ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS

DEPARTAMENTO DE POSTGRADOS

PROYECTO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN ESTADÍSTICA APLICADA

TEMA:

"MÉTODOS DE APRENDIZAJE ESTADÍSTICO PARA ESTIMAR PRONÓSTICOS DE VENTAS DE PRODUCTOS QUE AFECTAN A LA CADENA DE SUMINISTRO DE UNA EMPRESA DE CONSUMO MASIVO EN LA CIUDAD DE GUAYAQUIL"

Presentado por:

FRANCISCO PAÚL MORALES LUNA

Guayaquil – Ecuador

2022

RESUMEN

En la actualidad las venta del sector alimentario en el Ecuador está en constante crecimiento, especialmente en la industria de comercio minorista, debido a esto, los pronósticos de ventas constituyen un factor de alto interés para mejorar la competitividad dentro de las industrias. Para modelizar el comportamiento de compras individuales de los clientes y sus características de grupos, se recurre a las técnicas multivariantes de clusterización usando el método K-means y la caracterización mediante el método CHAID, que ayudan a simular el comportamiento real de ventas, que es uno de los problemas principales en los modelos de pronósticos. A partir de estas técnicas se identificaron se determinaron 7 segmentos de clientes homogéneos con características de compras similares que ayudaron a la identificación de variables relevantes para la generación de pronósticos de ventas más confiables y precisos a nivel de categorías por número de ordenes procesadas. Mediante la utilización de aprendizaje estadístico, utilizamos los modelos de pronósticos: regresión lineal, K-vecinos más cercanos (KNN) y Árboles de regresión. Los resultados de los pronósticos mostraron que el mejor modelo para ajustar los datos para las categorías de los distintos segmentos de clientes es el modelo de Regresión lineal, presentando medidas de errores más bajas en cuanto a MAPE y RMSE con relación a las medidas presentadas por los modelos KNN y Árboles de regresión, obteniendo resultados sobresalientes especialmente en laa categorías Papel Higiénico del segmento de clientes C con un MAPE apenas del 2,77% y un MAPE de 3,14% en la categoría Toallas Húmedas.

Palabras claves: segmentación, caracterización, pronósticos, consumo masivo

ABSTRACT

Currently, sales in the food sector in Ecuador are in constant growth, especially in the retail industry, therefore, sales forecasting is a factor of high interest to improve competitiveness within the industries. To model the behavior of individual customer purchases and their group characteristics, multivariate clustering, is need to use the K-means method and also the characterization through the CHAID method, which both help tos simulate the real sales behavior, which is one of the main problems in forecasting models. Based on these techniques, 7 homogeneous customer segments with similar purchasing characteristics were identified, which helped to identify relevant variables for the generation of more reliable and accurate sales forecasts at the category level and by number of orders, processed. Through the use of statistical learning, we used the following forecasting models: linear regression, K-nearest neighbors (KNN) and regression trees. The results of the forecasts showed that the best model to adjust the data for the categories of the different customer segments is the Linear Regression model, presenting lower error measures in terms of MAPE and RMSE in relation to the measures presented by the KNN and Regression Trees models, obtaining outstanding results especially in the Toilet Paper categories of customer segment C with a MAPE of only 2.77% and a MAPE of 3.14% in the Wet Towels category.

Keywords: segmentation, characterization, forecasts, mass consumption

DEDICATORIA

Dedico este trabajo a Dios, por haberme traído sano al mundo.

A mis padres por haberme dado la vida y sobre todo darme el apoyo en cada etapa de ella., con sus consejos y ejemplos de vida.

A mis hermanos por su grandiosa motivación en el camino que elegí.

A todos mis amigos que conocí en el trayecto de esta grandiosa etapa que me ayudaron de una u otra forma, ya sea a través de su buen compañerismo y que estuvieron presto con su ayuda cuando más lo necesité.

Paúl Morales

AGRADECIMIETO

A Dios, pilar fundamental de mi vida, por siempre estar conmigo y haberme dado más de lo que soñé cuando era un niño.

A mi madre Rosita Luna, por su amor desinteresado y por ser mi mayor fuente de inspiración en mi vida.

A mis hermanos por su apoyo incondicional y por sus valiosos consejos en estos estudios de postgrado.

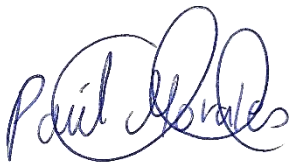
A dos grandes profesores Sergio Baúz y Johny Pambabay, por sus valiosos consejos y su valiosa colaboración en el desarrollo de este proyecto.

Paúl Morales

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría. El patrimonio intelectual del mismo corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.



FRANCISCO PAÚL MORALES LUNA

TRIBUNAL DE GRADUACIÓN

Firmado electrónicamente por:
OMAR HONORIO
RUIZ BARZOLA

Phd. Omar Ruiz

PRESIDENTE

Firmado electrónicamente por:
SERGIO ALEX
BAUZ OLVERA

Phd. Sergio Baúz

TUTOR

Firmado electrónicamente por:
FRANCISCO ANTONIO
MOREIRA VILLEGAS

Phd. Francisco Moreira

DOCENTE EVALUADOR

ABREVIATURAS O SIGLAS

CHAID	Detección Automática de Interacción Chi-cuadrado
CART	Árboles de Clasificación y Regresión
KNN	K-vecinos más Cercanos
MD	Minería de Datos
PAM	Particionamiento Alrededor de Medoids
WSS	Suma de Residuos Cuadrados Internos
WMAPE	Media del Error Absoluto en Porcentaje Ponderada
MAPE	Error Porcentual Absoluto Medio
WMPE	Media del Error en Porcentaje Ponderada
RMSE	Error Cuadrático Medio

TABLA DE CONTENIDO

Contenido

LISTADO DE FIGURAS

LISTADO DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Antecedentes

Las tecnologías de información debido a la alta competitividad empresarial de la mayoría de las industrias han tenido un desarrollo sumamente acelerado en los últimos años, lo que ha provocado que la cantidad de datos de ventas aumenten constantemente y ha generado una gran importancia de aprovechar estos datos que crecen continuamente. (Siwerz & Dahlén, 2017)

El crecimiento de estos datos generados por los clientes, durante la última década se ha convertido en un activo muy valioso para las industrias en la actualidad y a partir de allí se ha generado una oportunidad muy popular de poder obtener grandes beneficios del análisis de esa información. Por ejemplo: descubrir patrones que podrían usarse para guiar a una empresa sobre como tomar decisiones con respecto al marketing, organización y ventas. (Siwerz & Dahlén, 2017)

Un caso de uso muy relevante, usando estos recursos existentes y que se ha convertido en un enfoque muy importante que se utiliza para mejorar de forma eficiente las operaciones de cadena de suministro y comerciales es la predicción de ventas, que es un factor de alto interés que se usa para modelar el comportamiento de compra de consumidores individuales o para modelar la demanda de SKU´s en la industria de comercio minorista del sector alimentario. (Bajari et al., 2015b)

La predicción indistintamente sea la industria es muy importante y apreciada en todos los medios del mundo. Por lo tanto, en el caso de la industria de alimentos, realizar pronósticos de ventas que simulen el real comportamiento de la demanda es uno de los problemas principales de la cadena de suministro y la importancia radica en poder encontrar a través de un correcto enfoque matemático patrones de la demanda que ayuden a tratar de reducir el error de pronóstico actual que está basado en el hecho de que ofertar menos de lo demandado por el mercado, genera pérdidas potenciales de ventas; por lo contrario,

sobreestimar por encima de lo demandado ocasiona costos adicionales por exceso de producción, logística e inventarios. (Eguiguren Calisto, 2019)

En la actualidad el sector alimentario en el Ecuador presenta una alta competitividad existente, por lo que las empresas buscan mejorar sus procesos y disminuir considerablemente esos costos adicionales operativos con el fin de mantenerse competente dentro del mercado, lo que evidencia aún más la necesidad de tener una buena predicción de la demanda con un nivel de error lo más bajo posible. (Carreño et al., 2019)

Disponer de un pronóstico de la demanda acertado, es de suma importancia para una compañía en general pues está incluido en las distintas áreas, como: comercial, de pricing, abastecimiento, recursos humanos y en especial en el área de producción pues ayuda optimizar la utilización de recursos, minimizar los costos de inventarios, evitar los incumplimientos de la demanda, entre otros. (Carreño et al., 2019)

1.2. Descripción del problema

Históricamente el uso de técnicas de pronósticos de ventas dentro del sector alimentario se ha basado en modelos estadísticos tradicionales. Sin embargo, debido a la transformación digital que está experimentando actualmente este sector y que la tecnología ha creado nuevas oportunidades para comprender mejor los factores y patrones que influyen en las ventas diarias, el aprendizaje estadístico se ha convertido en un campo con una aplicación relativamente alta, debido a la flexibilidad que tienen sus técnicas para a partir de los datos aprender patrones que ayuden a predecir de una mejor manera los resultados de un evento futuro. (Siwerz & Dahlén, 2017)

Los estudios han evidenciado que los modelos de aprendizaje estadístico podrían ser provechosos en el sector alimentario minorista para comprender una determinada segmentación de clientes o correlación de productos demandados por un sector determinado del mercado. (Siwerz & Dahlén, 2017)

Tal es el caso práctico del comercio minorita del sector alimentario de Turquía que emplearon métodos de aprendizaje estadístico como redes neuronales artificiales (ANN) para estimar los ingresos por ventas y según sus resultados hay grandes semejanzas entre los datos

predichos y los reales. Pues los datos estimados presentan desviaciones a la alta o baja solo para el 10% por ciento de los datos reales. (Penpece & Elma, 2014)

En Ecuador el problema comercial al que se enfrentan cada día la mayoría empresas del sector alimentario es la falta de confiabilidad en los pronósticos de ventas de un determinado producto que les permita cubrir la demanda de sus clientes, como también realizar la toma de decisiones con respecto a la producción cualesquiera de un producto, con el fin de incrementar ingresos y disminuir costos de oportunidad como gastos de almacenamiento, debido a una sobreproducción. (Llumitasig Galarza, 2021)

Actualmente las empresas de consumo masivo nacional realizan predicciones de ventas de manera empírica (escasos procedimientos estadísticos), por lo que es necesario aplicar métodos de aprendizaje estadístico que se ajusten al comportamiento de los datos y que permitan estimar pronósticos confiables.

El análisis predictivo dentro del sector alimentario en el Ecuador está en constante desarrollo, y debido a la gran competencia empresarial que existe hoy en día, la mayoría de las empresas se encuentran en la búsqueda constante de aumentar sus ganancias y reducir sus costos; y uno de los mecanismo para lograr este objetivo es disponer pronósticos de ventas con un alto grado de precisión. Po lo que es de vital importancia dar a conocer los diferentes modelos de aprendizaje estadístico descritos en la literatura y su aporte significativo particularmente en este sector. (Burgaentzle Jarrín, 2016)

1.3. Objetivos

1.3.1. Objetivo general:

- Evaluar la predicción de ventas de productos de consumo masivo de una empresa perteneciente al sector alimentario de la ciudad de Guayaquil a través de métodos de aprendizaje estadístico; con el fin de generar pronósticos confiables que ayuden a reducir los gastos asociados a la cadena de suministro.

1.3.2. Objetivos específicos:

- Revisar la bibliografía científica aplicada a modelos de aprendizaje estadístico que coadyuve a la selección de la metodología apropiada.
- Obtener un conjunto de datos con variables relevantes para la aplicación de los modelos previamente analizados.
- Predecir las ventas de productos de consumo masivo de una empresa del sector alimentario con la ayuda de software especializado de aprendizaje estadístico para detectar posibles patrones de consumo.
- Validar las predicciones de productos de consumo masivo empleando validación cruzada para elegir el modelo más adecuado.

1.4. Alcance

El presente trabajo de investigación está enfocado en la búsqueda de métodos de aprendizaje estadístico que ayuden a generar pronósticos confiables de ventas de productos de consumo masivo para una empresa del sector alimentario con domicilio en la ciudad de Guayaquil. Para este finalidad se utilizarán datos oficiales de ventas de la empresa, los mismos que considerarán un horizonte de tiempo de los últimos 5 años (2017-2021).

CAPÍTULO 2

2. MARCO TEÓRICO

La empresa

La Empresa AAA es una empresa ecuatoriana cuya actividad económica consiste en la comercialización al por Mayor de Bienes No Duraderos Diversos (aceites, atún, azúcar jabones, detergentes y productos de cuidado personal). Trabaja actualmente con 45 proveedores, de los cuales provee 200 marcas con un portafolio de productos diversificados (62 categorías) a los clientes del canal tradicional más importantes dentro de la ciudad de Guayaquil. Fue fundada el 31 de mayo de 2000.

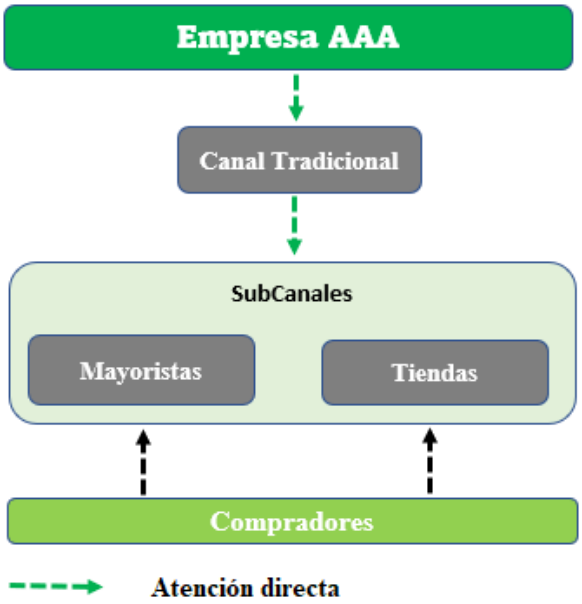
Algunos de sus principales competidores son Distribuidora Romero Reyes S.A, Asertia Comercial SA, dependiendo de la categoría de productos.

Canales de Distribución

En Guayaquil-Ecuador, Empresa AAA atiende exclusivamente el canal tradicional, específicamente mayoristas y tiendas. Estos subcanales se exponen en la Figura 2.1:

Figura 2. 1

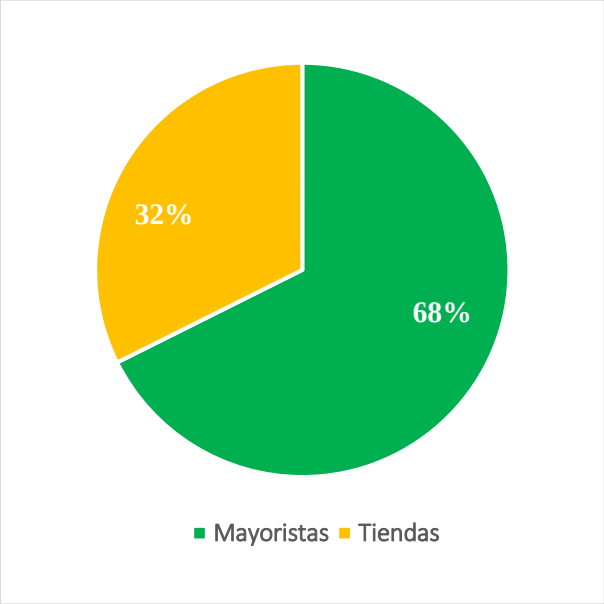
Subcanales de distribución de Empresa AAA en Guayaquil



Los ingresos provienen principalmente del canal Tradicional, siendo el subcanal mayorista el que mayor aporta con un 68% y las Tiendas con el 32%, del total de la ventas en 2021 como se muestra en la Figura 2.2:

Figura 2. 2

Distribución de ventas por Subcanal en 2021

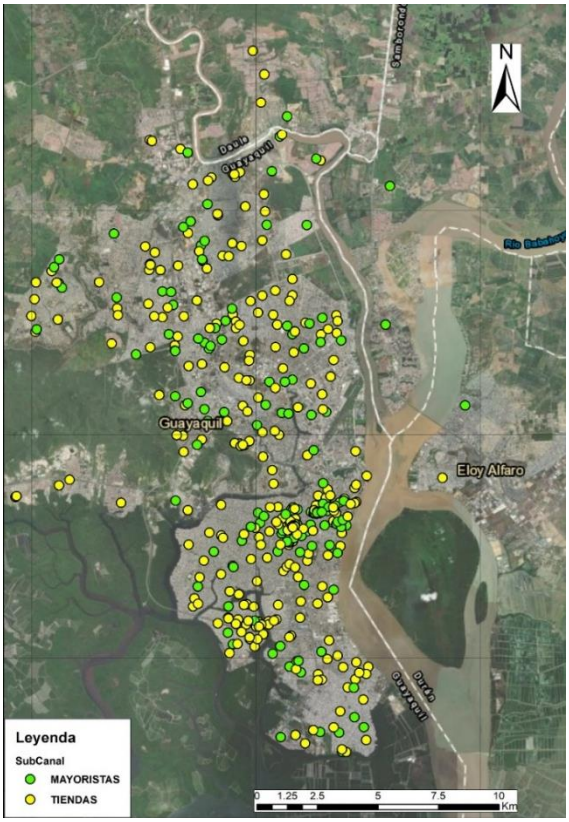


Fuente: Empresa AAA, Información de Ventas 2017-2021

Según los registros en la base de datos de la empresa, existen 1.218 clientes, abarcando el 71% el subcanal Tiendas (abarrotes, bazares, farmacias, minimarkets y otros) y el 29% al subcanal Mayoristas, los cuales se distribuyen en la ciudad de Guayaquil y se muestra en la Figura 2.3:

Figura 2. 3

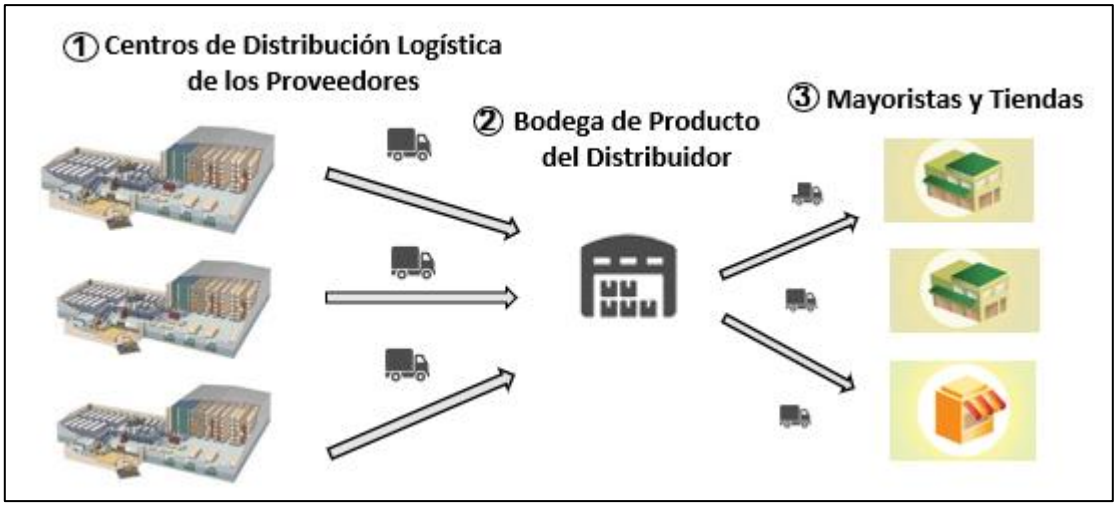
Mayoristas y Tiendas en Guayaquil



Los productos son enviados desde los centros de distribución de los proveedores a las Bodegas de Productos de la Empresa AAA, desde el cual se envía a los diferentes clientes mayoristas y tiendas a los cuales se debe abastecer dentro de la ciudad de Guayaquil, tal como lo evidencia la Figura 2.4:

Figura 2. 4

Diagrama de Flujo Proveedores-Distribuidor-Mayoristas y Tiendas



2.1. Minería de datos

Según Peacock (1998 citado por Rivero, 2012) define a la Minería de Datos (MD) como el "descubrimiento automático de patrones o modelos interesantes y no obvios ocultos en una base de datos, los cuales tienen un potencial para contribuir en los aspectos principales del negocio".

En un mundo de negocios altamente automatizado y competitivo la minería de datos ha tomado un rol relevante en la gestión de grandes volúmenes de datos, tal es el caso de un estudio realizado por (Femina & Sudheep, 2015), donde en el marco de trabajo para la minería de datos en la administración de la relación con el cliente (CRM), utiliza un modelo de clasificación para predecir el comportamiento de los clientes con el fin de mejorar los procesos de toma de decisiones para retener a clientes valiosos.

Así mismo, (Abadía et al., 2015) utilizó las técnica de minería de datos K para realizar la segmentación de clientes de una comercializadora de productos lácteos domiciliada en la ciudad de Popayán a través de la implementación del algoritmo K-means obteniendo como resultado 7 segmentos de clientes y a su vez permitiéndole a la empresa el uso de esta información para generar estrategias de marketing. En la misma línea (Cálad, 2015), aplicó técnicas de minería de datos para la segmentación de clientes de Tiendacol S.A una empresa del sector de modas ubicada en la ciudad de Medellín, a través de las técnicas: K-means y

árboles de decisión, obteniendo como resultado tres segmentos de clientes diferentes: los mejores, los intermedios y los peores. Los mejores que representan a los clientes más constantes con altos volúmenes de compras, con mayor frecuencia y altos créditos, los intermediarios que compran un número aceptable de veces al año y que no realizan muchos créditos y por último los peores, que son aquellos que compran bajos volúmenes y lo hacen de forma esporádica.

A continuación, se explican dos de las técnicas que se destacan por su alto uso como son: análisis de clusters y árboles de decisión:

2.1.1. Análisis de Clusters

El propósito del análisis de conglomerados(clusters) es agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Normalmente se agrupan las observaciones usando este análisis; pero también puede aplicarse para agrupar variables. (Peña, 2002)

Agrupamiento de particiones

Partición

Según Garey &Johnson (1969 citado por Delgado, 2016)), una partición de conjunto X es una colección de subconjuntos no vacíos $\{X_i: i \in I\}$ de X que son disjuntos dos a dos y cuya unión es X, donde llamamos a $\boldsymbol{X_i}$ como las partes de la partición y a $\boldsymbol{I}$ el conjunto de índice.

Según Kassambara (2015), el agrupamiento de particiones son métodos de agrupamientos que se utilizan para clasificar las observaciones, dentro de un conjunto de datos, en múltiples grupos en función de su similitud. Los algoritmos requieren que se especifique el número de clústers que se generarán.

Kassambara (2015), especifica dos métodos relevantes de agrupamiento de particiones, los cuales son: Agrupación K-means y Agrupación de K-medoids o PAM.

Agrupación de K-means

De acuerdo con MacQueen (1967 citado por Kassambara, 2015) es el algoritmo de aprendizaje en la que cada agrupación está representada por el centro o la media de los puntos que pertenecen a la agrupación.

Método clásico de partición: Algoritmo K-means

Fundamentos

Peña (2002), expresa que el objetivo del algoritmo K-means suponiendo una muestra n elementos con p variables es dividir esta muestra en un número de grupos predeterminados, G.

Este mismo autor explica que para su aplicabilidad este algoritmo requiere de cuatro etapas, las cuales son explicadas a continuación:

1. Seleccionar G 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 G 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 G 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.

Agrupación de K-medoids o PAM *(*Partitioning Around Medoids, Kaufman &Rousseeuw, 1990 citado por Kassambara, 2015), es el algo algoritmo de aprendizaje en el cada grupo está representado por uno de los objetos de los grupos (mediana). PAM es menos sensible a valores atípicos en comparación con k-medias.

Métodos para medir distancias

Según Kassambara (2015), la elección de las medidas de distancia es un paso crítico en la agrupación, ya que determina como se calcula la similitud de dos elementos (x, y) y a que su vez influye en la forma de los grupos.

Los métodos clásicos para medir distancias son las distancias Euclidiana y Manhattan, que expresan a continuación:

1. **Distancia Euclidiana**

$$d_{euc}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (x_i - x_j)^{\boldsymbol{2}}} \qquad (1)$$

2. **Distancia de Manhattan**

$$d_{man}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (x_i - x_j) \qquad (2)$$

Determinación del número óptimo de clusters

Determinar el número óptimo de conglomerados es un tema importante dentro del agrupamiento de particiones.

Según Kassambara (2015), se dividen en tres métodos: método del codo, método de silueta promedio y el estadístico de brecha, pero para el estudio solo se empleará el primero.

Método del codo

El método de codo analiza la suma de residuos cuadrados dentro del clusters (WSS) como una función de la cantidad de grupos y escoge como óptimo aquel valor a partir del cual agregar más clústers produce una mínima reducción de la varianza total del clústers.

Según Kassambara (2015), el número óptimo de clusters se puede definir de la siguiente manera:

1. Calcule el algoritmo de agrupamiento (agrupamiento de K-means o K-medoids) para diferentes valores de k.
2. Para cada k, calcule la suma total de cuadrados dentro del clusters (WSS)
3. Trazar la curva de WSS según el número de conglomerados k.

4. La ubicación de una curva (codo) en la parcela generalmente se considera como un indicador del número apropiado de conglomerados.

Una vez seleccionado el número de clusters adecuados y aplicado el algoritmo clustering es de suma importancia evaluar la calidad de los clusters. Una de las técnicas empleadas es:

Validación interna de los clusters

Según Kassambara (2015), se trata de un proceso no supervisado que solo utiliza información interna del proceso de agrupación de clusters para evaluar la bondad de las agrupaciones generadas.

Medidas internas para la validación de los clusters

Coeficiente de silueta

El análisis de silueta mide qué tan bien se agrupa una observación dentro del clusters generado y estima la distancia promedio entre los agrupamientos.

Para cada observación i , el coeficiente de silueta según Kassambara (2015), se obtiene la siguiente manera:

1. Para cada observación i, calcule la disimilitud promedio entre i y todos los demás puntos del grupo al que pertenece i.
2. Para todos los demás clústers C, a los que i no pertenece, calcule la disimilitud promedio $d(i, C)$ de i para todas las observaciones C. La más pequeña de estas $d(i, C)$ se define como $b_i = min_C d(i, C)$. El valor de b_i puede verse como la disimilitud entre i y su grupo "vecino", es decir el más cercano al que no pertenece.
3. Finalmente, el coeficiente de silueta de cada observación i viene determinado mediante la siguiente fórmula:

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \qquad (3)$$

El coeficiente de silueta se puede interpretar de la siguiente manera:

- Las observaciones con S_i grande (casi 1) están muy bien agrupadas.

- Un S_i pequeño (alrededor de 0) significa que la observación se encuentra entre dos grupos.
- Las observaciones con un S_i negativo probablemente estén agrupadas en el grupo incorrecto.

Índice de Dunn

El índice de Dunn es otra medida de validación de agrupamiento interno que se puede calcular de la siguiente manera:

1. Para cada grupo, calcule la distancia entre cada uno de los objetos del grupo y los objetos de los otros grupos.
2. Use el mínimo de esta distancia por pares como la separación entre grupos (separación mínima)
3. Para cada grupo, calcule la distancia entre los objetos en el mismo grupo.
4. Use la distancia máxima dentro-clusters (es decir, el diámetro máximo) como la compacidad dentro-clusters.
5. Calcular el índice de Dunn (D) de la siguiente manera:

$$D = \frac{min.separación}{max.diámetro}$$

El índice de Dunn se puede interpretar de la siguiente manera:

Si el conjunto de datos contiene grupos compactados y bien separados, se espera que el diámetro de los grupos sea pequeño y que la distancia entre los grupos sea grande. Por lo tanto, el índice de Dunn debe maximizarse.

A continuación, se presenta otra de las técnicas de minería de datos muy utilizada:

2.1.2. Árboles de decisión

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal forma que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta algunas hojas. (Rivero, 2012)

El algoritmo CHAID

Históricamente el análisis de clasificación se reducía al estudio de variables dependientes cuantitativas, utilizando el algoritmo presentado por Morgan y Sonquist (1963 citado por Escobar, 2002). Aquí, sin embargo, se centrará la atención en una derivación de esta técnica, pues en lugar de utilizar la suma cuadrática inter-grupos, se usa el estadístico χ^2 para la selección de los mejores pronosticadores.

De esta manera, en el siguiente apartado se expondrá los pasos lógicos a seguir cuando se considera la variable dependiente en escala nominal explicados por Escobar (2002):

a) Preparación de las variables. Tarea del analista, que debe seleccionar una variable dependiente que sea de interés para el análisis y elegir un conjunto de posibles pronosticadores relevantes (variables nominales, ordinales con pocas categorías, preferiblemente menos de diez, o incluso variables cuantitativas) que permitan realizar una descripción y pronóstico óptimo de la primera variable.
b) Agrupación de las categorías de las variables independientes en el caso de que éstas tengan un perfil similar de la variable dependiente.
c) Primera segmentación consiste en la selección de la variable que mejor prediga la variable dependiente.
d) Segunda segmentación. Para cada segmento formado en el paso anterior, se busca entre las variables cuyos valores han sido previamente agrupados de las misma forma que el paso b), la que tenga mayor poder pronosticador.
e) Sucesivas segmentaciones. Se procede de forma similar al paso anterior en cada grupo formado por la segmentación previa.

2.2. Revisión Bibliográfica (Estado del Arte)

A pesar de que históricamente el sector alimentario generalmente se ha basado en modelos estadísticos clásicos como series temporales y aunque computacionalmente son rápidos en la realización de predicciones, estos métodos pueden no funcionar tan bien en los problemas de pronósticos con patrones de datos altamente complejos. Es ahí donde los métodos de aprendizaje estadístico han ganado terreno en los últimos años con resultados impresionante para hacer este tipo de pronósticos.(Yu et al., 2011)

Un modelo de pronóstico preciso es de gran importancia en una empresa de alimentos porque puede aumentar los ingresos por ventas. Las investigaciones han evidenciado que las compañías que usan un enfoque basado en datos para la toma de decisiones son en, promedio un 6% más rentable de las que no lo hacen. Otros análisis realizados por varias compañías de alimentos mostraron que existe una utilidad de mejora en lo que se refiere a la práctica de pronóstico de ventas, pero también bosquejaron una oportunidad grandiosa sobre los beneficios que genera la recopilación de grandes cantidades de ventas. (Siwerz & Dahlén, 2017)

La aplicación de pronósticos de ventas es una herramienta que también ha sido sugerida para reducir la gestión de inventarios y los pedidos, que son de gran relevancias en estas industrias. (Siwerz & Dahlén, 2017)

Algunos de estos trabajos son los realizados por Bajari et al. (2015a), que en su artículo "Métodos de aprendizaje estadístico para la estimación de la demanda" prueban diferentes modelos para estimar la demanda: entre los cuales se encuentran dos modelos basados en árboles de regresión: embolsado (bagging) y bosques aleatorios(random forest), que son métodos flexibles para aproximar funciones arbitrarias, teniendo como resultado que estos modelos predicen la demanda fuera de la muestra en métricas estándar con mucha más precisión que un panel de datos o un modelo logístico.

Otros autores, Siwerz & Dahlén (2017) para un estudio de predicción de ventas en un departamento de tiendas de alimentos compara tres métodos de aprendizaje estadístico: perceptrón multicapa (MLP), SVM y Red de función de base radial (RBFN), teniendo como resultado que el SVM es el mejor modelo debido a que presenta medidas de errores menores que los otros dos métodos considerados. La comparabilidad se la realizó en función de la medidas Error Porcentual Absoluto Medio (MAPE) y el Error Cuadrático Medio (RMSE).

Para la previsión de la demanda en el comercio minorista de productos perecederos aplicado a un gran minorista con sede en Ecuador (Wang et al., 2019) , concluyó que para bienes específicos el SVM tienen una alta precisión predictiva para tratar productos perecederos. Sin embargo, el estudio también se puede mejorar considerando algunos factores económicos como el PIB, la tasa de inflación, el desempleo, etc.

En cuanto a la precisión de los pronósticos, un estudio realizado por (Huber & Stuckenschmidt, 2020) sobre la predicción de la demanda diaria de las diferentes categorías de productos a nivel de tienda de una cadena de panaderías con énfasis en días especiales de calendario, a través del uso de modelos de aprendizajes automáticos como árboles de regresión potenciados por gradiente, determinaron que este método no solo brinda una mayor precisión, sino que además en particular, para días especiales, el error puede reducirse en más de un 10% y hasta un 20% en comparación a modelos de series temporales con ajustes o modelos de regresión regularizados. Esto también fue demostrado por (Tanizaki et al., 2019), quienes utilizaron métodos como: regresión del árbol potenciado y regresión de bosques de decisiones para pronosticar el número de clientes que visitan a 5 restaurantes importantes de Japón, obteniendo como resultado que la tasa de pronóstico para cualquier tienda superó aproximadamente el 85% de precisión con respecto al número de clientes reales que visitaron los 5 restaurantes.

De igual manera, un estudio realizado por Penpece & Elma (2014) sobre la predicción de los ingresos por ventas del sector alimentario en Turquía mediante el uso de redes neuronales artificiales (ANN), obtuvieron como resultados que los datos estimados presentan desviaciones a la alta o baja solo para el 10 % de los datos reales. Como resultado de este estudio empresas de consumo masivo de Turquía emplean ANN como una herramienta fiable de pronóstico. Esto también fue demostrado por (Slimani, I., El Farissi, I., & Achchab, 2015) quienes utilizaron un MLP para la previsión de demanda para un producto de un supermercado marroquí, obteniendo como resultado un RMSE del 10% con respecto a los valores reales.

Así mismo, Palacios (2020) ha realizado un estudio donde determina que el método de árboles de clasificación y regresión (CART) presenta resultados de predicciones aceptables estando entre el 76% y 99% de eficacia para las predicciones de precios de bienes raíces y comerciales, siendo comparables con modelos robustos de predicción como MLP.

Para la previsión de los ingresos por ventas de un mismo producto para diferentes puntos de ventas de la empresa Big Mart (Shinde et al., 2022), aplica modelos de regresión lineal y algoritmo de árbol de decisiones y también propone un algoritmo basado en gradientes denominado XGBoot, obteniendo como resultado que tanto el método propuesto

como los modelos de árboles de decisión al no ser sus predicciones estadísticamente significativas, ambos modelos se pueden usar de forma eficiente para pronosticar la demanda.

Para la previsión de la demanda de productos para supermercado online (Evers et al., 2018), concluyeron que el modelo de árbol de regresión de decisión tiene una alta precisión que fue del 99.9% para predecir la cantidad de panes por día, dada la cantidad de clientes que van ordenar, teniendo como aporte significativo que sería factible pedir pan directamente al proveedor con un nivel significativo de confiabilidad, de forma desperdicio que el desperdicio debido a la sobreestimación y los pedidos incompletos debido a la subestimación pueden reducirse a niveles aceptables.

Para predecir la demanda de calzado de temporada de un minorista líder de calzado con sede EE. UU (Wing & Chan, 2018), utilizan dos modelos, un modelo general y un modelo de tres pasos, utilizando atributos de productos, calendario y precios para predecir la demanda. Los métodos de aprendizajes aplicados para ambos modelos de forma independiente fueron árboles de regresión (regression trees), bosques aleatorios (random forest), K-vecino más cercano (K-NN), regresión lineal y redes neuronales. El modelo de tres pasos se distingue del modelo general en que consta de tres etapas separadas: agrupación, clasificación y predicción; mientras que el modelo general realiza directamente predicciones individuales por cada método. Los resultados determinaron para el modelo general que los árboles de regresión y de regresión lineal brindaron el mejor rendimiento con una precisión de pronóstico del 31% y un 2% y 3% respectivamente de sesgo del pronóstico (sobre pronóstico).

Para predecir la demanda de calzado de temporada de un minorista líder de calzado con sede EE. UU (Wing & Chan, 2018), utilizan dos modelos, un modelo general y un modelo de tres pasos, utilizando atributos de productos, calendario y precios para predecir la demanda. Los métodos de aprendizajes aplicados para ambos modelos de forma independiente fueron árboles de regresión (regression trees), bosques aleatorios (random forest), K-vecino más cercano (K-NN), regresión lineal y redes neuronales. El modelo de tres pasos se distingue del modelo general en que consta de tres etapas separadas: agrupación, clasificación y predicción; mientras que el modelo general realiza directamente predicciones individuales por cada método. Los resultados determinaron para el modelo general que los

árboles de regresión brindaron el mejor rendimiento con una precisión de pronóstico del 31% y un 2% de sesgo del pronóstico (sobre pronóstico). Muy seguido de la regresión lineal que mostró un desempeño de precisión cercano al 31% y un sesgo del 3% (sobre pronóstico). Mientras que, para el modelo de tres pasos, los algoritmos de mejor rendimiento por grupo fueron la regresión lineal con WMAPE del 28% y un WMPE de menos el 11% para el grupo 1, para el grupo 3 fueron los bosques aleatorios con WMAPE del 32% y un WMPE de más 6% y para el grupo 4 los árboles de regresión presentaron un WMAPE del 39% y un WMPE del 0%.

Por otro lado (Díaz Sepúlveda & Correa, 2013), han realizado un estudio donde comparan a nivel predictivo los modelos de regresión lineal sea cuadrático o trigonométrico con CART mediante simulación, teniendo como resultado para diferentes muestras de datos que el error de predicción de la regresión lineal es menor que el de CART cuando se conoce la forma funcional de los datos. Y también concluyen que el modelo CART es una buena opción cuando se desconoce la forma funcional verdadera de los datos.

2.3. Modelos de aprendizajes estadísticos

Predicción

En muchas situaciones, disponer de un conjunto de variables de entradas X es algo que frecuentemente se encuentra disponible, no así de la variable de salida Y.

Según (James et al., 2013) , dado que el termino de error promedia cero, se puede predecir Y a través de la siguiente ecuación :

$$\hat{Y} = \hat{f}(X) \qquad (4)$$

Donde $\hat{f}$ representa la estimación para f,y $\hat{Y}$ representa el resultado de predicción de Y. En este entorno, $\hat{f}$ a menudo se trata como una caja negra, en el sentido de que uno no suele preocuparse por la forma exacta de $\hat{f}$, siempre que produzca predicciones precisas para Y.

La precisión de $\hat{Y}$ como predicción de Y depende de dos cantidades, a las que llamaremos error deducible y error irreducible

$$E(Y-\hat{Y})^2 = E\left[f(X)+\epsilon-\hat{f}(X)\right]^2$$

$$= E\left[f(X)-\hat{f}(X)\right]^2 + Var(\epsilon) \qquad (5)$$

Reducible Irreducible

Donde $E(Y-\hat{Y})^2$ representa el valor promedio, o esperado, de la diferencia del valor esperado al cuadrado entre el valor predicho y el real de Y, y $Var(\epsilon)$ representa la varianza asociada con el término de error.

En general el error reducible se puede mejorar usando métodos de aprendizajes estadísticos más apropiados para estimar f. Los cuales según (James et al., 2013) se pueden caracterizar como paramétricos y no paramétricos:

2.3.1 Métodos paramétricos: Modelo Regresión lineal

Según lo explica (James et al., 2013), un modelo lineal es un enfoque paramétrico que está basado en dos pasos:

1. Asume un modelo lineal, dado n observaciones de una variable observable Y en función de varias variables explicativas definido de la siguiente manera por (Cuadras, 2007):

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{im}\beta_m + \ldots + \epsilon_i\,,\ i = 1, \ldots, n \qquad (6)$$

que en notación matricial es

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Los elementos que intervienen en el modelo lineal son:

1. El vector de observaciones de Y

$$\mathrm{y} = (y_1, y_2, \ldots, y_n)'$$

2. El vector de parámetros

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)'$$

3. La matriz de diseño

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

4. El vector de desviaciones

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)'$$

La notación matricial compacta del modelo es:

$$y = X\boldsymbol{\beta} + \epsilon$$

Donde y y X son conocidas. Y en los modelos de regresión X contiene las observaciones de m variables explicativas.

Suposiciones básicas del modelo

Suponiendo que las desviaciones aleatorias o errores ϵ_i del modelo lineal se asimilan a n variables aleatorias con media 0, incorrelacionadas y con varianza común σ^2, es decir, satisfacen:

1. $\boldsymbol{E}(\epsilon_i) = 0,\ i = 1, \dots, n$
2. $\boldsymbol{E}(\epsilon_i \epsilon_j) = 0,\ i \neq 1, \dots, n$
3. $\boldsymbol{var}(\epsilon_i) = \sigma^2,\ i = 1, \dots, n$

Estas condiciones equivalen a decir que el vector de medias y la matriz de covarianzas del vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)'$ son:

$$E(\boldsymbol{\epsilon}) = 0, \quad \sum_{\boldsymbol{\epsilon}} = \sigma^2 I_p \qquad (7)$$

2. Después de seleccionar un modelo, se necesita un procedimiento que use los datos de entrenamientos para ajustar o entrenar el modelo lineal. Tal método es el de mínimos cuadrados ordinarios (MCO) que se usa para estimar el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m)'$, los cuales minimizan la suma cuadráticas de errores entre el i-ésimo valor de respuesta observado y el i-ésimo valor de respuesta estimado.

Ajuste del modelo lineal

Dos de las medidas más recurrentes para el ajuste del modelo son el error estándar residual (RSE) y Coeficiente de determinación R^2. Un valor R^2 cercano a 1 determina que el modelo explica gran parte de la varianza en la variable de respuesta y un valor cercano a cero indica que la regresión no explica gran parte de la variabilidad en la variable de respuesta. En cambio, el RSE se define como

$$RSE = \sqrt{\frac{1}{n-p-1} RSS} \tag{8}$$

donde RSS es la suma de residuos cuadrados y nos indica que si la disminución de este valor es mayor en relación con el aumento de p, entonces el RSE disminuye y mejor se ajusta el modelo a los datos.

2.3.2 Métodos no paramétricos: K-vecinos más cercanos (KNN) y Árboles de Regresión y Clasificación (CART)

Según (James et al., 2013), los métodos no paramétricos no hacen suposiciones explícitas sobre la forma funcional de f, y por lo tanto brindan un enfoque alternativo y más flexible para realizar la regresión. Sin embargo, los enfoques paramétricos presentan una gran desventaja: dado que no reducen el problema de estimar f a un pequeño número de parámetros, se necesita una gran cantidad de observaciones (mucho mayor de lo que se usa para un enfoque paramétricos) para obtener una estimación precisa de f. Dos ellos se expresan a continuación:

Algoritmo K-vecinos más cercanos (KNN)

Según (James et al., 2013), uno de los métodos más simples y conocidos de los métodos no paramétricos, es el método de regresión de K-vecinos más cercanos (KNN).

Dado un valor de K y un punto de predicción x_0, la regresión de KNN:

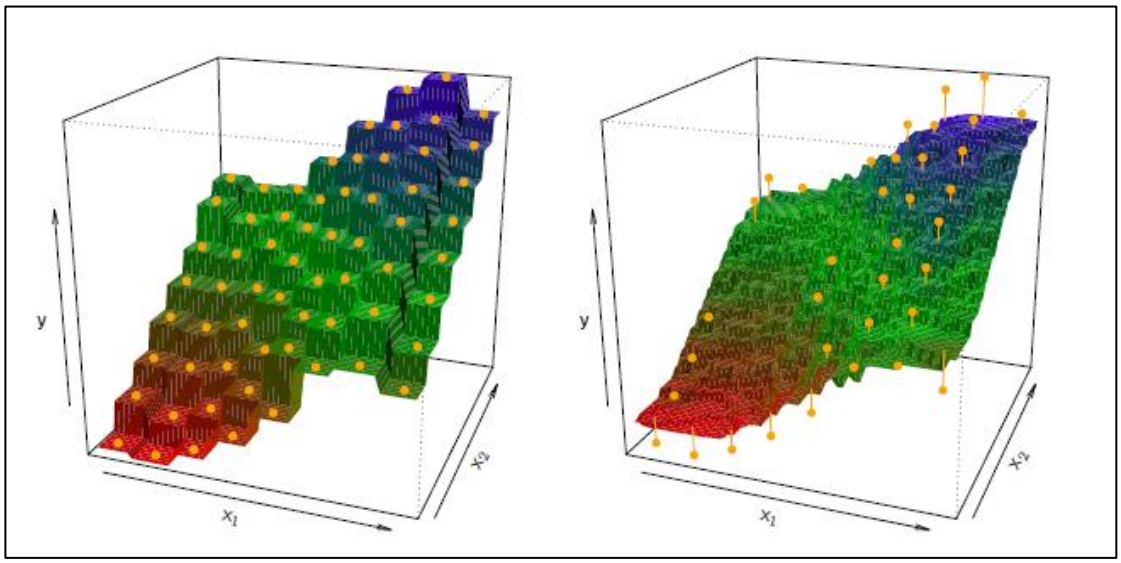
- Primero identifica las K observaciones de entrenamiento más cercanas a x_0, representadas por N_0.

- Luego estima $f(x_0)$ usando el promedio de todas las respuestas de entrenamiento en N_0. En otras palabras,

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_0 \epsilon N_0} y_i \quad (9)$$

Figura 2. 5

Gráficas de Y estimada utilizando la regresión KNN para diferentes K



Nota: En un conjunto de datos bidimensional con n observaciones (puntos naranjas). Izquierda: K=1 da como resultado un ajuste de función de paso aproximado. Derecha: K=9 produce un ajuste mucho más suave.
Fuente: (James et al., 2013)

La Figura 2.5 muestra dos ajustes KNN en un conjunto de datos con p=2 predictores. Observamos que cuando K=1 (panel izquierda), el ajuste interpola perfectamente las observaciones de entrenamiento, en consecuencia, toma la forma de una función escalonada. En cambio, cuando K=9, el ajuste KNN sigue siendo una función escalonada, pero el promedio de nueve observaciones da como resultado regiones mucho más pequeñas de predicción constante y, en consecuencia, un ajuste más suave. En general, el valor óptimo para K dependerá de la compensación sesgo-varianza. Un valor pequeño para K proporciona el ajuste más flexible, que tendrá un sesgo bajo, pero una varianza alta. Por el contrario, valores altos de K proporcionan ajustes más suaves y variables, a costo de un alto sesgo.(James et al., 2013)

Arboles de clasificación y regresión (CART)

La metodología CART según (Díaz Sepúlveda & Correa, 2013), utiliza datos históricos para construir arboles de clasificación o de regresión, los cuales son muy usados para clasificar o predecir nuevos datos. Entre las ventajas de los árboles CART está su robustez para trabajar con outliers, la invarianza en la estructura de sus árboles de clasificación yo de regresión a transformaciones monótonas de las variables independientes, y sobre todo su interpretabilidad.

El proceso para la construcción de un árbol de regresión según (Espinoza, E. N., & Bancalari, 2005), viene determinado de la siguiente manera:

Dado una regresión múltiple $y_i = f(x_{i1}, \dots, x_{ip}) + \epsilon_i,\ i = 1, \dots n,$ donde f es desconocido y no fácilmente parametrizado; los x_{ij} son variables independientes conocidas, y los ϵ_i son términos de error aleatorio con media cero. Un nodo N es un subgrupo de índices $\{1, \dots ,\ n\}$. La desviance de un nodo N se define como:

$$D(N) = \sum_{i \epsilon N}(y_i - \bar{y}(N))^2 \qquad (10)$$

Donde $\bar{y}(N)$ es la media de las observaciones en el nodo N.

- El nodo raíz consiste en todas las observaciones, y en cada paso, el nodo parental se divide recursivamente en dos nodos hijos, un nodo izquierdo (N_{izq}) y un nodo derecho (N_{der}), para así minimizar (N_{izq}) + (N_{der}).
- La partición de nodos se realiza considerando para el caso de variables continuas, todas las divisiones de la forma $N_{izq} = \{ i \epsilon N: x_{ij} \leq t\}$, $N_{der} = \{ i \epsilon N: x_{ij} > t\}$ para t constantes.
- Para cada variable independiente, se consideran todas las posibles particiones, calculando la desviance para el siguiente nodo a dividir D (N_{izq}) + D (N_{der}). Se calculan las particiones candidatas para cada variable independiente, y se seleccionan las variables que produzcan las mejores divisiones (con menor desviance) para particionar el nodo N. Se procede recursivamente hasta que no sea posible realizar la siguiente partición, de acuerdo con criterios predeterminados.

Elección del mejor árbol

La elección del mejor árbol respecto a otro dependerá de la estimación de la tasa de error $R(T)$. Una de las formas más destacadas para estimar dicha tasa, es la estimación por validación cruzada (cross-validation) que consiste en estimar $R(T)$ al estimador por muestra de validación, en forma reiterada y análoga. (Espinoza, E. N., & Bancalari, 2005)

$$R^{cv}(T) = \frac{R^1(T),\ldots,R^k(T)}{k} \quad (11)$$

Donde $R^1(T), \ldots, R^k(T)$ son la k estimaciones

2.4. Métricas de evaluación

Para medir la precisión de un modelo regresión, se usaron 3 métricas de evaluación: RMSE, Coeficiente de determinación R-Square y MAPE. Cuanto mayor sea el valor de R^2 y cuanto menor sea el valor de RSME y MAPE, mejor el rendimiento predictivo del modelo.

Error cuadrático medio (RMSE)

El RSME es el error cuadrático promedio del valor predicho y real

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}} \quad (12)$$

Donde $\boldsymbol{\hat{y}_i}$ es la predicción que da Y para la i-ésima observación. El RSME será pequeño si las respuestas predichas están muy próximas de las respuestas verdaderas, y será grande si para algunas de las observaciones, las respuestas pronosticadas y verdaderas difieren sustancialmente.

Coeficiente de determinación (R-Square)

El estadístico R^2 es una medida de ajuste alternativa y toma la forma de una proporción de la varianza explicada, por lo que siempre toma un valor entre 0 y 1, y es independiente de la escala de Y. Se define como:

$$\boldsymbol{R^2} = \frac{\boldsymbol{TSS - RSS}}{\boldsymbol{RSS}} = \mathbf{1} - \frac{\boldsymbol{RSS}}{\boldsymbol{TSS}} \quad (13)$$

donde $RSS = \sum_{t=1}^{n}(y_i - \hat{y}_i)^2$ es la suma de residuos cuadrados y $\boldsymbol{TSS} = \sum(y_i - \overline{y})^2$ es la suma total de los cuadrados.

Un valor R^2 cercano a 1 determina que el modelo explica gran parte de la varianza en la variable de respuesta y un valor cercano a cero indica que la regresión no explica gran parte de la variabilidad en la variable de respuesta.

Error porcentual absoluto medio (MAPE)

El valor MAPE es el porcentaje de errores promedio entre el valor real y el predicho:

$$\boldsymbol{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (14)$$

Donde y_i , $\hat{y}_i$ es el valor real y predicho respectivamente y n es el número de ejemplos de entrenamiento.

CAPÍTULO 3

3. METODOLOGÍA

3.1. Metodología de la investigación

3.1.1 Enfoque de la investigación

El enfoque del presente estudio es de carácter cuantitativo mayoritariamente ya que se buscó estimar pronósticos de ventas de productos a un grupo de clientes homogéneos que afectan a la cada de suministro de una empresa de consumo masivo de la ciudad de Guayaquil mediante la aplicación de métodos de aprendizajes estadísticos como: regresión lineal, K-vecinos más cercano (KNN) y Árboles de clasificación y regresión (CART) utilizando como variables de interés, variables comerciales y relevantes del sector económico y financiero del país.

3.1.2 Diseño de la investigación

El diseño de la investigación del presente trabajo corresponde a un estudio exploratorio, descriptivo, correlacional y explicativo de un conjunto de datos cuantitativos (variables de una base de información comercial y variables relevantes del sector económico y financiero del país) de una empresa nacional perteneciente al sector alimentario, de la cual se obtuvo información de los registros de ventas por productos de sus clientes, mediante el uso de bases de la base de datos de las transacciones comerciales de esta empresa de consumo masivo domiciliada en la ciudad de Guayaquil, en la que se busca primeramente segmentar un grupo de clientes con características de compras homogéneas y comparables, para luego a través de un método de aprendizaje estadístico generar un modelo de pronósticos que se ajuste a la estructura de ventas de los productos de consumo masivos de la Empresa AAA y ayude en la reducción de los gastos asociados a la cadena de suministro.

La recopilación de datos cuantitativos para realizar el presente trabajo de pronósticos de ventas producto de consumo masivo se lo determinó a partir de una muestra de 769 clientes activos y 783 productos activos, proporcionadas por fuentes de datos internas de la empresa AAA perteneciente al sector alimentario de la ciudad de Guayaquil.

3.1.3 Fuente

La información de las variables de ventas se la recopiló de una fuente primaria interna proporcionada por la Empresa AAA perteneciente al sector alimentario de la ciudad de Guayaquil.

3.2. Comprensión de los datos del negocio

La estructura del sistema de información comercial de ventas del negocio se encuentra definida a partir del modelo dimensional que se muestra en la Figura 3.1:

Figura 3. 1

Modelo dimensional de la Empresa AAA

Tal como se puede apreciar en la Figura 3.1 - Modelo dimensional de ventas, las ventas en el sistema de negocio de la Empresa AAA constituyen la tabla de hechos y sus asociaciones, constituidas por el cliente y producto, que representan las dimensiones. Para este caso, se detalla que los datos en el modelo dimensional de las ventas poseen una granularidad de un registro por cada ítem de factura por cada producto que realizó la Empresa AAA a un determinado cliente en el periodo del 2017 al 2021.

Detalle de las tablas involucradas en el modelo dimensional de ventas

Tabla de hechos

Tabla 3. 1

Diseño de la tabla de hechos ventas

Nombre del Atributo	Tipo de Dato
Factura	Integer
Fecha	Datetime
Código Bodega	Integer
Cantidad (Unidades)	Float
PVP Unitario	Float
Venta Parcial	Float
Porcentaje Descuento	Float
Descuento	Float
Venta Neta	Float
SKU(FC)	Integer
Código Cliente (FK)	Integer

Nota: FK= Foreign Key

Dimensiones

Tabla 3. 2

Diseño de la dimensión cliente

Nombre del Atributo	Tipo de Dato
Código Cliente (PK)	Integer
Sucursal	Varchar
RUC	Integer
Cliente	Varchar
Dirección	Varchar
Canal	Varchar
Subcanal	Varchar

Tipología del Cliente	Varchar
Ciudad	Varchar
País	Varchar
Zona	Varchar
Tipo Cédula o RUC	Integer
Cupo	Float
Latitud	Decimal
Longitud	Decimal
Naturaleza del Cliente	Varchar
Tiempo de Recuperación Cartera (Días)	Integer
Fecha Creación	Datetime

Nota: PK= Primary Key

Tabla 3. 3

Diseño de la dimensión producto

Nombre del Atributo	Tipo de Dato
SKU(PK)	Integer
Producto	Varchar
Categoría	Varchar
Segmento	Varchar
Unidades por Cajas	Integer

Nota: PK= Primary Key

3.3. Procesamiento de datos

Para el procesamiento de datos, como primer paso se obtuvo las bases de transacciones comerciales de la empresa de estudio, de la cual se consiguió la tabla de registros de ventas a nivel nacional del periodo 2017-2021, la tabla de clientes históricos dentro de la ciudad de Guayaquil y la tabla de productos disponibles.

3.3.1. Procesamiento de datos de la tabla de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil

Para la obtención de la tabla de registros de ventas diarias de clientes con productos disponibles dentro de la ciudad de Guayaquil se procedió de la siguiente manera:

En primera instancia se procedió a combinar a través del campo Cliente la tabla de registros de ventas a nivel nacional del periodo 2017-2021 (Ver Anexo 1) que contenía los registros de ventas diarias de 5.293 clientes con la tabla de clientes históricos de la ciudad de Guayaquil (Ver Anexo 2) que disponía de 1.866 registros únicos. Producto de esta combinación resultó la tabla de registros de ventas de 1.326 clientes de Guayaquil en el periodo 2017-2021 (Ver Anexo 3) con registros de ventas diarias correspondientes a 2.581 SKU en total.

Como segunda instancia se combinó mediante el campo SKU la tabla resultante anterior mencionada con la tabla de productos disponibles (Ver Anexo 4)**,** la cual disponía de 1.044 SKU únicos y como consecuencia de esta operación resultó tabla de registros de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil preliminar (Ver Anexo 5), la cual presenta registros de ventas diarias de 1.228 clientes de Guayaquil y de 901 SKU**.** La reducción de 98 clientes de la muestra de 1.326 que habían resultado de la primera combinación se debió a que estos no tenían registros de los SKU que estaban considerados dentro del listado de productos disponibles que distribuye la empresa AAA a nivel nacional.

Una vez determinada la tabla de registros de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil, se presentó tres escenarios de depuración por lo que a esta última se la denominó preliminar ya que estos escenarios debían ser revisados

y corregidos para tener un conjunto de dato limpio para el análisis, los cuales son explicados a continuación:

1. **Homologación de productos y de clientes**

Al añadir la descripción de los SKU a la tabla anterior mencionada, resultó que diferentes SKU tenían una misma descripción, por ejemplo: los SKU's "12345" y 5063020/R67 compartían la misma descripción del producto denominada "Ot.Toa Hum. Fisher Price 50x24pq", por lo que se procedió a homologarlo a un solo código en este caso el elegido fue el 12345. En total fueron 17 SKU´s que fueron homologados (Ver Anexo 6).

En cuanto a la homologación de clientes partían del evento que diferentes códigos de cliente compartían un mismo RUC o CEDULA, por ejemplo: los códigos de clientes6385 y 2360 compartían el mismo número de cédula:0201245859, por lo que se homologó al código 5893. En total fueron 6 códigos de clientes que fueron homologados (Ver Anexo 7).

Una vez realizado el primer ajuste la tabla de estudio, disminuyó su número de clientes a 1.222 y el número de productos únicos se contrajo a 884.

2. **Exclusión de los productos que tienen precios de:0, 0.001, 0.01 y 0.013**

La exclusión de los productos que mantienen estos precios es debido a que son promociones y reflejan un valor de cero en las ventas netas, no así en las unidades vendidas.

Una vez realizado el segundo ajuste la tabla de estudio preliminar, disminuyó su número de clientes a 1.218 y el número de productos únicos se contrajo a 825.

3. **Exclusión de clientes que no han realizado compras durante los dos últimos años (2020-2021)**

Al aplicar este tercer ajuste, se tiene como resultado la tabla de registros de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil final, la cual presenta el registro de ventas de 769 (Ver Anexo 8) clientes y de 783 productos (Ver Anexo 9), presentando una disminución de 449 clientes y 42 productos con respecto a la tabla preliminar que estaba siendo objeto de estudio.

En la siguiente Tabla 3.4 se realiza una descripción detallada de las variables originales consideradas luego de combinar y depurar la Tabla 3.1- Diseño de la tabla de hechos Ventas con las tablas de dimensiones: Tabla 3.2 – Diseño de la dimensión Cliente y Tabla 3.3 – Diseño de la dimensión Producto y adicionalmente se agrega la variable Cajas que resulta del cociente entre las variables Cantidad (Unidades)/ Unidades por Cajas.

Tabla 3. 4

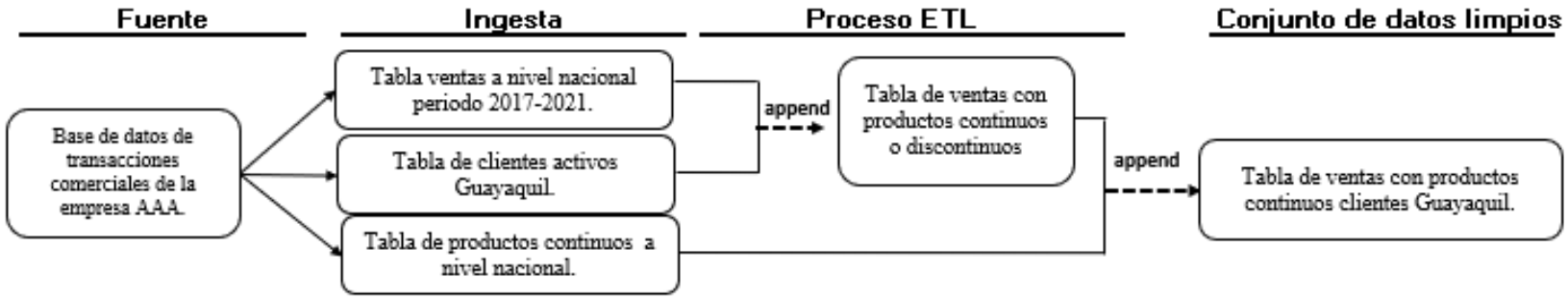
Registros de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil

Nombre del Atributo	Tipo de Dato	Descripción
Factura	Integer	El código de factura realizada por un determinado cliente.
Fecha	Datetime	La fecha en que se registró la venta de un producto a un determinado cliente en el periodo 2017-2021.
Venta Neta	Float	La compra en dólares de un determinado producto realizado por un cliente en el periodo 2017 y 2021.
Código Cliente (PK)	Integer	El código de identificación del cliente.
Cliente	Varchar	Nombre del cliente
Tipología del Cliente	Varchar	Representa la tipología del cliente: Mayorista, Tienda, Abarrotes, Minimarkets y Otros**.**
Zona	Varchar	Representa la zona donde se encuentra ubicado cada cliente dentro de la ciudad de Guayaquil: Norte, Noroeste, Sur, Suroeste y Centro.
Naturaleza del Cliente	Varchar	1: Hombre= H 2: Mujer= M 3: Persona Jurídica=PJ

Tiempo de Recuperación Cartera (Días)	Integer	El número de días máximo que se le otorga a cada cliente cuando se le realiza una venta a crédito.
Fecha Creación	Datetime	La fecha en que fue creado el cliente dentro del sistema de la Empresa AAA.
SKU(PK)	Varchar	El código de identificación del producto de la Empresa AAA.
Producto	Varchar	Nombre del producto
Categoría	Varchar	Categoría de los productos
Segmento	Varchar	Representa el segmento de compra al que pertenece cada producto de la Empresa AAA: 1: Cuidado Oral 2: Cuidado Hogar 3: Cuidado Personal 4: Cuidado Higiene 5: Alimentos
Cajas	Float	La compra en cajas de un determinado producto realizado por un cliente en el periodo 2017 y 2021.

Figura 3. 2

Estructura del procesamiento de datos



3.4. Preparación y determinación de las variables para la segmentación y caracterización de los clientes

Se procedió por entender el contexto de los clientes dentro de la empresa, para ello se identificó dos variables de interés relevantes como son las compras netas acumuladas y la cantidad de productos demandados por los clientes durante el periodo comprendido entre los años 2020 y 2021, para poder obtener los segmentos de clientes. Esta fase de preparación

La técnica de minería de datos usada fue Agrupación de particiones que utiliza algoritmos como: K-means y K-medoids para formar particiones de observaciones que al agruparse dentro de un mismo grupo son muy similares entre ellas y muy distintas a las observaciones de otros grupos.

Lo primero que se definió fue que la cantidad de grupos en los que se debían dividir los clientes serían 4 y esto se lo determinó a través de los resultados obtenido por tres métodos para determinar el número óptimo de clústers como lo son: el método del codo, método de silueta promedio y método del estadístico de brecha. Debido a que el tamaño de clientes agrupado en el cluster 4 era grande, se procedió a realizar una segunda clusterización para ese grupo, siendo elegido nuevamente 4 como el número óptimo escogido. Por lo que finalmente se formaron 7 clusters, considerando las dos variables de interés relevantes construidas a partir de las variables originales Venta Neta y SKU mencionadas en la Tabla 3.1 – Diseño de la tabla de hechos Ventas, cada una de ellas por cada cliente, que se muestra a continuación en la Tabla 3.5:

Tabla 3. 5

Tipología de la segmentación de los clientes

Variable	Tipo de Variable	Descripción de la variable
Montos de Compra	Cuantitativa Continua	Suma de las compras netas por cada cliente durante el periodo comprendido entre los años 2020 y 2021.
Cantidad de productos demandados	Cuantitativa Discreta	El número de productos total demandados por cada cliente durante el periodo comprendido entre los años 2020 y 2021.

Una vez realizada la segmentación de clientes determinada en 7 grupos se procedió a realizar la caracterización de los clientes, a través de la técnica Árboles de decisión para encontrar características adicionales en los grupos encontrados, ya que esta técnica hace posible dividir los datos en pequeños grupos de manera sucesiva encontrando que mientras más divisiones sucesivas se apliquen, más similares son los elementos de cada subgrupo.(Cálad, 2015)

De un total de 13 variables consideradas inicialmente, se decidió realizar el proceso con solo 12 de ellas, debido a que una de ellas aportaba información redundante para realizar el proceso de caracterización, como fue el caso de la variable Cantidad de Facturas Realizadas ya que mantenía una correlación fuerte positiva de 0.8362 con la variable Montos de Compra y de 0.8988 con la variable Frecuencia.

En la siguiente Tabla 3.6 se realiza una descripción detallada de las variables construidas y otras tomadas a partir Tabla 3.1 – Diseño de la tabla de hechos Ventas, cada una de ellas por cada cliente:

Tabla 3. 6

Tipología de la Caracterización de los Clientes

Variable	Tipo de Variable	Descripción de la variable
Segmento de Compra	Cualitativa Politómica	1: Cuidado Oral=CO 2: Cuidado Hogar=CHO 3: Cuidado Personal=CP 4: Cuidado Higiene=CH 5: Alimentos=A
Tipología de cliente	Cualitativa Politómica	Representa la tipología del cliente: 1: Mayorista=MAY 2: Tienda=TD 3: Abarrotes=AB 4: Minimarkets=MN 5: Otros=OT
Zona	Cualitativa Politómica	Representa la zona donde se encuentra ubicado cada cliente dentro de la ciudad de Guayaquil: 1: Norte=NT 2: Noroeste=NOE 3: Sur=SR 4: Suroeste=SOE 5: Centro=CT
Naturaleza del cliente	Cualitativa Politómica	1: Hombre= H 2: Mujer= M 3: Persona Jurídica=PJ

Montos de Compra	Cuantitativa Continua	Suma de los montos del mayor segmento de compra del cliente durante el periodo comprendido entre los años 2020 y 2021.
Frecuencia	Cuantitativa Discreta	Es el número total de transacciones efectuadas del mayor segmento de compra de cada cliente durante el periodo comprendido entre los años 2020 y 2021.
Recencia	Cuantitativa Discreta	El número de días transcurridos desde la última compra del mayor segmento del cliente, durante el periodo comprendido entre el 1 de enero de 2017 y el 3 de enero del 2022.
Portafolio	Cuantitativa Discreta	El número de productos del mayor segmento de compra del cliente durante el periodo comprendido entre los años 2020 y 2021.
Ticket Promedio	Cuantitativa Continua	El promedio de compra que cada cliente realiza de su mayor segmento de compra durante el periodo comprendido entre los años 2020 y 2021.
Antigüedad	Cuantitativa Discreta	El número de días de cada cliente desde su fecha de creación.

Frecuencia Anual del último año de compra.	Cuantitativa Discreta	El número de meses de compras que realizó el cliente el último año de compra de su mayor segmento.
Frecuencia Promedio Mensual del último año de compra.	Cuantitativa Discreta	El número de semanas promedio mensual de compras que realizó el cliente el último año de compra de su mayor segmento.

3.5. Modelización

Esta etapa tiene como objetivo principal realizar una primera aplicación de los algoritmos de aprendizajes estadísticos como: regresión lineal, K-vecinos más cercano (KNN) y Árboles de clasificación y regresión (CART). Para llevar a cabo es necesario definir las variables de interés, variables comerciales que aporten significativamente a los modelos. Una vez determinadas las variables relevantes, se ajustarán los modelos utilizando las diferentes técnicas de aprendizaje estadístico estudiados en la bibliografía mediante el enfoque de validación cruzada que garantice el equilibrio entre sesgo y varianza.

3.6. Evaluación y elección del mejor modelo

Una vez que se tengan las predicciones preliminares de los productos por clusters, lo que se hará con el fin de garantizar los pronósticos se dividirá el conjunto de datos en: conjunto de entrenamiento y conjunto de prueba, evitando así el sobreajuste de modelos (over-fitted). Se revisarán los resultados y se seleccionará el modelo que mejor se ajuste a los datos considerando la métricas de evaluación mencionadas en el apartado (2.4). Y finalmente se procederá a detallar las conclusiones y recomendaciones de la presente investigación.

CAPÍTULO 4

4. RESULTADOS

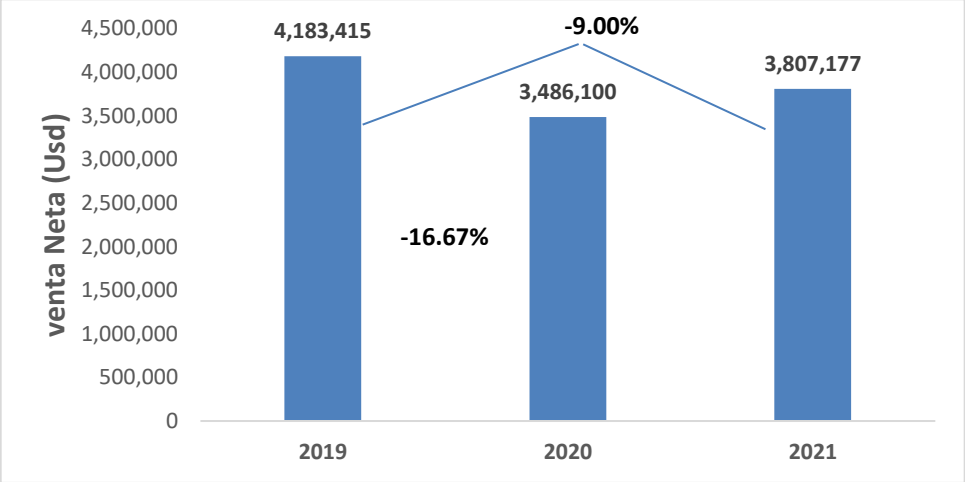
4.1. Análisis interno de la Empresa AAA

La empresa para el presente análisis está conformada por intermediarios minoristas que según Acosta(2017), se caracterizan por comprar productos o servicios a los mayoristas y revenderlos directamente al consumidor. Y por mayoristas generales que no se especializan en la distribución de ningún producto en particular y por tanto comercializan un sinfín de productos o servicios.

Demanda del canal tradicional

Figura 4. 1

Demanda del canal tradicional, 2019-2021



Fuente: Empresa AAA, Información de Ventas 2017-2021

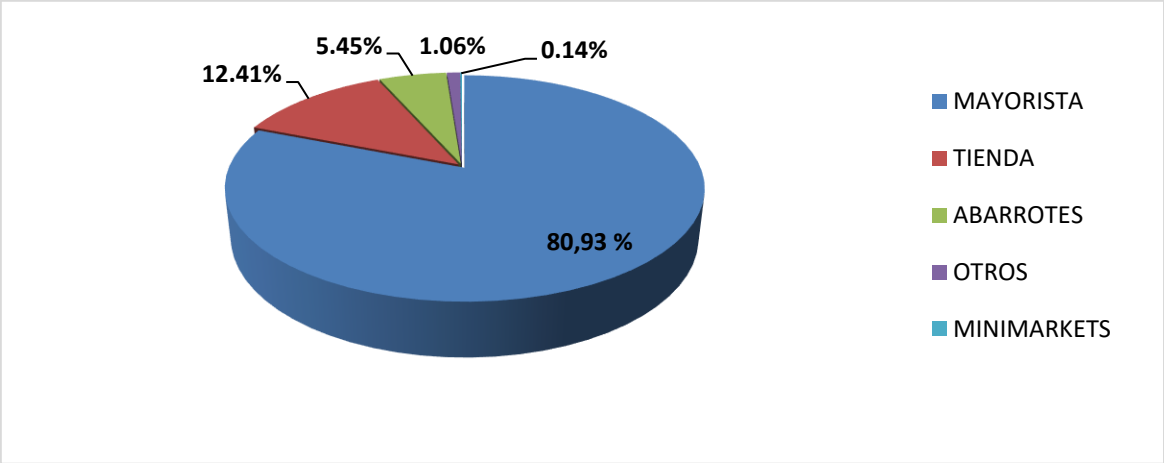
Según la Figura 4.1 en el año 2020 la Empresa AAA facturó 3’486.100 dólares, al canal tradicional. Si relacionamos con el año 2019 en el que se facturó 4’183.415, se puede observar un decrecimiento del 16,67%. En cambio, en el año 2021 la Empresa AAA facturó 3’807.177 dólares. Si relacionamos con el año 2020, se puede observar un crecimiento del 9,21%. Sin embargo, si comparamos el año 2021 con respecto al año 2019, se puede observar decrecimiento del 9,00%.

Demanda por Tipología de cliente del canal tradicional

La demanda realizada a la Empresa AAA por tipología de cliente en Guayaquil durante el año 2021 (Ver Anexo 10), tal como lo muestra la Figura 4.2 fue realizada por 5 tipos de clientes (Mayorista, Abarrotes, Tienda, Otros, Minimarkets); siendo el más representativo por su volumen de participación el Mayorista con el 80,93%, seguido por Abarrotes con el 12,41%, Tienda con el 5,45%, Otros con el 1,06% y Minimarkets con el 0,14% del total de ventas netas facturadas durante el año 2021 que fue de 3’807.177.

Figura 4. 2

Participación de la demanda por tipología de cliente, 2021



Fuente: Empresa AAA, Información de Ventas 2017-2021

En términos de variación porcentual por tipología de cliente a la Empresa AAA en Guayaquil durante el período 2019-2021 (Ver Anexo 11) se obtuvieron los siguientes resultados como muestra la Figura 4.3:

El tipo de cliente Mayorista demandó durante el año 2020, 2’982.758 dólares con una variación que representa el 22,23% de decremento con respecto al año 2019 que fue de 3’835.312 dólares. En cambio, en el año 2021 demandó 3’081.198 dólares. Si relacionamos con el año 2020, se puede observar un crecimiento del 3,30%; sin embargo, si comparamos con respecto al año 2019, se puede observar decrecimiento del 19,66%.

El tipo de cliente Tienda en cambio experimentó un crecimiento de 121,90%, ya que en el año 2020 demandó 260.133 dólares, comparado con el año 2019 que fue de 117.229.

Si se compara el año 2021 que fue de 472.622 con respecto al año 2020, se puede observar un crecimiento de 81.68%. y si lo relacionamos con el año 2019 también experimentó un incremento de 303.16%.

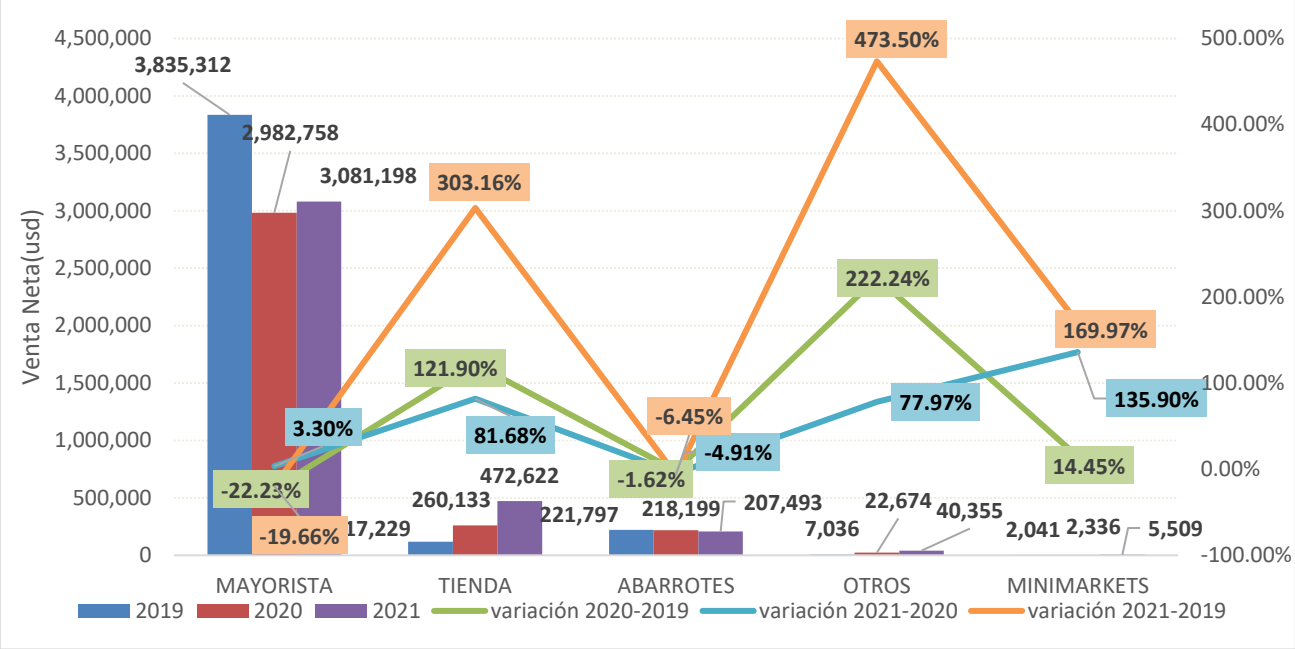
Por su lado Abarrotes durante el año 2020 demandó 218.199 dólares, obteniendo un decremento de 1,62%, comparado con el año 2019 en el que demandó 221.797 dólares de la venta neta. En cambio, en el año 2021 demandó 207.493. Si lo relacionamos con respecto al 2020, se produjo un caída de 4.91%, y el efecto de decrecimiento es aún mayor si lo relacionamos con respecto al año 2019, ya que se puede observar una caída de 6.45%.

Por otra parte, el tipo de cliente Otros evidenció un crecimiento de 222.24% ya que el año 2020 demandó 22.674 dólares, comparado con el año 2019 que fue de 7.036 dólares. También evidenció un crecimiento de 77.97% si se compara el año 2021 que fue de 40.355 con respecto al año 2020; y si lo relacionamos con el año 2019 el efecto de crecimiento es aún mayor con un valor de 473.50%.

Por último, el tipo de cliente Minimarkets durante el año 2020 demandó 2.336 dólares, evidenciando un crecimiento de 14.45%, comparado con el año 2019 en el que demandó 2.041 dólares de la venta neta. En el año 2021, demandó 2.336, presentando un crecimiento de 135.90% con respecto al año 2020 y si lo relacionamos con respecto al año 2019, se evidencia un incremento de 169.97%.

Figura 4. 3

Demanda por tipología de clientes, 2019-2021



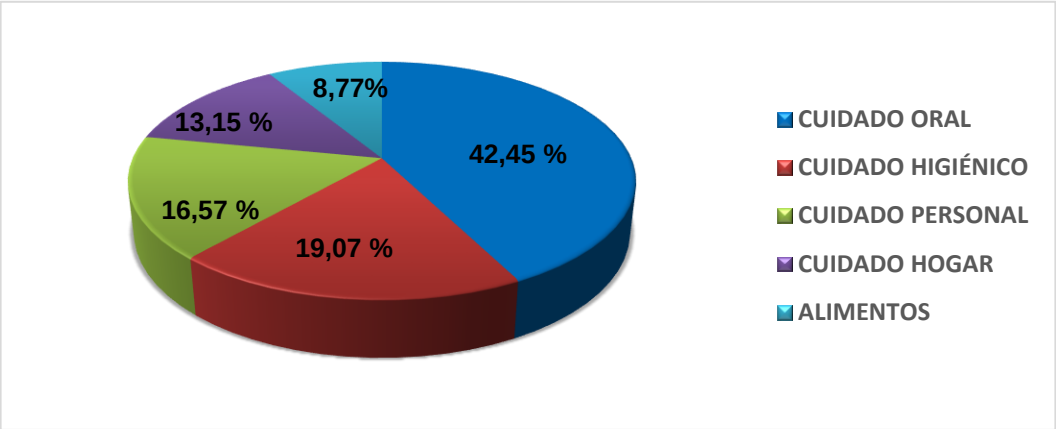
Fuente: Empresa AAA, Información de Ventas 2017-2021

Demanda por Segmento de Productos del canal tradicional

La demanda por segmento de productos a la Empresa AAA en Guayaquil durante el año 2021 (Ver Anexo 12), tal como lo muestra la Figura 4.4 está dividida en 5 segmentos (Cuidado Oral, Cuidado Higiénico, Cuidado Personal, Cuidado Hogar, Alimentos); siendo el más representativo por su volumen de participación el segmento de compra Cuidado Oral con el 42,45%, seguido por Cuidado Higiénico con el 19,07%, Cuidado Personal con el 16,57%, Cuidado Hogar con el 13,15% y Alimentos con el 8,77% del total de ventas netas facturadas durante el año 2021 que fue de 3’807.177.

Figura 4. 4

Participación de la demanda por segmento de productos, 2021



Fuente: Empresa AAA, Información de Ventas 2017-2021

En términos de variación porcentual de la demanda por segmento de productos del canal tradicional a la Empresa AAA en Guayaquil durante el período 2019-2021 (Ver Anexo 13) se obtuvieron los siguientes resultados como muestra la Figura 4.5:

El segmento Cuidado Oral reportó durante el año 2020 una valor de 1’492.834 dólares con una variación que representa el 5,74% de incremento con respecto al año 2019 que fue de 1’411.757 dólares. En cambio, en el año 2021 reportó 1’615.996 dólares. Si relacionamos con el año 2020, se puede observar un crecimiento del 8,25% y si lo comparamos con respecto al año 2019, también refleja un crecimiento de 14,47%.

El segmento Cuidado Higiénico en cambio experimentó un decrecimiento de 67,78%, ya que en el año 2020 demandó 207.646 dólares, comparado con el año 2019 que fue de 644.540. Si se compara el año 2021 que fue de 725.957 con respecto al año 2020, se puede observar un crecimiento exponencial de 249.61% y si lo relacionamos con respecto al año 2019, se observa un crecimiento menor de 12.63%.

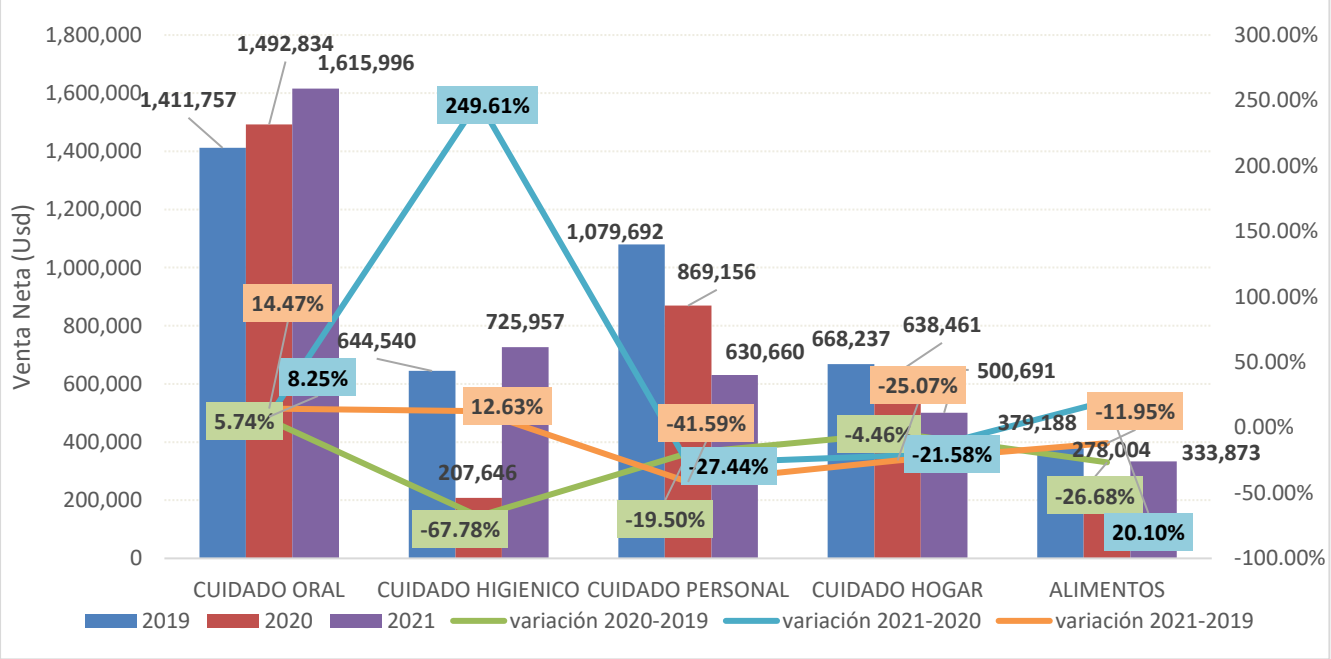
Por su lado Cuidado Personal durante el año 2020 reportó 869.156 dólares, obteniendo un decremento de 19,50%, comparado con el año 2019 en el que reportó un volumen de compra de 1’079.692 dólares de la venta neta. En cambio, en el año 2021 reportó 630.660 dólares. Si lo relacionamos con respecto al 2020, se produjo un caída de 27.44%, y el efecto de decrecimiento es aún mayor si lo relacionamos con respecto al año 2019, ya que se puede observar una caída de 41.59%.

Por otra parte, el segmento Cuidado Hogar evidenció un decremento de 4.46% ya que el año 2020 demandó 638.461 dólares, comparado con el año 2019 que fue de 668.237 dólares. También evidenció un decremento de 21.58% si se compara el año 2021 que fue de 500.691 con respecto al año 2020; y si lo relacionamos con el año 2019 el efecto de decrecimiento es aún mayor con un valor de 25.07%.

Por último, el segmento Alimentos durante el año 2020 reportó un volumen de compra de 278.004 dólares, evidenciando un decrecimiento de 26.68%, comparado con el año 2019 en el que demandó 379.188 dólares de la venta neta. En el año 2021, reportó un volumen de compra de 333.873 dólares, presentando un crecimiento de 20.10% con respecto al año 2020 y si lo relacionamos con respecto al año 2019, se evidencia un decrecimiento de 11.95%.

Figura 4. 5

Demanda por segmento de productos, 2019-2021



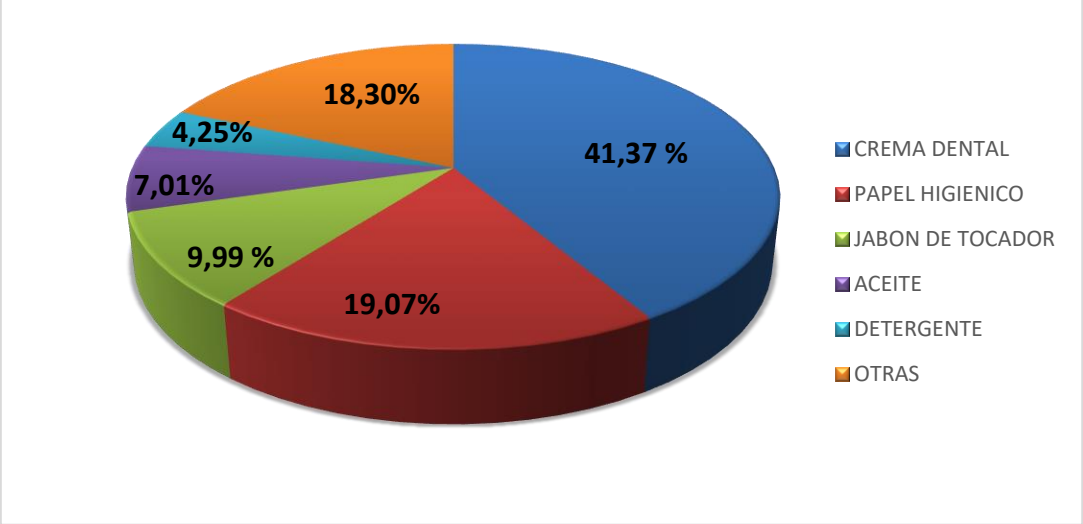
Fuente: Empresa AAA, Información de Ventas 2017-2021

Demanda por Categoría de Productos del canal tradicional

Las 5 categorías con mayor volumen de demanda realizada por el canal tradicional durante los tres últimos años de acuerdo a la Figura 4.6 son: Crema Dental, Papel Higiénico, Jabón Tocador, Aceite y Detergente que en conjunto representan el 81,70% del total de demanda del año 2021 que fue de 3'807.177 dólares, siendo la más representativa la categoría Crema Dental con una participación de 41,37%, seguido de Papel Higiénico con el 19,07%, Jabón Tocador con el 9,99%, Aceite con el 7,01% y Detergente con el 4,25%. El complemento 18.30% corresponde a 42 categorías (Ver Anexo 14) con un nivel de participación individual menor a 3.20% (para consultar cantidades exactas, remítase al Anexo 15).

Figura 4. 6

Participación de la demanda por categoría de productos, 2021



Nota: La categoría "OTRAS" contiene a 42 categorías con participaciones menores a 3,2% del total de demanda en el año 2021.
Fuente: Empresa AAA, Información de Ventas 2017-2021

En términos de variación porcentual de las 5 categorías de productos más demandadas por el canal tradicional a la Empresa AAA en Guayaquil durante el período 2019-2021 (Ver Anexo 16) se obtuvieron los siguientes resultados como muestra la Figura 4.7:

La categoría Crema Dental reportó durante el año 2020 un valor de 1'435.123 dólares con una variación que representa el 5,20% de incremento con respecto al año 2019 que fue de 1'364.156 dólares. En cambio, en el año 2021 reportó 1'575.210 dólares. Si

relacionamos con el año 2020, se puede observar un crecimiento del 9,76% y si lo comparamos con respecto al año 2019, también refleja un crecimiento de 15,47%.

La categoría Papel Higiénico en cambio experimentó un decrecimiento de 67,78%, ya que en el año 2020 demandó 207.646 dólares, comparado con el año 2019 que fue de 644.540. Si se compara el año 2021 que fue de 725.957 con respecto al año 2020, se puede observar un crecimiento exponencial de 249.61% y si lo relacionamos con respecto al año 2019, se observa un crecimiento menor de 12.63%.

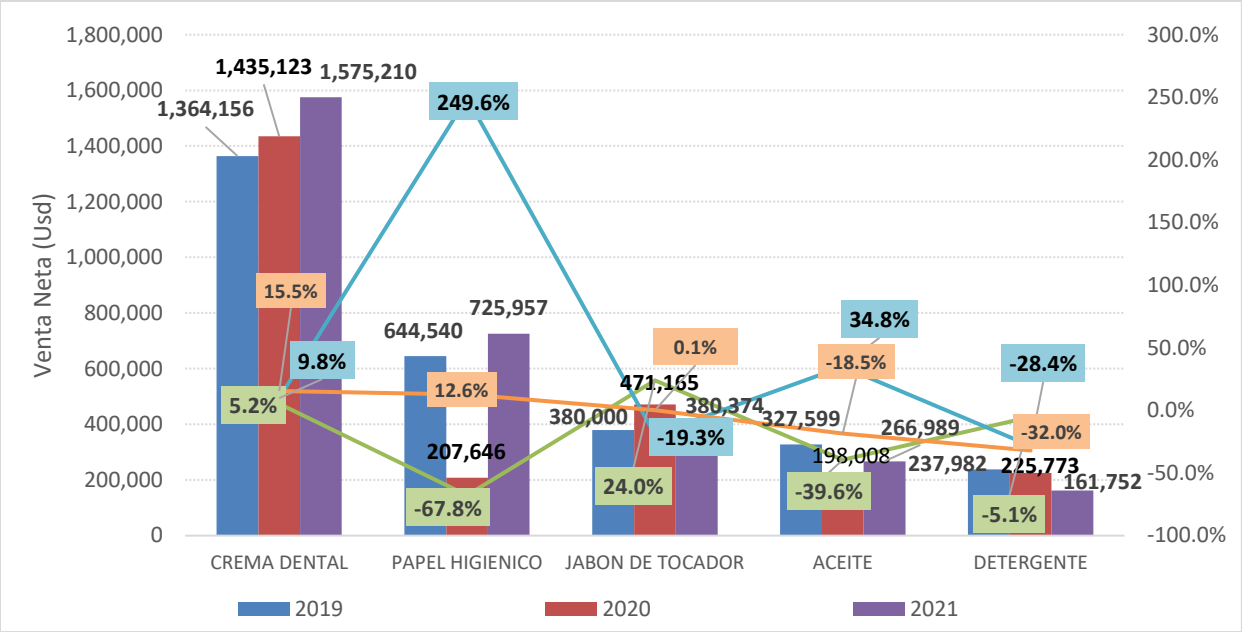
Por su lado Jabón de Tocador durante el año 2020 reportó 471.165 dólares, obteniendo un incremento de 23,99%, comparado con el año 2019 en el que reportó un volumen de compra de 380.00 dólares de la venta neta. En cambio, en el año 2021 reportó 380.374 dólares. Si lo relacionamos con respecto al 2020, se produjo un caída de 19.27%, y si lo comparamos con respecto al año 2019, se puede observar un crecimiento de 0.10%.

Por otra parte, la categoría Aceite evidenció un decremento de 39.56% ya que el año 2020 reportó un valor de 198.008 dólares, comparado con el año 2019 que fue de 327.599 dólares. En cambio, si comparamos el año 2021 que fue de 266.989 con respecto al año 2020 mostró un crecimiento de 34.84%; y si lo relacionamos con respecto al año 2019 se puede observar un decrecimiento de 18.50%.

Por último, la categoría Detergente reflejó un decremento de 5.13% ya que el año 2020 demandó 225.773 dólares, comparado con el año 2019 que fue de 237.982 dólares. También evidenció un decremento de 28.36% si se compara el año 2021 que fue de 161.752 con respecto al año 2020; y si lo relacionamos con el año 2019 el efecto de decrecimiento es aún mayor con un valor de 32.03%.

Figura 4. 7

Demanda de las 5 categorías de productos más representativas, 2019-2021



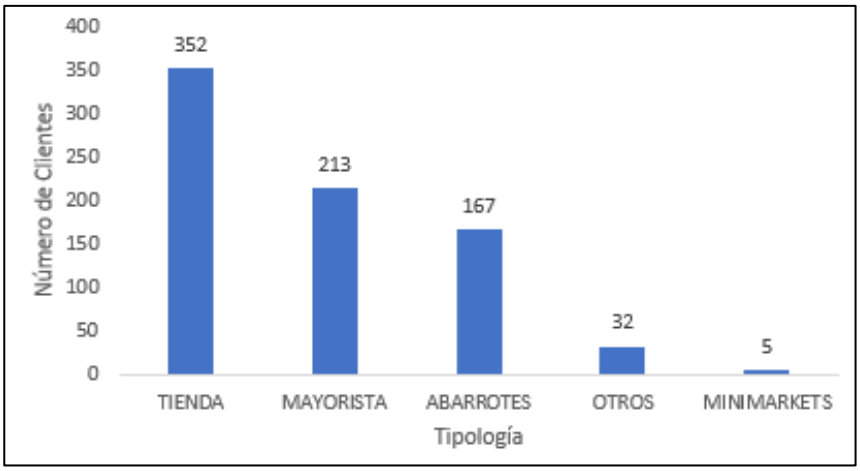
Fuente: Empresa AAA, Información de Ventas 2017-2021

4.2. Análisis descriptivo

En esta sección se utilizará las estadísticas descriptivas, para tener una visión preliminar de la naturaleza de los datos, de los 769 clientes seleccionados dentro de la muestra.

Figura 4. 8

Número de clientes según la tipología del cliente, 2020-2021



Fuente: Empresa AAA, Información de ventas 2017-2021

La figura 4.8 el número de clientes según su tipología, siendo las Tiendas de mayor proporción de clientes con un valor de 352, seguido los Mayoristas con un valor de 213, Abarrotes con 167 clientes, Otros con un valor de 32 y Minimarkets con 5.

Tabla 4. 1

Estadísticas descriptivas de la variable Monto de Compra según la Tipología de Cliente, 2020-2021

Tipología de cliente	n	X	Me	Mo	SD	Min	Max	R	Q3
Total	769	9.484	349	82	64.227	1	1.610.411	1.610.410	2.433
Mayorista	213	28.469	4.325	1.563	117.855	1	1.610.411	1.610.410	15.437
Tienda	352	2.082	215	82	17.824	7	297.532	297.526	340
Abarrotes	167	2.549	780	88	5.582	7	48.592	48.585	2.250
Otros	32	995	374	88	1.962	7	16.238	16.232	1.183
Minimarkets	5	1.569	675		2.452	59	5.877	5.818	3.519

Nota: n=muestra, X= media aritmética, Me=mediana, Mo=moda, SD= desviación estándar, Min= mínimo, Max= máximo, R= rango, Q3= cuartil 3.
Fuente: Empresa AAA, Información de Ventas 2017-2021

Una de las variables de interés para evaluar el comportamiento de compra del canal tradicional que atiende la Empresa AAA es el monto de compra que realizan los clientes durante un tiempo determinado.

La Tabla 4.1 evidencia que el 75% de los clientes (577) de la muestra están concentrados alrededor del intervalo de compra de 1 a 2.433, muy por debajo del valor promedio que es de 9.484. Además, se puede observar una desviación estándar alta de 64.227, lo cual refleja que existen montos de compra altos de clientes que producen esta fuerte desviación.

Al observar la Tabla 4.1, esta nos permite corroborar que la alta dispersión de los datos y desviación estándar se produce por el tipo de cliente Mayorista ya que 79 clientes que representa el 37% del total de la muestra de clientes mayoristas que es de 213, superan el valor promedio total de la muestra que es de 9.484. También se puede describir a partir del

cuartil 3 de la muestra total que representa el 75% de los clientes que compran dentro del intervalo de 1 a 2.433 que la mayoría de los clientes de las diferentes tipologías concentran sus montos de compras por debajo de ese valor excepto de minimarkets y mayorista; sin embargo al observar la mediana de esta última tipología de cliente que es de 4.325 y que representa el 50% de los clientes de su muestra que es 213, se puede interpretar que cierta participación de esa tipología también concentran clientes dentro del intervalo de compra de 1 a 2.433.

Por otro lado, el complemento del intervalo de compra del cuartil 3 de la muestra total que sería de 2.433 a 1'610.411, agrupa también diferentes nichos de clientes indiferente su tipología. Este análisis descriptivo nos permite bosquejar que a pesar de que ya existe una segmentación de clientes por su tipología de cliente, esta no discrimina grupos de clientes que se asemejen de forma particular por una o varias características de compra de cliente. Otras variables de interés que refuerzan que la tipología de cliente no es una variable suficiente para discriminar un determinado grupo de cliente es Cantidad de productos demandados (Tabla 3.5) y Frecuencia de compra (Ver Anexo 17).

Tabla 4. 2

Estadísticas descriptivas de la variable Cantidad de productos demandados según la Tipología de Cliente, 2020-2021.

Tipología de cliente	n	X	Me	Mo	SD	Min	Max	R	Q3
Total	769	17	14	14	19	1	261	260	19
Mayorista	213	21	10	1	29	1	261	260	28
Tienda	352	16	14	14	13	1	124	123	18
Abarrotes	167	16	13	4	14	1	71	70	23
Otros	32	14	11	1	14	1	71	70	20
Minimarkets	5	16	4		27	1	64	63	35

Nota: n=muestra, X= media aritmética, Me=mediana, Mo=moda, SD= desviación estándar, Min= mínimo, Max= máximo, R= rango, Q3= cuartil 3.
Fuente: Empresa AAA, Información de Ventas 2017-2021

La Tabla 4.2, refleja que el promedio de productos comprados por la tipología de cliente mayorista que fue de 21 supera al promedio de la muestra total que fue de 17. En cambio, las demás tipologías se encuentran por debajo de este valor, teniendo un valor igual de 16 productos comprados las tipologías Tienda, Abarrotes, Minimarkets y un valor menor la tipología Otros que fue de 14.

A partir del cuartil 3 de la muestra total que representan el 75% de los clientes que han demandados productos dentro del intervalo de 1 a 19, se puede observar en la tabla 8 que solamente la tipología de clientes Tienda concentra la mayoría de sus datos dentro de este intervalo, mientras que la demás tipologías concentran el 25% de sus datos restantes por encima de este valor.

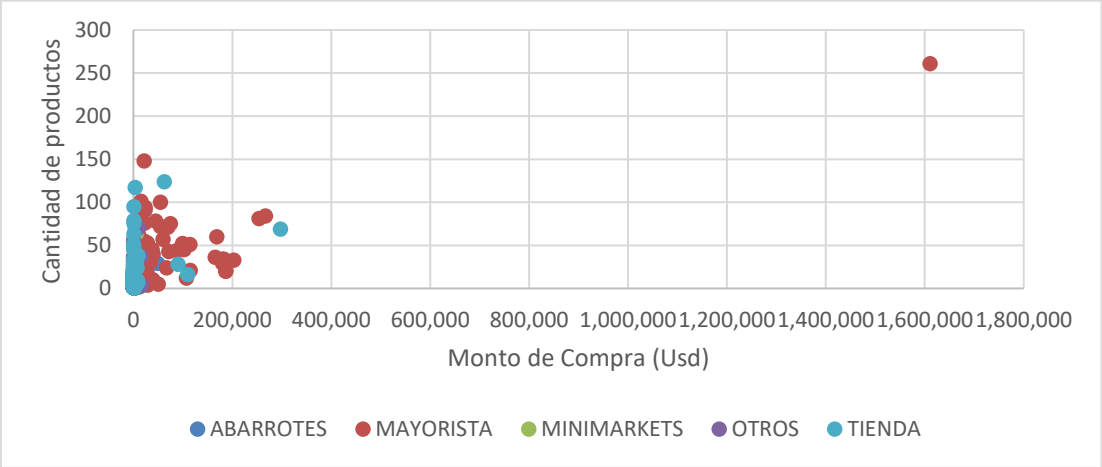
Al observar la mediana de la muestra total que representa el 50% de clientes que han demandados productos dentro del intervalo de 1 a 14, se puede observar que la mayoría de cliente de las distintas tipología han demandado productos por debajo a este valor, excepto Tienda que refleja una concentración del 50% igual a la mediana total que fue de 14.

Las figuras a continuación evidencian que la tipología de cliente no es una variable característica suficiente para dividir a los grupos de clientes de la muestra total con características de compras similares.

La Figura 4.9 que representa la dispersión entre las variables monto de compra y la cantidad de productos demandados de los 769 clientes considerados en la muestra, permiten observar que existe una alta dispersión de los datos dentro de cada tipología y que a su vez también muestran grados de asociaciones similares entre estas dos variables. Por ejemplo, el cliente mayorista 4394 presenta una compra de 1’610.411 dólares como consecuencia de la demanda de 261 productos. El cliente tendero 4466 presenta una compra de 297.532 dólares como resultado de la demanda de 69 productos, que es muy parecida a las compras del cliente mayorista 4357 que realizó una compra de 267.012 correspondiente de la demanda de 84 productos y el otro cliente más cercano es el cliente mayorista 4430 con una compra de 254.172 dólares a causa de la demanda de 81 productos.

Figura 4. 9

Monto de Compra (dólares por cantidad de productos demanda) según la tipología de cliente, 2020-2021



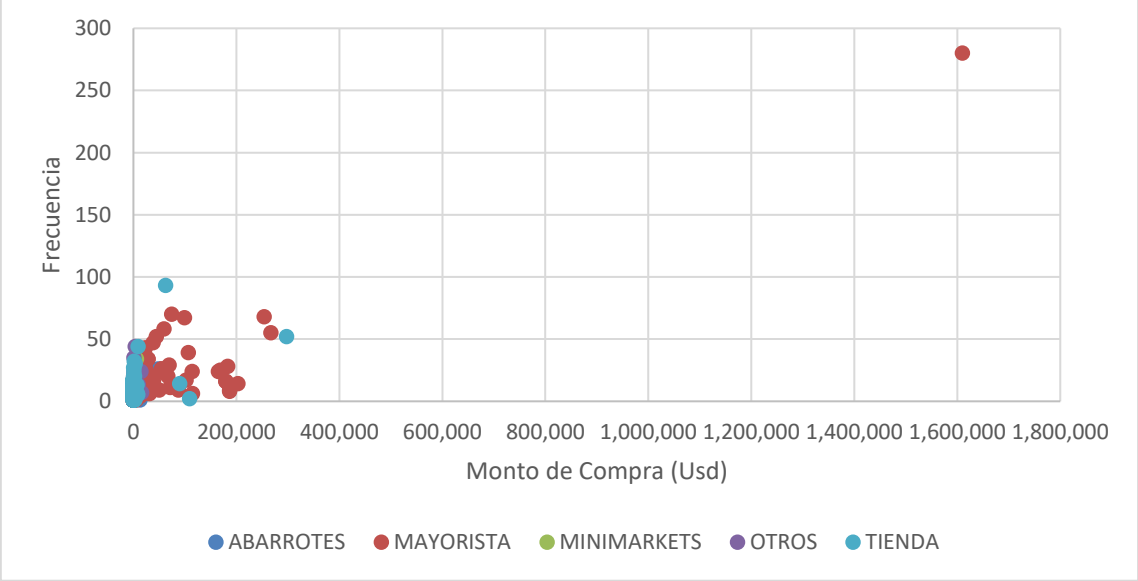
Fuente: Empresa AAA, Información de Ventas 2017-2021

La Figura 4.10 que representa la dispersión entre las variables monto de compra y la frecuencia de compra de los 769 clientes considerados en la muestra, permiten observar que existe una alta dispersión de los datos dentro de cada tipología y que a su vez también muestran grados de asociaciones similares entre estas dos variables. Por ejemplo, el cliente mayorista 4394 presenta una compra de 1’610.411 dólares como consecuencia de la demanda de 280 transacciones realizadas. El cliente tendero 4466 presenta una compra de 297.532 dólares como resultado de 52 transacciones hechas, que es muy parecida a las compras del cliente mayorista 4357 que realizó una compra de 267.012 por el motivo de 55 transacciones ejecutadas y el otro cliente más cercano es el cliente mayorista 4430 con una compra de 254.172 dólares a causa de 68 transacciones efectuadas.

Estas variables son de mucha utilidad para la segmentación de clientes ya que nos permitirán poder discriminar a los 769 clientes en grupos homogéneos con alguna característica particular de compra. En la siguiente sección se procederá a utilizar técnicas de análisis estadístico multivariante para poder explorar de una manera más técnica, la existencia estos grupos homogéneos.

Figura 4. 10

Monto de Compra (dólares por frecuencia de compra) según la tipología de cliente,2020-2021



Fuente: Empresa AAA, Información de Ventas 2017-2021

4.3. Minería de datos

4.3.1. Segmentación de los clientes mediante el método de Agrupación de particiones

Primera Clusterización

En esta sección se utilizará el método Agrupación de particiones (K-means, K-medoids) para segmentar la muestra de 769 clientes en función a dos variables de interés descritas anteriormente Monto de Compra y Cantidad de productos demandados que fueron calculadas en el periodo comprendido entre los años 2020-2021.

Como primer análisis mediante las técnicas de partición K-means y K-medoids se procederá a elegir el número óptimo de clústers haciendo uso de la técnica método del codo.

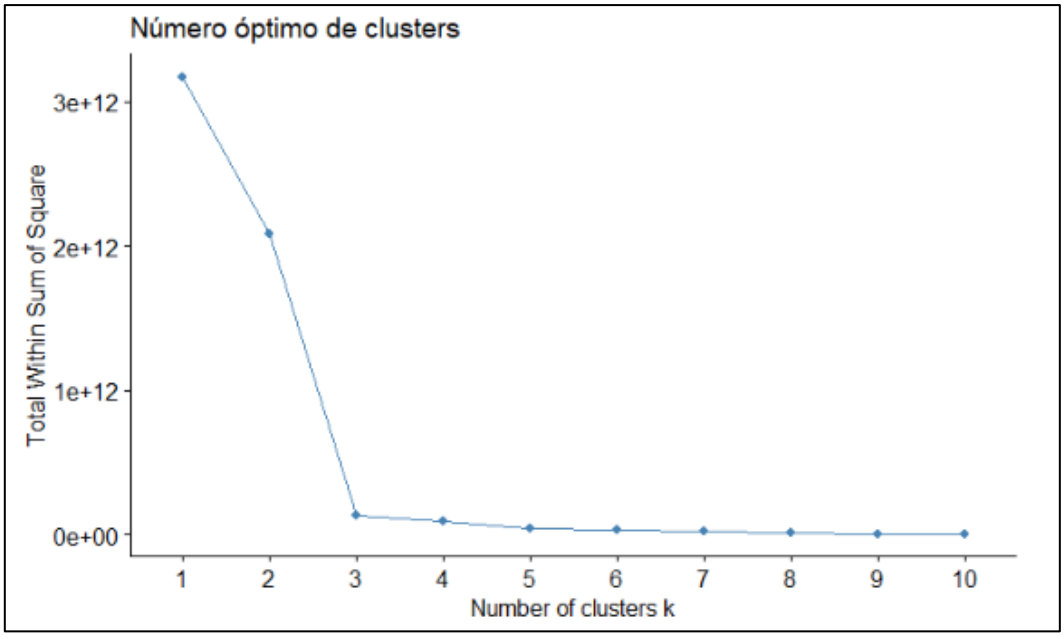
Elección del número óptimo de clusters

Método del codo

La figura 4.11 representa la evolución de la suma de cuadrado dentro del cluster utilizando el método de partición K-means, muestra que a partir de 4 clústers la reducción en la suma total de cuadrados internos parece estabilizarse, indicando que K = 4 es una buena opción. (Ver Anexo 18)

Figura 4. 11

Evolución de la suma de cuadrados dentro del cluster, usando el método K-means

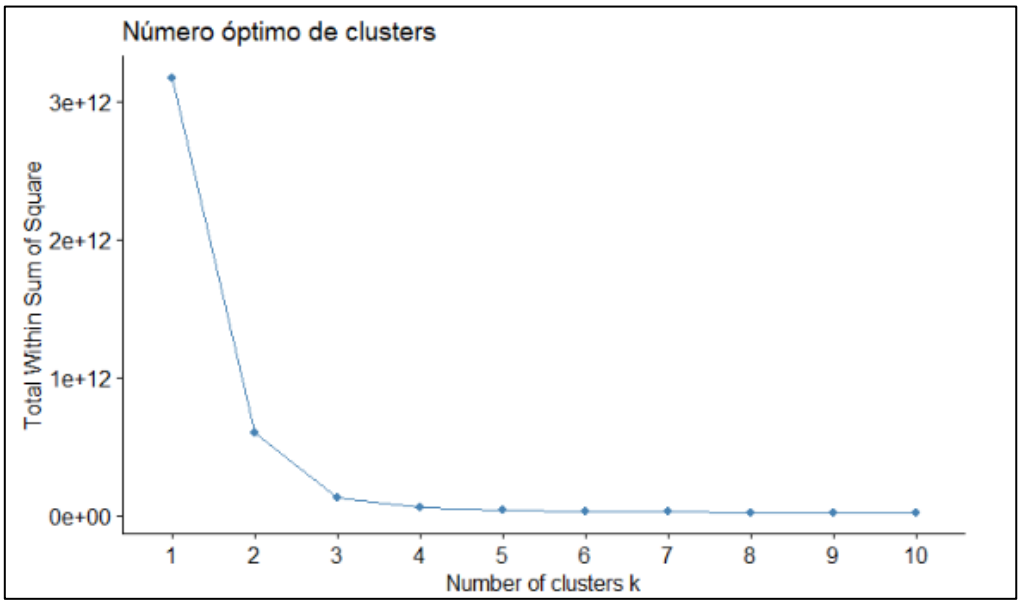


La Figura 4.12, también evidencia que a partir de 4 clusters la reducción en la suma total de los cuadrados internos utilizando el método de partición K-medoids presenta una pequeña reducción, por lo que se confirma que K=4 es una buena opción para discriminar el total de clientes que fueron considerados como muestra en el estudio. (Ver Anexo 18)

Una vez determinado el número de cluster óptimo que fue de K=4, es necesario realizar una validación interna para evaluar la calidad de los clusters generados.

Figura 4. 12

Evolución de la suma de cuadrados dentro del cluster, usando el método K-medoids



Validación interna de los clusters:

Medidas de Homogeneidad-Separación, silueta y Dunn

Tabla 4. 3

Resultados de los índices para la validación interna de los K=4 clusters

Métodos de Agrupación	Medidas de distancia	Cluster	n	Media silueta por clusters	Anchura media de la silueta	Dunn	Medidas de Homogeneidad y separación	
							X dentro.	X entre.
K-means	Euclidiana	1	11	0,644	0,9222	0,0267	5.192	160.283
		2	22	0,497				
		3	1	0				
		4	735	0,94				
K-medoids	Euclidiana	1	1	0	0,8201	0,0009	4.803	64.086
		2	17	0,475				
		3	91	0,035				
		4	660	0,938				
	Manhattan	1	1	0	0,82	0,001	4.817	64.116
		2	17	0,475				
		3	91	0,034				
		4	660	0,938				

Nota: n=Tamaño, X=Promedio

Se puede observar en la Tabla 4.3, que el modelo de partición K-means con distancia euclidiana es el que presenta una mayor bondad de ajuste interna de las 4 agrupaciones consideradas previamente.

El modelo K-means con distancia euclidiana tiene un mejor performance en 3 de las 4 medidas con respecto a los modelos K-medoids. La primer medida evaluada es la anchura media de la silueta que para el modelo K-means fue de 0.9222 y que al tener un valor más cercano a uno comparado con los valores de los modelos K-medoids con distancia euclidiana y manhattan que fueron de 0.8201 y 0.8200 respectivamente, demuestra que calidad de la estructura de clusters en su conjunto a partir del promedio de todos los índices silhoutte es excelente. La segunda medida evaluada es el índice de Dunn con un valor de 0.0267 para el modelo K-means, pues al tener un valor más alto versus los modelos K-medoids con distancia euclidiana y manhattan que fueron de 0.0009 y 0.010 respectivamente, determina que la estructura de sus clusters tiende a ser más compacta y separada. La tercer medida revisada es

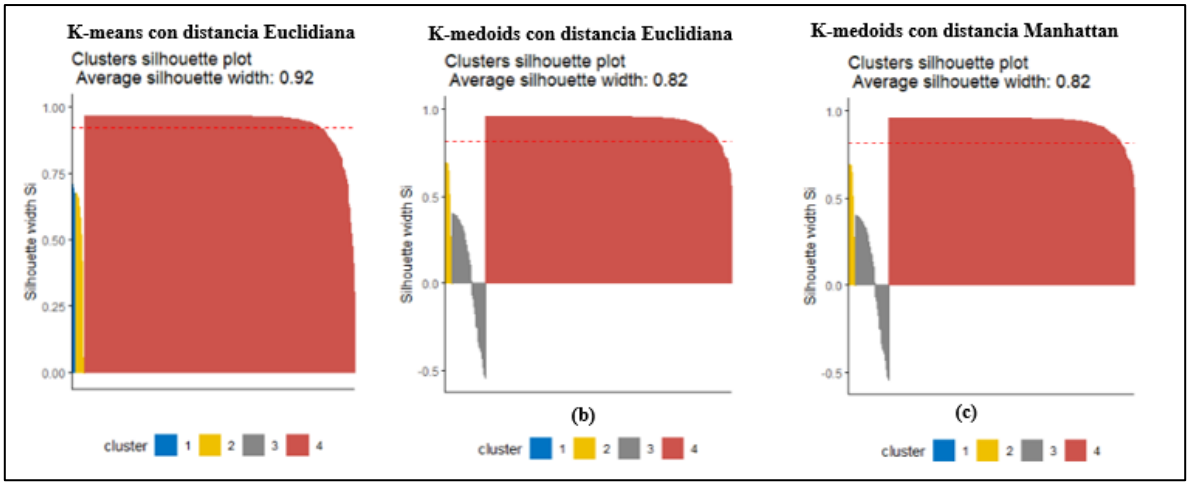
la de separación que viene representado por el valor promedio entre todos los clusters, y que de acuerdo a la Tabla 4.3, se puede observar que el valor del modelo K-means que fue de 160.283 es más alto comparado con los valores del modelo K-means con distancia euclidiana y manhattan que fueron de 64.086 y 64.116 respectivamente, corroborando que la distancia de los elementos que forman parte de un clusters comparado con los otros elementos de otros clusters tiende a ser más separados que los otros modelos y por tanto mejor.

El modelo K-means solamente tiende a ser superado por los modelos K-medoids con distancia euclidiana y manhattan en lo que respecta a la medida de homogeneidad que viene representada por el valor promedio dentro de todos los clusters, el cual debe ser el más pequeño posible y que para los modelos K-medoids con distancia euclidiana y manhattan fueron de 4.803 y 4.817 respectivamente comparado con el valor de K-means que fue de 5.192.

Otra forma de respaldar que el método de K-means con distancia euclidiana es la mejor opción comparado con los otros dos métodos descritos en la Tabla 4.3, es mediante la comparación de las siguientes gráficas presentadas en la Figura 4.13: Como se puede observar la Figura 4.13 (a) presenta un coeficiente de silueta promedio de anchura de 0.92 para el método de K-means con distancia euclidiana. El cual evidencia una alta calidad de la estructura de cluster en su conjunto, pero además constata que la calidad de asignación de cada observación por separado dentro de cada cluster también es buena (Ver el índice Media silueta por cluster en la Tabla 4.3). Este último detalle es el que mueve la balanza en favor del modelo K-means con distancia euclidiana, pues la Figura 4.13 (b) y (c) que representan a los modelos K-medoids con distancia euclidiana y manhattan respectivamente muestran una alta calidad de la estructura de cluster en su conjunto, no así por separado pues como se puede observar en ambas figuras no hay una buena asignación de cada observación en el cluster 3. (Ver el índice Media silueta por cluster en la Tabla 4.3)

Figura 4. 13

Gráfico de silueta promedio por grupos de clusters



A continuación, en la Figura 4.14, se presenta el mejor modelo que resultó luego de la validación interna realizada, el cual fue el modelo K-means con distancia euclidiana.

El modelo K-means con distancia euclidiana fue el modelo que obtuvo una mayor calidad en la asignación de cada observación dentro de cada clusters siendo agrupados de la siguiente manera como lo muestra la Figura 4.15: El cluster 1 está formado por 11 observaciones que representan el 1,4% de la muestra total de 765 clientes, dentro de los cuales se encuentran los clientes 5152 y 5563 (Ver Figura 4.14), el cluster 2 por 22 que representa el 2,9% dentro del cual se encuentra el cliente 6103(Ver Figura 4.14), el cluster 4 tiene un tamaño de 735 observaciones que representa el 95,6% dentro del cual se encuentran los clientes 4430,4357 y 4466 (Ver Figura 4.14). Y el cluster 3 está formado por una sola observación, el cual es el cliente 4394 (Ver Figura 4.14) que es representativo para la muestra de estudio y representa el 0.1% de la muestra.

Figura 4. 14

Modelo K-means usando la distancia euclidiana

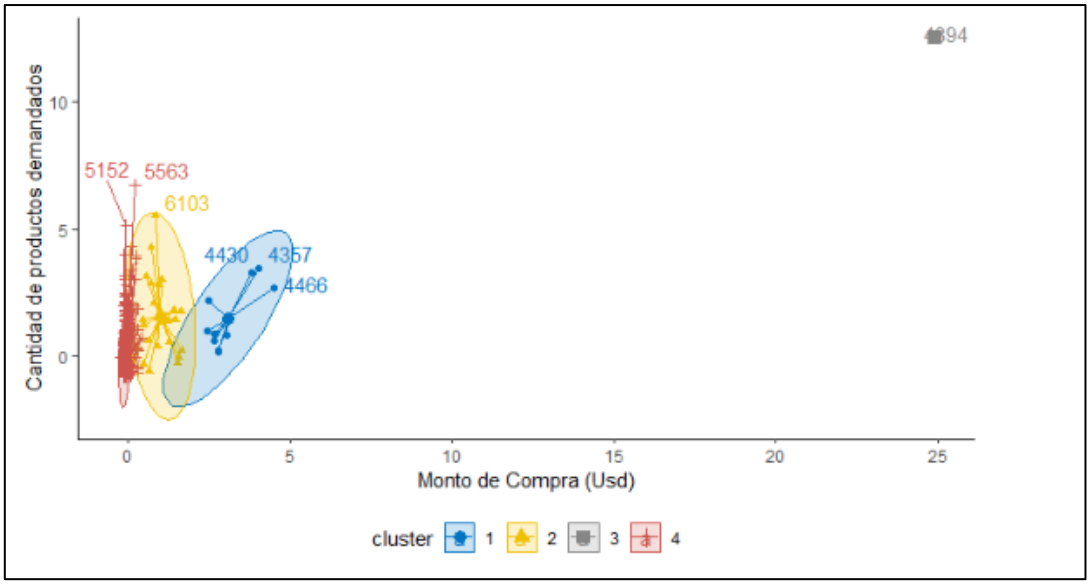
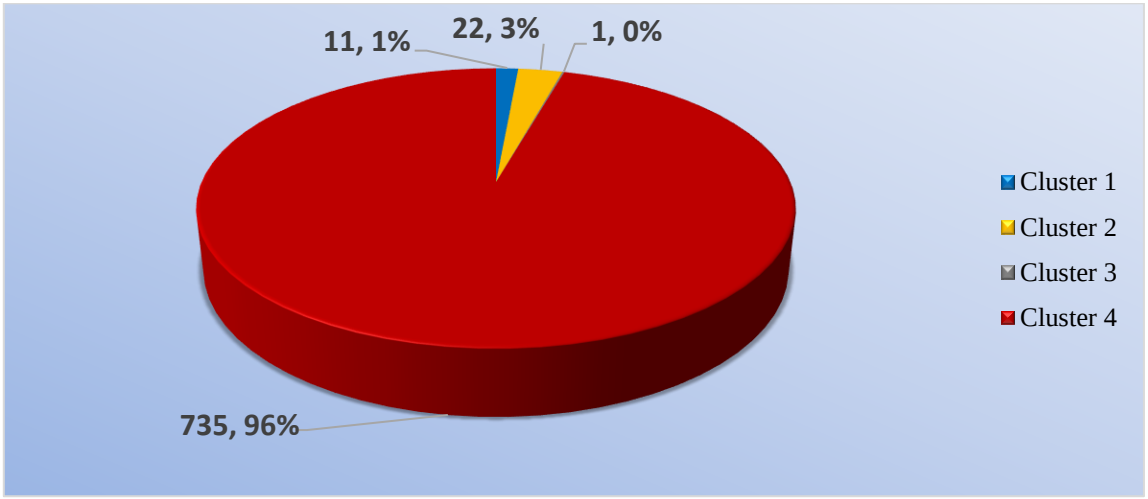


Figura 4. 15

Distribución porcentual de los 4 clusters producto de la primera clusterización



Segunda Clusterización

En esta sección se volverá hacer uso del método Agrupación de particiones (K-means, K-medoids), debido a que el tamaño de clientes agrupados en el cluster 4 tiene un valor alto que es de 735 (Ver Figura 4.14), por lo cual se procederá a realizar una segunda clusterización para ese grupo, con el fin de encontrar grupos de clientes con volúmenes de compra y cantidad de productos demandados más similares posibles.

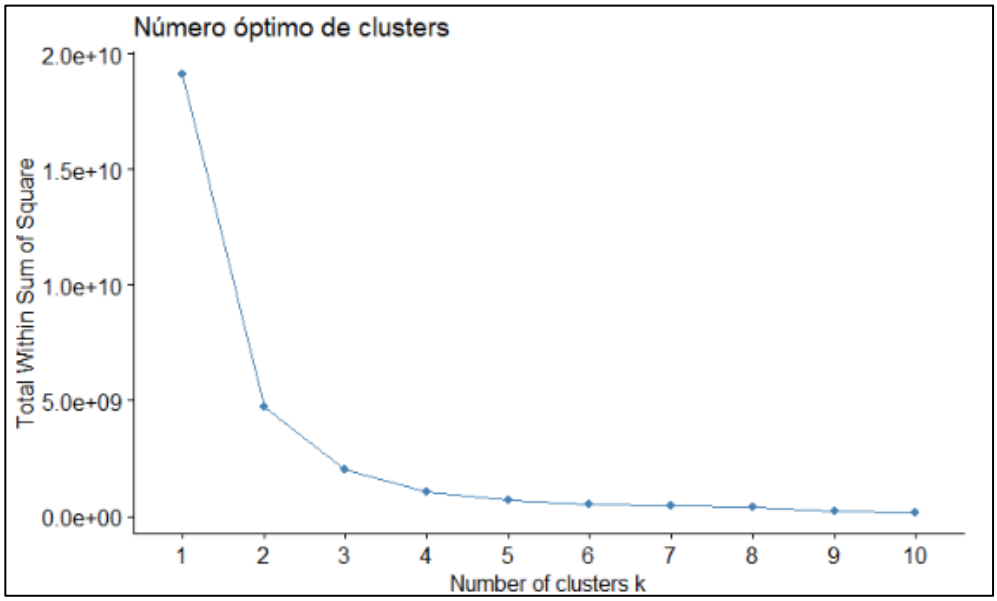
Elección del número óptimo de clusters

Método del codo

La curva de la Figura 4.16 indica que a partir de 4 clusters la suma de los cuadrados dentro de los clusters desciende de forma poco sustancial, por lo que K = 4 es una buena opción. (Ver Anexo 19)

Figura 4. 16

Evolución de la suma de cuadrados dentro de los clusters, usando el método K-means

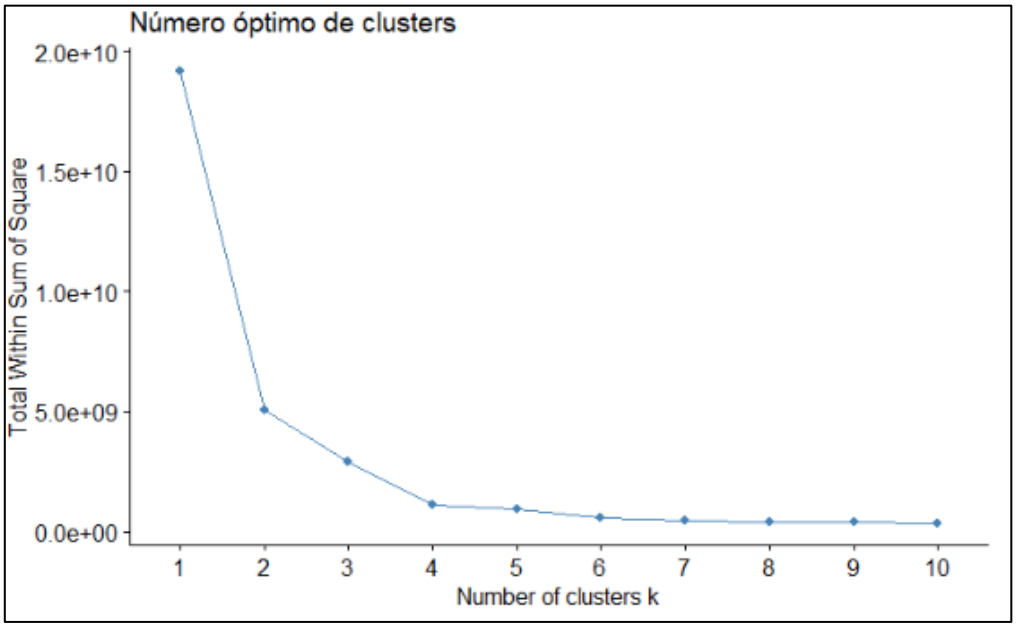


La Figura 4.17, indica también que 4 clusters es una buena opción para poder dividir al conjunto de clientes agrupados en el cluster 4, pues a partir de ese número la curva tiene una leve reducción en la suma de cuadrados dentro del cluster. (Ver Anexo 19)

Una vez determinado el número de cluster óptimo que fue de K=4, se procede a realizar una validación interna para evaluar la calidad de los clusters generados de igual forma que se lo realizó en la primera clusterización.

Figura 4. 17

Evolución de la suma de cuadrados dentro de los clusters usando el método K-medoids



Validación interna de los clusters: Medidas de Homogeneidad-Separación, Silueta y Dunn

Tabla 4. 4

Resultados de los índices para la validación interna de los K=4 clusters

Métodos de Agrupación	Medidas de distancia	Cluster	n	Media silueta por cluters	Anchura media de la silueta	Dunn	Medidas de Homogeneidad y Separación	
							X dentro.	X entre.
K-means	Euclidiana	1	47	0,529	0,7801	0,004	992	9.600
		2	85	0,471				
		3	20	0,634				
		4	583	0,85				
K-medoids	Euclidiana	1	20	0,661	0,7573	0,0026	940	8.362
		2	58	0,487				
		3	111	0,413				
		4	546	0,86				
	Manhattan	1	20	0,659	0,755	0,0036	953	8.381
		2	58	0,486				
		3	111	0,41				
		4	546	0,857				

Nota: n=Tamaño, X=Promedio

Se puede observar en la Tabla 4.4, que el modelo de partición K-means con distancia euclidiana nuevamente es el que presenta una mayor bondad de ajuste en los clusters internos de forma general.

El modelo K-means con distancia euclidiana tiene un mejor performance en 3 de las 4 medidas con respecto al modelo K-medoids con distancia euclidiana, las cuales son Anchura media de la silueta, el índice Dunn y la medida promedio entre todos los clusters con valores de 07801, 0.00040 y 9.600 respectivamente por sobre los valores del modelo K-medoids con distancia euclidiana que fueron de 0.7573, 0.0026 y 8.362.Sin embargo, este último presentó un mejor performance en la medida promedio dentro de todos los cluster ya que su valor fue de 940 comparado con el modelo de K-means con distancia euclidiana que fue de 992.

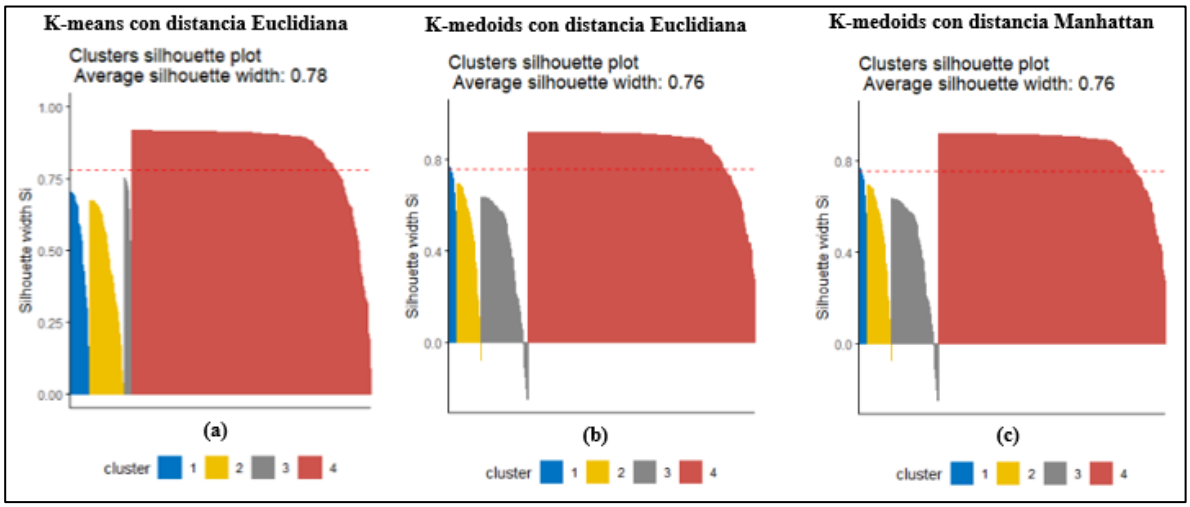
Análogamente si se compara el modelo K-means con distancia euclidiana con respecto al modelo K-medoids con distancia de manhattan, el primero supera al segundo en tres de las 4 medidas (Ver Tabla 4.4), las cuales son Anchura media de la silueta, el índice Dunn y la medida promedio entre todos los clusters. Solamente siendo superado por el modelo K-medoids con distancia de manhattan en la medida promedio dentro de todos los clusters que fue de 953, comparado con el modelo K-mean con distancia euclidiana que fue de 992.

El modelo de K-means con distancia euclidiana al igual que los otros métodos descritos en la Tabla 4.4 presentan una calidad parecida en lo que respecta a la estructura de cluster en conjunto tal como lo muestra la Figura 4.18.

La calidad de asignación de cada observación por separado dentro de cada cluster termina desequilibrando la balanza en favor del modelo de K-means con distancia euclidiana (Ver el índice Media silueta por cluster en la Tabla 4.4) con respecto a los modelos de K-medoids con distancia euclidiana y manhattan, pues estos últimos presentan una mala asignación de las observaciones en los clusters 2 y 3, tal como lo muestra la Figura 4.18 (b) y (c).

Figura 4. 18

Silueta promedio por grupos de clusters



A continuación, en la Figura 4.19 se presenta el mejor modelo que resultó luego de la validación interna realizada, el cual fue el modelo K-means con distancia euclidiana.

El modelo K-means con distancia euclidiana fue el modelo que obtuvo una mayor calidad en la asignación de cada observación dentro de cada cluster siendo agrupados de la siguiente manera como lo muestra la Figura 4.20: El cluster 1 está formado por 47 observaciones que representan el 6,4% de la muestra de 735 clientes, dentro del cual se encuentran por ejemplo los clientes 108,3044, y 6565 (Ver Figura 4.19), el cluster 2 por 85 que representan el 11,6% dentro del cual se encuentran los cliente 3416,3644 y 6218 entre otros(Ver Figura 4.19), el cluster 3 tiene un tamaño de 20 observaciones que representan el 2,7% dentro del cual se encuentran los clientes 6735,5065 y 7201 (Ver Figura 4.19). Y por último el cluster 4 está formado por 583, entre los cuales figuran los clientes 2493,5410, 4158, entre otros (Ver Figura 4.19) y representan el 79,3% de la muestra considerada en la segunda clusterización que fue de 735.

Figura 4. 19

Gráfico del modelo K-means usando la distancia euclidiana

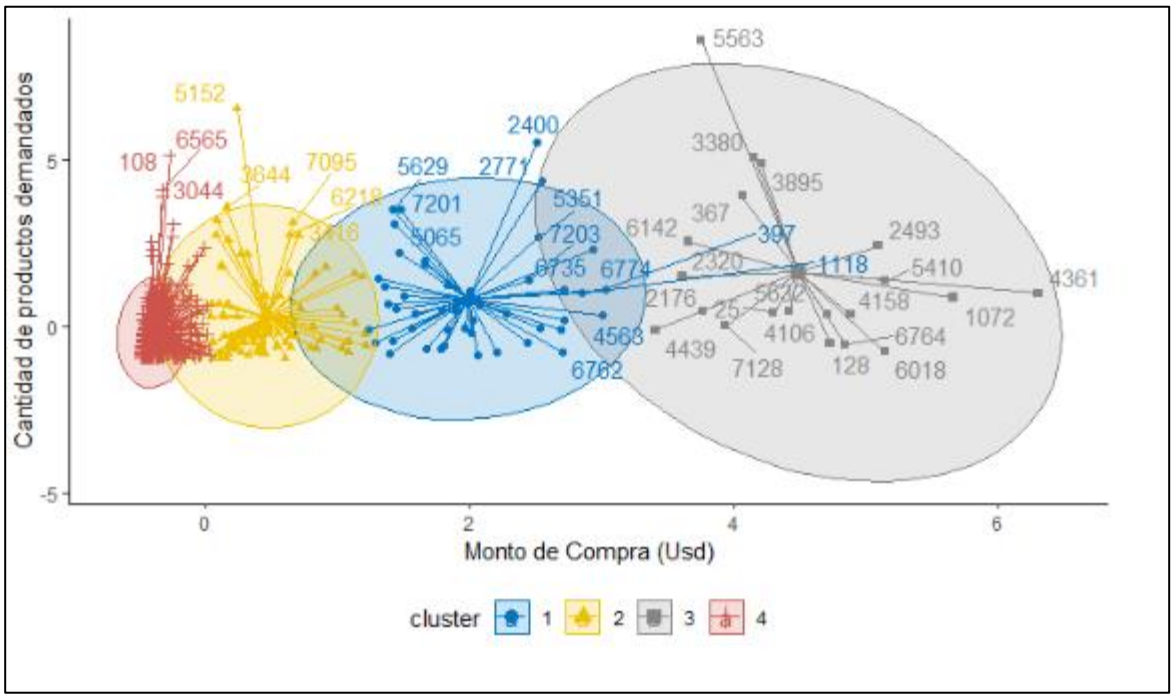
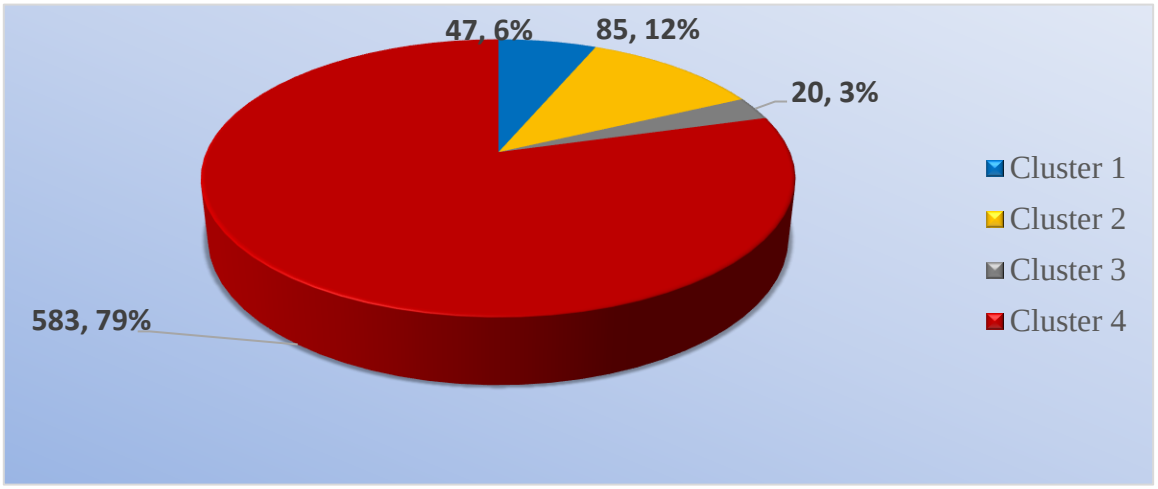


Figura 4. 20

Distribución porcentual de los 4 clusters producto de la segunda clusterización



Una vez realizada los dos procesos de clusterización, se obtuvo un total de 7 segmentaciones, los cuales se describen a continuación en la siguiente Tabla 4.5:

Tabla 4. 5

Segmentos de clientes encontrados en los dos procesos de clusterización, 2020-2021

Clusters	Tamaño	Monto de Compra	Media de Monto de Compra	Media Cantidad de productos demandados	% Participación Monto Compra	Segmentación
1	11	2.272.463	206.588	46	31,16%	A
2	22	1.599.356	72.698	47	21,93%	C
3	1	1.610.411	1.610.411	261	22,08%	B
4.1	47	593.054	12.618	29	8,13%	D
4.2	85	412.560	4.854	20	5,66%	F
4.3	20	507.116	25.356	41	6,95%	E
4.4	583	298.316	512	13	4,09%	G
Total	769	7.293.277	9.484	17	100,00%	

De acuerdo con la Figura 4.21, los segmentos de clientes están divididos de la siguiente manera:

El segmento de clientes A está formado por 11 clientes que pesan el 1,43% de la muestra de estudio, tienen un promedio de compra de 206.588 dólares como resultado de la demanda de 46 productos en promedio y tienen un monto de compra de 2’272.463 que representa el 31,2% del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

El segmento de clientes B está formado por 1 cliente que pesa el 0.13% de la muestra de 769 clientes, y tiene un monto de compra de 1’610.411 como resultado de la demanda de 261 productos que representa el 22,08% del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

El segmento de clientes C está formado por 22 clientes que pesan el 2,86% de la muestra de estudio, tienen un promedio de compra de 72.698 dólares como resultado de la demanda de 47 productos en promedio y tienen una participación del 21,93% que fue de 1’599.356 del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

El segmento de clientes D está formado por 47 clientes que pesan el 6,11% de la muestra de estudio, tienen un promedio de compra de 12.618 dólares como resultado de la demanda de 29 productos en promedio y tienen una participación del 8,13% que fue de 593.054 del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

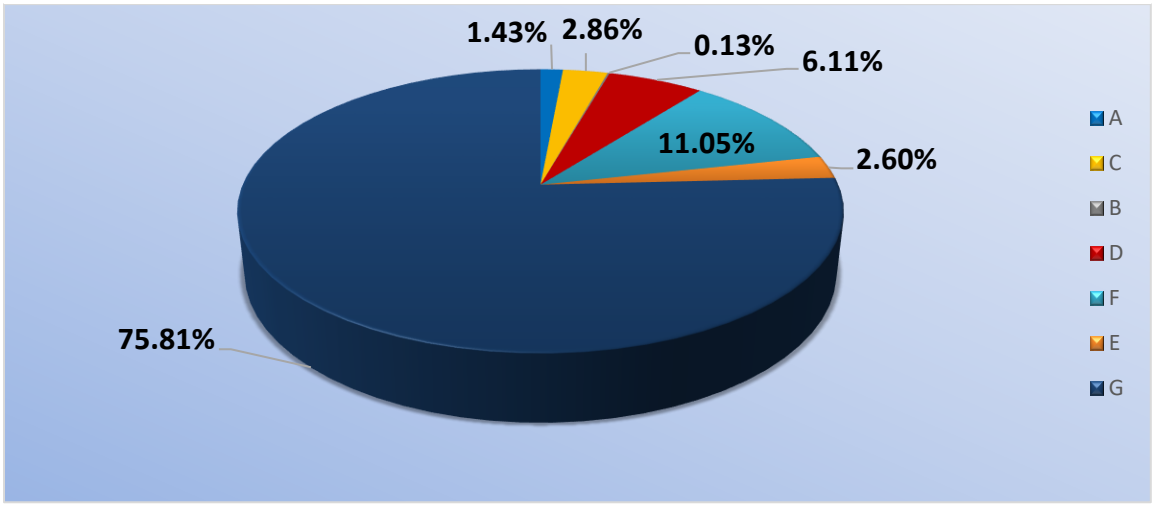
El segmento de clientes E está formado por 20 clientes que pesan el 2,60% de la muestra de estudio, tienen un promedio de compra de 25.356 dólares como resultado de la demanda de 41 productos en promedio y tienen una participación del 6,95% que fue de 507.116 del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

El segmento de clientes F está formado por 85 clientes que pesan el 11,05% de la muestra de estudio, tienen un promedio de compra de 4.854 dólares como resultado de la demanda de 20 productos en promedio y tienen una participación del 5,66% que fue de 412.560 del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

El segmento de clientes G está formado por 583 clientes que pesan el 75,81% de la muestra de estudio, tienen un promedio de compra de 512 dólares como resultado de la demanda de 13 productos en promedio y tienen una participación del 4,09% que fue de 298.316 del monto total de compra comprendido entre los años 2020-2021 que fue de 7’293.277.

Figura 4. 21

Distribución porcentual de los clientes para las 7 Segmentaciones



4.3.2. Caracterización de los clientes mediante el método CHAID

Antes de proceder a presentar los resultados correspondiente a la caracterización de los clientes, es necesario renombrar las variables definidas en la Tabla 3.6-Tipología de la Caracterización de los clientes que se encuentra en la sección 3.4 dentro del Marco Metodológico con el fin de realizar una mejor interpretación de los resultados obtenidos.

Tabla 4. 6

Nomenclatura de las variables elegidas para la caracterización de los clientes

Nomenclatura	Variable
Clusters	Segmentación
SegmComp	Segmento de Compra
TipClte	Tipología de cliente
Zona	Zona
NatClte	Naturaleza del cliente
MontComp	Montos de Compra
Frecuencia	Frecuencia
Recencia	Recencia
Portaf	Portafolio
TktProm	Ticket Promedio
Antigüedad	Antigüedad
FrecAnualUltAnoComp	Frecuencia Anual del último año de compra.
FrecPromMesUltAnoComp	Frecuencia Promedio Mensual del último año de compra.

El objetivo de esta sección es realizar la caracterización de los 7 grupos de segmentaciones de clientes determinados en la sección anterior a través de árboles de decisión usando el método CHAID, para de esta manera encontrar características particulares adicionales que permitan describir de una mejor forma a los grupos encontrados.

También es importante mencionar que el cluster B, debido a que está compuesto por un solo cliente representativo solo se procederá a realizar una interpretación de sus características más relevantes y no será sometido a las siguientes pruebas de significancia:

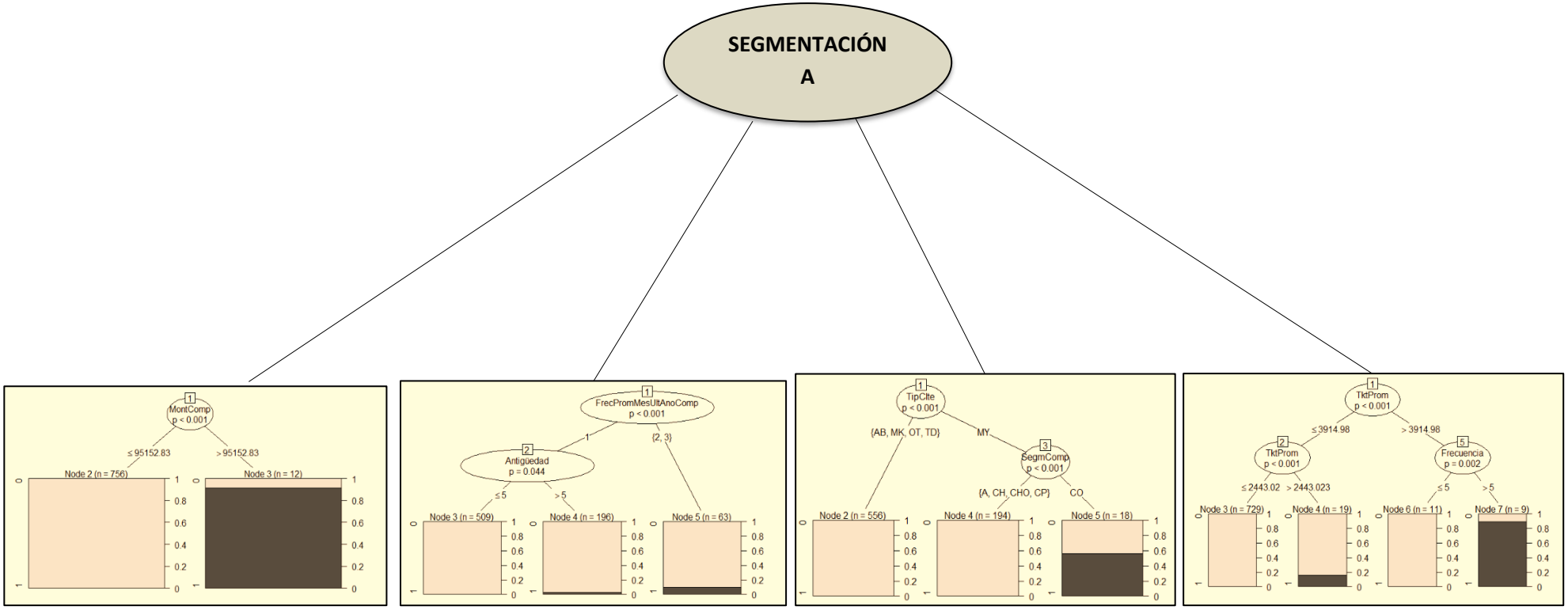
Significancia estadística de cada variable independiente en función del tipo de variable dependiente

En esta sección se procede a realizar la prueba de significancia entre las variables independientes y los clusters: A, C, D, E, F, G donde se construye una variable cualitativa dicotómica considerando al cluster A como 1 y al Resto de clusters como 0; proceso que se repite para los demás clusters para determinar que variables son significativas con respecto a cada cluster.

Para el conjunto de variables cualitativas **S**egmComp, TipClte, Zona, NatClte se realiza la prueba Chi-cuadrado de Independencia con la hipótesis nula: H_0: la variable independiente y cluster son independientes. Y para el conjunto de variables cuantitativas MontComp, Frecuencia, Recencia, Portaf, TktProm, Antigüedad, FrecAnualUltAnoComp, FrecPromMesUltAnoComp se realiza la prueba ANOVA con hipótesis nula: H_0: $\text{class}_1 = \text{class}_n$. (Ver Anexos 20 al 25)

Figura 4. 22

Árboles de decisión del segmento de clientes A



(a) **(b)** **(c)** **(d)**

Figura 4. 23

Caracterización del segmento de clientes A

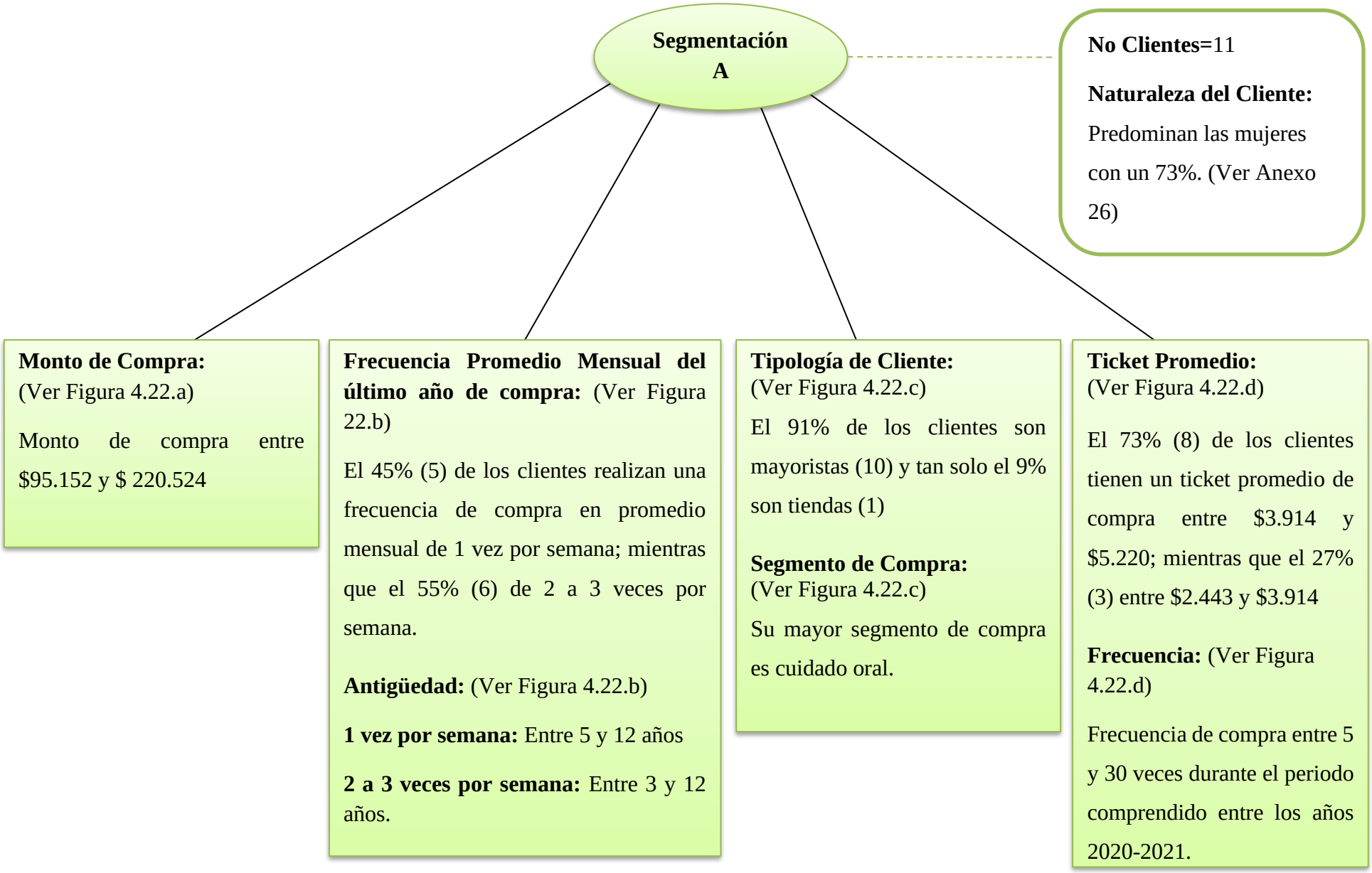


Figura 4. 24

Caracterización del segmento de clientes B

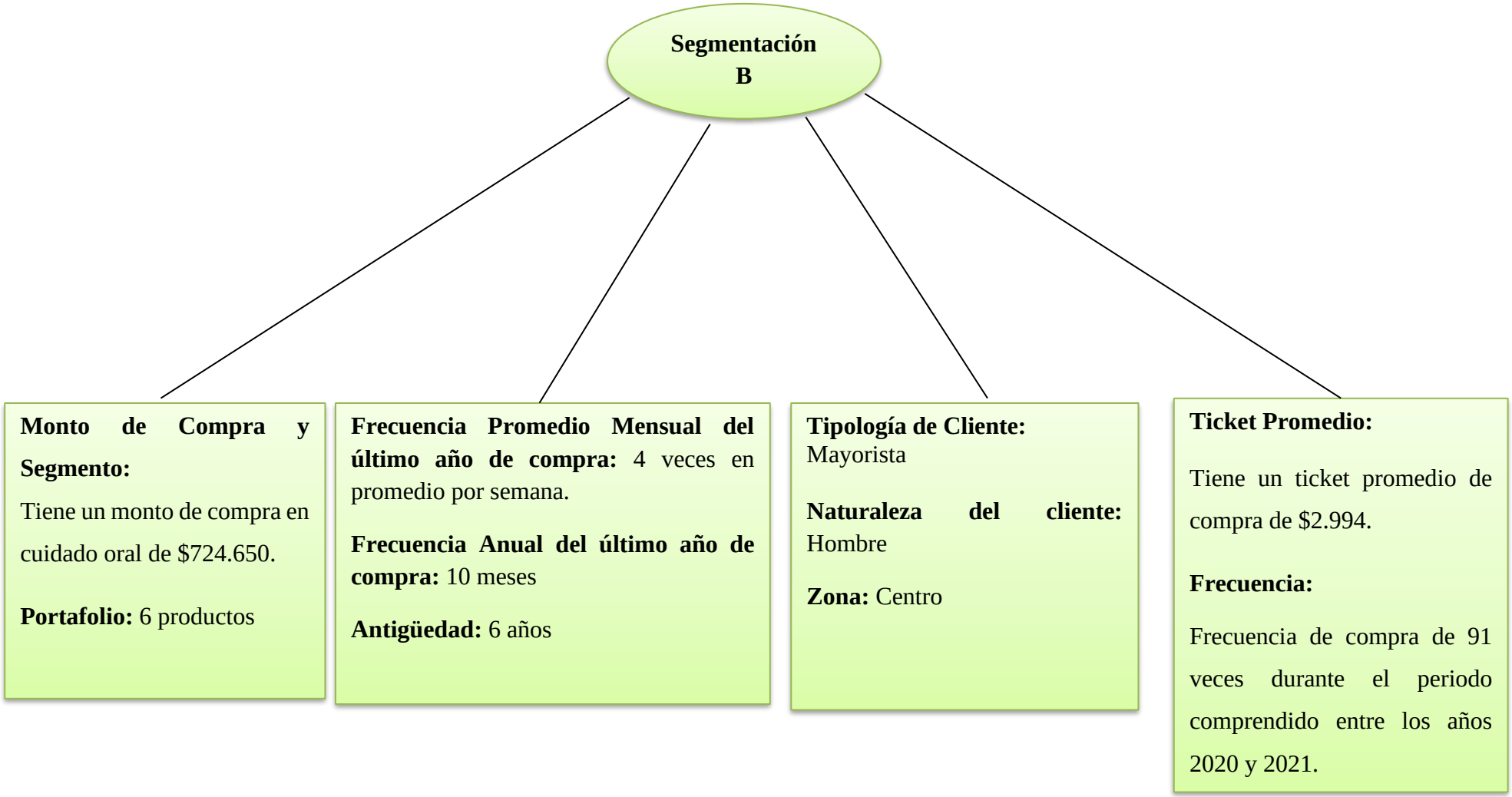
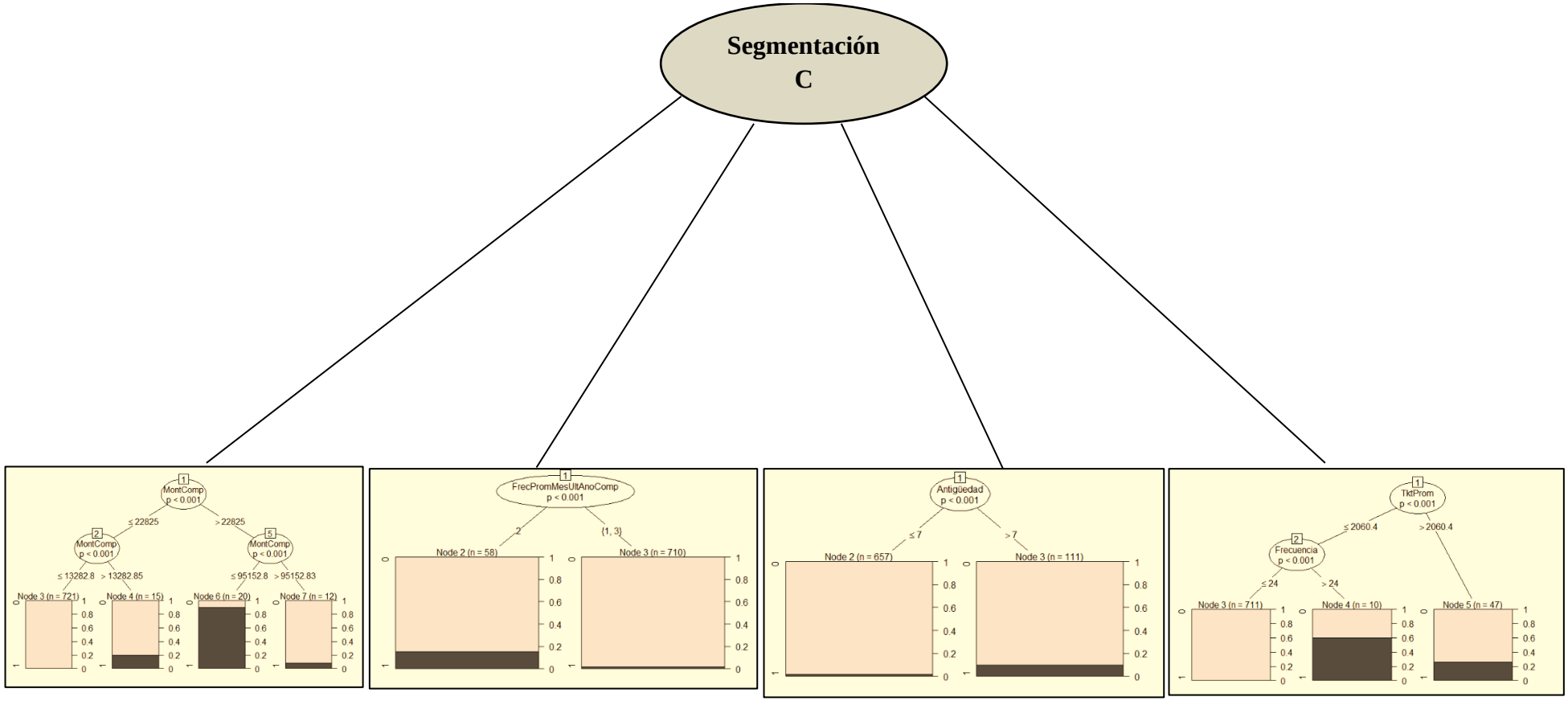


Figura 4. 25

Árboles de decisión del segmento de clientes C



(a) **(b)** **(c)** **(d)**

Figura 4. 26

Caracterización del segmento de clientes C

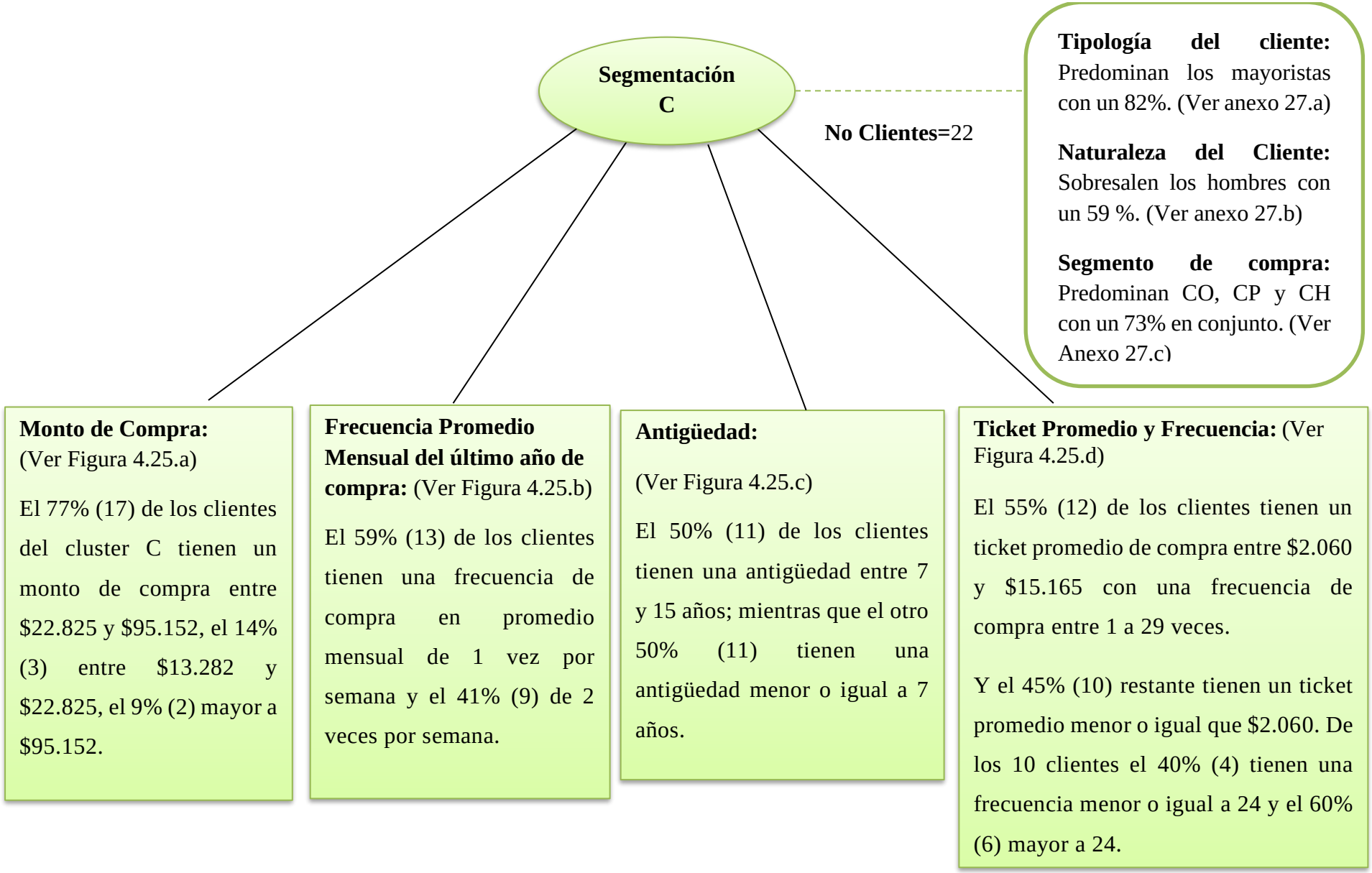


Figura 4. 27

Árboles de decisión del segmento de clientes D

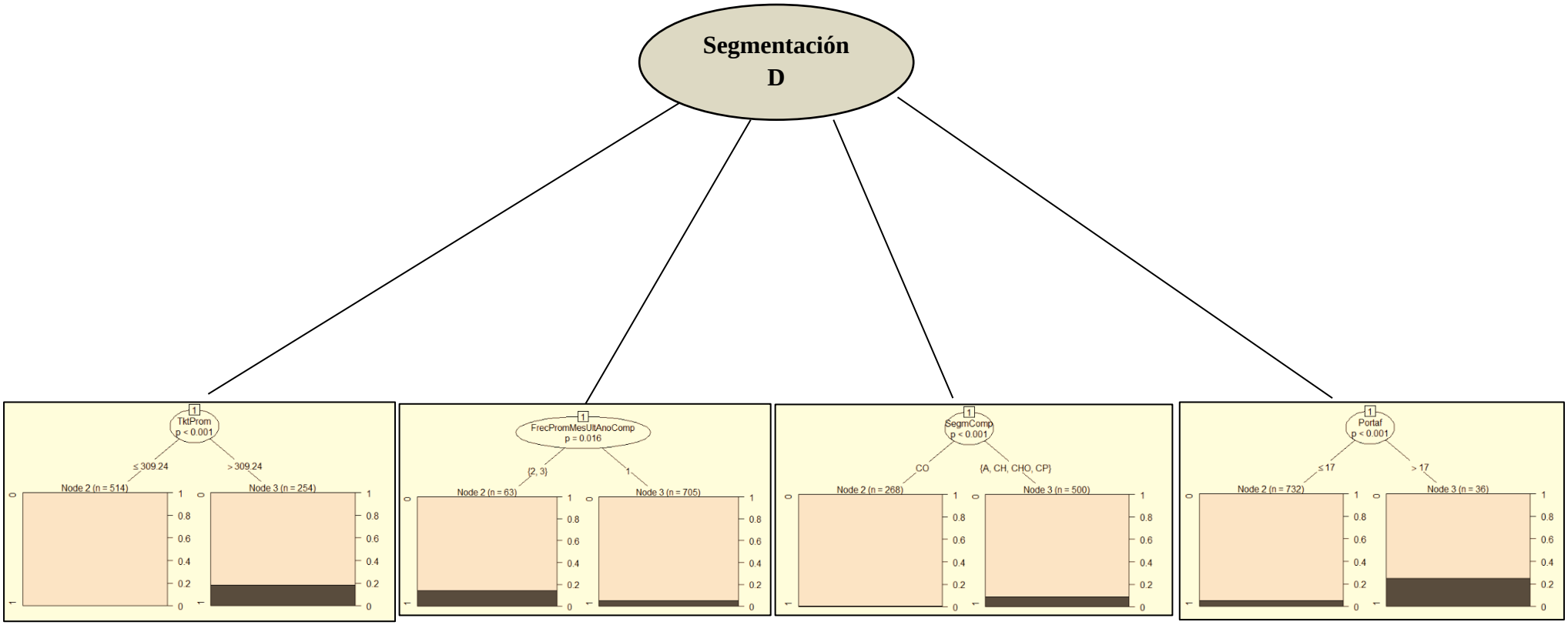


Figura 4. 28

Caracterización del segmento de clientes D

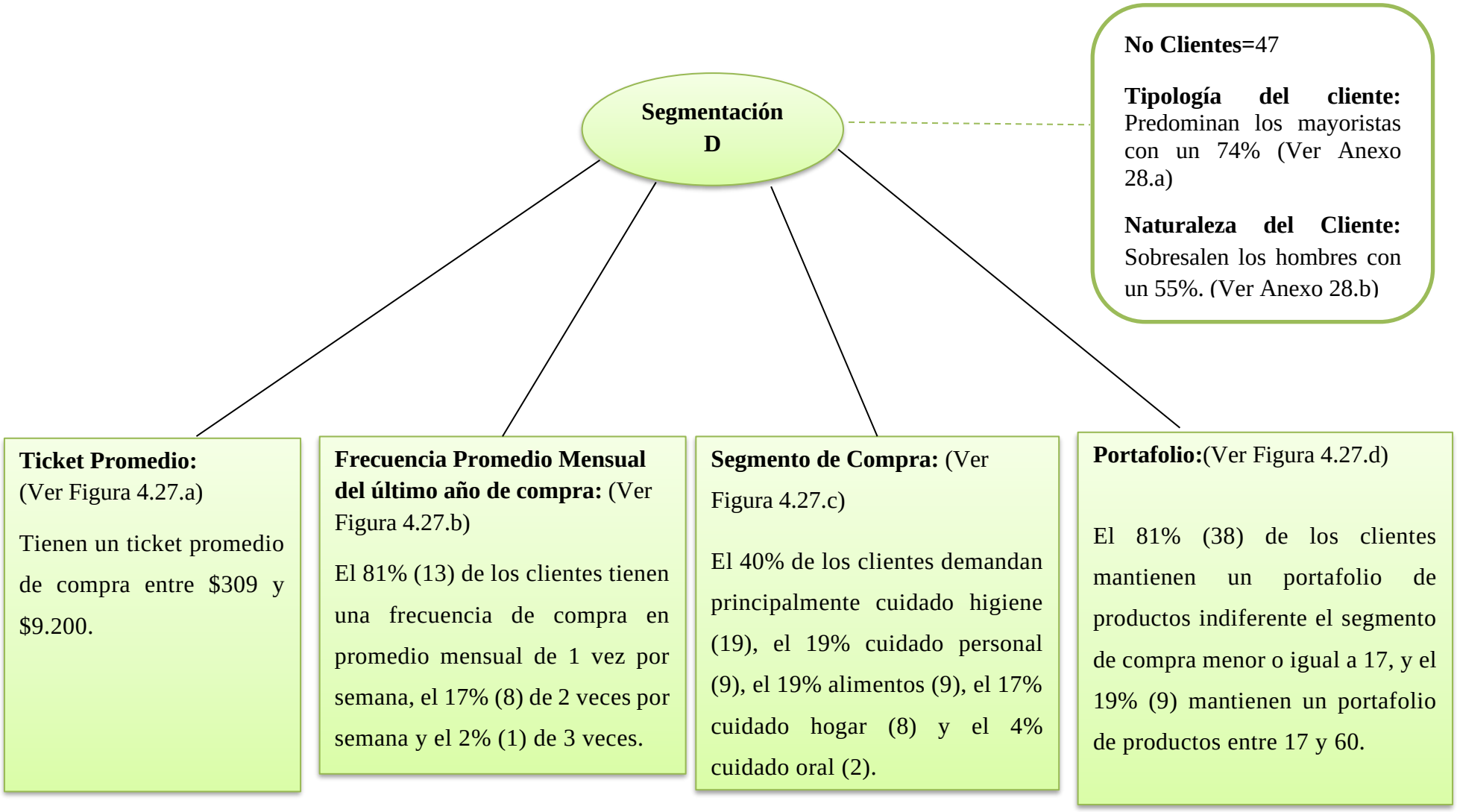


Figura 4. 29

Árboles de decisión del segmento de clientes E

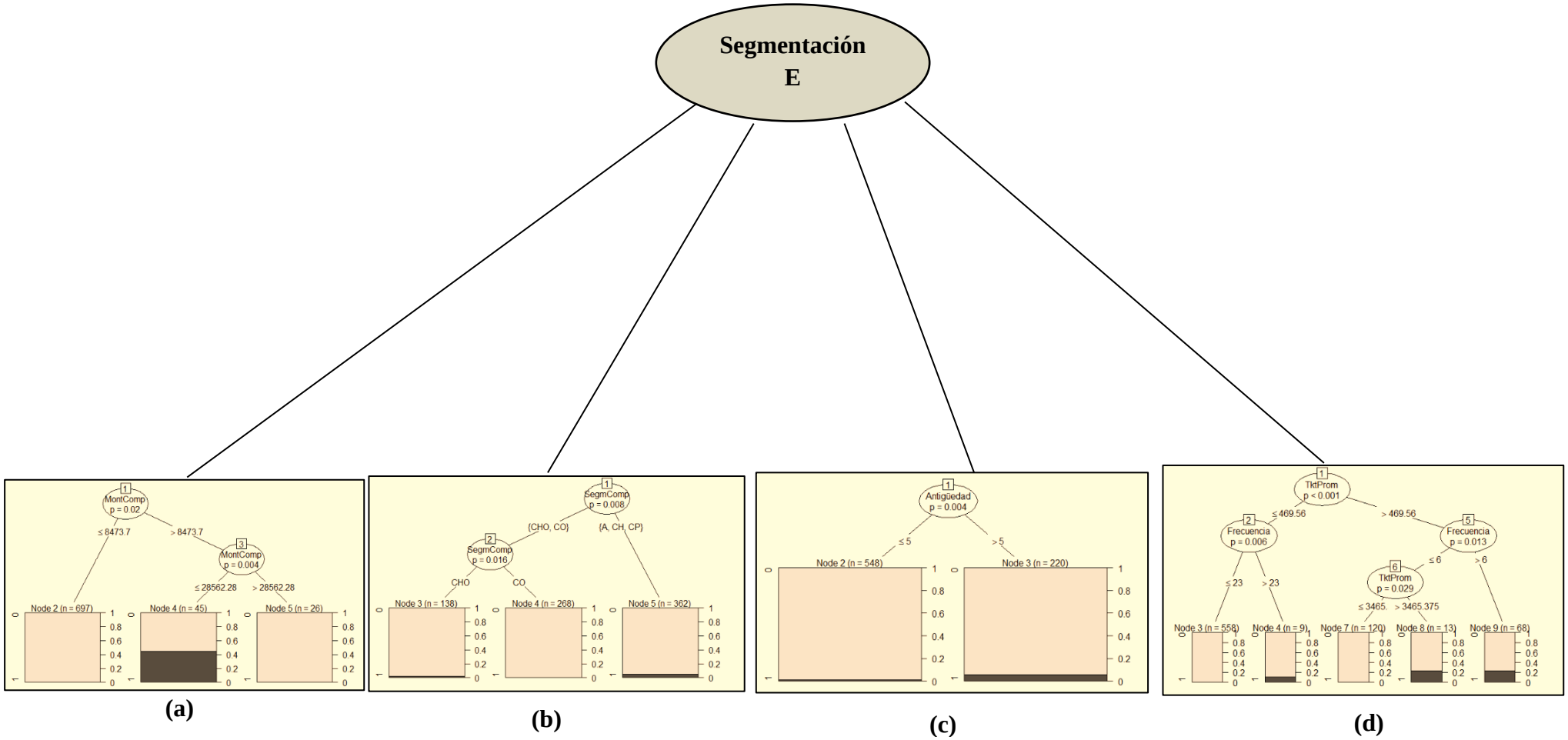


Figura 4. 30

Caracterización del segmento de clientes E

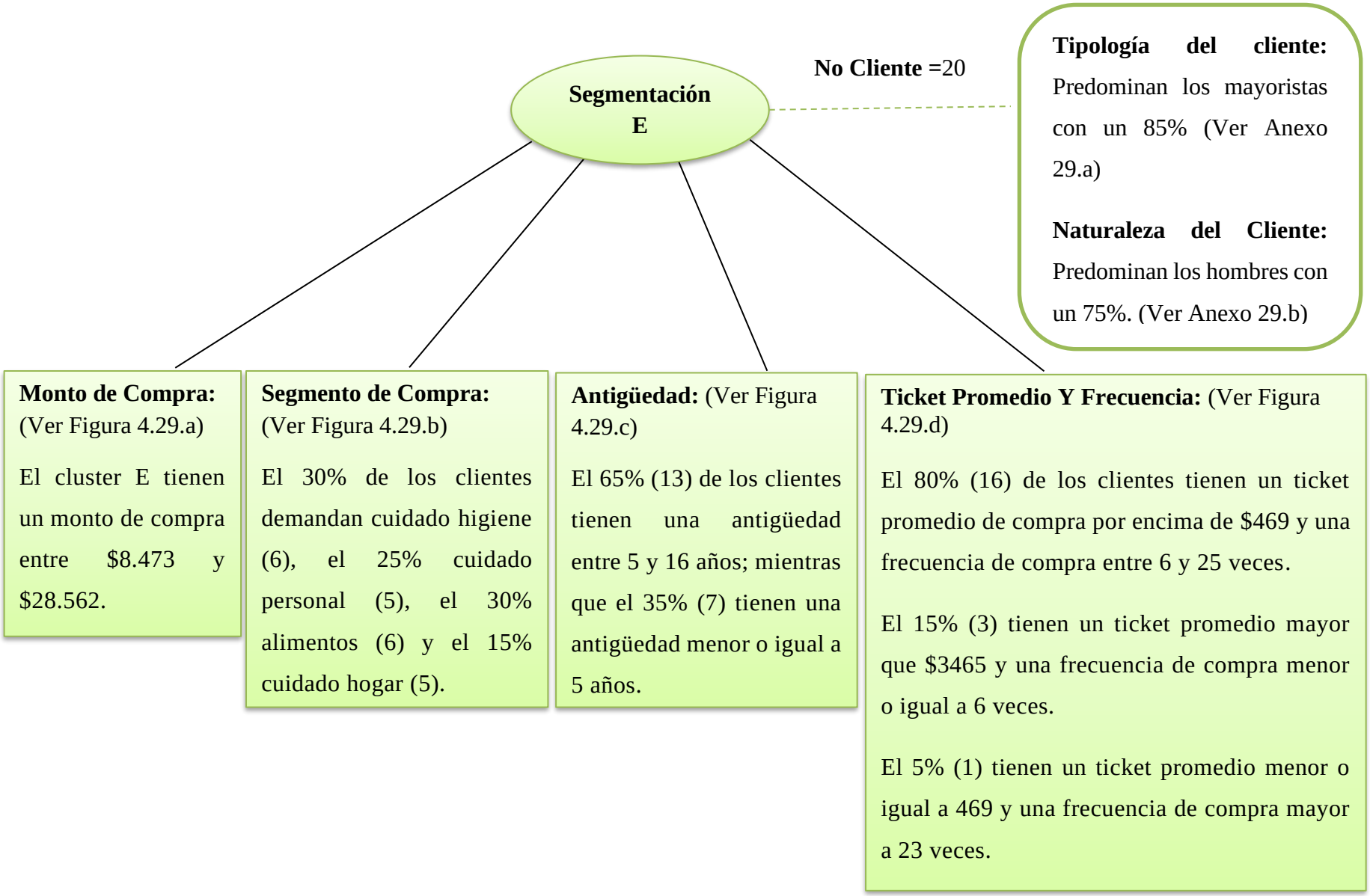
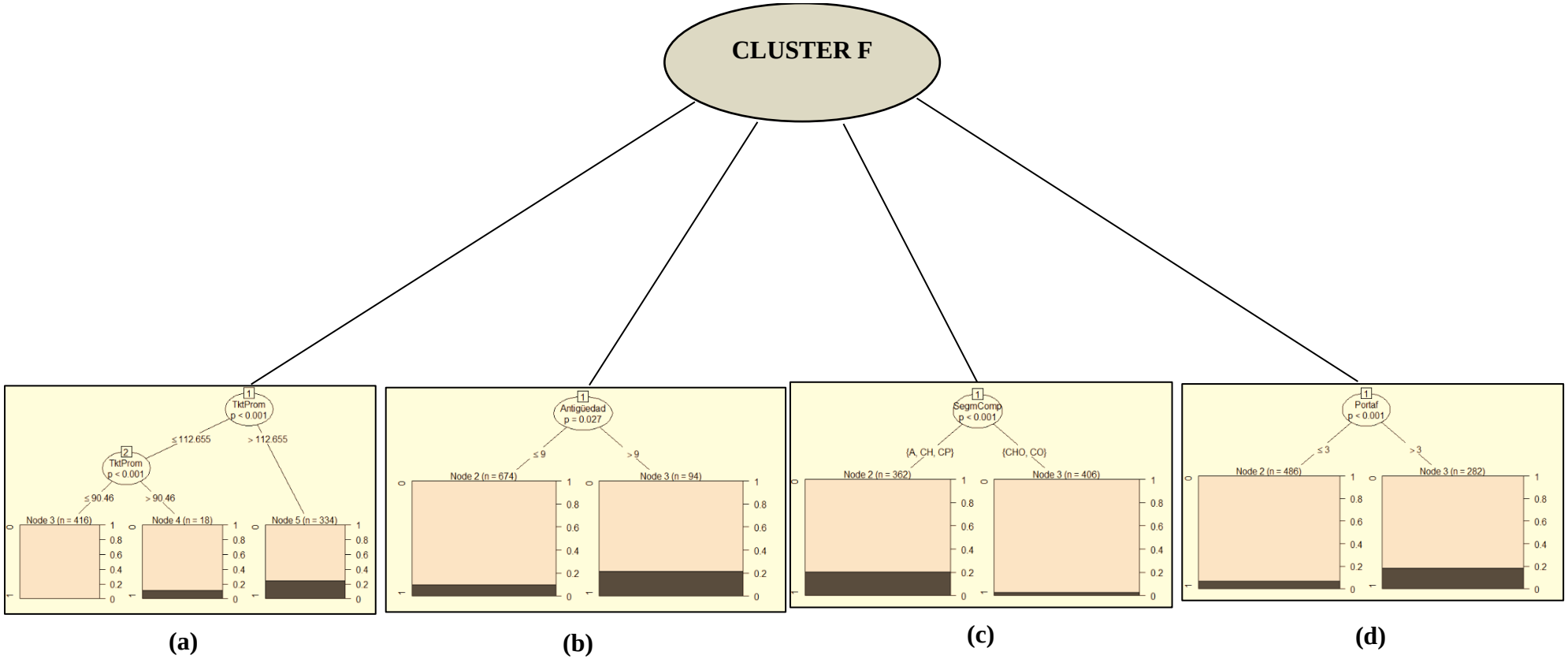


Figura 4. 31

Árboles de decisión del segmento de clientes F



(a) **(b)** **(c)** **(d)**

Figura 4. 32

Caracterización del segmento de clientes F

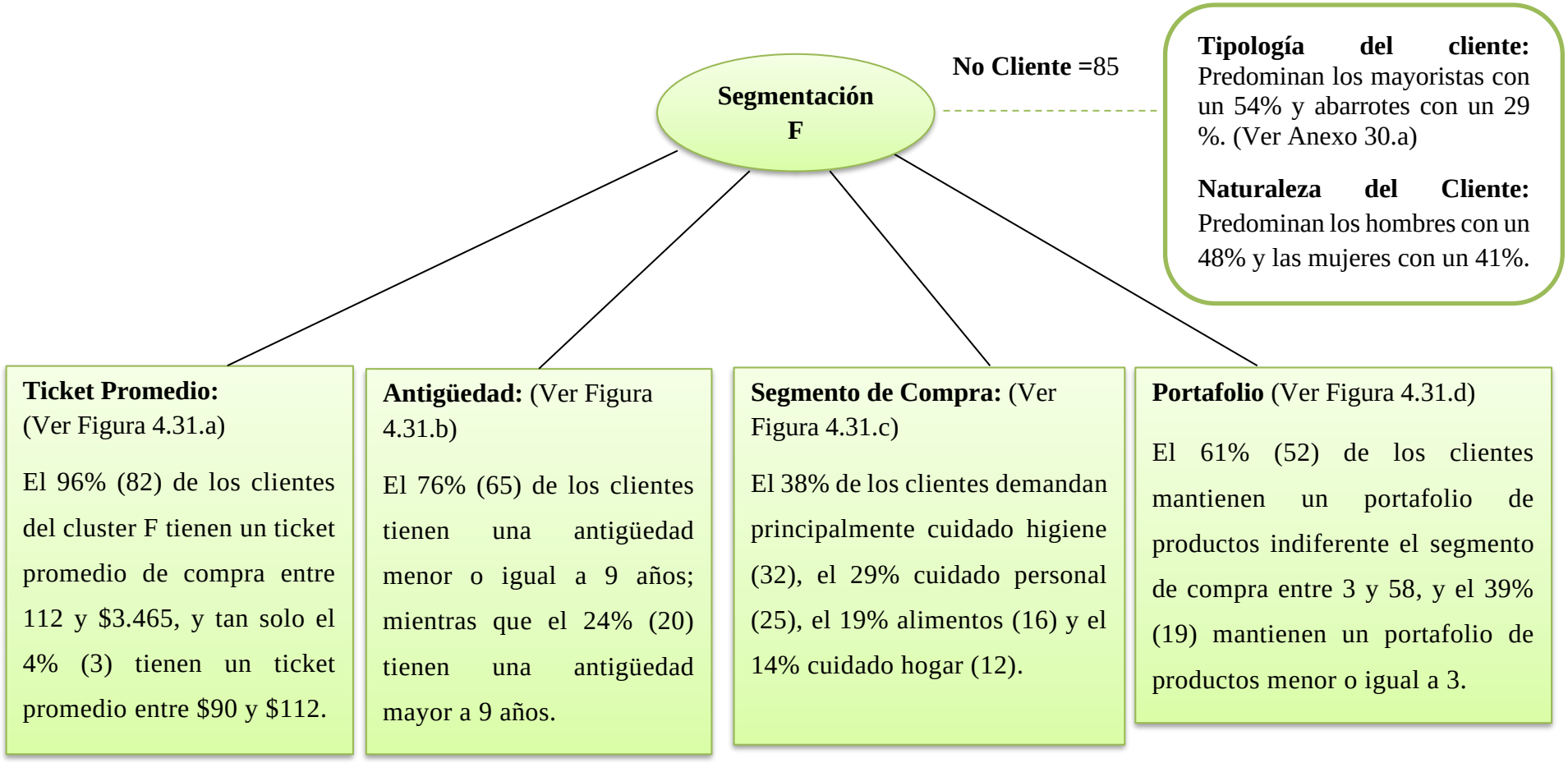
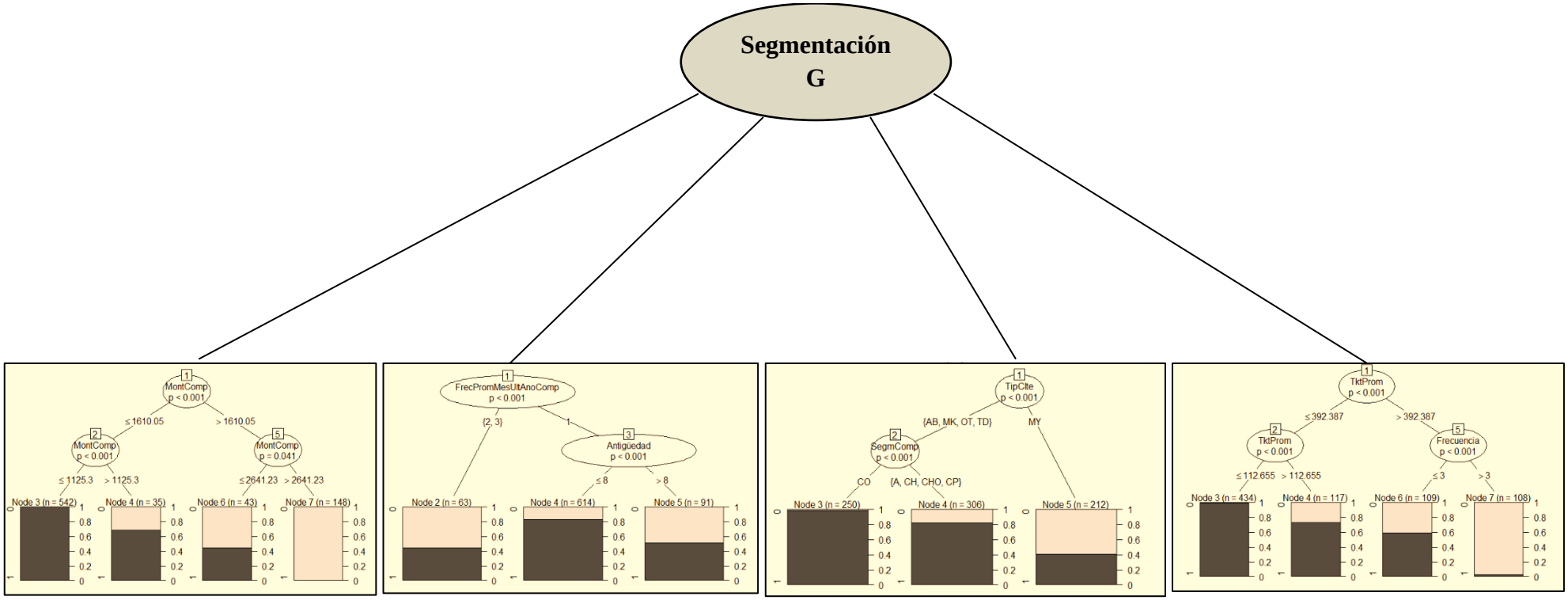


Figura 4. 33

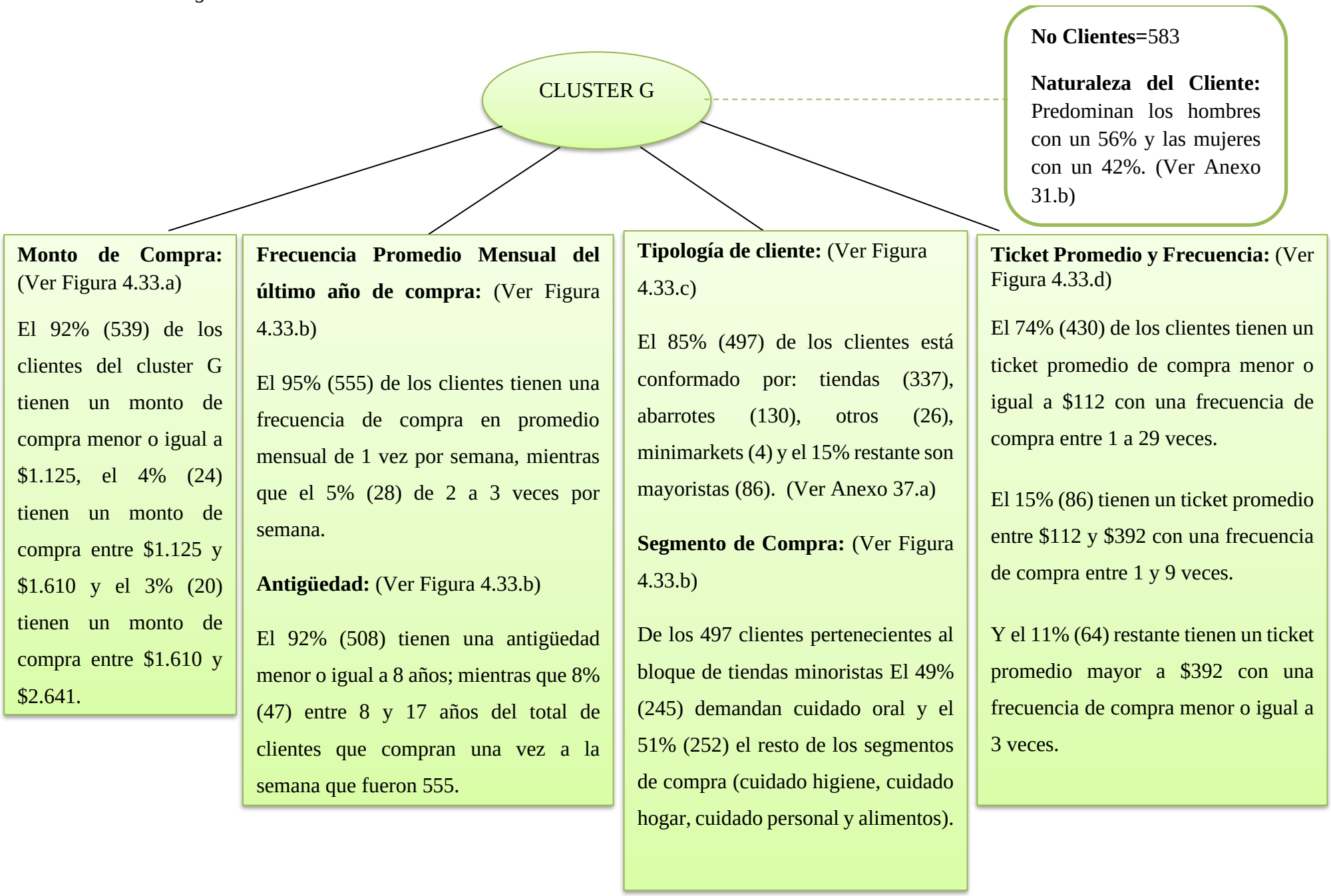
Árboles de decisión del segmento de clientes G



(a) **(b)** **(c)** **(d**

Figura 4. 34

Caracterización del segmento de clientes G



4.4. Modelos de aprendizajes estadísticos

Los segmentos de clientes A, B, C y D que representan en conjunto el 10.5 % de un total de 769 clientes, facturaron durante el periodo comprendido 2020 y 2021, el 83% de las ventas totales de la Empresa AAA que fueron $7.293.277.

Para el pronóstico de ventas se seleccionó las principales categorías de compras realizadas por las 4 segmentaciones de clientes determinadas en base a dos criterios: volumen de ventas y considerando el criterio por parte de la Gerencia Comercial de la Empresa AAA, las cuales fueron Crema Dental, Papel Higiénico, Jabón de Tocador y Aceite.

Las variables que se utilizaron para construir los pronósticos de las 4 categorías mencionadas anteriormente para los 4 segmentos de clientes elegidos (A, B, C y D) se las presentan en la Tabla 4.7, donde la granularidad del registro venta se la determinó por el número de órdenes procesadas por cada segmento de cliente dentro de cada categoría:

Tabla 4. 7

Listado de variables elegidas para la construcción de los modelos de aprendizajes estadísticos

Variable	Descripción de la variable
1. Venta	Suma de la venta en dólares de la categoría por número de órdenes procesadas.
2. Día del Año	Un número entre 1 y 366
3. Dia del Mes	Un número entre 1 y 31
4. Dia Semana	Un número entre 1 y 7
5. Año	Año del número de orden facturada.
6. Tipología de Cliente	Mayorista, Tiendas, Abarrotes, Minimarkets y Otros.
7. Naturaleza del cliente	Hombre, Mujer y Persona Jurídica.
8. Zona	Norte, Sur, Suroeste, Noroeste y Centro.
9. Cantidad de Productos	Número de productos vendidos en cada orden de factura
10. Precio Categoría	La suma del precio del conjunto de productos de la categoría.
11. Cajas Categoría	La suma de cajas del conjunto de productos de la categoría.

A continuación, se presentan los resultados de las categorías anteriormente mencionadas que mejor ajuste y rendimiento mostraron en las 4 segmentos de clientes elegidos, tanto en la selección del mejor modelo escogido en la fase de entrenamiento y la elección del mejor modelo con mayor rendimiento en la fase de prueba.

4.4.1. Pronósticos por categorías para el segmento de clientes A

Tabla 4. 8

Métricas de evaluación de la categoría Crema Dental para el segmento de clientes A

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		617,83	98,96%	82,64%	962,14
KNN	1	568,25	99,04%	8,88%	824,34
Árboles de regresión	5	1.759,58	93,16%	237,82%	1.974,81

Nota: RMSE= Error cuadrático medio, R^2= Coeficiente de determinación, MAPE= Error porcentual absoluto medio

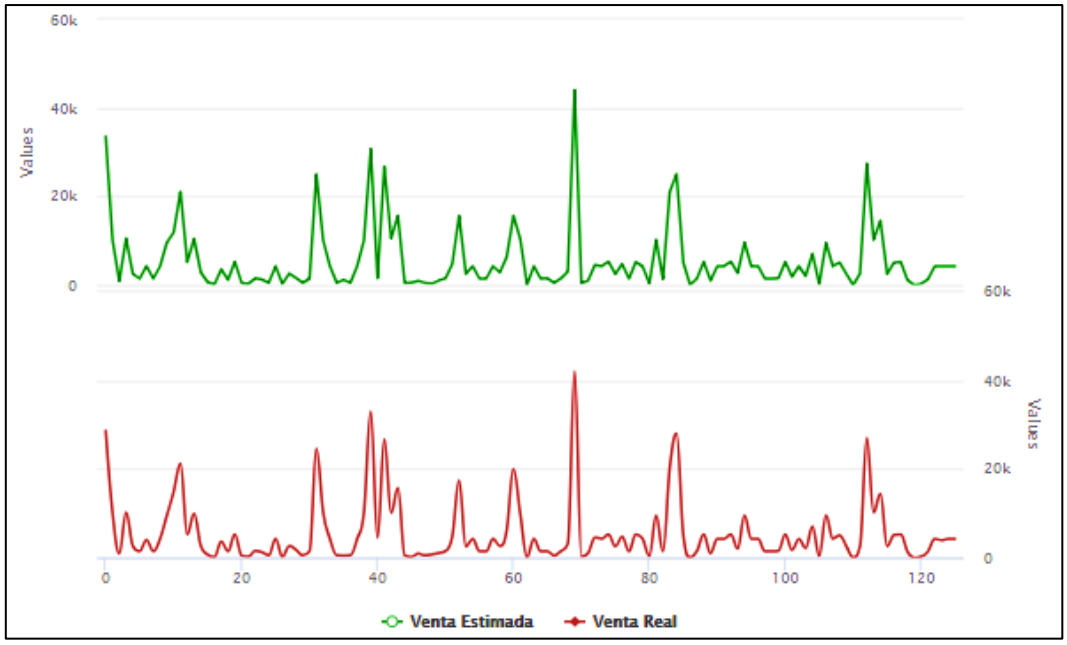
La Tabla 4.8 presenta los resultados para la categoría Crema Dental correspondiente al segmento de clientes A, donde se puede observar que el modelo KNN es el que tuvo un mayor ajuste en la fase de entrenamiento con un R^2 de 99,04% en relación con los modelos de Regresión lineal y Árboles de regresión que presentaron los valores de 98,96% y 93,16% respectivamente. Así mismo, en la medida RMSE de entrenamiento el modelo KNN presenta el valor más bajo que es de 568,25 en comparación con los modelos de Regresión lineal y Árboles de regresión que presentaron los valores 617,83 y 1.759,58 respectivamente. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 32).

La Tabla 4.8 en la fase de prueba también evidencia que el modelo KNN tiene las mejores medidas para ser considerado el mejor modelo tanto en RMSE como MAPE. El modelo KNN presenta el menor MAPE con un valor de 8,88% en relación con los modelos de Regresión lineal y Árboles de regresión que mostraron los valores 82,64% y 237,82% respectivamente. En lo que acontece al RMSE de prueba también se puede apreciar que el modelo KKN presenta el menor valor 824,34, en relación con los valores presentados por los modelos KNN y Árboles de regresión que fueron 962,14 y 1.974,81 respectivamente.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría crema dental del segmento de clientes A:

Figura 4. 35

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Crema Dental del segmento de clientes A.



La Figura 4.35 muestra que la venta estimada (color verde) a través del modelo KNN sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El pronóstico de venta del número de orden 39, presenta un MAPE de 6,87% (subestimado) como resultado de la comparación entre el valor estimado de 30.738 y el valor real de 33.005. (para ver más resultados remítase a ver el Anexo 43)

4.4.2. Pronósticos por categorías para el segmento de clientes B

Tabla 4. 9

Métricas de evaluación de la categoría Crema Dental para el segmento de clientes B

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		229,87	98,11%	8,51%	178,26
KNN	1	387,59	93,93%	11,02%	558,51
Árboles de regresión	5	411,19	94,14%	20,50%	520,74

La Tabla 4.9 presenta los resultados para la categoría Crema Dental correspondiente al segmento de clientes B, donde se puede observar que el modelo de Regresión lineal es el que tuvo un mayor ajuste en la fase de entrenamiento con un R^2 de 98,11% y un RMSE de 229,87 en relación con los modelos KNN y Árboles de regresión que mostraron los valores 93,93% y 94,14% en la medida R^2 respectivamente y de 387,59 y 411,19 en la medida de ajuste RMSE respectivamente. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 33).

En la fase de prueba la Tabla 4.9, también evidencia que el modelo de Regresión lineal es el mejor modelo por las tasas bajas correspondientes a las medidas RMSE y MAPE. En la medida MAPE el modelo de Regresión lineal presenta un valor de 8,51% en relación con los modelos KNN y Árboles de regresión que mostraron los valores: 11,02% y 20,50% respectivamente. En lo que se refiere al RMSE de prueba también se puede apreciar que el modelo de Regresión lineal presenta el menor valor que es de 178,26 en relación con los valores mostrados por los modelos KNN y Árboles de regresión que fueron 558,51 y 520,74 respectivamente.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Crema Dental del segmento de clientes B:

Figura 4. 36

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Crema Dental del segmento de clientes B

La Figura 4.36 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El pronóstico de venta del número de orden 25, presenta un MAPE de 1,42% (sobrestimado) como resultado de la comparación entre el valor estimado que es 4.072 y el valor real de 4.015 (para ver más resultados remítase a ver el Anexo 44)

Tabla 4. 10

Métricas de evaluación de la categoría Toallas Húmedas para el segmento de clientes B

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		75,25	99,34%	3,14%	55,33
KNN	1	379,42	93,62%	14,62%	518,75
Árboles de regresión	10	557,47	85,17%	30,67%	531,74

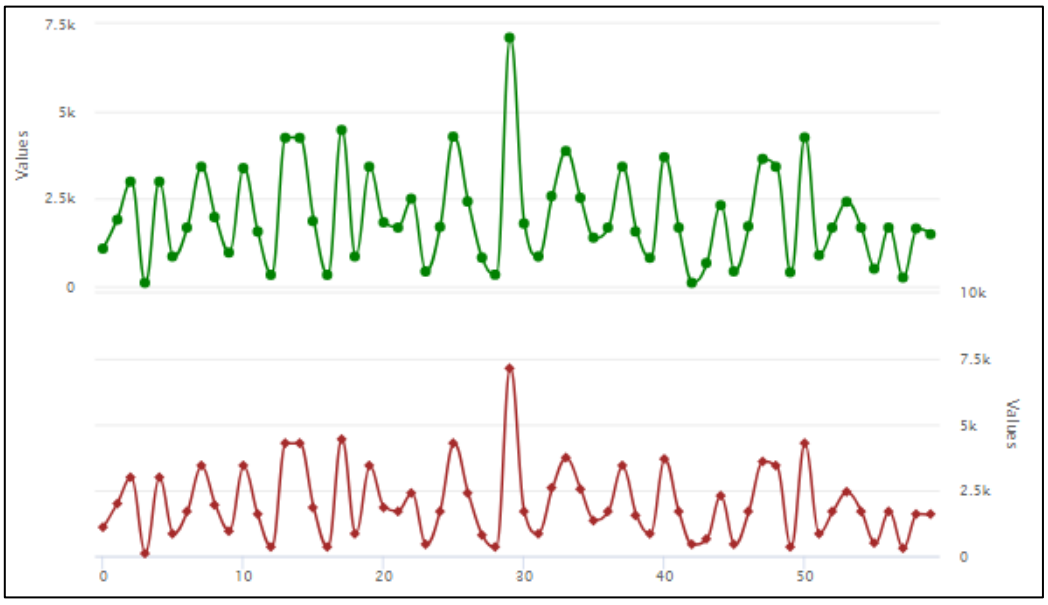
En lo que respecta a la categoría de Toallas Húmedas para el segmento de clientes B se puede observar en la Tabla 4.10, que el modelo de Regresión lineal presenta las mejores medidas de ajustes tanto en la fase de entrenamiento como de prueba. En la fase de entrenamiento se puede observar que el modelo de Regresión lineal es el que tuvo un mayor R^2 que es de 99,34% y el menor RMSE que es de 75,25% en relación con los mejores modelos KNN y Árboles de regresión que presentaron los valores 93,62% y 85,17% respectivamente en la medida R^2 y de 379,42 y 557,47 respectivamente en la medida RMSE. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 34).

En la fase de prueba el modelo de Regresión lineal es el que tiene el menor MAPE con un valor de 3,14% en relación con los modelos KNN y Árboles de regresión que presentan los valores 14,62% y 30,67% respectivamente. De la misma forma si se relaciona el RMSE de prueba del modelo de Regresión lineal con los presentados por los otros dos modelos de estudios.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Toallas Húmedas del segmento de clientes B:

Figura 4. 37

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Toallas Húmedas del segmento de clientes B



La Figura 4.37 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El pronóstico de venta del número de orden 10, presenta un MAPE de 1,19% (subestimado) como resultado de la comparación entre el valor estimado que es 3.388 y el valor real de 3.429. (para ver más resultados remítase a ver el Anexo 45)

Tabla 4. 11

Métricas de evaluación de la categoría Jabón de Tocador para el segmento de clientes B

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		103,16	98,77%	5,85%	93,09
KNN	1	277,27	93,41%	18,09%	401,75
Árboles de regresión	5	372,53	82,02%	56,54%	306,77

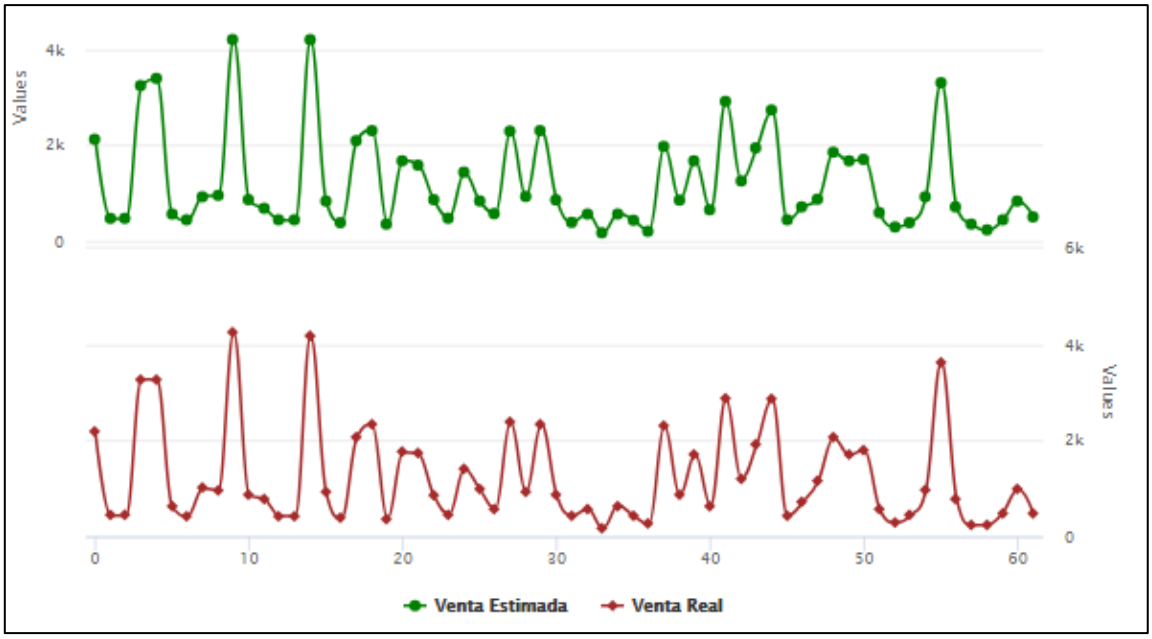
En lo que compete a la categoría de Jabón de Tocador para el segmento de clientes B se puede observar en la Tabla 4.11, que el modelo de Regresión lineal se ajusta de mejor manera a los datos de entrenamiento con respecto a los modelos KNN y Árboles de regresión con valores de 98.77% y 103,16 en la medida R^2 y RMSE respectivamente. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 35).

En la fase de prueba el modelo de Regresión lineal, tal como lo muestra la Tabla 17 presenta el menor MAPE que es de 5,85% y RMSE con un valor de 93,09 en relación con los modelos KNN y Árboles de regresión que presentaron los valores 401,75 y 306.77 respectivamente.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Jabón de Tocador del segmento de clientes B:

Figura 4. 38

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Jabón de Tocador del segmento de clientes B



La Figura 4.38 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El pronóstico de venta del número de orden 44, presenta un MAPE de 5,10% (subestimado) como resultado de la comparación entre el valor estimado que es 2.723 y el valor real de 2.870. (para ver más resultados remítase a ver el Anexo 46)

4.4.3. Pronósticos por categorías para el segmento de clientes C

Tabla 4. 12

Métricas de evaluación de la categoría Crema Dental para el segmento de clientes C

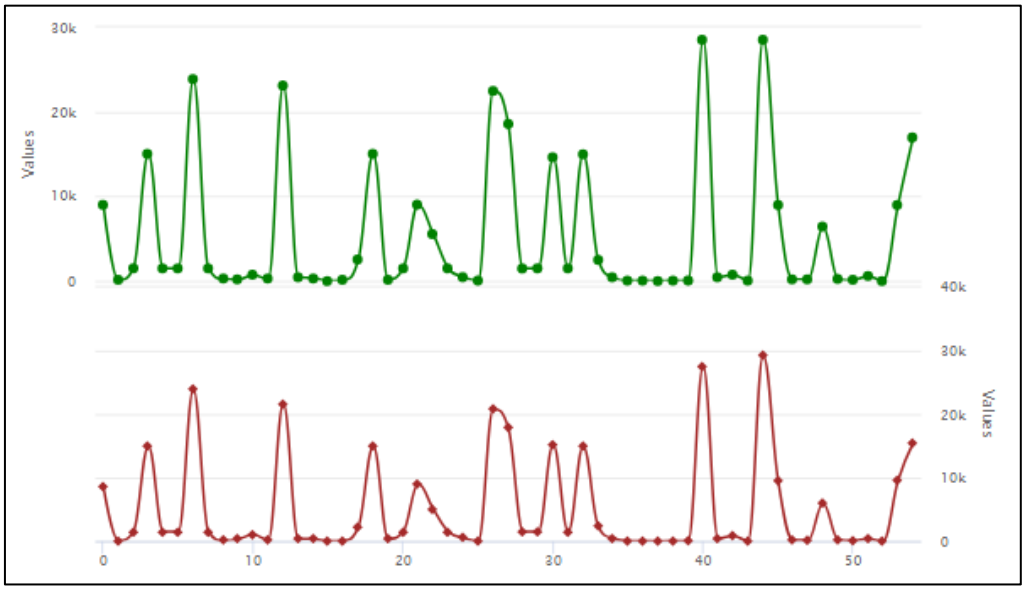
		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		898,04	98,13%	662,52%	1.094,00
KNN	3	887,18	98,29%	9,72%	447,90
Árboles de regresión	3	1.522,95	94,45%	565,85%	1.845,82

El mejor modelo de predicción para la categoría Crema Dental para el segmento de clientes C es el KNN, según la Tabla 4.12 este modelo presenta las mejores medidas de ajustes en la fase de entrenamiento donde obtiene el mayor R^2 que es de 98,29% y el menor RMSE que es de 447,90,70 en relación con los valores de los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 36). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE y RMSE con los valores: 9,72% y 447,90 respectivamente en relación con los valores presentados por los otros dos modelos analizados.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Crema Dental del segmento de clientes C:

Figura 4. 39

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Crema Dental del segmento de clientes C



La Figura 4.39 muestra que la venta estimada (color verde) a través del modelo KNN sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. Si consideramos como punto de evaluación el pronóstico de venta del número de orden 26, presenta un MAPE de 7,78% (sobreestimado) como resultado de la comparación entre el valor estimado que es 22.456 y el valor real de 20.835 (para ver más resultados remítase a ver el Anexo 47)

Tabla 4. 13

Métricas de evaluación de la categoría Jabón de Tocador para el segmento de clientes C

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		144,32	96,58%	14,61%	348,41
KNN	5	665,10	63,85%	112,22%	1.020,69
Árboles de regresión	5	674,29	57,70%	161,07%	1.000,46

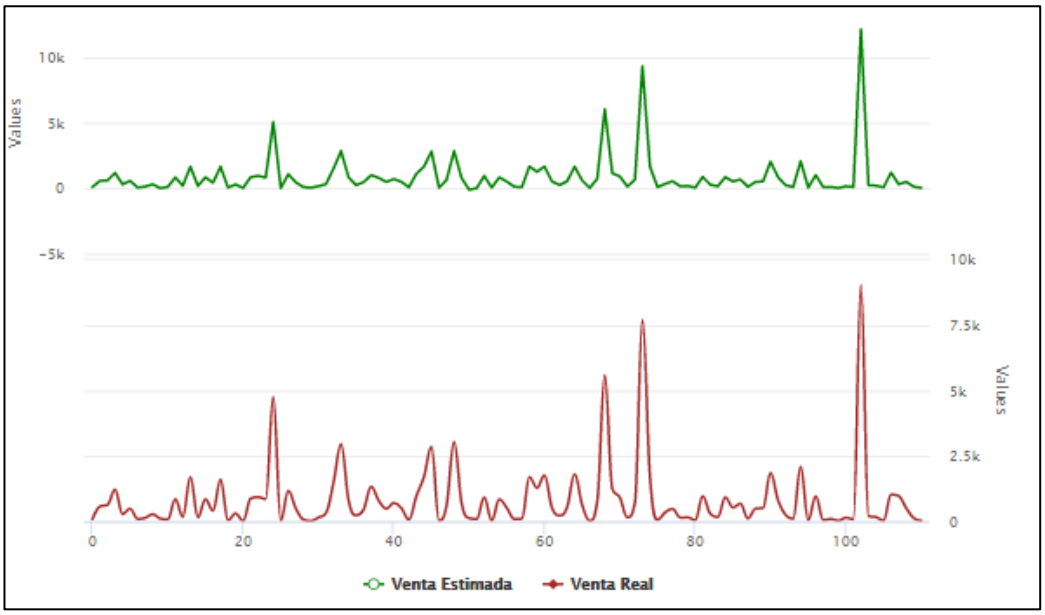
El mejor modelo de predicción para la categoría Jabón de Tocador para el segmento de clientes C es el modelo de Regresión lineal, pues según la Tabla 4.13 presenta las mejores

medidas de ajustes en la fase de entrenamiento donde obtiene el mayor $\mathbf{R^2}$ que es 96,58% y el menor RMSE que es de 144,32, en relación con los valores mostrados por los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 37). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE y RMSE con los valores:14,61% y 348,41 respectivamente en relación con los otros dos modelos estudiados.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Jabón de Tocador del segmento de clientes C:

Figura 4. 40

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Jabón de Tocador del segmento de clientes C



La Figura 4.40 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. Si evaluamos la precisión del pronóstico de venta del número de orden 62, este, presenta un MAPE de 6,92% (sobreestimado) como resultado de la comparación entre valor estimado que es 246 y el valor real 230. (para ver más resultados remítase a ver el Anexo 48)

Tabla 4. 14

Métricas de evaluación de la categoría Papel Higiénico para el segmento de clientes C

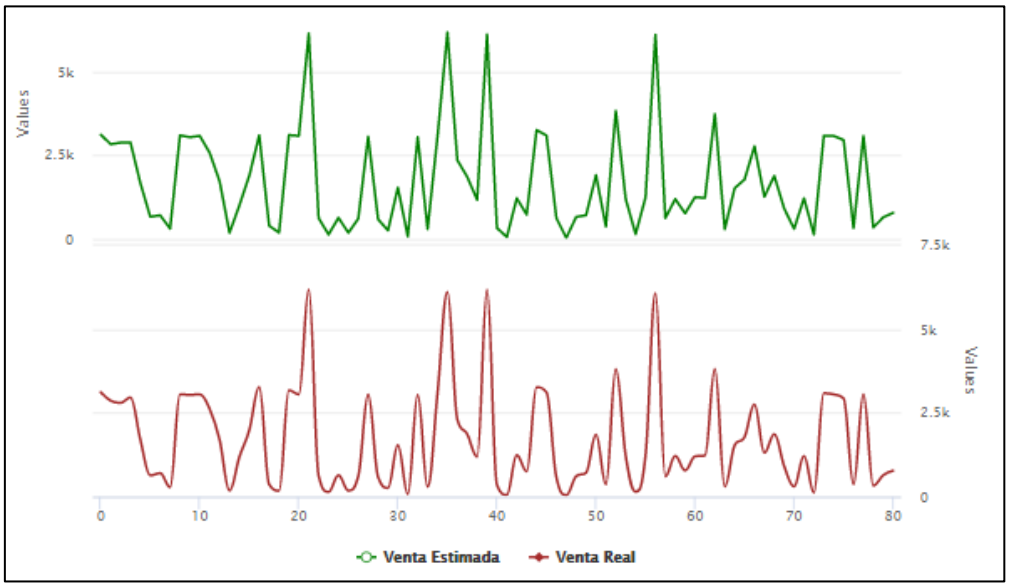
		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		60,48	99,65%	2,77%	43,02
KNN	1	281,74	93,72%	10,50%	343,18
Árboles de regresión	10	608,05	77,52%	65,83%	707,08

El modelo de predicción que se ajusta de mejor manera para la categoría Papel Higiénico para el segmento de clientes C es el modelo de Regresión lineal, pues según la Tabla 20 presenta el mayor R^2 que es 99,65% y el menor RMSE que es 60,48 en relación con los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 38). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE que es 2,77% y RMSE que es 43,02. Y si se lo relaciona estos valores con las medidas de los otros dos modelos, son más bajas.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Papel Higiénico del segmento de clientes C:

Figura 4. 41

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Papel Higiénico del segmento de clientes C



La Figura 4.41 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El número de orden 3 presenta un MAPE de 2,66% (subestimado) como resultado de la comparación entre el valor estimado que es 2.886 y el valor real 2.965. (para ver más resultados remítase a ver el Anexo 49)

Figura 4. 42

Métricas de evaluación de la categoría Aceite para el segmento de clientes C

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		315,09	92,09%	27,71%	235,02
KNN	1	547,87	86,46%	56,96%	582,88
Árboles de regresión	3	830,98	82,96%	129,08%	1.179,44

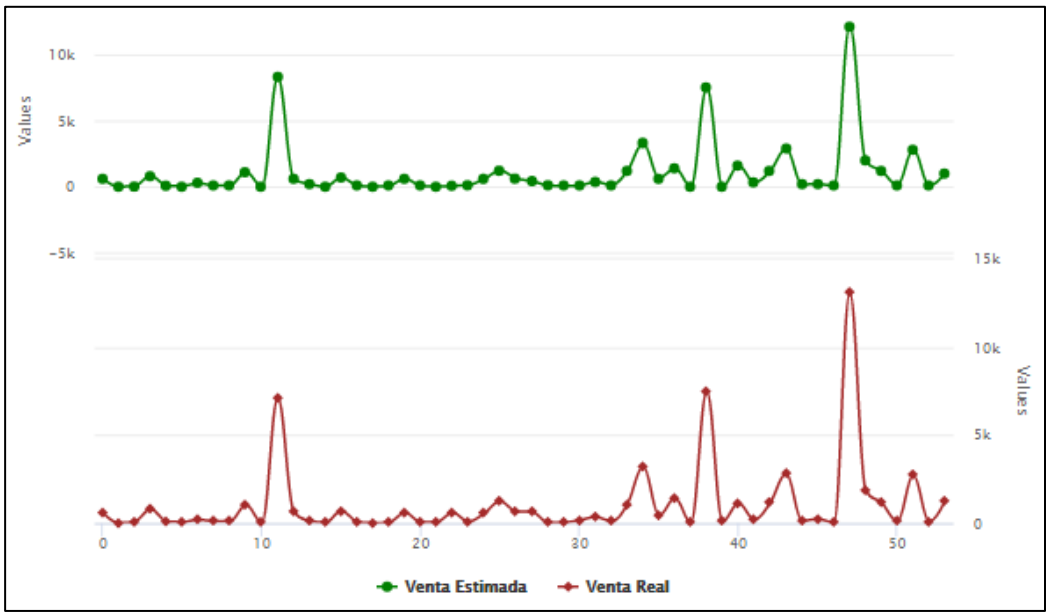
El mejor modelo de predicción para la categoría Aceite para el segmento de clientes C es el modelo de Regresión lineal, pues según la Tabla 21 presenta las menores medidas de ajustes tanto en la fase de entrenamiento como de prueba. En la fase de entrenamiento, el modelo de regresión lineal obtiene el mayor $\mathbf{R^2}$ que es 92,09% y el menor RMSE, el cual es

315,09 en relación con los valores de los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y árboles de regresión remítase al Anexo 39). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE y RMSE que fueron de 27,71% y 235,02 respectivamente con respecto a los otros dos modelos.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Aceite del segmento de clientes C:

Figura 4. 43

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Aceite del segmento de clientes C



La Figura 4.43 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El número de orden 40, presenta un MAPE de 40,04% (sobreestimado) como resultado de la comparación entre el valor estimado y real que mostraron los valores de 1.583 y 1.130 respectivamente. (para ver más resultados remítase a ver el Anexo 50)

4.4.4. Pronósticos por categorías para el segmento de clientes D

Tabla 4. 15

Métricas de evaluación de la categoría Papel Higiénico para el segmento de clientes D

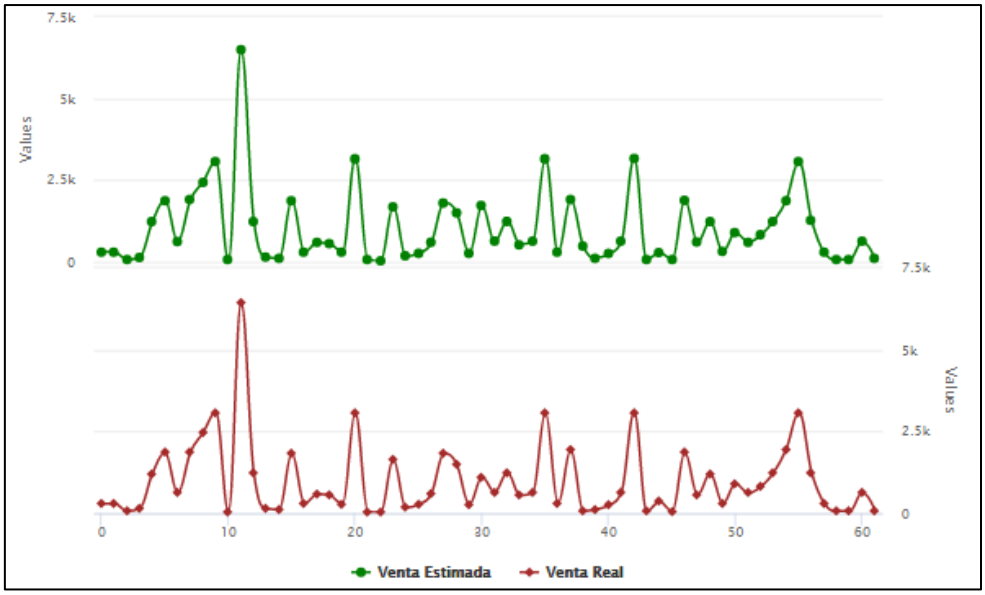
		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		111,65	97,69%	13,83%	97,09
KNN	1	231,51	94,41%	27,30%	210,21
Árboles de regresión	10	461,68	77,38%	119,54%	485,60

El mejor modelo de predicción para la categoría Papel Higiénico para el segmento de clientes D es el modelo de Regresión lineal, pues según la Tabla 4.15 presenta las mejores medidas de ajustes en la fase de entrenamiento donde obtiene el mayor R^2 que fue 97,69% y el menor RMSE que fue 111,65 con respecto a los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 40). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE y RMSE que fueron de 27,71% y 235,02 respectivamente con respecto a los valores presentados por los otros dos modelos.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Papel Higiénico del segmento de clientes D:

Figura 4. 44

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Papel Higiénico del segmento de clientes D



La Figura 4.44 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. El número de orden 15, presenta un MAPE de 27,74% (sobreestimado) como resultado de la comparación entre el valor estimado y real que mostraron los valores de 1.883 y 1.833 respectivamente. (para ver más resultados remítase a ver el Anexo 51)

Tabla 4. 16

Métricas de la categoría Aceite para el segmento de clientes D

		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		171,56	96,26%	11,81%	111,51
KNN	10	486,27	88,05%	32,56%	433,36
Árboles de regresión	3	333,34	84,44%	22,47%	316,11

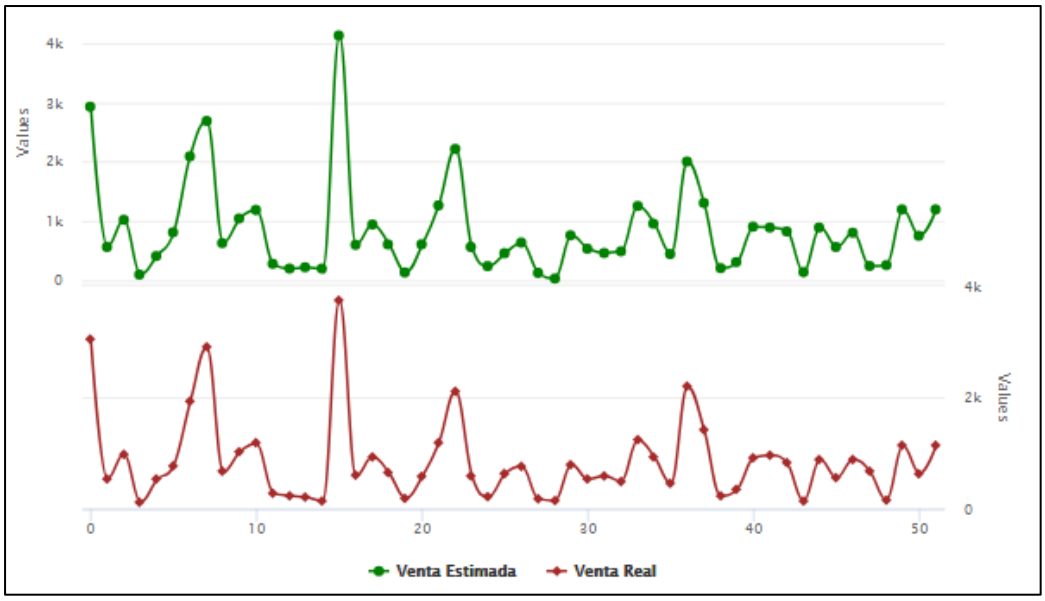
El mejor modelo de predicción para la categoría Aceite para el segmento de clientes D es el modelo de Regresión lineal, pues según la Tabla 4.16 presenta las mejores medidas de ajustes en la fase de entrenamiento donde obtiene el mayor R^2 que fue 96,26% y el menor

RMSE que fue 171,56 con respecto a los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 41). En la fase de prueba el modelo KNN es el que el obtiene el mayor rendimiento, pues presenta el menor MAPE y RMSE que fueron de 11,81% y 111,51 respectivamente con respecto a los otros dos modelos.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Aceite del segmento de clientes D:

Figura 4. 45

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Aceite del segmento de clientes D



La Figura 4.45 muestra que la venta estimada (color verde) a través del modelo Regresión lineal sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. Si consideramos para análisis el número de orden 37, presenta un MAPE de 8,05% (subestimado) como resultado de la comparación entre el valor estimado y real que mostraron los valores de 1.307 y 1.421 respectivamente. (para ver más resultados remítase a ver el Anexo 52)

Tabla 4. 17

Métricas de evaluación de la categoría Jabón de Tocador para el segmento de clientes D

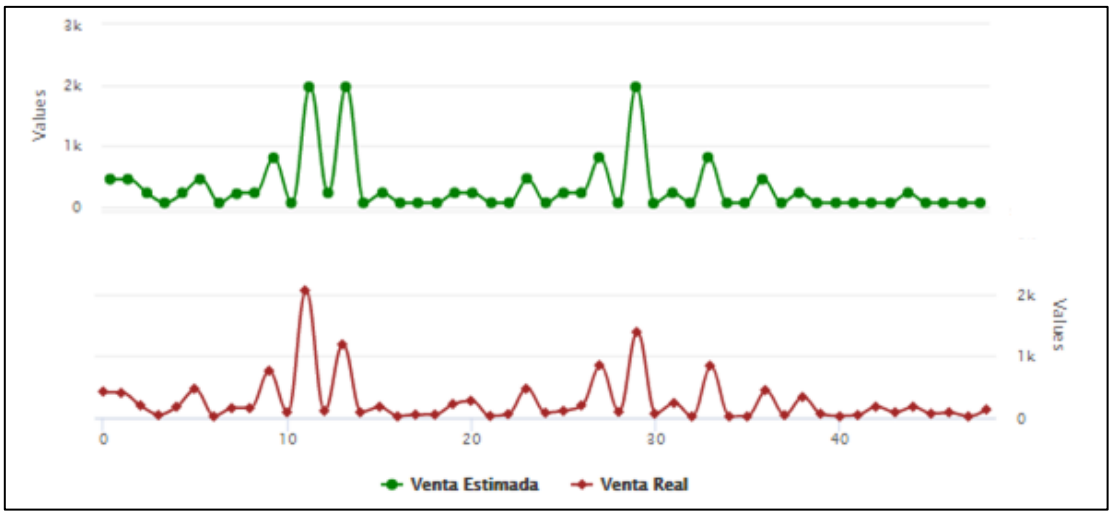
		Entrenamiento		Prueba	
Modelos	**K**	**RMSE**	**R^2**	**MAPE**	**RMSE**
Regresión lineal		200,39	90,80%	97,50%	141,28
KNN	5	254,35	84,96%	56,54%	139,81
Árboles de regresión	3	256,84	81,22%	49,39%	147,41

Según la Tabla 4.17 el modelo de Regresión lineal es el que presenta las mejores medidas de ajustes en la fase de entrenamiento donde obtiene el mayor **R^2** que fue 90,80% y el menor RMSE que fue 200,39 con respecto a los otros dos modelos analizados. (para ver los modelos para diferentes k de KNN y Árboles de regresión remítase al Anexo 42). Sin embargo, en la fase de prueba este modelo no presenta las mejores medidas de ajustes, siendo superado por Árboles de regresión, que muestra un mayor rendimiento, tanto en MAPE y RMSE, con valores 49,39% y 147,41% respectivamente.

A continuación, se presenta la gráfica del modelo con mayor ajuste y rendimiento para la categoría Jabón de Tocador del segmento de clientes D:

Figura 4. 46

Pronóstico de las ventas netas estimadas vs reales por número de orden procesadas para la categoría Jabón de Tocador del segmento de clientes D



La Figura 4.46 muestra que la venta estimada (color verde) a través del modelo de Árboles de regresión sigue un comportamiento similar a la venta real (color rojo) del conjunto de datos de prueba. Si consideramos para análisis el número de orden 29, presenta un MAPE de 42,06% (sobreestimado) como resultado de la comparación entre el valor estimado y real que mostraron los valores de 1.968 y 1.386 respectivamente. (para ver más resultados remítase a ver el Anexo 53)

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

- Considerando que la variable tipología del cliente expresa un criterio de segmentación a priori, esta, no es suficiente para discriminar de manera adecuada grupos de clientes con similares características de compras, por tal motivo se recurrió a la técnica de segmentación K-means que mostró un mejor desempeño para dicha tarea.

- Con la técnica K-means se determinó el número adecuado de 7 clusters con características de compras realistas, para lo cual se utilizó las variables montos de compras y cantidad de productos demandados, esto, independientemente de que ya existe una tipología de cliente.

- La caracterización de clientes a través del modelo de Árbol de decisión CHAID, reforzó la agrupación que se había considerado de 7 clusters, ya que, al considerar otras variables como frecuencia de compra, segmento de compra, antigüedad del cliente, ticket promedio de compra mensual, etc. Se pudo clasificar a clientes mayoristas, abarrotes, minimarkets y tiendas que estaban agrupados en el mismo cluster y compartían dichas características en segmentos más específicos.

- El análisis de segmentación mostró que el 11% de los clientes de un total de 769, representados por los segmentos A, B, C y D, alcanzan el 83,30% del monto total de ventas consideradas en el periodo comprendido entre los años 2020-2021.

- Con los segmentos de clientes anteriores identificamos variables con características relevantes para la generación de pronósticos de ventas más precisos y confiables dentro cada clusters en las categorías Crema Dental, Papel Higiénico, Jabón de Tocador y Aceite y además facilita la oportunidad de mejora

en la calidad del servicio exclusivo que se les puede ofrecer a dicho grupo de clientes.

- El modelo de Regresión lineal fue el modelo que mayormente se ajustó a los datos de pruebas de las categorías de los distintos segmentos de clientes presentando medidas de errores más bajas en cuanto a MAPE y RMSE con relación a las medidas presentadas por los modelos KNN y Árboles de regresión, obteniendo resultados sobresalientes especialmente en la categoría Papel Higiénico del segmento de clientes C con un MAPE apenas del 2,77% y un MAPE de 3,14% en la categoría Toallas Húmedas.

- El segmento de clientes A obtuvo una mayor precisión en la fase de prueba por intermedio del modelo KNN, logrando un MAPE del 8,88% en la categoría Crema Dental, constituyendose como un modelo fiable para la generación de pronósticos para las ventas por números de ordenes procesadas para dicho grupo.

- El segmento de clientes B a través de modelo de Regresión lineal obtuvo una alta precisión en la fase de prueba con respecto a los prónosticos por número de orden procesadas tanto para la categoría Crema Dental, Toallas Húmedas y Jabón Tocador con un MAPE de 8.51%, 3,14% y 5,85% respectivamente, generando predicciones bastante confiables para dicho segmento de clientes.

- El segmento de clientes C presentó predicciones aceptables en el conjunto de prueba con un MAPE por debajo del 10%, en la cetagoría Crema Dental por medio del modelo KNN y Papel Higiénico por intermedio del modelo Regresión lineal. La categoría Jabón de Tocador tuvo un rendimienro aceptable desde el punto de vista comercial ya que presentó un MAPE de 14,61%, no así la categoría Aceite que tuvo un MAPE de 27,71% siendo la categoría con menor ajuste dentto de dicho segmento de clientes.

- El segmento de clientes D mostró un rendimiento aceptable de pronósticos en la categoría Papel Higiénico y Aceite con un MAPE menor a 14% para ambas

categorías por intermedio de su mejor modelo regresión lineal, no asi para la categoría Jabón de Tocador que a pesar de tener altas medidas de ajustes en la fase de entrenamiento, presentó un MAPE superior a 49% para todos los modelos en el conjunto de prueba, concluyendo que ningún modelo presenta un buen rendimiento a la hora de estimar el número de órdenes facturadas dentro de esta categoría.

- Los beneficios entregados para la empresa en general son altamente confiables debido a que una mejora en la predicción de la demanda puede contribuir al área de producción a reducir sus costos operativos y al área comercial a establecer estrategias para construir una adecuada planificación de acciones comerciales.

5.2. RECOMENDACIONES

- Se recomienda incluir otras variables que tengan influencia dentro de los modelos propuestos para corroborar la veracidad de los datos debido a que obtuvieron R^2 altos, esto con el fin de evitar sobreajustes.

- De acuerdo a los resultados obtenidos por las categorías antes mencionadas, sería factible que la empresa implemente una herramienra análitica interna automarizada para la generación de pronósticos de los diferentes grupos de clientes.

- Contar con información actualizada con la misma estructura y vigente para poder realizar pronósticos periódicos de las categorías en los segmentos de clientes identificados.

- Ampliar el grupo de variables propuestas dentro de la caracterización de clientes para poder crear subcluster con rasgos similares de los grupos ya formados.

6. REFERENCIAS

Abadía, Palacios, a., & Pastor, n. (2015). Segmentacion de clientes de una empresa comercializadora de productos de consumo masivo en la ciudad de popayán soportado en machine learning y analisis rfm (Recency, Frecuency Y money). *Paper Knowledge . Toward a Media History of Documents*, *3*(April), 49–58.

Acosta, A. L. (2017). *Canales de distribucion*. https://digitk.areandina.edu.co/bitstream/handle/areandina/1270/Canales de Distribución.pdf?sequence=1&isAllowed=y

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015a). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–485. https://doi.org/10.1257/aer.p20151021

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015b). NBER WORKING PAPER SERIES DEMAND ESTIMATION WITH MACHINE LEARNING AND MODEL COMBINATION Demand Estimation with Machine Learning and Model Combination. *NBER Working Paper Series*, *20955*. http://www.nber.org/papers/w20955

Burgaentzle Jarrín, F. (2016). *Pronósticos y modelos de inventarios en las industrias de alimentos: caso de estudio de una empresa láctea ecuatoriana.* [(Bachelor's thesis, Quito: USFQ, 2016).]. https://repositorio.usfq.edu.ec/bitstream/23000/6226/1/128660.pdf

Cálad, F. (2015). Segmentación de clientes automatizada a partir de técnicas de minería de datos (K-Means Clustering). *Trabajo de Grado - Pregrado, Universidad EIA*, 88.

Carreño Carreño, C. A., & Troncoso Tirapegui, J. J. (2019). Diseño de metodología para el pronóstico de demanda con modelos de series de tiempo para las familias de longanizas y madurados de PF S.A. *Universidad de Talca (Chile). Escuela de Ingenieria Civil Industrial*. http://dspace.utalca.cl/handle/1950/12269

Cuadras, C. M. (2007). Nuevos Metodos de Analisis Multivariante. *Revista Española de Quimioterapia : Publicación Oficial de La Sociedad Española de Quimioterapia*, *20*(3), 249. http://www.ncbi.nlm.nih.gov/pubmed/19406528

Delgado, J. (2016). *P ARTICIÓN P AR -I MPAR es NP-difícil*. 1–5. https://www.ime.usp.br/~jdelgado/misc/particionParImpar.pdf

Díaz Sepúlveda, J. F., & Correa, J. C. (2013). Comparación entre árboles de regresión CART y regresión lineal. *Comunicaciones En Estadística*, *6*(2), 175. https://doi.org/10.15332/s2027-3355.2013.0002.05

Eguiguren Calisto, P. A. (2019). Modelos de proyección de demanda para productos de alta volatilidad y bajo volumen en ventas dentro de una empresa de alimentos. *Tesis (Ingeniero Industrial), Universidad San Francisco de Quito, Colegio de Ciencias e Ingenierías; Quito, Ecuador*, *2019*, 1–62. https://repositorio.usfq.edu.ec/bitstream/23000/8320/1/142766.pdf

Escobar Mercado, R. M. (2002). Las aplicaciones del análisis de segmentación: el procedimiento Chaid. *Empiria. Revista de Metodología de Ciencias Sociales*, *0*(1), 13. https://doi.org/10.5944/empiria.1.1998.706

Espinoza, E. N., & Bancalari, M. E. (2005). Aplicación de métodos CART en el modelamiento de la densidad de la madera: un enfoque práctico. In M. A. Ortíz Jorge (Ed.), *Modelamiento Estadístico* (p. 335). Departamento de Estadística de la Universidad Nacional de Colombia,. https://books.google.com.ec/books?hl=es&lr=&id=1UGDhYbvzYAC&oi=fnd&pg=PA335&dq=Aplicación+de+métodos+CART+en+el+modelamiento+de+la+densidad+de+la+madera:+un+enfoque+práctico&ots=2vYhEhIEh2&sig=hbcLU3M9xkZbYVcOcfxIf9JR_hg&redir_esc=y#v=onepage&q=Aplicación

Evers, J. M., Tavasszy, L., van Duin, R., Schott, D. L., & Gorte, F. (2018). Demand forecast models for online supermarkets. *The Network on European Communications and Transport Activity Research Conference*, 1–13. https://repository.tudelft.nl/islandora/object/uuid:76257f4e-55fd-45e7-b3b0-38a59e873455?collection=research

Femina, B. T., & Sudheep, E. M. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*(Icict 2014), 725–731. https://doi.org/10.1016/j.procs.2015.02.136

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine

learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420–1438. https://doi.org/10.1016/j.ijforecast.2020.02.005

James, G., DanielaWitten, Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning wth application in R*.

Kassambara, A. (2015). *Alboukadel Kassambara Practical Guide To Cluster Analysis in R*. 1–187.

Llumitasig Galarza, M. C. (2021). *Simulación de pronósticos de ventas en la empresa IMPACTEX mediante redes neuronales*. (Master's thesis.Universidad Técnica de Ambato.Facultad de Ingeniería en Sistemas, Electrónica e Industrial. Maestría en Matemática Aplicada).

Palacios, C. (2020). Análisis y predicción de las tendencias de venta en el mercado usando árboles de regresión. In *Universidad San Francisco De Quito. Bachelor's thesis, Quito)*. https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=Análisis+y+predicción+de+las+tendencias+de+venta+en+el+mercado+usando+árboles+de+regresión&btnG=

Peña, D. (2002). Analisis de Datos Multivariantes, 2002. *McGraw-Hill, Madrid, December*, 201–226. https://www.researchgate.net/profile/Daniel_Pena4/publication/40944325_Analisis_de_Datos_Multivariantes/links/549154880cf214269f27ffae/Analisis-de-Datos-Multivariantes.pdf?origin=publication_detail

Penpece, D., & Elma, O. E. (2014). Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey. *International Journal of Trade, Economics and Finance*, *5*(5), 435–440. https://doi.org/10.7763/ijtef.2014.v5.411

Rivero, C. (2012). *Clasificación De Clientes Mediante Técnicas De Minería De Datos Para La Empresa Comertex S.a.*

Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R. (2022). Loan Prediction System Using Machine Learning. *ITM Web of Conferences*, *44*(2), 03019. https://doi.org/10.1051/itmconf/20224403019

Siwerz, R., & Dahlén, C. (2017). Predicting sales in a food store department using machine learning. *Degree Project Computer Engineering*.

Slimani, I., El Farissi, I., & Achchab, S. (2015). Artificial Neural Networks/or Demand Forecasting: Application Using Moroccan Supermarket Data. *15th International Conference on Intelligent Systems DeSign and Applications (ISDA)*, 266–271.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*(ii), 679–683. https://doi.org/10.1016/j.procir.2019.02.042

Wang, J., Liu, G. Q., & Liu, L. (2019). A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry. *2019 4th IEEE International Conference on Big Data Analytics, ICBDA 2019*, 317–320. https://doi.org/10.1109/ICBDA.2019.8713196

Wing, V., & Chan, K. (2018). *Forecasting Seasonal Footwear Demand Using Machine Learning*. The Hong Kong Polytechnic University.

Yu, Y., Choi, T. M., & Hui, C. L. (2011). An intelligent fast sales forecasting model for fashion products. *Expert Systems with Applications*, *38*(6), 7373–7379. https://doi.org/10.1016/j.eswa.2010.12.089

7. APÉNDICES Y ANEXOS

Anexo 1

Tabla Registros de ventas a nivel nacional, periodo 2017-2021

SKU	Factura	Fecha	Cliente	Venta Neta
001151	028939	20/1/2017	Cachupud Carrillo Cirilo	18,94
001151	029774	31/1/2017	Ordoñez Armijos Edwin Geovanny	18,94
001151	047990	18/1/2018	Yaucan Paca Manuela	38,13
1180005	048885	31/1/2018	Jara Cuzco Vicente Leonardo	55,79
04020035	068282	17/1/2019	Sarango Ramírez Edwin Fernando	1,44
080129	069352	31/1/2019	Mayolema Paguay Aida Piedad	17,12
1180005	089384	17/1/2020	López Maldonado José Rodrigo	33,48
12863	089987	31/1/2020	Reyes Buri German Enrique	277,6
1TTEC000149	108330	31/1/2021	Bastidas Arciniega Ángel	356,03

Anexo 2

Tabla de registros de clientes históricos Guayaquil.

Cliente	Código Cliente
Cachupud Carrillo Cirilo	4290
Ordoñez Armijos Edwin Geovanny	3698
Yaucan Paca Manuela	5351
Jara Cuzco Vicente Leonardo	1181
Sarango Ramírez Edwin Fernando	2377
Mayolema Paguay Aida Piedad	1118
López Maldonado José Rodrigo	1687
Reyes Buri German Enrique	4361
Chipantiza Punguil Zoila Rosa	367
Bastidas Arciniega Ángel	4394

Anexo 3

Tabla de registros de ventas de los clientes de Guayaquil, periodo 2017-2021

SKU	Factura	Fecha	Cliente	Venta Neta	Código Cliente
1151	28939	20/1/2017	Cachupud Carrillo Cirilo	18,94	4290
1151	29774	31/1/2017	Ordoñez Armijos Edwin Geovanny	18,94	3698
1151	47990	18/1/2018	Yaucan Paca Manuela	38,13	5351
1180005	48885	31/1/2018	Jara Cuzco Vicente Leonardo	55,79	1181
4020035	68282	17/1/2019	Sarango Ramírez Edwin Fernando	1,44	2377
80129	69352	31/1/2019	Mayolema Paguay Aida Piedad	17,12	1118
1180005	89384	17/1/2020	López Maldonado José Rodrigo	33,48	1687
12863	89987	31/1/2020	Reyes Buri German Enrique	277,6	4361
30228955	107045	13/1/2021	Chipantiza Punguil Zoila Rosa	608,6	367
1TTEC000149	108330	31/1/2021	Bastidas Arciniega Ángel	356,03	4394

Anexo 4

Tabla de registros de productos disponibles de la Empresa AAA

SKU	Descripción	Categoría	Unidades por Cajas
000147	P. Lasaña De Carne 220gr X 24und	Lasaña	24
001151	Jw. Jab.Macho Blanco Suav 220g X50	Jabón De Ropa	50
001223	Jw. Lava Antibacter Crema 250g X 36	Lavaplatos	36
001230	Jw. Lava Limón Crema 500g X 18	Lavaplatos	18

Anexo 5

Tabla de registros de ventas de clientes con productos disponibles dentro de la ciudad de Guayaquil preliminar periodo 2017-2021.

SKU	Factura	Fecha	Cliente	Venta Neta	Código Cliente	Producto
1151	28939	20/1/2017	Cachupud Carrillo Cirilo	18,94	4290	Jw.Jab.Macho Blanco Suav 220g X50
1151	29774	31/1/2017	Ordoñez Armijos Edwin Geovanny	18,94	3698	Jw.Jab.Macho Blanco Suav 220g X50

1151	47990	18/1/2018	Yaucan Paca Manuela	38,13	5351	Jw.Jab.Macho Blanco Suav 220g X50
1180005	48885	31/1/2018	Jara Cuzco Vicente Leonardo	55,79	1181	Hada.Ph Inst.Jumbo 180mtsx4r
12345	70938	28/2/2019	Villamar Laje Julio	2,07	3044	Ot.Toa Hum. Fisher Price 50x24pq
5063020/R67	110340	10/3/2021	Barre Mora Angela Tania	23,46	5152	Ot.Toa Hum. Fisher Price 50x24pq
1180005	89384	17/1/2020	López Maldonado José Rodrigo	33,48	1687	Hada.Ph Inst.Jumbo 180mtsx4r
12863	89987	31/1/2020	Reyes Buri German Enrique	277,6	4361	Ot.Jab.Family Avena 270grx24 Flowpack
3E+07	107045	13/1/2021	Chipantiza Punguil Zoila Rosa	608,6	367	Kim.Ph Flor Clásico 2pl 6x8pq
1TTEC000149	108330	31/1/2021	Bastidas Arciniega Ángel	356,03	4394	So.Toa Ladysoft Nocturna C/A 8x24pq

Anexo 6

Homologación de códigos de productos de la Empresa AAA.

SKU	SKU Homologado
12345	12345
5063020/R67	
12250	12250
5063031/Q20	
12863	12863
5051614/Q46	
12864	12864
5051642/Q46	
14144	14144
5063041/R63	
14153	14153
5063043/R63	

14154	14154
5063043/R96	
1TTEC000503	1TTEC000503
1TTEC000638	
1TTEC000549	1TTEC000549
1TTEC000632	
1TTEC000570	1TTEC000570
ET-0570	
1TTEC000610	1TTEC000610
1TTEC000682	
1TTEC000792	1TTEC000792
1TTEC001008	
5051348/G43	5051348/G43
5051348/S82	
5071077/B35	LF-0007
LF-0007	
5071077/B37	LF-053
LF-053	
5071081/B35	LF-045
LF-045	
FCO22514A	FCO22514A
FMX05334	

Anexo 7

Homologación de código de clientes de la Empresa AAA

Código Cliente	Código Homologado Cliente
4720	
4687	4720
6599	

5893	5893
6397	
6385	6385
2360	
6323	5080
5080	
6386	6386
2425	

Anexo 8

Muestra de los 769 clientes seleccionados para el estudio

Código Cliente
4720
7060
1665
2122
2289
2553
2630
3187
⋮
7200
7202
7205
7208
7212
7213
7215

Anexo 9

Muestra de los 783 productos seleccionados para el estudio

SKU
12345
12250
12863
12864
14144
14153
14154
1TTEC000503
⋮
001151
001223
001230
001255
001462
001471
001472

Anexo 10

Participación de la demanda por tipología del cliente, año 2021

Tipología	Venta Neta	% Participación 2021
Mayorista	3.081.198	80,93
Tienda	472.622	12,41
Abarrotes	207.493	5,45
Otros	40.355	1,06
Minimarkets	5.509	0,14
Total	3.807.177	100,00

Anexo 11

Demanda por tipología de clientes, 2019-2021

Tipología	2019	2020	2021	% Var. 2020-20219	% Var. 2021-2020	% Var. 2021-2019
Mayorista	3.835.312	2.982.758	3.081.198	-22,23	3,30	-19,66
Tienda	117.229	260.133	472.622	121,90	81,68	303,16
Abarrotes	221.797	218.199	207.493	-1,62	-4,91	-6,45
Otros	7.036	22.674	40.355	222,24	77,97	473,50
Minimarkets	2.041	2.336	5.509	14,45	135,90	169,97
Total	4.183.415	3.486.100	3.807.177	-16,67	9,21	-8,99

Anexo 12

Participación de la demanda por segmento de producto, 2021

Segmento	Venta Neta	% Participación 2021
Cuidado Oral	1.615.996	42,45
Cuidado Higiene	725.957	19,068113
Cuidado Personal	630.660	16,565039
Cuidado Hogar	500.691	13,15125
Alimentos	333.873	8,7695617
Total	3.807.177	100,00

Anexo 13

Demanda por Segmento de Productos, 2019-2021

Segmento	2019	2020	2021	% Variación 2020-20219	% Variación 2021-2020	% Variación 2021-2019
Cuidado Oral	1.411.757	1.492.834	1.615.996	5,74	8,25	14,47
Cuidado Higiene	644.540	207.646	725.957	-67,78	249,61	12,63
Cuidado Personal	1.079.692	869.156	630.660	-19,50	-27,44	-41,59
Cuidado Hogar	668.237	638.461	500.691	-4,46	-21,58	-25,07
Alimentos	379.188	278.004	333.873	-26,68	20,10	-11,95
Total	4.183.415	3.486.100	3.807.177	-16,67	9,21	-8,99

Anexo 14

Otras categorías de productos, 2021

Segmento	Categoría
Alimentos	Aderezos
Alimentos	Arroz
Alimentos	Atún
Alimentos	Azúcar
Alimentos	Café
Alimentos	Caramelo
Alimentos	Conservas
Alimentos	Harina
Alimentos	Leche
Alimentos	Manteca
Alimentos	Margarina
Alimentos	Tapioca
Cuidado Hogar	Alcohol
Cuidado Hogar	Alimento Mascotas
Cuidado Hogar	Ambiental
Cuidado Hogar	Betún
Cuidado Hogar	Bolígrafos
Cuidado Hogar	Cepillo De Ropa
Cuidado Hogar	Cera
Cuidado Hogar	Cloro
Cuidado Hogar	Desinfectante
Cuidado Hogar	Encendedor
Cuidado Hogar	Escoba
Cuidado Hogar	Esponja
Cuidado Hogar	Foco
Cuidado Hogar	Funda
Cuidado Hogar	Guantes

Cuidado Hogar	Insecticida
Cuidado Hogar	Jabón De Ropa
Cuidado Hogar	Jabón Liquido
Cuidado Hogar	Lavaplatos
Cuidado Hogar	Limpieza
Cuidado Hogar	Mascarillas
Cuidado Hogar	Paños Limpieza
Cuidado Hogar	Papel Aluminio
Cuidado Hogar	Pastilla Ambiental
Cuidado Hogar	Pastilla Baño
Cuidado Hogar	Pegamento
Cuidado Hogar	Pinzas
Cuidado Hogar	Recogedor
Cuidado Hogar	Servilletas
Cuidado Hogar	Suavizante
Cuidado Oral	Cepillo Dental
Cuidado Personal	Acondicionador
Cuidado Personal	Afeitadoras
Cuidado Personal	Crema Corporal
Cuidado Personal	Crema Peinar
Cuidado Personal	Desodorante
Cuidado Personal	Pañales
Cuidado Personal	Shampoo

Cuidado Personal	Toallas Húmedas
Cuidado Personal	Toallas Sanitarias
Cuidado Personal	Vaporal

Anexo 15

Participación de la demanda por categoría de producto, año 2021

Categoría Producto	2021	% Participación 2021
Crema Dental	1.575.210	41,37
Papel Higiénico	725.957	19,07
Jabón de Tocador	380.374	9,99
Aceite	266.989	7,01
Detergente	161.752	4,25
Otras	696.895	18,30
Total	3.807.177	100,00

Anexo 16

Demanda de las 5 categorías de productos más representativas, 2019-2021

Categoría Producto	2019	2020	2021	% Variación 2020-20219	% Variación 2021-2020	% Variación 2021-2019
Crema Dental	1.364.156	1.435.123	1.575.210	5,20	9,76	15,47
Papel Higiénico	644.540	207.646	725.957	-67,78	249,61	12,63
Jabón de Tocador	380.000	471.165	380.374	23,99	-19,27	0,10

Aceite	327.599	198.008	266.989	-39,56	34,84	-18,50
Detergente	237.982	225.773	161.752	-5,13	-28,36	-32,03
Otras	1.229.139	948.386	696.895	-22,84	-26,52	-43,30
Total	4.183.415	3.486.100	3.807.177	-16,67	9,21	-8,99

Anexo 17

Estadísticas descriptivas de la variable Frecuencia de compra según la Tipología de Cliente.

Frecuencia	n	X	Me	Mo	**SD**	Min	Max	R	Q3
Total	769	10	8	1,00	13,00	1	280	279	14
Mayorista	213	12	6	1,00	22,00	1	280	279	14
Tienda	352	11	13	14,00	8,00	1	93	92	15
Abarrotes	167	7	5	1,00	6,00	1	31	30	8
Otros	32	6	4	1,00	6,00	1	44	43	7
Minimarkets	5	8	2	2,00	14,00	1	34	33	19

Nota: n=muestra, X= media aritmética, Me=mediana, Mo=moda, SD= desviación estándar, Min= mínimo, Max= máximo, R= rango, Q3= cuartil 3

Anexo 18

Evolución de la suma de cuadrados dentro-clusters usando el método de codo para la primera clusterización.

Métodos	Clusters									
K	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
K-means	1,91E+10	4,74 E+09	2,01 E+09	1,07 E+09	7,21 E+08	5,37 E+08	4,70 E+08	3,97 E+08	3,30 E+08	1,75 E+08
K-medoids	2.300	1.271	933	669	541	444	399	347	318	283

Anexo 19

Evolución de la suma de cuadrados dentro-cluster, utilizando el método del codo, para la segunda clusterización.

Métodos	Clusters									
K	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
K-means	3,17E+12	6,02E+11	1,30E+11	5,36E+10	3,57E+10	2,76E+10	2,43E+10	2,34E+10	2,28E+10	2,21E+10
K-medoids	9.314	6.302	4.433	3.178	2.255	1.915	1.625	1.307	1.040	936

Anexo 20

Considerando a la variable dependiente Cluster: A=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	21	0,0003	SI
TipClte	chi-Sq	4	22	0,0002	SI
Zona	chi-Sq	4	27	0,0249	SI
NatClte	chi-Sq	2	63	0,0429	SI
MontComp	F test	1.766	2.611	0,0000	SI
Frecuencia	F test	1.766	16	0,0001	SI
Recencia	F test	1.766	1	0,4320	NO
Portaf	F test	1.766	0	0,8020	NO
TktProm	F test	1.766	135	0,0000	SI
Antigüedad	F test	1.766	7	0,0085	SI
FrecAnualUltAnoComp	F test	1.766	2	0,1370	NO
FrecPromMesUltAnoComp	F test	1.766	52	0,0000	SI

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 21

Considerando a la variable dependiente Cluster: C=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	2	0,7292	NO
TipClte	chi-Sq	4	33	0,0010	SI
Zona	chi-Sq	4	32	0,0018	SI
NatClte	chi-Sq	2	0	0,8780	NO
MontComp	F test	1.766	91	0,0000	SI
Frecuencia	F test	1.766	83	0,0000	SI
Recencia	F test	1.766	1	0,2400	NO
Portaf	F test	1.766	21	0,0000	SI
TktProm	F test	1.766	150	0,0000	SI
Antigüedad	F test	1.766	13	0,0003	SI
FrecAnualUltAnoComp	F test	1.766	2	0,1300	NO
FrecPromMesUltAnoComp	F test	1.766	26	0,0000	SI

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 22

Considerando a la variable dependiente Cluster: D=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	14	0,0082	SI
TipClte	chi-Sq	4	35	0,0004	SI
Zona	chi-Sq	4	18	0,0012	SI
NatClte	chi-Sq	2	72	0,0269	SI
MontComp	F test	1.766	5	0,0204	SI
Frecuencia	F test	1.766	9	0,0036	SI
Recencia	F test	1.766	2	0,137	NO
Portaf	F test	1.766	39	0	SI
TktProm	F test	1.766	25	0	SI
Antigüedad	F test	1.766	8	0,0043	SI
FrecAnualUltAnoComp	F test	1.766	3	0,0727	NO
FrecPromMesUltAnoComp	F test	1.766	1	0,364	NO

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 23

Considerando a la variable dependiente Cluster: E=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	31	0,0028	SI
TipClte	chi-Sq	4	62	0	SI
Zona	chi-Sq	4	28	0,0108	SI
NatClte	chi-Sq	2	4	0,1667	NO
MontComp	F test	1.766	1	0,352	NO
Frecuencia	F test	1.766	2	0,193	NO
Recencia	F test	1.766	2	0,187	NO
Portaf	F test	1.766	21	0	SI
TktProm	F test	1.766	36	0	SI
Antigüedad	F test	1.766	0	0,899	NO
FrecAnualUltAnoComp	F test	1.766	0	0,508	NO
FrecPromMesUltAnoComp	F test	1.766	8	0,0041	SI

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 24

Considerando a la variable dependiente Cluster: F=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	72	0	SI
TipClte	chi-Sq	4	52	0	SI
Zona	chi-Sq	4	24	0,0765	SI
NatClte	chi-Sq	2	13	0,0012	SI
MontComp	F test	1.766	1	0,375	NO
Frecuencia	F test	1.766	2	0,152	NO
Recencia	F test	1.766	0	0,497	NO
Portaf	F test	1.766	15	0,0001	SI
TktProm	F test	1.766	14	0,0002	SI
Antigüedad	F test	1.766	5	0,0269	SI
FrecAnualUltAnoComp	F test	1.766	7	0,0083	NO
FrecPromMesUltAnoComp	F test	1.766	0	0,859	NO

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 25

Considerando a la variable dependiente Cluster: G=1, Resto=0

Variable	Test	df	Est.	Prob.	Significancia
SegmComp	chi-Sq	4	95	0	SI
TipClte	chi-Sq	4	221	0	SI
Zona	chi-Sq	4	100	0	SI
NatClte	chi-Sq	2	21	0,023	SI
MontComp	F test	1.766	144	0	SI
Frecuencia	F test	1.766	28	0	SI
Recencia	F test	1.766	6	0,0135	SI
Portaf	F test	1.766	97	0	SI
TktProm	F test	1.766	330	0	SI
Antigüedad	F test	1.766	25	0	SI
FrecAnualUltAnoComp	F test	1.766	1	0,474	NO
FrecPromMesUltAnoComp	F test	1.766	37	0	SI

df=grados de libertad, Est=Estadístico, Prob=Probabilidad

Anexo 26

Participación de la Naturaleza del cliente del Cluster A

Naturaleza del Cliente	Número Cliente	% Participación
M	8	73%
H	2	18%
PJ	1	9%
Total	11	100%

Anexo 27

Participación de la Tipología, Naturaleza y Segmento de Compra del Cluster C

Tipología Cliente	Número Cliente	%Participación
MY	18	82%
TD	3	14%
AB	1	5%
Total	22	100%
	(a)	

Naturaleza del Cliente	Número Cliente	% Participación
H	13	59%
M	8	36%
PJ	1	5%
Total	22	100%
	(b)	

Segmento de Compra	Número Cliente	%Participación
CO	7	32%
CP	5	23%
CH	4	18%
A	4	18%
CHO	2	9%
Total	22	100%
	(c)	

Anexo 28

Participación de la Tipología, y Naturaleza del cliente del Cluster D

Tipología de Cliente	Número de Cliente	% Participación
MY	35	74%
AB	8	17%
OT	3	6%
TD	1	2%
Total	47	100%
	(a)	

Naturaleza del Cliente	Número Cliente	% Participación
H	26	55%
M	17	36%
PJ	4	9%
Total	47	100%
	(b)	

Anexo 29

Participación de la Tipología, y Naturaleza del cliente del Cluster E

Tipología de Cliente	Número Cliente	% Participación
MY	17	85%
AB	3	15%
Total	20	100%

(a)

Naturaleza del Cliente	Número Cliente	% Participación
H	15	75%
M	3	15%
PJ	2	10%
Total	20	100%

(b)

Anexo 30

Participación de la Tipología, y Naturaleza del cliente del Cluster F

Tipología de Cliente	Número de Cliente	% Participación
MY	46	54%
AB	25	29%
TD	10	12%
OT	3	4%
MK	1	1%
Total	85	100%

(a)

Naturaleza del Cliente	Número Cliente	% Participación
H	41	48%
M	35	41%
PJ	9	11%
Total	85	100%

(b)

Anexo 31

Participación de la Tipología y Naturaleza del cliente del Cluster G

Tipología de Cliente	Número de Cliente	% Participación
TD	337	58%
AB	130	22%
MY	86	15%
OT	26	4%
MK	4	1%
Total	583	100%

(a)

Naturaleza del Cliente	Número Cliente	% Participación
H	325	56%
M	247	42%
PJ	11	2%
Total	583	100%

(b)

Anexo 32

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Crema Dental del segmento de clientes A

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	568,25	99,04%	Mejor
	5	928,96	97,78%	
	10	1.038,92	97,49%	
	30	1.852,71	94,86%	
	100	3.702,49	82,46%	
Árboles de regresión	3	1.615,78	93,09%	
	5	1.759,58	93,16%	Mejor
	10	1.759,58	93,16%	
	20	1.759,58	93,16%	

Anexo 33

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Crema Dental del segmento de clientes B

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	387,59	93,93%	Mejor
	5	489,95	91,06%	
	10	735,25	82,36%	
	30	1.127,61	67,92%	
	100	1.503,75	45,30%	
Árboles de regresión	1	952,97	69,40%	
	5	411,19	94,14%	Mejor
	10	411,19	94,14%	
	20	411,19	94,14%	

Anexo 34

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Toallas Húmedas del segmento de clientes B

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	379,42	93,62%	Mejor
	5	479,29	91,04%	
	10	560,59	88,37%	
	30	771,07	81,21%	
	100	1.124,88	66,23%	
Árboles de regresión	1	867,82	64,63%	
	5	656,08	80,24%	
	10	557,47	85,17%	Mejor
	20	557,47	85,17%	

Anexo 35

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Jabón de Tocador para el segmento de clientes B

		Entrenamiento		
Modelo	**K**	**RMSE**	$\mathbf{R^2}$	**Mejor ajuste**
KNN	1	240,65	93,32%	
	5	277,27	93,41%	Mejor
	10	350,28	89,32%	
	30	431,49	84,71%	
	100	611,96	80,28%	
Árboles de regresión	1	602,67	52,61%	
	5	372,53	82,02%	Mejor
	10	372,82	82,88%	
	20	372,82	82,88%	

Anexo 36

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Crema Dental del segmento de clientes C

		Entrenamiento		
Modelo	**K**	**RMSE**	$\mathbf{R^2}$	**Mejor ajuste**
KNN	1	715,59	98,72%	
	3	887,18	98,29%	Mejor
	5	952,81	98,19%	
	10	1.308,53	96,38%	
	30	2.254,04	94,21%	
Árboles de regresión	1	2.940,41	80,54%	
	3	1.522,95	94,45%	Mejor
	5	1.536,18	94,22%	
	10	1.536,18	94,22%	

Anexo 37

Elección del mejor modelo de KNN y Árboles de regresión para la categoría de Jabón de Tocador del segmento de clientes C

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	493,02	69,00%	
	5	665,10	63,85%	Mejor
	10	717,11	56,88%	
	30	815,29	42,59%	
	100	882,32	18,62%	
Árboles de regresión	1	823,65	34,69%	
	5	674,29	57,70%	Mejor
	10	674,29	57,70%	
	20	674,29	57,70%	

Anexo 38

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Papel Higiénico del segmento de clientes C

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	281,74	93,72%	Mejor
	5	401,38	88,91%	
	10	525,29	83,16%	
	30	804,43	73,10%	
	100	1.230,52	52,27%	
Árboles de regresión	1	1.018,07	36,79%	
	3	763,27	62,57%	
	5	715,15	67,86%	
	10	608,05	77,52%	Mejor
	20	608,05	77,52%	

Anexo 39

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Aceite del segmento de clientes C

		Entrenamiento		
Modelo	**K**	**RMSE**	$\mathbf{R^2}$	**Mejor ajuste**
KNN	1	547,87	86,46%	Mejor
	5	651,72	90,21%	
	10	820,33	90,54%	
	30	1.141,99	88,31%	
	100	1.414,84	74,99%	
Árboles de regresión	1	1.188,15	63,89%	
	3	830,98	82,96%	Mejor
	5	873,76	82,10%	
	10	873,76	82,10%	
	20	873,76	82,10%	

Anexo 40

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Papel Higiénico del segmento de clientes D

		Entrenamiento		
Modelo	**K**	**RMSE**	$\mathbf{R^2}$	**Mejor ajuste**
KNN	1	231,51	94,41%	Mejor
	5	310,19	93,23%	
	10	397,44	91,66%	
	30	733,53	77,92%	
	100	1.002,97	71,14%	
Árboles de regresión	1	735,13	58,99%	
	5	515,54	71,38%	Mejor
	10	461,68	77,38%	
	20	461,68	77,38%	

Anexo 41

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Aceite del segmento de clientes D

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	1	418,00	75,22%	
	5	430,39	87,08%	
	10	486,27	88,05%	Mejor
	30	640,82	78,19%	
	100	753,30	56,96%	
Árboles de regresión	1	590,91	70,78%	
	3	333,34	84,44%	Mejor
	5	426,72	80,90%	
	10	426,72	80,90%	
	20	426,72	80,90%	

Anexo 42

Elección del mejor modelo de KNN y Árboles de regresión para la categoría Jabón de Tocador del segmento de clientes D

		Entrenamiento		
Modelo	**K**	**RMSE**	**R^2**	**Mejor ajuste**
KNN	5	254,35	84,96%	Mejor
	7	276,50	83,37%	
	9	285,27	83,88%	
Árboles de regresión	1	359,33	59,92%	
	3	256,84	81,22%	Mejor
	5	284,17	75,88%	
	10	284,17	75,88%	
	20	284,17	75,88%	

Anexo 43

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Crema Dental del segmento de clientes A

Muestra	Número de Orden	Venta Real	Venta Estimada	MAPE
1	115	2.562	2.564	0,1%
2	11	21.286	21.062	1,1%
3	6	4.041	4.301	6,4%
4	11	21.286	21.062	1,1%
5	39	33.005	30.738	6,87%
6	118	1.357	1.121	17,4%
7	21	235	328	39,2%
8	104	7.039	7.101	0,9%
9	78	5.266	5.240	0,5%
10	2	956	956	0,0%

Nota: El número total del conjunto de prueba fue de 126

Anexo 44

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Crema Dental del segmento de clientes B

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	12	1.644	1.252	23,86%
2	24	3.635	3.700	1,78%
3	65	4.365	4.361	0,07%
4	8	4.302	4.417	2,68%
5	4	469	466	0,63%
6	25	4.015	4.072	1,42%
7	48	4.508	4.436	1,61%
8	96	4.406	4.239	3,79%
9	28	4.035	4.139	2,56%
10	68	4.015	4.072	1,42%

Nota: El número total del conjunto de prueba fue de 100

Anexo 45

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Toallas Húmedas del segmento de clientes B

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	28	338	342	1,28%
2	14	4.286	4.253	0,78%
3	15	1.837	1.888	2,81%
4	51	857	897	4,66%
5	13	4.286	4.252	0,80%
6	40	3.696	3.697	0,02%
7	10	3.429	3.388	1,19%
8	49	357	398	11,53%
9	43	634	666	4,92%
10	56	1.715	1.700	0,86%

Nota: El número total del conjunto de prueba fue de 60

Anexo 46

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Jabón de Tocador del segmento de clientes B

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	30	860	862	0,28%
2	24	1.404	1.429	1,79%
3	44	2.870	2.723	5,10%
4	19	358	372	3,76%
5	12	430	440	2,45%
6	29	2.337	2.302	1,50%
7	6	430	437	1,73%
8	46	716	722	0,83%
9	31	431	395	8,44%
10	43	1.916	1.944	1,51%

Nota: El número total del conjunto de prueba fue de 62

Anexo 47

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Crema Dental del segmento de clientes C

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	21	9.035	9.026	0,10%
2	43	40	39	1,17%
3	51	394	576	46,04%
4	24	538	467	13,25%
5	45	9.559	9.000	5,86%
6	12	21.651	23.149	6,92%
7	53	9.559	9.000	5,86%
8	26	20.835	22.456	7,78%
9	36	40	39	1,17%
10	6	23.913	23.802	0,46%

Nota: El número total del conjunto de prueba fue de 55

Anexo 48

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Jabón de Tocador del segmento de clientes C

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	30	151	153	1,24%
2	110	31	31	0,38%
3	10	77	98	27,90%
4	27	492	446	9,34%
5	32	1.491	1.504	0,90%
6	62	230	246	6,92%
7	67	757	725	4,25%
8	103	204	227	10,96%
9	28	70	91	29,18%
10	71	151	128	15,28%

Nota: El número total del conjunto de prueba fue de 111

Anexo 49

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Papel Higiénico del segmento de clientes C

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	13	183	190	3,84%
2	55	1.230	1.242	0,98%
3	51	365	375	2,80%
4	6	705	712	1,03%
5	26	615	610	0,77%
6	72	122	139	14,50%
7	21	6.181	6.138	0,70%
8	3	2.965	2.886	2,66%
9	22	620	624	0,64%
10	31	78	81	3,64%

Nota: El número total del conjunto de prueba fue de 81

Anexo 50

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Aceite del segmento de clientes C

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	29	50	53	6,61%
2	30	180	81	54,84%
3	40	1.130	1.583	40,04%
4	39	126	33	74,21%
5	49	1.179	1.202	1,92%
6	43	2.822	2.861	1,37%
7	32	132	114	13,90%
8	3	804	836	3,96%
9	6	213	262	23,21%
10	28	47	94	100,97%

Nota: El número total del conjunto de prueba fue de 54

Anexo 51

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Papel Higiénico del segmento de clientes D

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	6	619	624	0,81%
2	33	548	538	1,83%
3	57	304	287	5,84%
4	49	310	321	3,82%
5	17	586	594	1,40%
6	45	56	77	37,59%
7	15	1.833	1.883	2,74%
8	7	1.871	1.898	1,44%
9	40	251	250	0,25%
10	25	276	268	3,20%

Nota: El número total del conjunto de prueba fue de 62

Anexo 52

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Aceite del segmento de clientes D

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	19	187	141	24,72%
2	25	632	462	26,99%
3	22	2.107	2.217	5,19%
4	14	146	200	37,23%
5	34	932	963	3,23%
6	9	1.028	1.047	1,77%
7	8	664	637	4,14%
8	37	1.421	1.307	8,05%
9	1	538	574	6,65%
10	16	594	605	1,94%

Nota: El número total del conjunto de prueba fue de 52

Anexo 53

Muestra de 10 registros de órdenes de compras seleccionadas de forma aleatoria del conjunto de prueba de la categoría Jabón de Tocador del segmento de clientes D

Muestra	Número de orden	Venta Real	Venta Estimada	MAPE
1	3	53	74	39,74%
2	37	48	74	52,82%
3	29	1.386	1.968	42,06%
4	48	144	74	48,75%
5	46	84	74	12,45%
6	25	117	233	99,71%
7	23	481	460	4,44%
8	47	25	74	190,91%
9	31	241	233	3,12%
10	39	58	74	26,18%

Nota: El número total del conjunto de prueba fue de 49