DIVISIÓN DE POSGRADO

Título del trabajo:

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING EN EL DESARROLLO DE UN SISTEMA DE APOYO EN LA TOMA DE DECISIONES PARA LA SEGMENTACIÓN DE CLIENTES EN UNA CADENA DE CAFETERÍAS

Línea de Investigación:

TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

Modalidad de titulación:

PROYECTO DE DESARROLLO

MAESTRÍA EN SISTEMAS DE INFORMACIÓN
MENCIÓN EN INTELIGENCIA DE NEGOCIOS

Título a obtener:

MASTER EN SISTEMAS DE INFORMACIÓN CON MENCIÓN EN INTELIGENCIA DE NEGOCIOS

Autor:

Johnny Gustavo Idrovo Granizo

Tutor:

Mgtr. Pedro Fabricio Echeverría Briones

Samborondón – Ecuador
2019

TRIBUNAL DE SUSTENTACION

DEDICATORIA

La presente investigación está dedicada a DIOS por todas las bendiciones recibidas y que sin su gracia no hubiera alcanzado esta meta, a mi querida madre la SRA. YOLANDA GRANIZO que siempre ha sido mi apoyo y sustento espiritual y a la que estoy eternamente agradecido por lo que soy.

Ing. Johnny Idrovo Granizo

AGRADECIMIENTOS

Agradezco a mi Dios por sobre todas las cosas, por la salud y trabajo. A mi madre y mis seres queridos por haberme comprendido el tiempo que no pude compartir con ellos. A todos mis profesores por sus excelentes enseñanzas que enriquecieron la carrera.

Ing. Johnny Idrovo Granizo

Universidad Ecotec

TRABAJO DE TITULACIÓN

CERTIFICACION DE REVISION FINAL

QUE EL PRESENTE PROYECTO DE DESARROLLO TITULADO:

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING EN EL DESARROLLO DE UN SISTEMA DE APOYO EN LA TOMA DE DECISIONES PARA LA SEGMENTACIÓN DE CLIENTES EN UNA CADENA DE CAFETERÍAS,

FUE REVISADO, SIENDO SU CONTENIDO ORIGINAL EN SU TOTALIDAD, ASÍ COMO EL CUMPLIMIENTO DE LOS REQUERIMIENTOS QUE SE DICTAN EN LA GUÍA DE ELABORACIÓN DEL TRABAJO DE TITULACIÓN, POR LO QUE SE AUTORIZA A: JOHNNY GUSTAVO IDROVO GRANIZO 0915216584**, QUE PROCEDA A SU PRESENTACION.**

Samborondón, 24 de Julio de 2019

TUTOR: Pedro Fabricio Echeverría Briones

Universidad Ecotec

DIVISIÓN DE POSGRADO
PROCESO DE TITULACIÓN
CERTIFICADO DEL PORCENTAJE DE COINCIDENCIAS

Habiendo sido nombrado PEDRO FABRICIO ECHEVERRIA BRIONES tutor del trabajo de titulación " APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING EN EL DESARROLLO DE UN SISTEMA DE APOYO EN LA TOMA DE DECISIONES PARA LA SEGMENTACIÓN DE CLIENTES EN UNA CADENA DE CAFETERÍAS" elaborado por JOHNNY GUSTAVO IDROVO GRANIZO, 0915216584, con mi respectiva supervisión como requerimiento parcial para la obtención del título de **Maestría en Sistemas de Información Mención Inteligencia de Negocios**.

Se informa que el mismo ha resultado tener un porcentaje de coincidencias (10 %) mismo que se puede verificar en el siguiente link:
https://secure.urkund.com/view/53127284-433756-491207#DcI7DsIwEEXRvbi+QvPGn7GzFUSBoiRykTQpEXuHo/NJ552Wp6F/IUcZVdRQoI4GbnjGC17xhgeFSiMYL9I9j2vuc31f65YWe9jI5rm30SNKdyvfHw==

Adicional se adjunta print de pantalla de dicho resultado.

URKUND

Documento	JOHNNY GUSTAVO IDROVO GRANIZO.docx (D54561252)
Presentado	2019-07-24 18:49 (-05:00)
Presentado por	joidrovo@mgs.ecotec.edu.ec
Recibido	nvillena.ecotec@analysis.urkund.com
Mensaje	Fwd: Tesis Mostrar el mensaje completo
	10% de estas 21 páginas, se componen de texto presente en 9 fuentes.

Pedro Fabricio Echeverría Briones
PROFESOR TUTOR DE LA CIENCIA

RESUMEN

El aprendizaje de máquina no supervisado es el campo de *machine learning* que demanda mayor esfuerzo para los científicos de datos y una amplia aplicación para distintas áreas de negocio. En la actual propuesta de desarrollo se analizan los algoritmos más utilizados para la segmentación de clientes desde un enfoque de negocio para los administradores de los departamentos de marketing en empresas de alimentos y bebidas en especial las cafeterías que poseen gran cantidad de clientes.

El agrupamiento o clustering es la técnica de *machine learning* utilizada para este tipo de análisis. Esta técnica se basa en analizar los datos para encontrar similitudes y agruparlos en segmentos o cluster con características similares entre sí y diferentes con respecto a los otros segmentos.

Siguiendo la metodología CRISP-DM para desarrollar el modelo de *machine learning* en base al análisis RFM (Recencia, Frecuencia, Monto de consumo), se aplicaron los algoritmos jerárquicos, mapas auto organizados basado en redes neuronales y *k-means*. En todas las etapas de desarrollo del Trabajo de Titulación se ha utilizado la herramienta RStudio de libre distribución. Para validar el resultado de los segmentos encontrados se ha aplicado los dos índices de validación de agrupamiento comúnmente utilizados: el ancho de la silueta y el índice Dunn.

Palabras clave: Segmentación de Clientes, Técnicas de Aprendizaje de Máquina, CRISP-DM, Análisis de Agrupamiento, Cadenas de Cafeterías.

ABSTRACT

Unsupervised machine learning is the field of machine learning that demands more effort for data scientists and a wide application for different business areas. The current development proposal analyzes the most used algorithms for customer segmentation from a business approach for the administrators of the marketing departments in food and beverage companies, especially coffee shops that have a large number of customers.

Clustering is the machine learning technique used for this type of analysis. This technique is based on analyzing the data to find similarities and grouping them into segments or clusters with similar characteristics to each other and different with respect to the other segments.

Following the CRISP-DM methodology to develop the machine learning model based on the RFM analysis (Reception, Frequency, Amount of consumption), the hierarchical algorithms, self-organized maps based on neural networks and k-means were applied. In all stages of development of the Degree Work, the RStudio tool of free distribution has been used. To validate the result of the segments found, the two commonly used clustering validation indices have been applied: the width of the silhouette and the Dunn index.

Keywords: Customer Segmentation, Machine Learning Techniques, CRISP-DM, Clustering Analysis, Coffee Shop Chains.

Índice de contenido

INTRODUCCIÓN 14
I. Antecedentes 15
II. Descripción del problema 16
III. Pregunta problémica 17
IV. Idea a defender 17
V. Objetivo General 17
VI. Objetivos Específicos 17
VII. Justificación 18
VIII. Procedimiento metodológico 20
IX. Método científico 20
CAPÍTULO 1 22
MARCO TEÓRICO 22
1.1. Segmentación 22
1.1.1. Criterios de segmentación. 22
1.2. Segmentación RFM 23
1.3. Machine Learning 23
1.3.1. Aprendizaje supervisado. 24
1.3.2. Aprendizaje no supervisado. 24
1.3.3. Diferencias entre el *data mining*, el *machine learning* y el *deep learning*. 25
1.4. Funcionamiento de *Machine Learning* 25
1.4.1. Pasos usados en *Machine Learning*. 26
1.5. Algoritmos de segmentación o de agrupamiento 28
1.5.1. Posibles aplicaciones en problemas reales. 28
1.5.2. Cluster jerárquico 29
1.5.3. K-means clustering. 29
1.5.4. Particionamiento alrededor de medoides (PAM). 31
1.5.5. Mapa de Kohonen (SOM). 32
1.5.6. Algoritmo Two-step. 32
1.6. Etapas a seguir en el desarrollo del Análisis de Cluster. 32
1.7. Determinación del número k de clústeres 33
1.8. Marco Metodológico 34
1.9. Sustento Legal 36
CAPÍTULO 2 37
SOLUCIÓN PROPUESTA 37
2.1. Comprensión del negocio 37
2.1.1. Objetivos del Negocio. 37
2.1.2. Entendimiento de los objetivos. 38
2.1.3. Elaborar plan del proyecto. 38
2.2. Entendimiento de datos 39
2.2.1. Verificación de la calidad de los datos. 40
2.2.2. Exploración de los datos. 41

2.3. Preparación de datos....41
2.3.1. Selección de los datos. 41
2.3.2. Construcción de nuevos datos.....43
2.3.3. Integración de datos. 44
2.3.4. Formato de datos.....45
2.4. Modelado....45
2.4.1. Elección de la técnica de modelado. 45
2.4.2. Generación de un diseño de comprobación. 46
2.4.3. Generación de los modelos. 46
2.4.4. Evaluación del modelo.....52
2.5. Evaluación....52
2.5.1. Evaluación de los clústeres. 52
2.5.2. Revisión del proceso. 53
2.5.3. Determinación de los próximos pasos. 54
2.6. Despliegue....54
2.6.1. Planificación de la Implantación.....54
2.7. Definición de requerimientos funcionales y no funcionales de la solución a desarrollar 54
CAPÍTULO 3....56
VALIDACIÓN DE LOS RESULTADOS....56
3.1. Esquema de validación....56
3.2. Aplicación y descripción de los resultados 56
3.3. Impacto....59
3.3.1. Perspectiva tecnológica.....59
3.3.2. Perspectiva económica.....59
CONCLUSIONES 60
RECOMENDACIONES....62
BIBLIOGRAFIA 63
ANEXOS 69
Anexo 1: Historias de Usuario....69
Anexo 2. Análisis estadístico mediante histogramas y diagramas de cajas de las variables RFM 70
Anexo 3. Resultados y comparativa de distintos métodos de agrupamiento jerárquico y sus dendogramas. 75
Anexo 4. Visualización de clústers con el método de agrupamiento jerárquico. 77
Anexo 5. Calcular la distancia al punto cero. 80
Anexo 6. Determinar los parámetros del mapa auto-organizado de Kohonen 81
Anexo 7. Visualización de los mapas auto-organizados mediante distintos gráficos.....82

Índice de tablas

Tabla 1 Plan del Proyecto 39

Tabla 2 Tabla Clientes 39

Tabla 3 Tabla Facturas 40

Tabla 4 Tamaño inicial de los datos seleccionados 42

Tabla 5 Resumen de los datos seleccioados.....42

Tabla 6 Resumen de los Atributos Seleccionados 43

Tabla 7 Estadística de los datos seleccionados 44

Tabla 8 Resultado de 7 clúster creados por el algoritmo Jerárquico.....48

Tabla 9 Expertos evaluados 56

Tabla 10 Preguntas para los expertos.....57

Tabla 11 Escala de valoración.....57

Tabla 12 Respuestas dadas por los expertos por cada aspecto relevante.....58

Índice de figuras

Figura 1 Pasos usados en machine learning. Fuente: elaboración propia 26

Figura 2 Ejemplo de agrupamiento con el algoritmo K-means. 31

Figura 3 El ciclo vital de CRISP-DM 34

Figura 4 Dendograma del método jerárquico separación de segmentos. 47

Figura 5 Número óptimo de clúster usando el método de suma de cuadrados. 50

Figura 6 Mapa de clústers usando algoritmo k-means. Fuente: elaboración propia. 51

Figura 7 Evaluación de estabilidad de clústeres. Fuente: elaboración propia. 53

Figura 8 Índice Porcentual de Concordancia de los expertos. 58

Figura 9 Histograma de frecuencia de consumo por Cliente. 71

Figura 10 Histograma de recencia por cliente. Fuente: elaboración propia. 72

Figura 11 Histograma de promedio de consumo por cliente. Fuente: elaboración propia. 73

Figura 12 Histograma promedio de consumo por cliente con transformación logarítmica. Fuente: elaboración propia. 74

Figura 13 Visualización de valores atípicos de las variables RFM. 75

Figura 14 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace completo. Fuente: elaboración propia. 76

Figura 15 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace promedio. Fuente: elaboración propia. 76

Figura 16 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace Ward.D2. Fuente: elaboración propia. 77

Figura 17 Visualización de clústeres promedio de consumo vs frecuencia. 78

Figura 18 Visualización de clústeres promedio consumo vs recencia. 79

Figura 19 Visualización de clústeres usando los primeros dos componentes principales. Fuente: elaboración propia. 80

Figura 20. Fórmula para calcular la distancia al punto cero (Ching-Hsue & You-Shyang, 2009) 81

Figura 21 Progreso de entrenamiento para encontrar los clústeres. 83

Figura 22 Mapa de calor de la variable recencia. Fuente: elaboración propia.....84

Figura 23 Mapa de calor de la variable frecuencia. Fuente: elaboración propia. 84

Figura 24 Mapa de calor de la variable promedio de consumo. 85

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING EN EL DESARROLLO DE UN SISTEMA DE APOYO EN LA TOMA DE DECISIONES PARA LA SEGMENTACIÓN DE CLIENTES EN UNA CADENA DE CAFETERÍAS

INTRODUCCIÓN

Según un post en las redes, Netflix no mató a Blockbuster sino la excesiva mora que éste cobraba. Uber no mató el negocio de taxis, el acceso limitado, el mal servicio y el control de las tarifas lo hicieron. Apple no mató a la industria de la música, sino el forzar a comprar álbumes completos. Amazon no mató a otros minoristas, sino el mal servicio y experiencia del cliente fue el que lo hizo. La tecnología por sí misma no es la disruptora. No estar centrado en el cliente es la mayor amenaza a cualquier negocio (Boza, 2019).

Y de qué manera las compañías pueden centrarse en el cliente si cada vez existen más canales de interacción, más registros. A quiénes poner más énfasis, cómo tomar la decisión si los presupuestos son limitados y se debe ser certero en las decisiones de marketing y ventas. Como conclusión, es primordial utilizar los datos para generar mayor competitividad en los negocios.

La presente investigación de desarrollo tecnológico pretende demostrar una alternativa de segmentación de clientes usando algoritmos de *machine learning*. El agrupamiento o clustering representa una de las técnicas no supervisadas de *machine learning* más utilizada para este tipo de análisis, es la técnica de agrupar un conjunto de objetos tales que los objetos en el mismo grupo (clúster) son más similares entre sí que a los de otros grupos.

Esta herramienta puede ser de mucha ayuda para los administradores gerenciales de marketing en ofrecer un apoyo para la toma de decisiones a través de promociones o nuevos productos de una manera personalizada o para cada perfil de clientes.

Además, con los algoritmos de *machine learning* se pueden generar modelos para predecir futuras tendencias y comportamientos de los clientes, utilizando información conocida en la empresa como características demográficas, comportamiento de consumo, facturaciones, entre otros.

En el ámbito académico, la realización del presente proyecto permitirá adquirir nuevas habilidades y conocimientos, que pueden ser aplicados para dar solución a otros problemas de este tipo.

I. Antecedentes

Los métodos tradicionales en el marketing *Business to Consumer* (B2C) segmentan a los clientes de acuerdo a características demográficas como edad, género, estado civil, ubicación u otros (Shopify, s.f.).

Los sistemas de relación de clientes para la toma de decisiones, en sus siglas en inglés CRM, analizan los datos históricos e identifican las tendencias de los datos.

En la actualidad, estos métodos no son suficientes, cada vez más se está generalizando el uso de aplicaciones inteligentes de análisis de datos. *Machine Learning* puede constituir una herramienta esencial para las empresas. Ayuda a liberar talento humano para dedicarlo a otros fines, sino que puede ofrecer a los clientes una experiencia más personalizada, gracias al análisis de su historial de compras y a la proyección de sus intereses futuros.

Cuando se realiza un consumo se tiene registro de un cliente con diferentes atributos que lo identifican (Nombres, Apellidos, Dirección), estos datos además se ven enriquecidos por las variaciones (modificadores) en las bebidas de café dependiendo de los

gustos del cliente; por ejemplo, puede escoger el tipo de leche (entera, descremada, almendra o soya) o café descafeinado, con o sin crema, normal o bien caliente, etc.

El *cross-selling* es una estrategia esencial para mejorar la lealtad de marca, aumentar los ingresos, y maximizar el valor de relaciones con los clientes. La base para el éxito del *cross-selling* es el conocimiento del cliente. Un profundo conocimiento de los gustos de sus clientes le pone en una posición de recomendar productos adicionales.

El departamento de marketing en la cadena de cafeterías objeto de estudio de esta investigación, no tiene un conocimiento real de sus clientes y sus estrategias son generales para fidelizar y atraer nuevos. Esto conlleva a un bajo incentivo dirigido a sus clientes más fieles y a los menos fieles atraerlos con estrategias personalizadas que contribuirán de manera efectiva al negocio.

II. Descripción del problema

- Clientes cada vez son más exigentes y tienen necesidades insatisfechas.
- Gran cantidad de datos transaccionales impide a los usuarios del departamento de marketing realizar una segmentación de clientes con las herramientas actuales.
- Se desconoce el nivel de lealtad de los clientes con la empresa.
- Se tiene inconvenientes al momento de preparar una estrategia de retención de clientes y de aumentar su consumo promedio por no tener herramientas y técnicas adecuadas.

Los equipos de marketing no están en la capacidad por sí solos de extraer y procesar el gran volumen de datos que se generan diariamente en las tiendas de la cadena para poder elaborar sus estrategias de retención y recompensa a la velocidad que requiere la empresa. *Machine Learning* puede ayudar a encontrar patrones de consumo y crear automáticamente

perfiles de clientes que no eran evidentes antes. Ante estas circunstancias, cabe plantearse la siguiente pregunta de investigación:

III. Pregunta problémica

¿Cómo contribuir a la segmentación de los clientes en una cadena de cafeterías como apoyo al proceso de toma de decisiones?

IV. Idea a defender

Si se desarrolla un sistema de apoyo para la toma de decisiones utilizando técnicas de *machine learning* se podrá segmentar la base de datos de clientes en una cadena de cafeterías.

V. Objetivo General

Aplicar técnicas de *machine learning* en el desarrollo de un sistema de apoyo en la toma de decisiones para la segmentación de clientes en una cadena de cafeterías.

VI. Objetivos Específicos

1. Analizar el estado del arte en técnicas de *machine learning* para la segmentación de clientes en una cadena de cafeterías.
2. Analizar la estrategia actual de segmentación de clientes en una cadena de cafeterías.
3. Modelar distintos algoritmos de agrupamiento.
4. Validar la solución propuesta con la opinión de expertos.

VII. Justificación

En la actualidad una de las tendencias más creciente constituye la interconexión y digitalización de los procesos organizacionales en beneficio a la mejora de la calidad de sus productos y servicios para mantener su posición en el mercado. Como consecuencia de este fenómeno, los clientes se vuelven más exigentes y requieren que las marcas les provean productos y servicios más afines a sus necesidades y preferencias apostando por un mayor interacción e identificación con ellos. El no centrarse en el cliente es la mayor amenaza de cualquier negocio.

El escenario planteado obliga a las empresas soluciones innovadoras que les permitan tomar decisiones de forma proactiva y acertada para mantenerse en las preferencias de sus consumidores. Entre las áreas que se encuentran actualmente emergiendo cada vez con mayor fuerza se encuentra la inteligencia de negocios, la inteligencia artificial aplicada a las diferentes áreas de la empresa como marketing, publicidad, la producción y las ventas. Esta realidad no se materializa en el Ecuador, fundamentalmente en las cadenas de cafeterías donde las tomas de decisiones se basan en la experiencia, análisis superficiales, datos históricos, comentarios en las redes sociales y en tendencias que muchas veces no responden a la realidad de este sector.

Estos factores no resultan suficientes para poder sustentar con planteamiento científico y técnico la realidad y las preferencias de los clientes, por lo que muchas ocasiones se evidencias los siguientes problemas:

- No se hacen campañas de marketing mediante una estrategia más personalizada basada en los comportamientos de consumo.
- Se desperdician recursos en el desarrollo de nuevos productos.

En la presente investigación se mostrará técnicamente cómo m*achine learning*, ayuda a entender una realidad compleja que hay detrás de las transacciones de los clientes en una cadena de cafeterías. Los modelos de *machine learning* permiten maximizar la eficacia de los esfuerzos de marketing. Entre más datos tenga que analizar, identificará de mejor manera los patrones de consumo del cliente. (Insights, 2017)

La segmentación divide a los clientes en grupos que compartan características comúnes. La segmentación de clientes puede ser realizada en función de:

- Región geográfica
- Demografía (edad, sexo, estado civil, ingresos, etc.)
- Psicográficas (valores, intereses, estilo de vida, pertenencia a grupos, etc.)
- Comportamiento de compra (compras previas, preferencias de despacho, páginas vistas en el *website*, etc.)

En el caso de estudio de una cadena de cafeterías, contiene datos de comportamiento, que son en general los más fáciles de obtener. La meta será encontrar segmentos que sean de alto y bajo valor, desde la perspectiva de acciones de marketing a realizar.

Con la presente propuesta se pretende contribuir a una manera más efectiva en la segmentación de clientes para una cadena de cafeterías y así puedan obtener mejores resultados en la aplicación de sus estrategias de marketing en la fidelización o compañas de promoción de nuevos productos y servicios. En resumen, se puede resaltar estas ventajas:

- Mayor precisión. Con la ayuda de *machine learning*, los profesionales del marketing tienen la capacidad de optimizar las campañas, incluyendo la selección del contenido adecuado, a la audiencia correcta.

- A medida que los datos explotan aún más en volumen, velocidad y complejidad, el *machine learning* puede mostrar nuevos *customer journeys* y segmentos de clientes que no eran evidentes antes.
- Descubrir nuevas oportunidades para el cliente. Aprendiendo de lo que les gusta y prediciendo lo que están más inclinados a comprar.

VIII. Procedimiento metodológico

La presente investigación tendrá como desarrollo los siguientes puntos:

- Levantamiento de información sobre la estructura de datos de la base de datos transaccional.
- Creación de los procesos de extracción, transformación y carga de datos por medio de la herramienta RStudio.
- Comparación de distintos modelos *machine learning* de segmentación usando R.
- Capacitación del usuario para el uso de la herramienta.

IX. Método científico

La investigación no experimental es aquella que se realiza sin manipular deliberadamente variables. Es decir, es investigación donde no se hace variar intencionalmente las variables independientes. Se observa y describe el comportamiento de la información al relacionarla con el objetivo que se pretende alcanzar (Kerlinger, 1979). Desde la metodología de la investigación científica, un modelo puede definirse como una representación simplificada de la realidad (Denning, 2012). El desarrollo de un modelo implica revelar desde una perspectiva de análisis, una manifestación que permite una comprensión más plena del objeto de estudio para resolver el problema y representarlo con

un enfoque sistémico (Reina-Romo, 2012). La forma descriptiva y exploratorio permite indagar y exponer todos aquellos datos que necesitan ser aclarados y delimitados del objetivo propuesto (Jiménez Paneque, 1998).

Sobre la base de los anterior, en el desarrollo de la investigación se utilizó el método inductivo, para la comprensión de los objetivos del departamento de marketing de una cadena de cafeterías mediante el análisis del problema detectar la necesidad del agrupamiento de los clientes basado en sus consumos.

El método científico se utilizó para analizar la literatura sobre técnicas no supervisadas de *machine learning* para la segmentación de clientes en base al comportamiento de sus consumos. El tipo de investigación aplicada a la presente investigación es no experimental de carácter descriptivo y exploratorio. En el desarrollo del modelo se empleó como base el marco teórico referenciado, las experiencias de los profesionales del área de marketing y los resultados del proyecto de investigación llevado a cabo en una cadena de cafeterías. Luego, dicho conocimiento se validó con expertos en marketing con el objeto de demostrar la aplicabilidad y validación del conocimiento creado.

Para el desarrollo de la solución propuesta se utilizó la metodología CRISP-DM para el establecimiento de las tareas y los diferentes pasos en la obtención del modelo.

CAPÍTULO 1

MARCO TEÓRICO

1.1. Segmentación

Segmentación de mercado, consiste en dividir el mercado en segmentos o grupos de compradores que tienen características semejantes que demandan productos separados o diferentes mezclas de marketing. (Escudero, 2017)

Según Ortiz (2014): “La segmentación es la identificación a través de diversas variables de características similares que posee un subgrupo de mercado y que las diferencia del resto.” (pág. 136)

1.1.1. Criterios de segmentación.

Según Braulio & Curto (2015), la creación de perfiles de clientes depende principalmente de tres tipos de datos:

- **Demográficos**, como edad, genero, etnia, país, raza, estado civil, nivel de educación, nivel de ingresos, …
- **Psicográficos**, que describen la personalidad, opiniones, creencias, intereses, estilo de vida, gustos, …
- **Comportamentales**, que se derivan de las interacciones con la organización, como por ejemplo, frecuencia de compra.

1.2. Segmentación RFM

La segmentación RFM (Recencia, Frecuencia, Valor Monetario) es un tipo de segmentación muy común en las empresas de retail para entender y monitorear el comportamiento del consumo de acuerdo con desde cuándo, cuán a menudo y cuánto gastan. (Chorianopoulos, 2016)

RFM ayuda a dividir a los clientes en varias categorías o grupos para identificar a los clientes que tienen más probabilidades de responder a las promociones y también para futuros servicios de personalización.

Como menciona Christy, A.J., et al. (2018), RFM es un proceso científicamente probado. En primer lugar, como menciona Srivastava (2016), se basa en el Principio de Pareto, comúnmente conocido como la regla 80-20, que dice que el 80% de los resultados provienen del 20% de las causas.

Un extenso estudio realizado por Blattberg et al. (2008) demostró la efectividad de RFM cuando se aplica a bases de datos de marketing. Otros investigadores tales como Sarvari et al. (2016), Cuadros L. & Gonzales C. (2017), Mozafari (2018), Chiang (2018) también han comprobado la vigencia del modelo RFM en la toma de decisiones logrando reducir los costos de mercadeo y aumentar los retornos al negocio.

1.3. Machine Learning

Machine Learning es una disciplina científica del entorno de la inteligencia artificial y la ciencia de la computación, cuyo objetivo reside en que los sistemas aprendan automáticamente (sin intervención humana). Toman la información y aprenden por sí mismos, luego aplican ese conocimiento y entrenamiento de un gran conjunto de datos

para reconocer patrones y hacer predicciones, como reconocimiento facial y de objetos, traducción, reconocimiento de voz y más. (Isson, 2015)

Machine learning no sólo pretende construir estos algoritmos para los datos que han sido dados, sino que también pretende mejorar el rendimiento de la tarea de aprendizaje, así como poder dar respuesta a situaciones desconocidas a partir de las observaciones previas.

Swamynathan (2017, pág. 67) clasifica *machine learning* en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje con refuerzo. Sarkar, Bali, & Sharma (2018, pág. 37) hacen una clasificación más amplia de los distintos métodos de *machine learning*, agregando una cuarta categoría llamada aprendizaje semi-supervisado. Sin embargo, los dos tipos de *machine learning* más importante son el aprendizaje supervisado y el aprendizaje no supervisado.

1.3.1. Aprendizaje supervisado.

Theobald, O. (2018, pág. 37) afirma: **El aprendizaje supervisado** es una técnica en la que toma una muestra de datos (conocidos como datos de entrenamiento) y resultados asociados (conocidos como etiquetas o respuestas) con cada muestra de datos durante el proceso de entrenamiento del modelo.

1.3.2. Aprendizaje no supervisado.

Theobald, O. (2018, pág. 37) afirma: En el caso del **aprendizaje no supervisado** no todas las variables y patrones de datos están clasificados, más bien el modelo revelar los patrones escondidos y crear etiquetas a través del uso de algoritmos de aprendizaje no

supervisado. El algoritmo K-means *clustering* es un ejemplo muy popular de aprendizaje no supervisado.

1.3.3. Diferencias entre el *data mining*, el *machine learning* y el *deep learning*.

Aunque todos estos métodos tienen la misma meta – obtener *insights*, patrones y relaciones que se puedan usar para tomar decisiones – tienen diferentes enfoques y habilidades.

Data Mining es el proceso de extracción de información o conocimiento útil (patrones o ideas no tan obvios) de enormes conjuntos de datos. Este conocimiento también podría usarse para aplicaciones comerciales como análisis de mercado, gestión de riesgos, detección de fraudes, etc. La minería de datos se puede realizar mediante análisis, visualización o modelado de aprendizaje de máquina.

Machine Learning es el proceso de obtener conocimiento de datos pasados y usar ese conocimiento para hacer predicciones futuras.

Deep Learning es un tipo de técnica de *machine learning* con más capacidades, ya que trata de imitar las neuronas del cerebro humano. Intenta aprender un fenómeno como una jerarquía anidada de conceptos, con cada concepto definido en relación con conceptos más simples. Ejemplo de algoritmo: redes neuronales de convolución.

1.4. Funcionamiento de *Machine Learning*

Los sistemas de *machine learning* se componen de tres partes principales, que son:

- **Modelo:** el sistema que realiza predicciones o identificaciones.

- **Parámetros:** las señales o factores utilizados por el modelo para tomar sus decisiones.
- **Aprendiz:** el sistema que ajusta los parámetros (y a su vez el modelo) al observar las diferencias en las predicciones frente al resultado real.

1.4.1. Pasos usados en *Machine Learning*.

Según Lantz (2015) hay 5 pasos básicos utilizados para realizar un proyecto de *machine learning* tal como se ilustra en la figura 1.

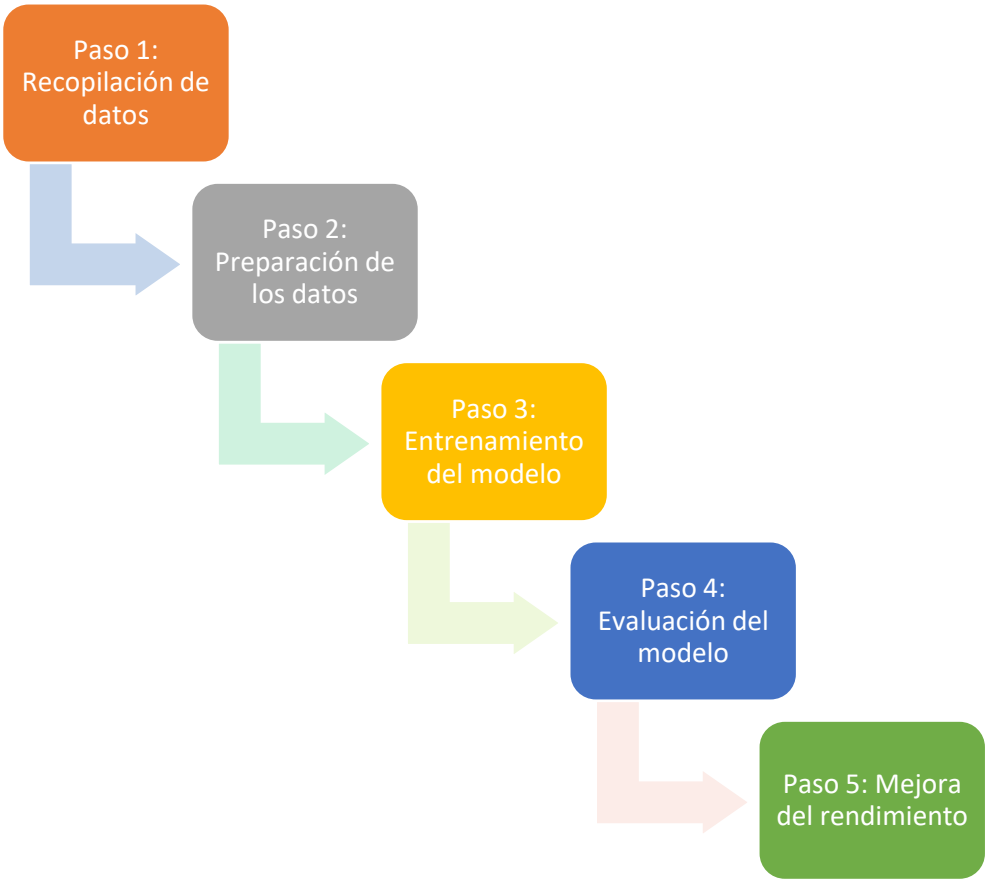


Figura 1 Pasos usados en machine learning. Fuente: elaboración propia

Paso 1 - Recopilación de datos: ya sean los datos sin procesar de Excel, acceso, archivos de texto, etc., este paso (recopilar datos del pasado) constituye la base del

aprendizaje futuro. Cuanto mejor sea la variedad, densidad y volumen de datos relevantes, mejores serán las perspectivas de aprendizaje de la máquina.

Paso 2 - Preparación de los datos: cualquier proceso analítico prospera con la calidad de los datos utilizados. Es necesario, dedicar tiempo a determinar la calidad de los datos y luego tomar medidas para solucionar problemas como la falta de datos y el tratamiento de los valores atípicos. El análisis exploratorio es quizás un método para estudiar los matices de los datos en detalle, lo que a su vez aumenta el contenido nutricional de los datos.

Paso 3 – Entrenamiento del modelo: este paso implica elegir el algoritmo apropiado y la representación de los datos en la forma del modelo. Los datos limpiados se dividen en dos partes: entrenamiento y prueba (la proporción dependerá de los prerrequisitos); la primera parte (datos de entrenamiento) se utiliza para desarrollar el modelo. La segunda parte (datos de prueba), se utiliza como referencia.

Paso 4 - Evaluación del modelo: para probar la precisión, se utiliza la segunda parte de los datos (datos de prueba). Este paso determina la precisión en la elección del algoritmo en función del resultado. Una mejor prueba para verificar la precisión del modelo es ver su desempeño en datos que no se usaron en absoluto durante la construcción del modelo.

Paso 5 - Mejora del rendimiento: este paso puede implicar elegir un modelo diferente por completo o introducir más variables para aumentar la eficiencia. Es por eso que se necesita invertir una cantidad significativa de tiempo en la recopilación y preparación de datos.

1.5. Algoritmos de segmentación o de agrupamiento

El agrupamiento (en inglés, *clustering*) es una tarea de aprendizaje no supervisado, no dispone de una etiqueta de clase a predecir. Puede considerarse como un problema de clasificación, pero donde no existen un conjunto de clases predefinidas, y estas se descubren de forma autónoma por el método o algoritmo de agrupamiento, basándose en patrones de similitud identificados en los datos (Gironés, Casas, Minguillón, & Caihuelas, 2017, pág. 37).

El objetivo del clustering no es clasificar, estimar o predecir una variable, sino entender la estructura macroscópica y relaciones entre objetos, considerando las maneras en las que estos son similares y diferentes. En otras palabras, se enfoca en segmentar el conjunto completo de datos en subgrupos homogéneos (Brunetta, 2014).

Patanjali Kashyap (2017) afirma que el agrupamiento es la actividad de recolectar muestras en grupos de acuerdo con algunas medidas de similitud o disimilitud predefinidas (pág. 94).

1.5.1. Posibles aplicaciones en problemas reales.

Los algoritmos de agrupamiento pueden ser aplicados en muchos campos, por ejemplo:

- Medicina: en la investigación del cáncer para clasificar a los pacientes en subgrupos según su perfil de expresión génica. Esto puede ser útil para identificar el perfil molecular de pacientes con buen o mal pronóstico, así como para comprender la enfermedad.

- Marketing: encontrar grupos de clientes con comportamiento similar dada una gran base de datos de datos de clientes que contiene sus propiedades y registros de compras anteriores.
- Biología: clasificación de plantas y animales dada sus características.
- Seguros: identificar grupos de titulares de pólizas de seguros de automóviles con un alto costo promedio de reclamo; identificación de fraudes.
- Urbanismo: para identificar grupos de casas según su tipo, valor y ubicación.
- Web: clasificación de documentos; agrupación de datos de weblog para descubrir grupos de patrones de acceso similares.

1.5.2. Cluster jerárquico

Méndez (2018, pág. 122) afirma:

> El cluster jerárquico se basa en el cálculo de la distancia euclidiana entre cada uno de los casos. Su ventaja es que permite realizar un gráfico llamado dendograma en el cual estas distancias se pueden visualizar y puede ayudar a decidir el número de grupos a crear. Su inconveniente es que genera una matriz de distancias que es directamente proporcional al número de casos, de forma que para grandes conjuntos de datos el cálculo puede ser muy intensivo en computación.

1.5.3. K-means clustering.

El algoritmo K-means, es uno de los métodos de clustering iterativos más usados. Es destinado a situaciones en las cuales todas las variables son de tipo cuantitativo, y la distancia euclidiana es generalmente escogida como medida de disimilitud.

Méndez (2018, pág. 122) afirma:

El método asigna aleatoriamente un individuo a cada grupo y mide la distancia de cada individuo al centroide como la media de la posición de cada uno de los individuos. A partir de aquí se establece un mecanismo de iteración en el que se reclasifican los individuos al grupo en el que están más cerca de su centroide y se vuelven a recalcular los nuevos centroides. Estos pasos se repiten hasta que no hay disminución entre la distancia de cada individuo a su centroide.

Aunque el algoritmo termina siempre, no se garantiza el obtener la solución óptima. En efecto, el algoritmo es muy sensible a la elección aleatoria de los K centros iniciales. Esta es la razón por la que, se utiliza el algoritmo del K-Means numerosas veces sobre un mismo conjunto de datos para intentar minimizar este efecto, sabiendo que a centros iniciales lo más espaciados posibles dan mejores resultados.

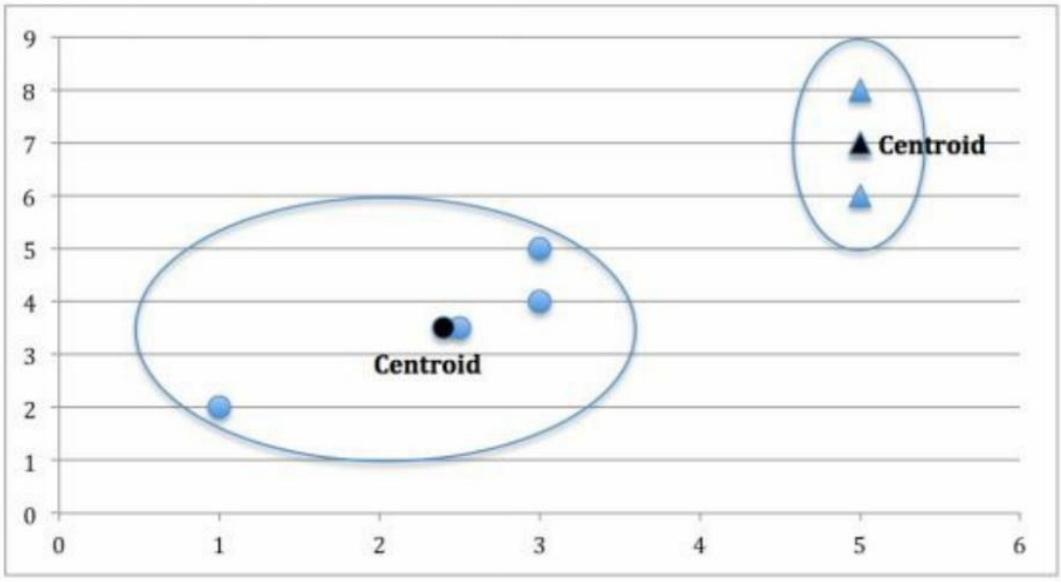


Figura 2 Ejemplo de agrupamiento con el algoritmo K-means.
Fuente: Theobald, O. (2018)

1.5.4. Particionamiento alrededor de medoides (PAM).

El algoritmo k-medoides es un método de agrupamiento por particionamiento relacionado con k-medias para agrupar un conjunto de datos en k grupos o clústeres durante el proceso de entrenamiento. El término medoides se refiere a un objeto dentro de un clúster para el cual la disimilitud promedio entre este y todos los demás miembros del clúster son mínimos (2017, pág. 48). Es una alternativa robusta frente a k-means porque es menos sensitivo a los valores atípicos o ruido en los datos.

El método más común del agrupamiengo k-medoides es el algoritmo de Particionamiento Alrededor de Medoides o conocido por sus siglas en inglés PAM (Partitioning Around Medoids) (Kaufman & Rousseeuw, 1990).

1.5.5. Mapa de Kohonen (SOM).

El modelo fue descrito por primera vez como una red neuronal artificial por el profesor finlandés Teuvo Kohonen, debido a lo cual en ocasiones son llamadas redes o mapas de Kohonen. El mapa de Kohonen, SOM (self- organizing map, mapa autoorganizativo) o SOFM (self-organizing feature map, mapa autoorganizativo de características) es un tipo de red neuronal artificial (ANN por sus siglas en inglés), que es entrenada usando aprendizaje no supervisado para producir una representación discreta del espacio de las muestras de entrada, llamado mapa (Ayundyah Kesumawati, 2016).

Al igual que la mayoría de las redes neuronales artificiales, los SOMs operan en dos modos: entrenamiento y mapeo. En el entrenamiento construye el mapa usando ejemplos entrenantes, mientras que en el mapeo clasifica una nueva entrada.

1.5.6. Algoritmo Two-step.

Consiste en la ejecución de un procedimiento de partición y luego un procedimiento jerárquico. Resuelve las desventajas con las ventajas de cada método. Soporta grandes cantidades de datos de tipo binario y/o categoría. Este procedimiento ha sido usado en una amplia variedad de aplicaciones incluyendo tipos de acceso a internet de los usuarios de teléfonos móviles, segmentando potenciales turistas basados en factores temporales, identificando y caracterizando potenciales adoptantes de vehículos eléctricos y segmentando riesgos relacionados con los viajes (Dolnicar, Grün, & Leisch, 2018).

1.6. Etapas a seguir en el desarrollo del Análisis de Cluster.

Según Parra (2019) se puede distinguir tres pasos básicos para todos los estudios realizados con análisis de cluster:

1) Selección de la muestra que se pretende dividir en grupos.

   El conjunto o muestra de datos que se van a analizar por ejemplo ciudades del Ecuador, compradores de una cafetería, etc.

2) Selección de las variables que se van a utilizar para realizar el análisis.

   Este es uno de los pasos más críticos dentro del proceso. Se debe elegir las variables más representativas para el análisis, determinar si es necesario transformarlas o estandarizarlas sobre todo para algoritmos sensibles a los datos numéricos.

3) Cálculo de las similitudes o disimilitudes entre los casos o sujetos y validación de los resultados obtenidos.

   Una vez calculada la similitud, se realiza el proceso de partición de los datos, pero para ello hay que escoger el algoritmo de segmentación. Los algoritmos de clustering se subdividen en:

   - Jerárquicos: encadenamiento simple, encadenamiento completo, encadenamiento medio, método de Ward, método del centroide.
   - No Jerárquicos: umbral secuencial, umbral paralelo, k-means.
   - O una combinación de los dos.

1.7. Determinación del número k de clústeres

Kassambara (2017, págs. 128-130) afirma: El número k de clústeres es determinado con dos métricas.

- Valor de la silueta: el valor más alto es el mejor.
- La suma de los errores al cuadrado (en inglés sum of squared errors SSE): el valor más pequeño es el mejor pero no siempre, por lo tanto, se usa el

método del codo (en inglés Elbow Method) donde se busca caídas significativas del valor SSR antes de que comience a aplanarse.

1.8. Marco Metodológico

La metodología CRISP-DM nació en el seno de dos empresas Daimler Chrysler y SPSS, que en su día fueron pioneras en la aplicación de técnicas de minería de datos (en inglés, *data mining*) en los procesos de negocio. CRISP-DM se ha convertido de facto en la metodología del sector. Su éxito se debe a que está basada en la práctica y experiencia real de analistas de minería de datos que han contribuido activamente al desarrollo de la misma (Gironés, Casas, Minguillón, & Caihuelas, 2017).



Figura 3 El ciclo vital de CRISP-DM
Fuente: Tomado del manual CRISP-DM de IBM SPSS Modeler

El ciclo vital del modelo CRISP-DM (ver figura 3) contiene seis fases (IBM, 1994, 2012):

- [BU] Comprensión del negocio
    - Determinar objetivos de negocio.
    - Evaluar la situación
    - Determinar los objetivos de minería de datos
    - Producir el plan de proyecto
- [DU] Comprensión de los datos
    - Recoger datos iniciales
    - Describir datos
    - Explorar datos
    - Verificar la calidad de los datos
- [DP] Preparación de los datos
    - Seleccionar datos
    - Datos limpios
    - Datos de construcción
    - Integrar datos
- [M] Modelado
    - Selecciona la técnica de modelado.
    - Generar diseño de prueba
    - Construir modelo
    - Modelo de evaluación
- [E] Evaluación
    - Evaluar resultados
    - Proceso de revisión
    - Determinar los siguientes pasos
- [D] Despliegue

- Despliegue del plan
- Plan de seguimiento y mantenimiento.
- Producir informe final
- Revisar proyecto

Se opta por el uso de la metodología de desarrollo a CRISP-DM por la independencia de esta metodología con respecto a las herramientas tecnológicas a utilizar en la presente investigación de desarrollo tecnológico. (Espinoza, 2017)

1.9. Sustento Legal

Debido a que los datos para esta investigación han sido cedidos por una entidad privada, y que los datos que contiene son de personas naturales, de acuerdo con la Ley Orgánica de la Gestión de la Identidad y Datos Civiles (2016), se ha prestado a borrar todos los rastros de datos personales que existan en estos registros para asegurar que sean anónimas.

Así también, de acuerdo a la misma ley y con los requisitos específicos de la empresa que ha cedido los datos, debido a la firma de un contrato de confidencialidad, no se incluirá en esta investigación ninguna información personal sobre la empresa que ha cedido los datos.

CAPÍTULO 2

SOLUCIÓN PROPUESTA

2.1. Comprensión del negocio

2.1.1. Objetivos del Negocio.

La cadena de cafeterías al momento requiere mejorar la experiencia de sus clientes y ser la mejor opción frente a la competencia. Compite con franquicias internacionales y marcas locales con muchos años en el mercado.

La empresa requiere segmentar o categorizar los clientes para elaborar estrategias de marketing directo de acuerdo con la información de cada cliente, de sus necesidades y sus hábitos de consumo. Por lo que ha planteado un objetivo general para que mediante la inteligencia de negocios se cumpla con mejorar la experiencia de sus clientes:

1. Obtener un modelo de segmentación de clientes utilizando técnicas de *machine learning* sobre los hábitos de consumo de sus clientes actuales.

Este trabajo de investigación se enfoca en la primera tarea: identificar subgrupos o segmentos válidos dentro de los datos del cliente. Desde un punto de vista gerencial, los grupos deben caracterizarse por una similitud suficiente dentro del grupo y la disimilitud entre los grupos para justificar un tratamiento distinto.

Esta es una tarea de aprendizaje no supervisado. El objetivo no es predecir un resultado dado sino descubrir patrones útiles en los datos.

Objetivo de rendimiento: Es imposible establecer un objetivo de rendimiento específico en este contexto. Puede guardar, quizás, para asegurarse de que los grupos identificados son realmente válidos, es decir, no aleatorios.

2.1.2. Entendimiento de los objetivos.

Del primer objetivo de negocio, los datos aquí contenidos corresponden a clientes que realizaron consumos comprendidos entre el periodo enero a junio del 2019, los datos se entregaron en un archivo plano, se realiza limpieza sobre ellos y se estructura en un modelo de datos transformado para el modelo de *machine learning*. Para recabar esta información se hará uso de un único recurso, el autor, con disponibilidad total pues se encuentra enfocado 100% al presente trabajo de investigación.

Los datos recopilados para la realización del proyecto se han categorizado de la siguiente manera:

- **Clientes:** Esta categoría abarca los datos personales de los clientes, proporciona descriptores geográficos y demográficos, como cédula o ruc, dirección, edad, género, estado civil, teléfono y correo electrónico.
- **Ventas:** Esta categoría abarca los registros diarios de facturación en los puntos de venta, que proporcionan la descripción de cada consumo realizado por los clientes durante el periodo de enero a junio del 2019.

2.1.3. Elaborar plan del proyecto.

En el plan del proyecto presentado en la Tabla 1 se describen las fases a realizar con sus respectivas tareas, recursos, duración, dependencias, entradas y salidas.

Tabla 1 Plan del Proyecto

Fases	Tareas	Recursos	Duración estimada	Entradas	Salidas
Comprensión del Negocio	Objetivos del negocio. Evaluación de la situación. Plan del Proyecto.	Investigador	2 semanas	Información sobre la empresa.	Análisis de la información del negocio.
Comprensión de los datos	Recopilación de datos. Descripción de datos. Exploración de los datos. Verificación de la calidad de los datos.	Investigador	1 semana	Recolección inicial de datos de clientes y transaccionales de la empresa.	Obtención de la base de datos con la que se va a trabajar.
Preparación de los datos	Selección de datos a analizar. Limpieza de los datos. Construcción de nuevos datos. Integración y formato de los datos.	Investigador	2 semanas	Preprocesamiento de los datos proporcionados por la empresa.	Datos seleccionados y preparados para aplicar *machine learning*.
Modelado	Seleccionar las técnicas de modelado. Diseñar un modelo de comprobación. Construir el modelo. Evaluar el modelo.	Investigador	4 semanas	Técnicas de *machine learning* basadas en el enfoque de segmentación.	Selección de técnicas a aplicar y generación del modelo.
Evaluación	Evaluar los resultados obtenidos.	Investigador	2 semanas	Modelo de m*achine learning*.	Evaluación e interpretación de los resultados obtenidos.

Fuente: elaboración propia.

2.2. Entendimiento de datos

La base de datos que maneja la empresa se encuentra en formato MSSQL Server versión Estándar 2014. Para explorar los datos se ha utilizado la herramienta de administración de bases de datos SQL MSS. A continuación, se describen las tablas:

Tabla Clientes: Contiene información personal los clientes de la empresa. La tabla clientes cuenta con un total de 464560 registros de clientes. Esta tabla contiene un gran número de atributos, en la Tabla 2 se describen los atributos más representativos para el análisis a realizar:

Tabla 2 Tabla Clientes

Atributos	Descripción	Tipo	Keys
codcliente	Código de identificación que se asigna a cada cliente.	int	PK
nombre	Nombres y apellidos de clientes.	varchar (100)	
dirección	Indica la dirección de domicilio del cliente.	varchar (200)	
ciudad	Indica la ciudad de residencia del cliente.	varchar (40)	
teléfono	Número de teléfono convencional del cliente.	varchar(25)	
e_mail	Correo electrónico personal del cliente.	varchar(150)	
regimfact	Código de identificación del tipo de cliente.	varchar(12)	

Fuente: elaboración propia.

Tabla Factura: Contiene toda la información de facturación realizada por la empresa en el periodo Junio 2018 – Agosto 2018. Cuenta con un total de 136278 registros de facturas. Esta tabla contiene un gran número de atributos, en la Tabla 3 se describen los atributos más representativos para el análisis a realizar (Ver Tabla 3):

Tabla 3 Tabla Facturas

Atributos	**Descripción**	**Tipo**	**Keys**
numserie	Serie de identificación de factura.	varchar(14)	PK
numfactura	Número de identificación de factura.	Int	PK
codcliente	Código de identificación del cliente que realiza la compra.	Int	FK->clientes
fecha	Fecha en que ha sido emitida la factura.	Date	
fecha_anulacion	Fecha en que ha sido anulada una factura, en caso de haber sido anulada.	datetime	
hora	Hora en que ha sido emitida la factura.	Varchar(8)	
totalbruto	La cantidad en dólares que representa el subtotal de la factura.	Double	
totdescuento	La cantidad en dólares del descuento que se realizará a la factura.	Double	
impuestos	La cantidad en dólares del impuesto.	Double	
totalneto	La cantidad en dólares que representa el total de la factura.	Double	

Fuente: elaboración propia.

2.2.1. Verificación de la calidad de los datos.

Antes de continuar con el desarrollo de la investigación y, asumiendo que los datos que proporciona la empresa “Café y Dulces” son veraces, se revisará que no existan incoherencias en los datos y que todos los atributos necesarios tengan un valor correcto.

La base de datos posee una información transaccional muy amplia, en cambio existen muy pocos datos sociodemográficos de los clientes de la empresa.

La tabla Clientes contiene errores en los datos, específicamente para los atributos género, estado civil y fecha de nacimiento, ya que la información que se encuentra en estos campos no corresponde a información real. El campo género no describe correctamente el género de los clientes, el campo estado civil tampoco describe el estado correcto de los clientes, y el campo de nacimiento contiene una fecha estándar para todos los clientes, por lo tanto, no es un dato real.

2.2.2. Exploración de los datos.

Tomando en cuenta que el propósito de la empresa es descubrir los niveles de lealtad de sus clientes, lo cual repercute en la rentabilidad, en el presente trabajo de investigación se ha realizado un análisis de los datos transaccionales, para conocer cómo se encuentra la actividad de compra de los clientes.

2.3. Preparación de datos

La preparación y muestreo de datos abarca varias actividades de: selección, limpieza, construcción de nuevos datos, integración y formato de los datos.

2.3.1. Selección de los datos.

La selección de los datos se realizó tomando en cuenta los objetivos del negocio y entendimiento de los objetivos (Ver sección 2.1.1 Objetivos del negocio y 2.1.2 Entendimiento de los objetivos).

Se han seleccionado los siguientes atributos de la tabla Factura: codcliente, fecha y monto. De los 1360278 registros de facturación se seleccionaron las facturas que correspondían a los clientes finales, quedando un total de 474346.

Se han seleccionado los siguientes atributos de la tabla Clientes: codcliente, nombre. De los 464560 clientes, se han seleccionado los registros correspondientes a clientes finales, ya que el análisis se realizará sobre este tipo de clientes dejando fuera a sociedades y extranjeros, como ha sugerido la empresa, y de estos clientes finales se han seleccionado a aquellos que poseen transacciones realizadas entre Enero de 2019 y Junio 2019, quedando un total de 29769 clientes. En la Tabla 4, se presenta un resumen de los datos seleccionados.

Tabla 4 Tamaño inicial de los datos seleccionados

DATOS CAFÉ Y DULCES	
Periodo a analizar	6 meses (Enero a Junio 2019)
Número de clientes	29769
Número de transacciones	474346

Fuente: elaboración propia.

Con el siguiente código se realiza la carga de los datos seleccionados al ambiente de RStudio:

```
dsInvoice = read.csv2("C://Users//User//Desktop//TesisML//input//dataconsumo.csv",
                      header = TRUE,
                      sep = ",",
                      dec = ".",)
```

Con la función *summary()* se puede revisar un resumen general de los datos seleccionados tal como se muestra en la Tabla 5.

Tabla 5 Resumen de los datos seleccioados

```
> summary(dataclientes)
   CodCliente        FecFactura           Consumo
 Min.   : 283799   Min.   :2019-01-01   Min.   :  1
 1st Qu.:3018928   1st Qu.:2019-02-17   1st Qu.:  3
 Median :3275623   Median :2019-04-05   Median :  5
 Mean   :3624675   Mean   :2019-04-03   Mean   :  7
 3rd Qu.:3858642   3rd Qu.:2019-05-19   3rd Qu.:  9
 Max.   :6838737   Max.   :2019-07-01   Max.   :360
```

Fuente: elaboración propia.

2.3.2. Construcción de nuevos datos.

Para realizar la segmentación se construyeron los atributos Recencia, Frecuencia y Monto Promedio.

- **Recencia:** Para este atributo se tomó la última fecha de compra para cada cliente y luego se calcula el intervalo de tiempo entre la última fecha de compra de cada cliente y la fecha en que se tomaron los datos para el análisis (2019-07-01).
- **Frecuencia:** Este atributo contiene el número de visitas por cada cliente durante el periodo de análisis. Para ello se cuentan las distintas fechas de todos los consumos.
- **Monto Promedio:** El atributo Monto promedio se construye calculando el gasto o consumo promedio por cada cliente en todas sus transacciones durante el periodo de análisis.

Resumen de los atributos seleccionados.

En la Tabla 6 se muestran los atributos seleccionados para la realización del análisis.

Tabla 6 Resumen de los Atributos Seleccionados

Categoría de atributo	Atributo	Tipo	Descripción
Identificador de cliente	CodCliente	Continuo	Este es el atributo se utilizó para identificar a cada cliente en el conjunto de datos.
Comportamiento de compra	Recencia	Discreto	Número de días transcurridos desde la última fecha de consumo del cliente hasta la fecha en se tomaron los datos para el análisis.
	Frecuencia	Discreto	La suma total de visitas o consumos de cada cliente.

	Monto_prom	Continuo	El promedio de dinero gastado en consumos.

Fuente: elaboración propia.

Para el cálculo de estas variables se utilizó el lenguaje R con la herramienta RStudio:

```
## Crear variable Recencia
dataclientes$dias_desde <- as.numeric(difftime(time1 = "2019-07-01",
                                               time2 = dataclientes$FecFactura,
                                               units = "days"))

## Crear conjunto de datos para el análisis incluyendo variable 'frecuencia'
clientesRFM <- dataclientes %>%
  group_by(CodCliente) %>%
  summarise(recencia = min(dias_desde),
            frecuencia = n(),
            monto_prom = mean(Consumo))
```

En la Tabla 7 se observa un resumen de los datos para iniciar la fase de entrenamiento con aprendizaje no supervisado.

Tabla 7 Estadística de los datos seleccionados

```
> stat.desc(select(customers, -CodCliente))
                recencia frecuencia prom_consumo
nbr.val         29769.00  29769.000    29769.000
nbr.null            0.00      0.000        0.000
nbr.na              0.00      0.000        0.000
min                 0.21      1.000        0.600
max               181.21    181.000      182.000
range             181.00    180.000      181.400
sum          1765315.88 160186.000   254936.613
median            44.21      2.000        6.633
mean              59.30      5.381        8.564
SE.mean            0.30      0.062        0.041
CI.mean.0.95       0.58      0.122        0.081
var             2650.90    115.268       50.843
std.dev           51.49     10.736        7.130
coef.var           0.87      1.995        0.833
```

Fuente: elaboración propia.

2.3.3. Integración de datos.

Se realizó fusión del conjunto de datos de clientes: CodCliente y el conjunto de las nuevas variables creadas: Recencia, Frecuencia y Monto_prom utilizando la herramienta

RStudio, ahora se tiene un solo conjunto de datos que contiene todas las variables necesarias para el análisis.

2.3.4. Formato de datos.

Los algoritmos que más adelante se utilizan en la etapa de modelado son muy sensibles a un conjunto de datos que contiene valores atípicos y valores que son de escalas o magnitudes incomparables, es por ellos que las variables Recencia, Frecuencia y Monto_prom se analizaron mediante gráficos estadísticos, para determinar cómo se distribuyen cada una de estas variables a través de histogramas y visualizar la cantidad de valores atípicos mediante diagramas de cajas (ver Anexo 2).

Se identificó que existen algunos casos que presentan valores muy alejados de la mayoría, y por lo general estos son considerados valores atípicos para el análisis y se separan del resto, pero en este caso el método de normalización de las variables RFM permite que se pueda mantener todos los valores.

Normalización de las variables RFM.

Para equilibrar la influencia de las variables medidas en diferentes escalas y / o con variaciones muy diferentes, se debe normalizarlas. Se utilizará la función *scale().*

2.4. Modelado

2.4.1. Elección de la técnica de modelado.

Como se trata de técnicas de segmentación se seleccionaron los siguientes algoritmos:

- Herárquico
- DBSCAN

- K-means
- SOM

2.4.2. Generación de un diseño de comprobación.

Para la selección del número de grupos se aplicarán las técnicas de evaluación interna: suma de error al cuadrado y el índice de la silueta.

Luego de aplicar los algoritmos de segmentación, se determinará cuál de ellos proporciona mejores resultados en base al método de evaluación en cascada. Tomando en cuenta que resulta difícil medir los resultados del aprendizaje no supervisado.

2.4.3. Generación de los modelos.

A continuación, se procedió aplicar los algoritmos seleccionados sobre las variables RFM (Recencia, Frecuencia y Monto promedio).

Aplicación del algoritmo jerárquico

Se aplicó la técnica de clustering jerárquico para segmentar los clientes, mediante el método de Ward, y que de acuerdo a un análisis realizado ofrece una mayor precisión de agrupación frente a otros métodos de agrupación jerárquica (ver Anexo 3). El número de grupos seleccionado mediante la visualización del dendograma (ver Figura 2) fue 7. A continuación, se presenta el código utilizado.

```
## Eliminar la columna CodCliente y establecer como nombres de fila
row.names(customers) <- customers$CodCliente
customers <- select(customers, -CodCliente)

## Normalizar para equilibrar la influencia de las variables medidas en diferentes escalas
clust_data <- scale(customers)

set.seed(11)
hc.complete =hclust (dist(clust_data), method = "complete")
hc.average =hclust (dist(clust_data), method = "average")
hc.ward =hclust (dist(clust_data), method = "ward.D2")
```

En la Figura 4 se muestra el dendograma creado con el siguiente código, con la separación de los 7 grupos.

```
plot(hc.ward, main = "Cluster Tree", xlab = "", sub = "", cex = .9)

rect.hclust(hc.ward, k = 9, border = "darkred")
```

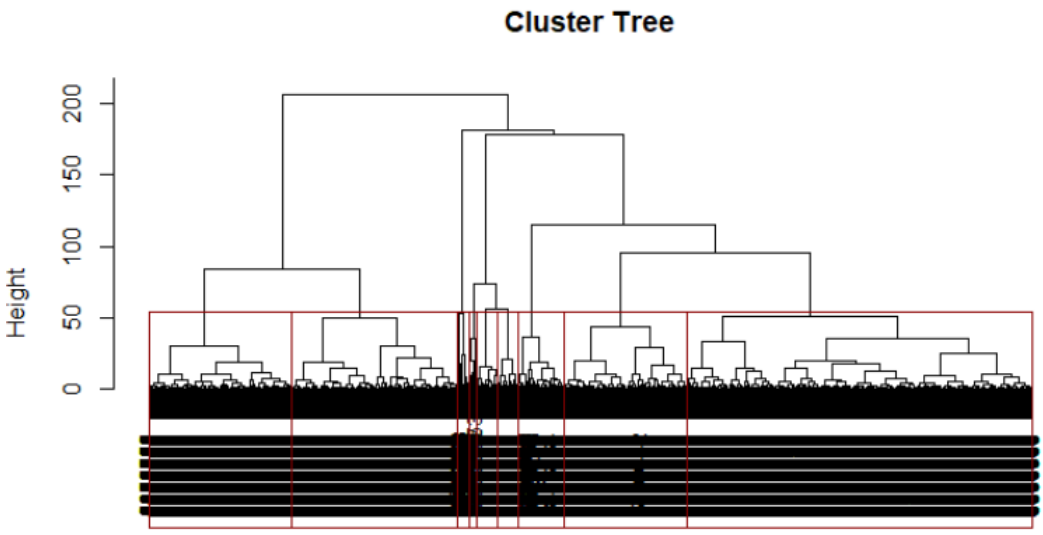


Figura 4 Dendograma del método jerárquico separación de segmentos.
Fuente: elaboración propia.

El Ward.D2 proporciona el dendrograma más equilibrado comparado con los otros dos métodos analizadas (ver Anexo 3) y se usará para obtener los clústers.

Se puede observar en el dendrograma 7 grupos. Esta elección se basa en parte en la evaluación de la validez del clúster por los paquetes R {NbClust} y {clValid}. Las pruebas desafortunadamente no son reproducibles aquí debido a limitaciones de memoria.

Para obtener los centroides de cada clúster se utiliza el siguiente código:

```
## Cortar el dendrograma para producir 7 clústers
clusters <- cutree(hc.ward, 7)
table(clusters)


## Asignar los segmentos a los clientes
customers$cluster <- clusters

## Obtener los centroides de los segmentos
profiles <- customers %>%
  group_by(cluster) %>%
  summarise_each(funs(mean))

cluster_counts <- customers %>%
  group_by(cluster) %>%
  summarise(counts = n())

profiles <- mutate(profiles, counts = cluster_counts$counts)
profiles
```

A continuación, en la Tabla 8 se muestra los segmentos de clientes con el número de clúster que corresponde a cada uno.

Tabla 8 Resultado de 7 clúster creados por el algoritmo Jerárquico

Clúster	Recencia	Frecuencia	Monto Prom	Fidelidad	Nro, obs
1	139.4	1.4	7.2	Muy Baja	5594
2	20.4	5	6.3	Media	11656
3	82	2.2	4.8	Baja	4810
4	8.7	27.1	6.5	Alta	1541
5	84	2.1	29.5	Baja	1656
6	47.9	2.4	13.9	Baja	4155
7	4.5	78.8	5.3	Muy Alta	357

Fuente: elaboración propia.

Es posible obtener información más detallada sobre los segmentos a través de análisis más detallados con paquetes R, como factomineR () y PCA (ver Anexo 4).

Aplicación de mapas auto organizados de Kohonen (SOM).

Determinar el número de nodos o neuronas que se van a utilizar es una tarea difícil (Costa & Netto, 1999). Para determinar el tamaño de nuestra red se aplicó una regla descrita en el Anexo 9, además se seleccionó el número de iteraciones del algoritmo, este fue un proceso iterativo de prueba y error. Finalmente, el tamaño de la red que se utilizó fue de 300 neuronas, con dimensión: 30x30, y el número de iteraciones seleccionado fue 100 (ver Anexo 7).

La función som() está disponible en el paquete "Kohonen" (Wehrens & Kruisselbrink, 2018) de R. El conjunto de datos de entrenamiento debe ser una matriz, en este caso la lista RFM previamente normalizada y luego se utiliza el siguiente código:

```
# Crear un grid SOM
set.seed(100)
ads.grid <- somgrid(xdim = 10, ydim = 10, topo = "hexagonal")

# Crear un modelo SOM
set.seed(100)
som_model <- som(clientesRFM_train,
                 ads.grid,
                 rlen = 500,
                 radius = 2.5,
                 keep.data = TRUE,
                 dist.fcts = "euclidean")
```

Aplicación del algoritmo K-means Clustering.

Antes de aplicar el algoritmo k-means clustering, se debe determinar la cantidad de clúster que se quiere crear. En este caso, se usó el método de suma de cuadrados (*Elbow)* para determinar la cantidad de clúster aplicando el siguiente código:

```
## K-means Clustering
require(factoextra)
set.seed(100)
fviz_nbclust(som_model$codes[[1]], kmeans, method = "wss")
```

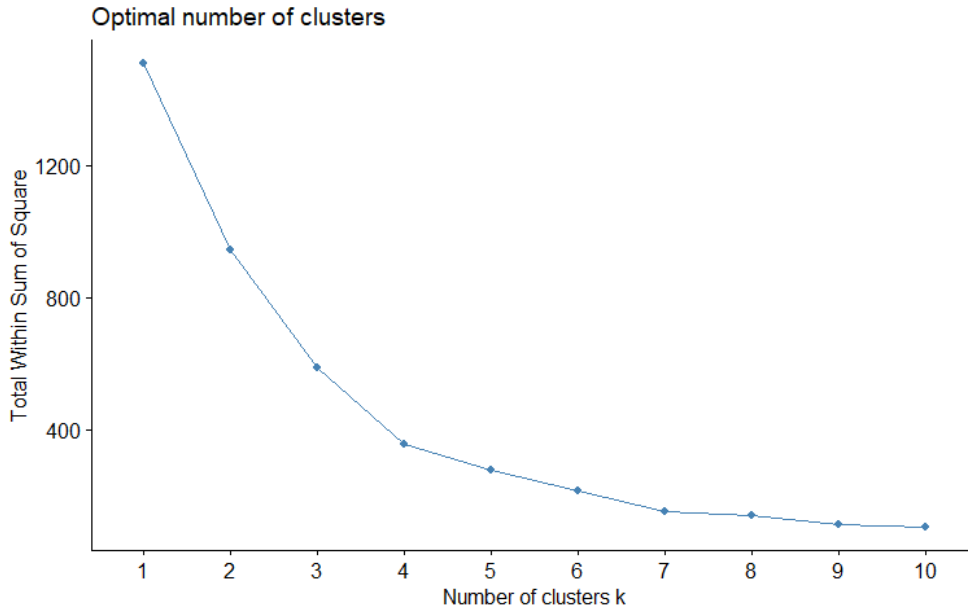


Figura 5 Número óptimo de clúster usando el método de suma de cuadrados. Fuente: elaboración propia.

Se puede observar en la figura 5 que el número óptimo de clúster es 6 según el método WSS. Por tanto, con el siguiente código se aplica el algoritmo k-means con k=6:

```
set.seed(100)
clust <- kmeans(som_model$codes[[1]], 6)

plot(som_model, type = "codes", bgcol = rainbow(9)[clust$cluster], main = "Cluster Map")
add.cluster.boundaries(som_model, clust$cluster)
```

Para visualizar los clúster se puede utilizar la función plot(). El gráfico que se obtiene se muestra en la figura 6.

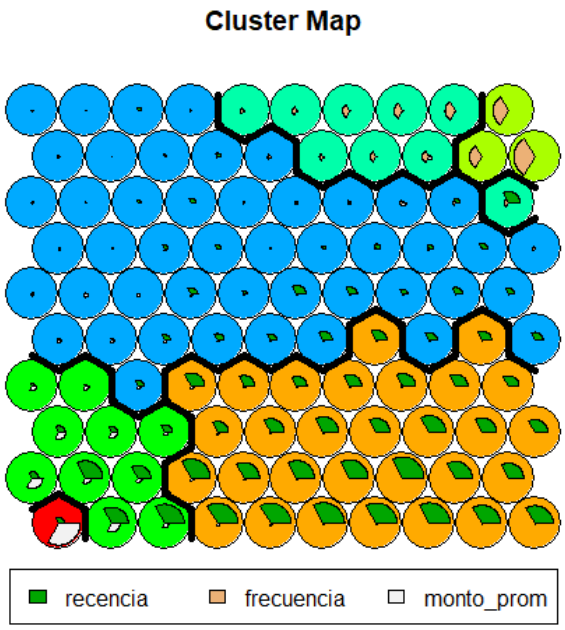


Figura 6 Mapa de clústers usando algoritmo k-means. Fuente: elaboración propia.

En esta parte, para hacer el perfil de los clústeres encontrados, no es necesario mirar los resultados descriptivos del clúster o mirar más profundamente en los datos, solo se tiene que observar el mapa de clústeres anterior.

- Cluster 1 (rojo), promedio de consumo alto, recencia media y frecuencia bajo.
- Cluster 2 (verde), recencia alta, frecuencia baja y promedio de consumo medio.
- Cluster 3 (naranja), recencia alta y frecuencia y promedio de consumo bajo.
- Cluster 4 (azul), recencia media, frecuencia y promedio de consumo bajo.
- Cluster 5 (verde suave), frecuencia media, recencia y promedio de consumo bajo.
- Cluster 6 (verde aceituno), frecuencia alta, recencia y promedio de consumo bajo.

2.4.4. Evaluación del modelo.

2.5. Evaluación

2.5.1. Evaluación de los clústeres.

El algoritmo *clusterboot* intenta responder a esta pregunta evaluando la robustez de cada cluster a las perturbaciones en los datos. La estabilidad del clúster se evalúa midiendo la similitud entre conjuntos de 100 muestras.

Para el cálculo del método de agrupamiento jerárquico, el código de evaluación de estabilidad del clúster es el siguiente:

```
## Cluster evaluation
clust_no <- 7

set.seed(12)
cboot_hclust <- clusterboot(clust_data,
                            clustermethod = hclustCBI,
                            method = "ward.D2",
                            k = clust_no)

bootMean_data <- data.frame(cluster = 1:7, bootMeans = cboot_hclust$bootmean)
```

Para visualizar el resultado (ver figura 7), se utiliza el siguiente código:

```
## Visualize evaluation results
ggplot(data = bootMean_data, aes(x = cluster, y = bootMeans)) +
  geom_point(aes(colour = "darkred", size = 1)) +
  geom_hline(yintercept = c(0.6, 0.8)) +
  labs(y = "stability", title = "Stability evaluation") +
  theme(legend.position="none")
```

Una regla general para interpretar los valores de estabilidad sugiere que los clústers con una puntuación inferior a 0,6 deben considerarse inestables. Aquellos entre 0,6 y 0,75 están midiendo patrones en un grado razonable. Las puntuaciones superiores a 0,8 indican grupos altamente estables.

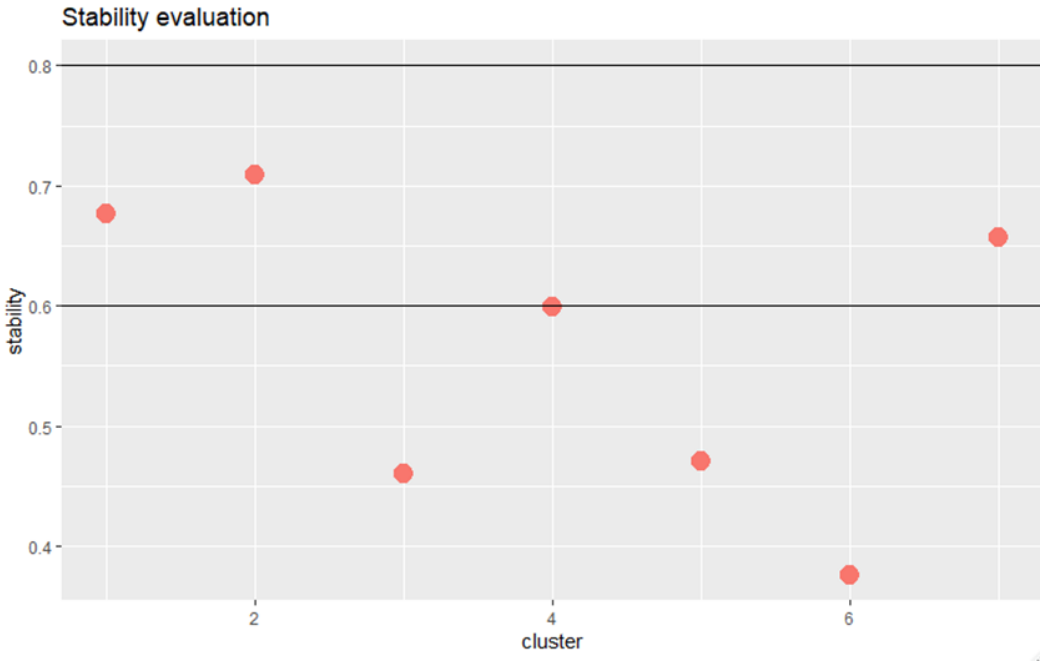


Figura 7 Evaluación de estabilidad de clústeres. Fuente: elaboración propia.

Por lo que se obtiene:

- Clúster 1, 2, y 7 son razonables
- Clúster 4 está en el borde
- Clúster 3 y 5 son inestables, y
- Clúster 6 es altamente inestable. Es probable que esté compuesto de casos inusuales que no encajan cómodamente en ningún otro lugar.

2.5.2. Revisión del proceso.

Durante el proceso de construcción del modelo se tuvo que evaluar varios métodos como por ejemplo en el caso de la red neuronal se usó el método neuralnet pero se tuvo que descartar debido a que tomaba mucho tiempo en ejecutarse. Luego se usó el método h2o.deeplearning, el cual resultó mucho más rápido, pero finalmente se optó por usar los

métodos que posee el paquete “caret” ya que tiene los tres modelos de predicción escogidos en este trabajo.

2.5.3. Determinación de los próximos pasos.

El siguiente paso consiste en ejecutar la etapa de distribución para cumplir con los objetivos comerciales definidos en la sección 2.1.1.2.

2.6. Despliegue

El despliegue es un proceso que consiste en utilizar nuevos conocimientos, como por ejemplo de minería de datos, para implementar las mejoras de la organización. (IBM, 2012).

2.6.1. Planificación de la Implantación.

Para la implantación del presente trabajo de investigación, es necesario tener acceso a los datos, y preferiblemente de las transacciones históricas de más de un año.

2.7. Definición de requerimientos funcionales y no funcionales de la solución a desarrollar

En las entrevistas realizadas a los expertos del Departamento de Marketing se determinó los requisitos funcionales que nacen de las historias de usuario (ver Anexo 1). En resumen, estas son:

1. Identificar subgrupos gerencialmente relevantes en una base de datos de clientes (segmentación).
2. Distinguir los clientes que más consumen y los que menos consumen.

3. Clasificar los clientes por monto de compra y frecuencia de compra.

Para desarrollar la presente propuesta se necesita además del siguiente hardware y software:

1. Hardware necesario para para desarrollar el modelo propuesto: Dell Inspiron 5000 con Windows 10 Home Edition, procesador Intel Core i5 1.6GHz, 16GB de RAM y 500GB disco duro.
2. Software: lenguaje R y RStudio.

CAPÍTULO 3

VALIDACIÓN DE LOS RESULTADOS

3.1. Esquema de validación

Para la validación de la propuesta de segmentación de clientes con técnicas de *machine learning* se utilizó el método de criterio de expertos empleando el escalamiento de Likert (Herrnández et al., 2010). Su opinión y alto grado de conocimiento acerca de la importancia de enfocar esfuerzos y estrategias a segmentos objetivos, asegura la validación de la presente propuesta. Esta validación permitirá identificar si la solución propuesta es viable de ser implementada.

La tabla 9 recoge el perfil de los dos expertos con conocimientos en el área de marketing y muchos años de experiencia liderando equipos de marketing de grandes empresas de consumo y minorista.

Tabla 9 Expertos evaluados

Empresa	Cargo	Instrucción	Experiencia
Dulcafe S.A.	Gerente de Marketing	Magister en Marketing	10 años
Juan Marcet Cia. Ltda.	Gerente Comercial	Magíster en Marketing MBA	15 años

Fuente: elaboración propia

3.2. Aplicación y descripción de los resultados

Se realizó la presentación de la propuesta en las oficinas de los expertos, donde se expuso el modelo con la muestra seleccionada, los grupos de clientes encontrados y posteriormente analizar y evaluar los resultados. Luego de la reunión, se entrevistó a los expertos en base a cinco preguntas detalladas en la Tabla 10.

Tabla 10 Preguntas para los expertos

Pregunta	Descripción
1	En base a lo expuesto: ¿Cómo evalúa el modelo desarrollado para mejorar la toma de decisiones para la segmentación de clientes en una cadena de cafeterías?
2	En base a lo expuesto: ¿Cómo valora los métodos desarrollados como parte del modelo y los principios que lo rigen?
3	En base a lo expuesto: ¿Considera que el modelo presentado segmenta el mercado en grupos que tengan sentido al negocio?
4	En base a su análisis: ¿Considera viable la solución BI expuesta?
5	En base a su experiencia como líder organizacional: ¿Recomendaría la implementación del modelo de segmentación propuesto a ?
6	De su apreciación. ¿Debería una empresa destinar un fondo a los proyectos de incremento de innovación tecnológica a fin de hacer crecer su imagen?

Fuente: elaboración propia

El método criterio de expertos proporciona valoraciones de expertos sobre asuntos vinculados con la propuesta de solución. Como método para el procesamiento estadístico de estas opiniones o evaluaciones fue aplicada la escala psicométrica creada por Rensis Likert en 1932 (Likert, 1932) (Boone & Boone, 2012). Fue aplicada en esta investigación a través de un cuestionario con el objetivo de conocer el nivel de acuerdo o desacuerdo con la solución propuesta.

Tabla 11 Escala de valoración

Criterios	Puntuación
Muy de acuerdo	5
De acuerdo	4
Ni de acuerdo ni en desacuerdo	3
En desacuerdo	2
Completamente en desacuerdo	1

Fuente: elaboración propia

Se tabularon los datos luego del resultado de las entrevistas a los expertos que se observa en la tabla 12.

Tabla 12 Respuestas dadas por los expertos por cada aspecto relevante

Experto	Aspectos relevantes					
	P1	**P2**	**P3**	**P4**	**P5**	**P6**
E1	5	5	5	4	5	5
E2	5	4	5	5	4	5

Fuente: elaboración propia

La Figura 8 muestra que el índice porcentual relacionado con la valoración de los expertos, sobre los aspectos planteados, es superior a 90 en todos los casos:

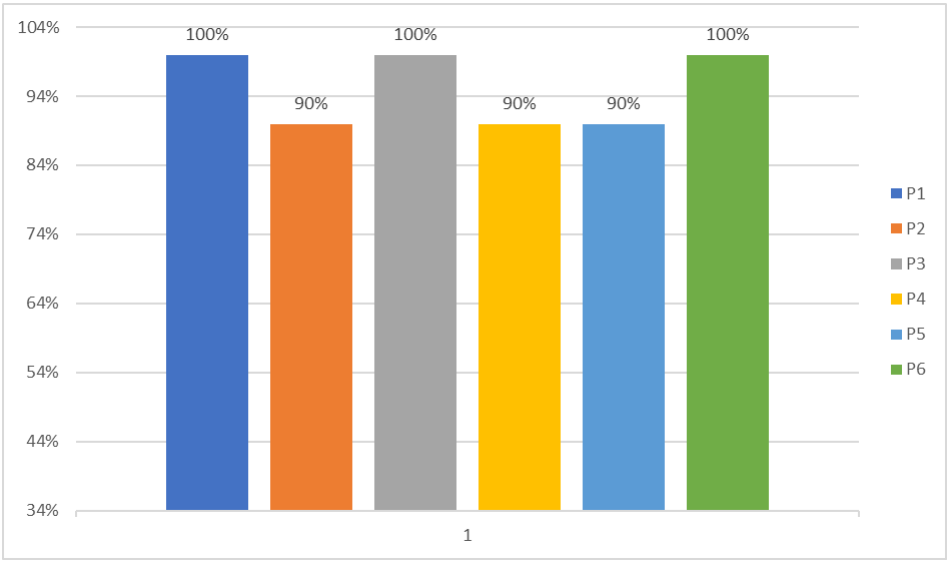


Figura 8 Índice Porcentual de Concordancia de los expertos.
Fuente: elaboración propia.

El procesamiento elaborado a través del escalamiento de Likert demuestra que tanto los elementos teóricos como las características de los componentes del modelo, tienen una alta valoración por parte de los expertos. Durante el proceso se cotejaron criterios favorables para el uso y aplicación del modelo de segmentación de clientes para apoyar la toma de decisiones a los profesionales de marketing en una cadena de cafeterías.

3.3. Impacto

3.3.1. Perspectiva tecnológica.

El desarrollo de la presente investigación se realizó basándose en la metodología CRISP-DM y mediante la aplicación de técnicas descriptivas y auxiliares, se descubrieron segmentos representativos de clientes, utilizando algoritmos de aprendizaje no supervisado de *machine learning*, con un alto grado de aceptación según los expertos en marketing entrevistados.

3.3.2. Perspectiva económica.

Los resultados obtenidos del Trabajo de Titulación representan un aporte hacia la mejora en la toma de decisiones de los profesionales de marketing, el mismo que también sirve de base para futuras investigaciones.

El desarrollo del proyecto no conllevó a una inversión económica mayor, puesto que se basó en herramientas Open Source como RStudio. En la parte de recursos humanos se contó con la persona investigadora y el tutor de Trabajo de Titulación

CONCLUSIONES

La segmentación utilizando técnicas de *machine learning* es un enfoque poderoso y útil para encontrar subgrupos en las bases de datos de clientes. Claro que está que los métodos tradicionales estadísticas siguen siendo ampliamente utilizados en las empresas. Pero con el auge del *big data*, este trabajo de investigación presente demostrar los avances y la gran variedad de algoritmos para solucionar esta clase de problemas de una forma más eficiente. Por un lado, estos algoritmos identifican patrones naturales en conjuntos de datos multivariados, a diferencia de las alternativas no estadísticas más utilizadas que se basan en gran medida en el juicio. Los últimos enfoques pueden verse seriamente comprometidos por varios tipos de sesgos, agendas personales y limitaciones cognitivas en el procesamiento de datos complejos.

En términos prácticos, el enfoque demostrado aquí se puede utilizar para obtener una descripción fiel de los subgrupos de clientes en un momento determinado en el tiempo y podría servir para formular hipótesis, guiar investigaciones adicionales e informar en general soluciones de segmentación gerencial no estadística.

En la presente investigación de desarrollo tecnológico, se escogieron los modelos *machine learning* de agrupación y redes neuronales de mayor uso para llevar a cabo la segmentación de clientes. En el mundo real, determinar una clase de valor del cliente no es una tarea sencilla. La clase de valor del cliente a menudo no está disponible o es difícil de predefinir; y es importante involucrar a personas del lado del negocio en el proceso.

El algoritmo SOM resultó ser el de menor tiempo de ejecución para el conjunto de datos de entrenamiento en comparación con los demás algoritmos. No fue necesario recurrir a

otros métodos para hallar el valor k de subgrupos de segmentos como sí hay que hacerlo para el caso de K-means.

El algoritmo jerárquico resultó el más lento de todos en la fase de entrenamiento. Su principal ventaja es observar fácilmente los subgrupos de segmentos a través de la visualización en forma de árbol o dendograma. El método Ward.D2 se comprobó es el que mejores resultados balanceados entregó.

Para los profesionales de marketing de esta cadena de cafeterías, el desarrollo de este trabajo de investigación le ha brindado un enfoque diferente para solucionar el problema de clasificar al universo de clientes por sus hábitos de consumo, demostrando con los resultados obtenidos que la propuesta es económicamente viable si desean desarrollar una solución completa para sus estrategias de comunicación vía correo electrónico, tomando como base la presente investigación de desarrollo tecnológico.

Se demostró que los algoritmos de *machine learning* dan resultados altamente precisos en todos los pasos que involucran la construcción del modelo, incluida la detección de valores atípicos, el tratamiento de valores perdidos, la minería de datos y las predicciones. Esto se debe a que los patrones en los datos se observan a un nivel muy granular de lo que tradicionalmente era posible.

RECOMENDACIONES

Es muy recomendable entender los objetivos del negocio para seleccionar el conjunto de variables que se a analizar, que haga sentido de negocio a los usuarios tomadores de decisiones. Como se ha comprobado en este proyecto, el análisis basado en RFM sigue siendo muy útil para el objetivo de negocio planteado.

En los algoritmos de segmentación por particionamiento como k-means, se requiere especificar el valor k de segmentos a clasificar. Se recomienda utilizar la función *NbClust* () del paquete del mismo nombre. Con una sola llamada evalúa con treinta índices el valor óptimo de segmentos distintos en un conjunto de datos.

Para la segmentación de clientes es recomendable probar diferentes algoritmos de clustering. Ninguno arrojará los mismos resultados. Para la validación del clustering es recomendable utilizar el coeficiente de la silueta o el índice Dunn y compararlos entre los distintos algoritmos utilizados. El mejor algoritmo dependerá de la distribución de los datos y si los segmentos encontrados hacen sentido al negocio.

Sobre la base del trabajo de investigación de segmentación con técnicas no supervisadas de *machine learning* se puede identificar los grupos de clientes más leales y menos leales; esperando que los profesionales del marketing puedan decidir estrategias para retener a los clientes frecuentes y estimular mayores consumos a los menos frecuentes con el objetivo de aumentar el nivel de ventas.

BIBLIOGRAFIA

Alipour Sarvari, P., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes, 45*(7), 1129-1157.

Ayundyah Kesumawati, D. S. (December de 2016). A Segmentation Group by Kohonen Self Organizing Maps (SOM) and K -Means Algorithms. *Int. J. Advance Soft Compu. Appl, 8*(3), 103.

Blattberg, R. e. (2008). *RFM Analysis. In: Database Marketing. International Series in Quantitative Marketing, vol 18.* New York: Springer.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of extension, 50*(2), 1-5.

Boubeta, A. I. (2010). *Organización en El Punto de Venta.* Ideaspropias.

Boza, B. (2 de Enero de 2019). *Semanaeconomica.com*. Obtenido de http://semanaeconomica.com/article/management/gobierno-corporativo/324810-cinco-pilares-para-el-2019/

Braulio, N., & Curto, J. (2015). *Customer Analytics: mejorando la inteligencia del cliente mediante los datos.* Barcelona: UOC.

Brunetta, H. A. (2014). *Del marketing relacional al CRM.* Buenos Aires: Ediciones Todo Management.

Chiang, W.-Y. (2018). Identifying high-value airlines customers for strategies of online marketing systems. An empirical case in Taiwan. *Kybernetes, 47*(3), 525-538.

Ching-Hsue, C., & You-Shyang, C. (Abril de 2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications: An International Journal, 36*(3), 4176-4184.

Chorianopoulos, A. (2016). *Effective CRM Using Predictive Analytics.* Chichester: John Wiley & Sons, Ltd.

Christy, A. e. (2018). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University – Computer and Information Sciences*.

Costa, J. A., & Netto, M. L. (Junio de 1999). Estimating the number of clusters in multivariate data by self-organizing maps. *International Journal of Neural Systems, 9*(3), 195-202.

Cuadros L., A., & Gonzales C., C. y. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura,, 21*(54), 41-51.

Denning, P. J. (2012). Closing Statement: What Have We Said. *The Computer Journal, 55*(7), 863-865.

Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful.* Springer.

Escudero, J. (2017). *Comunicación y atención al cliente.* Madrid: Ediciones Paraninfo.

Espinoza, M. (2017). CRISP-DM: Conocimiento y Comunicación de una Metodología para Minería de Datos. En A. Colina, *Retos y Perspectivas de las Tecnologías de la Información* (págs. 42-53). Samborondón: Universidad Ecotec.

Gartner. (2017, May 31). *Let Machine Learning Boost Your Business Intelligence*. Retrieved from https://www.gartner.com/smarterwithgartner/let-machine-learning-boost-your-business-intelligence/

Gironés, J., Casas, J., Minguillón, J., & Caihuelas, R. (2017). *Mierí a de Datos: Modelos y Algoritmos.* Barcelona: UOC.

Herrnández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2010). *Metodología de la investigación.* México: Mc Graw Hill.

IBM. (1994, 2012). *Manual CRISP-DM de IBM SPSS Modeler.* EEUU: IBM.

IBM, C. (2012). *Manual CRISP-DM de IBM SPSS Modeler.* Obtenido de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Insights, R. (9 de Agosto de 2017). *3 razones para usar machine learning en marketing*. Obtenido de http://insights.la/2017/08/09/3-razones-incluir-machine-learning-estrategia/

Isson, J. P. (2015). *Unstructured Data Analytics.* New Jerse: Wiley.

Jiménez Paneque, R. (1998). *Metodología de la Investigación. Elementos Básicos para la investigación clínica.* La Habana: Ciencias Médicas.

Kashyap, P. (2017). *Machine Learning for Decision Makers.* Bangalore: Apress.

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning.*

Kaufman, ,. L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken: John Wiley & Sons, Inc.

Lantz, B. (2015). *Machine Learning with R* . Birmingham: Packt Publishing Ltd.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 1-55.

Lynn, S. (14 de Enero de 2014). *Self-Organising Maps for Customer Segmentation using R.* Recuperado el 15 de Mayo de 2019, de https://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/

Méndez, M. (2018). *Análisis de datos con R: Una aplicación a la investigación de mercados.* Madrid: Esic Editorial.

Mozafari, A. (2018). A Method for Clusterink Bank Customers using RFM Model and three parameter grey numbers. *International Journal of Information, Business and Management, 10*(4), 289-298.

Ortiz, M. (2014). *Marketing. Conceptos y Aplicaciones.* Barranquilla: Editorial Universidad del Norte.

Parra, F. (25 de Enero de 2019). *Estadística y Machine Learning con R.* Obtenido de https://bookdown.org/content/2274/agrupacion-de-la-informacion.html

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning.* Cham: Springer.

Reina-Romo, E. e. (2012). *In-Silico Models as a Tool for the Design of Specific Treatments: Applications in Bone Regeneration. In Technologies for Medical Sciences.* Springer.

Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with Python. A Problem-Solver's Guide to Building Real-World Intelligent Systems.* Bangalore: Apress.

Shopify. (s.f.). *Enciclopedia de negocios para empresas*. Obtenido de Segmentación de clientes: http://es.shopify.com/enciclopedia/segmentacion-de-clientes

Srivastava, R. (2016). Identification of customer clusters using RFM model: a case of diverse purchaser classification. *Int. J. Bus. Anal. Intell., 4*(2), 45-50.

Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps.* Bangalore: Apress.

Theobald, O. (2018). *Machine Learning For Absolute Beginners.* Independently Published.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. (IEEE, Ed.) *IEEE Transactions on Neural Networks, 11*(3), 586 - 600.

Wehrens, R., & Kruisselbrink, J. (17 de Diciembre de 2018). *Pakage Kohonen.* Obtenido de https://cran.r-project.org/web/packages/kohonen/kohonen.pdf

Wikipedia. (19 de Enero de 2010). *Aprendizaje no supervisado*. Obtenido de https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado

Wikipedia. (27 de Mayo de 2010). *Aprendizaje supervisado*. Obtenido de http://es.wikipedia.org/wiki/Aprendizaje_supervisado

ANEXOS

Anexo 1: Historias de Usuario

Historia de Usuario: 1. Experiencia del cliente	
Como: Líder de Marketing **Quiero:** Entender mejor a nuestros clientes **Para poder:** Realizar acciones específicas para mejorar la experiencia del cliente	
Validación: - Tipos de segmentación por hábitos de compra y niveles de gasto.	**Valor:** 500 **Prioridad:** 1 **Estimación:** 20 h **Calidad:** INVEST

Historia de Usuario: 2. Segmentación de clientes	
Como: Líder de Marketing **Quiero:** Tener un claro criterio para segmentar a nuestros clientes **Para poder:** Identificar los distintos grupos relevantes al giro de negocio	

Validación:
- Clientes clasificados por las variables RFM
  o Recencia
  o Frecuencia
  o Monto de compra
	Valor: 500
	Prioridad: 1
	Estimación: 20 h
	Calidad: INVEST

Historia de Usuario: 3. Segmentación de cliente por Recencia	
Como: Líder de Marketing **Quiero:** Clasificar a los clientes por una variable de medición de recencia **Para poder:** Organizar estrategias de marketing por cada segmento obtenido	
Validación: - Clientes clasificados de acuerdo a cuándo fue la última vez que realizó un consumo tomando la diferencia en días desde la última orden respecto al último día en que se realiza el análisis.	**Valor:** 250 **Prioridad:** 2 **Estimación:** 12 h **Calidad:** INVEST

Historia de Usuario: 4. Segmentación de cliente por Frecuencia	
Como: Líder de Marketing **Quiero:** Clasificar por una variable de medición de frecuencia **Para poder:** Organizar estrategias de marketing por cada segmento obtenido	
Validación: - Clientes clasificados en base al número de visitas:	**Valor:** 250 **Prioridad:** 3 **Estimación:** 10 h

- La fecha única de la orden de consumo se considera una visita.	**Calidad:** INVEST

Historia de Usuario: 5. Segmentación de cliente por valor monetario	
Como: Líder de Marketing **Quiero:** Clasificar por una variable de medición de valor monetario **Para poder:** Organizar estrategias de marketing por cada segmento obtenido	

Validación:
- Clientes clasificados en base al valor monetario consumido:
  o 50% más bajo valor de vida de un cliente
  o 51-80% la mitad valor de vida de un cliente
  o El 20% más alto valor de vida de un cliente
- Identificar los clientes por cada nivel de valor monetario
	Valor: 250
	Prioridad: 4
	Estimación: 15 h
	Calidad: INVEST

Anexo 2. Análisis estadístico mediante histogramas y diagramas de cajas de las variables RFM

Se realizó un análisis de los valores RFM, mediante histogramas que permitieron visualizar cómo se encuentran distribuidos. También se visualizaron los valores atípicos mediante diagramas de cajas.

Las figuras 8, 9 y 10 muestran la distribución de las variables, y en base a estas se observa que todas las variables tienen una distribución con sesgo hacia la izquierda, aunque en menor proporción para la variable Recencia.

```
ggplot(customers, aes(x = frecuencia)) +
  geom_histogram(bins = 20, fill = "steelblue") +
  theme_classic() +
  labs(x = "", title = "Frecuencia Consumo") +
  scale_x_continuous(breaks = c(0, 10, 20, 30, 40, 50, 60))
```

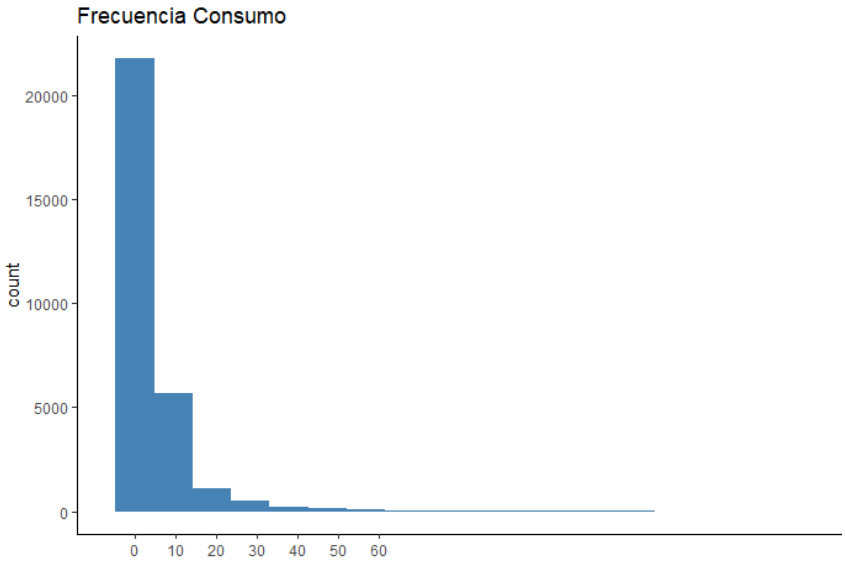


Figura 9 Histograma de frecuencia de consumo por Cliente.
Fuente: elaboración propia.

Aproximadamente el 40% de los clientes hacen solo un consumo.

```
dim(filter(customers, frecuencia == 1))[1]/dim(customers)[1]
```

```
## [1] 0.40
```

Recencia

Aparte de un pico en el extremo izquierdo, la distribución es bastante uniforme (ver figura 11).

```
ggplot(customers, aes(x = recencia)) +
  geom_histogram(fill = "darkred") +
  theme_classic() +
  labs(x = "", y = "", title = "Recency in Days") +
  scale_x_continuous(breaks = c(0, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200))
```

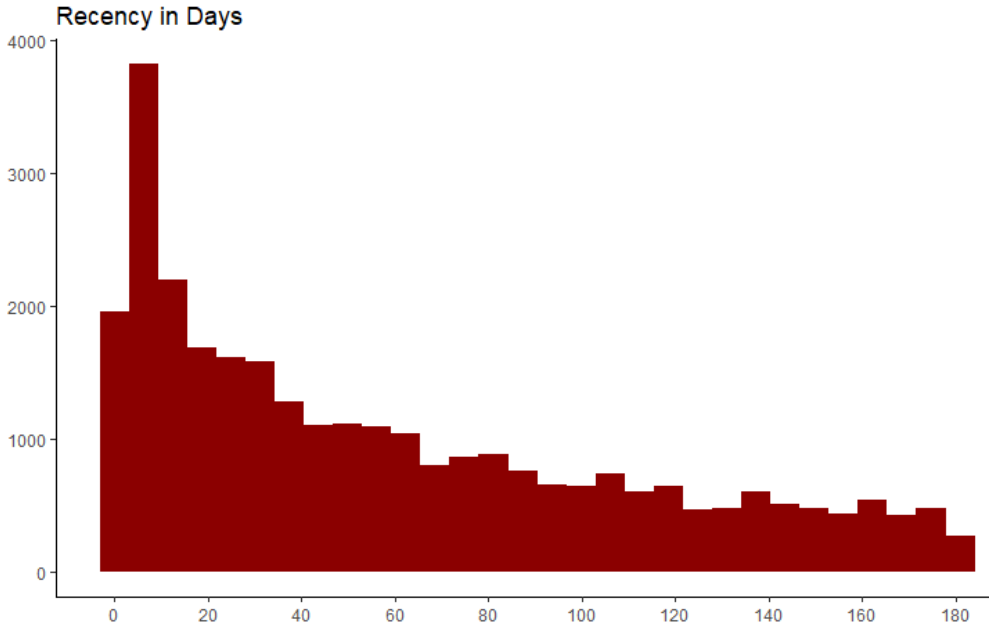


Figura 10 Histograma de recencia por cliente. Fuente: elaboración propia.

El 40% de los clientes han estado activo en el último mes del periodo de análisis.

```
dim(filter(customers, recencia < 30))[1]/dim(customers)[1]
## [1] 0.40
```

Monto Promedio

La distribución tiene una media de 9 con un 50% entre 4 y 10. La media se sitúa por encima de le mediana por el sesgo causado por algunos valores extremos. La distribución se extiende hasta 182.

```
summary(customers$prom_consumo)
##  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1       4       7       9      10     182
```

```
ggplot(customers, aes(x = prom_consumo)) +
  geom_histogram(bins = 20, fill = "darkred") +
  theme_classic() +
  labs(x = "average amount", title = "Promedio de Consumos")+
  scale_x_continuous(breaks = c(0, 10, 20, 30, 40, 50, 60))

ggplot(customers, aes(x = prom_consumo)) +
  geom_histogram(bins = 20, fill = "darkred") +
  theme_classic() +
  labs(x = "amount (log10)",
       title = "Promedio de Consumos\n (Data are log transformed to ease visualization)") +
  scale_x_log10()
```

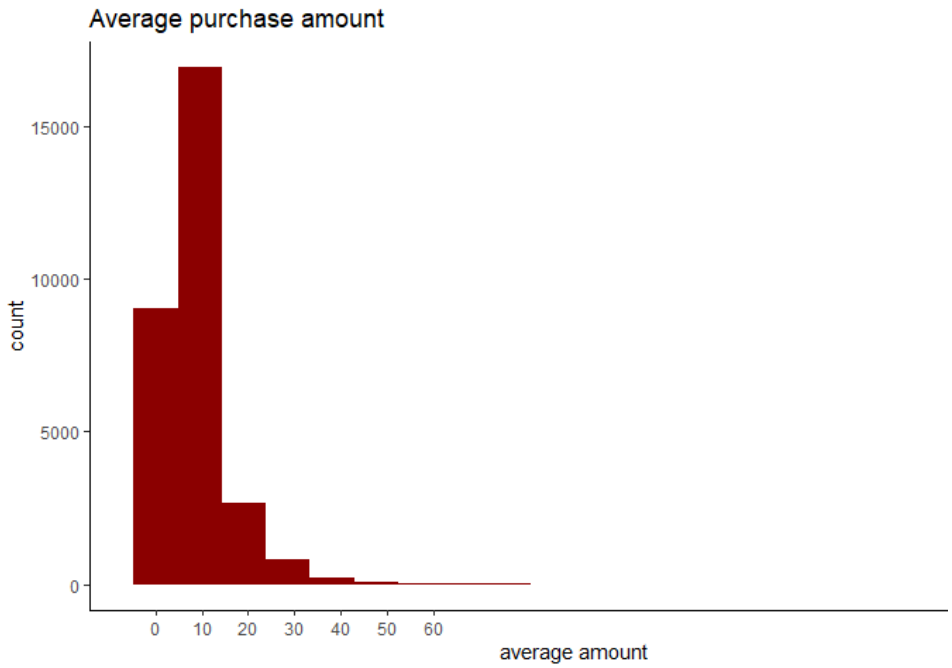


Figura 11 Histograma de promedio de consumo por cliente. Fuente: elaboración propia.

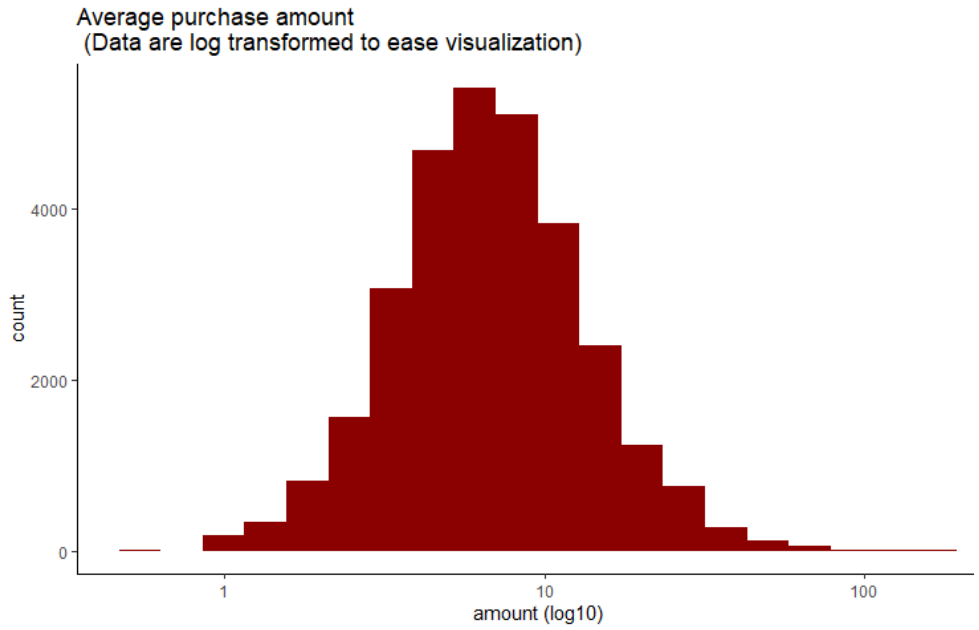


Figura 12 Histograma promedio de consumo por cliente con transformación logarítmica. Fuente: elaboración propia.

Se presentan los diagramas de caja en la figura 13 que muestran de forma gráfica los valores atípicos del conjunto de datos para las variables RFM. Se puede observar que el atributo Recencia no presenta ningún valor atípico, mientras que en las variables frecuencia y valor monetarios se identificaron una gran cantidad de valores atípicos.

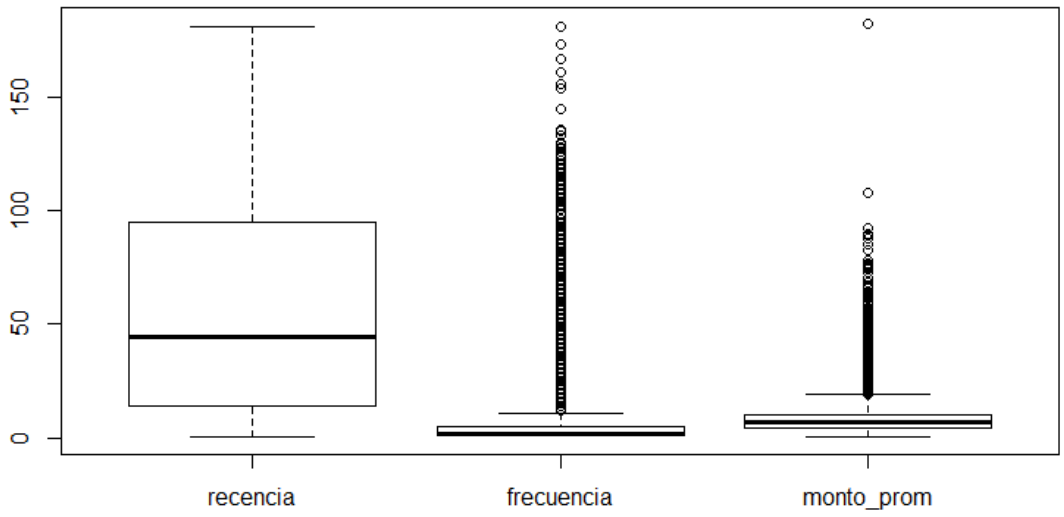


Figura 13 Visualización de valores atípicos de las variables RFM.
Fuente: elaboración propia

Anexo 3. Resultados y comparativa de distintos métodos de agrupamiento jerárquico y sus dendogramas.

Implementando la segmentación jerárquica con 3 tipos de enlace y utilizando la distancia euclidiana y su respectiva visualización con el siguiente código como se observa en las figuras 14, 15 y 16:

```
set.seed(11)

hc_complete =hclust (dist(clust_data), method = "complete")
hc_average =hclust (dist(clust_data), method = "average")
hc_ward =hclust (dist(clust_data), method = "ward.D2")
plot(hc.complete, main = "Complete Linkage", xlab = "", sub = "", cex = .9)
plot(hc.average, main = "Average Linkage", xlab = "", sub = "", cex = .9)
plot(hc.ward, main = "Cluster Tree", xlab = "", sub = "", cex = .9)
```

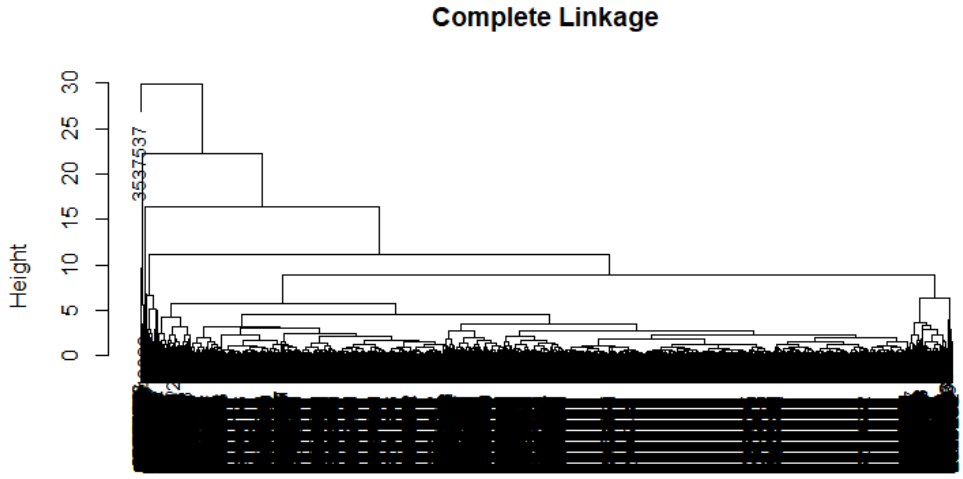


Figura 14 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace completo. Fuente: elaboración propia.

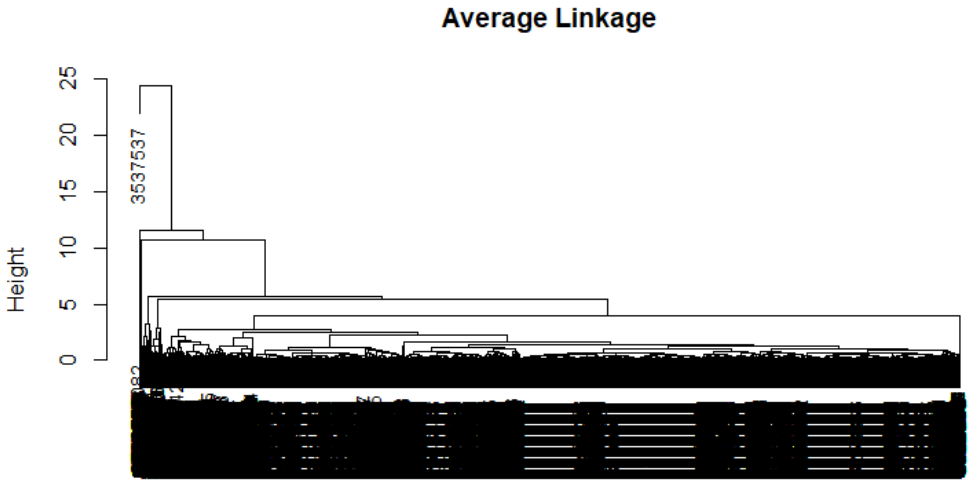


Figura 15 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace promedio. Fuente: elaboración propia.

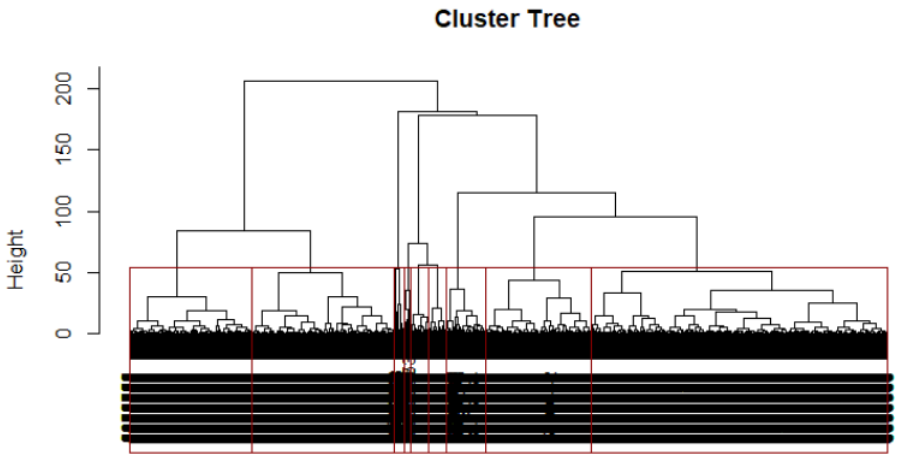


Figura 16 Dendograma de segmentación de resultados de red neuronal SOM con tipo de enlace Ward.D2. Fuente: elaboración propia.

Anexo 4. Visualización de clústers con el método de agrupamiento jerárquico.

Con el siguiente código se puede crear la visualización de la Frecuencia vs Monto Promedio con clúster como variable de agrupamiento como se observa en la figura 17.

```
customers$cluster <- as.factor(customers$cluster)
ggplot(data = customers,
       aes(x = frecuencia, y = prom_consumo, colour = cluster, shape = cluster)) +
  geom_point() +
  theme_dark() +
  labs(title = "promedio consumo vs. frecuencia", shape = "cluster", colour = "cluster")
```

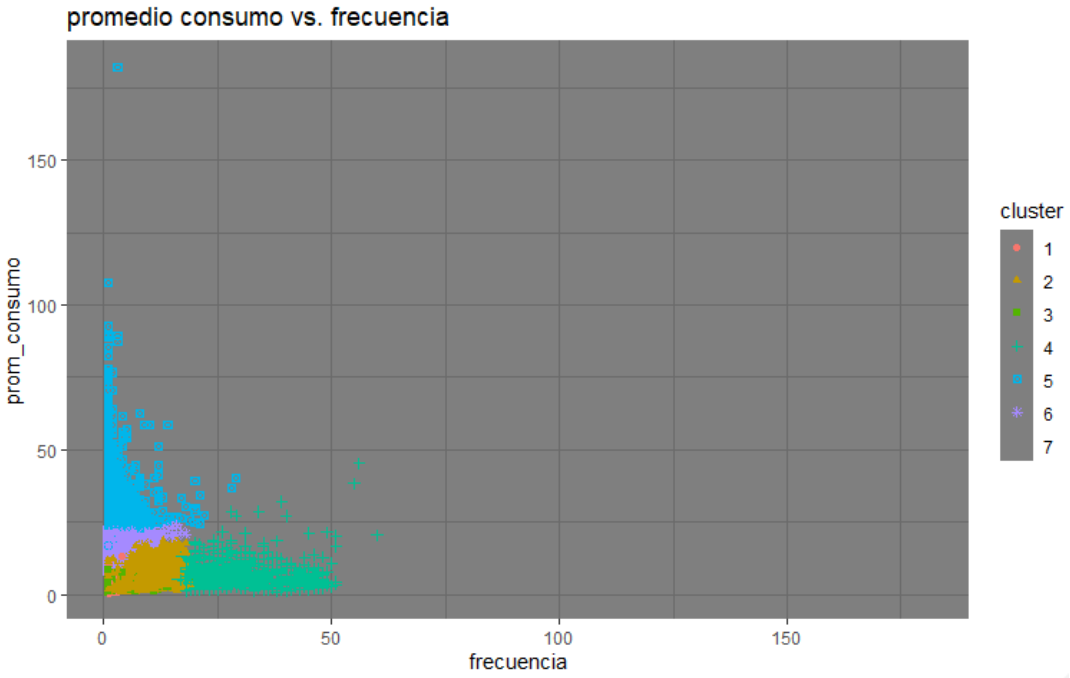


Figura 17 Visualización de clústeres promedio de consumo vs frecuencia. Fuente: elaboración propia.

Con el siguiente código se puede crear la visualización Monto Promedio vs Recencia con clúster como variable de agrupamiento como se observa en la figura 16.

```
ggplot(data = customers,
       aes(x = recencia, y = prom_consumo, colour = cluster, shape = cluster)) +
  geom_point() +
  theme_dark() +
  labs(title = "promedio consumo vs recencia",  shape = "cluster", colour = "cluster")
```

Componentes Principales.

Los primeros dos componentes principales describes en un plano N dimensiones de datos que captura tanta información como sea posible sobre el espacio de características. Es una visualización bidimensional más representativa de los grupos que las gráficas anteriores que solo muestran combinaciones de datos por pares.

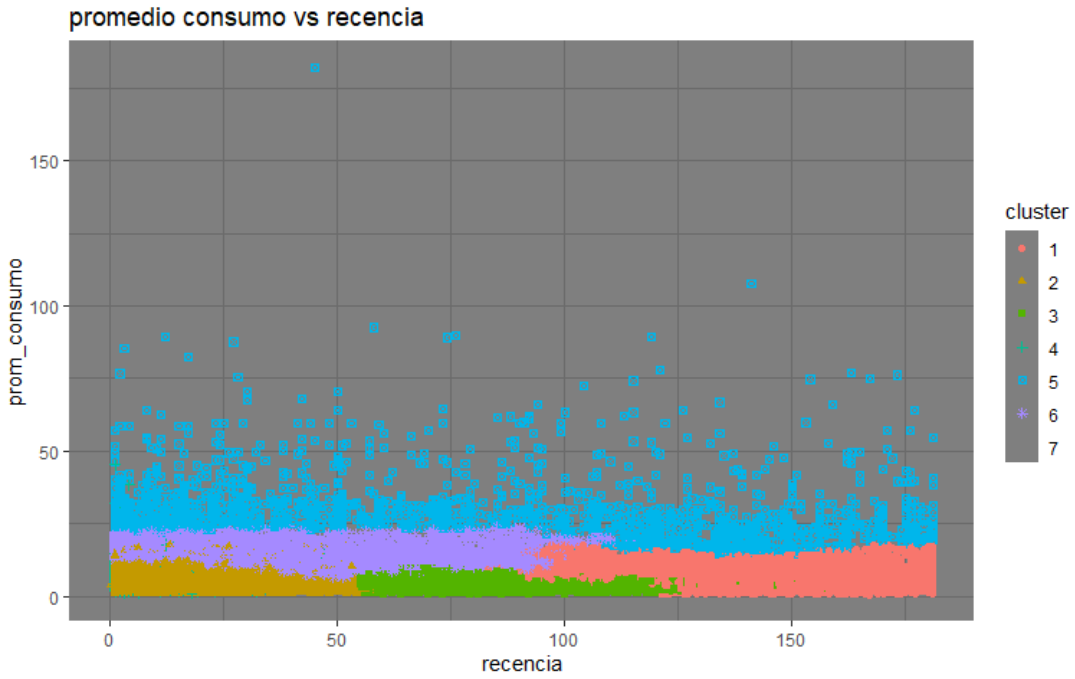


Figura 18 Visualización de clústeres promedio consumo vs recencia. Fuente: elaboración propia.

Con la función prcomp() en el siguiente código se realiza un análisis de componentes principales sobre el conjunto de datos entrenados.

```
## Proyectar los segmentos con los primeros dos componentes principales
prin_comp <- prcomp(clust_data)
nComp <- 2
project <- predict(prin_comp, newdata=clust_data)[,1:nComp]

## visualizar los segmentos con los datos transformados
project_data <- cbind(as.data.frame(project),
                      cluster = as.factor(clusters))

ggplot(project_data, aes(x=PC1, y=PC2)) +
  geom_point(aes(shape=cluster, colour = cluster)) +
  theme_dark() +
  labs(x = "Componente Principal 1", y = "Componente Principal 2")
```

Esta es una visualización bidimensional más fiel de los grupos que las gráficas anteriores que solo muestran combinaciones de datos por pares tal como se observa en la figura 19.

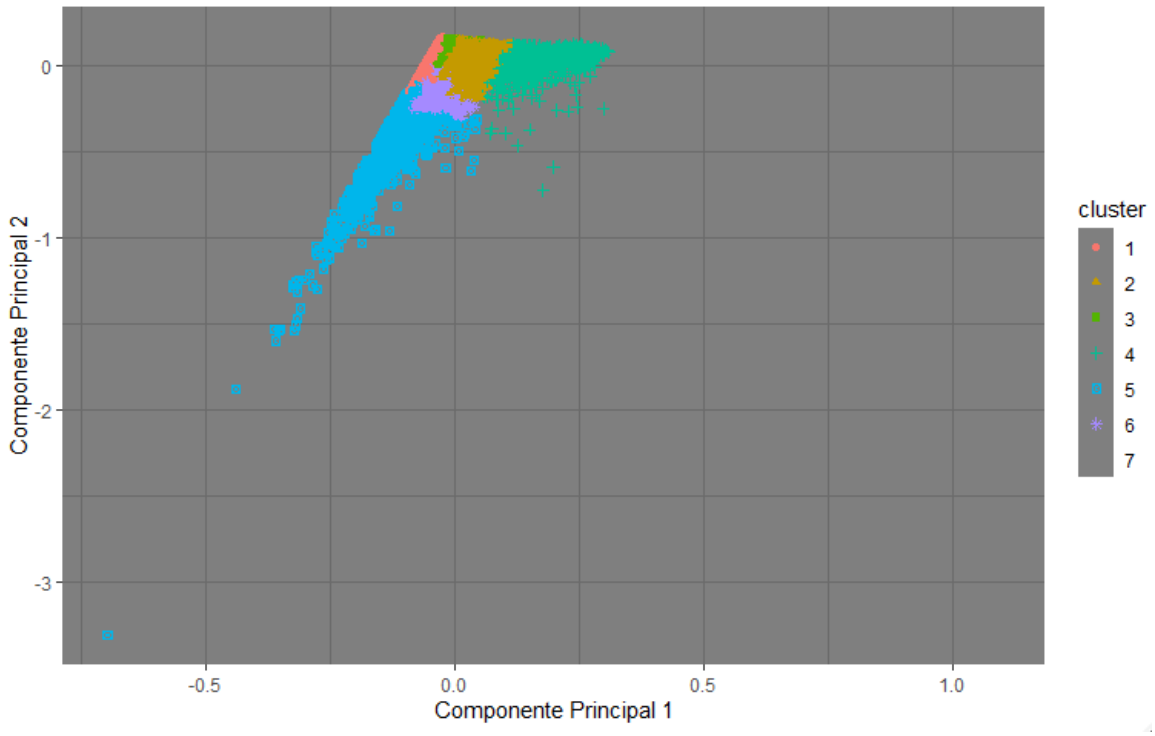


Figura 19 Visualización de clústeres usando los primeros dos componentes principales. Fuente: elaboración propia.

Anexo 5. Calcular la distancia al punto cero.

Calcular la distancia al punto cero, es muy importante porque ayuda a determinar el valor de cada grupo, los clientes con una distancia cero más alta son los clientes más leales y los clientes con distancia cero más baja son los menos leales.

$$D_1 = \sqrt{(v_{11} - 0)^2 + (v_{12} - 0)^2 + (v_{13} - 0)^2},$$
$$D_2 = \sqrt{(v_{21} - 0)^2 + (v_{22} - 0)^2 + (v_{23} - 0)^2},$$
$$\vdots$$
$$D_5 = \sqrt{(v_{51} - 0)^2 + (v_{52} - 0)^2 + (v_{53} - 0)^2}.$$

Figura 20. Fórmula para calcular la distancia al punto cero (Ching-Hsue & You-Shyang, 2009)

El código usado en R para calcular la distancia de cada clúster al punto cero es el siguiente:

```
## tomamos los centros de cada grupo
centros<- aggregate(customers,by=list(customers$cluster), FUN=mean)

## calculando la distancia al punto cero de cada grupo
for (i in c(1:nrow(centros))){
  centros$distanciaCero[i]<-sqrt(sum((centros$recencia[i]-0)^2,
                                     (centros$frecuencia[i]-0)^2,
                                     (centros$prom_consumo[i]-0)^2))
}
```

Anexo 6. Determinar los parámetros del mapa auto-organizado de Kohonen

El tamaño del mapa fue determinado en base a la siguiente regla $\mathbf{5} * \sqrt{\boldsymbol{N}}$ sugerida en (Vesanto & Alhoniemi, 2000) donde N es el número de la muestra de datos, el valor obtenido será el número de nodos a utilizar.

De acuerdo con la fórmula mencionada, para la muestra analizada se necesitan 900 nodos, por lo que se estableció el tamaño del mapa como 30x30.

El número de iteraciones se estableció realizando algunas pruebas con diferentes valores, y en base a la verificación de la estabilidad de los grupos se fijó el valor de 100 iteraciones.

Anexo 7. Visualización de los mapas auto-organizados mediante distintos gráficos.

La inspección visual es uno de los pasos más interesantes del análisis de conglomerados SOM ya que revela los patrones ocultos del conjunto de datos original.

La visualización proporciona una visión instantánea de las formas de los grupos, número de grupos, las relaciones entre grupos, así como la existencia de subgrupos. Al mirar los mapas, se pueden encontrar datos interesantes sobre el conjunto de datos. Hay una gran cantidad de presentaciones diferentes para los mapas. A continuación, se muestran algunos que corresponden a la segmentación realizada para el conjunto de datos RFM.

Progreso del entrenamiento:

A medida que avanzan las iteraciones de entrenamiento de SOM, se reduce la distancia de los pesos de cada nodo a las muestras representadas por ese nodo. Idealmente, esta distancia debería alcanzar una meseta mínima. Esta opción de gráfico muestra el progreso en el tiempo. Si la curva disminuye continuamente, se requieren más iteraciones (Lynn, 2014). Ejecutando el siguiente código se obtiene la tendencia del progreso de entrenamiento para encontrar los clústeres tal como se observa en la figura 21.

```
plot(som_clientes, type = "changes")
```

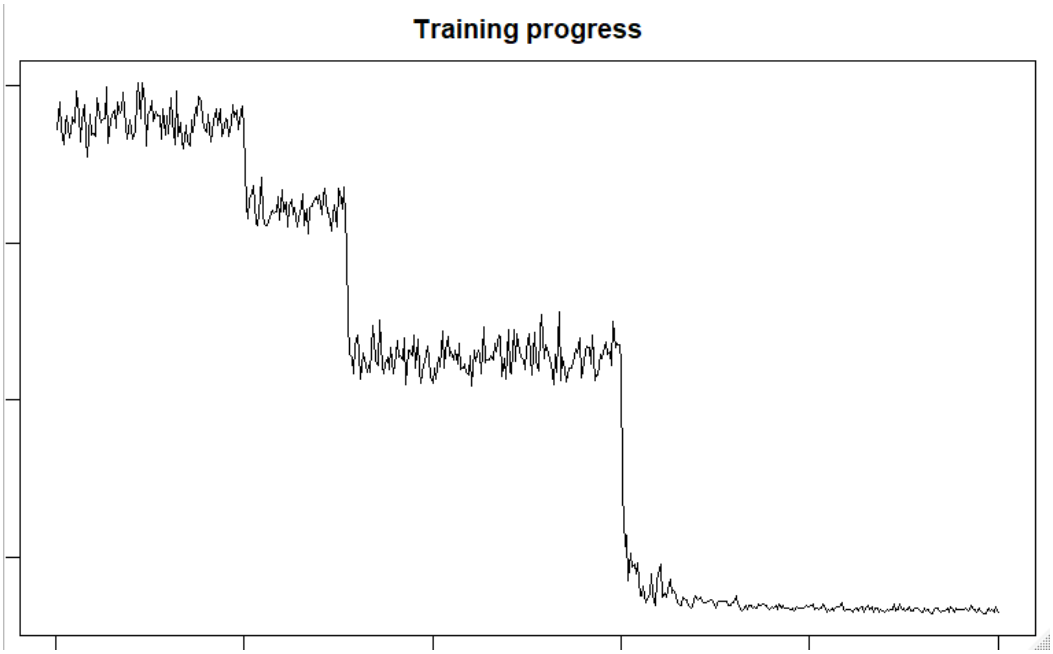


Figura 21 Progreso de entrenamiento para encontrar los clústeres.
Fuente: elaboración propia.

Mapas de calor

Un mapa de calor SOM permite visualizar la distribución de una sola variable a través del mapa. Las figuras 22, 23 y 24 muestran la distribución de las variables RFM, respectivamente usando el siguiente código:

```
## Crear Mapas de calor Kohonen

heatmap.som <- function(model){
  for (i in 1:3) {
    plot(model, type = "property", property = getCodes(model)[,i],
         main = colnames(getCodes(model))[i])
  }
}
heatmap.som(som_model)
```

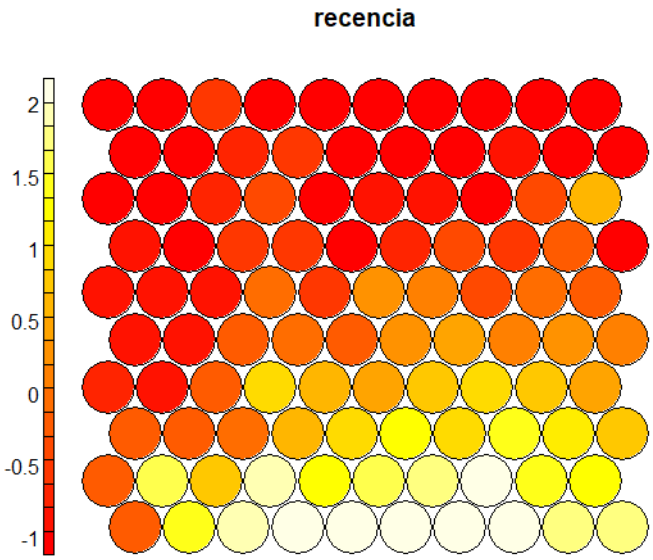


Figura 22 Mapa de calor de la variable recencia. Fuente: elaboración propia

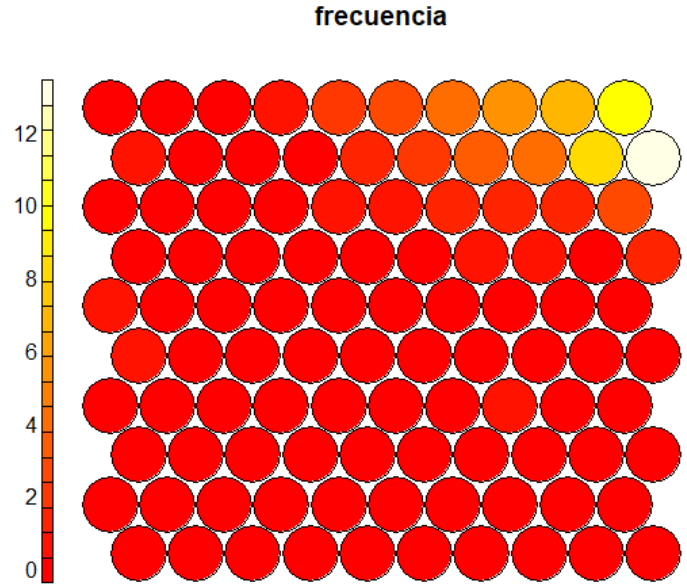


Figura 23 Mapa de calor de la variable frecuencia. Fuente: elaboración propia.

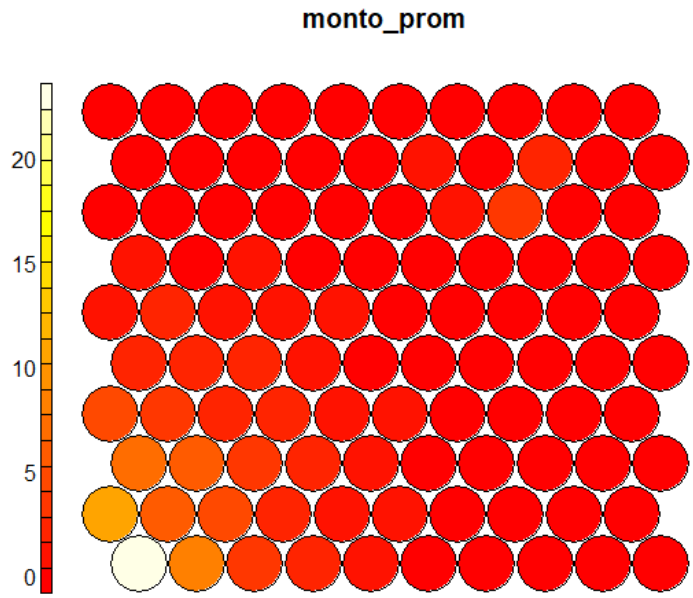


Figura 24 Mapa de calor de la variable promedio de consumo.
Fuente: elaboración propia.

A partir del mapa de calor que se forma, se puede saber qué nodos tienen las características de cada variable cuyo valor es alto y el valor bajo.