UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS

ESCUELA DE MATEMÁTICA

TRABAJO DE GRADO TITULADO:

ESTUDIO E IDENTIFICACIÓN DE VARIABLES QUE DETERMINAN LA CLUSTERIZACIÓN DE CLIENTES-APLICACIÓN A DATOS REALES

PRESENTADO POR:

FRANKLIN IVÁN ARGUETA BERMÚDEZ

PARA OPTAR AL TITULO DE:

LICENCIADO EN MATEMÁTICA

CIUDAD UNIVERSITARIA, 12 DE DICIEMBRE DE 2019

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS

ESCUELA DE MATEMÁTICA

TRABAJO DE GRADO TITULADO:

ESTUDIO E IDENTIFICACIÓN DE VARIABLES QUE DETERMINAN LA CLUSTERIZACIÓN DE CLIENTES-APLICACIÓN A DATOS REALES

PRESENTADO POR:

FRANKLIN IVÁN ARGUETA BERMÚDEZ

PARA OPTAR AL TITULO DE:

LICENCIADO EN MATEMÁTICA

ASESOR:

M.SC. WALTER OTONIEL CAMPOS GRANADOS

CIUDAD UNIVERSITARIA, 12 DE DICIEMBRE DE 2019

AUTORIDADES DE LA UNIVERSIDAD DE EL SALVADOR

RECTOR:
MSC. ROGER ARMANDO ARIAS ALVARADO

VICERECTOR ACADÉMICO:
PHD. RAÚL ERNESTO AZCÚNAGA LÓPEZ

VICERECTOR ADMINISTRATIVO:
ING. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:
ING. FRANCISCO ALARCÓN

AUTORIDADES DE LA FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS

DECANO:
LIC. MAURICIO HERNÁN LOVO CÓRDOVA

VICEDECANO:
M.SC. ZOILA GUERRERO

SECRETARIA:
LIC. DAMARYS MELANY HERRERA TURCIOS

AUTORIDADES DE LA ESCUELA DE MATEMÁTICA

DIRECTOR:
M.SC. PORFIRIO ARMANDO RODRÍGUEZ

SECRETARIO:
M-SC. CARLOS ERNESTO GÁMEZ RODRÍGUEZ

DEDICATORIA

A mi madre que me apoyo todos estos años de estudio, realizando un gran esfuerzo por sacarnos adelante haciéndome sentir orgulloso de ser su hijo, te amo y no va haber manera de devolverte tanto que me has ofrecido; no sé en donde me encontraría de no ser por tus ayudas, tu compañía, y tu amor.

Índice General

Página

Índice de Figuras

Índice de Tablas

RESUMEN

El análisis de conglomerados por variables es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión, como el análisis de componentes principales. La idea es construir una matriz de distancias o similitudes entre variables y aplicar a esta matriz un algoritmo jerárquico de clasificación con el objetivo de agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos, para luego aplicar los métodos de clasificación basado en árboles de decisión o bosques aleatorios los cuales estratifican o segmentan el espacio del predictor en un número simple de regiones; para ello, se parte del registro histórico de los datos reales proporcionados por cierta institución dichos datos están compuesto por las observaciones de los clientes de los últimos 4 años (2015 - 2018) y se busca a través del estudio identificar las variables que determinan la clusterización de clientes, analizando la influencia de las variables mediante dos enfoques importantes para la generación de modelos. Lo que se pretende con los modelos es desarrollar una estrategia de negocios, la cual consiste en generar movilidad positiva para cada uno de los clientes, mejorando su clasificación básica predefinida.

INTRODUCCIÓN

El análisis de conglomerados es una técnica de análisis exploratorio, definido dentro de los métodos multivariantes de clasificación, que permite separar en diferentes clases o grupos a un conjunto de objetos o individuos, de modo que todos los que pertenecen a una misma clase son homogéneos entre si y diferentes de aquellos objetos que pertenecen a una clase distinta.

La presente investigación aborda esta técnica para estudiar e identificar las variables que determinan la clusterización de clientes de una institución financiera, donde cada variable representa los productos financieros, actividad económica, periodos de tiempo, ciclos de rentabilidad, transacciones, entre otros servicios financieros.

En la investigación se parte de una clasificación básica, que la institución financiera ha realizado sobre sus clientes, ya que los ha separado por su tamaño corporativo en Clientes Tipo #1 y Clientes Tipo #2. La clusterización de estos clientes se realizó con el propósito de agrupar a los clientes que sean lo más homogéneos entre si. Esto permitió identificar las variables que influyen en la formación de cada clúster, lo que viene resultando en identificar los servicios que ofrece la institución, los cuales determinan la clusterización de sus clientes. Analizar e identificar estas variables involucro aplicar métodos de aprendizaje no supervisado.

Durante la investigación se hace un estudio teórico de los métodos tradicionales del análisis multivariante en cuanto a segmentación de poblaciones de interés y se trabaja con datos históricos reales proporcionados por cierta institución financiera (que nos ha solicitado el anonimato y el cuido celoso de los datos); donde cada variable está de forma enmascarada, y solamente se conoce el tipo de variable (*cualitativa* o *cuantitativa*), siendo esto un obstáculo a la hora de formar y completar las Bases de Datos de manera que solamente se puede añadir a cada cliente las variables que se encuentran relacionadas a él a través de su código de referencia sin tener la libertad de completar algún dato faltante durante los cuatro años de registro que se tiene a disposición.

En el primer capítulo se realizo el planteamiento de la investigación la cual busca identificar las variables que determinan el pertenecer a un determinado clúster y desarrollar modelos de clasificación de clientes a partir de los datos históricos que se posean, con el propósito de indicar los servicios financieros que han hecho crecer a un cliente dentro de la institución en el transcurso del tiempo.

En el segundo capítulo se desarrolla la investigación teórica del Análisis de Multivariante, el cual involucra el estudio del Conglomerados por variable que trata sobre el análisis estadístico de forma unidimensional, bidimensional y multidimensional, así como el análisis en componentes principales buscando reducir la dimensión de los datos para observar la formación de los clúster de forma gráfica, el análisis de conglomerados que involucra los métodos clásicos y modernos de partición de los datos y los métodos predictivos a utilizar como lo son árboles de decisión y bosques aleatorios.

En el tercer capítulo se detalla el tipo de investigación a realizar, el diseño de la investigación mostrando cronológicamente el camino que se siguió para la culminación del estudio.

En el cuarto capítulo se muestran los resultados de la investigación, la cuál se realizó sobre dos Bases de Datos que son formadas por los Clientes de Tipo #1 y los de Tipo #2, los resultados que se muestran en el trabajo son los del estudio de la Base de Datos para los Clientes de Tipo #1. Los resultados del estudio de la Base de Datos para los Clientes de Tipo #2 no se muestran pero si se mencionan los resultados más relevantes identificados en la investigación.

En el quinto capítulo se detallan las conclusiones generales que se obtienen con los resultados de la investigación.

En el sexto capítulo se elaboran recomendaciones a tomar en cuenta para replicar está investigación para nuevas bases de datos de clientes y se detallan líneas a seguir para realizar futuros estudios tomando como base esta investigación.

Capítulo 1

PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1. Identificación del problema

En el presente trabajo se abordan aplicaciones de la minería de datos para la segmentación, clasificación y fidelización de clientes en el área de productos financieros.

En la presente investigación se trabaja con datos históricos reales proporcionados por cierta institución financiera (que nos ha solicitado el anonimato y el cuido celoso de los datos).

En este contexto se hace un estudio teórico de los métodos tradicionales del análisis multivariante en cuanto a segmentación y clasificación de poblaciones de interés.

Para la manipulación de los datos se usará el software estadístico libre R y las diferentes instancias de concatenación que éste tiene con los software utilizados para el manejo de base de datos.

En este trabajo se parte de una clasificación básica que la institución financiera ha hecho sobre sus clientes:

1. Clientes A,
2. Clientes M,
3. Clientes B,
4. Clientes I

Por lo que se busca a través del análisis de los datos históricos indicar cuales son las variables que determinan el pertenecer a un tipo de cliente de los anteriores y generar modelos que permitan identificar productos financieros, épocas del año, ciclos y tipos de abordaje propicios que se debe recomendar a los clientes para llevarles a una clase mejor a la que pertenecen.

1.2. Objetivos de la investigación

1.2.1. General

Desarrollar modelos de clasificación de clientes a partir del análisis de los datos históricos que se posean.

1.2.2. Específicos

- Aplicar técnicas de minería de datos como métodos de aprendizaje no supervisado para la clusterización de clientes, a través del análisis de los datos históricos que se posean.
- Determinar el mejor modelo que permita indicar que productos, época del año y el periodo de tiempo, son propicios para recomendar a los clientes que han sido clasificados fuera del clúster de valor alto.

1.3. Justificación

Es de suma importancia que la teoría matemática sea una herramienta que nos permita resolver los problemas a los que la sociedad se enfrenta cada día.

Con especial interés debe estudiarse los métodos de Análisis Multivariantes y las técnicas de Minería de Datos, específicamente en la segmentación de clientes a través de perfiles similares, para la generación de algoritmos de recomendación, de tal manera que estos modelos matemáticos - estadísticos permitan predecir el comportamiento de los clientes, dependiendo el clúster al que pertenezcan.

En este sentido, en esta investigación, trabajamos con datos reales de una institución financiera y se busca crear modelos que permitan formular recomendaciones para la fidelización, optimización y generación de clientes con valor alto para dicha institución.

Considerando que es de suma importancia que en la Escuela de Matemática se promueva una cultura de vinculación de la matemática con los sectores productivos del país, públicos o privados.

1.4. Importancia

La mejora de la calidad de los servicios financieros es uno de los retos más importantes que actualmente deben afrontar todos los agentes implicados en el ámbito económico y, en especial, los responsables de su dirección y gestión empresarial. El sistema financiero en nuestro país a sufrido cambios importantes, tanto en términos cuantitativos como cualitativos, ya que se ha tenido que hacer frente a una movilidad económica irregular, asi como la exigencia de servicios de calidad. Una de las maneras de evaluar la calidad del servicio lo constituye la percepción del cliente, el cual es componente importante de cada institución.

En el sistema financiero de nuestro país no se tienen modelos que permitan medir la importancia de los productos financieros tanto en la rentabilidad como en la aplicación de los servicios. Nuestro estudio pretende hacer un diagnostico en ambos aspectos para identificar las variables que intervienen en la actividad económica de los clientes, con el fin de que los analistas financieros, propongan cambios para enmendar las falencias encontradas.

El análisis de datos en el área financiera constituye uno de los elementos más importantes en la construcción de modelos de calidad, el cual es llevado a cabo mediante la manipulación de los datos reales, recopilados en un determinado periodo de tiempo.

Es de vital importancia para el crecimiento empresarial el lograr avances en las áreas que se desempeña y ofrece la institución. Sin embargo muchas veces la oferta financiera no es acorde a las demandas o necesidades del cliente. Por lo cual debe tomarse en cuenta las condiciones de los clientes para mejorar los productos financieros. De igual manera muchas instituciones se quejan de que reciben clientes en condiciones complicadas, como consecuencia de la dinámica económica del país. Por lo que es necesario hacer un diagnostico exhaustivo de la problemática en cuestión.

Mientras la instituciones financieras no mejoren sus modelos, seguirán obteniendo clientes con deficiencias económicas que a largo o mediano plazo las terminaran afectando. De ahí que nuestro estudio pretende dar un paso importante en la identificación de variables que permita crear una movilidad positiva en los clientes.

Capítulo 2

MARCO TEÓRICO

2.1. Antecedentes de estudio

Una de las actividades más primitivas, comunes y básicas del ser humano consiste en clasificar objetos. La clasificación o identificación es el proceso de asignar un nuevo objeto en su lugar correspondiente dentro de un conjunto de categorías establecidas.

A partir de la segunda mitad del Siglo XX se ha visto un aumento en las técnicas numéricas disponible para la clasificación. Este crecimiento ha ido paralelo con el desarrollo de los ordenadores, que son necesarios para poder realizar el gran número de operaciones que se precisan. Asimismo, un desarrollo similar ha tenido lugar en las áreas de aplicación. Como por ejemplo, en Biología se usa taxonomía numérica, en Inteligencia Artificial se usa la técnica de reconocimiento de patrones, entre otras.

El problema de la clasificación puede ser complicado debido a varios factores, como la presencia de clases definidas de forma imperfecta, la existencia de categorías solapadas y posibles variaciones aleatorias en las observaciones. Una forma de tratar estos problemas, desde el punto de vista estadístico, sería encontrar la probabilidad que tiene cada nueva observación de pertenecer a cada categoría. En este sentido, el criterio de clasificación más simple sería elegir la categoría más probable, mientras que pueden necesitarse reglas más sofisticadas si las categorías no son igualmente probables o si los costos de mala clasificación varían entre las categorías.

En la escuela de matemática se han realizado investigaciones orientadas en la aplicación directa de las herramientas matemáticas:

- Orellana Romero, José Luis (2012) *Modelación y pronóstico de la demanda de ener-*

gía eléctrica de mediano plazo de El Salvador. Tesis de Maestría, Universidad de El Salvador.

- Rosa Alvarado, Welman del Carmen (2011) *Modelo geoestadístico espacio-temporal del crimen en El Salvador: análisis estructural y predictivo*. Tesis de Maestría, Universidad de El Salvador.

- Rivas Morales, Milton Arnoldo (2017) *Estudio de la Geometría fractal con aplicaciones a finanzas y Vulcanología*. Tesis de Licenciatura, Universidad de El Salvador.

2.2. Conglomerados por variable

El análisis de conglomerados por variable es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión, como el análisis factorial o los métodos de correlación canónica y el análisis de componentes principales. La idea es construir una matriz de distancias o similitudes entre variables y aplicar a esta matriz un algoritmo jerárquico de clasificación.

Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión (análisis en componentes principales) tratan de eliminarla. Es decir, tratan de describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

Además el análisis en componentes principales permite pasar a un nuevo conjunto de variables, las componentes principales, que gozan de la ventaja de estar incorrelacionadas entre sí. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquélla que tenga mayor varianza, mientras que, por el contrario, la última es la de menor varianza.

En general, la extracción de componentes principales se efectúa sobre variables tipificadas para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en desviaciones respecto a la media. Si p variables están tipificadas, la suma de las varianzas es igual a p, ya que la varianza de una variable tipificada es por definición igual a 1. El nuevo conjunto de variables que se obtienen por el método de componentes principales, es igual en número al de variables originales. Cuando las variables originales están muy correlacionadas

entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. La suma de las varianzas de las variables (inercia total de la nube de puntos) es igual a la suma de las varianzas de las componentes principales e igual a la suma de los valores propios de la matriz de varianzas y covarianzas.

La aplicación del método de componentes principales puede abordarse desde tres perspectivas equivalentes:

1. *Enfoque descriptivo*: Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible.

2. *Enfoque estadístico*: Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno, que es equivalente a sustituir las p variables originales por una nueva variable z_1 que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión.

3. *Enfoque geométrico*: Los puntos se sitúan siguiendo una trayectoria o forma y se puede describir su orientación dando otra dirección y la posición de los puntos por su proyección sobre esta dirección, lo cual logra minimizar las distancias ortogonales.

2.2.1. Análisis estadístico

2.2.1.1. Análisis estadístico univariado

Es un resumen numérico y gráfico de la variable. Si la variable a analizar es cuantitativa, se medirán su tendencia central y su dispersión. Si $x_1, x_2, \ldots, x_n$ son las observaciones de la variable cuantitativa x y $p_1, p_2, \ldots, p_n$ son las ponderaciones de los individuos con $p_i = 1/n$, es usual denotar su media por $\bar{x}$:

$$\bar{x} = \sum_{i=1}^{n} p_i x_i \tag{2.1}$$

su desviación estándar σ_x:

$$\sigma_x = \sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} = \sqrt{\sum_{i=1}^{n} p_i x_i^2 - \bar{x}^2} \tag{2.2}$$

y su varianza $var(x)$:

$$var(x) = \sigma_x^2 \tag{2.3}$$

Si la variable a analizar es cualitativa o binaria, se calculan las frecuencias (absolutas y relativas) de cada categoría, y en caso de ser ordinal la variable, es usual calcular también las frecuencias acumuladas. Los principales gráficos asociados a una variable cuantitativa son generalmente los histogramas, las cajas de dispersión y los diagramas de tallo-hoja. En el caso de una variable cualitativa, se usan los gráficos de barras, circulares, que representan proporcionalmente a las frecuencias.

2.2.1.2. Análisis estadístico bivariado

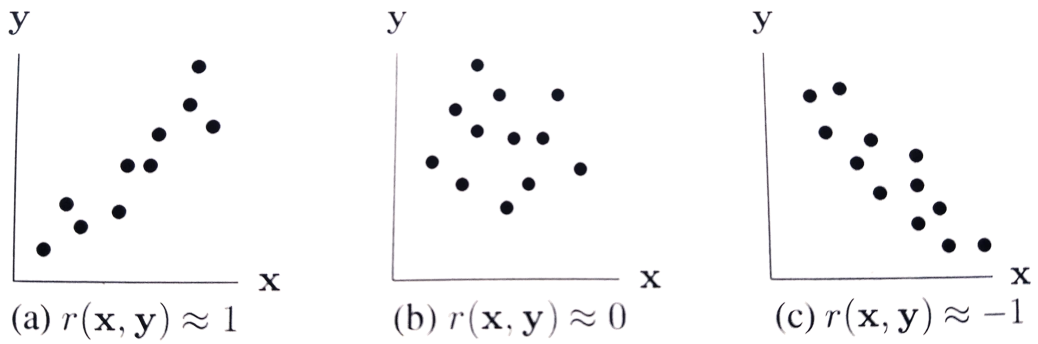
Consiste en el estudio de las relaciones entre parejas de variables, y también forma parte de la descripción simple de una tabla de datos.

En el caso de tener dos variables cuantitativas, se suele hacer el diagrama de dispersión, el cual gráfica en ejes de abscisas y de ordenadas a las dos variables, y permite ver la asociación entre ellas. El **coeficiente de correlación lineal** también llamado coeficiente de correlación de Pearson denotado por r, es una cuantificación de la relación entre dos variables cuantitativas x e y.

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{2.4}$$

donde la covarianza es:

$$cov(x,y) = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\bar{y} \tag{2.5}$$



(a) $r(\mathbf{x}, \mathbf{y}) \approx 1$ (b) $r(\mathbf{x}, \mathbf{y}) \approx 0$ (c) $r(\mathbf{x}, \mathbf{y}) \approx -1$

Figura 2.1: El coeficiente de correlación lineal muestra el tipo de relación entre dos variables cuantitativas

Si las dos variables son cualitativas, entonces se suele estudiar la independencia entre las categorías de las dos variables mediante un índice de asociación, que usualmente es el índice de chi-cuadrado (denotado χ^2).

2.2.1.3. Medidas de distancia y similitud entre variables

Las medidas habituales de asociación entre variables continuas son la covarianza y la correlación. Estas medidas tienen en cuenta únicamente las relaciones lineales. Así con el propósito de encontrar una clasificación de filas o de columnas, el primer problema a tomar en cuenta es cómo cuantificar la similitud entre objetos o entre grupos de objetos. Alternativamente, podríamos construir una medida de distancia entre dos variables x_j y x_h representando cada variable como un punto en $\mathbb{R}^n$ calculando la distancia euclídea entre los dos puntos. Esta medida es:

$$d_{jh}^2 = \sum_{i=1}^{n}(x_{ij} - x_{ih})^2 \tag{2.6}$$

$$= \sum x_{ij}^2 + \sum x_{ih}^2 - 2\sum x_{ij}x_{ih} \tag{2.7}$$

Para que la distancia no dependa de las unidades, las variables deben estar estandarizadas.

Cuando las variables tienen varianzas muy desiguales, la magnitud del término $(x_{ij} - x_{ih})^2$ puede depender de la varianza de la variable x_j, haciendo depender la distancia entre filas, de la estructura de varianzas más que de la estructura de correlaciones. Para corregir este efecto se usa la fórmula:

$$d_{jh}^2 = \sum_{i=1}^{n}\frac{1}{\sigma_j^2}(x_{ij} - x_{ih})^2 \tag{2.8}$$

Obsérvese que lo anterior equivale a dividir cada columna x_j por su desviación estándar σ_j y usar la distancia euclídea clásica sobre los datos así transformados.

En otro caso la distancia entre dos variables podría alterarse arbitrariamente mediante transformaciones lineales de éstas. (Por ejemplo, midiendo las estaturas en metros, en lugar de en cm. y en desviaciones respecto a la media poblacional en lugar de con carácter absoluto). Suponiendo, por tanto, que trabajamos con variables estandarizadas de media cero y varianza uno, se obtiene que la ecuación (2.6) se reduce a:

$$
\begin{aligned}
d_{jh}^2 &= \sum_{i=1}^{n}(x_{ij}-x_{ih})^2 \\
&= \sum x_{ij}^2 + \sum x_{ih}^2 - 2\sum x_{ij}x_{ih} \\
&= n + n - 2\sum x_{ij}x_{ih} \\
&= 2n - 2nr_{jh} \\
d_{jh}^2 &= 2n(1-r_{jh}) \qquad (2.9)
\end{aligned}
$$

Observemos que:

(a) si $r_{jh} = 1$, la distancia es cero, indicando que las dos variables son idénticas.

(b) si $r_{jh} = 0$, las dos variables están incorreladas y la distancia es $d_{jh} = \sqrt{2n}$.

(c) si $r_{jh} < 0$, las dos variables tienen correlación negativa, y la distancia tomará su valor máximo, $\sqrt{4n}$, cuando las dos variables tengan correlación -1.

Para variables cualitativas binarias se puede construir una medida de similitud construyendo *una tabla de asociación entre variables*.

2.2.2. Estadística Multidimensional

Cuando se dispone de muchas observaciones para cada individuo, los análisis univariados y bivariados tiene la limitación de contemplar las interrelaciones entre todas las variables y cómo esas interrelaciones afectan al conjunto de individuos. El objetivo de las técnicas multivariadas o multidimensionales es el de proveer descripciones de esas interrelaciones, tomando las variables en su conjunto. Tales descripciones son hechas, en la visión del Análisis de Datos, a partir de representaciones geométricas, para las que se usa como principal herramienta el Álgebra Lineal.

2.2.2.1. Los espacios vectoriales asociados a las tablas de datos

Se tiene una matriz de datos X con n individuos y p variables. La i-ésima fila de X se denota x_i y se ve que está representada por el vector de p dimensiones. Si bien es cierto que x_i es una fila de la matriz X, como vector se representa como una columna.

$$
x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}
$$

Entonces x_i pertenece al espacio vectorial de $\mathbb{R}^p$. Por ello, $\mathbb{R}^p$ se llama el **espacio de los individuos**.

Por otro lado, a cada variable observada le corresponde una columna de X. La j-ésima columna se denota x^j y está representada por el vector de n dimensiones:

$$x^i = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

Entonces x^j está en el espacio vectorial $\mathbb{R}^n$. Por esta razón, a $\mathbb{R}^n$ se le llama el **espacio de variables**. Una **distancia** sobre el espacio $\mathbb{R}^p$ es una aplicación $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ tal que:

1. $d(x, x) = 0$ para todo x en $\mathbb{R}^p$.
2. $d(x, y) = d(y, x)$, para todo x, y en $\mathbb{R}^p$
3. $d(x, z) \leq d(x, y) + d(y, z)$, para todo x, y, z en $\mathbb{R}^p$

Un caso típico de distancia es la *distancia Euclídea clásica*:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \tag{2.10}$$

donde x y y son dos elementos de $\mathbb{R}^p$ con p componentes.

La distancia Euclídea clásica puede formularse según el siguiente producto matricial:

$$Si\, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix},$$

Entonces $d(x, y) = \sqrt{(x-y)^t (x-y)}$. Esto es, $d(x, y) = \|x - y\|$, donde $\|\cdot\|$ denota la norma matricial clásica, y también se puede escribir $d(x, y) = \sqrt{(x-y)^t I_p (x-y)} = \|x - y\|_{I_p}$ donde I_p es la matriz identidad $p \times p$ con elementos en $\mathbb{R}$

Una norma en un espacio vectorial es una aplicación de $\mathbb{R}^p$ en $\mathbb{R}^+$, denotada $\|\cdot\|$, tal que

a) $\|x\| = 0 \Leftrightarrow x = 0$,

b) para todo vector x y todo $\lambda \in \mathbb{R} : \|\lambda x\| = |\lambda| \|x\|$,

c) para cualesquiera vectores $x, y : \|x + y\| \leq \|x\| + \|y\|$

Los conceptos de norma y distancia Euclídea se puede generalizar para otro tipo de matrices. Sea M una matriz simétrica, definida positiva de dimensiones $p \times p$. Entonces el producto matricial $x^t Mx$ permite definir una norma sobre $\mathbb{R}^p$, que se denotará $\|\cdot\|_M$, así:

$$\|x\|_M = \sqrt{x^t Mx} \tag{2.11}$$

Se dice que M es:

a) Simétrica: $x^t My = y^t Mx$ para cualesquiera par de vectores p-dimensionales x, y

b) Definida: $x^t Mx = 0 \Leftrightarrow x = 0$

c) Positiva: $\forall x : \quad x^t Mx \geq 0$

Se llama **métrica** sobre $\mathbb{R}^p$ a una matriz $p \times p$ que sea simétrica, definida positiva.

Una métrica define un *producto interno* sobre $\mathbb{R}^p$: $\langle x, y \rangle_M = x^t My$, que es una función bilineal, simétrica, definida positiva.

Se dice que una función es bilineal cuando es lineal en ambos argumentos de la función $\langle \cdot, \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$

Si x y y son dos vectores del espacio entonces el coseno del ángulo θ que forman se puede determinar a partir de:

$$\cos\theta = \frac{\langle x, y \rangle_M}{\|x\|_M \|y\|_M} \tag{2.12}$$

Se dirá que dos vectores x, y son *ortogonales* si $\langle x, y \rangle = 0$

2.2.2.2. Métricas de pesos en $\mathbb{R}^n$

Sobre le espacio de variables $\mathbb{R}^n$ se define una métrica sobre la proximidad entre las variables, se trata entonces de una matriz de orden $n \times n$ simétrica, definida positiva. Salvo que se indique lo contrario, se usará la *métrica de pesos D* cuya matriz tiene en la diagonal los pesos de los individuos y ceros en las otras entradas:

$$D = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{bmatrix} \tag{2.13}$$

donde se supone que para todo i se tiene $p_i > 0$ y $\sum_{i=1}^{n} p_i = 1$. En el caso en que todos los pesos de los individuos son iguales, entonces $D = \frac{1}{n} I_n$ con I_n la matriz identidad de dimensión n. En el caso de variables cualitativas, las métricas de pesos se definirán a partir de la tabla de contingencia por medio de los perfiles marginales.

2.2.2.3. Matriz de covarianza

Se tienen p variables cuantitativas centradas $x^1, x^2, \cdots, x^p$, que definen una matriz X. Se define la matriz de covarianzas, también llamada matriz de varianzas-covarianzas, como la matriz V de dimensiones $p \times p$ tal que en la entrada (i, j) de la diagonal contiene la varianza de la variable $x^j : var(x^j)$, y en la entrada (j, k), con $j \neq k$, la covarianza entre x^j y $x^k : cov(x^j, x^k)$.

Así,

$$V = \begin{bmatrix} var(x^1) & cov(x^1, x^2) & \cdots & cov(x^1, x^p) \\ cov(x^1, x^2) & var(x^2) & \cdots & cov(x^2, x^p) \\ \vdots & & \ddots & \vdots \\ cov(x^1, x^p) & cov(x^2, x^p) & \cdots & var(x^p) \end{bmatrix} \tag{2.14}$$

Entonces, V puede calcularse matricialmente así:

$$V = X^t D X \tag{2.15}$$

Si x^j, x^k son dos de las variables centradas, entonces su covarianza es:

$$cov(x^j, x^k) = (x^j)^t D x^k \tag{2.16}$$

Asimismo, la varianza de x^j es:

$$var(x^j) = (x^j)^t D x^j \tag{2.17}$$

La varianza puede ser vista como la norma al cuadrado de un vector de $\mathbb{R}^n$ en el caso de variables centradas: $var(x^j) = \|x^j\|_D^2$

Para variables centradas, la correlación por su lado puede ser vista como el ángulo entre dos vectores de $\mathbb{R}^n$:

$$r(x^j, x^k) = \frac{cov(x^j, x^k)}{\sqrt{var(x^j) var(x^k)}} = \frac{(x^j)^t D x^k}{\|x^j\|_D \|x^j\|_D} = \cos\theta \tag{2.18}$$

donde θ es el ángulo en $\mathbb{R}^n$ formado por los vectores x^j y x^k, siempre que las

variables tengan media 0.

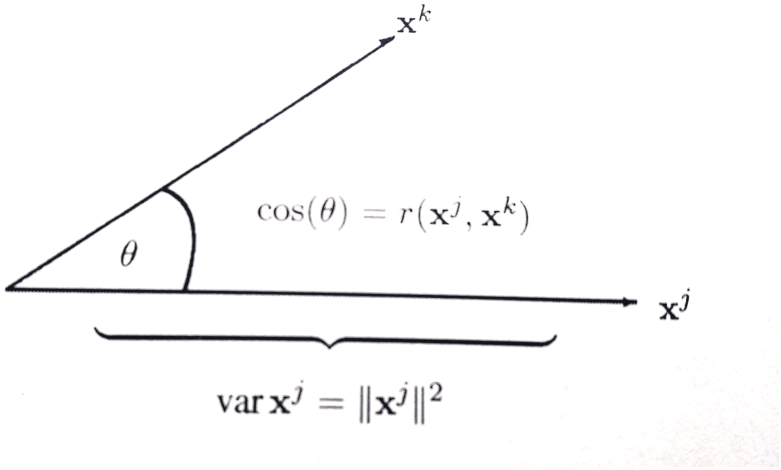


Figura 2.2: Caso de variables centradas: la norma es una varianza y la correlación es un coseno.

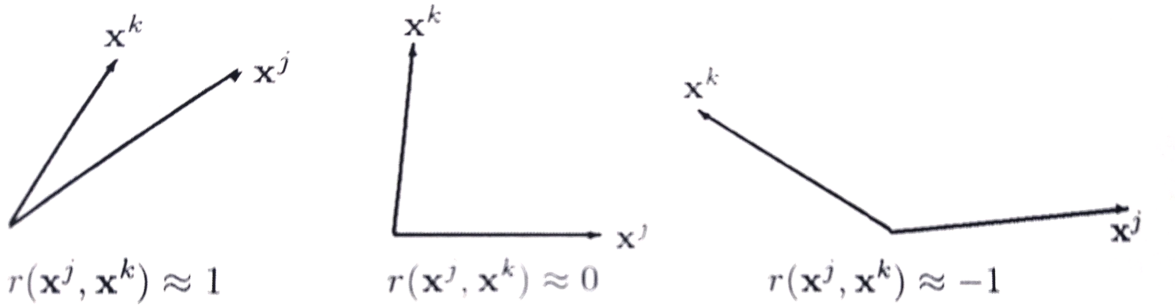


Figura 2.3: Tres situaciones típicas para la correlación entre dos variables centradas x^j y x^k

2.2.2.4. Métricas en $\mathbb{R}^p$

Se define otra métrica usual en $\mathbb{R}^p$ a través de la diagonal de las inversas de las varianzas:

$$D_{1/\sigma^2} = \begin{bmatrix} \frac{1}{var(x^1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{var(x^2)} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{var(x^p)} \end{bmatrix} \tag{2.19}$$

Está métrica se recomienda cuando las unidades de medida para las variables son diferentes (por ejemplo, algunas variables son medidas en metros, otras en ki-

logramos, otras son notas,etc).

El usar la métrica D_{1/σ^2} para medir distancias, corresponde en la practica a estandarizar las variables y usar luego la distancia Euclídea clásica.

2.2.2.5. Matriz de correlaciones

Se define la matriz de correlaciones R tal que contiene en su entrada (i, j) la correlación $r(x^j, x^k)$

$$R = \begin{bmatrix} 1 & r(x^1, x^2) & \cdots & r(x^1, x^p) \\ r(x^1, x^2) & 1 & \cdots & r(x^2, x^p) \\ \vdots & & \ddots & \vdots \\ r(x^1, x^p) & r(x^2, x^p) & \cdots & 1 \end{bmatrix} \tag{2.20}$$

Matricialmente se puede escribir $R = D_{1/\sigma} V D_{1/\sigma}$ donde $D_{1/\sigma} = \text{diag}(1/\sigma_j)$

2.2.3. Análisis en Componentes Principales

El Análisis en Componentes Principales constituye la técnica base en Análisis Multivariado de Datos. Su principal objetivo es el de encontrar, a partir de una tabla de datos con variables cuantitativas, un conjunto de variables sintéticas cuya información sea lo más parecida a la de las variables originales. Es por lo tanto, una técnica de *reducción de las dimensiones* de un problema puesto que de un conjunto inicial de variables, que pueden ser muchas, se trata de encontrar un conjunto reducido de variables que contengan prácticamente la misma información que las variables originales.

En general, las tablas de datos definen nubes de puntos en espacios vectoriales con dimensiones muy grandes, por lo que la visualización de las relaciones entre los puntos es imposible cuando la dimensión del espacio es mayor que 3.

2.2.3.1. Nubes de puntos

Sea X una tabla de datos definida con variables cuantitativas, y sean $\mathbb{R}^p$ el espacio de individuos y $\mathbb{R}^n$ el de variables. Si M es la métrica sobre $\mathbb{R}^p$ y D la métrica de pesos sobre $\mathbb{R}^n$, entonces se denota con $\mathcal{N} = (X, M, D)$ la *nube de puntos*, esto es, los n puntos ponderados del espacio vectorial $\mathbb{R}^p$, junto con la medida de proximidad y angular definidas por M, y las medidas de tendencia central y de dispersión asociadas a D.

El término nube de puntos es entonces un concepto geométrico, cuya forma se tratará de describir y sintetizar mediante métodos estadísticos.

2.2.3.2. Inercia en un punto

Sea $\mathbb{R}^p$ provisto de una métrica M, se llama *inercia en un punto a* de $\mathbb{R}^p$ a la cantidad:

$$I_a(\mathcal{N}) = \sum_{i=1}^{n} p_i \|x_i - a\|_M^2 \tag{2.21}$$

Este valor mide la *dispersión* de la nube de puntos $\mathcal{N}$ alrededor del punto a. El *centro de gravedad* g de la nube de puntos se define como:

$$g = \sum_{i=1}^{n} p_i x_i \tag{2.22}$$

La inercia o dispersión es mínima cuando es medida respecto el centro de gravedad o término medio, tal como lo establece el teorema de Huygens.

Teorema 2.1 (Teorema de Huygens) *Para todo $a \in \mathbb{R}^p$ se tiene*

$$I_a(\mathcal{N}) = I_g(\mathcal{N}) + \|a - g\|_M^2 \tag{2.23}$$

Demostración: Sea $a \in \mathbb{R}^p$, entonces:

$$\begin{aligned}
I_a(\mathcal{N}) &= \sum_{i=1}^{n} p_i (x_i - a)^t M (x_i - a) \\
&= \sum_{i=1}^{n} p_i (x_i - g + g - a)^t M (x_i - g + g - a) \\
&= \sum_{i=1}^{n} p_i (x_i - g)^t M (x_i - g) + 2 \sum_{i=1}^{n} p_i (x_i - g)^t M (g - a) \\
&+ \sum_{i=1}^{n} p_i (g - a)^t M (g - a) \\
&= I_g(\mathcal{N}) + 2(g - a)^t M \sum_{i=1}^{n} p_i (x_i - g) + \|g - a\|_M^2
\end{aligned}$$

donde se ha usado el hecho que M es simétrica, que un número real puede ser visto como una matriz 1×1 y por tanto es igual a su transpuesta, y que la suma de los pesos es 1.

Ahora bien, $\sum_{i=1}^{n} p_i (x_i - g) = 0$ por definición de g, por lo que se obtiene el resultado.

□

$I_g(\mathcal{N})$ es llamada la **inercia total** de la nube $\mathcal{N}$ y se suele denotar $I(\mathcal{N})$.

Si se dispone de solamente una variable x, entonces el centro de gravedad es $\bar{x}$, y por lo tanto la inercia $I(\mathcal{N})$ es exactamente $var(x)$.

2.2.3.3. Objetivo del Análisis en Componentes Principales (A.C.P)

Supóngase que se está en presencia de n individuos $x_1, x_2, \ldots, x_n$ sobre los que se han medido p variables cuantitativas $x^1, x^2, \ldots, x^p$. Por lo tanto se define una tabla de datos X, con n filas y p columnas. Cada fila de la matriz se puede ver como un punto de $\mathbb{R}^p$, así el conjunto de n individuos define una nube de n puntos de $\mathbb{R}^p$, denotada $\mathcal{N} = (X, M, D)$, con M la métrica $p \times p$ sobre el espacio de individuos y D la métrica de pesos (matriz diagonal $n \times n$) sobre el espacio de variables.

Se supondrá que las variables x^j están centradas. Esto significa que el centro de gravedad de la nube de los n puntos en $\mathbb{R}^p$ está en el origen de coordenadas: $g = 0$.

El A.C.P se plantea como una técnica de reducción de la dimensión del problema, se busca un espacio de dimensión q, menor que p, de manera que las posiciones relativas de los puntos - individuos sean lo más similares posibles a sus posiciones en el espacio $\mathbb{R}^p$. Esto significa que hay una pérdida mínima de información al proyectar los n individuos sobre un espacio de dimensión menor, de esta forma, su dispersión en el espacio proyectado $\mathbb{R}^q$ debe ser máxima, de manera que la forma de la nube se asemeja lo mejor posible a su forma original.

Dada la tabla de datos X, se busca un conjunto de q variables sintéticas $c^1, c^2, \ldots, c^q$, donde $q < p$, que más adelante se llamarán *componentes principales*, tal que:

1. Cada componente principal c^k debe ser combinación lineal de las variables originales x^j; esto significa que la información contenida en las x^j también está reflejada en las c^k.
2. Las componentes principales deben ser no correlacionadas dos a dos; esto significa que las c^k no tienen información redundante.
3. Las componentes principales deben tener varianzas máximas; esto significa que contendrán el máximo de información posible.

2.2.3.4. Cálculo de las componentes

Cálculo del primer componente

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima. Los valores de este primer componente en los n individuos se representarán por un vector z^1, dado por

$$z^1 = Xa^1 \tag{2.24}$$

Como las variables originales tienen media cero también z^1 tendrá media cero. Su varianza será:

$$\frac{1}{n}(z^1)^t z^1 = \frac{1}{n}(a^1)^t X^t X a^1 = (a^1)^t V a^1 \tag{2.25}$$

donde V es la matriz de varianzas y covarianzas de las observaciones. Podemos maximizar la varianza sin limite aumentando el módulo del vector a_1. Para que la maximización de (2.25) tenga solución debemos imponer una restricción al módulo del vector a^1, en este caso, impondremos que $(a^1)^t a^1 = 1$

Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$L = (a^1)^t V a^1 - \lambda((a^1)^t a^1 - 1) \tag{2.26}$$

y maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de a^1 e igualando a cero. Entonces

$$\frac{\partial L}{\partial a^1} = 2Va^1 - 2\lambda a^1 = 0 \tag{2.27}$$

Cuya solución es:

$$Va^1 = \lambda a^1 \tag{2.28}$$

que implica que a^1 es un vector propio de la matriz V, y λ su correspondiente valor propio. Para determinar qué valor propio de V es la solución de la ecuación (2.28) tendremos en cuenta que, multiplicando por la izquierda por $(a^1)^t$ esta ecuación:

$$(a^1)^t V a^1 = \lambda (a^1)^t a^1 = \lambda \tag{2.29}$$

y concluimos, por (2.25), que λ es la varianza de z^1. Como esta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz V. Su vector asociado, a^1, define los coeficientes de cada variable en el primer componente principal.

Cálculo del segundo componente

Para obtener el mejor plano de proyección de las variables X. Se establece como función objetivo que la suma de las varianzas de $z^1 = Xa^1$ y $z^2 = Xa^2$ sea máxima, donde a^1 y a^2 son los vectores que definen el plano. La función objetivo será:

$$\phi = (a^1)^t V a^1 + (a^2)^t V a^2 - \lambda_1((a^1)^t a^1 - 1) + \lambda_2((a^2)^t a^2 - 1) \tag{2.30}$$

que incorpora las restricciones de que las direcciones deben de tener módulo unitario $((a^i)^t a^i) = 1, i = 1, 2$. Derivando e igualando a cero:

$$\frac{\partial \phi}{\partial a^1} = 2Va^1 - 2\lambda_1 a^1 = 0 \tag{2.31}$$

$$\frac{\partial \phi}{\partial a^2} = 2Va^2 - 2\lambda_2 a^2 = 0 \tag{2.32}$$

La solución de este sistema es:

$$Va^1 = \lambda_1 a^1 \tag{2.33}$$

$$Va^2 = \lambda_2 a^2 \tag{2.34}$$

que indica que a^1 y a^2 deben ser vectores propios de V. Tomando los vectores propios de norma uno y sustituyendo en (2.30), se obtiene que, en el máximo, la función objetivo es:

$$\phi = \lambda_1 + \lambda_2 \tag{2.35}$$

Es claro que λ_1 y λ_2 deben ser los dos autovalores mayores de la matriz V y a^1 y a^2 sus correspondientes autovectores. Observemos que la covarianza entre z^1 y z^2, dada por $(a^1)^t V a^2$ es cero ya que $(a^1)^t a^2 = 0$, y las variables z^1 y z^2 estarán incorrelacionadas.

Generalización

El espacio de dimensión q que mejor representa a los puntos viene definido por los vectores propios asociados a los q mayores autovalores de V. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables por ellas definidas componentes principales.

En general, la matriz X (y por tanto la V) tiene rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los

valores propios o raíces características, $\lambda_1, \lambda_2, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables V mediante:

$$|V - \lambda I| = 0 \tag{2.36}$$

y sus vectores propios asociados

$$(V - \lambda_i I)a^i = 0 \tag{2.37}$$

Los términos λ_i son reales y positivos, al ser la matriz V simétrica, definida positiva.

Por ser V simétrica, si λ_i y λ_h son dos raíces distintas sus vectores asociados son ortogonales. Si V fuese semidefinida positiva de rango $q < p$, lo que ocurriría si $p - q$ variables fuesen combinación lineal de las demás, habría solamente q raíces características positivas y el resto serían ceros.

Llamando Z a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$Z = XA \tag{2.38}$$

donde $A^t A = I$.

Calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables X (ejes originales) para obtener unas nuevas variables Z incorrelacionadas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los ejes naturales de los datos.

2.2.3.5. Propiedades de las componentes

Las componentes principales como nuevas variables tienen las propiedades siguientes:

1) **Conservan la variabilidad inicial**: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.

   Comprobemos el primer punto. Como $var(z^h)$ y la suma de las raíces características es la traza de la matriz:

$$tr(L) = var(x^1) + \cdots + var(x^p) = \lambda_1 + \cdots + \lambda_p \tag{2.39}$$

Por tanto, $\sum_{i=1}^{p} var(x^i) = \sum \lambda_i = \sum_{i=1}^{p} var(z^i)$. Las nuevas variables z^i tienen conjuntamente la misma variabilidad que las variables originales, la suma de varianzas es la misma, pero su distribución es muy distinta en los dos conjuntos.

Para comprobar que los componentes principales también conservan la *varianza generalizada*, el valor del determinante de varianzas y covarianzas de las variables. Como el determinante es el producto de las raíces características, tenemos que, llamando V_z a la matriz de covarianzas de los componentes, que es diagonal con términos λ_i :

$$|V_x| = \lambda_1 + \cdots + \lambda_p = \prod_{i=1}^{p} var(z^i) = |V_z| \tag{2.40}$$

2) **La proporción de variabilidad explicada por un componente es el cociente entre su varianza el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz**.

   Como la varianza del componente h es λ_h, el valor propio que define el componente, y la suma de todas las varianzas de las variables originales es $\sum_{i=1}^{p} \lambda_i$, donde la suma de las varianzas de los componentes, la proporción de variabilidad total explicada por el componente h es $\dfrac{\lambda_i}{\sum \lambda_i}$

3) **Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por el valor propio**:

   $$cov(z^i; x^1, \cdots, x^p) = \lambda_i a^i = (\lambda_i a_{i1}, \cdots, \lambda_i a_{ip}) \tag{2.41}$$

   donde a^i es el vector de coeficientes de la componente z^i.

   Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

   $$cov(z, x) = \frac{1}{n} Z^t X \tag{2.42}$$

   y su primera fila proporciona las covarianzas entre la primera componente y las p variables originales. Como $Z = XA$, sustituyendo en la ecuación (2.42), tenemos que:

$$cov(z, x) = \frac{1}{n} A^t X^t X = A^t V = P A^t \tag{2.43}$$

donde A contiene en columnas los vectores propios de V y P es la matriz diagonal de los valores propios.

En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $a^t V$, es decir $(a^1)^t V$ o también $\lambda_1 (a^1)^t$, donde $(a^1)^t$ es el vector de coeficientes de la primera componente principal.

4) **Las correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable**:

   Para comprobarlo:

$$r(x^i, x^j) = \frac{cov(z^i, x^j)}{\sqrt{var(z^i) var(x^j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij} \frac{\sqrt{\lambda_i}}{s_j} \tag{2.44}$$

5) **Las q componentes principales $(q < p)$ proporcionan la predicción lineal óptima con q variables del conjunto de variables X**:

   Esta afirmación puede expresarse de dos formas. La primera demostrando que la mejor predicción lineal con q variables de las variables originales se obtiene utilizando las q primeras componentes principales. La segunda demostrando que la mejor aproximación de la matriz de datos que puede construirse con una matriz de rango q se obtiene construyendo esta matriz con los valores de los q primeros componentes principales.

6) **Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales**.

   Estandarizando los componentes Z por sus desviaciones típicas, se obtienen las nuevas variables

$$Y_c = Z P^{-1/2} = X A P^{-1/2} \tag{2.45}$$

2.3. Análisis de conglomerados

El análisis de conglomerados (clústers) tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos, es similar a la clasificación (discriminación), excepto que los grupos no son predefinidos. El objetivo es particionar o segmentar un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similaridad de los datos o individuos en ciertas variables. Como los grupos no son dados a priori el experto debe dar una interpretación de los grupos que se forman.

Este método se conoce también con el nombre de método de clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión, la cual tiene por objetivo reconocer grupos de individuos homogéneos, de tal forma que los grupos queden bien separados y bien diferenciados. Éstos individuos pueden estar descritos en una tabla de datos de individuos y variables, con variables cuantitativas o cualitativas, o por una tabla de proximidades. El nombre de no supervisado se aplica para distinguirlos del análisis discriminante.

El análisis de conglomerados estudia tres tipos de problemas:

Partición de los datos. Disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:

(1) Cada elemento pertenezca a uno y solo uno de los grupos.

(2) Todo elemento quede clasificado.

(3) Cada grupo sea internamente homogéneo.

Construcción de jerarquías. Deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Sin embargo, la jerarquía construida permite obtener también una partición de los datos en grupos.

Clasificación de variables. En problemas con muchas variables es interesante hacer un estudio exploratorio inicial para dividir las variables en grupos. Este estudio puede orientarnos para plantear los modelos formales para reducir la dimensión. Las variables pueden clasificarse en grupos o estructurarse en una jerarquía.

Los métodos de partición utilizan la matriz de datos, pero los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. Para agrupar

variables se parte de la matriz de relación entre variables; para variables continuas suele ser la matriz de correlación, y para variables discretas, se construye a partir de la distancia chi-cuadrado.

2.3.1. Métodos clásicos de partición

La Minería de Datos es la extracción de información o de patrones no trivial, implícita, previamente desconocida y potencialmente útil de grandes bases de datos. Es analizar datos para encontrar patrones ocultos usando medios automatizados. Es un proceso no elemental de búsqueda de relaciones, correlaciones, dependencias, asociaciones, modelos, estructuras, tendencias, clases (clústers), segmentos, los cuales se obtienen de grandes juegos de datos, los cuales generalmente están almacenados en bases de datos (relacionales o no).

Esta búsqueda se lleva a cabo utilizando métodos matemáticos, estadísticos o algorítmicos. Se considera la Minería de Datos como el proceso más automatizado, que va de los datos elementales disponibles en una Bodega de Datos a la decisión. Dentro de los objetivos de la Minería de Datos está crear un proceso automatizado que toma como punto de partida los datos y cuya meta es la ayuda a la toma de decisiones.

2.3.1.1. Fundamentos de algoritmos de *k*-means

Existe un poco de confusión en la literatura acerca del método de las *k*-means, ya que hay dos métodos distintos que son llamados con el mismo nombre. Originalmente, Forgy propuso en 1965 un primer método de reasignación-recentraje que consiste básicamente en la iteración sucesiva, hasta obtener convergencia, de las dos operaciones siguientes:

(a) Representar una clase por su centro de gravedad, esto es, por su vector de promedios.

(b) Asignar los objetos a la clase del centro de gravedad más cercano.

Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado K, la cual requiere las cuatro etapas siguientes:

(1) Seleccionar K puntos como centros de los grupos iniciales. Esto puede hacerse:

a) Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados.

b) Tomando como centros los K puntos más alejados entre sí.

c) Construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

(2) Calcular las distancias euclídeas de cada elemento al centro de los K grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

(3) Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

(4) Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

Implementación del algoritmo

El criterio de homogeneidad que se utiliza en el algoritmo de k-means es *la suma de cuadrados dentro de los grupos (SCDG)* para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2 \tag{2.46}$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k y $\bar{x}_{jk}$ es la media de esta variable en el grupo. El criterio se escribe

$$\text{mín}\, SCDG = \text{mín} \sum_{k=1}^{K} \sum_{j=1}^{p} n_k \sigma_{jk}^2 \tag{2.47}$$

donde n_k es el número de elementos del grupo k y σ_{jk}^2 es la varianza de la variable j en dicho grupo. La varianza de cada variable en cada grupo es claramente una medida de la heterogeneidad del grupo y al minimizar las varianzas de todas las variables en los grupos obtendremos grupos más homogéneos. Un posible criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclídea, este criterio se escribe:

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)'(x_{ik} - \bar{x}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,k) \tag{2.48}$$

donde $d^2(i,k)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo k y su media de grupo. Para comprobar que ambos criterios son idénticos escribimos este último criterio en términos de su traza, como:

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} tr[d^2(i,k)] = \text{mín}\, tr \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)' \right] \tag{2.49}$$

y llamando W a la matriz de la suma de los cuadrados dentro de los grupos,

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)' \tag{2.50}$$

tenemos que:

$$\text{mín}\, tr(W) = \text{mín}\, SCDG \tag{2.51}$$

Como la traza es la suma de los elementos de la diagonal principal ambos criterios coinciden. Este criterio se denomina criterio de la traza.

La maximización de este criterio requeriría calcularlo para todas las posibles particiones, labor claramente imposible, salvo para valores de n muy pequeños. El algoritmo de k-means busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. El algoritmo funciona de la siguiente manera:

(1) Partir de una asignación inicial.

(2) Comprobar si moviendo algún elemento se reduce W.

(3) Si es posible, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible, terminar.

En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. Conviene siempre repetir el algoritmo desde distintos valores iniciales y permutando los elemento de la muestra. El efecto del orden de las observaciones suele ser pequeño, pero conviene asegurarse en cada caso de que no esta afectando.

El criterio de la traza tiene dos propiedades importantes. La primera es que no es invariante ante cambios de medida en las variables. Cuando las variables vayan en unidades distintas conviene estandarizarlas, para evitar que el resultado del algoritmo de k-means dependa de cambios irrelevantes en la escala de medida. Cuando vayan en las mismas unidades suele ser mejor no estandarizar, ya que es posible que una varianza mucho mayor que el resto sea precisamente debida a que existen dos grupos de observaciones en esa variable, y si estandarizamos podemos ocultar la presencia de los grupos.

La segunda propiedad del criterio de la traza es que minimizar la distancia euclídea produce grupos aproximadamente esféricos. Por otro lado este criterio esta pensado para variables cuantitativas, aunque puede aplicarse si existe un pequeño número de variables binarias.

Número de Grupos

En la aplicación habitual del algoritmo de *k*-means hay que fijar el número de grupos, K. Es necesario hacer notar que, cuando se quiere obtener una partición en K grupos de un conjunto con n individuos, no tiene sentido examinar todas las posibles particiones del conjunto de individuos en K grupos.

Es claro que este número no puede estimarse con un criterio de homogeneidad ya que la forma de conseguir grupos muy homogéneos y minimizar la $SCDG$ es hacer tantos grupos como observaciones, con lo que siempre $SCDG = 0$. Se han propuesto distintos métodos para seleccionar el número de grupos. Un procedimiento aproximado que se utiliza bastante, aunque puede no estar justificado en unos datos concretos, es realizar un test F aproximado de reducción de variabilidad, comparando la $SCDG$ con K grupos con la de $K+1$, y calculando la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test es:

$$F = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n-K-1)} \tag{2.52}$$

y compara la disminución de variabilidad al aumentar un grupo con la varianza promedio.

2.3.1.2. Métodos Jerárquicos

Distancias y Similaridades

Distancias Euclídeas

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en una distancia. Si todas las variables son continuas, la distancia más utilizada es la distancia euclídea entre las variables estandarizadas.

Si no estandarizamos, la distancia euclídea dependerá sobre todo de las variables con valores más grandes, y el resultado del análisis puede cambiar completamente al modificar su escala de medida. Si estandarizamos, estamos dando a priori un peso semejante a las variables, con independencia de su variabilidad original, lo

que puede no ser siempre adecuado.

Cuando en la muestra existen variables continuas y atributos el problema se complica. Supongamos que la variable x_1 es binaria. La distancia euclídea entre dos elementos de la muestra en función de esta variable es $(x_{i1} - x_{h1})^2$ que tomará el valor cero si $x_{i1} = x_{h1}$, es decir cuando el atributo está, o no está, en ambos elementos. Sin embargo, la distancia entre dos elementos correspondiente a una variable continua estandarizada, $(x_{i1} - x_{h1})^2/\sigma_1^2$, puede ser mucho mayor que uno con lo que las variables continuas van en general a pesar mucho más que las binarias. Esto puede ser aceptable en muchos casos pero cuando por la naturaleza del problema esta situación no sea deseable la solución es trabajar con similaridades.

Similaridades

El coeficiente de similaridad según la variable $j = 1, \ldots, p$ entre dos elementos muéstrales (i, h), se define como una función, s_{jih}, no negativa y simétrica:

(1) $s_{jii} = 1$

(2) $0 \leq s_{jih} \leq 1$

(3) $s_{jih} = s_{jhi}$

Si obtenemos las similaridades para cada variable entre dos elementos podemos combinarlas en un coeficiente de similaridad global entre los dos elementos:

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih} s_{jih}}{\sum_{j=1}^{p} w_{jih}} \tag{2.53}$$

donde w_{jih} es una variable ficticia que es igual a uno si la comparación de estos dos elementos mediante la variable j tiene sentido, y será cero si no queremos incluir esa variable en la comparación entre los elementos. Por ejemplo, la variable x_1 es si una persona ha pedido $(x_1 = 1)$ o no $(x_1 = 0)$ un crédito y la x_2 si lo ha devuelto o no, si una persona no ha pedido crédito, tiene $x_1 = 0$, no tiene sentido preocuparse de x_2. En este caso al comparar individuos (i, j) si uno cualquiera de los dos tiene un valor cero en x_1, asignaremos a la variable w_{2ij} el valor cero.

Las similaridades entre elementos en función de las variables cualitativas pueden construirse individualmente o por bloques. La similaridad entre dos elementos por una variable binaria será uno, si ambos tienen el atributo, y cero en caso

contrario. Alternativamente, podemos agrupar las variables binarias en grupos homogéneos y tratarlas conjuntamente. Si suponemos que todos los atributos tienen el mismo peso, podemos construir una medida de similaridad entre dos elementos A y B respecto a todos estos atributos contando el número de atributos que están presentes:

1. En ambos (a).
2. En A y no en B, (b).
3. En B y no en A, (c).
4. En ninguno de los dos elementos, (d).

Estas cuatro cantidades forman una tabla de asociación entre elementos, y servirán para construir medidas de similitud o similaridad entre los dos elementos comparados. En esta tabla se verifica que $n_a = a + b + c + d$, donde n_a es el número de atributos.

1. *Proporción de coincidencias*. Se calcula como el número total de coincidencias sobre el número de atributos totales:

$$s_{ij} = \frac{a+d}{n_a} \tag{2.54}$$

2. *Proporción de apariciones*. Cuando la ausencia de un atributo no es relevante, podemos excluir las ausencias y calcular sólo la proporción de veces donde el atributo aparece en ambos elementos. El coeficiente se define por:

$$s_{ij} = \frac{a}{a+b+c} \tag{2.55}$$

Aunque las dos propuestas anteriores son las más utilizadas puede haber situaciones donde sean recomendables otras medidas. Por ejemplo, podemos querer dar peso doble a las coincidencias, con lo que resulta $s_{ij} = 2(a+d)/(2(a+d)+b+c)$, o tener sólo en cuenta las coincidencias y tomar $s_{ij} = a/(b+c)$. Finalmente los coeficientes de similitud o similaridad para una variable continua se construye mediante

$$s_{jih} = 1 - \frac{|x_{ij} - x_{hj}|}{rango(x_j)} \tag{2.56}$$

de esta manera el coeficiente resultante estará siempre entre cero y uno. Cuando tenemos varias variables estos coeficientes pueden combinarse como indica la ecuación (2.53).

Una vez obtenida la similaridad global entre los elementos, podemos transformar los coeficientes en distancias. Lo más simple es definir la distancia mediante $d_{ij} = 1 - s_{ij}$, pero está relación puede no verificar la propiedad triangular, pero si la matriz de similaridades es definida positiva y calculamos las similitudes por (2.54) o (2.55), y definimos la distancia por:

$$d_{ij} = \sqrt{2(1 - s_{ij})} \tag{2.57}$$

entonces sí se verifica la propiedad triangular

Algoritmos Jerárquicos

Dada una matriz de distancias o de similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos:

1. De *aglomeración*. Parten de los elementos individuales y los van agregando en grupos.

2. De *división*. Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Los algoritmos de aglomeración requieren menos tiempo de cálculo y son los más utilizados.

Métodos Aglomerativos

Los algoritmos aglomerativos que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Su estructura es:

(1) Comenzar con tantas clases como elementos. Las distancias entre clases son las distancias entre elementos originales.

(2) Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.

(3) Sustituir los dos elementos utilizados en (2) para definir la clase por un nuevo elemento que represente la clase construida.

(4) Volver a (2) y repetir (2) y (3) hasta que tengamos todos los elementos agrupados en una clase única.

Criterios para definir distancias entre grupos

Supongamos que tenemos un grupo A con n_a elementos, y un grupo B con n_b elementos, y que ambos se fusionan para crear un grupo (AB) con $n_a + n_b$ elementos. La distancia del nuevo grupo, (AB), a otro grupo C con n_c elementos, se calcula habitualmente por alguna de las cinco reglas siguientes:

1. *Encadenamiento simple o vecino más próximo*. La distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = \text{mín}(d_{CA}, d_{CB}) \tag{2.58}$$

   Una forma simple de calcular con un ordenador el mínimo entre las dos distancias es utilizar que

$$\text{mín}(d_{CA}, d_{CB}) = 1/2(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|) \tag{2.59}$$

   En efecto, si $d_{CB} > d_{CA}$ el término en valor absoluto es $d_{CB} - d_{CA}$ y el resultado de la operación es d_{CA}, la menor de las distancias. Si $d_{CA} > d_{CB}$ el segundo término es $d_{CA} - d_{CB}$ y se obtiene d_{CB}.

   Como este criterio sólo depende del orden de las distancias será invariante ante transformaciones monótonas; obtendremos la misma jerarquía aunque las distancias sean numéricamente distintas.

2. *Encadenamiento completo o vecino más alejado*. La distancia entre los dos nuevos grupos es la mayor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = \text{máx}(d_{CA}, d_{CB}) \tag{2.60}$$

   y puede comprobarse que

$$\text{máx}(d_{CA}, d_{CB}) = 1/2(d_{CA} + d_{CB} + |d_{CA} - d_{CB}|) \tag{2.61}$$

   Este criterio será también invariante ante transformaciones monótonas de las distancias al depender, como el anterior, del orden de las distancias.

3. *Media de grupos*. La distancia entre los dos nuevos grupos es la media ponderada entre las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB} \tag{2.62}$$

Como se ponderan los valores de las distancias, este criterio no es invariante ante transformaciones monótonas de las distancias.

4. *Método del centroide*. Se aplica generalmente sólo con variables continuas. La distancia entre dos grupos se hace igual a la distancia euclídea entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Cuando se unen dos grupos se pueden calcular las nuevas distancias entre ellos en la que el cuadrado de la distancia euclídea de un grupo C a la unión de los grupos A, con n_a elementos y B con n_b es

$$d^2(C;AB) = \frac{n_a}{n_a+n_b}d^2_{CA} + \frac{n_b}{n_a+n_b}d^2_{CB} - \frac{n_a n_b}{(n_a+n_b)^2}d^2_{AB} \tag{2.63}$$

El método de Ward

Un proceso algo diferente de construir el agrupamiento jerárquico ha sido propuesto por Ward. La diferencia con los métodos anteriores es que ahora se parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define una medida global de la heterogeneidad de una agrupación de observaciones en grupos. Esta medida es W, la suma de las distancias euclídeas al cuadrado entre cada elemento y la media de su grupo:

$$W = \sum_k \sum_{i\in k} (x_{ik} - \bar{x}_k)'(x_{ik} - \bar{x}_k) \tag{2.64}$$

donde $\bar{x}_k$ es la media del grupo k. El criterio comienza suponiendo que cada dato forma un grupo, $k = n$ y por tanto W en (2.64) es cero. A continuación se unen los elementos que produzcan el incremento mínimo de W. Obviamente esto implica tomar los más próximos con la distancia euclídea. En la siguiente etapa tenemos $n-1$ grupos, $n-2$ de un elemento y uno de dos elementos. Decidimos de nuevo que dos grupos unir para que W crezca lo menos posible, con lo que pasamos a $n-2$ grupos y así sucesivamente hasta tener un único grupo. Los valores de W van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos.

Puede demostrarse que, en cada etapa, los grupos que debe unirse para minimizar W son aquellos tales que:

$$\text{mín}\ \frac{n_a n_b}{(n_a+n_b)^2}(x_{ik} - \bar{x}_k)'(x_{ik} - \bar{x}_k) \tag{2.65}$$

Comparación

Es difícil dar reglas generales que justifiquen un criterio sobre otro, aunque los más utilizados son los tres últimos. Aunque es conveniente analizar que criterio es más razonable para los datos que se quieren agrupar y en caso de duda probar con varios y comparar los resultados.

El dendrograma

El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias tienen la propiedad de que si consideramos tres grupos, A, B, C, se tiene que:

$$d(A, C) \leq \text{máx}\{d(A, B), d(B, C)\} \tag{2.66}$$

y una medida de distancia que tiene esta propiedad se denomina ultramétrica. Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto si $d^2(A, C)$ es menor o igual que el máximo de $d^2(A, B), d^2(B, C)$ forzosamente será menor o igual que la suma $d^2(A, B) + d^2(B, C)$. El dendrograma es la representación de una ultramétrica, y se construye como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.
2. Las uniones entre elementos se representan por tres lineas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están conectados por lineas rectas.

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.

2.3.2. Métodos modernos de partición

En el caso de métodos no supervisado, la información correspondiente a los grupos es desconocida. Los algoritmos de los métodos no supervisado incluyen reglas de asociación y métodos de agrupamiento. El método de agrupamiento más

utilizado es el *k*-means. Sin embargo, este algoritmo tiene varias desventajas: el número de clusters es desconocido, es sensible a los valores atípicos y también a los puntos iniciales. Por otra parte, *k*-means generalmente genera agrupaciones que son relativamente uniformes en tamaño y de forma esférica. Debido a que *k*-means es el rey de los algoritmos de agrupamiento, tiene muchas variantes y debido a sus desventajas algunos de ellos son mejores para el análisis de grandes conjuntos de datos y otros dan mejores resultados para encontrar grupos con formas arbitrarias. Algunas alternativas son *k*-medoids que es menos sensible a Outliers, CLARA y CLARANS. Estos métodos pueden aplicarse a datos de alta dimensión. Otra debilidad del uso de *k*-means es la forma de los grupos, pero esto puede resolverse utilizando DBSCAN.

2.3.2.1. Método *k*-medoids

Este método es una variación del algoritmo *k*-means (*k*-medias), el objetivo del algoritmo *k*-medoids, está en encontrar una solución de agrupamiento la cual permita minimizar una función objetivo predefinida. Es decir, minimizar la suma de las diferencias por parejas en lugar de minimizar la suma de los cuadrados de las distancias euclidianas.

La diferencia entre *k*-means y *k*-medoides es análoga a la diferencia entre la media y la mediana; donde la media indica el valor promedio de todos los elementos de datos recopilados, mientras que la mediana indica el valor alrededor del cual todos los elementos de datos se distribuyen uniformemente a su alrededor. La idea básica de este algoritmo es calcular primero los K objetos representativos que se denominan medoides; después de encontrar el conjunto de medoides, cada objeto del conjunto de datos se asigna al medoide más cercano. Es decir, el objeto i se coloca en el grupo k, cuando el medoid x_k está más cerca que cualquier otro medoid x_w.

k-means clustering encuentra los k centroides, donde la coordenada de cada centroide es el medio de las coordenadas de los objetos en el clúster y asigna cada objeto al centroide más cercano. El algoritmo se puede resumir de la siguiente manera:

- Paso 1: Seleccionar K objetos al azar. Estos objetos representan los centroides del grupo inicial.
- Paso 2: Asigne cada objeto al grupo que tenga el centroide más cercano.
- Paso 3: Cuando todos los objetos hayan sido asignados, vuelva a calcular las posiciones de los K centroides.
- Paso 4: Repita los pasos 2 y 3 hasta que los centroides ya no se muevan.

Desafortunadamente, el agrupamiento k-means es sensible a los valores atípicos y un conjunto de objetos más cercanos a un centroide puede estar vacío, en cuyo caso los centroides no pueden actualizarse. Por esta razón, a veces se usan agrupaciones de k-medoids, donde se consideran objetos representativos llamados medoides en lugar de centroides. Debido a que utiliza el objeto ubicado más centralmente en un clúster, es menos sensible a los valores atípicos en comparación con el clúster k-means.

Algoritmo k-medoid

Supongamos que tenemos n objetos que tienen p variables que se clasificarán en $k(k < n)$ agrupamientos (se supone que se obtienen k grupos). Definamos para cada objeto i con variable j-ésima como $X_{ij}(i = 1, \ldots, n; \quad j = 1, \ldots, p)$. El algoritmo propuesto se compone de los siguientes tres pasos:

Paso 1: (Seleccione los medoides iniciales)

I) Usando la distancia euclidiana como una medida de disimilitud, calcula la distancia entre cada par de todos los objetos de la siguiente manera:

$$d_{ij} = \sqrt{\sum_{\alpha=1}^{p} (X_{i\alpha} - X_{j\alpha})^2} \qquad i = 1, \ldots, n; \quad j = 1, \ldots, n \tag{2.67}$$

II) Calcular p_{ij} para hacer una estimación inicial de los centros de los grupos.

$$p_{ij} = \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \qquad i = 1, \ldots, n; \quad j = 1, \ldots, n \tag{2.68}$$

III) Calcule $\sum_{i=1}^{n} p_{ij} \quad (j = 1, \ldots, n)$, en cada objeto y ordénelos en orden ascendente. Seleccione k objetos que tengan el valor mínimo como grupo inicial medoids.

IV) Asignar cada objeto al medoid más cercano.

V) Calcular el valor óptimo actual, la suma de la distancia de todos los objetos a sus medoids.

Paso 2: (Encontrar nuevos medoids)
Reemplace el medoid actual en cada grupo por el objeto que minimiza la distancia total a otros objetos en su grupo.

Paso 3: (Nueva asignación)

i) Asignar cada objeto al nuevo medoid más cercano.

ii) Calcular el nuevo valor óptimo, la suma de la distancia de todos los objetos a sus nuevos medoids. Si el valor óptimo es igual al anterior, detenga el algoritmo. De lo contrario, vuelve al paso 2.

El algoritmo anterior se ejecuta de la misma manera que *K*-means. El rendimiento del algoritmo puede variar según el método de selección de los medoids iniciales.

Ventajas

- Es fácil de entender y fácil de implementar.
- El algoritmo *k*-medoid es rápido y converge en un número fijo de pasos.

Desventajas

- *k*-mediods es más costoso que el método *k*-means debido a su complejidad de tiempo.
- No se escala bien para grandes conjuntos de datos.
- Los resultados y el tiempo total de ejecución dependen de las particiones iniciales.

2.3.2.2. Método PAM

La **partición alrededor del medoid** es conocido por ser el más poderoso de los algoritmos de partición. Sin embargo, PAM también tiene el inconveniente de que funciona de manera ineficiente para grandes conjuntos de datos debido a su complejidad.

La principal ventaja del algoritmo es la robustez del método en presencia de ruido u Outliers, pues el cálculo del medoid está menos influido por ellos u otros valores extremos. Aunque comienza a ser muy costoso a medida que el tamaño muestral n y el número de iteraciones k aumenta, siendo una de las principales desventajas de este algoritmo, razón por la cual es eficiente solo para bases de datos pequeñas.

El Algoritmo PAM toma k grupos deseados y se considera que un conjunto aleatorio de p elementos es el conjunto de medoids. Luego, en cada paso, todos los elementos del conjunto de datos de entrada que no son medoids se examinan uno

por uno para ver si deben ser medoids. Es decir, el algoritmo determina si hay un elemento que debe reemplazar uno de los medoids existentes. Al observar todos los pares de objetos medoids y no medoids, el algoritmo elige el par que mejora la calidad general de la agrupación y los intercambia. La calidad aquí se mide por la suma de todas las distancias desde un objeto considerado como no medoid hasta el medoid para el grupo en el que se encuentra. Un elemento se asigna al grupo representado por el medoid al que está más cercano (distancia mínima o distancia directa euclidiana entre los elementos y el centro del clúster al que pertenecen).

2.3.2.3. Método CLARA

El **Agrupamiento para aplicaciones grandes** es un algoritmo de partición que ha entrado en vigor para resolver el problema de Partition Around Medoids (PAM). CLARA extiende su enfoque de k-medoids para una gran cantidad de objetos. Esta técnica selecciona arbitrariamente los datos utilizando PAM.

Los pasos del algoritmo son:

1. Dibuja una muestra de $40 + 2k$ objetos al azar de todo el conjunto de datos, y llame al algoritmo PAM para encontrar k medoid de la muestra.
2. Para cada uno de los objetos, determine el k medoid específico que es similar al objeto dado.
3. Calcula la disimilitud media de la agrupación así obtenida. Si el valor así obtenido es menor que el mínimo actual, podemos usarlo y conservar el k-medoid que se encuentra en el segundo paso como el mejor medoid.
4. Podemos repetir los pasos para la siguiente iteración.

Ventajas

- El algoritmo CLARA trata con conjuntos de datos más grandes que PAM.

Desventajas

- El rendimiento eficiente de CLARA depende del tamaño del conjunto de datos.
- Una muestra sesgada de datos puede dar como resultado una confusión y una mala agrupación de conjuntos de datos completos.

2.3.2.4. Método CLARANS

El **Agrupación de aplicaciones grandes basadas en búsquedas aleatorias** es un método de partición utilizado para grandes bases de datos. Es más eficiente y escalable que PAM y CLARA. En el cual se recomienda que se realicen los siguientes pasos:

1. Parámetros de entrada numlocal y maxneighbour.
2. Seleccione k objeto al azar de la base de datos.
3. Marca los k objetos como S_i *seleccionado* y todos los demás como S_i *no seleccionados*.
4. Calcule el costo T para el S_i *seleccionado*.
5. Si T es un conjunto de medoides de actualización negativa. De lo contrario selecciono al medoid como el óptimo local.
6. Reinicia la selección de otro conjunto de medoides y encuentre otro óptimo local.
7. Se detiene hasta que devuelve lo mejor.

CLARANS utiliza dos parámetros: numlocal y maxneighbour. Donde **numlocal** significa el número mínimo local obtenido y **maxneighbhour** significa el número máximo de vecinos examinados.

Cuanto más alto sea el valor de este último, más cerca estará CLARANS de PAM y más larga será cada búsqueda de mínimos locales. Esto es una ventaja porque la calidad de los mínimos locales es mayor y se puede encontrar un número menor de mínimos locales.

Ventajas

- Es fácil manejar los valores atípicos.
- El resultado de CLARANS es más efectivo si se compara con el método PAM y CLARA.

Desventajas

- No garantiza dar búsqueda a un área localizada.
- Utiliza muestras aleatorias para los vecinos.
- No es muy eficiente para grandes conjuntos de datos.

2.3.2.5. Método DBSCAN

Los **Métodos de agrupamiento basados en densidad** suponen que los puntos que pertenecen a cada grupo se extraen de una distribución de probabilidad específica. Este algoritmo se puede usar solo para campos esféricos en forma de agrupaciones. El mérito de tal agrupación es que tienen una densidad de puntos considerablemente mayor que fuera del agrupamiento. Este método puede ser efectivo para manejar el ruido hasta cierto punto siempre que podamos escanear el conjunto de datos de entrada, solo necesita un escaneo del conjunto de datos de entrada. La condición previa de este algoritmo es que los parámetros de densidad deben inicializarse apriori. Permite que el grupo dado crezca continuamente siempre que la densidad del vecindario supere un cierto umbral.

El algoritmo agrupa el área de alta densidad a un clúster. Hay dos parámetros importantes en el algoritmo: radio vecino del objeto Eps y el Número mínimo de vecinos del objeto MinPts. El objetivo del algoritmo DBSCAN es encontrar el conjunto de objetos unidos a la densidad, en otras palabras, descubrir la diferentes agrupaciones. Un clúster es el área de objeto de alta densidad dividida por el área de objeto de baja densidad en el espacio de datos. La idea básica de DBSCAN es que para encontrar un área de densidad combinada se van agrupando a partir de un objeto p tomado al azar. Si p es el objeto del núcleo, es decir p es el centro de un círculo, Eps es el radio y el número de objetos en el círculo es igual de grande que MinPts, entonces el algoritmo devuelve un conjunto unido a la densidad. Luego, identifica todos los objetos que se establecen como en el mismo grupo; si p no es un objeto del núcleo no se puede alcanzar ningún otro objeto desde la densidad de p, entonces p es identificado como el ruido. Los objetos unidos por densidad se agrupan en el mismo grupo, y el objeto que no pertenece a ningún clúster es ruido.

La característica sobresaliente del algoritmo DBSCAN es que puede descubrir cualquier agrupamiento de formas arbitrarias, se rompe la limitación de otros algoritmos que solo pueden descubrir el conjunto de forma fija. No es sensible al ruido, por lo que mejora enormemente la flexibilidad de agrupamiento y el resultado de la agrupación no actúa sobre el ruido.

Mejora del algoritmo

Al escanear los datos se debe establecer alguna distribución aproximada con la cual se obtenga todo el conjunto de datos de entrada, y todos los datos se dividen en varias particiones de acuerdo con la distribución aproximada. Los diferentes Eps son determinados en particiones diferentes, y se debe aplicar el algoritmo DBSCAN en diferentes particiones utilizando el Eps respectivamente hasta que el mayor nú-

mero de datos se identifique a un grupo. De esta manera, todo el procedimiento de agrupamiento es supervisado, y la mejora del resultado está en el agrupamiento de densidad.

2.4. Métodos predictivos

Los métodos para clasificación basado en árboles de decisión estratifican o segmentan el espacio del predictor en un número simple de regiones, y para obtener las predicciones se suele usar la media o moda de las observaciones de entrenamiento en la región en la que cada observación a predecir pertenece.

2.4.1. Árboles de decisión

Los árboles de decisión pueden aplicarse tanto para problemas de regresión como de clasificación, y también pueden contener predictores tanto cuantitativos como cualitativos. Son modelos de clasificación que dividen los datos en subconjuntos basados en categorías de variables de entrada. Esto es de gran ayuda a la hora de determinar las decisiones a lo largo del camino. Los árboles de decisión tienen la forma de un árbol en el que cada rama representa una elección entre el número de alternativas, y cada hoja representa una clasificación o decisión. Este es un modelo que al buscar en los datos trata de encontrar la variable que permita dividir el conjunto de datos en grupos lógicos que son más diferentes entre sí. Se usan bastante porque son fáciles de entender e interpretar. Permite controlar bien los valores que faltan y son útiles para la selección preliminar de variables.

2.4.1.1. Construcción del árbol de decisión

Los árboles de clasificación son muy similares a los de regresión, con la diferencia de que se usan para predecir una variable respuesta cualitativa, asignando la predicción para cada observación como la clase más común (moda) de observaciones de entrenamiento en la región o nodo terminal al que pertenece dicha observación de test. Para construir un árbol de clasificación se emplea el mismo la división binaria recursiva para generar el árbol. Sin embargo, como la variable respuesta es cualitativa existen varias alternativas a tomar como criterio de selección de las divisiones óptimas, todas ellas con el objetivo de encontrar nodos lo más puros/homogéneos posible. Las más empleadas son:

Error de clasificación

Se define como la proporción de observaciones que no pertenecen a la clase más común en el nodo.

$$E_m = 1 - max_k(\hat{p}_{mk}) \tag{2.69}$$

donde $\hat{p}_{mk}$ representa la proporción de observaciones del nodo m que pertenecen a la clase k.

El error de clasificación no suele ser lo suficientemente sensible en medir la pureza de los nodos para crear el árbol, por lo que en la práctica se opta por otras dos medidas, el ***índice de Gini*** o el ***cross-entropy***. Aun así, cualquiera de los tres puede usarse para la poda del árbol. El error de clasificación es preferible si el objetivo es conseguir la máxima predicción del árbol en la poda final.

Índice de Gini

Es una medida de la varianza total en el conjunto de las K clases del nodo m. Se considera una medida de pureza del nodo.

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.70}$$

Cuando $\hat{p}_{mk}$ es cercano a 0 o a 1 (el nodo contiene mayoritariamente observaciones de una clase), el término $\hat{p}_{mk}(1 - \hat{p}_{mk})$ es muy pequeño. Como consecuencia, cuanto mayor sea la pureza del nodo, menor el valor del índice Gini G_m.

Cross-entropy

La entropía es otra forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se corresponde con la impureza. Si un nodo es puro, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo de 1.

$$D = -\sum_{i=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}) \tag{2.71}$$

Independientemente de la medida empleada como criterio de selección de las divisiones, el proceso siempre es el mismo:

1. Para cada posible división se calcula el valor de la medida en cada uno de los dos nodos resultantes.

2. Se suman los dos valores ponderando cada uno por la fracción de observaciones que contiene cada nodo. Este paso es muy importante, ya que no es lo mismo dos nodos puros con 2 observaciones, que dos nodos puros con 100 observaciones.

$$\frac{n \quad \text{observaciones nodo A}}{n \quad \text{observaciones totales}} \times \text{pureza A} + \frac{n \quad \text{observaciones nodo B}}{n \quad \text{observaciones totales}} \times \text{pureza B}$$

3. La división con menor o mayor valor (dependiendo de la medida empleada) se selecciona como división óptima.

2.4.1.2. Controlar el tamaño del árbol

El tamaño final que adquiere un árbol puede controlarse mediante reglas que detengan la división de los nodos dependiendo de si se cumplen o no determinadas condiciones. El nombre de estas condiciones puede variar dependiendo del software o librería empleada, pero suelen estar presentes en todos ellos.

- **Observaciones mínimas para división**: define el número mínimo de observaciones que debe tener un nodo para poder ser dividido. Cuanto mayor el valor, menos flexible es el modelo.
- **Observaciones mínimas de nodo terminal**: define el número mínimo de observaciones que deben tener los nodos terminales. Su efecto es muy similar al de observaciones mínimas para división.
- **Profundidad máxima del árbol**: define la profundidad máxima del árbol, entendiendo por profundidad máxima el número de divisiones de la rama más larga (en sentido descendente) del árbol.
- **Número máximo de nodos terminales**: define el número máximo de nodos terminales que puede tener el árbol. Una vez alcanzado el límite, se detienen las divisiones. Su efecto es similar al de controlar la profundidad máxima del árbol.
- **Reducción mínima de error**: define la reducción mínima de error que tiene que conseguir una división para que se lleve a cabo.

2.4.1.3. Ventajas y desventajas de los árboles

Una vez visto el procedimiento de construcción del árbol se describen algunos aspectos deseables y no deseables de los modelos.

Entre las características deseables se pueden identificar las siguientes:

- Puede ser aplicado para cualquier estructura de datos a través de una formulación apropiada del conjunto de cuestiones.
- La clasificación final tiene una forma simple que puede ser almacenada de manera compacta y clasifica eficientemente nuevos datos.
- La selección de variables se hace paso a paso, automático y reduciendo el coste de complejidad. Se busca nodo a nodo hasta conseguir la división más significativa. En cada etapa se intenta extraer la información más relevante de la parte del espacio que se está trabajando.
- Proporciona, no sólo una clasificación si no también una estimación de la probabilidad de clasificar un objeto erróneamente.
- Es muy robusto respecto a los outliers y los puntos mal clasificados.
- Es muy fácil su interpretación.
- Los árboles toma decisiones muy cercanas a las que tomaría un humano.
- Se pueden visualizar gráficamente.
- Pueden manejar fácilmente predictores cualitativos sin la necesidad de crear variables ficticias y las posibles interacciones se incluyen automáticamente.
- En conjunto de datos grandes puede revelar estructuras complejas.
- Al tratarse de métodos no paramétricos, no es necesario que se cumpla ningún tipo de distribución específica.
- Son muy útiles en la exploración de datos, permiten identificar de forma rápida y eficiente las variables más importantes.
- Son capaces de seleccionar predictores de forma automática.

Entre las características no deseables se pueden identificar las siguientes desventajas:

- Clasifica de manera aleatoria cuando tenemos valores perdidos.
- Aunque la optimalidad se aplique a cada división, esto no significa que el árbol sea óptimo.
- Las variables predictoras continuas han de ser discretizadas.

- Es posible que las interacciones débiles se impongan a las más fuertes.
- Los árboles grandes tienen tendencias a sobre ajustarse a los datos.
- Son inestables, es decir, pequeños cambios en los datos iniciales pueden producir árboles muy distintos.

2.4.2. Bosques aleatorios

El algoritmo de bosque aleatorio es la clave que se esconde en actividades de personalización automatizada y segmentación automática. Con bosque aleatorio se combinan cientos de árboles de decisión para conseguir una mejor predicción que la que se conseguiría con un solo árbol.

El objetivo de un árbol de decisión es desglosar todos los datos de vistas disponibles en los que un sistema puede aprender y agruparlos de modo que las vistas de cada grupo sean lo más parecidas posibles las unas a las otras con relación a la métrica objetivo.

Entre grupos, sin embargo, las vistas son lo más diferentes posibles con relación a la métrica objetivo (por ejemplo, la tasa de conversión). El árbol de decisión tiene en cuenta las diferentes variables existentes en el conjunto de formación para determinar cómo dividir los datos de forma MECE (mutuamente exclusiva, colectivamente exhaustiva) en estos grupos (u "hojas") para maximizar este objetivo.

La construcción de cada árbol, dado un conjunto de entrenamiento de tamaño n con p variables predictoras, se realiza según las siguientes indicaciones:

1. Seleccionar una muestra con reemplazamiento de tamaño n de la muestra de entrenamiento.
2. En cada nodo del árbol construido en cada muestra se eligen aleatoriamente $m < p$ variables predictoras, y se elige la mejor división entre esas m variables.
3. Cada árbol se construye hasta alcanzar un tamaño razonablemente grande, sin realizar poda.

La tasa de error del modelo final depende de dos elementos:

a) La correlación entre dos árboles cualesquiera del bosque. A mayor correlación, menor error.

b) La fuerza de cada árbol en el bosque. Un árbol con una tasa de error reducida es un clasificador fuerte. Al aumentar la fuerza de los árboles individuales, disminuye la tasa de error del bosque.

Disminuir m reduce tanto la correlación como la fuerza, y viceversa. Este es el único parámetro a ajustar respecto al cual Bosques Aleatorios es sensible, puede ser ajustado con la ayuda de procedimientos de validación cruzada.

A continuación, se detallan algunas de las ventajas de la aplicación de los modelos Bosques Aleatorios.

1. Proporciona muy buenos resultados en los estudios empíricos.
2. Se ejecuta de forma eficiente sobre grandes bases de datos.
3. Puede tratar miles de variables sin tener que eliminar ninguna.
4. Proporciona estimaciones de la importancia de cada variable.
5. Dispone de un método efectivo de estimación de valores perdidos.
6. Calcula proximidades entre pares de casos que pueden emplearse en análisis de conglomerados, identificación de outliers, o escalamiento de los datos para obtener representaciones gráficas.

Capítulo 3

METODOLOGÍA

3.1. Tipo de investigación

Esta investigación tiene las siguientes características:

- **Aplicativo:** Se pretende tener una base de datos a partir de la cual se pueda generar el modelo probabilístico, las simulaciones a realizar se trabajarán en el software estadístico libre R.
- **Bibliográfico:** Ya que se ha hecho una recolección de libros impresos y digitales, artículos académicos, entre otros, para contar con el material necesario y suficiente que nos permita cubrir las necesidades de la presente investigación.
- **Deductivo:** Se parten de resultados generales para aplicarlos en áreas concretas como las finanzas.
- **Sistemático:** Se ha empleado una estructura coherente y ordenada de lo que se debe seguir en la presente investigación para obtener los resultados esperados.

3.2. Diseño de la investigación

3.2.1. Forma de Trabajo

- Revisión de la bibliografía a seguir.
- Reuniones semanales con el docente asesor de la investigación para discutir asuntos de la teoría como de la parte aplicada de la investigación, así como las correcciones que se le deban hacer al trabajo escrito.
- Reuniones semanales en la institución financiera que ha proporcionado los datos para fines estrictamente académicos, para lo cual ha pedido estricta confidencialidad y ha dispuesto un área de trabajo dentro dicha institución para el uso y control unilateral de los datos.

3.2.2. Cronograma de actividades

ACTIVIDADES	MESES											
AÑO 2019	E	F	M	A	M	J	J	A	S	O	N	D
Elaboración de Perfil	X											
Revisión y corrección de Perfil	X	X										
Defensa de Perfil		X										
Elaboración del Capítulo 2			X	X	X	X						
Revisión y corrección del Capítulo 2						X	X					
Elaboración del Capítulo 3							X					
Revisión y corrección del Capítulo 3								X				
Elaboración del Capítulo 4	X	X	X	X	X	X	X	X	X			
Revisión y corrección del Capítulo 4					X	X	X	X	X	X	X	
Elaboración del Capítulo 5 y 6											X	
Revisión y corrección del Capítulo 5 y 6											X	
Elaboración de la Introducción											X	
Defensa Pública Final												X

Tabla 3.1: Cronograma de actividades en la Investigación

3.3. Recolección y procesamiento de información

3.3.1. Recopilación de información

La información bibliográfica utilizada para el desarrollo de la investigación se obtuvo a partir del análisis documental. Además, se emplearon fichas de trabajo para recolectar la información y hacer anotaciones importantes de fuentes bibliográficas; como libros, revistas científicas y electrónicas. La bibliografía consultada se encontró, en su mayoría, en las bibliotecas de la Universidad de El Salvador.

Las fuentes de información que se utilizaron en esta investigación facultaron el sustento teórico y metodológico del trabajo. Asimismo permitieron el acceso y ampliación del conocimiento sobre el tema en estudio.

También se tuvo en cuenta los documentos y registros de las bases de datos seleccionadas y analizadas para la realización de la investigación.

3.3.2. Procesamiento de información

El tratamiento de las fuentes antes mencionadas se realizó mediante el análisis crítico de toda la información adquirida, para ampliar conocimientos y argumentar de forma teórica y practica el trabajo.

Capítulo 4

RESULTADOS

En este capítulo se presenta el estudio realizado para la identificación de variables que determinan la clusterización de clientes, analizando la influencia de las variables mediante dos enfoques importantes para la generación de modelos. Lo que se pretende con los modelos es desarrollar una estrategia de negocios, la cual consiste en generar movilidad positiva para cada uno de los clientes, mejorando su clasificación básica predefinida. En el primer enfoque se aplican los métodos de aprendizaje no supervisado, que nos permiten determinar como se agrupan los datos según el valor de sus variables, lo cual incluye un análisis exploratorio de datos, comenzando con un análisis en componentes principales para comprobar si es posible reducir la dimensión de los datos según el número de componentes a utilizar para observar gráficamente la formación de los grupos y aplicar seguidamente los métodos de clusterización para analizar las variables que influyen en la formación de cada clúster.

El segundo enfoque tiene que ver con métodos de aprendizaje supervisado, como árboles de decisión y bosques aleatorios. Estás herramientas son las más adecuados para la creación de modelos en el que se describen las rutas a seguir, así como también, para determinar las variables a utilizar si queremos maximizar la representación de los datos o minimizar el error de la predicción. Para nuestro caso, este tipo de enfoque se aplica a la identificación de variables, y que tiene como objetivo primordial el estudio estructural y predictivo, utilizando árboles de decisión para la creación de rutas a través del valor de las variables que influyen en la clusterización de los clientes.

En esta investigación se parte de una clasificación básica que la institución financiera ha hecho sobre sus clientes: clientes A (***Cuadrante 1***), clientes M (***Cuadrante 2***), clientes B (***Cuadrante 3***) y clientes I (***Cuadrante 4***), se ha realizado una segmentación previa con varias variables que la institución considera que son las más importantes para clasificar a cada cliente en ***16 Grupos*** por mes según los valores

que posean en dichas variables, de esta manera se podrá graficar a los clientes en el plano cartesiano según los valores de las variables. Por lo que se busca a través del análisis de los datos históricos indicar cuantos grupos o clúster se pueden formar, cuales son las variables que determinan el pertenecer a cada clúster y comparar los clientes de cada clúster con los clientes de la segmentación ya antes realizada por ellos. Teniendo esos resultados se buscara realizar su respectivo análisis y generar modelos que permitan identificar rutas que involucren productos financieros, épocas del año, ciclos y tipos de abordajes propicios que se debe recomendar a los clientes para generar una movilidad positiva.

4.1. Base de Datos

Se parte del registro histórico de los datos reales proporcionados por cierta institución financiera (que nos ha solicitado el anonimato y el cuido celoso de los datos), dichos datos están compuesto por los registros de los clientes de los últimos 4 años (2015 - 2018), donde cada observación consta de un código (número de identificación) que diferencia un cliente de otro; y para todos los clientes se tienen las mismas observaciones (105 variables). Es importante mencionar que cada variable está enmascarada por lo que no se conoce el significado de cada variable pero si el tipo de variables (cualitativa o cuantitativa), además la institución determinó las condiciones para realizar la investigación. Es decir, se realizo dentro de sus instalaciones y con la supervisión de uno de sus colaboradores.

En la formación de la base de datos se tenía que cada cliente aparecía en diferentes tablas, donde cada tabla estaba compuesta por varias variables las cuales describen la relación Cliente - Institución que se tiene. Por lo que fue necesario realizar una concatenación de las tablas disponibles para tener un registro único de clientes. Realizando lo anterior se tiene una Base de Datos compuesta por todos los clientes y el valor registrado en cada variable durante los doce meses de los cuatro años disponibles.

Los clientes según su tamaño corporativo se dividen en Clientes Tipo #1 y Clientes Tipo #2. Además para el estudio es importante diferenciar a los clientes por su tamaño corporativo y variable de clasificación. Es así, que los datos se pueden dividir en 12 Clases, en los cuales únicamente se han tomado en cuenta aquellos clientes que aparecen en los 12 meses de cada año de estudio.

Por lo que, en la Tabla 4.1 se presenta el número de clientes de Tipo #1 que se tiene según las Clases para cualquier mes de estudio.

Nombre de las Clases	Número de Clientes
VS2_CLI1	473
VS2_CLI2	3
VS2_CLI3	197
VS2_CLI4	18
VS2_CLI5	207
VS2_CLI6	262
VS2_CLI7	30
VS2_CLI8	60
VS2_CLI9	197
VS2_CLI10	654
VS2_CLI11	1118
VS2_CLI12	304

Tabla 4.1: Distribución de clientes de Tipo #1 según su Clase

También, en la Tabla 4.2 se presenta el número de clientes de Tipo #2 que se tiene según las Clases para cualquier mes de estudio.

Nombre de las Clases	Número de Clientes
VS2_CLI1	373
VS2_CLI2	1
VS2_CLI3	68
VS2_CLI4	3
VS2_CLI5	165
VS2_CLI6	151
VS2_CLI7	14
VS2_CLI8	32
VS2_CLI9	60
VS2_CLI10	392
VS2_CLI11	148
VS2_CLI12	274

Tabla 4.2: Distribución de clientes de Tipo #2 según su Clase

Por lo tanto, se manejan dos Bases de Datos; en la que una incluye a todos los Clientes de Tipo #1 la cual posee 3604 Clientes por mes y otra que incluye a todos los Clientes de Tipo #2 que poseen 1717 Clientes por mes; para ambas bases de datos se manejan 105 variables de las cuales solo se utilizarán las siguientes:"$IDCLI$","$V1DS$","$V2DS$","$V3DS$","$V4DS$","$V1FN$","$V2FN$","$V3FN$", "$VRCP1$","$VRCP2$","$VRCP3$","$VRCP1entreVRCP3$","$PRO1$","$PRO2$","$PRO3$",

"PRO4","PRO5","PRO6","PRO7","PRO8","PRO9","PRO10","PRO11","PRO12",
"PRO13","PRO14","PRO15","PRO16","PRO17","PRO18","PRO19","PRO20",
"PRO21","PRO22","PRO23","PRO24","PRO25","PRO26","PRO27","PRO28",
"PRO29","PRO30","PRO31","PRO32","PRO33","PRO34","PRO35","PRO36",
"PRO37","PRO38","CANTIDAD_PRO","TOTAL_PRO_DIF","BC1","BC2",
"BC3","BC4","BC5","BC6","BC7","BC8","BC9","BC10","BC11","BC12","BC13",
"BC14","INDICE_BC","TV1","TV2","TV3","TV4","R1","R2","R3","R4","R6",
"P1","P2","P3","P4","Cy","Cx","Grupo","Cuadrante".

4.2. Enfoque Descriptivo

La finalidad es examinar los datos previamente a la aplicación de cualquier técnica estadística, además de explorar y organizar la información que nos dan los datos ha manera de detectar algún patrón de comportamiento, de modo tal que sobresalga su estructura y las relaciones existentes entre las variables analizadas.

Por lo tanto, se ha seleccionado la Base de Datos de los Clientes de Tipo #1 y la Clase VS2_CLI11, por el hecho de que es el sector que posee mayor cantidad de clientes y se tomará en cuenta únicamente 76 variables.

4.2.1. Análisis en Componentes Principales

Con el conjunto de datos se aplica un análisis en componentes principales para determinar que tanto se puede reducir la dimensión transformando las variables en componentes

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 1	7.564801526	17.59256169	17.59256169
comp 2	4.73396907	11.0092304	28.60179208
comp 3	3.650648757	8.489880829	37.09167291
comp 4	3.041048222	7.072205167	44.16387808
comp 5	2.616287477	6.084389481	50.24826756
comp 6	2.211500673	5.143024822	55.39129238
comp 7	2.119663954	4.929451056	60.32074344
comp 8	2.03982586	4.74378107	65.06452451

(Continúa en la página siguiente)

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 9	1.752946719	4.076620277	69.14114479
comp 10	1.425158956	3.314323154	72.45546794
comp 11	1.322105732	3.074664493	75.53013243
comp 12	1.120741026	2.606374479	78.13650691
comp 13	1.033944122	2.404521215	80.54102813
comp 14	0.948739302	2.206370471	82.7473986
comp 15	0.881673303	2.05040303	84.79780163
comp 16	0.818705233	1.903965657	86.70176728
comp 17	0.736946387	1.713828807	88.41559609
comp 18	0.643435234	1.49636101	89.9119571
comp 19	0.606542621	1.410564234	91.32252133
comp 20	0.525810795	1.222815802	92.54533714
comp 21	0.44315148	1.030584836	93.57592197
comp 22	0.42853167	0.996585278	94.57250725
comp 23	0.412152844	0.958494985	95.53100224
comp 24	0.390419985	0.907953454	96.43895569
comp 25	0.375734419	0.873800975	97.31275667
comp 26	0.287387019	0.668341905	97.98109857
comp 27	0.286686115	0.666711896	98.64781047
comp 28	0.221494683	0.515103915	99.16291438
comp 29	0.159436632	0.370782864	99.53369725
comp 30	0.111610721	0.259559817	99.79325706
comp 31	0.064281852	0.149492679	99.94274974
comp 32	0.019477735	0.045297058	99.9880468
comp 33	0.005123001	0.011913956	99.99996076
comp 34	1.69E-05	3.92E-05	100
comp 35	2.93E-10	6.80E-10	100

Tabla 4.3: Porcentaje de varianza acumulada por Componentes de los datos de Enero 2015 para la Clase VS2_CLI11

En la tabla 4.3 se muestra el porcentaje acumulado de las componentes en el mes de Enero del año 2015, considerando para efecto de análisis que podemos seleccionar 34 componentes y estaríamos utilizando el 100 % de la representación de los datos. Lo que nos permitiría pasar de 76 variables a trabajar con 34 componentes, es decir, reduciríamos más de la mitad la dimensión de trabajo.

Es importante mencionar que los Clientes no se comportan de la misma manera, por lo que no podemos asumir que se trabajara con el mismo número de componentes para todos los años en el mes de Enero e incluso para los demás meses obteniendo el mismo porcentaje de representatividad. Por lo que, en la siguientes tablas se muestra el porcentaje de varianza acumulado para el mes de Enero en los años 2016 - 2018 respectivamente.

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 1	6.55473708	15.2435746	15.2435746
comp 2	3.86561165	8.98979454	24.2333692
comp 3	3.7266451	8.66661651	32.8999857
comp 4	3.15513527	7.33752388	40.2375095
comp 5	2.92519529	6.80277974	47.0402893
comp 6	2.25894627	5.25336342	52.2936527
comp 7	1.94462313	4.52237938	56.8160321
comp 8	1.69539035	3.94276826	60.7588003
comp 9	1.53047899	3.55925346	64.3180538
comp 10	1.38801347	3.22793829	67.5459921
comp 11	1.2204376	2.83822699	70.3842191
comp 12	1.18290943	2.75095216	73.1351712
comp 13	1.07024271	2.48893654	75.6241078
comp 14	1.05803467	2.46054574	78.0846535
comp 15	1.01179616	2.35301431	80.4376678
comp 16	0.97993169	2.27891091	82.7165787
comp 17	0.94388644	2.19508474	84.9116635
comp 18	0.89010413	2.0700096	86.9816731
comp 19	0.78212556	1.81889666	88.8005697
comp 20	0.69274226	1.6110285	90.4115982
comp 21	0.6484233	1.50796116	91.9195594
comp 22	0.57590201	1.339307	93.2588664
comp 23	0.51048029	1.18716347	94.4460299
comp 24	0.4390903	1.02114023	95.4671701
comp 25	0.42433096	0.98681618	96.4539863
comp 26	0.34675698	0.80641159	97.2603979
comp 27	0.33000863	0.76746194	98.0278598
comp 28	0.25903234	0.6024008	98.6302606
comp 29	0.21049513	0.48952355	99.1197842

(Continúa en la página siguiente)

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 30	0.13703768	0.31869229	99.4384765
comp 31	0.11411848	0.26539182	99.7038683
comp 32	0.07559464	0.17580148	99.8796698
comp 33	0.01789661	0.04162003	99.9212898
comp 34	0.01693474	0.03938313	99.9606729
comp 35	0.01652738	0.03843577	99.9991087
comp 36	0.00037886	0.00088106	99.9999898
comp 37	4.40E-06	1.02E-05	100
comp 38	5.28E-10	1.23E-09	100

Tabla 4.4: Porcentaje de varianza acumulada por Componentes de los datos de Enero 2016 para la Clase VS2_CLI11

La Tabla 4.4 nos indica que podemos trabajar con 37 componentes. En este caso vemos que ha comparación de Enero del año 2015 tenemos que se ha aumentado en 3 componentes.

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 1	7.33421384	14.1042574	14.1042574
comp 2	4.6957723	9.03033134	23.1345887
comp 3	3.72563813	7.16468871	30.2992774
comp 4	3.03392237	5.83446609	36.1337435
comp 5	2.84179766	5.4649955	41.598739
comp 6	2.69164148	5.17623361	46.7749726
comp 7	2.27730086	4.37942474	51.1543974
comp 8	2.06291961	3.96715309	55.1215505
comp 9	1.77434869	3.41220903	58.5337595
comp 10	1.67174325	3.21489087	61.7486504
comp 11	1.51111787	2.90599591	64.6546463
comp 12	1.48678827	2.85920822	67.5138545
comp 13	1.33704273	2.57123602	70.0850905
comp 14	1.18587974	2.28053796	72.3656285
comp 15	1.08494501	2.08643271	74.4520612

(Continúa en la página siguiente)

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 16	1.05661086	2.03194397	76.4840051
comp 17	1.04792149	2.01523363	78.4992388
comp 18	0.99846398	1.92012303	80.4193618
comp 19	0.96383566	1.85353012	82.2728919
comp 20	0.89810663	1.72712813	84.00002
comp 21	0.81067224	1.55898507	85.5590051
comp 22	0.73553504	1.41449046	86.9734956
comp 23	0.65326551	1.25627983	88.2297754
comp 24	0.62219996	1.19653839	89.4263138
comp 25	0.59140872	1.13732447	90.5636383
comp 26	0.5592688	1.07551693	91.6391552
comp 27	0.49231861	0.94676656	92.5859217
comp 28	0.46986786	0.90359203	93.4895138
comp 29	0.43604341	0.83854502	94.3280588
comp 30	0.4041565	0.77722403	95.1052828
comp 31	0.38496344	0.74031431	95.8455971
comp 32	0.36918109	0.70996364	96.5555608
comp 33	0.34055303	0.65490966	97.2104704
comp 34	0.31978481	0.61497078	97.8254412
comp 35	0.27645538	0.53164496	98.3570862
comp 36	0.2415652	0.46454846	98.8216346
comp 37	0.16664659	0.32047421	99.1421089
comp 38	0.15272309	0.29369825	99.4358071
comp 39	0.11402667	0.21928205	99.6550892
comp 40	0.08071657	0.15522417	99.8103133
comp 41	0.0484908	0.09325154	99.9035649
comp 42	0.02403431	0.04621982	99.9497847
comp 43	0.01292159	0.02484921	99.9746339
comp 44	0.00773722	0.01487926	99.9895132
comp 45	0.00541198	0.01040766	99.9999208
comp 46	3.42E-05	6.58E-05	99.9999866
comp 47	6.61E-06	1.27E-05	99.9999993
comp 48	3.60E-07	6.92E-07	100
comp 49	1.48E-10	2.84E-10	100
comp 50	1.42E-14	2.74E-14	100

(Continúa en la página siguiente)

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada

Tabla 4.5: Porcentaje de varianza acumulada por Componentes de los datos de Enero 2017 para la Clase VS2_CLI117

La Tabla 4.5 nos indica que podemos trabajar con 48 componentes. En este caso vemos que ha comparación de Enero del año 2016 tenemos que se ha aumentado en 11 componentes.

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 1	7.01049259	12.7463502	12.7463502
comp 2	4.75312842	8.64205168	21.3884018
comp 3	3.79134825	6.89336046	28.2817623
comp 4	3.71648851	6.75725184	35.0390141
comp 5	3.14993487	5.72715432	40.7661684
comp 6	2.82773394	5.14133443	45.9075029
comp 7	2.40142209	4.36622198	50.2737249
comp 8	1.98595778	3.61083232	53.8845572
comp 9	1.9146236	3.48113381	57.365691
comp 10	1.89029998	3.43690906	60.8026
comp 11	1.62856282	2.96102331	63.7636234
comp 12	1.48148502	2.69360913	66.4572325
comp 13	1.30179409	2.36689834	68.8241308
comp 14	1.13937351	2.07158821	70.895719
comp 15	1.10118568	2.00215578	72.8978748
comp 16	1.07381184	1.95238516	74.85026
comp 17	1.02779621	1.86872039	76.7189804
comp 18	1.01200823	1.84001496	78.5589953
comp 19	0.99179677	1.80326686	80.3622622
comp 20	0.92329833	1.67872423	82.0409864
comp 21	0.88665843	1.61210624	83.6530927
comp 22	0.85554072	1.55552857	85.2086212
comp 23	0.80620611	1.46582929	86.6744505
comp 24	0.74347649	1.35177544	88.026226

(Continúa en la página siguiente)

Componentes	Varianza	Porcentaje de Varianza	Porcentaje de Varianza Acumulada
comp 25	0.69062397	1.25567995	89.2819059
comp 26	0.64254357	1.16826103	90.4501669
comp 27	0.60347068	1.09721941	91.5473863
comp 28	0.58771141	1.06856621	92.6159526
comp 29	0.55897196	1.01631266	93.6322652
comp 30	0.49131001	0.89329093	94.5255561
comp 31	0.43286114	0.78702025	95.3125764
comp 32	0.38256527	0.69557322	96.0081496
comp 33	0.35804688	0.65099433	96.6591439
comp 34	0.33193153	0.60351187	97.2626558
comp 35	0.29674277	0.53953232	97.8021881
comp 36	0.28013491	0.5093362	98.3115243
comp 37	0.2320123	0.42184054	98.7333649
comp 38	0.20099384	0.36544335	99.0988082
comp 39	0.1460483	0.26554236	99.3643506
comp 40	0.09862798	0.17932361	99.5436742
comp 41	0.08711347	0.15838813	99.7020623
comp 42	0.05710342	0.1038244	99.8058867
comp 43	0.04340355	0.07891554	99.8848023
comp 44	0.03176032	0.05774603	99.9425483
comp 45	0.01255979	0.02283598	99.9653843
comp 46	0.00944293	0.01716896	99.9825532
comp 47	0.00793547	0.01442813	99.9969814
comp 48	0.00164758	0.00299559	99.9999769
comp 49	1.08E-05	1.95E-05	99.9999965
comp 50	1.22E-06	2.21E-06	99.9999987
comp 51	7.10E-07	1.29E-06	100
comp 52	3.34E-10	6.07E-10	100

Tabla 4.6: Porcentaje de varianza acumulada por Componentes de los datos de Enero 2018 para la Clase VS2_CLI11

La Tabla 4.6 nos indica que podemos trabajar con 51 componentes. En este caso vemos que ha comparación de Enero del año 2017 tenemos que se ha aumentado en 3 componentes.

Dado lo anterior, para el análisis de los datos utilizando componentes principa-

les no es necesario utilizar el 100 % de representatividad, en el caso de trabajar con base de datos más grandes podemos utilizar el número de componentes principales que nos de un porcentaje de representatividad aceptable. En nuestro caso si tomaremos el número de componentes que acumule el 100 %. Es importante mencionar que con la cantidad de componentes a seleccionar no es posible mostrar gráficamente la formación de los grupos, lo cual se mostrara más adelante.

4.2.2. Clusterización

La clusterización o clustering consiste en dividir la base de datos en grupos diferentes, la meta principal es encontrar grupos que son diferentes entre ellos y que sus miembros sean similares entre si dentro del grupo para analizar la correlación entre las variables. La herramienta que permite identificar tales grupos es la clusterización jerárquica o *k*-means. Es de mencionar además, que para aplicar cualesquiera de estas herramientas es necesario conocer el número de clúster óptimo y para ello simulamos utilizando *k*-means la corrida de 30 modelos en los cuales se va aumentando el número de grupos y graficamos la inercia intra clase, para conocer dicho número óptimo. En primer lugar lo que se busca es verificar si se forma la misma cantidad de segmentos que la institución a hecho sobre sus clientes, de ser así, comparar los clientes que lo forman y las variables que influyen. En caso contrario, formar los grupos, observar los clientes que los forman y analizar las variables que influyen.

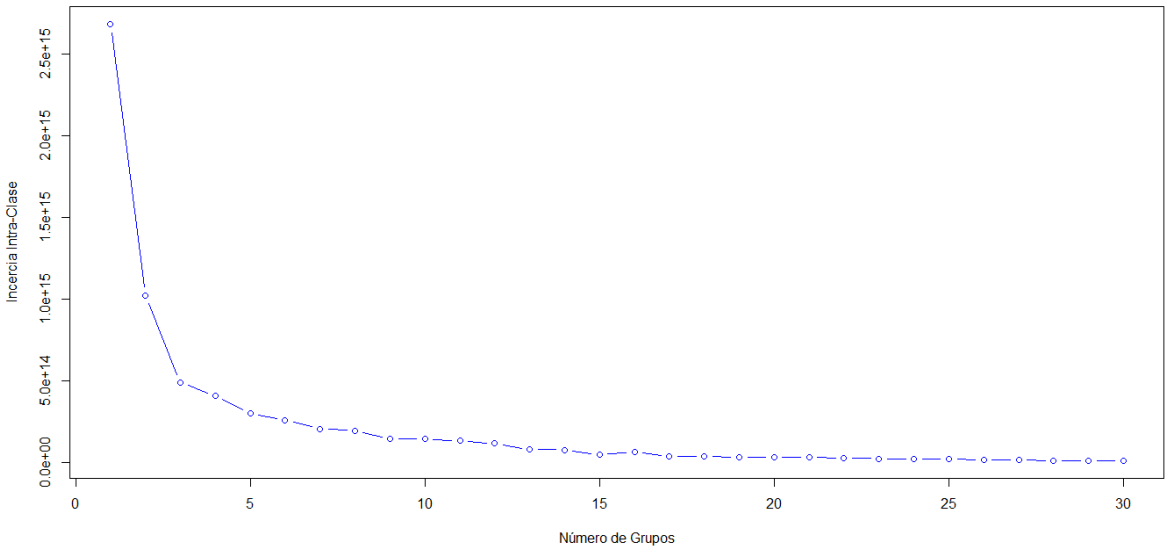


Figura 4.1: Número de Clúster óptimo para la Clase VS2_CLI11 Enero 2015

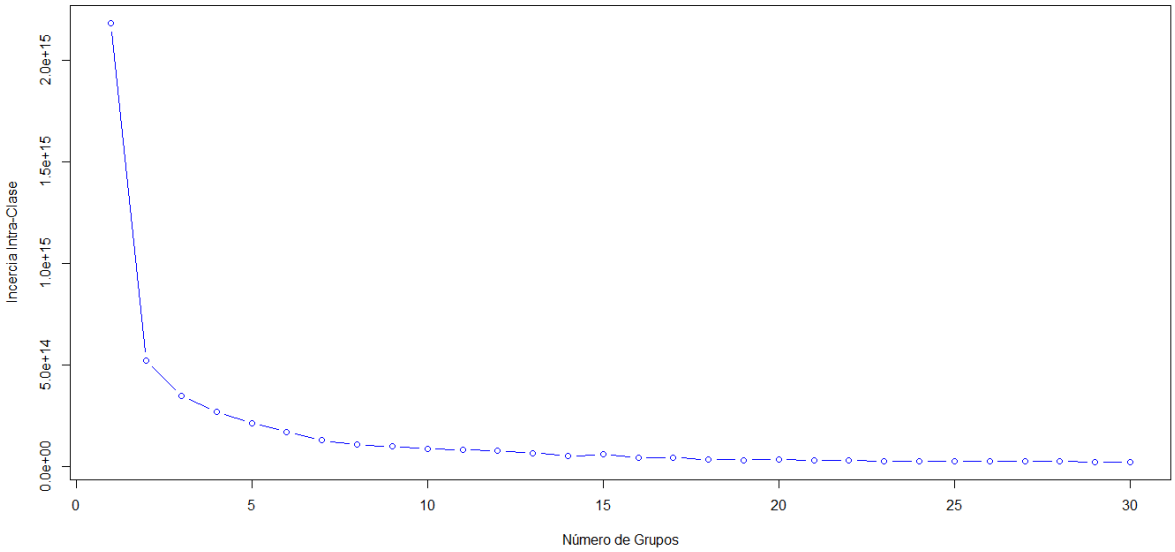


Figura 4.2: Número de Clúster óptimo para la clase VS2_CLI11 Enero 2016

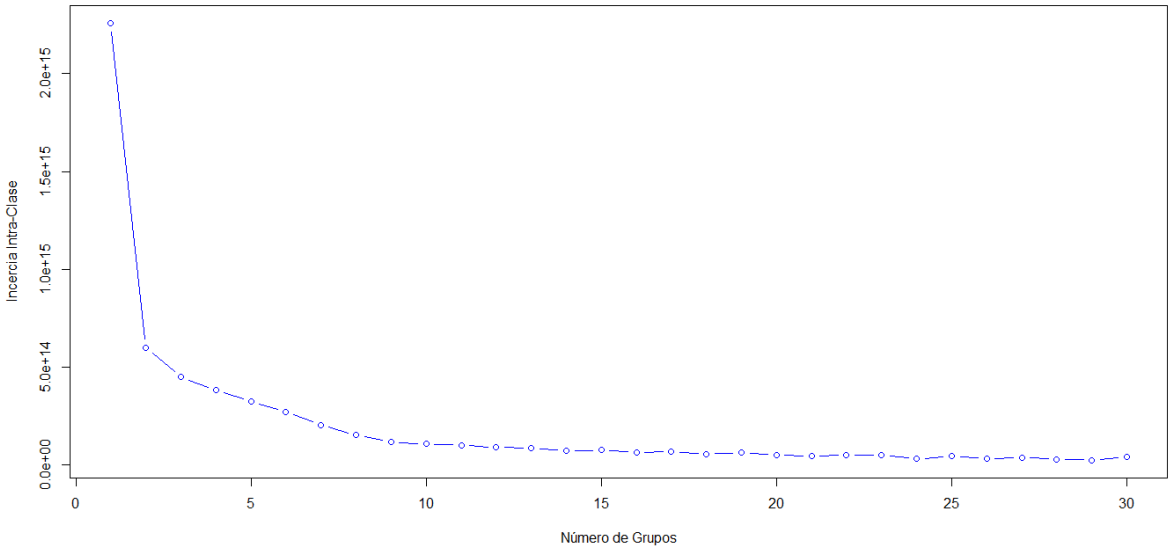


Figura 4.3: Número de Clúster óptimo para la Clase VS2_CLI11 Enero 2017

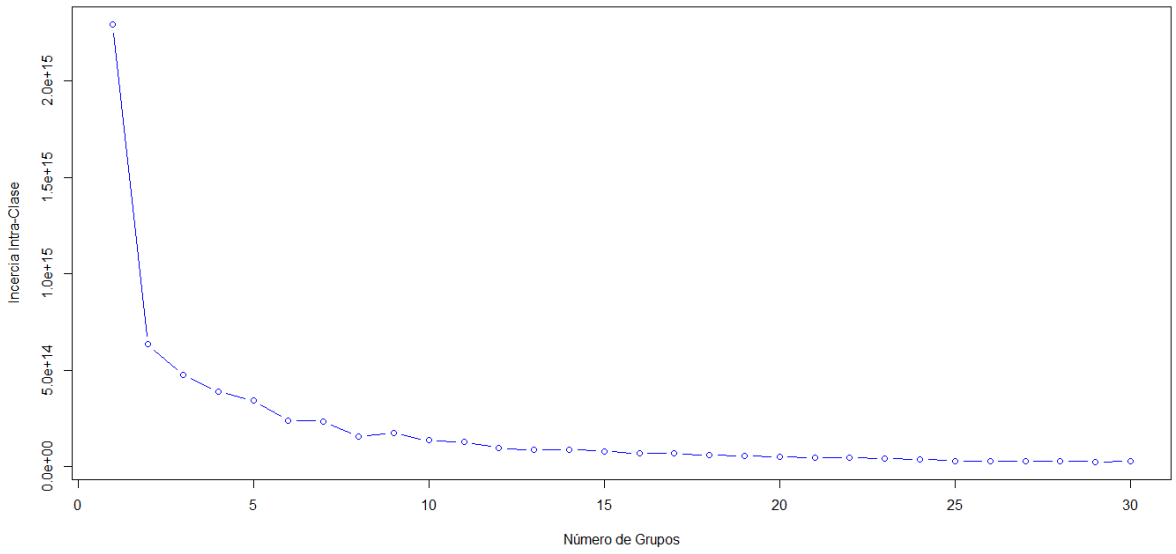


Figura 4.4: Número de Clústers óptimo para la Clase VS2_CLI11 Enero 2018

Al analizar las cuatro figuras anteriores se determina que el número óptimo para formar los clúster es 3, aunque claramente las gráficas son diferentes lo que se busca es crear el mismo número de clúster en cada mes con el objetivo de comparar la formación de los grupos para el mes de Enero en los años 2015 - 2018 y determinar las variables que influyen en cada uno de ellos.

En la Figura 4.5, la Figura 4.6, la Figura 4.7 y la Figura 4.8, se muestra como quedan distribuidos los grupos en cada uno de los meses y podemos decir que no existe una relación de similitud para cada uno de los clúster en cada mes con respecto al del siguiente año, es decir el clúster 3 de Enero 2015 es diferente al clúster 3 de Enero 2016 y así sucesivamente para los demás meses y los demás clúster. Lo anterior se verifica analizando como quedan compuesto los clúster. Además, la representación gráfica de los datos es muy mala con respecto a cada mes ya que la suma de la representación de los ejes no supera el 30 %. En tal sentido, para poder realizar un mejor análisis de los clúster gráficamente se tendrían que utilizar al menos unas 16 componentes para observar a detalle cada clúster los cual no se podría, por lo que es necesario presentar de manera independiente las variables que influyen en la formación de cada clúster, determinar la relación con respecto a los otros meses y la importancia que tiene cada una de ellas.

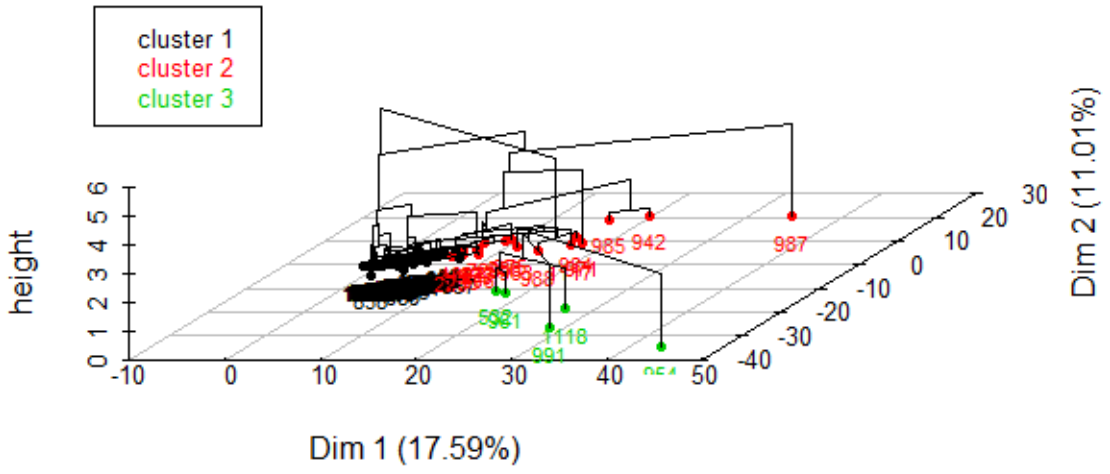


Figura 4.5: Clusterización Jerárquica para la Clase VS2_CLI11 Enero 2015

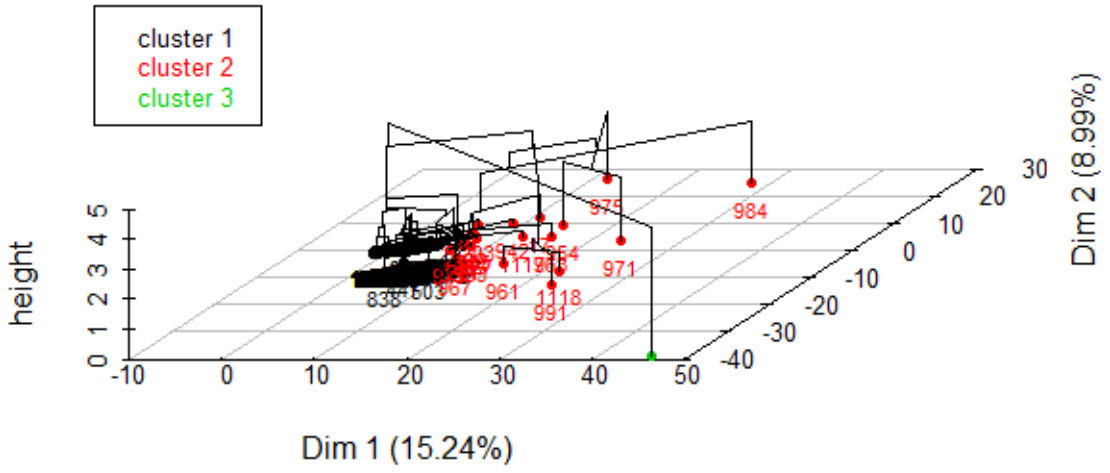


Figura 4.6: Clusterización Jerárquica para la Clase VS2_CLI11 Enero 2016

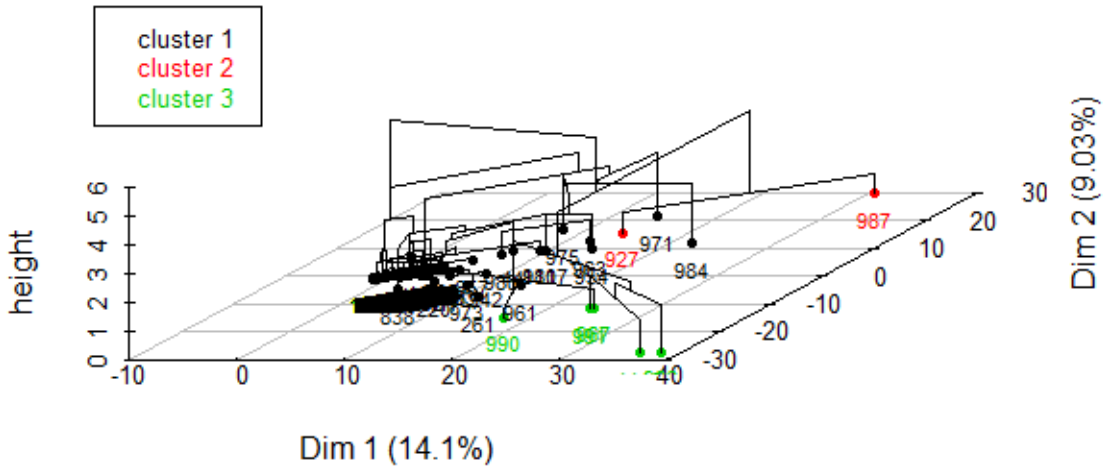


Figura 4.7: Clusterización Jerárquica para la Clase VS2_CLI11 Enero 2017

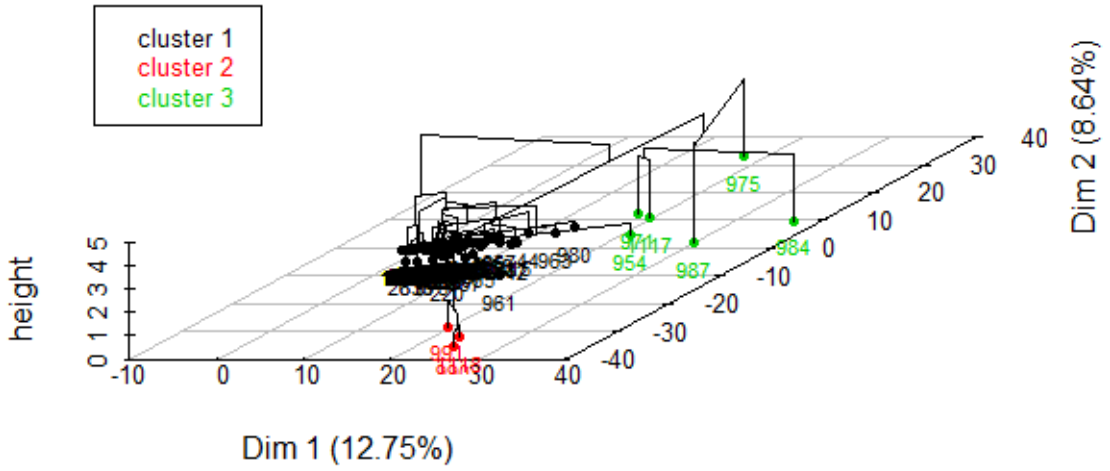
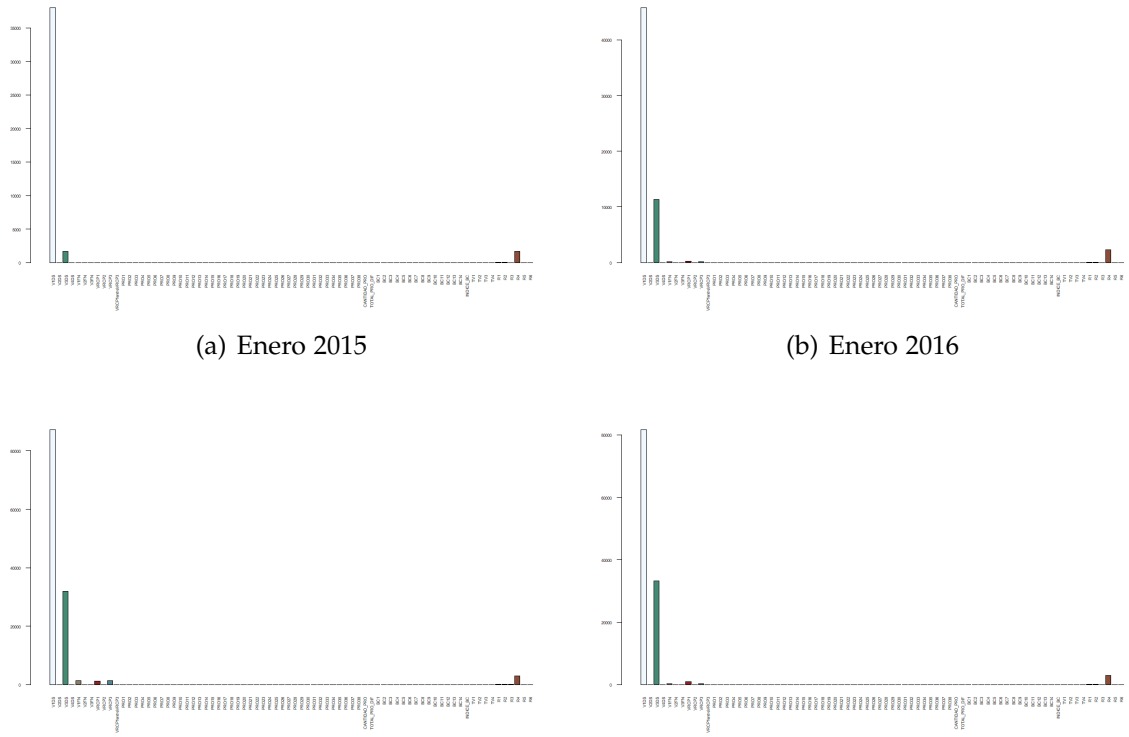


Figura 4.8: Clusterización Jerárquica para la Clase VS2_CLI11 Enero 2018

En la Figura 4.9 se muestra las variables que influyen en la creación del Clúster 1 para el mes de Enero en los años 2015 - 2018, tomando en cuenta que estamos analizando la Clase VS2_CLI11, para los clientes de Tipo #1. Se observa que los

clientes que pertenecen al clúster 1 del año 2015 utilizan más las variables ***V1DS, V3DS, R1, R2, R3, R4*** para su actividad económica, los del año 2016 utilizan más las variables ***V1DS, V3DS, V1FN, V2FN, V3FN, VRCP1, VRCP3, R1, R2, R4***, los del año 2017 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP3, R1, R2, R3, R4*** y los del año 2018 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2, VRCP3, R1, R2, R4***, que tan importante, rentable o conveniente es que los clientes del clúster 1 utilicen estas variables lo determina la institución dependiendo de lo que representa para ellos estas variables y el valor de influencia que tiene para cada clúster, ya que cambia el valor que alcanzan las variables para los diferentes años.



(a) Enero 2015

(b) Enero 2016

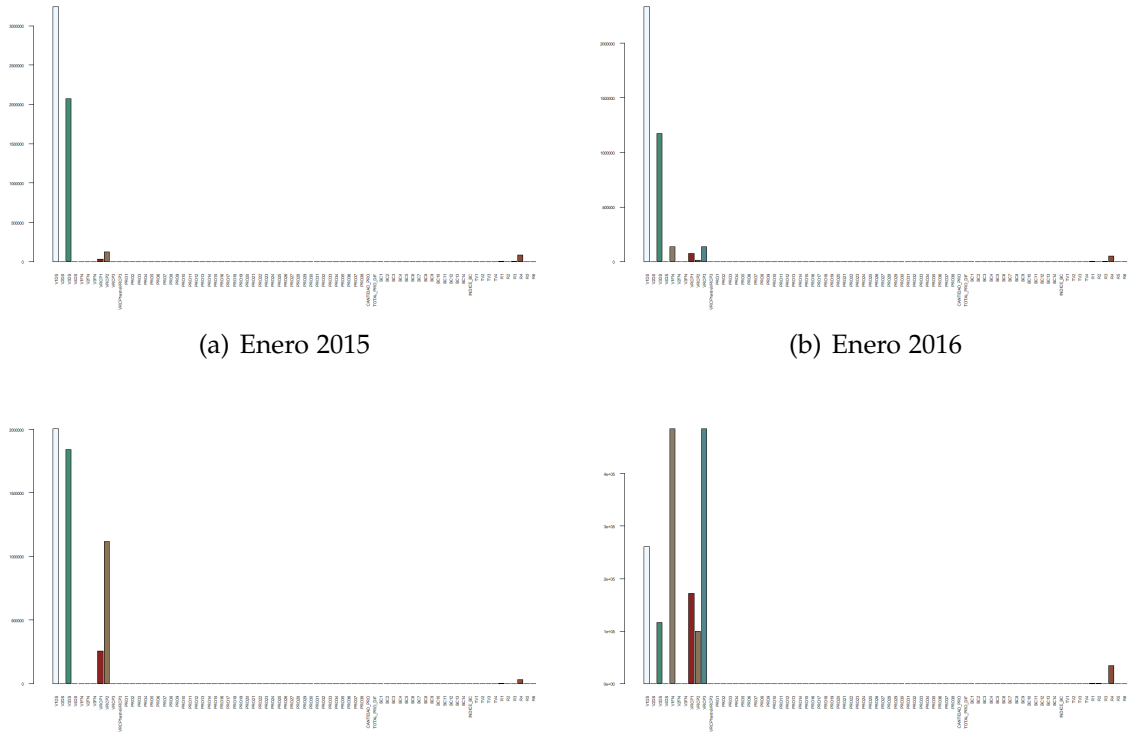
(c) Enero 2017

(d) Enero 2018

Figura 4.9: Importancia de las variables en la formación del Clúster 1 para la Clase VS2_CLI11

Siguiendo el mismo análisis que se hizo para el Clúster 1, tenemos que en la Figura 4.10 se muestran las variables que influyen en la creación del Clúster 2 para el mes de Enero en los años 2015 - 2018, para la Clase VS2_CLI11, para los clientes de Tipo #1. Se observa que los clientes que pertenecen al clúster 2 del año 2015 utilizan más las variables ***V1DS, V3DS, VRCP1, VRCP2, R1, R3, R4***, los del año 2016 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2, VRCP3, R1, R3, R4***, los del año 2017 utilizan más las variables ***V1DS, V3DS, VRCP1, VRCP3, R1, R4*** y los del año 2018 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2,***

VRCP3, R1, R2, R3, R4 para su actividad económica de ese mes.



(a) Enero 2015

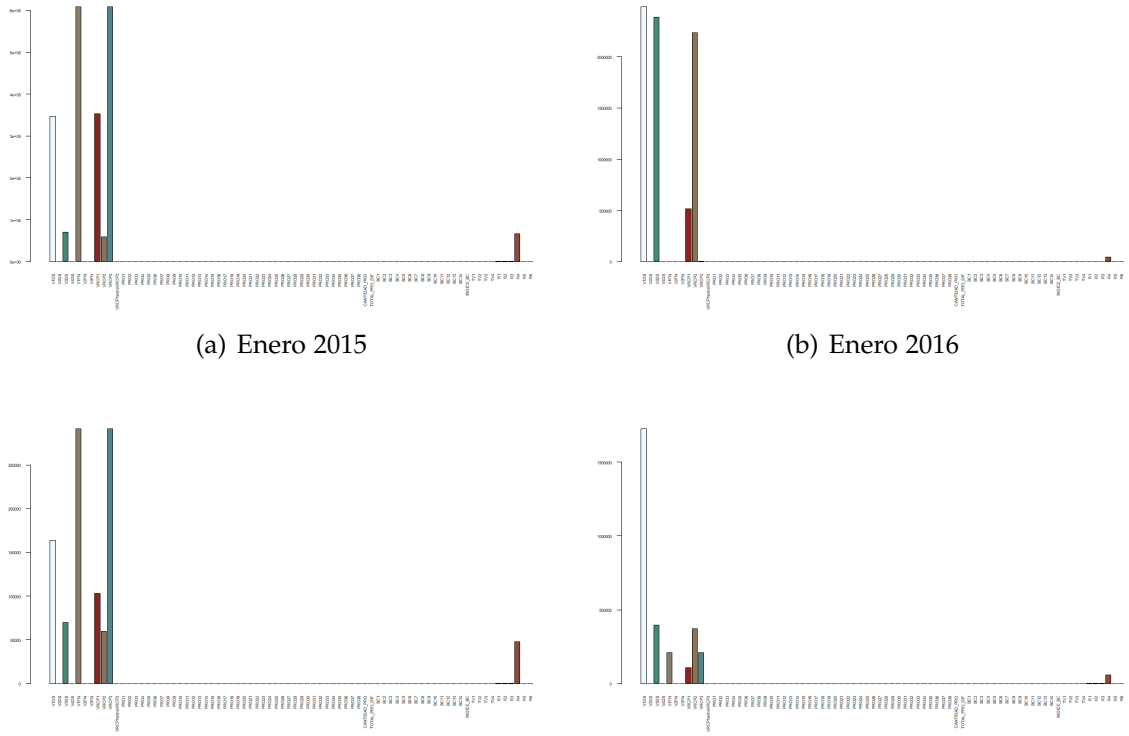
(b) Enero 2016

(c) Enero 2017

(d) Enero 2018

Figura 4.10: Importancia de las variables en la formación del Clúster 2 para la Clase VS2_CLI11

Para finalizar, en la Figura 4.11 se muestran las variables que influyen en la creación del Clúster 3 y se observa que los clientes que pertenecen al clúster 3 del año 2015 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2, VRCP3, R1, R2, R3, R4***, los del año 2016 utilizan más las variables ***V1DS, V3DS, VRCP1, VRCP2, VRCP3, R4***, los del año 2017 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2, VRCP3, R1, R2, R3, R4*** y los del año 2018 utilizan más las variables ***V1DS, V3DS, V1FN, VRCP1, VRCP2, VRCP3, R1, R2, R3, R4*** para su actividad económica de ese mes.



(a) Enero 2015

(b) Enero 2016

(c) Enero 2017

(d) Enero 2018

Figura 4.11: Importancia de las variables en la formación del Clúster 3 para la Clase VS2_CLI11

Por lo que, es de recordar que con la clusterización se busca formar los grupos y determinar la influencia de las variables para cada clúster; determinar dicha influencia es necesaria para desarrollar la estrategia de negocio que se pretende. Es por ello que para cada uno de los clientes de la base de datos se guarda el clúster al cual pertenecen en cada mes, ya que esto ayudará para analizar cada clúster con el propósito de establecer los valores que alcanza cada una de esas variables.

Es así, que la base de datos queda de la siguiente manera:

IDCLI	V1DS	V2DS	V3DS	V4DS	V1FN	V2FN	V3FN	VRCP1	VRCP2	VRCP3	VRCP1entreVRCP3	PRO1	PRO2	PRO3	PRO4	PRO5	PRO6	PRO7
1049371	4493.38	0.00	0	2	0.0	0	0.00	0.0	0.0	0.0	0.0000000	2	0	0	0	0	0	0
1049282	0.00	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	0	0	0	0	0	0	0
1045158	4001.13	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	0	0	0	0	0	0	0
1042274	5225.81	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	1	0	0	0	0	0	0
1041509	0.47	0.00	0	2	0.0	0	0.00	0.0	0.0	0.0	0.0000000	2	0	0	0	0	0	0
1041500	939.94	0.00	0	2	0.0	0	0.00	0.0	0.0	0.0	0.0000000	2	0	0	0	0	0	0
1040969	0.00	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	0	0	0	0	0	0	0
1040819	0.00	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	1	0	0	0	0	0	0
137179	10692.72	0.20	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	0	0	0	0	0	0	0
56561	17836.38	0.00	0	2	0.0	0	0.00	0.0	0.0	0.0	0.0000000	2	0	0	0	0	0	0
49579	9978.83	0.00	0	1	0.0	0	0.00	0.0	0.0	0.0	0.0000000	1	0	0	0	0	0	0
44223	74819.16	0.50	0	2	0.0	0	0.00	0.0	0.0	0.0	0.0000000	0	2	0	0	0	0	0
1003	664697.10	1.18	350000	4	655756.9	0	5.50	407382.1	298387.1	655756.9	0.6212395	1	1	1	0	1	0	0

(a)

PRO8	PRO9	PRO10	PRO11	PRO12	PRO13	PRO14	PRO15	PRO16	PRO17	PRO18	PRO19	PRO20	PRO21	PRO22	PRO23	PRO24	PRO25	PRO26
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	4	1	2	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

(b)

PRO27	PRO28	PRO29	PRO30	PRO31	PRO32	PRO33	PRO34	PRO35	PRO36	PRO37	PRO38	CANTIDAD_PRO	TOTAL_PRO_DIF	BC1	BC2	BC3	BC4
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	11	6	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	6	6	0	0	0	0

(c)

BC5	BC6	BC7	BC8	BC9	BC10	BC11	BC12	BC13	BC14	INDICE_BC	TV1	TV2	TV3	TV4	R1	R2	R3	R4	R5	R6	Grupos
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.45	2.06	2.39	190.95	0.279654400	0.150196400	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.56	2.88	0.68	267.44	0.159736800	0.030511520	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-27.62	0.76	-28.38	71.05	-4.664884000	-4.793244000	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-7.91	0.89	-8.80	82.42	-1.151662000	-1.281242000	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3.15	0.09	-3.24	8.10	-4.666667000	-4.800000000	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4.69	0.16	-4.85	15.34	-3.668840000	-3.794003000	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.000000000	0.000000000	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.72	8.54	-3.82	791.43	0.071566660	-0.057920470	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.78	5.85	3.93	542.74	0.216236100	0.086892440	1
0	0	0	0	0	0	0	0	0	0	0	155	9	17	1	57.07	22.65	34.42	2101.63	0.325861400	0.196533200	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1105.55	664.02	441.53	61608.66	0.215336600	0.086000250	3

(d)

Figura 4.12: Base de Datos de Clientes Tipo #1 para el mes de Enero del 2015 para la Clase VS2_CLI11

Ahora, se desea comparar la formación de los clúster con la segmentación previa que ha realizado la institución clasificando a sus clientes en ***4 Cuadrantes*** buscando siempre la influencia de las variables en la formación de los clúster. Por lo que se

selecciona el mes de Enero del año 2018, para la base de datos de Tipo #1 de la Clase VS2_CLI11. Obteniendo los siguientes resultados.

En la Figura 4.13 se tiene que de los 1118 Clientes 1109 pertenecen al Clúster 1, 1 pertenece al Clúster 2, 3 pertenecen al Clúster 3 y 5 pertenecen al Clúster 4, de los cuales 4 pertenecen al Cuadrante 1, 308 pertenecen al Cuadrante 2, 4 pertenecen al Cuadrante 3 y 802 pertenecen al Cuadrante 4. En este mes vemos que 2 Clientes pertenece al Cuadrante y Clúster 1, 1 Cliente pertenece al Cuadrante y Clúster 2, 2 Clientes pertenece al Cuadrante y Clúster 3 y 2 Clientes pertenece al Cuadrante y Clúster 4, a partir de esto se analizan los valores de las variables que determinan esta paridad, así como las otras agrupaciones que resultan.

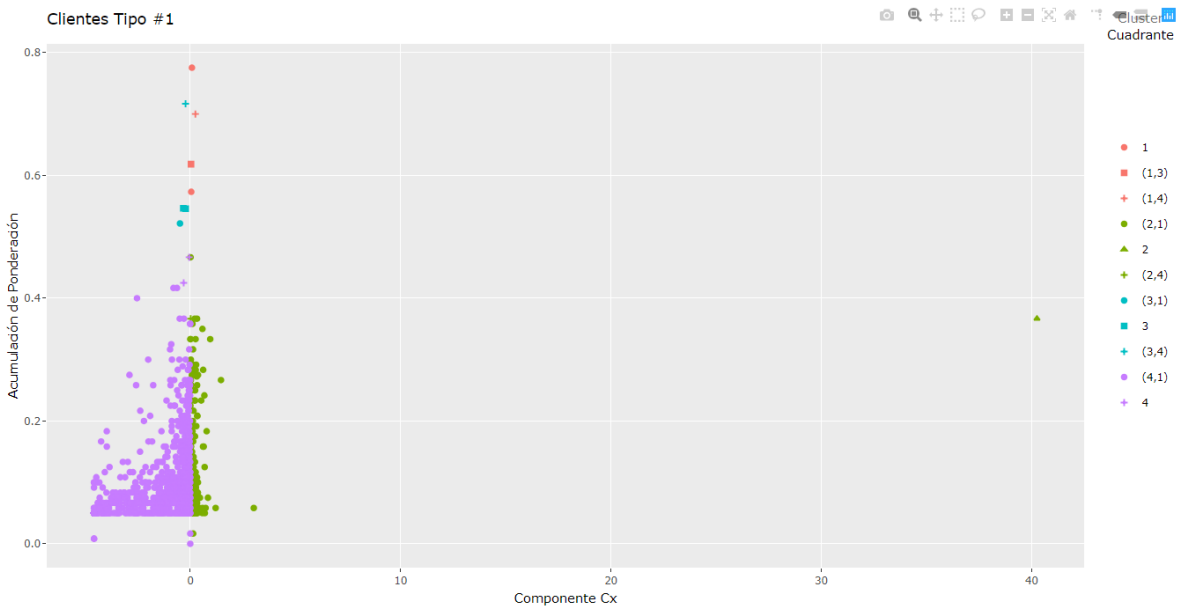


Figura 4.13: Distribución gráfica de los clientes del mes de Enero 2018 y Clase VS2_CLI11, comparando Clúster y Cuadrante

Por lo tanto, en la Figura 4.14 se muestran las variables que influyen en la formación del Clúster 1, de las cuales destacan las variables: ***V1DS, V3DS, V1FN, VRCP1, VRCP3, BC1, BC3, BC4, BC8, BC9, BC10, BC14, R4***, de estas variables cuantitativas podemos realizar un análisis lineal.

Realizando los cálculos tomando en cuenta únicamente los clientes que pertenecen al Clúster 1 se obtienen los resultados mostrados en la Tabla 4.7. En la cuál se detallan los valores mínimos y máximos que toman las variables.

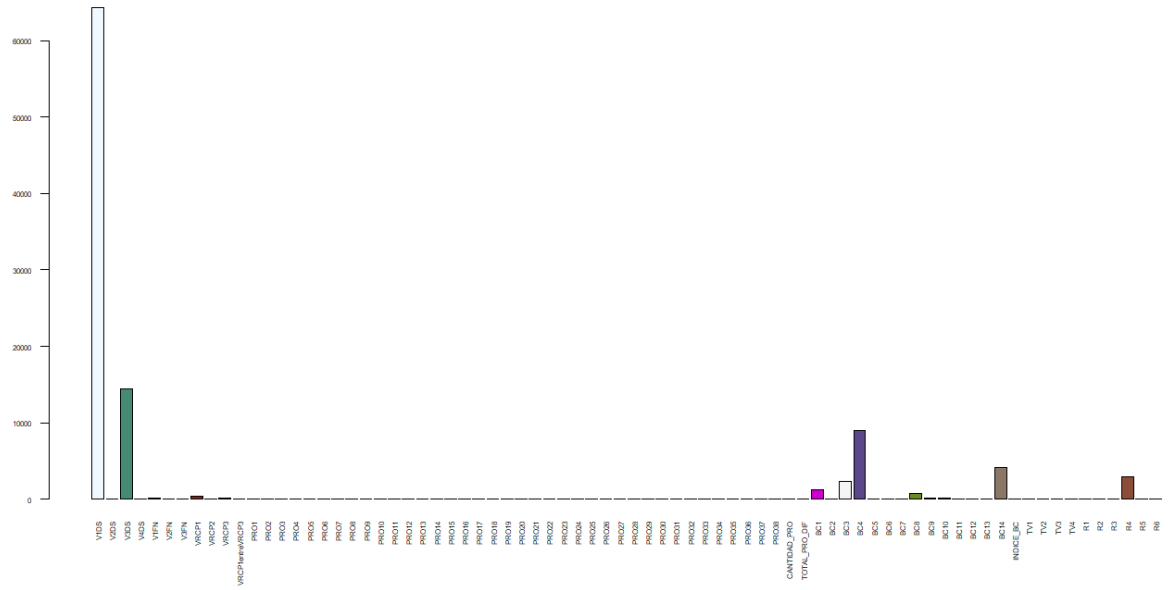


Figura 4.14: Importancia de las Variables en la Formación del Clúster 1 Enero 2018

V1DS	V3DS	V1FN	VRCP1	VRCP3
Min. : 0	Min. : 0	Min. : 0.0	Min. : 0.0	Min. : 0.0
1st Qu.: 230	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 3915	Median : 0	Median : 0.0	Median : 0.0	Median : 0.0
Mean : 64297	Mean : 14427	Mean : 120.1	Mean : 423.8	Mean : 126.3
3rd Qu.: 18005	3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.: 0.0	3rd Qu.: 0.0
Max. :7155879	Max. :4459000	Max. :121520.8	Max. :431729.9	Max. :122424.7

BC1	BC3	BC4	BC8
Min. : 0	Min. : 0	Min. : 0	Min. : 0.0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0	1st Qu.: 0.0
Median : 0	Median : 0	Median : 0	Median : 0.0
Mean : 1198	Mean : 2275	Mean : 9026	Mean : 700.5
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 0.0
Max. :1193879	Max. :2501715	Max. :9999832	Max. :776851.9

BC9	BC10	BC14	R4
Min. : 0.0	Min. : 0	Min. : 0	Min. : 0.00
1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 0	1st Qu.: 58.58
Median : 0.0	Median : 0	Median : 0	Median : 282.10
Mean : 112.9	Mean : 123	Mean : 4194	Mean : 2968.97
3rd Qu.: 0.0	3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 942.37
Max. :116914.1	Max. :136442	Max. :1893072	Max. :310006.99

Tabla 4.7: Resultados estadísticos lineales de las variables de importancia para el Clúster 1 Enero 2018

Para el Clúster 2, en la Figura 4.15 se muestran las variables que influyen en la formación de las cuales destacan las variables: ***TV1, R1, R2, R3, R4, R5, R6***, de estas variables cuantitativas podemos realizar también un análisis lineal.

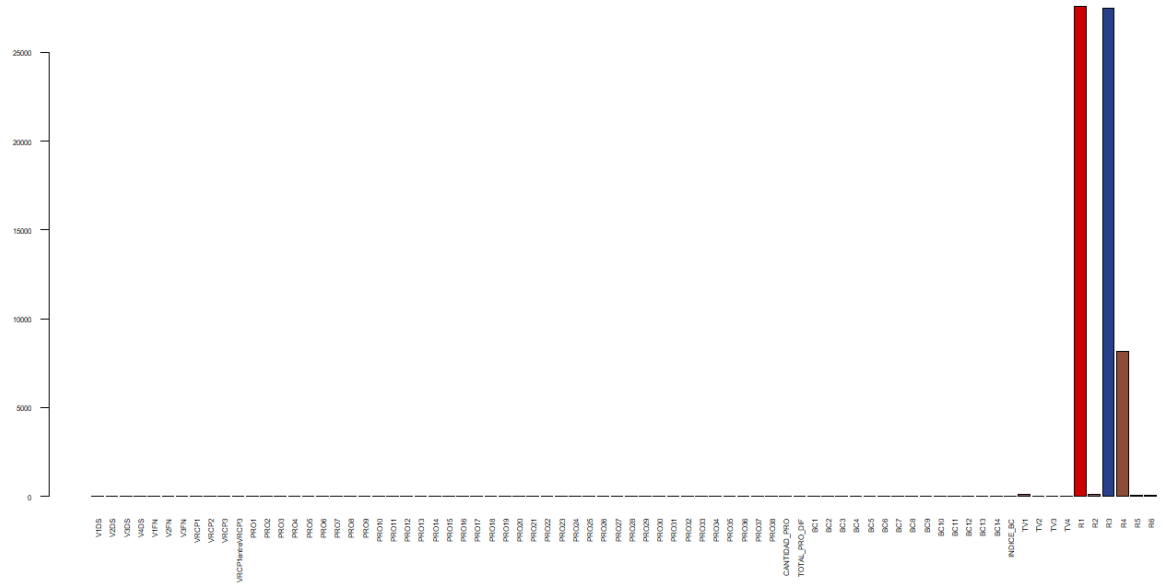


Figura 4.15: Importancia de las Variables en la Formación del Clúster 2 Enero 2018

Realizando los cálculos tomando en cuenta únicamente los clientes que pertenecen al Clúster 2, se detallan los valores que toman las variables en la Tabla 4.8

TV1	R1	R2	R3
Min. :98	Min. :27568	Min. :97.29	Min. :27471
1st Qu.:98	1st Qu.:27568	1st Qu.:97.29	1st Qu.:27471
Median :98	Median :27568	Median :97.29	Median :27471
Mean :98	Mean :27568	Mean :97.29	Mean :27471
3rd Qu.:98	3rd Qu.:27568	3rd Qu.:97.29	3rd Qu.:27471
Max. :98	Max. :27568	Max. :97.29	Max. :27471

R4	R5	R6
Min. :8188	Min. :40.4	Min. :40.26
1st Qu.:8188	1st Qu.:40.4	1st Qu.:40.26
Median :8188	Median :40.4	Median :40.26
Mean :8188	Mean :40.4	Mean :40.26
3rd Qu.:8188	3rd Qu.:40.4	3rd Qu.:40.26
Max. :8188	Max. :40.4	Max. :40.26

Tabla 4.8: Resultados estadísticos lineales de las variables de importancia para el Clúster 2 Enero 2018

En la Figura 4.16 se muestran las variables que influyen en la formación del

Clúster 3, las cuales son: ***V1DS, V1FN, VRCP1, VRCP3, BC1, BC2, BC14, R1, R3, R4***.

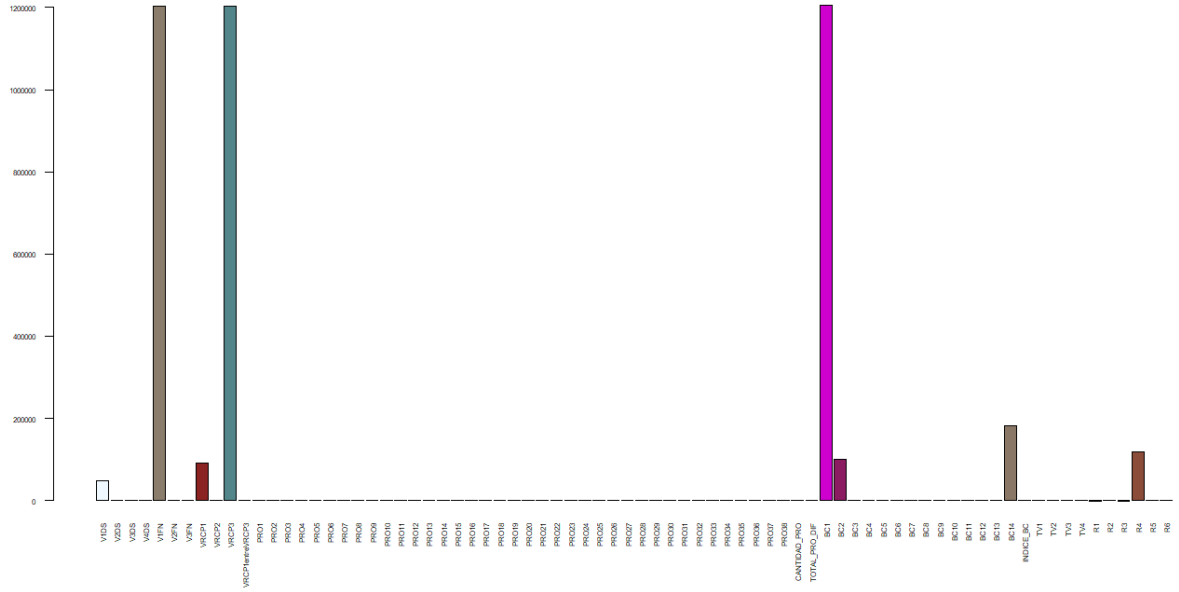


Figura 4.16: Importancia de las Variables en la Formación del Clúster 3 Enero 2018

A continuación en la Tabla 4.9 se detallan los valores que toman las variables que influyen en la formación del Clúster 3.

V1DS	V1FN	VRCP1	VRCP3	BC1
Min. : 0	Min. : 860263	Min. : 0	Min. : 860263	Min. : 862233
1st Qu.:33569	1st Qu.:1024341	1st Qu.: 40394	1st Qu.:1024341	1st Qu.:1028056
Median :67138	Median :1188419	Median : 80788	Median :1188419	Median :1193879
Mean :48918	Mean :1202058	Mean : 90668	Mean :1202058	Mean :1204951
3rd Qu.:73377	3rd Qu.:1372956	3rd Qu.:136001	3rd Qu.:1372956	3rd Qu.:1376310
Max. :79615	Max. :1557492	Max. :191215	Max. :1557492	Max. :1558741

BC2	BC14	R1	R3	R4
Min. : 0	Min. : 0	Min. :-3655.3	Min. :-6421.0	Min. : 49904
1st Qu.: 1081	1st Qu.: 0	1st Qu.:-1988.9	1st Qu.:-3668.2	1st Qu.: 60675
Median : 2163	Median : 0	Median : -322.5	Median : -915.5	Median : 71446
Mean : 99601	Mean :182536	Mean : -972.8	Mean :-2375.4	Mean :118039
3rd Qu.:149401	3rd Qu.:273804	3rd Qu.: 368.4	3rd Qu.: -352.6	3rd Qu.:152107
Max. :296639	Max. :547607	Max. : 1059.3	Max. : 210.4	Max. :232767

Tabla 4.9: Resultados estadísticos lineales de las variables de importancia para el Clúster 3 Enero 2018

Para finalizar, en la Figura 4.17 se muestran las variables que influyen en la formación del Clúster 4, las cuales son: ***V1DS, V3DS, VRCP1, VRCP2, BC1, BC10, R4***.

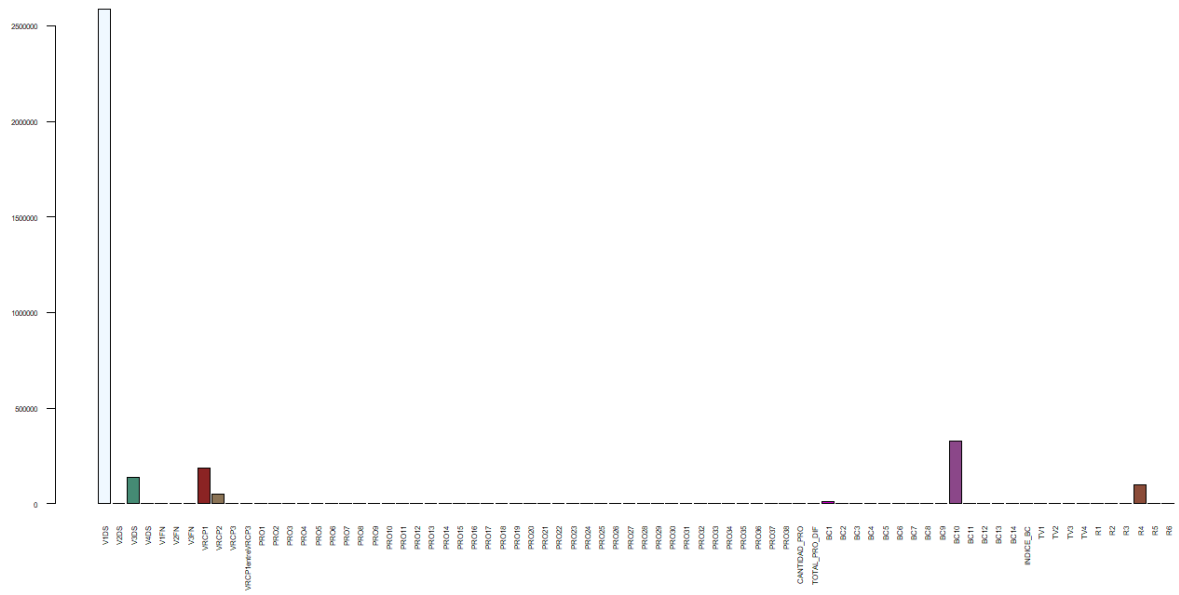


Figura 4.17: Importancia de las Variables en la Formación del Clúster 4 Enero 2018

Y en la Tabla 4.10 se muestran los valores que toman las variables más importancia para el Clúster 4.

V1DS	V3DS	VRCP1	VRCP2
Min. : 238105	Min. : 0	Min. : 0	Min. : 0
1st Qu.: 721724	1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 763839	Median : 50000	Median : 0	Median : 0
Mean : 2585653	Mean :140000	Mean :188292	Mean : 50000
3rd Qu.: 1075455	3rd Qu.:250000	3rd Qu.: 0	3rd Qu.: 0
Max. :10129140	Max. :400000	Max. :941457	Max. :250000

BC1	BC10	R4
Min. : 0	Min. : 0	Min. : 14079
1st Qu.: 0	1st Qu.: 0	1st Qu.: 20137
Median : 0	Median : 0	Median : 25996
Mean : 9448	Mean : 328166	Mean : 99799
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 43349
Max. :47238	Max. :1640829	Max. :395433

Tabla 4.10: Resultados estadísticos lineales de las variables de importancia para el Clúster 4 Enero 2018

Entonces tomando en cuenta los resultados anteriores, tenemos que la variable que influye en la formación de todos los clúster es la variable ***R4***, para poder determinar su importancia comparamos en la Tabla 4.11 los valores obtenidos como el mínimo, primer cuartil, mediana, media aritmética, tercer cuartil y el máximo.

R4				
MES	Clúster 1	Clúster 2	Clúster 3	Clúster 4
Min. : 0.0	Min. : 0.00	Min. :8188	Min. : 49904	Min. : 14079
1st Qu.: 61.2	1st Qu.: 58.58	1st Qu.:8188	1st Qu.: 60675	1st Qu.: 20137
Median : 289.0	Median : 282.10	Median :8188	Median : 71446	Median : 25996
Mean : 3715.5	Mean : 2968.97	Mean :8188	Mean :118039	Mean : 99799
3rd Qu.: 963.6	3rd Qu.: 942.37	3rd Qu.:8188	3rd Qu.:152107	3rd Qu.: 43349
Max. :395432.6	Max. :310006.99	Max. :8188	Max. :232767	Max. :395433

Tabla 4.11: Comparación de los valores que toma la variable R4 en Enero 2018 para los clúster 1, clúster 2, clúster 3 y clúster 4

De la tabla anterior podemos concluir que, la variable ***R4*** en el Clúster 1 posee valores más cercanos a los generales del mes gracias a que la mayoría de los clientes pertenecen a dicho clúster, en el Clúster 2 tenemos que existe solo un cliente y la variable ***R4*** es una de las variables que posee valores más altos para este único cliente del clúster lo cuál también nos dice que este cliente es atípico con el resto de los clientes.

En este sentido, se puede seguir analizando de la misma manera las otras variables, junto con las que no tienen influencia en la formación de estos clúster para determinar que tan importante es que los clientes posean esos servicios, y dependiendo de la importancia encontrada dejarlos de la misma manera o mejorarlos.

Por lo tanto, ahora nos interesa comparar las variables que podrían influir en la formación de ***16 clúster*** y verificar la relación que existe con los ***16 Grupos*** predefinidos por la Institución. En la Figura 4.18 se tiene que de los 1118 Clientes; 0 pertenecen al Grupo 1, 0 pertenecen al Grupo 2, 1 pertenece al Grupo 3, 3 pertenecen al Grupo 4, 0 pertenecen al Grupo 5, 33 pertenecen al Grupo 6, 0 pertenecen al Grupo 7, 4 pertenecen al Grupo 8, 0 pertenecen al Grupo 9, 275 pertenecen al Grupo 10, 0 pertenecen al Grupo 11, 0 pertenecen al Grupo 12, 37 pertenecen al Grupo 13, 765 pertenecen al Grupo 14, 0 pertenecen al Grupo 15 y 0 pertenecen al Grupo 16.

Ahora, si vemos la distribución de los 1118 Clientes por clúster tenemos que: 919 pertenecen al Clúster 1, 174 pertenecen al Clúster 2, 1 pertenece al Clúster 3, 2 pertenecen al Clúster 4, 1 pertenece al Clúster 5, 1 pertenece al Clúster 6, 3 pertenecen al Clúster 7, 7 pertenecen al Clúster 8, 2 pertenecen al Clúster 9, 1 pertenece al Clúster 10, 1 pertenece al Clúster 11, 1 pertenece al Clúster 12, 1 pertenece al Clúster 13, 2 pertenecen al Clúster 14, 1 pertenece al Clúster 15 y 1 pertenece al Clúster 16. Por lo tanto, para hacer un análisis primero debemos tener claro que el Clúster 1, no es sinónimo del Grupo 1, ya que según la Figura 4.19 podemos decir que el Grupo 14 contiene los clientes del Clúster 1, así que podríamos relacionarlos.

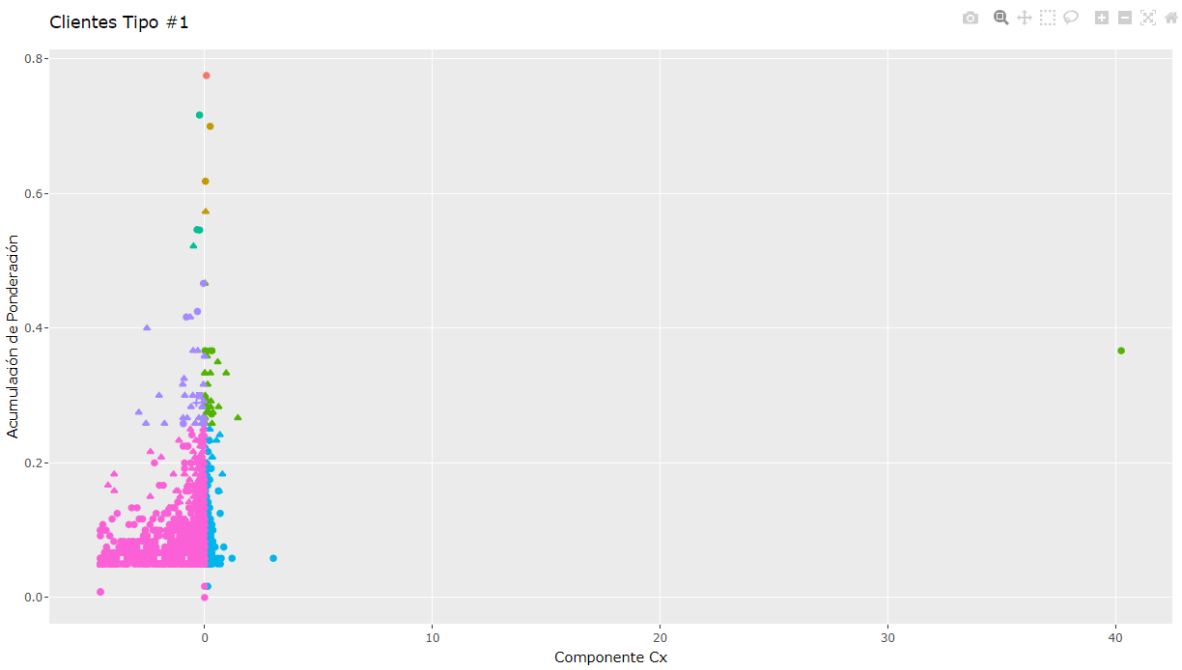
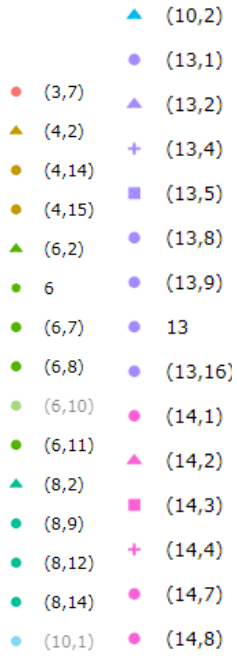


Figura 4.18: Distribución gráfica de los clientes del mes de Enero 2018 y Clase VS2_CLI11, comparando Grupo y Clúster



(a) (b)

Figura 4.19: Distribución gráfica de los clientes comparando (Grupo, Clúster) para el mes de Enero 2018

4.3. Enfoque Predictivo

Una vez realizado el análisis exploratorio el paso siguiente es la predicción mediante los métodos de aprendizaje supervisado como lo son árboles de decisión y bosques aleatorios. Los datos con los que se trabaja en este enfoque son todos los Clientes de Tipo #1 en el mes de Enero para el año 2018 sin hacer distinción entre las Clases y se agregan las variables que indica el Cuadrante y Grupo al que pertenece cada cliente.

Por lo que, se aplica el criterio de la institución en cuanto a los cálculos que realizan con las variables que ellos consideran que son las de importancia para clasificar a sus clientes. Dicha clasificación se guardan en la variable ***Cuadrante*** en donde se tiene que **Cuadrante 1** son los *Clientes A*, **Cuadrante 2** son los *Clientes M*, **Cuadrante 3** son los *Clientes B* y **Cuadrante 4** son los *Clientes I*.

4.3.1. Métodos Predictivos

Lo que se busca es aplicar la herramienta ***Bosques Aleatorios*** con la finalidad de determinar cuales son las variables que más influyen en la creación del modelo predictivo, el cuál depende de que es lo que se busca: si minimizar el error de predicción o maximizar la representación de los datos en el modelo. Luego de determinar cuales son dichas variables, se le aplicara ***Árboles de Decisión*** para generar el modelo y representar gráficamente esos caminos. Esto se hará con respecto a cada mes de los años de estudio. Realizando lo anterior se procederá a calcular cual es la probabilidad de que un Cliente cambie de Cuadrante.

Por lo tanto se procede a utilizar la herramienta de Árboles de Decisión para generar las rutas que clasifiquen a los Clientes en uno de los Cuadrantes y se obtiene como resultado la Figura 4.20. En la cual tenemos que las variables que intervienen en el modelo son ***R6, INDICE_BC, V1DS y TV2***.

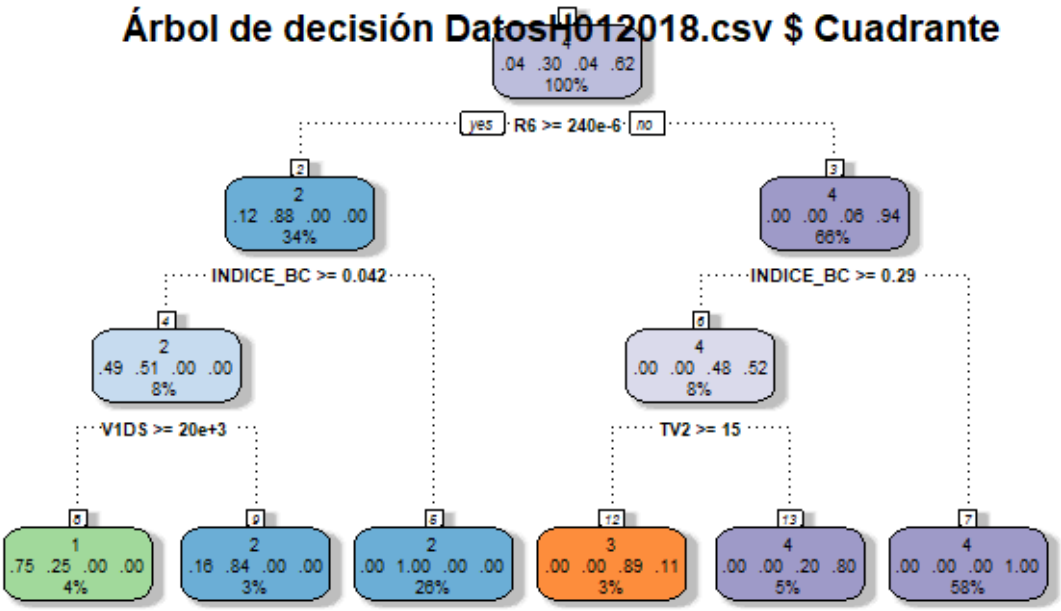


Figura 4.20: Diagrama del Modelo Predictivo Árboles de decisión para Enero 2018

Lo que el modelo nos dice es lo siguiente:

1. Si el valor de la Variable **R6 >= 240**e^{-6} entonces el Cliente tiene 4 % de pertenecer al Cuadrante 1 y 30 % de pertenecer al Cuadrante 2.

    1.1. Si el valor de la Variable **INDICE_BC >= 0.042** entonces el Cliente tiene 12 % de pertenecer al Cuadrante 1.

        1.1.1. Si el valor de la Variable **V1DS >= 20**e^{3} entonces el Cliente tiene 49 % de pertenecer al Cuadrante 1.

        1.1.2. Si el valor de la Variable **V1DS <20**e^{3} entonces el Cliente tiene 51 % de pertenecer al Cuadrante 2.

    1.2. Si el valor de la Variable **INDICE_BC <0.042** entonces el Cliente tiene 88 % de pertenecer al Cuadrante 2.

2. Si el valor de la Variable **R6 <240**e^{-6} entonces el Cliente tiene 4 % de pertenecer al Cuadrante 3 y 62 % de pertenecer al Cuadrante 4.

    2.1. Si el valor de la Variable **INDICE_BC >= 0.29** entonces el Cliente tiene 6 % de pertenecer al Cuadrante 3.

        2.1.1. Si el valor de la Variable **TV2 >= 15** entonces el Cliente tiene 48 % de pertenecer al Cuadrante 3.

        2.1.2. Si el valor de la Variable **TV2 <15** entonces el Cliente tiene 52 % de pertenecer al Cuadrante 4.

2.2. Si el valor de la Variable **INDICE_BC <0.29** entonces el Cliente tiene 94 % de pertenecer al Cuadrante 4.

Construyendo la matriz de Error para evaluar el modelo Árboles de decisión tenemos que:

		Predicción				Error
		1	2	3	4	
Actual	1	119	24	0	0	16.8
	2	40	1051	0	0	3.7
	3	0	0	101	40	28.4
	4	0	0	12	2217	0.5

Tabla 4.12: Matriz de Error del modelo Árboles de decisión Enero 2018 para Cuadrantes

En la Tabla 4.12 se muestra los resultados de aplicar Árboles de decisión a los 3604 Clientes del mes de Enero del año 2018 con 79 variables, además se muestra la matriz de confusión con un error total aproximado del 3.2 % y los errores dependiendo de cada clase. Es decir, de 143 Clientes que pertenecen al Cuadrante 1 el modelo predijo bien 119 y se equivoco en 24 que los clasifico en el Cuadrante 2, de 1091 Clientes que pertenecen al Cuadrante 2 el modelo predijo bien 1051 y se equivoco en 40 que los clasifico en el Cuadrante 1, de 141 Clientes que pertenecen al Cuadrante 3 el modelo predijo bien 101 y se equivoco en 40 que los clasifico en el Cuadrante 4, de 2229 Clientes que pertenecen al Cuadrante 4 el modelo predijo bien 2217 y se equivoco en 12 que los clasifico en el Cuadrante 3 y dependiendo de lo antes mencionado se calcula el error de cada clase.

Además, En la Tabla 4.13 se muestra los resultados de aplicar Bosques Aleatorios a los 3604 Clientes del mes de Enero del año 2018 con 79 variables, además se muestra la matriz de confusión con un error total aproximado del 2 % y los errores dependiendo de cada clase. Es decir, de 143 Clientes que pertenecen al Cuadrante 1 el modelo predijo bien 121 y se equivoco en 22 que los clasifico en el Cuadrante 2, de 1091 Clientes que pertenecen al Cuadrante 2 el modelo predijo bien 1080 y se equivoco en 11 que los clasifico en el Cuadrante 1, de 141 Clientes que pertenecen al Cuadrante 3 el modelo predijo bien 111 y se equivoco en 30 que los clasifico en el Cuadrante 4, de 2229 Clientes que pertenecen al Cuadrante 4 el modelo predijo bien 2220 y se equivoco en 9 que los clasifico en el Cuadrante 3 y dependiendo de lo antes mencionado se calcula el error de cada clase.

		Predicción				Error
		1	2	3	4	
Actual	1	121	22	0	0	15.38
	2	11	1080	0	0	1.00
	3	0	0	111	30	21.27
	4	0	0	9	2220	0.40

Tabla 4.13: Matriz de Confusión del modelo Bosques Aleatorios Enero 2018 para Cuadrantes

Ahora se verifica la importancia de las variables lo cual queda reflejado en la Figura 4.21, donde las variables de *MeanDecreaseAccuracy* minimizan el error en el modelo y las variables de *MeanDecreaseGini* maximizan la representación de los clientes en el modelo.

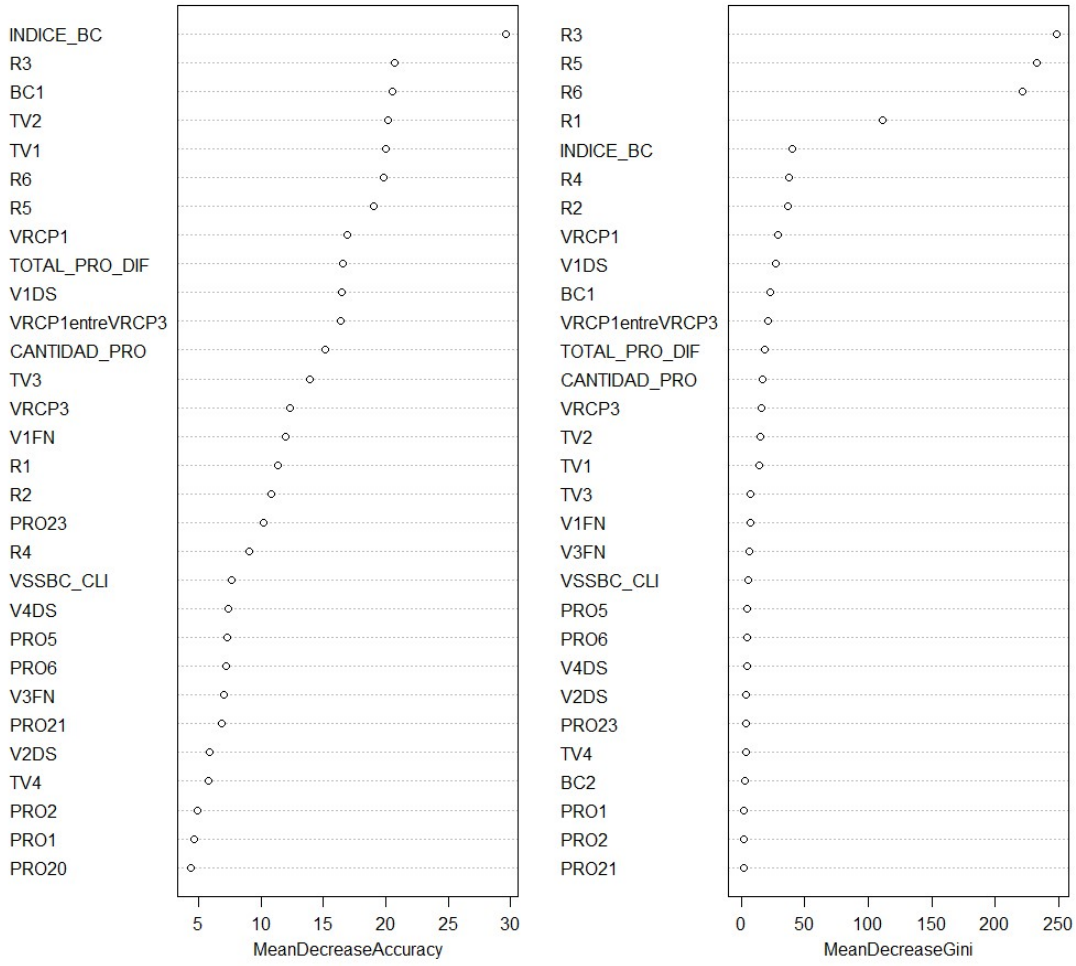


Figura 4.21: Importancia de las variables en Enero 2018 para mejorar el modelo predictivo por Cuadrantes

Por lo tanto aplicamos Árboles de Decisión solamente para las primera 10 variables de *MeanDecreaseAccuracy* para minimizar el error en el modelo y se obtiene

la Figura 4.22 que representa como queda la clasificación en el modelo.

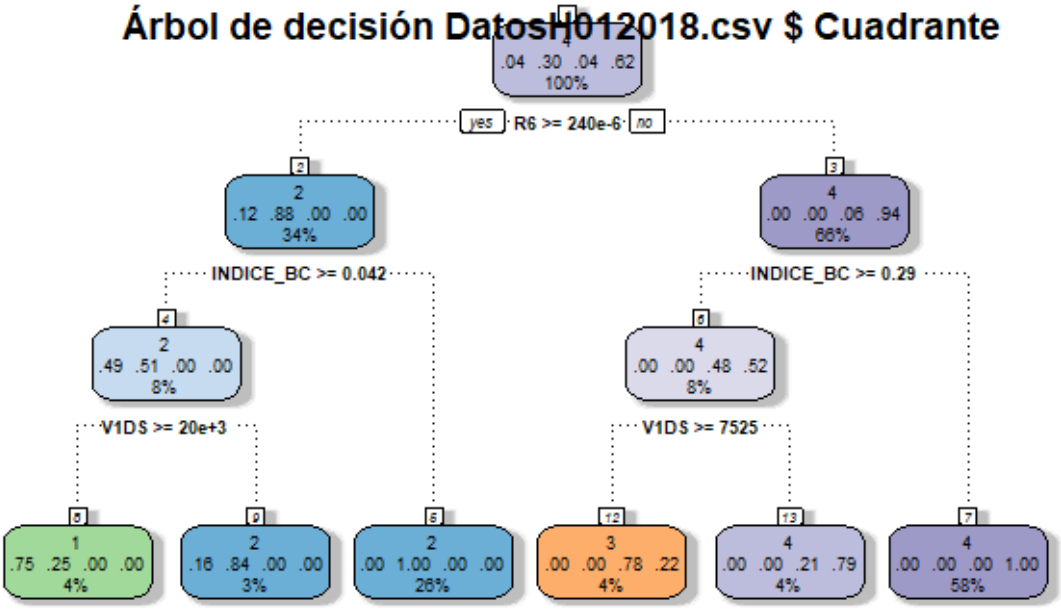


Figura 4.22: Diagrama del Modelo Predictivo por Cuadrantes para Enero 2018 tomando las 10 primeras variables de *MeanDecreaseAccuracy*

Lo que el modelo nos dice es lo siguiente:

1. Si el valor de la Variable **R6 >= 240**e^{-6} entonces el Cliente tiene 4 % de pertenecer al Cuadrante 1 y 30 % de pertenecer al Cuadrante 2.

    1.1. Si el valor de la Variable **INDICE_BC >= 0.042** entonces el Cliente tiene 12 % de pertenecer al Cuadrante 1.

        1.1.1. Si el valor de la Variable **V1DS >= 20**e^{3} entonces el Cliente tiene 49 % de pertenecer al Cuadrante 1.

        1.1.2. Si el valor de la Variable **V1DS <20**e^{3} entonces el Cliente tiene 51 % de pertenecer al Cuadrante 2.

    1.2. Si el valor de la Variable **INDICE_BC <0.042** entonces el Cliente tiene 88 % de pertenecer al Cuadrante 2.

2. Si el valor de la Variable **R6 <240**e^{-6} entonces el Cliente tiene 4 % de pertenecer al Cuadrante 3 y 62 % de pertenecer al Cuadrante 4.

    2.1. Si el valor de la Variable **INDICE_BC >= 0.29** entonces el Cliente tiene 6 % de pertenecer al Cuadrante 3.

2.1.1. Si el valor de la Variable **V1DS >= 7525** entonces el Cliente tiene 48 % de pertenecer al Cuadrante 3.

2.1.2. Si el valor de la Variable **V1DS <7525** entonces el Cliente tiene 52 % de pertenecer al Cuadrante 4.

2.2. Si el valor de la Variable **INDICE_BC <0.29** entonces el Cliente tiene 94 % de pertenecer al Cuadrante 4.

Para finalizar en la Tabla 4.14 se muestra la matriz de confusión con un error total aproximado del 1.94 % y los errores dependiendo de cada clase. Es decir, de 143 Clientes que pertenecen al Cuadrante 1 el modelo predijo bien 119 y se equivoco en 24 que los clasifico en el Cuadrante 2, de 1091 Clientes que pertenecen al Cuadrante 2 el modelo predijo bien 1078 y se equivoco en 12 que los clasifico en el Cuadrante 1 y 1 que lo clasifico en el Cuadrante 4, de 141 Clientes que pertenecen al Cuadrante 3 el modelo predijo bien 118 y se equivoco en 23 que los clasifico en el Cuadrante 4, de 2229 Clientes que pertenecen al Cuadrante 4 el modelo predijo bien 2019 y se equivoco en 10 que los clasifico en el Cuadrante 3 y dependiendo de lo antes mencionado se calcula el error de cada clase.

		Predicción				Error
		1	2	3	4	
Actual	1	119	24	0	0	16.78
	2	12	1078	0	1	1.19
	3	0	0	118	23	16.31
	4	0	0	10	2219	0.44

Tabla 4.14: Matriz de Confusión del modelo Arboles de decisión Enero 2018 por Cuadrantes utilizando las 10 primeras variables de *MeanDecreaseAccuracy*

Por lo tanto, a través de este camino analizamos los demás meses y determinamos cuales son las variables que intervienen en la predicción, y como va cambiando con respecto a cada mes.

Para finalizar se analiza el mismo mes, creando el modelo pero para predecir los Grupos. En la Figura 4.23 se muestra los resultados de aplicar Árboles de decisión a los 3604 Clientes del mes de Enero del año 2018.

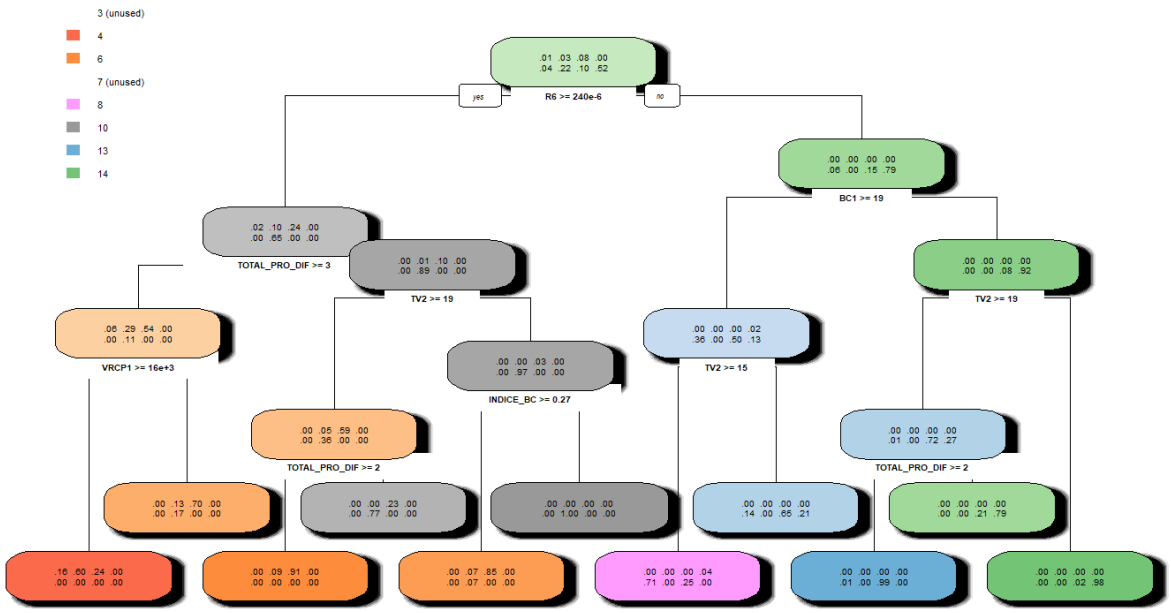


Figura 4.23: Gráfica del Modelo Predictivo utilizando Árboles de decisión para Enero 2018 por Grupos

En la Tabla 4.15 se muestra la matriz de confusión con un error total aproximado del 7.7 % y los errores dependiendo de cada clase. Es decir, de 22 Clientes que pertenecen al Grupo 3 el modelo predijo 0 y se equivoco en todos clasificando 22 en el Grupo 4, de 121 Clientes que pertenecen al Grupo 4 el modelo predijo bien 91 y se equivoco en 30 que los clasifico en el Grupo 6, de 294 Clientes que pertenecen al Grupo 6 el modelo predijo bien 244 y se equivoco en 16 que los clasifico en el Grupo 4 y 34 que los clasifico en el Grupo 10, de 7 Clientes que pertenecen al Grupo 7 el modelo predijo 0 y se equivoco en todos clasificando 6 en el Grupo 8 y 1 en el Grupo 13, de 134 Clientes que pertenecen al Grupo 8 el modelo predijo bien 102 y se equivoco en 32 que los clasifico en el Grupo 13, de 797 Clientes que pertenecen al Grupo 10 el modelo predijo bien 790 y se equivoco en 7 que los clasifico en el Grupo 6, de 352 Clientes que pertenecen al Grupo 13 el modelo predijo bien 270 y se equivoco en 36 que los clasifico en el Grupo 8 y 46 en el Grupo 14, de 1877 Clientes que pertenecen al Grupo 14 el modelo predijo bien 1830 y se equivoco en 47 que los clasifico en el Grupo 13; dependiendo de lo antes mencionado se calcula el error de cada clase.

		Predicción							Error
		3	4	6	7	8	10	13	14	
Actual	3	0	22	0	0	0	0	0	0	100
	4	0	91	30	0	0	0	0	0	24.8
	6	0	16	244	0	0	34	0	0	17
	7	0	0	0	0	6	0	1	0	100
	8	0	0	0	0	102	0	32	0	23.9
	10	0	0	7	0	0	790	0	0	0.9
	13	0	0	0	0	36	0	270	46	23.3
	14	0	0	0	0	0	0	47	1830	2.5

Tabla 4.15: Matriz de Confusión del modelo Árboles de decisión Enero 2018 por Grupos

Luego se aplica Bosques aleatorios y se muestran los resultados a través de la Tabla 4.16 donde se presenta la matriz de confusión con un error total aproximado del 4.3 % y los errores dependiendo de cada clase. Es decir, de 22 Clientes que pertenecen al Grupo 3 el modelo predijo bien 13 y se equivoco clasificando 9 en el Grupo 4, de 121 Clientes que pertenecen al Grupo 4 el modelo predijo bien 99 y se equivoco en 1 que los clasifico en el Grupo 3 y 21 que los clasifico en el Grupo 6, de 294 Clientes que pertenecen al Grupo 6 el modelo predijo bien 263 y se equivoco en 15 que los clasifico en el Grupo 4 y 16 que los clasifico en el Grupo 10, de 7 Clientes que pertenecen al Grupo 7 el modelo predijo 0 y se equivoco en todos clasificando los 7 en el Grupo 8, de 134 Clientes que pertenecen al Grupo 8 el modelo predijo bien 107 y se equivoco en 27 que los clasifico en el Grupo 13, de 797 Clientes que pertenecen al Grupo 10 el modelo predijo bien 782 y se equivoco en 15 que los clasifico en el Grupo 6, de 352 Clientes que pertenecen al Grupo 13 el modelo predijo bien 325 y se equivoco en 11 que los clasifico en el Grupo 8 y 16 en el Grupo 14, de 1877 Clientes que pertenecen al Grupo 14 el modelo predijo bien 1860 y se equivoco en 17 que los clasifico en el Grupo 13; dependiendo de lo antes mencionado se calcula el error de cada clase.

Actual	Predicción								Error
	3	4	6	7	8	10	13	14	
3	13	9	0	0	0	0	0	0	40.9
4	1	99	21	0	0	0	0	0	18.18
6	0	15	263	0	0	16	0	0	10.54
7	0	0	0	0	7	0	0	0	100
8	0	0	0	0	107	0	27	0	20.15
10	0	0	15	0	0	782	0	0	1.88
13	0	0	0	0	11	0	325	16	7.67
14	0	0	0	0	0	0	17	1860	0.91

Tabla 4.16: Matriz de Confusión del modelo Bosques Aleatorios Enero 2018 por Grupos

Ademas, aplicando Bosques aleatorios se verifica la importancia de las variables lo cual queda reflejado en la Figura 4.24, donde hay que recordar que las variables de *MeanDecreaseAccuracy* minimizan el error en el modelo y las variables de *MeanDecreaseGini* maximizan la representación de los clientes en el modelo.

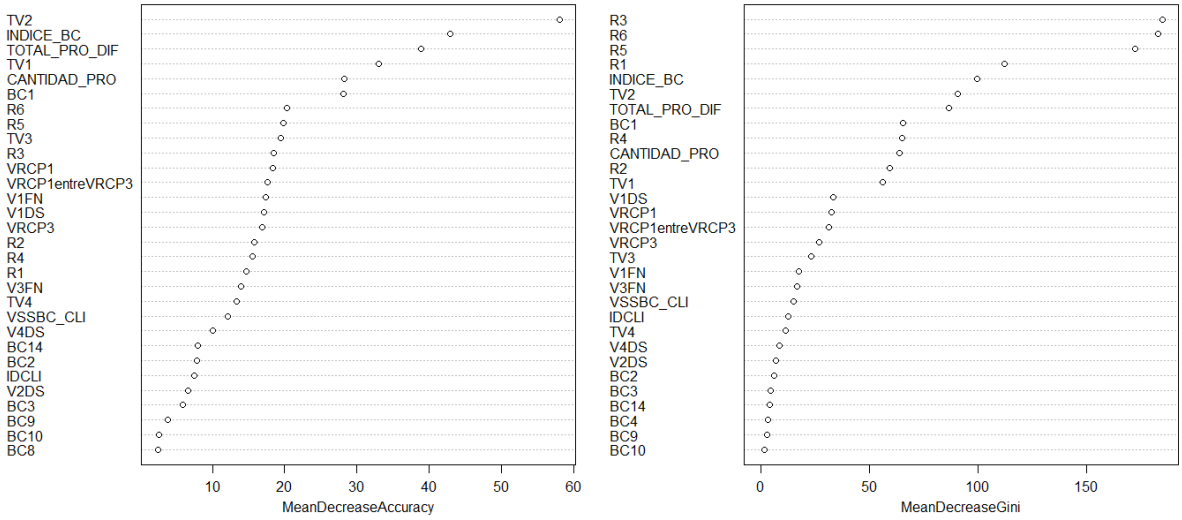


Figura 4.24: Importancia de las variables Enero 2018 para mejorar el modelo predictivo por Grupos

Capítulo 5

CONCLUSIONES

1. Aplicar un análisis en componentes principales no es suficiente como para visualizar los clúster gráficamente, ya que reduciendo la dimensión todavía quedan demasiadas componentes (ejes), lo cual nos indica que existen muchas variables independientes entre si.

2. El número de clúster óptimo es tres, lo cual nos indica que al formar un número mayor de clúster no estamos reduciendo la inercia intra-clase, pero nos garantiza que los clúster quedan muy bien definidos. Es decir, que los clientes que quedan en cada clúster son homogéneos entre si por lo que las variables que se determinen que influyen en la clusterización estarán representando a todos los clientes del clúster.

3. El alto de la barra de las variables que influyen en la formación de cada clúster no indica cual variable influye más, sino que indica el promedio del valor que toma esa variable para los clientes que conforman cada clúster representado.

4. Las variables que más influyen en la formación de los cuatro clúster de la base de datos de los clientes de Tipo #1 filtrados por mes y Clase VS2_CLI11 son: ***V1DS, V3DS, V1FN, V2FN, V3FN, VRCP1, VRCP2, VRCP3, BC1, BC2, BC3, BC4, BC8, BC9, BC10, BC14, TV1, R1, R2, R3, R4, R5, R6***

5. Para el enfoque predictivo, lamentablemente se cuenta con una base de datos desbalanceada, ya que todos los cuadrantes se ven representados, pero no por la misma cantidad de clientes y si lo vemos por Grupos tenemos que existen 5 Grupos que no tienen representación de clientes y los que si la tienen son de forma diferente en cuanto a número de clientes para cada Grupo.

6. En el modelo predictivo por Cuadrantes las variables que mas intervienen son: ***R6, VRCP1, INDICE_BC, TV2, V1DS***.

7. En el modelo predictivo por Grupos las variables que mas intervienen son: ***R6, INDICE_BC, TV2, TOTAL_PRO_DIF, TV1, VRCP1, BC1***.

Capítulo 6

RECOMENDACIONES

1. Es necesario conocer el significado de cada variables para determinar la influencia y la importancia que tiene para la actividad económica del cliente y de esta manera desarrollar conclusiones más profundas de los resultados que ayuden a la elaboración de los modelos para el desarrollo de la estrategia establecida.

2. Se debe modificar los métodos que se tienen para capturar la información de los clientes a manera que ayuden a formar bases de datos más solidas.

3. Si se quiere tener bases de datos más completas y balanceadas para el estudio es necesario aplicar inferencia estadística a las variables que tienen datos vacíos, de tal manera que se represente de manera correcta la actividad económica del cliente.

4. Se debe analizar aquellos clientes que se podrían catalogar como atípicos dentro de cada clase, ya que estos datos influyen y afectan en gran manera en el análisis en componentes principales y en la formación de los clúster.

5. Se debe estudiar las variables y reglas que se tiene para asignar a cada cliente en su Cuadrante o Grupos respectivo, buscando obtener una representación más balanceada en la distribución de los cliente para cada una de esas categorías.

REFERENCIAS BIBLIOGRÁFICAS

[1] Anderberg, G.M.R. (1973), *Cluster Analysis for Applications*, New York, Academic Press.

[2] Escudero, L. F. (1977), *Reconocimiento de patrones*, Paraninfo.

[3] Everitt, B.S. (1993), *Cluster Analysis*, Oxford University Press.

[4] Gordon A. D. (1981), *Classification*, Chapman and Hall.

[5] Hartigan, J.A. (1975), *Clustering Algorithms*, New-York, Wiley.

[6] Mirkin, B. (1996), *Mathematical Classification and Clustering*, Kluwer Academic Publishers

[7] Spath, H. y Bull, U. (1980), *Cluster Analysis of Algorithms for Data Reduction and Classifications of Objects*, New York, Wiley.

[8] Seber, G.A.F. (1984), *Multivariate Observations*, New York, Wiley.

[9] Spath, H. (1985), *Cluster Dissection and Analysis*, Chichester: Ellis Horwood.