20	43	2192

Figura 196 Diseño del reporte de Influencia en las atenciones.

La Figura 196 muestra el reporte de influencia en las atenciones. La primera tabla muestra la influencia que ha tenido la Defensoría del Consumidor, la segunda tabla muestra la matriz de confusión, que es un indicador de rendimiento para los algoritmos de clasificación. Por último, el gráfico de pastel muestra los porcentajes totales según su influencia.

Figura 197 Diseño informe de casos a recibir en fechas futuras por tipo y oficina.

La Figura 197 muestra el reporte de casos a recibir en fechas futuras por tipo y oficina. Este reporte muestra un gráfico de líneas una representa el pronóstico generado mediante series

temporales. El pronóstico se realiza para aproximadamente 8 semanas a partir de la última fecha en la que se tiene datos.

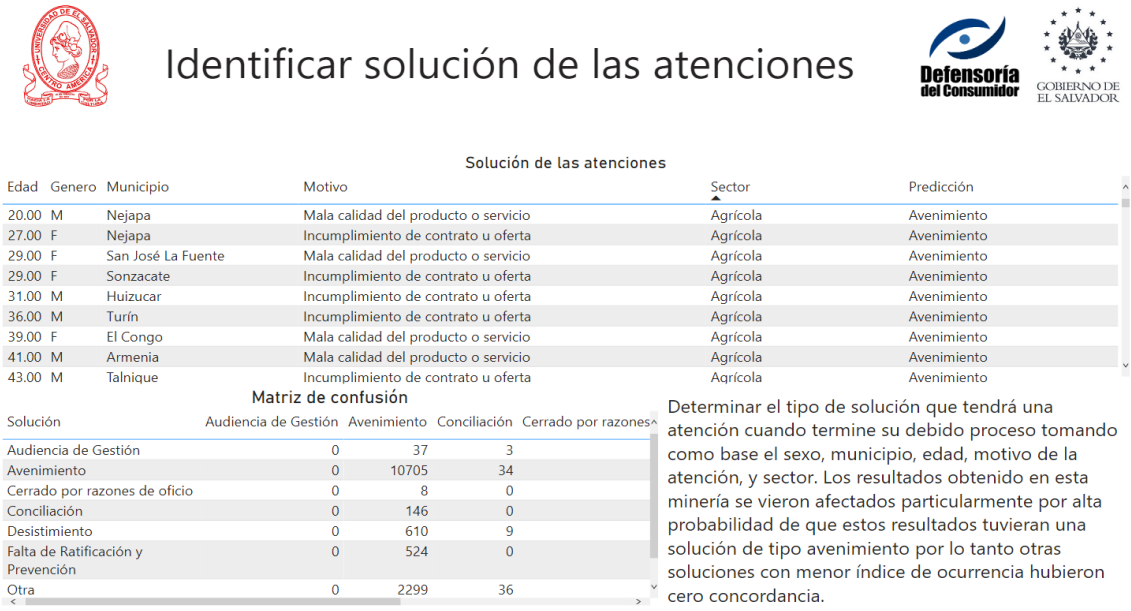


Figura 198 Diseño del reporte para identificar solución de las atenciones.

La Figura 198 muestra el reporte para identificar la solución que tendrán las atenciones. En la primera tabla se muestra los datos obtenidos y en la segunda la matriz de confusión, que es un indicador de rendimiento para el algoritmo de clasificación.

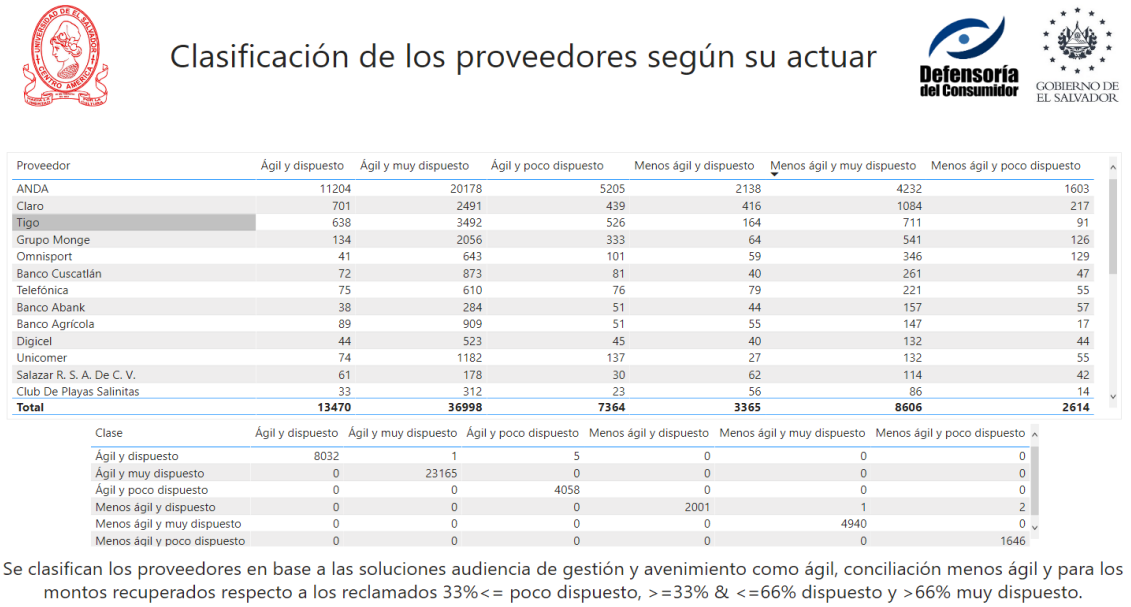


Figura 199 Diseño del reporte de clasificación de los proveedores según su actuar.

La Figura 199 muestra el reporte de la clasificación de los proveedores según su actuar. Nos muestra la forma de actuar antes las reclamaciones. Los proveedores y la clasificación se muestran mediante una tabla. Además de la matriz de confusión como su indicador de rendimiento.

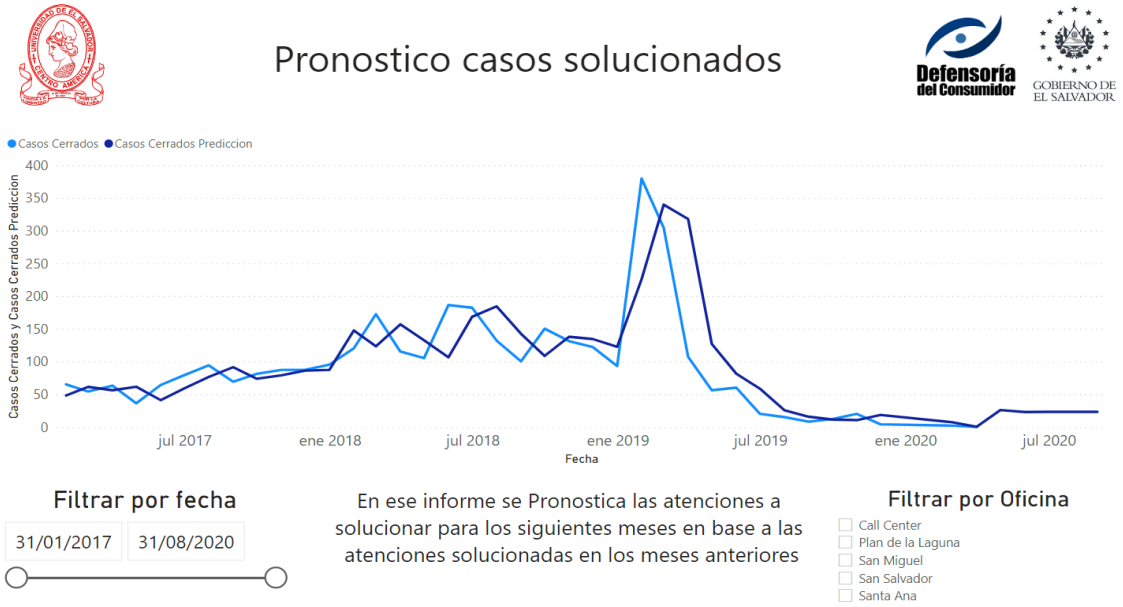


Figura 200 Diseño del reporte de Pronósticos atenciones cerradas.

La Figura 200 muestra el reporte del pronóstico de casos solucionados. Este reporte muestra un gráfico de líneas dónde se muestra un pronóstico generado mediante técnicas de minería de datos como son las series temporales. El pronóstico se realiza para aproximadamente 8 semanas a partir de la última fecha en la que se tiene datos.

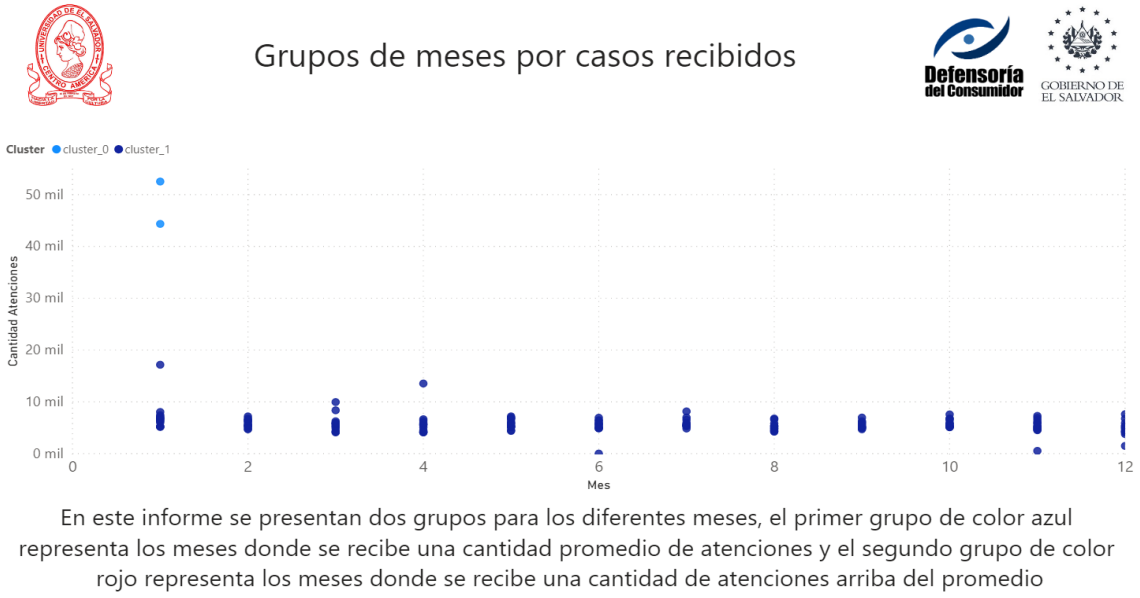


Figura 201 Diseño del reporte de segmentación de meses.

La Figura 201 presenta el reporte de segmentación de meses por casos recibidos. Este reporte nos muestra mediante un gráfico de dispersión el agrupamiento realizado mediante minería de datos de los meses que registran más casos.

19 Entregables

Entregable	Descripción
Archivos de Talend Open Studio	Archivos que contiene las migraciones de datos para el staging area y Modelo Multidimensional.
Exploración de Minería de Datos	Contiene los Workflows de Knime.
Informes de Minería de Datos e Inteligencia de Negocios	Visualizaciones de Minería de Datos e informes para la UACM
Manual de Usuario	Manual que explica el uso de los informes de Power BI
Manual Técnico	Manual donde se explican aspectos técnicos sobre los ETLs, Minería de Datos e Informes de Power BI
Manual de Instalación	Manual donde se explica la instalación de la solución y de las diferentes herramientas como PostgresSQL, Talend Open Studio Power BI etc.

Tabla 97 Entregables del Proyecto

20 Conclusiones

- Identificar los orígenes de datos es una parte fundamental de los procesos ETL y al utilizar herramientas especializadas en esta tarea permite de manera rápida y eficaz tener la información de la organización de forma centralizada facilitando así el acceso a la información.

- Los Datawarehouse o modelos multidimensionales son de gran importancia para las organizaciones ya que nos permite integrar y depurar distintos orígenes de información en un solo componente para que posteriormente esta información sea analizada desde diferentes puntos de vista.

- La visualización de información es una parte importante en las organizaciones ya que a partir de estas permite a las diferentes áreas realizar análisis de los datos, además de ayudar a la efectiva toma de decisiones, asimismo, utilizar una herramienta especializada de visualización nos permite mejorar la calidad de las visualizaciones y nos brinda la capacidad de poder visualizarlos desde diferentes dispositivos.

- A medida que las organizaciones crecen y su información se aumenta con los años se vuelve cada día más difícil analizar por eso en los últimos años la minería de datos ha ganado bastante terreno ya que permite explorar y analizar los datos en búsqueda de patrones, correlaciones y predicciones empleando una amplia variedad de técnicas en diferentes ámbitos de la organización.

21 Recomendaciones

- Se recomienda para la carga completa del modelo multidimensional deshabilitar los índices de las tablas para mejorar el rendimiento al momento de cargar la información.

- Se recomienda mantener actualizados los diferentes catálogos configurables de campos derivados, equivalentes, agrupados que no forman parte de los datos de origen para un correcto funcionamiento del modelo multidimensional.

- Se recomienda hacer una inversión en la herramienta de visualización ya que está permite publicar informes y trabajarlos de forma colaborativa permitiendo así mejorar los tiempos en las actividades del personal.

- Se recomienda promover cursos, seminarios y congresos orientados a impulsar las tecnologías para aumentar el uso de la Minería de datos.

- Todas las entidades públicas deben de reconocer las ventajas e importancia en la aplicación de la minería de datos para la obtención de conocimiento para la toma de decisiones.

22 Glosario

A

Accuracy (Exactitud): Se refiere a la cercanía de las mediciones a un valor específico.

Agrupación Difusa: Es una clase de algoritmos de agrupamiento donde cada elemento tiene un grado de pertenencia difuso a los grupos.

Agrupamiento: Es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares.

Algoritmo: Conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un tipo de problemas.

Almacén de Datos: Es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

Análisis de Supervivencia: Es una técnica inferencial que tiene como objetivo esencial modelizar el tiempo que se tarda en que ocurra un determinado suceso.

Antimonotónica: Es una restricción C tal que, si un conjunto de elementos S satisface a C, entonces cualquier subconjunto de S también satisface a C.

ARIMA: Es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro.

Association Rule Learner (Borgelt): Nodo de Knime que implementa el algoritmo A priori.

Azure: Es una creciente colección de servicios en la nube para crear, implementar y administrar aplicaciones inteligentes.

B

Backends: La parte que se encarga del acceso a los datos.

Base de Datos: Es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

Big data: Término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día.

Business Intelligence: Es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

C

Clave Subrogada: Es un campo numérico de una tabla cuyo único requisito es almacenar un valor numérico único para cada fila de la tabla, actuando como una clave sustituta.

Clustering: Técnica de la minería de datos que consiste en agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo (llamado clúster) sean más similares, en algún sentido u otro.

Código Abierto: Es el software cuyo código fuente y otros derechos que normalmente son exclusivos para quienes poseen los derechos de autor, son publicados bajo una licencia de código abierto o forman parte del dominio público.

Coeficiente Kappa: Refleja la concordancia inter-observador y puede ser calculado en tablas de cualquier dimensión, siempre y cuando se contrasten dos observadores, puede tomar valores entre -1 y +1. Mientras más cercano a +1, mayor es el grado de concordancia inter-observador, por el contrario, mientras más cercano a -1, mayor es el grado de discordancia inter-observador. Un valor de K = 0 refleja que la concordancia observada es precisamente la que se espera a causa exclusivamente del azar.

Curvas ROC: Presentan la sensibilidad de una prueba diagnóstica que produce resultados continuos, en función de los falsos positivos.

D

Data Integration: Es el proceso que implica combinar datos desde distintas fuentes en una única visión unificada.

Data Mart: Es un almacén de datos orientado a un área específica, como, por ejemplo, Ventas, Recursos Humanos u otros sectores.

Data Mining: Es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Data Preparation: Es el proceso de limpieza y transformación de datos para su uso posterior.

Data Quality: Hace referencia a una percepción o una evaluación de la idoneidad de los datos para cumplir su propósito en un contexto dado.

Data Science (Ciencia De Datos): Es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados

Data Warehouse: Es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

Dataset: Es una colección de datos habitualmente tabulada

Diagrama BPMN: Es una notación gráfica estandarizada que permite el modelado de procesos de negocio, en un formato de flujo de trabajo

Diagrama Conceptual: Identifica las relaciones de más alto nivel entre las diferentes entidades.

Diagrama de Componentes: Representa cómo un sistema de software es dividido en componentes y muestra las dependencias entre estos componentes.

Diagrama de Dominio: Es un artefacto de la disciplina de análisis, construido con las reglas de UML durante la fase de concepción, en la tarea construcción del modelo de dominio, presentado como uno o más diagramas de clases y que contiene, no conceptos propios de un sistema de software sino de la propia realidad física.

Diagrama de Despliegue: Diagrama que muestra las relaciones físicas de los distintos nodos que componen un sistema y el reparto de los componentes sobre dichos nodos.

Diagramas Multidimensionales: Extensión de UML propuesta por Sergio Luján-Mora y Juan Trujillo.

Diccionario de Datos: Es un listado organizado de todos los datos pertinentes al sistema, con definiciones precisas y rigurosas.

Diccionario de Datos: Es un repositorio centralizado de información sobre datos tales como significado, relación con otros datos, origen, uso y formato.

Dimensión Lentamente Cambiante: Son Dimensiones en las cuales sus datos tienden a modificarse a través del tiempo, ya sea de forma ocasional o constante.

Discretización: Es el proceso de transferir funciones continuas, modelos, variables y ecuaciones a contrapartes discretas.

Diseño Físico Base de Datos: Es un proceso que forma parte diseño de bases de datos, y su resultado final es un esquema físico de la base de datos. Es una descripción de la implementación de una base de datos en memoria secundaria.

Dispersión: Sirve como indicador de la variabilidad de los datos.

Distribución Normal: Es la distribución continua que se utiliza más comúnmente en estadística.

E

Eclipse: Software compuesto por un conjunto de herramientas de programación de código abierto multiplataforma para desarrolla.

Ecuación Lineal: Igualdad en la que intervienen términos acompañados de una variable con exponente uno.

Entropía (o medida del desorden): Mide la incertidumbre de una fuente de información

Eliminación lógica: Es aquella que ocurre al activar una marca de "eliminado" al registro.

Equiespaciados: Conserva una misma distancia entre un elemento y otro.

Esquema Estrella: Un *esquema de estrella* es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central rodeada de tablas de dimensiones.

Esquema Estrella: Es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central rodeada de tablas de dimensiones.

Estacionalidad (Seasonal): Es la variación periódica y predecible de la misma con un periodo inferior o igual a un año

ETL: Extract, Transform and Load («extraer, transformar y cargar», frecuentemente abreviado ETL) es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar.

ETS: Es un método de pronóstico de series de tiempo para datos univariados que se puede extender para respaldar datos con una tendencia sistemática o un componente estacional.

F

Forecast: Consiste en la estimación y previsión de la demanda futura de un producto o servicio

Fuente de Datos: Información de diversos tipos de documentos que contienen datos útiles para satisfacer una demanda de información o conocimiento.

Fuerza de Correlación: Técnica estadística que nos indica si dos variables están relacionadas o no.

G

GNU: Es un sistema operativo de tipo Unix, así como una gran colección de programas informáticos que componen al sistema.

Gráficos de Dispersión: Es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.

Homogenización: Hacer que los elementos de un conjunto o los miembros de un grupo tengan una serie de características iguales o uniformes.

I

Indicadores de Rendimiento: Son instrumentos esenciales para medir el resultado para así poder evaluar el grado de credibilidad que le asignamos a cualquier decisión que dependa de la minería de datos.

Intervalo de Confianza: Un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido con un determinado nivel de confianza.

Item: Unidad u objeto de un conjunto.

Itemsets Frecuentes: Conjunto de Objetos que suelen aparecen juntos en las transacciones.

J

Job (Talend Open Studio): Es un espacio gráfico, de uno o más componentes conectados entre sí, que le permite configurar y ejecutar procesos de gestión de flujo de datos.

L

Lenguajes de Programación: Están formados por un conjunto de símbolos, reglas gramaticales (léxico/morfológicas y sintácticas) y semánticas.

Linear Modeling (Modelo lineal): Los modelos lineales describen una variable de respuesta continua en función de una o más variables predictoras. Pueden ayudarlo a comprender y predecir el comportamiento de sistemas complejos o analizar datos experimentales, financieros y biológicos.

Log: Es un archivo de texto en el que constan cronológicamente los acontecimientos que han ido afectando a un sistema informático.

M

Machine Learning: Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente.

Máquinas de Vectores de Soporte: Son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T.

Matrices de Clasificación: Ordena todos los casos del modelo en categorías, determinando si el valor de predicción coincide con el valor real.

Matriz de Confusión: Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.

Measure o Medida: Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.

Modelo Relacional: Consistente en el almacenamiento de datos en tablas compuestas por filas, o tuplas, y columnas o campos.

N

Normalización: Es un proceso que consiste en designar y aplicar una serie de reglas a las relaciones obtenidas tras el paso del modelo entidad-relación al modelo relacional.

Normalización min-max: Transformación lineal de todos los valores, de modo que el mínimo y el máximo en cada columna sean los indicados.

Normalización Z-score: Transformación lineal de modo que los valores en cada columna estén distribuidos en Gauss (0,1), es decir, la media es 0.0 y la desviación estándar es 1.0.

Normalización por escala decimal: El valor máximo en una columna (tanto positivo como negativo) se divide j-veces por 10 hasta que su valor absoluto sea menor o igual a 1. Todos los valores en la columna se dividen por 10 a la potencia de j.

O

Outliers: Es una observación que es numéricamente distante del resto de los datos.

P

Parámetros: Es cualquier característica que pueda ayudar a definir o clasificar un sistema particular, Es decir, es un elemento de un sistema que es útil o crítico al identificar el sistema o al evaluar su rendimiento, estado, condición, etc.

Pivot: Permiten intercambiar los resultados de filas por columnas (referencias cruzadas).

Pivote: Presente información en un informe con referencias cruzadas con formato de planilla de cálculo a partir de cualquier tabla relacional usando código SQL sencillo, y almacene datos de una tabla con referencias cruzadas en una tabla relacional.

Postgresql: Es un sistema de gestión de bases de datos relacional orientado a objetos y de código abierto.

Prueba de Box-Ljung: Es utilizada para comprobar si una serie de observaciones en un período de tiempo específico son aleatorias e independientes. Si las observaciones no son independientes, una observación puede estar correlacionada con otra observación *k* unidades de tiempo después, una relación que se denomina autocorrelación.

Pruebas de Independencia: Se utiliza cuando se tiene una muestra de n individuos que se clasifican respecto a dos variables, preferentemente cualitativas (nominales dicotómicas o politómicas) y se desea conocer a partir de datos muestrales, si existe asociación de estas a nivel poblacional.

R

Regresión Logística: Es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras.

S

Serie Temporal: Es una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados.

SQL: Es un lenguaje de dominio específico utilizado en programación, diseñado para administrar, y recuperar información de sistemas de gestión de bases de datos relacionales. Sus siglas en inglés Structured Query Language; en español lenguaje de consulta estructurada.

Sql Server: Es un sistema de gestión de base de datos relacional, desarrollado por la empresa Microsoft. Cuenta con un lenguaje de desarrollo llamado Transact-SQL.

SPRS: Sistema de Recolección de Precios de Supermercados.

Staging Área Es un área intermedia de almacenamiento de datos utilizada para el procesamiento de los mismos durante procesos de ETL. Esta área se encuentra entre la fuente de los datos y su destino, que a menudo son almacenes de datos, data marts u otros repositorios de datos.

Stitch Data Loader: Producto de Talend que simplifica la carga de datos a partir de decenas de fuentes en cloud a almacenes de datos y data lakes acelerando los procesos de desarrollo.

T

Tabla de Dimensión: Proporciona una forma uniforme de mantener una versión actualizada de los datos asociados con entidades.

Tabla de Hechos: Son claves externas que apuntan a las claves primarias para entradas relacionadas en tablas de dimensiones.

Tendencia: Es la dirección o rumbo del mercado. Pero es importante entender que los mercados no se mueven en línea recta en ninguna dirección. Los movimientos en los precios se caracterizan por un movimiento zigzagueante.

Teorema de Bayes: Asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

Tests Estadísticos: Son el instrumento para validar o rechazar las hipótesis de modelación probabilistas. Ellos tratan de distinguir lo que es plausible de lo que es muy poco verosímil, en el marco de un modelo dado.

Transacción: Es una interacción compuesta por varios procesos que se han de aplicar uno después del otro. La transacción debe realizarse de una sola vez.

Transformada de Fourier: Es una transformación de una señal que está dada por una función de x, que nos permite calcular la contribución de cada valor de frecuencia a la formación de la señal. El sistema separador de Fourier que es capaz de expresar una serie temporal en la suma de ondas.

TSLM: Es un método de pronóstico de series de tiempo que se utiliza para ajustar modelos lineales a series temporales que incluyen componentes de tendencia y estacionalidad.

U

UML: Es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad; está respaldado por el Object Management Group (OMG). Sus siglas en inglés Unified Modeling Language (Lenguaje Unificado de Modelado). Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema.

Unpivot: Permite transformar datos que están dispuestos como columnas en filas.

V

Variable Dependiente: Representa una cantidad cuyo valor depende de cómo se modifica la variable independiente.

Variable Independiente o Explicativas: Es una variable que representa una cantidad que se modifica en un experimento.

23 Referencia Bibliográfica

Amat Rodrigo, J. (2018, Junio). *Reglas de asociación y algoritmo Apriori con R*. Retrieved from cienciadedatos.net: https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion

Balanced Insight, Inc. (2000, Junio 14). *Information Package Design.* (B. Insight, Ed.) Retrieved from Information Package Design: https://www.balancedinsight.com/wp-content/files/BIWhitepaper_InformationPackageDesign.pdf

Berrios, C. D. (2014, Julio 1). *APLICACIÓN DE ÁRBOLES DE DECISIÓN PARA LA ESTIMACIÓN DEL ESCENARIO ECONÓMICO Y LA ESTIMACIÓN DE MOVIMIENTO LA TASA DE INTERÉS EN CHILE*. Retrieved from http://repositorio.uchile.cl: http://repositorio.uchile.cl/bitstream/handle/2250/117556/Dupouy%20Berrios%20Carlos.pdf?sequence=1&isAllowed=y

Bouza, C. N. (2014, Marzo 11). *LA MINERÍA DE DATOS: ARBOLES DE DECISIÓN Y SU APLICACIÓN EN ESTUDIOS MÉDICOS*. Retrieved from researchgate.net: https://www.researchgate.net/publication/268516570_LA_MINERIA_DE_DATOS_ARBOLES_DE_DECISION_Y_SU_APLICACION_EN_ESTUDIOS_MEDICOS

Cascos, I. (n.d.). Regresión lineal múltiple. Madrid: Universidad Carlos III.

colaboradores de Wikipedia. (2019, Diciembre 25). *KNIME*. (L. e. Wikipedia, Editor) Retrieved from Wikipedia: https://es.wikipedia.org/w/index.php?title=KNIME&oldid=122269532

colaboradores de Wikipedia. (2019, Noviembre 8). *Power BI*. Retrieved from Wikipedia: https://es.wikipedia.org/w/index.php?title=Power_BI&oldid=121173163

colaboradores de Wikipedia. (2020, Enero 11). *R (lenguaje de programación)*. Retrieved from Wikipedia: https://es.wikipedia.org/w/index.php?title=R_(lenguaje_de_programaci%C3%B3n)&oldid=122663781

de la Fuente Fernández, S. (2011). Regresión Múltiple. Madrid: Universidad Autonoma de Madrid.

EcuRed. (n.d.). *EcuRed*. Retrieved from https://www.ecured.cu/Minería_de_Datos

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques (Third Edition).* Morgan Kaufmann.

Logicalis. (2014, Septiembre 7). *Blog Logicalis.* Retrieved from https://blog.es.logicalis.com/analytics/mineria-de-datos-aplicaciones-que-ya-son-una-realidad

Luján-Mora, S., & Trujillo, J. (2003, 12 14). Un método global basado en UML para el diseño. Alicante, España: Universidad de Alicante.

McCulloch, W., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, 115-133.

Monteserin, A. (2018, Abril 13). *Reglas de asociación.* Retrieved from La Universidad Nacional del Centro de la Provincia de Buenos Aires: http://www.exa.unicen.edu.ar/catedras/optia/public_html/2018 Reglas de asociación.pdf

Profe. (2016, Abril 12). *Mi Profe*. Retrieved from https://miprofe.com/minimos-cuadrados/

Talend Inc. (2020). *Talend*. Retrieved from Talend | Productos: https://es.talend.com/products/talend-open-studio/

24 Anexos

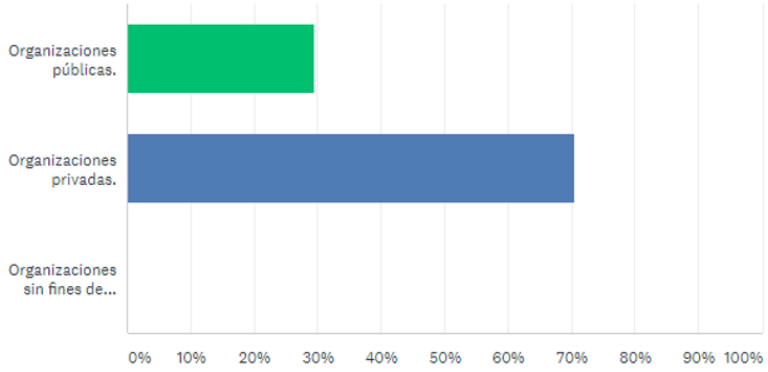
24.1 Anexo 1: Encuesta sobre Desarrollo de Proyectos de Exploración de la Información para la Generación de Nuevo Conocimiento

La herramienta de recolección de datos se realizó en una plataforma web llamada SurveyMonkey. Esta se compartió con una muestra de expertos en el tema de minería de datos en El Salvador y se obtuvieron los siguientes resultados.

Objetivo: Determinar la importancia de la minería de datos y su impacto en los negocios adicionalmente explorar el grado de conocimiento que poseen los profesionales en minería de datos de las diferentes organizaciones del sector público y privado.

Indicaciones: Conteste a su conocimiento las siguientes preguntas.

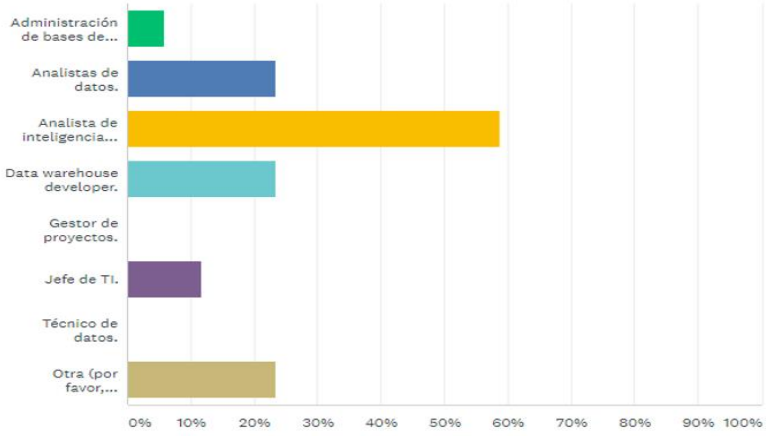
1. ¿Cuál de las siguientes organizaciones describe donde labora?



OPCIONES DE RESPUESTA	RESPUESTAS	
Organizaciones públicas.	29,41%	5
Organizaciones privadas.	70,59%	12
Organizaciones sin fines de lucro, Autónomas o Semiautónomas.	0,00%	0
TOTAL		17

Figura 202: Pregunta 1.

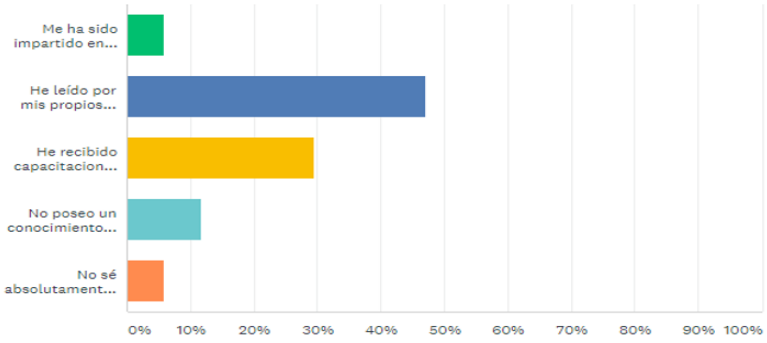
2. ¿Cuál de las siguientes opciones describe mejor su trabajo?



OPCIONES DE RESPUESTA		RESPUESTAS	
Administración de bases de datos.		5,88%	1
Analistas de datos.		23,53%	4
Analista de inteligencia de negocios.		58,82%	10
Data warehouse developer.		23,53%	4
Gestor de proyectos.		0,00%	0
Jefe de TI.		11,76%	2
Técnico de datos.		0,00%	0
Otra (por favor, especifique)	Respuestas	23,53%	4
Total de encuestados: 17			

Figura 203: Pregunta 2.

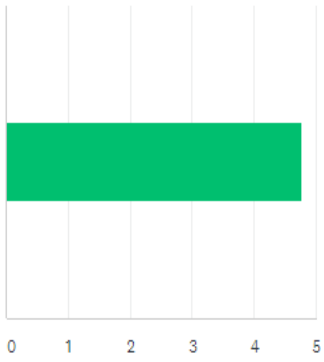
3. Si posee conocimientos de minería de datos, ¿de cuál de las siguientes fuentes los ha obtenido?



OPCIONES DE RESPUESTA	RESPUESTAS	
Me ha sido impartido en la Universidad.	5,88%	1
He leído por mis propios medios, (Libros, papers, noticias, artículos etc).	47,06%	8
He recibido capacitaciones, certificaciones, diplomados al respecto.	29,41%	5
No poseo un conocimiento formal, pero tengo la noción de lo que es.	11,76%	2
No sé absolutamente nada al respecto.	5,88%	1
TOTAL		**17**

Figura 204: Pregunta 3.

4. En la escala de 1 a 5, ¿considera usted que la minería de datos puede crear valor agregado a la toma de decisiones de las empresas? (Siendo 1 muy bajo y 5 muy alto)

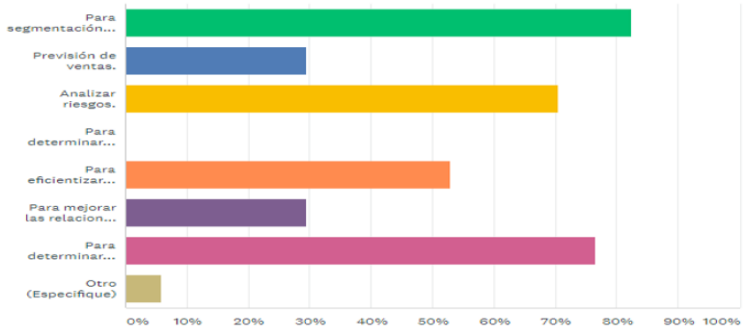


OPCIONES DE RESPUESTA	CANTIDAD PROMEDIO	CANTIDAD TOTAL	RESPUESTAS
Respuestas	5	81	17

Total de encuestados: 17

Figura 205: Pregunta 4.

5. ¿Cuál cree usted que es el principal o los principales motivos por el cual las empresas acuden a utilizar minería de datos? (Puede seleccionar una opción o más).

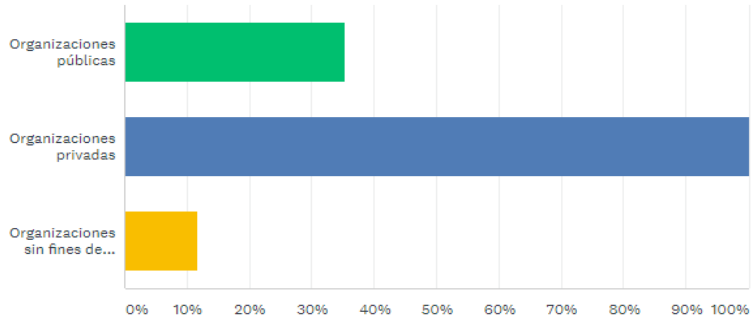


OPCIONES DE RESPUESTA		RESPUESTAS	
Para segmentación de mercado de clientes.		82,35%	14
Previsión de ventas.		29,41%	5
Analizar riesgos.		70,59%	12
Para determinar fallas de seguridad.		0,00%	0
Para eficientizar la toma de decisiones.		52,94%	9
Para mejorar las relaciones existentes en la cadena de valor del negocio.		29,41%	5
Para determinar patrones de comportamientos en los clientes.		76,47%	13
Otro (Especifique)	Respuestas	5,88%	1

Total de encuestados: 17

Figura 206: Pregunta 5.

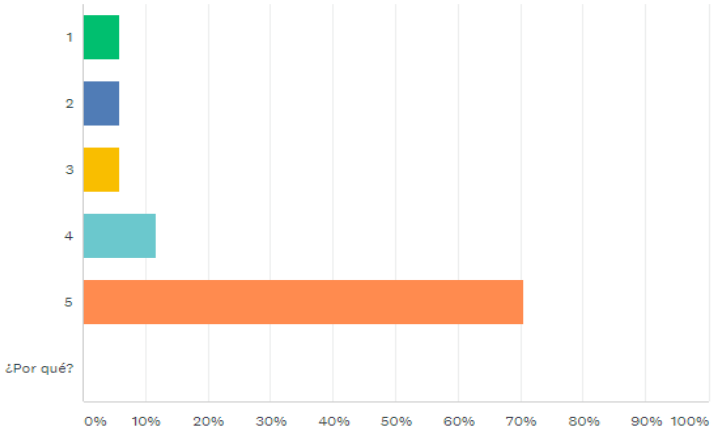
6. ¿Qué tipos de organizaciones considera usted que optan por implementar minería de datos con mayor frecuencia? (Puede seleccionar una opción o más).



OPCIONES DE RESPUESTA	RESPUESTAS	
Organizaciones públicas	35,29%	6
Organizaciones privadas	100,00%	17
Organizaciones sin fines de lucro, Autónomas o Semiautónomas.	11,76%	2
Total de encuestados: 17		

Figura 207: Pregunta 6.

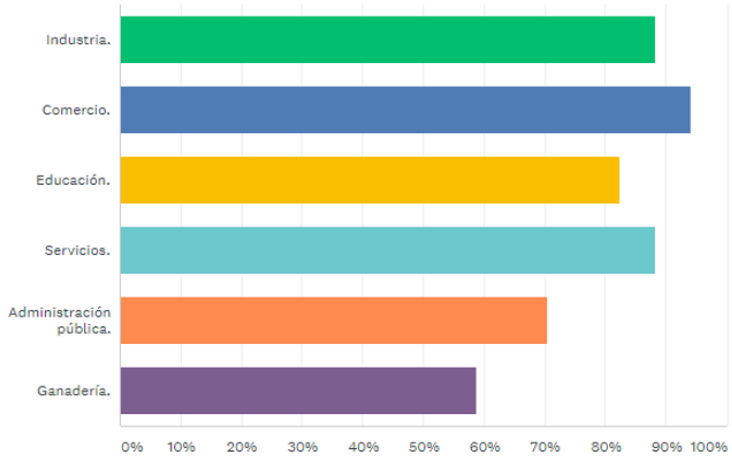
7. ¿Cuánto considera usted que la minería de datos sería de utilidad en los ministerios públicos, instituciones autónomas y semiautónomas? (Siendo 1 muy bajo y 5 muy alto)



OPCIONES DE RESPUESTA		RESPUESTAS	
1		5,88%	1
2		5,88%	1
3		5,88%	1
4		11,76%	2
5		70,59%	12
¿Por qué?	Respuestas	0,00%	0
TOTAL			**17**

Figura 208: Pregunta 7.

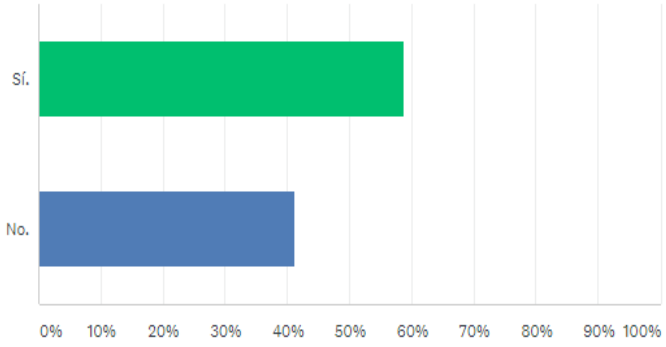
8. ¿En qué sectores considera usted que tiene aplicabilidad la minería de datos? (Puede seleccionar una opción o más).



OPCIONES DE RESPUESTA	RESPUESTAS	
Industria.	88,24%	15
Comercio.	94,12%	16
Educación.	82,35%	14
Servicios.	88,24%	15
Administración pública.	70,59%	12
Ganadería.	58,82%	10
Total de encuestados: 17		

Figura 209: Pregunta 8.

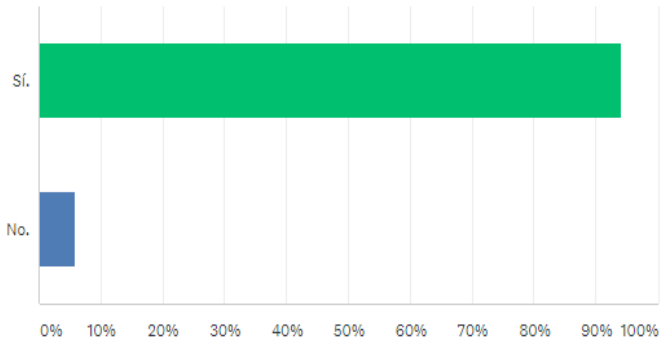
9. ¿Conoce alguna organización, ya sea pública, privada o sin fines de lucro, que haya implementado algún proyecto de minería de datos en el país?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	58,82%	10
No.	41,18%	7
TOTAL		17

Figura 210: Pregunta 9.

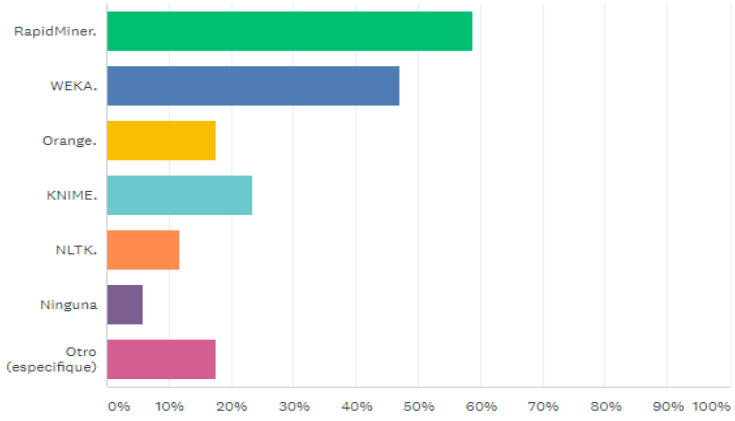
10. A su criterio, ¿La minería de datos se vuelve una parte vital en las organizaciones para ser competitivas en el mercado?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	94,12%	16
No.	5,88%	1
TOTAL		17

Figura 211: Pregunta 10.

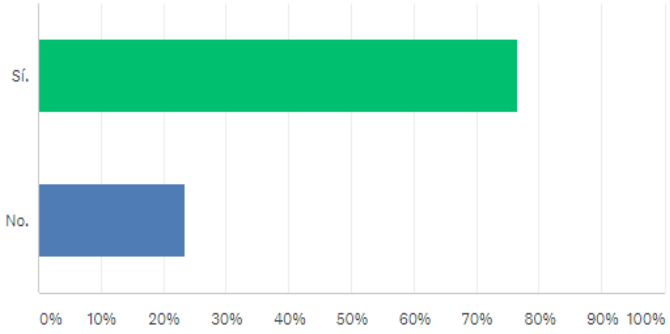
11. ¿Conoce alguna de las siguientes herramientas de minería de datos? (Puede seleccionar una opción o más).



OPCIONES DE RESPUESTA		RESPUESTAS	
RapidMiner.		58,82%	10
WEKA.		47,06%	8
Orange.		17,65%	3
KNIME.		23,53%	4
NLTK.		11,76%	2
Ninguna		5,88%	1
Otro (especifique)	Respuestas	17,65%	3
Total de encuestados: 17			

Figura 212: Pregunta 11.

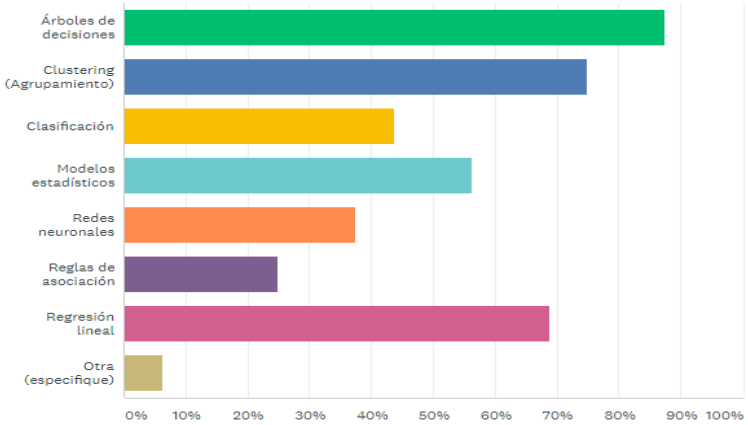
12. ¿Conoce usted alguna técnica que se utilice en la minería de datos?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	76,47%	13
No.	23,53%	4
TOTAL		17

Figura 213: Pregunta 12.

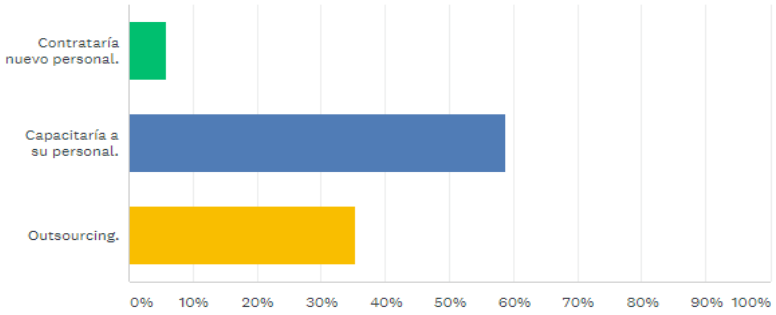
13. ¿Cuáles de estas técnicas conoce?



OPCIONES DE RESPUESTA		RESPUESTAS	
Árboles de decisiones		87,50%	14
Clustering (Agrupamiento)		75,00%	12
Clasificación		43,75%	7
Modelos estadísticos		56,25%	9
Redes neuronales		37,50%	6
Reglas de asociación		25,00%	4
Regresión lineal		68,75%	11
Otra (especifique)	Respuestas	6,25%	1
Total de encuestados: 16			

Figura 214: Pregunta 13.

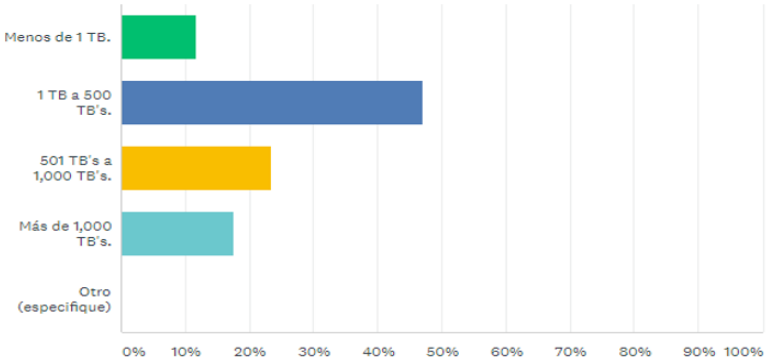
14. Si su empresa deseara implementar un proyecto de minería de datos, ¿cómo lo haría?



OPCIONES DE RESPUESTA	RESPUESTAS	
Contrataría nuevo personal.	5,88%	1
Capacitaría a su personal.	58,82%	10
Outsourcing.	35,29%	6
TOTAL		17

Figura 215: Pregunta 14.

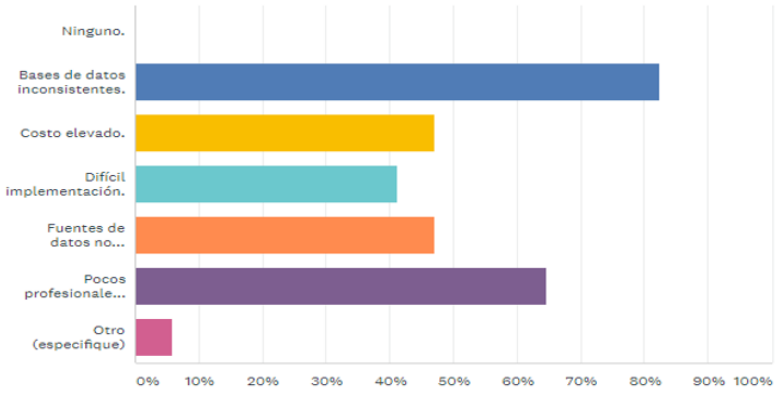
15. ¿De qué tamaño (GB's o TB's) considera que es la base de datos de los sistemas que tiene su organización?



OPCIONES DE RESPUESTA		RESPUESTAS	
Menos de 1 TB.		11,76%	2
1 TB a 500 TB's.		47,06%	8
501 TB's a 1,000 TB's.		23,53%	4
Más de 1,000 TB's.		17,65%	3
Otro (especifique)	Respuestas	0,00%	0
TOTAL			17

Figura 216: Pregunta 15.

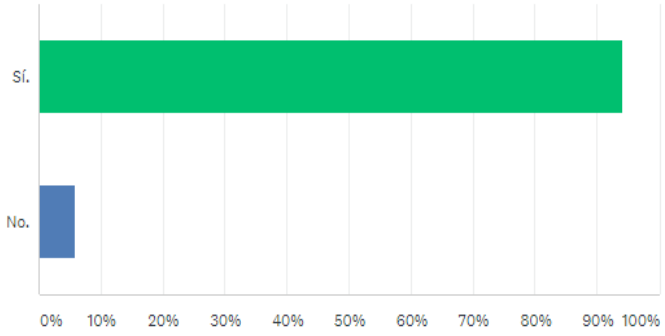
16. ¿Cuál considera que es el desafío principal al momento de implementar un proyecto de minería de datos?



OPCIONES DE RESPUESTA		RESPUESTAS	
Ninguno.		0,00%	0
Bases de datos inconsistentes.		82,35%	14
Costo elevado.		47,06%	8
Difícil implementación.		41,18%	7
Fuentes de datos no integrados.		47,06%	8
Pocos profesionales en el área.		64,71%	11
Otro (especifique)	Respuestas	5,88%	1
Total de encuestados: 17			

Figura 217: Pregunta 16.

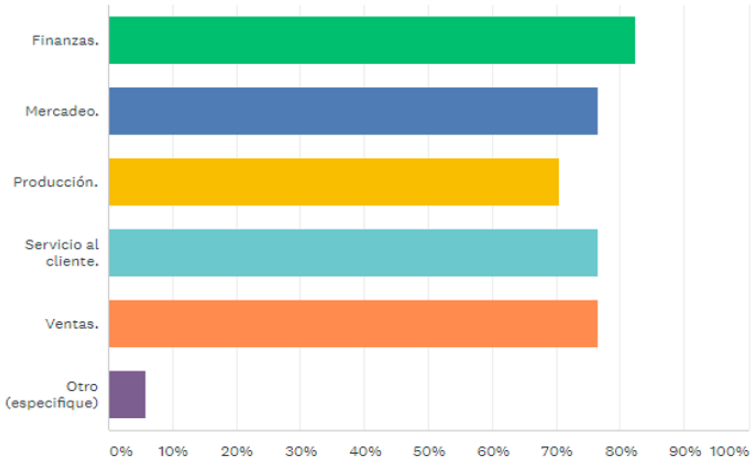
17. ¿Recomendaría el uso de minería de datos a otras organizaciones?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	94,12%	16
No.	5,88%	1
TOTAL		17

Figura 218: Pregunta 17.

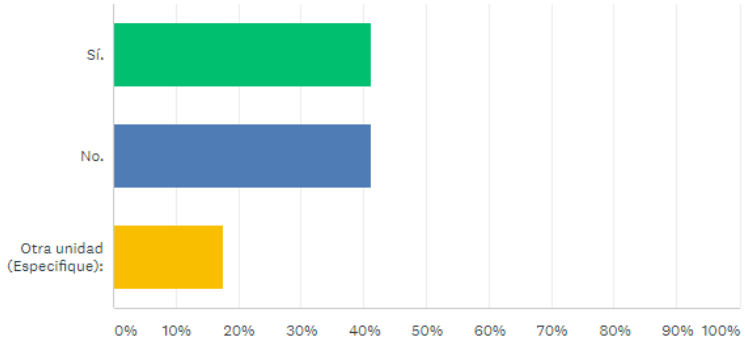
18. ¿Cuál es el área de la organización que cuenta con el mayor beneficio al hacer uso de la minería de datos? (Puede seleccionar una opción o más).



OPCIONES DE RESPUESTA		RESPUESTAS	
Finanzas.		82,35%	14
Mercadeo.		76,47%	13
Producción.		70,59%	12
Servicio al cliente.		76,47%	13
Ventas.		76,47%	13
Otro (especifique)	Respuestas	5,88%	1
Total de encuestados: 17			

Figura 219: Pregunta 18.

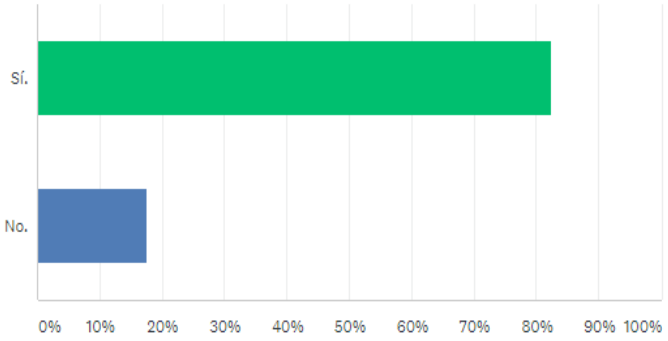
19. Si se ha implementado un proyecto de minería de datos, ¿es la unidad de informática la encargada de dicho proceso?



OPCIONES DE RESPUESTA		RESPUESTAS	
Sí.		41,18%	7
No.		41,18%	7
Otra unidad (Especifique):	Respuestas	17,65%	3
TOTAL			**17**

Figura 220: Pregunta 19.

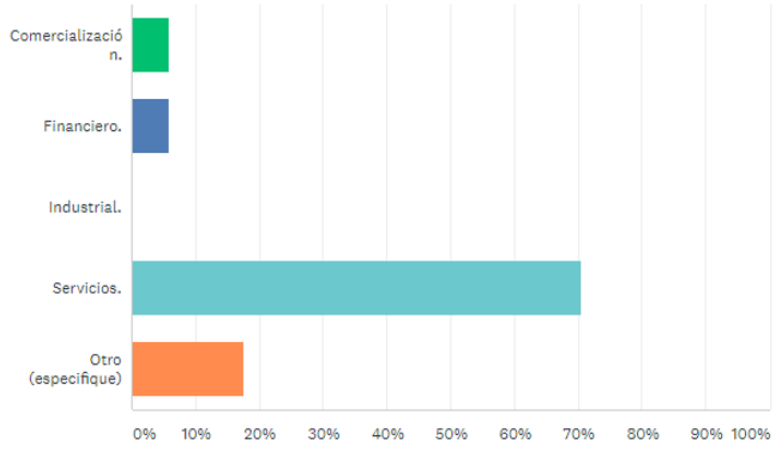
20. ¿Cuenta su organización con personas capacitadas en el área de minería de datos?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	82,35%	14
No.	17,65%	3
TOTAL		17

Figura 221: Pregunta 20.

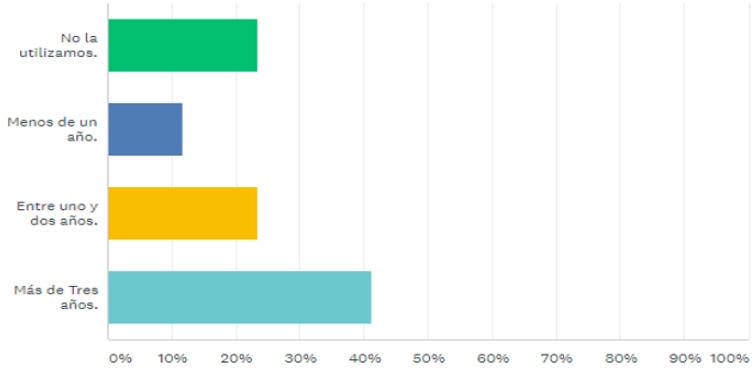
21. ¿Cuál es el rubro de su organización?



OPCIONES DE RESPUESTA		RESPUESTAS	
Comercialización.		5,88%	1
Financiero.		5,88%	1
Industrial.		0,00%	0
Servicios.		70,59%	12
Otro (especifique)	Respuestas	17,65%	3
TOTAL			17

Figura 222: Pregunta 21.

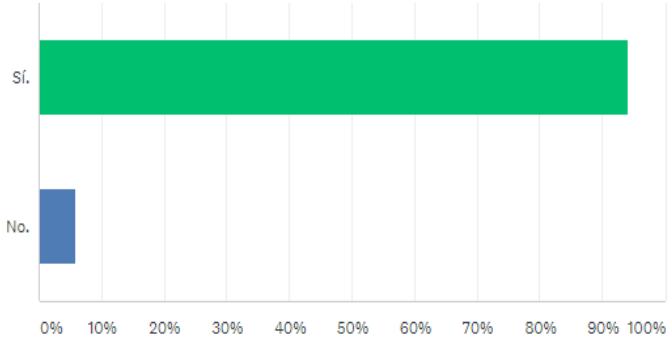
22. ¿Cuánto tiempo lleva utilizando minería de datos su organización?



OPCIONES DE RESPUESTA	RESPUESTAS	
No la utilizamos.	23,53%	4
Menos de un año.	11,76%	2
Entre uno y dos años.	23,53%	4
Más de Tres años.	41,18%	7
TOTAL		17

Figura 223: Pregunta 22.

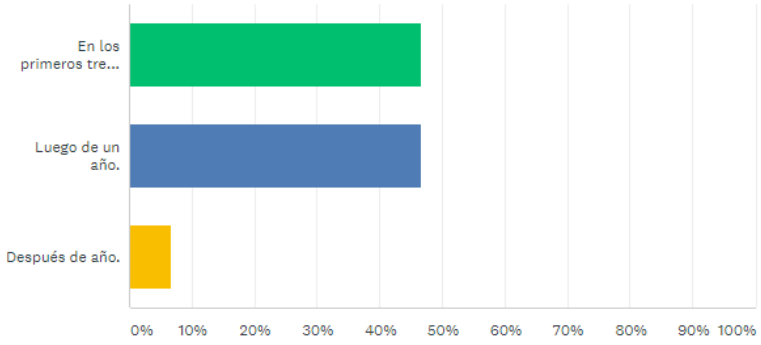
23. ¿Considera usted que la minería de datos facilita la toma de decisiones en su organización?



OPCIONES DE RESPUESTA	RESPUESTAS	
Sí.	94,12%	16
No.	5,88%	1
TOTAL		17

Figura 224: Pregunta 23.

24. ¿Después de cuánto tiempo de usar minería de datos se obtuvieron beneficios?



OPCIONES DE RESPUESTA	RESPUESTAS	
En los primeros tres meses.	46,67%	7
Luego de un año.	46,67%	7
Después de año.	6,67%	1
TOTAL		15

Figura 225: Pregunta 24.

24.2 Anexo 2: Definición, Evaluación y Selección de la Herramienta de Minería de Datos

RapidMiner

Conocido inicialmente como YALE, es un software de minería de datos gratuito distribuido bajo licencia GPL e implementado en Java. Desarrollada por la Universidad de Dortmund en el año 2001. Proporciona más de 500 operadores orientados al análisis de datos. Produce sus resultados en archivos XML y cuentan con la interface gráfica del mismo programa. Existiendo tres módulos: Rapid Miner Studio, Rapid Miner Server, Rapid Miner Radoop. Puede usarse a través de una interfaz gráfica, línea de comandos, batch o incluso desde otros programas a través de llamadas a sus bibliotecas. Incluye gráficos y herramientas para la visualización de los datos. Soporta el lenguaje de programación R.

IBM SPSS Modeler

Es una aplicación de software de análisis de texto y minería de datos. Nace en el año 1994 por ISL luego fue adquirido por parte de IBM en el año 2009. Se utiliza para construir modelos predictivos y realizar otras tareas analíticas. Cuenta con una interfaz visual que permite a los usuarios aprovechar los algoritmos estadísticos y de minería de datos sin programación. Es multiplataforma plataformas y distribuido bajo licencia propietaria. Permite el uso de R, Python, Spark y Hadoop para amplificar el poder de la analítica.

KNIME

Konstanz Information Miner (KNIME) es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. Está construido bajo la plataforma Eclipse. Desarrollado por la Universidad de Constanza. Es una herramienta escrita en su mayor parte en Java por lo cual la convierte en multiplataforma. Es una herramienta de software libre pero la empresa brinda soporte en varios niveles. Es altamente funcional a nivel gráfica, pero además permite el uso de script de lenguajes de programación como lo es Python, R, Groovy y Matlab.

SAS Data Mining

Statistical Analysis System (SAS) es un producto de SAS Institute desarrollado para el análisis y la gestión de datos. SAS puede extraer datos, alterarlos, administrar datos de diferentes fuentes y realizar análisis estadísticos. Distribuida bajo licencia propietaria, tiene una arquitectura de procesamiento de memoria distribuida que es altamente escalable. Es adecuado para la minería de datos, la minería de textos y la optimización. Hace uso de la estadística, machine learning e inteligencia artificial.

H2O

Es un software open-source para el análisis de Big Data. Desarrollada por H2O.ai en el año 2011. Permite tener acceso a lenguajes como Python, Scala, R y otros. Permite a la vez entradas como sistemas distribuidos como Hadoop, bases de datos SQL, entre otras. La herramienta está escrita en mayor parte en Java, Python y R. El software se puede ejecutar en las plataformas más populares y debido a que explora datos almacenados en la nube, brinda un servicio web siendo compatible con los navegadores Chrome, Safari, Firefox e Internet Explorer. Se puede obtener soporte técnico por parte de H2O.ai.

Cuadro Comparativo

Criterios	Peso	RapidMiner		IBM SPSS Modeler		KNIME		SAS Data Mining		H2O	
		Calificación	Puntuación	Calificación	Puntuación	Calificación	Puntuación	Calificación	Puntuación	Calificación	Puntuación
Performance (Rendimiento)	0,30										
Plataforma	0,10	3,00	0,30	3,00	0,30	3,00	0,30	3,00	0,30	3,00	0,30
Arquitectura	0,05	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15
Acceso a datos	0,20	3,00	0,60	1,00	0,20	3,00	0,60	3,00	0,60	3,00	0,60
Tamaño de datos	0,30	3,00	0,90	3,00	0,90	4,00	1,20	4,00	1,20	3,00	0,90
Eficiencia	0,15	3,00	0,45	3,00	0,45	3,00	0,45	3,00	0,45	3,00	0,45
Robustez	0,20	3,00	0,60	4,00	0,80	3,00	0,60	3,00	0,60	3,00	0,60
Puntuacion performance		3,00		2,80		3,30		3,30		3,00	
Funcionalidad	0,20										
Algoritmos	0,20	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60
Metodologías	0,05	3,00	0,15	3,00	0,15	3,00	0,15	5,00	0,25	3,00	0,15
Modelos	0,05	3,00	0,15	4,00	0,20	3,00	0,15	4,00	0,20	3,00	0,15
Variedad de tipos de datos	0,20	3,00	0,60	2,00	0,40	3,00	0,60	3,00	0,60	3,00	0,60
Modificabilidad de algoritmos	0,05	3,00	0,15	2,00	0,10	3,00	0,15	3,00	0,15	4,00	0,20
Muestreo de datos	0,05	3,00	0,15	3,00	0,15	4,00	0,20	3,00	0,15	3,00	0,15
Informes	0,20	3,00	0,60	4,00	0,80	3,00	0,60	3,00	0,60	3,00	0,60
Exportar modelos	0,20	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60	2,00	0,40
Puntuacion funcionalidad		3,00		3,00		3,05		3,15		2,85	
Usabilidad	0,20										
Interfaz de usuario	0,20	3,00	0,60	3,00	0,60	4,00	0,80	3,00	0,60	3,00	0,60
Curva de aprendizaje	0,20	3,00	0,60	1,00	0,20	3,00	0,60	1,00	0,20	2,00	0,40
Visualización de datos	0,20	3,00	0,60	4,00	0,80	4,00	0,80	3,00	0,60	2,00	0,40
Reporte de errores	0,20	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60
Historial de acciones	0,15	3,00	0,45	3,00	0,45	3,00	0,45	3,00	0,45	3,00	0,45
Tipos de negocios	0,05	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15
Puntuacion usubilidad		3,00		2,80		3,40		2,60		2,60	
Apoyo a las tareas	0,10										
Limpieza de datos	0,50	3,00	1,50	3,00	1,50	3,00	1,50	3,00	1,50	3,00	1,50
Sustituir valores	0,20	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60
Filtrado de datos	0,20	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60	3,00	0,60
Aleotorización	0,05	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15
Manipulación de metadatos	0,05	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15	3,00	0,15
Puntuacion apoyo a las tareas		3,00		3,00		3,00		3,00		3,00	
Otros	0,20										
Costo	0,75	3,00	2,25	1,00	0,75	3,00	2,25	1	0,75	3,00	2,25
Soporte técnico	0,25	3,00	0,75	5,00	1,25	3,00	0,75	5	1,25	3,00	0,75
Puntuacion otros		3,00		2,00		3,00		2,00		3,00	
Peso promedio		3,00		2,70		3,18		2,84		2,89	

Tabla 98: Tabla comparativa de las 5 herramientas seleccionadas en base al cuadrante mágico de Gartner, tomando como referencia el formato de "A methodology for evaluating and selecting data mining software.