UNIVERSIDAD NACIONAL DE SAN MARTÍN

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO

CONCURSO DE PROYECTOS DE INVESTIGACIÓN PARA TESIS A NIVEL DE PREGRADO 2020

Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA

Tesis para optar el Título Profesional de Ingeniero de Sistemas e Informática

AUTOR:

Jaime Hugo Chacaliaza Almeyda

ASESOR:

Ing. Richard Enrique Injante Ore

Tarapoto - Perú

2021

UNIVERSIDAD NACIONAL DE SAN MARTÍN

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO

CONCURSO DE PROYECTOS DE INVESTIGACIÓN PARA TESIS A NIVEL DE PREGRADO 2020

Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA

AUTOR:

Jaime Hugo Chacaliaza Almeyda

Sustentada y aprobada el 29 de octubre del 2021, ante el honorable jurado:

.....

Ing. Dr. Carlos Enrique López Rodríguez

Presidente

.....

Ing. M.Sc. Andy Hirvyn Rucoba Reátegui

Secretario

.....

Ing. M.Sc. Pedro Antonio Gonzales Sanchez

Vocal

Declaratoria de autenticidad

Jaime Hugo Chacaliaza Almeyda, con DNI N° 46835939, bachiller de la Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de San Martín, autor de la tesis titulada: **Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA**.

Declaro bajo juramento que:

1. La tesis presentada es de mi autoría.

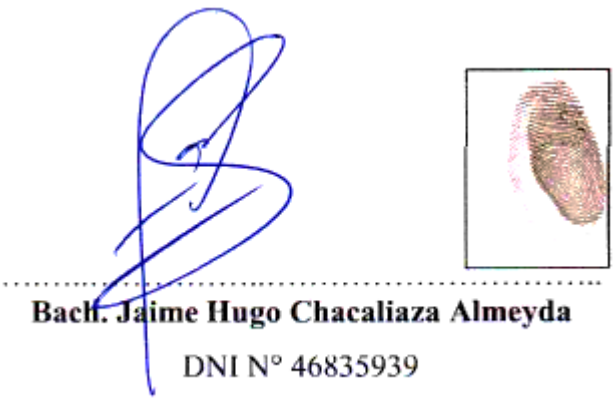
2. La redacción fue realizada respetando las citas y referencias de las fuentes bibliográficas consultadas.

3. Toda información que contiene la tesis no ha sido autoplagiada.

4. Los datos presentados en los resultados son reales, no han sido alterados ni copiados, por lo tanto, la investigación debe considerarse como parte a la realidad investigada.

Por lo antes mencionado, asumo bajo responsabilidad las consecuencias que deriven mi accionar, sometiéndome a las leyes de nuestro país y normas vigentes de la Universidad Nacional de San Martín – Tarapoto.

Tarapoto, 29 de octubre del 2021.



Bach. Jaime Hugo Chacaliaza Almeyda

DNI N° 46835939

Formato de autorización NO EXCLUSIVA para la publicación de trabajos de investigación, conducentes a optar grados académicos y títulos profesionales en el Repositorio Digital de Tesis

1. Datos del autor:

Apellidos y nombres:	Cracaliaza Almeyda Jaime Hugo
Código de alumno : 087103	**Teléfono:** 975725818
Correo electrónico : joch791@gmail.com	**DNI:** 46835739

(En caso haya más autores, llenar un formulario por autor)

2. Datos Académicos

Facultad de:	Ingeniería de Sistemas e Informática
Escuela Profesional de:	Ingeniería de Sistemas e Informática

3. Tipo de trabajo de investigación

Tesis	(X)	Trabajo de investigación	( )
Trabajo de suficiencia profesional	( )		

4. Datos del Trabajo de investigación

Titulo :	Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del Oriente S.A.
Año de publicación:	2021

5. Tipo de Acceso al documento

Acceso público *	(X)	Embargo	( )
Acceso restringido **	( )		

Si el autor elige el tipo de acceso abierto o público, otorga a la Universidad Nacional de San Martín – Tarapoto, una licencia **No Exclusiva**, para publicar, conservar y sin modificar su contenido, pueda convertirla a cualquier formato de fichero, medio o soporte, siempre con fines de seguridad, preservación y difusión en el Repositorio de Tesis Digital. Respetando siempre los Derechos de Autor y Propiedad Intelectual de acuerdo y en el Marco de la Ley 822.

En caso que el autor elija la segunda opción, es necesario y obligatorio que indique el sustento correspondiente:

6. Originalidad del archivo digital.

Por el presente dejo constancia que el archivo digital que entrego a la Universidad Nacional de San Martín - Tarapoto, como parte del proceso conducente a obtener el título profesional o grado académico, es la versión final del trabajo de investigación sustentado y aprobado por el Jurado.

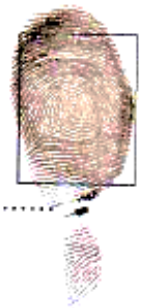
7. Otorgamiento de una licencia *CREATIVE COMMONS*

Para investigaciones que son de acceso abierto se les otorgó una licencia *Creative Commons*, con la finalidad de que cualquier usuario pueda acceder a la obra, bajo los términos que dicha licencia implica

https://creativecommons.org/licenses/by-nc-sa/2.5/pe/

El autor, por medio de este documento, autoriza a la Universidad Nacional de San Martín - Tarapoto, publicar su trabajo de investigación en formato digital en el Repositorio Digital de Tesis, al cual se podrá acceder, preservar y difundir de forma libre y gratuita, de manera íntegra a todo el documento.

Según el inciso 12.2, del artículo 12° del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales - RENATI **"Las universidades, instituciones y escuelas de educación superior tienen como obligación registrar todos los trabajos de investigación y proyectos, incluyendo los metadatos en sus *repositorios institucionales precisando si son de acceso abierto o restringido*, los cuales serán posteriormente recolectados por el Repositorio Digital RENATI, a través del Repositorio ALICIA"**.



Firma del Autor

8. Para ser llenado en el Repositorio Digital de Ciencia, Tecnología e Innovación de Acceso Abierto de la UNSM - T.

Fecha de recepción del documento.

19/11/2021

UNIVERSIDAD NACIONAL DE SAN MARTÍN
Repositorio Digital de Ciencia, Tecnología e Innovación de Acceso Abierto – UNSM.

Ing. M.Sc. Alfredo Ramos Perea
Responsable

***Acceso abierto**: uso lícito que confiere un titular de derechos de propiedad intelectual a cualquier persona, para que pueda acceder de manera inmediata y gratuita a una obra, datos procesados o estadísticas de monitoreo, sin necesidad de registro, suscripción, ni pago, estando autorizada a leerla, descargarla, reproducirla, distribuirla, imprimirla, buscarla y enlazar textos completos (Reglamento de la Ley No 30035).

** **Acceso restringido:** el documento no se visualizará en el Repositorio.

Dedicatoria

Lleno de regocijo, amor y esperanza, dedico esta investigación a mis adorados padres: José y Mercedes, por acompañarme y siempre desear lo mejor para mí, gracias por creer en mí y poder cumplir con excelencia el desarrollo de esta tesis.

A la memoria de mi tío Jesús, te recordaré cada día que reste de mi vida, con tu sonrisa y tu alegría.

También a mi hermana Joselin por formar parte importante en mi crecimiento.

A Eva Margarita, gracias por aparecer en mi vida, y ser el motor de mi día a día.

Y sin dejar atrás a todos mis amigos, docentes de mi facultad, por formar parte de mi desarrollo profesional.

Jaime Hugo Chacaliaza Almeyda.

Agradecimientos

Gracias a Dios por permitirme tener y disfrutar a mis padres y a mi hija, gracias a ellos por apoyarme en cada decisión y proyecto.

No ha sido sencillo el camino hasta ahora, pero gracias a sus aportes, a su amor, a su inmensa bondad y apoyo, lo complicado de lograr esta meta se ha notado menos.

Un agradecimiento especial a mi asesor Ing. Richard Injante, gracias por su paciencia y dedicación con mi persona.

Jaime Hugo Chacaliaza Almeyda.

Índice general

Índice de tablas

Índice de figuras

Resumen

La presente investigación titulada: “Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA”, tuvo como objetivo general: Diseñar un modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del Oriente SA. Con los métodos usados se tuvo una investigación de tipo aplicada a nivel descriptivo y diseño descriptivo - correlacional. Para el diseño del modelo se utilizó la metodología KDD para todo el proceso de minería de datos combinando también el análisis RFM y el enfoque de cuartiles, en los datos extraídos del sistema ERP de la organización mediante el lenguaje R. Concluyendo que el modelo basado en técnicas de minería de datos segmentó a los clientes de la empresa distribuidora Suministros del Oriente SA (Hi).

Palabras clave: Minería de datos, técnicas de segmentación, modelo KDD

Abstract

This research entitled: "Model based on data mining techniques for customer segmentation in the distribution company Suministros del Oriente SA", had as general objective: To design a model based on data mining techniques for customer segmentation in the distribution company Suministros del Oriente SA. The methods used were applied research at a descriptive level and descriptive-correlational design. For the design of the model the KDD methodology was used for the whole data mining process combining also the RFM analysis and the quartiles approach, in the data extracted from the ERP system of the organization using the R language. Concluding that the model based on data mining techniques segmented the customers of the distribution company Suministros del Oriente SA (Hi).

Keywords: Data mining, segmentation techniques, KDD model.

Introducción

La competencia entre empresas para posicionarse en el mercado ha generado que las mismas busquen diversas estrategias, con la finalidad de mantener y atraer nuevos clientes, una alternativa que trae beneficios a las empresas es la implementación de un modelo de segmentación de clientes.

De acuerdo al principio de Pareto (Srivastava, 2016), se evidencia que es el 20% de clientes quienes más aportan más ingresos a la empresa que el 80% restante. En el proceso de segmentación, podemos utilizar una variedad de características únicas de los clientes que permitan a los empresarios a personalizar planes de marketing, tendencias, campañas publicitarias. La división o segmentación de clientes proporciona una mejor comprensión de las necesidades de ellos y a su vez identificar aquellos que la empresa considera potenciales (Christy et. al, 2018).

Muchas investigaciones a nivel internacional mencionan la importancia de aplicar técnicas de clustering para encontrar segmentos de clientes. Las revisiones literarias de referencia en esta investigación apuntaron a la definición de características de los grupos de cliente. Autores como: Dursun y Caber (2016), Cuadros et al. (2017), Bachtiar (2018) y Aryuni et al. (2018), combinan las técnicas de clustering más el análisis RFM para la agrupación de clientes mientras que Flores et al. (2019) y Qadadeh y Abdallah (2018) solo emplearon técnicas de clustering.

En el ámbito nacional, se tienen referencia de investigaciones para lograr el posicionamiento de marca de materiales de construcción con la utilización un método multivariante para segmentar a los clientes (Reyes, 2018) y en la personalización de ventas de servicios de agencias turísticas utilizando K-means (De la Cruz, 2017).

En la presente investigación titulada “Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA”, parte a base de la problemática de la necesidad de las organizaciones en identificar los grupos de clientes y sus características comerciales de manera que faciliten la toma de decisiones para la aplicación adecuada de estrategias comerciales. Esta situación no es ajena a la realidad en nuestra región, tal es el caso de la empresa Suministros del Oriente, donde la utilización de forma convencional para segmentar los clientes con los reportes que brinda el sistema

comercial, el poco conocimiento de las características comerciales de los grupos de clientes, y el desconocimiento de herramientas informáticas que sirvan como soporte en el proceso de segmentación; han generado que el proceso de segmentación no sea adecuado a la realidad del mercado competitivo en el que se encuentran.

La inexistencia de un modelo de segmentación en la mencionada empresa proporciona información insuficiente sobre los grupos clientes para facilitar la toma de decisiones, además genera una demora en la ejecución de planes de acción para mejorar la participación de la empresa en el mercado y un nivel bajo de lealtad de los clientes más importantes en la empresa. Por esta situación se formula la siguiente interrogante: ¿Cómo mejorar el proceso de segmentación de clientes en la empresa Suministros del Oriente SA?; con el objetivo general: Diseñar un modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del Oriente SA. El trabajo de investigación se basa en la siguiente hipótesis: El modelo basado en técnicas de minería de datos segmentará a los clientes en la empresa distribuidora Suministros del Oriente SA, El modelo de segmentación está basado en la metodología KDD para minería de datos, utilizando el análisis RFM y el enfoque de cuartiles, obteniendo una secuencia de varios pasos en la ejecución del modelo, conllevando a una mejor identificación de los grupos de clientes y productos para una mejor implementación de estrategias y toma de decisiones.

El presente trabajo de investigación se divide en tres capítulos fundamentales:
Capítulo I: denominado Revisión Bibliográfica, en donde se expone el planteamiento, antecedentes y formulación del problema, también se justifica la investigación. Adicionalmente, se incluye el marco teórico, que comprenden todo el conglomerado de teorías relacionados al tema. Capítulo II: denominado Materiales y Métodos, comprende la metodología realmente aplicada, las técnicas y herramientas empleadas, donde además se hace la prueba de hipótesis. Capítulo III: denominado Resultados y Discusión, respectivamente, en donde se describen el modelo y se exponen los resultados obtenidos.

Finalmente se presentan las conclusiones que vienen a ser las consecuencias lógicas, las deducciones y los logros más importantes del trabajo de investigación; y las recomendaciones, en donde se indican un conjunto de sugerencias.

CAPÍTULO I
REVISIÓN BIBLIOGRÁFICA

1.1 Planteamiento del problema

Al-Hagery et al. (2015) mencionan que las empresas guardan y generan gran cantidad de datos todos los días, pero del modo del que están almacenadas; por lo general, no suministra un beneficio directo a la organización. El valor radica en la información que se puede obtener al analizarlos, es decir, cuando participan en el proceso de toma de decisiones o ayudan a comprender el fenómeno que los origina. En ese sentido, Aryuni et al. (2019) señala que el rápido crecimiento y magnitud de las bases de datos comerciales, exigen que los a todos los profesionales vinculados a la empresa asuman el compromiso de entender quiénes y cómo son cada uno de sus clientes, debido que para la empresa los clientes tienen diversas prioridades. En el límite, cada cliente individual tiene necesidades e intereses únicos que la empresa puede aprovechar dentro de sus estrategias comerciales. Por otro lado, Dursun y Caber (2016) señalan que comprender y analizar a los clientes es de suma importancia para la organización ya que con esta información pueden ofrecer productos personalizados.

Adicionalmente, Qadadeh y Abdallah (2018) indican que las empresas comerciales necesitan detectar las características de los clientes y analizarlas por segmentos para establecer estrategias y campañas de marketing personalizadas en lugar de ofrecer un solo plan para todos. Para Cuadros et al. (2017) identificar el valor de los clientes, nos suministra información básica para idear planes de marketing exitosos, considerando que deben enfocarse los esfuerzos en entablar relaciones competitivas y de éxito. Además, Bachtiar (2018) indica que las características de cada cliente, permiten diferenciarlos y comprender lo que realmente necesita cada segmento en específico.

La empresa Suministros del oriente SA está ubicada en ciudad de Tarapoto y cuenta con 5 sucursales en el Oriente del país. Además de contar con una amplia gama de productos y a pesar de tener más de 20 años en la región, cuenta con una deficiente identificación de las características de sus clientes, debido a que realizan sus procesos de segmentación de forma empírica lo cual impide aplicar estrategias comerciales y diferenciación de precios.

Al no contar con un segmento objetivo identificado, se complica al área comercial elaborar acciones para incrementar el volumen de las ventas o generar lealtad para alguna de las marcas o categorías de productos que ofrece, es así que el segmento del mercado o de clientes al que se dirige puede ser diferente cada vez. Cabe mencionar que, la empresa emplea un sistema ERP, el cual solo cuenta con reportes gerenciales básicos, ya que existe cierto desconocimiento sobre las tecnologías de información y específicamente el uso de técnicas de minería de datos que le permitan aprovechar todo el potencial de esta tecnología.

Según los problemas planteados se identificó la existencia de una deficiente segmentación de clientes por lo cual, surge la necesidad de proveer información estratégica a la empresa para que con ella puedan tomar de decisiones que tengan impacto a nivel comercial, para ello se propone un modelo basado técnicas de minería de datos para la organización, de forma que puedan convertirse en un valor agregado para el área comercial.

1.2 Antecedentes de la investigación

Internacional

Dursun y Caber (2016) mencionan que la creación de relaciones a largo plazo con los clientes más rentables mejora el éxito de las empresas, para ello los datos que se almacenan diariamente tienen que ser transformados en información útil, por ello se plantearon como objetivo perfilar a los clientes más rentables asumiendo que estudios previos en el sector servicios reportan que el 15% de los clientes generan un 45% de ingresos y un 70% de ganancias, para ello utilizaron el análisis RFM, los mapas autoorganización- SOM (detectar el número de clústers) y K-means (definir los clústers) para segmentar a clientes de 3 grandes cadenas hoteleras en Turquía, obteniendo como resultado la identificación de 8 categorías de clientes y sus respectivas características.

Por su parte, Cuadros et al. (2017) indica que la identificación del verdadero valor de los clientes proporciona información básica para implementar estrategia de Marketing más dirigidas y personalizadas, por ello su objetivo fue incluir nuevas variables, que puede tener un portafolio de clientes, a las que emplea el análisis RFM

justificando su inclusión con técnicas de análisis multivariable para segmentar a los clientes, resultando 5 grupos de clientes y la descripción de sus características.

De la misma forma, Qadadeh y Abdallah (2018) afirman que la detección de caracateristicas comerciales en los clientes permite establecer estrategias y campañas de marketing más efectivas, su objetivo fue descubrir grupos significativos de clientes utilizando técnicas de clustering a los datos de alta dimensión de la CRM de T.H.E INSURANCE COMPANY (TIC), para ello utilizaron técnicas de minería de datos, (K-means y SOM-Kmeans), el método del codo (para definir el número de clúster) y el índice de Davies Bouldin (para determinar el grado de agrupación de los elementos), los resultados indicaron una mejor segmentación cuando se empleo la combinación de las técnicas SOM y K-MEANS.

Adicionalmente, Bachtiar (2018) prioriza la identificacion de características principales en los clientes para comprender los productos y beneficios que nececsitan cada segmento en específico, para ello tuvieron como objetivo implementar un método de minería de datos en dos pasos basado en RFM y la gestión de relación con el cliente, para ello aplicaron el modelo RFM y K-means , luego utilizaron el algoritmo a priori para una mejor descripción de los segmentos creados, logrando una interpretación más descriptiva de los grupos encontrados.

Del mismo modo, Aryuni et al. (2018) señala la importancia de construir modelos de agrupación de datos del perfil del cliente en función de uso en las transacciones bancarias por internet, teniendo como objetivo aplicar la segmentación de clientes utilizando el métódo K-means y K-medoides, en función de su puntaje RFM de transacciones de banca por internet, con la finalidad de comparar el rendimiento de ambos métodos de agrupación, adicionalmente a las técnicas utilizadas se empleo la metodología KDD, obteniendo Kmeans como mejores resultados en la comparación de los rendimientos de agrupamiento AWC y DBI.

Finalmente, Flores et al. (2019), afirman que el crecimiento y la magnitud del mercado electrónico exigen conocer el comportamiento y características de los consumidores, para ello fijaron como objetivo segmentar a los compradores online en base a la frecuencia de consumo en internet y definir los perfiles basados en características demográficas y conductuales de cada miembro utilizando Kmean, resultando tres grupos de consumidores electrónicos con sus respectivas caracteristicas.

En la Tabla 1, vemos que las investigaciones a nivel internacional, muestran una tendencia de uso del algoritmo K-meas y RFM, para poder realizar la segmentación de clientes que permita definir las características de cada grupo.

Tabla 1

Resumen de Antecedentes Internacionales.

Autor(es)	Técnica(s) de Clustering utilizadas	RFM	Definición de características
Dursun y Caber (2016)	SOM , K-means	SI	SI
Cuadros et al. (2017)	K-means	SI	SI
Qadadeh y Abdallah (2018)	SOM, K-means	NO	SI
Bachtiar (2018)	K-means	SI	SI
Aryuni et al. (2018)	K-means, K-medoids	SI	SI
Flores et al. (2019)	K-means	NO	SI

Fuente: Elaboración Propia

Nacional

En el ámbito nacional, Laura et al. (2016) en su artículo afirma que la utilización de técnicas de no supervisadas de minería de datos contribuye a una correcta segmentación de alumnos, por ello el objetivo fue analizar diversas técnicas de minería de datos comparando tres técnicas clustering jerárquico aglomerativo, K-means y PAM en un grupo de alumnos a partir de datos académicos, obteniendo como resultado tres agrupaciones correspondiente a BAJO, MEDIO, ALTO y eligiendo al algoritmo K-means por representar mayor homogeneidad.

Mientras que, Reyes (2018) especifica en su investigación que la segmentación de clientes de materiales de construcción se determina de acuerdo a los atributos y marcas de los mismos, además de las características socioeconómicas del cliente, su objetivo fue determinar las características que determinan la segmentación de clientes y el posicionamiento de marcas de materiales de construcción, utilizando el método multivariante de clasificación jerárquica (Análisis de clúster) para segmentar clientes y el método de valoraciones ponderadas para determinar el posicionamiento, obteniendo como resultado 2 grupos de clientes con las siguientes variables: edad, categoría ocupacional e ingreso mensual.

Finalmente, De la Cruz (2017) indica que segmentando los clientes se consigue una mayor satisfacción, compromiso y lealtad del mismo, el objetivo en su investigación fue implementar un modelo de inteligencia analítica basado en redes artificiales K-means que permita identificar factores para segmentar y definir el perfil de clientes que utilizan servicios turísticos, obteniendo como resultado la descripción de las características de los turistas de acuerdo a las dimensiones de lealtad: cognitiva, conativa y de acción o conducta.

Local

En el ámbito local no se encontraron investigaciones relacionadas al tema de estudio.

1.3 Formulación del problema

¿Cómo mejorar el proceso de segmentación de clientes en la empresa Suministros del Oriente SA?

1.4 Objetivos

1.4.1. Objetivo general

Diseñar un modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del Oriente SA.

1.4.2. Objetivos específicos

OE1: Identificar las características de los hábitos de compra de los clientes.

OE2: Proponer un modelo basado en KDD y técnicas de minería de datos.

OE3: Recomendar estrategias comerciales de segmentación de clientes.

1.5 Justificación de la investigación

1.5.1. Conveniencia

La presente investigación resultó ser conveniente producto de 2 motivos, la primera de ellas fue porque permitió efectuar un análisis en la cartera de los clientes con la cual está trabajando la empresa Suministros del Oriente SA que no se encuentran diferenciado por segmentos, esto es ocasionado por los

mecanismos tradicionales al momento de ejecutar la categorización de los clientes; el segundo motivo tiene que ver con una nueva alternativa para ejecutar estrategias comerciales con nueva información que antes no estaba al alcance de la Gerencia. Conllevando a una mejor gestión de los clientes y productos en la empresa.

1.5.2. Relevancia social

La investigación reviste de importancia social, debido a que sirve como alternativa en las organizaciones del rubro comercial para la segmentación de clientes con el enfoque de minería de datos, que conllevará a una mejor gestión en la cartera de clientes, logrando mejores resultados en la aplicación de estrategias comerciales.

1.5.3. Valor teórico

La investigación pretende cubrir vacíos de conocimiento en base a la problemática descrita, mediante el análisis de las variables en estudio con la finalidad de responder a los objetivos que se han manifestado.

1.5.4. Implicancia práctica

Se realiza la investigación debido a que existe una necesidad de segmentar clientes en la empresa, para ello se propone un modelo basado en técnicas de minería de datos que brinde reglas de clasificación para los grupos de clientes Adicionalmente permitirá remarcar la revisión conceptual en el campo de segmentación de clientes que servirá de guía al área comercial para la toma de decisiones dentro de la empresa.

1.5.5. Utilidad metodológica

Esta investigación se la presenta un modelo de segmentación de clientes mediante técnicas de minería de datos basado en la metodología KDD, una vez demostrada su validez y confiabilidad puede ser utilizado en otros trabajos de investigación u otras organizaciones.

1.6 Bases teóricas

1.6.1. Segmentación de mercado

Westwood (2016), afirma que clientes diferentes, tienen diferentes necesidades, el producto no precisa el mismo beneficio para todos, y de manera individual, cada cliente lo compra por diferente motivo, por ello la segmentación permite considerar los mercados en los que la empresa en los que tiene y debe tener presencia.

Bernal (2017) señala también que en el mercado se encuentran diferentes empresas con variedades de líneas y productos, para ello es importante identificar los segmentos de mercado que le puedan servir de forma eficaz. La segmentación de mercado tal como lo indica su nombre consiste en la separación de grupos homogéneos en cuanto a necesidades, deseos y comportamientos similares, con la finalidad de conocer realmente a los consumidores. Los segmentos deben tener las siguientes condiciones para poder ser considerados útiles:

a) **Medibles:** Para medir el volumen de compra y características de cada uno de ellos.

b) **Rentables:** Brindar la posibilidad de obtener ganancias

c) **Accesibles:** Contar con los recursos suficientes para llegar al segmento deseado a un costo razonable.

d) **Diferenciables:** Se distinguen por su concepto y atienden de diferente manera a cada uno de los elementos y programas de marketing.

e) **Susceptibles de acción:** Con la posibilidad de proponer programas que puedan ser eficaces para servir y atraer a cada uno de sus elementos.

1.6.1.1. Criterios o variables de segmentación

Marmol y Ojeda (2016), mencionan que un criterio o variable o de segmentación es una característica de los individuos que componen el mercado con la finalidad de identificar grupos que sean lo más homogéneos internamente y diferentes entre sí. La selección de

variables o criterios dependerá de los objetivos perseguidos, los mercados pueden segmentarse a partir de:

a) Una sola variable: Tiene por ventaja ser simple y fácil de usar.

b) Varias Variables: Es más precisa.

Ahora bien, Bernal (2017) explica los siguientes criterios a tener en consideración para la selección de variables:

- Segmentación geográfica: Se basa en que la necesidad de los consumidores varía de acuerdo a su región (países, ciudades, barrios, localidades, densidad, clima, etc.).
- Segmentación demográfica: Es la base más popular para segmentar clientes (edad, sexo, tamaño de familia, estado civil, ingresos, ocupación, religión, etc.).
- Segmentación conductual: Se basa en el conocimiento del cliente sobre el producto, la forma que usan el producto o la forma que responden al mismo (los beneficios del producto, momentos de compra, la situación de usuario, el nivel de consumo, el grado de lealtad, la etapa de preparación del consumidor y su actitud ante ello).
- Segmentación psicográfica: Especifica las características y las reacciones de un individuo ante su medioambiente (pasividad o agresividad o, firmeza o facilidad al cambio, etc.), por lo general describen el estilo de vida, personalidad y valores.

1.6.1.2. Estrategias de cobertura al mercado

Alvarez (2018) explica que, al existir diferentes tipos de clientes, a la organización no le conviene implementar una sola táctica de mercadotecnia a esas diferencias, para ello es importante las 4 P´s de mercadotecnia: (producto, precio, plaza y promoción).

Adicionalmente, Marmol y Ojeda (2016) detallan que luego de realizar la segmentación, la empresa debe generar estrategias para cada uno de los mercados identificados, por lo general se pueden aplicar cuatro tipos:

a) **Estrategia indiferenciada**: o también llamada masiva, emplea solo un marketing-mix (estrategia de negocio, precio de venta, logística, y mercadotecnia), por igual en todos los segmentos, de forma que intenta satisfacer todas las necesidades con una única oferta comercial.

b) **Estrategia diferenciada:** o también llamada segmentada, emplea un marketing-mix distinto para cada segmento objetivo detectado, y ofrece productos de acorde a sus necesidades, intentando conseguir la máxima cobertura de estos

c) **Estrategia concentrada:** o estrategia de nicho, se centra en un solo segmento concreto del mercado, esta estrategia es apropiada para la pequeña y micro empresa, porque fomenta una ventaja comercial competitiva. Por lo general, el producto se adapta totalmente a las necesidades del segmento elegido.

d) **Estrategia de micro-marketing:** o micro-segmentación, la empresa ajusta sus ofertas a las necesidades y preferencias de los consumidores, clientes locales, o individuos específicos.

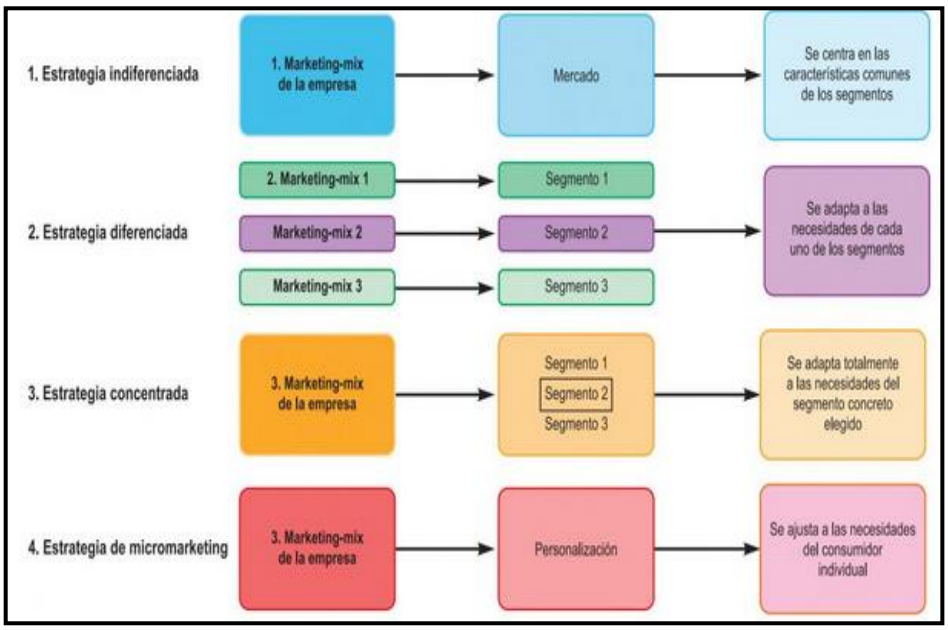


Figura 1. Estrategias de cobertura de mercado de Marmol y Ojeda (2016).

Bernal (2017), añade que algunas funciones de la segmentación ayudan en la creación de las estrategias de marketing, por ejemplo:

- Proporciona descripciones de segmentos para idear programas novedosos y eficaces de mercadeo.
- Da mayor exactitud de la descripción de las características del bien y/o servicio.
- Permite idear un servicio u oferta adecuado y el precio de acuerdo al público que se espera llegar.
- Proporciona la elección de canales de comercialización y mercadotecnia.

1.6.2. Inteligencia de mercados

Arroyo y Borja (2018) indican que la inteligencia de mercados tiene como objetivo reunir, clasificar y distribuir información confiable y oportuna del mercado, además del impacto que tiene sobre ella; la información que es debidamente procesada y analizada, facilita la interacción entre los involucrados, la sistematización de los hechos, el estudio de relaciones y la comprensión de situaciones logrando apoyar de manera efectiva en el proceso de toma de deciones.

Además, las fuentes de donde se obtiene la esta información provienen de cuatro fuentes principales: la propia organización, su cadena de abastecimiento, la competencia y el consumidor. Para ello se requiere contar con procedimientos para que las variables se puedan organizar, analizar y permitan estimar modelos de asociación. En la figura 2, se describen las actividades que comprenden la inteligencia de mercados.

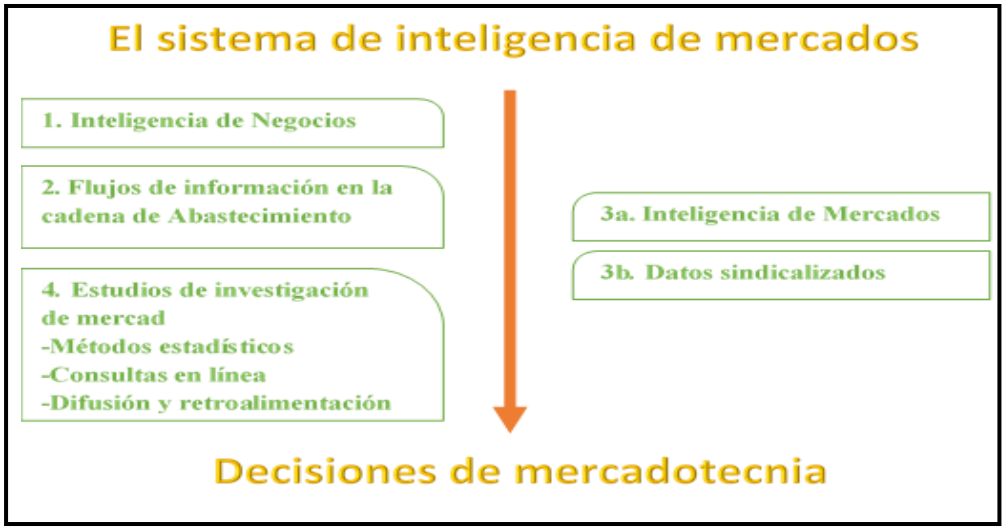


Figura 2. Sistema de inteligencia de mercados de Arroyo y Borja (2018)

Por su parte, Tang (2015) explica que la inteligencia de mercados es la convergencia de la evolución de la inteligencia competitiva con el surgimiento del marketing como un área estrategica en la empresa. Esta comprende todas las actividades fundamentales de las empresas privadas e involucra también las organizaciones públicas, los ámbitos de estudios más importantes son la inteligencia de negocios (BI) y la inteligencia competitiva. En la figura 3, podemos ver algunas clásficaciones de y intelingencia y sus interacciones.

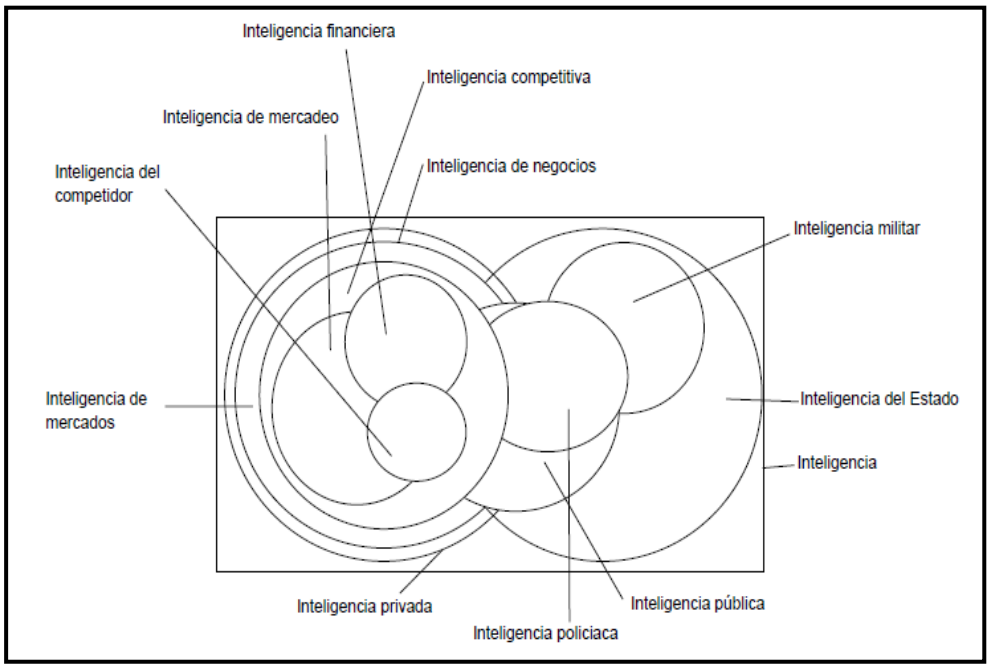


Figura 3. Clasificación de estudios de inteligencia de Tang (2015).

Bernal (2017) señala que la inteligencia de mercados es un proceso de exploración de las variables del cual se explican el comportamiento actual y la tendencia de la oferta, es importante para las organizaciones realizar un estudio de este tipo para incursionar o mejorar su participación en el comercio nacional y/o internacional. Se considera también como un instrumento de información y seguimiento estratégico que vincula variables de marketing empleando diversas herramientas y metodologías para optimizar la toma de decisiones implicando una alternativa para:

- Añadir valor agregado a los esquemas de precios que existen, con análisis de tendencias y tendencias de su comportamiento en un futuro.

- Describir breve y concisamente los datos del estudio de mercado obtenidos en la investigación.
- Obtener información para tomar decisiones acertadas en base a los datos proporcionados por el mercado.
- Establecer relaciones entre las oportunidades, créditos, asesoría técnica y venta.

Bernal (2017) menciona que existen 03 componentes involucrados en la inteligencia de mercados:

a) Marketing estratégico: El origen de la inteligencia de mercados se centra en el planeamiento estratégico de la venta, en el cual los lineamientos estratégicos deben ser coherentes con la visión que se pretende alcanzar.

b) Scoring aplicado al marketing: Busca establecer los indicadores y elaborar los formatos adecuados para el estudio cualitativo y cuantitativo del mercado.

c) Levantar la información correcta: La captura de información relevante y en el tiempo estimado, minimiza los errores muestrales y no muestrales para minimizar el riesgo de las decisiones que se toman en la organización.

1.6.3. Minería de datos

Joyanes (2016) lo define de la siguiente manera: "*Es un proceso que utiliza técnicas estadísticas, matemáticas, inteligencia artificial y de aprendizaje de máquinas para extraer e identificar información útil que convierte en conocimiento a partir de grandes bases de datos, data warehouses o data mart.*"

Ruiz (2017) añade que la minería de datos comprende un conjunto de técnicas, que adecuadamente procesadas permiten obtener conocimiento, que se encuentran implícito en la base de datos. La inteligencia artificial y el análisis estádistico son los pilares fundamentales para abordar problemas de predicción, clasificación y segmentación.

Mientras que, Bernal (2017) señala que el datamining es muy relevante en los negocios porque contribuye a las estrategias y tácticas de la empresa, los modelos de minería de datos pueden aplicarse en los siguientes escenarios:

- **Pronóstico:** Cálculo y predicción de las ventas.
- **Riesgo y probabilidad:** Elección de mejores clientes, punto de equilibro, diagnóstico de aceptación de nuevos productos y/o servicios.
- **Recomendaciones:** Determinación de mix de productos y/o servicios.
- **Búsqueda de secuencias:** Se emplean en el análisis de la compra de los productos y/o servicios artículos que los clientes han adquirido y si estás presentan algún tipo de relación.
- **Agrupación:** Segmentación de clientes o acciones en grupos de que tienen elementos vinculados.

1.6.3.1. Técnicas de minería de datos

Escobar et al. (2016), describen algunas técnicas más usadas de minería de datos, tal como se muestra en la Tabla 2:

Tabla 2

Técnicas más utilizadas de minería de datos

Técnica	Objetivo
Clasificación	Diseñar un modelo para distribuir un caso de clase desconocida a una clase concreta.
Regresión	Obtener un modelo que sirva en la predicción de un valor numérico de una variable. (Regresión logística)
Agrupamiento (clustering)	Encontrar conjunto de datos con características similares, diferenciando el grupo completo de una serie de categorías.
Resumen	Representar de forma compacta el subconjunto de los datos de entrada
Modelado de Dependencias	Obtener descripciones de dependencias existentes en variables.
Análisis de secuencias	Modelar la evaluación temporal de alguna variable, con fines descriptivos o predictivos

Fuente: Escobar et al. (2016)

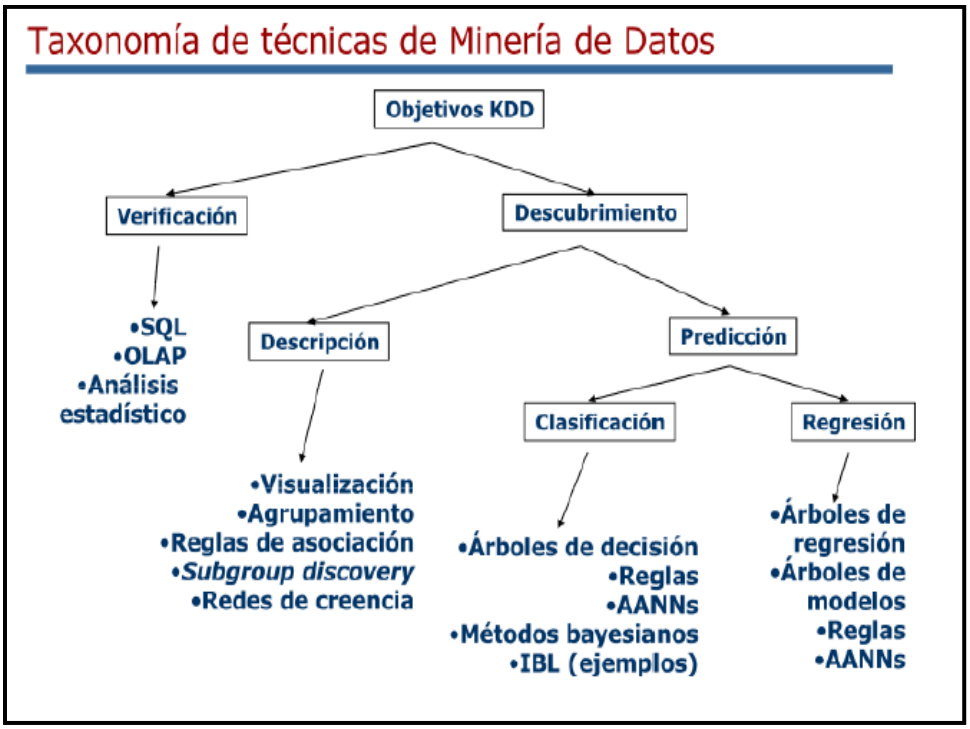


Figura 4. Taxonomía de técnicas de minería de datos de Escobar et al. (2016)

1.6.3.2. Agrupamiento o Clustering

Cabeza (2016), en su artículo científico menciona que agrupar un almacén de datos es trascendental para entender el comportamiento de la población, de la cual solo se conoce una cantidad "n" de sus características, para esto se hace uso de algoritmos de agrupamiento.

Además de ello, Vicente y Mateos (2018) afirman que los algoritmos de clustering se dividen en dos grandes grupos: algoritmos jerárquicos y los algoritmos particionales.

a) **Algoritmos jerárquicos:** Presuponen la existencia de una estructura ramificada de forma que pertenecen a un tronco común que se divide según una función de similitud o distancia. El objetivo es dibujar una estructura ramificada que recibe el nombre de dendrograma.

b) **Algoritmos particionales:** Dividen el espacio en una colección de subconjuntos disjuntos dos a dos que lo recubra, de forma que dos puntos que pertenezcan a la misma clase sean más parecidos o cercanos que dos elementos de distintas clases.

1.6.3.3. Algoritmos de agrupamiento o clustering

a) K-means

Vicente y Mateos (2018) definen que el algoritmo de k medias (también conocido como el algoritmo de Lloyd, es aquel que trata de minimizar las diferencias intraclase mediante la minimización de la distancia de cada elemento a su centro de gravedad, es decir busca "r" clases disjuntas dos a dos que recubran el espacio de forma que se minimice:

$$\sum_{i=1}^{r} \sum_{x_j \in G_i} \frac{d(x_j, c_i)}{n},$$

Dónde: "d" es una función de distancia o disimilaridad.

Méndez (2018) indica que el K-means es una técnica no supervisada ya que no existe ningún resultado para predecir, se basa en asignar aleatoriamente un individuo a cada grupo y mide la distancia de cada individuo al centroide (media de la posición de cada uno de los individuos) del clúster. A partir de aquí, se establece un mecanismo de iteración en el que se re-clasifican los individuos al grupo en el que estén más cerca de sus centroides y éstos se vuelven a recalcular.

Rojas y Sebastián (2017), representan al algoritmo K-means de la siguiente forma:

1. **Entrada:**
    - **K:** el número de segmentos (clústeres),
    - **D:** conjunto de datos que contiene "n" objetos.
2. **Salida: Conjunto de "k" segmentos o clústeres.**
3. **Método**
    - Elegir arbitrariamente "k" objetos de D como los centros de agrupación inicial;
    - Repetir

- (Re)asignar cada objeto al clúster al que el objeto es el más similar, basado en el valor medio de los objetos del clúster
- Actualizar los medios de agrupamiento, es decir, calcular el valor de los objetos para cada grupo.
- Hasta que no cambie.

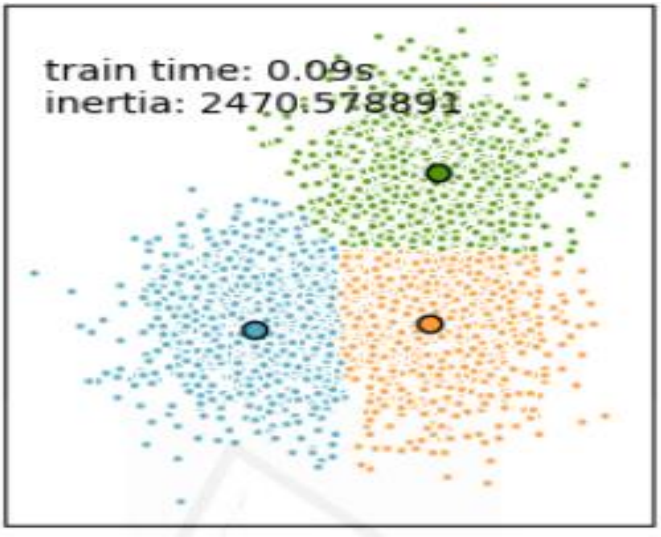


Figura 5. Ejemplo de K-means de Méndez (2018)

b) K-medoides.

Hernandez y Francisco (2017) definen que es una técnica de agrupación tradicional que aglomera el conjunto de datos de "n" objetos en "k" segmentos previamente conocidos, pero que en lugar de tomar el valor medio de los objetos (k-medios) utiliza el objeto más centrico de un clúster (medoide).

Ruiz (2019) lo considera como: "*Una variación del k-means, el objetivo es determinar al mejor representante del centro de cada clúster (medoide). Trabaja con una métrica arbitraria de distancia entre puntos (observaciones), minimiza la suma de diferencia entre los puntos etiquetados para estar en un grupo y el punto designado como el centro"*

1. Entrada:

   - K: número de segmentos,
   - D: Conjunto de datos que contiene "n" objetos.

2. Salida: conjunto de “k” clústeres.

3. Método

- Seleccionar “k” objetos aleatoriamente.
- Calcular C_{ij}: coste O_i ,k_h
- Se asocia cada O_i al k_h medoide más cercano.
- Se establece el Coste Total (CT): suma de la distancia de los puntos a sus medoides.

Mientras Coste configuración disminuye **hacer**:

- Para cada k_h , para cada O_i.
- Intercambiar k_h y O_i, recalcular costo
- Si costo aumentó deshacer cambio.

Fin

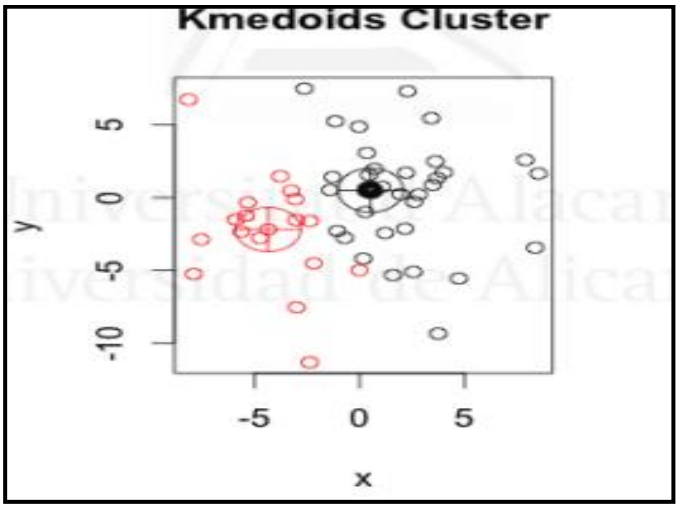


Figura 6. Ejemplo de K-medoides de Méndez (2018)

b.1) Algoritmo PAM

El algorimto PAM es un método tipo k-medoides, que intenta determinar *k* particiones de *n* objetos determinado los objetos representativos de cada conglomerado (medoides). Después de encontrarlos, los grupos se construyen asignando cada observación al medoide más cercano, luego se intercambian cada medoide “m” seleccionado y cada punto de datos no medoideado, se calcula la función objetivo, que corresponde a la suma de las diferencias de todos los objetos de su medoide más cercano.

El paso de intercambio intenta mejorar la agrupación mediante el intercambio de objetos seleccionados y objetos no seleccionados. Si la función se puede reducir intercambiando un objeto seleccionado con un objeto no seleccionado, entonces se realiza el intercambio. Esto continua hasta que la función objetivo ya no se pueda disminuir.

b.2) Algoritmo CLARA

El algoritmo CLARA, separa múltiples muestras y aplica el algoritmo PAM sobre cada una de ellas; luego, encuentra los conjuntos de k-medoides de las muestras. Uno de los principales motivos de Kauffman y Rousseuw (1990) para proponer este algoritmo fue debido a la deficiencia del algoritmo pam para trabajar con base de datos con grandes volúmenes de información. Según esos autores, los resultados de sus experimentos indican que 5 muestras con $(40 + 2k)$ objetos cada una, producen resultados satisfactorios.

La efectividad de Clara depende tanto del tamaño de la muestra como su calidad. Pam busca los mejores *k*-medoides entre un conjunto total de datos, mientras que Clara busca los mejores *k*-medoides entre las muestras seleccionadas del conjunto total de datos. Clara no podría encontrar la mejor agrupación, si los mejores k-medoides no son seleccionados en las muestras (Leiva-Valdebenito y Torres-Avilés, 2010).

c) Algoritmo K- Nearest NeighBors

También llamado K- vecinos más cercanos, es un algoritmo sencillo y local, que necesita especificar la métrica adecuada para medir la proximidad. Consiste en un entrenamiento mediante ejemplos cercanos al espacio de los elementos, es un tipo de algoritmo Lazy Learning, donde la función se aproxima solo localmente, y todo el cómputo es diferido a la clasificación, es decir, un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus vecinos más cercanos. (Ruiz, 2019)

1. **Entrada:**

   - **K: el número de clúster,**
   - **D: un conjunto de datos que contiene n objetos.**
   - **T: un conjunto de datos de test.**

2. **Salida: Un conjunto de "k" clústeres.**

3. **Método**

   Para todo objeto xi e T hacer:

   - Calcular $d_i = d(x_i,x)$

   Fin

   - Ordenar ascendetemente di(i=1, …., N)
   - Escoger los "K" casos $\boldsymbol{D_n^k}$ ya clasificados más cercanos a x
   - Asignamos x a la clase más frecuente en $\boldsymbol{D_n^k}$

A continuación, se muestra un ejemplo utilizando KNN, esta imagen es el resultado de aplicar el algoritmo con valor de parámetro K=7, y utilizar el valor del parámetro weigths = uniform que significa, que algoritmo asigna pesos uniformes a cada vecino.

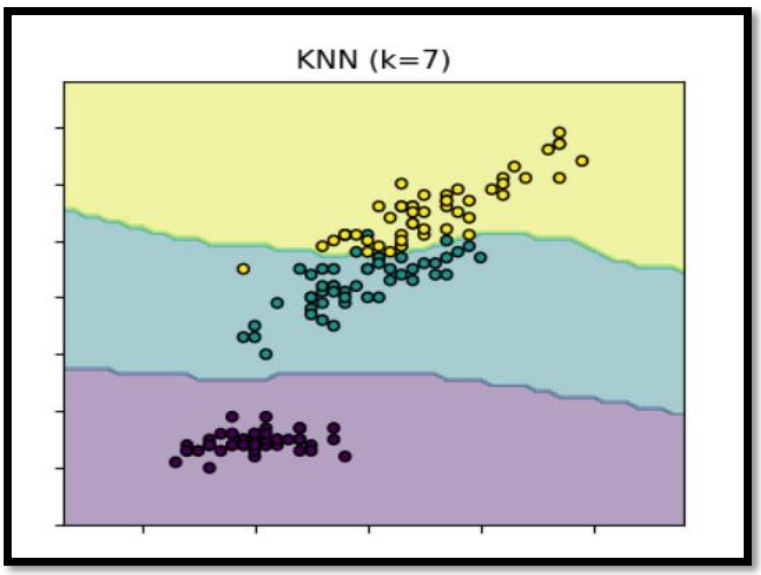


Figura 7. Ejemplo KNN de Méndez (2018)

d) Mapas autoorganizados de Kohonen (SOM)

Peña et al. (2015), afirma que el mapa autoorganizado de Kohonen (en inglés Self Organizing Map) es una red neuronal de entrenamiento no supervisado que tiene la capacidad de representar la estructura de los datos de entrada por medio de la autorganización de sus neuronas. La primera capa lleva los datos a la segunda capa de procesamiento y salida, que se conforma de una malla rectangular. Cada neurona tiene conexión a todos los elementos de entrada por medio de pesos sinápticos. Las neuronas se activan colectivamente con cierta intesidad ante patrones de entrada, describiendo relaciones subyacentes entre ellos. Tal intesidad es representada por los pesos que las neuronas ajustan a través del entrenamiento.

1. Entrada:
    - K: el número de clúster,
    - D: un conjunto de datos que contiene n objetos.
2. Salida: Un conjunto de "k" clústeres.
3. Método

    - Inicializamos los pesos w_{ij}
    - Introducir: Ek =(e_1,…..,e_N), e_i valores continuos

    Mientras t ≤ 500 **hacer**:

    Para *todo objeto* $e_i \in E_k$ hacer:

    Para j ∈ [1, M] hacer:

    - Calcular $d_j = \sum_{i=1}^{N}(e_i - w_{ji})$

    fin

    - Encontramos neurona vencedora j*
    - Actualizamos peso de $zona_j$
    - Calculamos $\beta(t) = \frac{1}{t}$
    - Calculamos $w_{ji}(t+1) = w_{ji}(t) + \beta(t)[e_i^{(k)} - w_{j*i}(t)]$

    fin

    - Incrementar "t"

    Fin

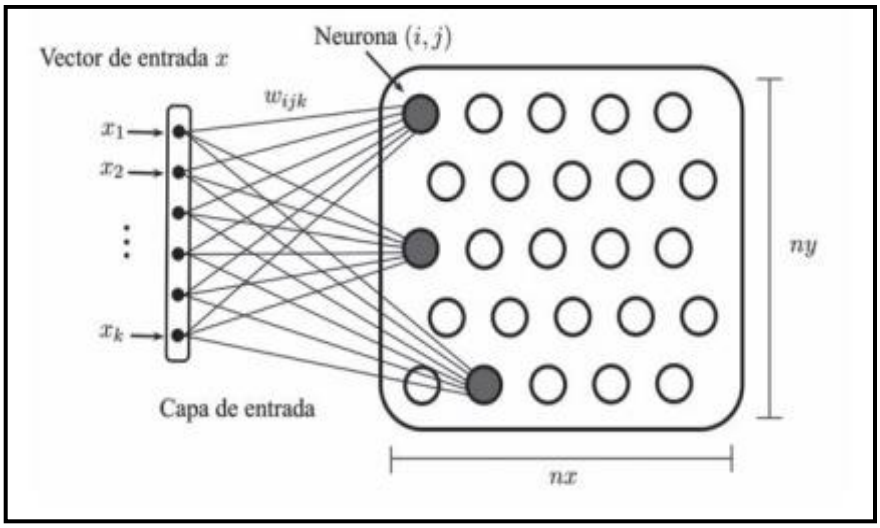


Figura 8. Descripción gráfica de un SOM de Peña et al. (2015)

e) Algoritmo DBSCAN

Vicente y Mateos (2018) definen al algoritmo de agrupamiento espacial basado en densidad de aplicaciones con ruido (conocido por sus siglas en inglés density based spatial clustering of applications with noise). Este algoritmo calcula la densidad de cada región del plano calculando el número de puntos que caen dentro de las esferas de radio *eps* de cada elemento de la población. A partir de esos valores, un entorno tiene densidad adecuada si el número de puntos que yacen en dicho entorno es al menos un valor prefijado *MinPT.*

Ruiz (2019) precisa que el algoritmo DBSCAN determina de manera automática el número de agrupaciones en los que se organizan los datos de entrada. Sin embargo, zonas del espacio con baja densidad se clasifican como ruido y son omitidos, generando que en algunas situaciones no se genere un clustering completo. Además, sirve para identificar cluster en grandes conjuntos de datos espaciales observando la densidad local de elementos base y utilizando solo un parámetro de entrada.

1. **Entrada:**
    - **D: un conjunto de datos que contiene “n” objetos.**
2. **Salida: Un conjunto de “k” clústeres.**

3. Método

- Definimos *MinPT, eps,C=0*
- Etiquetamos los puntos como Central, frontera o ruido

Para cada punto $p_i \in D$ **hacer**:

- Etiquetamos a p_i como visitado
- Verificamos si:
- p_i es punto central: $\exists \boldsymbol{MinPT}$ en un radio $\leq eps$
- p_i es punto frontera: sí la $\boldsymbol{d(p_{central}, p_i) = eps}$
- p_i es punto ruido: $\boldsymbol{d(p_{central}, p_i) > eps}$

fin

Para cada punto central $p_i \in D$ **hacer**:

Para puntos $p_j \in D$ **hacer**:

Si $\boldsymbol{d_i(p_j - p_i \leq eps)}$ entonces

- $\boldsymbol{p_j}$ pertenece al cluster$_i$

Sino

- C=nextcluster

Fin

Fin

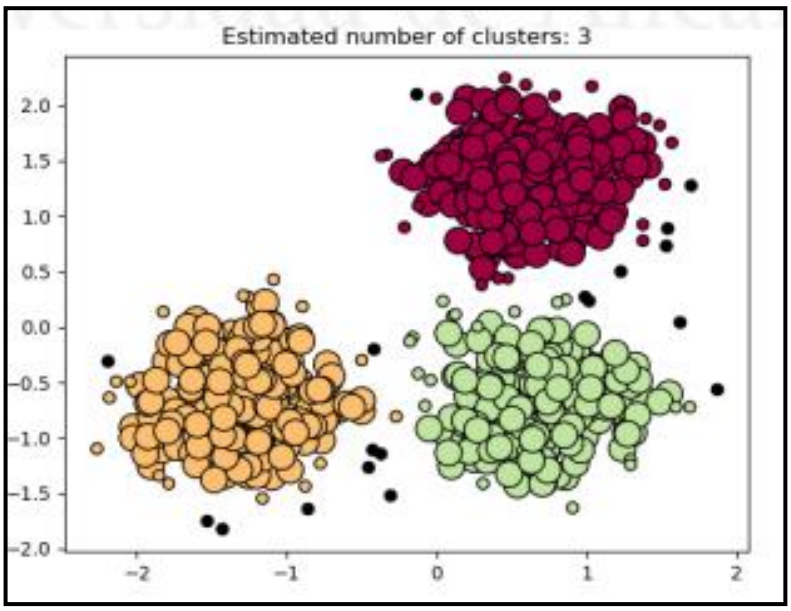


Figura 9. Ejemplo DBSCAN de Méndez (2018)

1.6.3.4. Análisis de Conglomerados o análisis clúster

El análisis de conglomerados se refiere a la creación de grupos con patrones de similitud en sus variables observadas, la idea clave es la representación de "N" objetos o entidades en "K" grupos homogéneos. Una vez caracterizado los grupos, es posible diseñar productos y estrategias de mercadotecnia según el perfil de los segmentos identificados (Arroyo y Borja, 2018).

También, Aldas y Uriel (2017) lo definen como una técnica que emplea distintas observaciones y los clasifica en grupos, consiguiendo:

a) Cada grupo sea homogéneo respecto a las variables utilizadas para caracterizarlo, para especificar que cada observación sea parecida con todas las que estén incluidas en ese grupo.

b) Los grupos sean lo más distintos posible entre ellos, respecto a las variables consideradas.

1.6.3.4.1. Medidas de distancia y Similitud

Según Arroyo y Borja (2018) el primer paso en el análisis de conglomerados es definir una medida de similitud entre los objetos que se desea agrupar. Esta medida debe permitir representar a los objetos como puntos en el espacio, en donde se les puede observar cómo distancias métricas entre ellos y deben cumplir las siguientes propiedades:

- Simetría: la similitud (distancia) entre el objeto "x" y el "y", es igual en forma inversa, es decir: $d(x, y) = d(y, x)$.

- Desigualdad del triángulo: La menor distancia entre 2 objetos está definida por su distancia diagonal, que es menor que la suma de las distancias de los objetos respecto a un tercero (considerando el plano bidimensional), esto es: $d(x, y) < d(x, z) + d(y, z)$.

- Distinción: Si la medida de similitud es diferente de 0, los objetos son diferentes $d(x, y) \neq 0$, entonces $x \neq y$.

- Identidad: Si la medida similitud es de 0, quiere decir que los objetos son idénticos, se expresa como x y x' idénticos, entonces d (x, x') = 0.

a) Distancia Euclideana: Calcula la raíz de la diferencia cuadrática entre coordenadas de un par de puntos u objetos. Se define por la siguiente Formula

$$D_{xy} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

b) Distancia de Manhattan: La distancia entre dos puntos es la suma de las diferencias absolutas de sus coordenadas.

$$D_{xy} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

c) Distancia de Minkowski: Esta distancia se utiliza típicamente con p siendo 1 o 2, que corresponden a la distancia Manhattan y la distancia euclidiana, respectivamente. En el caso límite de p alcanzar el infinito, obtenemos la distancia Chebyshev:

$$D_{xy} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

d) Distancia de Chebyshev: La distancia es la mayor de sus diferencias a lo largo de cualquier dimensión de coordenadas.

$$D_{xy} = \max_i \{|x_i - y_i|\}$$

1.6.3.4.2. Determinación del número óptimo de clústeres

Según Moya 2016, uno de los problemas más frecuentes que se encuentra a la hora de aplicar algunas de las técnicas de clustering es la elección del número de clúster. Se han implementado diferentes métodos de que nos ayudan a elegir el número apropiado de clústeres para la agrupación de datos.

Para Kassambara (2017), existen métodos directos y de prueba estadística:

- Métodos directos: consiste en optimar un criterio, como la suma de cuadrados dentro del clúster o el promedio de la silueta. A estos se le denominan elbow method (método del codo) y average silhouette method

- Métodos de prueba estadística: Consiste en comparar el error esperado bajo una distribución de referencia nula. Un ejemplo es Gap Statistics. (Estadística de la brecha)

a) **Método del Codo**: La idea básica en los métodos de partición, es definir las agrupaciones de tal manera que la variación total dentro de la agrupación se minimice. La suma total de cuadrados dentro de la agrupación (within-cluster sum of square, WSS) mide el espacio compacto del agrupamiento, debiendo ser este el más pequeño posible (Kassambral, 2017), se define por la siguiente fórmula:

$$WSS = \sum_{i=1}^{k} \sum_{x \in C_i} d\left(x, \bar{x}_{C_i}\right)^2$$

Donde:

$$C_i = Clúster\ , k = número\ de\ Clúster, \bar{x}_{C_i} = centroide$$

El número óptimo de clúster se puede definir de la siguiente manera:

- Indicar el algoritmo de agrupación para diferentes valores de k.
- Para cada k, calcule la suma total del cuadrado dentro del grupo (WSS)
- Trazar la curva de WSS de acuerdo con el número de grupos
- La ubicación de una curva (codo) en la gráfica generalmente se considera como un indicador apropiado de grupos,

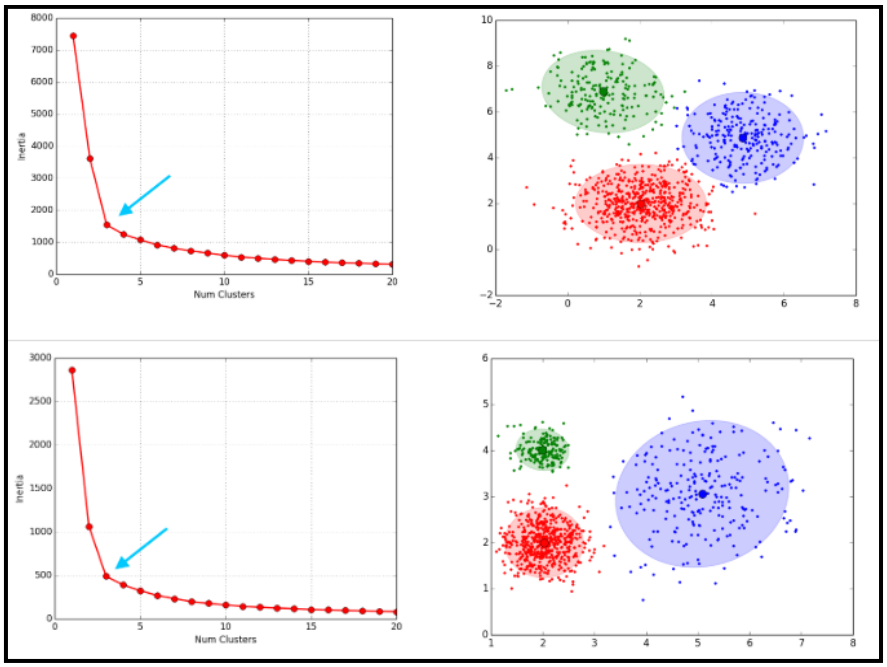


Figura 10. Agrupaciones utilizando el método del codo Moya (2016)

b) **Método de la silueta promedio:** Este método calcula la silueta promedio de las observaciones para valores diferentes de K, midiendo la calidad de una agrupación. El algoritmo es similar al método del codo y se puede calcular de la siguiente manera:

- Indicar el algoritmo de agrupación para diferentes valores de k.
- Para cada k, calculamos la silueta promedio de las observaciones.
- Trazar la curva de acuerdo con el número de clústeres.
- El máximo valor promedio se considerar como el número apropiado de clústeres.

El coeficiente de la silueta para un objeto se define:

$$S(i) = \frac{b-a}{\max(a,b)}$$

Dónde:

a: Es la distancia media entre los objetos y todos los otros objetos de la misma clase.

b: Es la distancia media entre el objeto y todos los objetos del clúster más próximo.

El valor S(i) puede ser obtenido combinado los valores a(i) y b(i) como se muestra a continuación:

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, si\ a(i) < b(i) \\ 0, si\ a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1, si\ a(i) > b(i) \end{cases}$$

De la definición anterior, se obtiene que el coeficiente de la silueta es: -1 ≤ S(i) ≤ 1. Para que el valor de s(i) sea próximo a uno entonces a(i) < b(i). Un valor cercano a cero indica que el objeto i está en la frontera de los otros clústeres, por el contrario, si el valor de s(i) es negativo, entonces dicho objeto debería ser asignado al clúster más cercano.

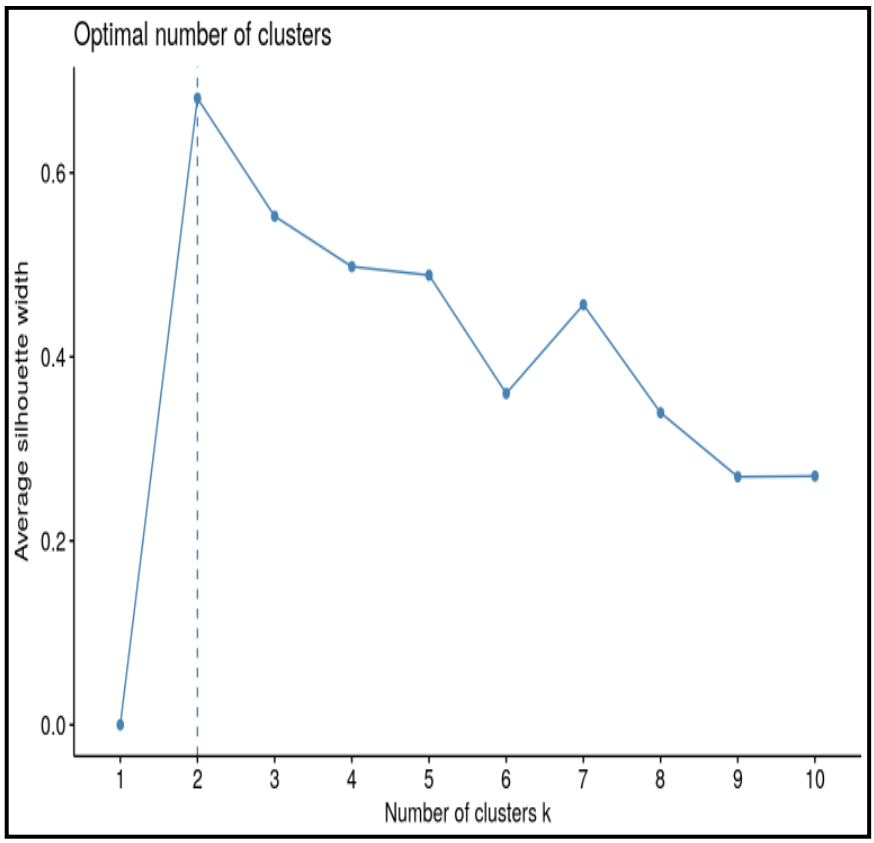


Figura 11. Método de la silueta en R

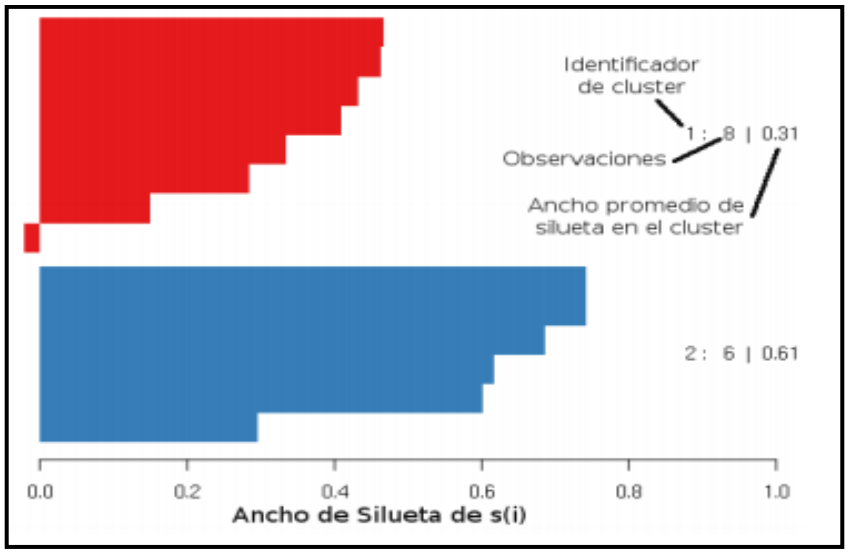


Figura 12. Interpretación visual del cálculo de silueta.

c) **Estadística de la brecha (Gap Statistics)**: Este enfoque se puede aplicar a cualquier método de agrupación, similar al método del codo, su finalidad es encontrar la mayor distancia o diferencia entre los diferentes grupos de objetos. Consiste en generar una medida de error frente a diferentes propuestas de agrupación y localizar donde el error se vuelve asintótico y estima su ubicación como el número de grupos con el valor de brecha más grande. Por lo tanto, el número óptimo de las agrupaciones se produce con el valor de brecha más grande dentro de un rango de tolerancia (Estupiñan et. al, 2016). Es definido como:

$$Gap_n(k) = E_n\{\log(W_k)\} - \log(W_k) \qquad [1]$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \qquad [2]$$

Donde n es el tamaño de la muestra, k es el número de clústeres, n_r es el número de puntos en el clúster r, n_r es la suma de las distancias de pares de todos los puntos del clúster r y En(log(W_k)) denota el valor esperado de log(W_k).

Generar conjunto de datos de referencia B con una distribución aleatoria uniforme. Agrupar cada uno de estos conjuntos de datos de referencia con número variable de grupos k=1 hasta k_n y calculamos el total de la variación dentro del grupo W_{kb}

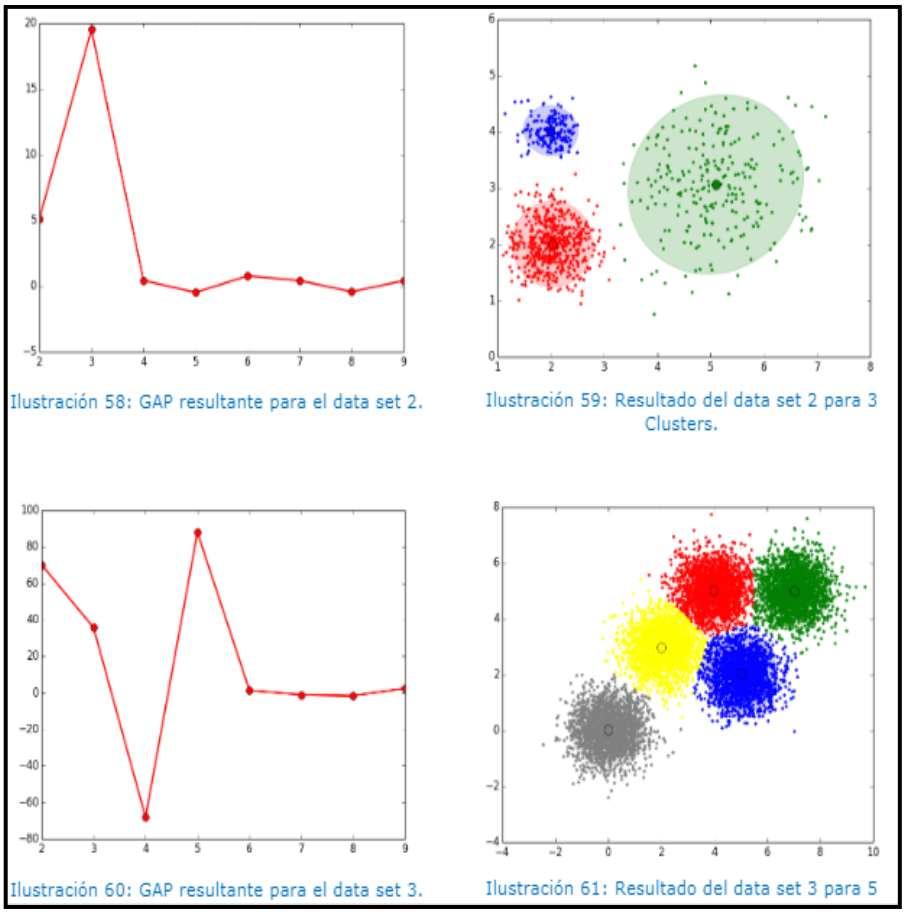


Figura 13. Agrupaciones utilizando la estadística de la brecha en Python. (Moya, 2016)

1.6.3.4.3. Índices de validación de clúster

Como resultado de la aplicación de los algoritmos de segmentación, se obtienen grupos diferentes, algunos de los cuales están mejor conformados que otros, para ello se realizan pruebas para verificar los resultados, se les conoce como índice de validación interna (Rodríguez y García, 2016). Algunos de los más conocidos son:

a) **Índice de Ball y Hall (1965):** La dispersión media de un grupo es la media de las distancias al cuadrado de los puntos del grupo con respecto a su baricentro. El índice Ball – Hall es la media, a través de todos los grupos, de su dispersión media (Desgraupes, 2017):

$$C = \frac{1}{K}\sum_{K=1}^{K}\frac{1}{n_k}\sum_{i\in I_k}\left\|M_i^{\{k\}} - G^{\{k\}}\right\|^2$$

b) **Índice de Calinski y Harabaz (1974):** Conocido también como el índice "pseudo F" muestra la razón entre varianza inter-grupos con la varianza intra-grupos. Siendo n el número de casos y K el número de clústeres en cualquier fase del proceso jerárquico de conglomeración. GSS: la suma de cuadrados inter-grupos y WSS: la suma de cuadrados intra-grupos, entonces: (Alaminos et al., 2015).

$$\boldsymbol{Pseudo\ F} = \frac{\boldsymbol{GSS}}{(\boldsymbol{K-1})} \Big/ \frac{\boldsymbol{WSS}}{(\boldsymbol{n-K})}$$

c) **Índice de Dunn (1974):** Se define como el valor mínimo de la razón entre la medida de disimilaridad de los clústeres y el diámetro del clúster. (Alaminos et al., 2015)

$$\boldsymbol{DI_m} = \frac{\min\limits_{1 \le i < j \le m} \boldsymbol{\partial(C_i, C_j)}}{\max\limits_{1 \le k \le m} \boldsymbol{\Delta_k}}$$

d) **Índice de Davies Bouldin (1979):** Cuantifica la similaridad media entre un clúster y los que están próximos a él. De acuerdo a la formación del índice cuanto menor es su valor mejor es la solución. (Alaminos et al., 2015)

$$\boldsymbol{C} = \frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{M_k} = \frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \max_{\boldsymbol{k' \neq k}} \left(\frac{\boldsymbol{\delta_k + \delta_{k'}}}{\boldsymbol{\Delta_{kk'}}} \right)$$

1.6.3.5. KDD

Giraldo y Mejía (2019) en su artículo científico define al proceso de KDD como la utilización de métodos de minería de datos (algoritmos) para extraer conocimiento con la especificación de ciertos parámetros de una base de datos. Además, indica que, en el gran volumen de datos almacenados en la actualidad, se encuentra información "oculta" que es posible descubrir gracias a la minería de datos, pero es el Descubrimiento de Conocimiento (KDD, por sus siglas en inglés), el que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos.

Adicionalmente, Ñaupas (2016) afirma que KDD es una metodología para descubrir información de una gran base de datos y generar conocimiento, el objetivo principal de esta metodología es automatizar el procesamiento de los datos, para que los usuarios puedan dedicar más tiempo al análisis y el descubrimiento de patrones.

En la siguiente figura se muestran las etapas del KDD de Fayyad (1996):

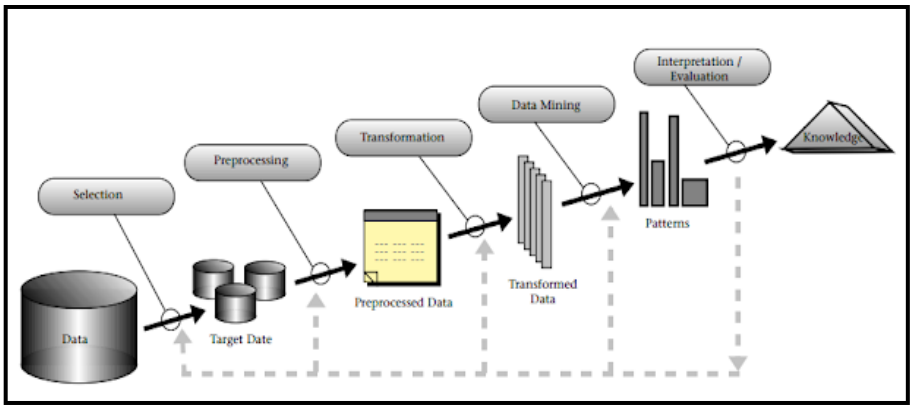


Figura 14. Etapas de KDD de Fayyad (1996)

- **Fase 1. Ubicación y elección de los datos:**

  En esta fase se busca responder a las siguientes preguntas: ¿En qué tipo de almacenamiento se encuentran los datos?, ¿Cuál es su estructura?, ¿Cuál es su significado?, y verificar si los datos están relacionados con el objetivo del proyecto.

- **Fase 2. Limpieza y preprocesado:**

  Se aplican estrategias para manejar el ruido en los datos, valores incompletos, secuencias de tiempo. También en esta fase, se pueden utilizar muestras al azar para reducir el volumen de datos, y también concretar variables y registros apropiados como datos de entrada para la minería de datos.

- **Fase 3. Transformación:**

  En esta etapa, se transforman los datos al tipo de formato requerido por la técnica de minería de datos que se va a utilizar, por ejemplo,

Reglas de Asociación, Clasificación, Regresión Logística, Clustering, entre otras.

- **Fase 4. Minería:**

  Se emplea la técnica de minería para encontrar patrones que puedan expresarse como un modelo o simplemente que expresen dependencia entre los mismos datos.

- **Fase 5. Interpretación y validación:**

  Interpretar los datos puede implicar repetir otra vez el proceso, quizás con otros datos, otros datos, otras estrategias y/u otras metas. Este es un paso importante, en donde se requiere dominio del tema. La interpretación puede beneficiarse de procesos de visualización como la regresión, y sirve para borrar patrones redundantes o irrelevantes obtenidos en técnicas como el Clustering o las Reglas de Asociación.

  El conocimiento que se obtiene permite realizar acciones dentro de un sistema de desempeño o simplemente para almacenarlos, y luego disponer de él, por parte de los usuarios involucrados en el dominio específico.

1.6.3.6. Software Weka

Conocido por sus siglas en inglés (Waikato Environment for knowledge Analysis) es una herramienta de sofware libre para el aprendizaje automático y minería de datos diseñado a base de JAVA y desarrollado en la universidad de Waikato en Nueva Zelanda en el año 1993. Weka contiene una colección de algoritmos para realizar análisis y modelado predictivo, también tiene herramientas para la visualización de datos, adicionalmente una interfaz gráfica que facilita la mejor disposición del contenido. Soporta muchas tareas estándar de la minería de datos en especial de procesamiento de datos, regresión, clasificación, clustering, etc. Todas las técnicas en WEKA están basadas en la sunción de datos que estén disponibles en un fichero plano o una relación, en donde cada

registro de datos está descrito por un número fijo de atributos nominales o numéricos, permite el acceso a otras instancias de la base de datos por medio de SQL, gracias al JDBC, además puede procesar un resultado generado de una consulta hecha a una base de datos. (Nuñez et al., 2019)

1.6.3.7. Lenguaje R

Según Murillo y Saavedra (2017), mencionan que el R es un lenguaje de programación de código abierto, el cual tiene similitud con otros lenguajes como C o C++, pero su utilidad principal es mezclar diferentes características de otros lenguajes y paradigmas de programación. Adicionalmente, está orientado hacia la minería y el análisis de datos por lo estar compuesto de librerías o paquetes que realizan esas funciones.

1.7 Definición de términos básicos

- **Algoritmo A priori:** El algoritmo a priori es un algoritmo utilizado para encontrar reglas de asociación en los registros, de manera que se puedan formar relaciones y obtener atributos ocultos en los mismos. Este algoritmo se ha aplicado grandemente en el análisis de transacciones comerciales y en problemas de predicción (Amador et al., 2015).

- **Análisis RFM:** Es una herramienta que ofrece una vista que permite identificar a los mejores clientes de una empresa mediante la medición de ciertos factores. El modelo RFM se basa en tres factores cuantitativos: Recientemente (Qué tan recientemente un cliente ha hecho una compra), Frecuencia (Con qué frecuencia un cliente hace una compra) Valor monetario (Cuánto dinero gasta un cliente en compras) (Cuadros et al., 2017).

- **Base de datos:** Es un conjunto de datos dispuestos con la finalidad de proporcionar información y permitir transacciones como inserción, eliminación y actualización de los datos (Benitez y Arias, 2015).

- **Datamart:** Son almacenes de datos que contienen datos específicos de un área o departamento de una organización. Se caracteriza por disponer la estructura

óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento (Ramos et al., 2017).

- **Datawarehouse:** Es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta (Ramos, 2016).

- **Dendrograma:** La forma más simple en la cual una estructura jerárquica de datos puede ser representada, y se basa en un gráfico con forma de árbol (dendro=árbol) (Castillo et al., 2017).

- **ERP**: Es un sistema de planeación de recursos empresariales (ERP por su acrónimo en inglés) es una herramienta de información integrada que enlaza funciones intra e inter empresariales con la finalidad de proveer flujos ininterrumpidos de información a través de la cadena de suministro (Álvarez y Sánchez, 2016).

- **Inteligencia de negocios:** Disciplina encargada de alimentar nuestro sistema de inteligencia de mercado con información del negocio (Bernal, 2017).

- **Inteligencia mercado:** Disciplina que brinda información sobre el consumidor (Bernal, 2017).

- **Java:** Lenguaje de programación orientado a objetos. Consta de una o más clases interdependientes, las clases permiten describir las clases habilidades y propiedades de los objetos (Prieto et al., 2016).

- **Marketing: S**istema total de actividades de negocios ideado para planear productos satisfactores de necesidades, asignarles precios, promover y distribuirlos a los mercados meta, a fin de lograr los objetivos de la organización (Etzel et al., 2004).

- **Mix de Productos:** Conjunto de herramientas de empleadas en el marketing, que la empresa combina con la finalidad de producir una respuesta en el

mercado objetivo y cubrir en cierto aspecto las exigencias del consumidor (Muñoz, 2015).

- **Scoring:** La calificación crediticia o scoring, utilizan la información del consumidor en función de una puntuación generada por el sistema de puntuación. (Albrecht y Savio, 2015).

- **Sistema gestor de base de datos:** Programa informático donde el objetivo principal es evitar la manipulación directa por el usuario y establecer un marco estándar para organizar los datos (Benitez y Arias, 2015).

- **SQL:** Lenguaje utilizado para base de datos desarrollado a principios de los años 70. La mayoría de sistemas gestores de base de datos relacional, tratan de seguir el estándar SQL para formalizar sus consultas y otras operaciones (Benitez y Arias, 2015).

1.8 Hipótesis

Hi: El modelo basado en técnicas de minería de datos segmentará a los clientes en la empresa distribuidora Suministros del Oriente SA.

Ho: El modelo basado en técnicas de minería de datos no segmentará a los clientes de la empresa distribuidora Suministros del oriente SA.

1.9 Sistema de variables

Variable independiente: Modelo basado en técnicas de minería de datos.

Variable dependiente: Segmentación de clientes.

1.10 Operacionalización de variables

VARIABLES	DEFINICIÓN CONCEPTUAL	DIMENSIONES	INDICADORES	ESCALA DE MEDICIÓN
Modelo basado en técnicas de minería de datos.	"Proceso que utiliza técnicas estadísticas, matemáticas, inteligencia artificial y de aprendizaje de máquinas para extraer e identificar información útil que convierte en conocimiento a partir de grandes bases de datos, data warehouses o data mart". (Joyanes, 2016)	Clustering o Agrupamiento	Número de clústeres de clientes	Razón
Segmentación de clientes	Clientes diferentes, tienen diferentes necesidades, el producto no precisa el mismo beneficio para todos, y de manera individual, cada cliente lo compra por diferente motivo, por ello la segmentación permite considerar los mercados en los que la empresa tiene y debe tener presencia. (Westwood, 2016)	Clientes	Número de Clientes	Nominal
		Productos	Categoría de Productos	Nominal
		Características Comerciales	Número de características comerciales	Razón

Fuente: Elaboración propia

CAPÍTULO II
MATERIAL Y MÉTODOS

2.1. Materiales

2.1.1. Modelo para segmentación de clientes

2.2.1.1. Diseño del modelo

En la revisión literaria, no se ha encontrado un modelo para segmentar clientes con características RFM, por tal se presenta el siguiente modelo de segmentación. En la figura 15, se muestra el diseño general de la propuesta.

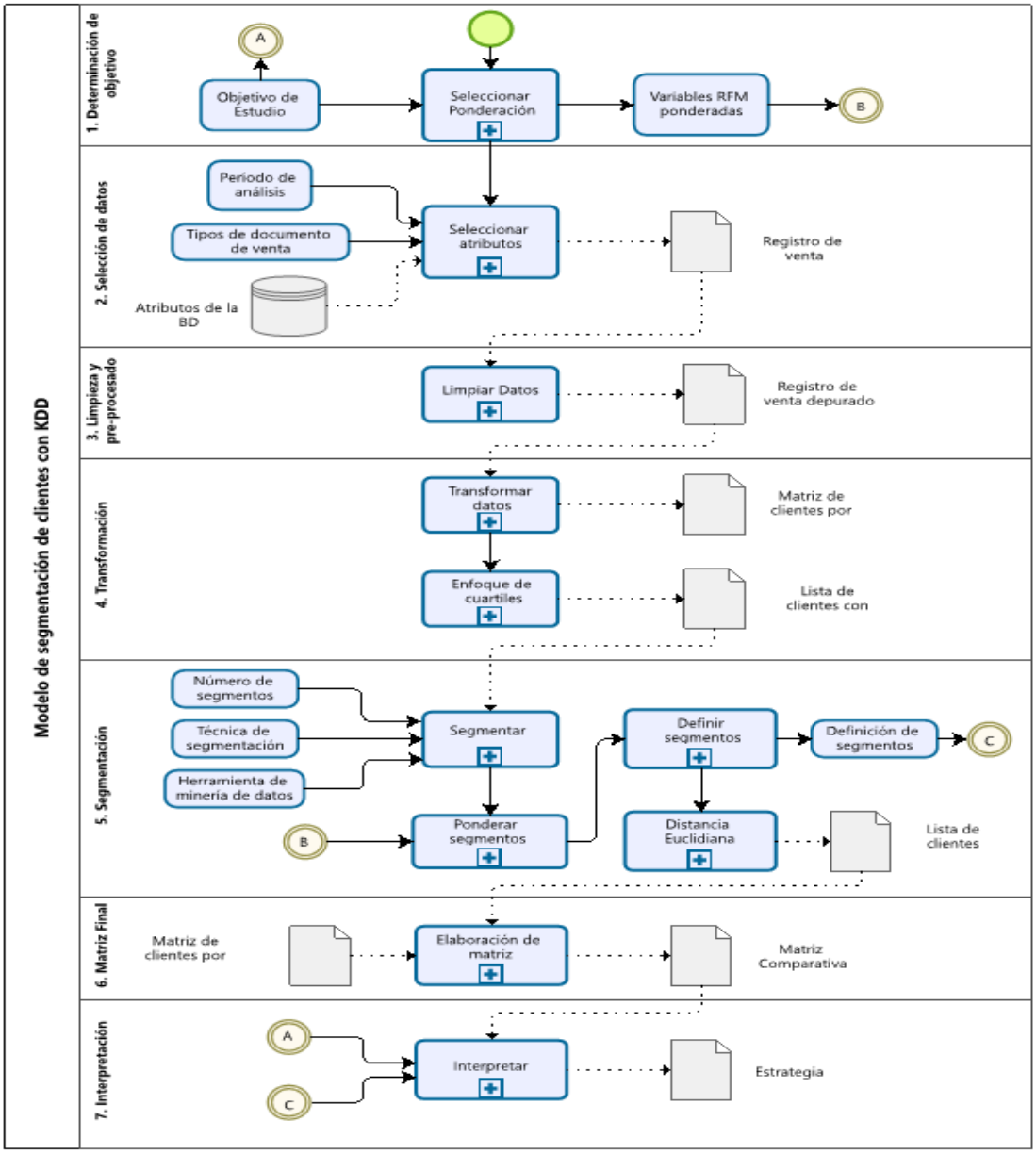


Figura 15. Diseño de modelo propuesto (Elaboración propia)

2.2.1.2. Modelo

A continuación, se describen los pasos del modelo propuesto, la cual contiene tablas con entrada, procedimiento y salida.

Paso 1. Determinación del objetivo

En este paso se define la característica RFM predominante, y asigna peso según prioridad de estudio. La Tabla 3, muestra las entradas, herramientas y técnicas, así como la salida correspondiente.

Tabla 3
Paso 1: Determinación de Objetivo

1. Determinación de Objetivo		
Entrada	Procedimientos	Salida
- Objetivo de Estudio	- Seleccionar ponderación	- Variables RFM ponderadas

Fuente: Elaboración propia

Entrada:

Objetivo de Estudio**:**

Seleccionar uno de los objetivos siguientes:

- Objetivo 1: Identificar los grupos de productos con mayor **recencia** de ventas.
- Objetivo 2: Identificar los grupos de productos con mayor **frecuencia** de ventas.
- Objetivo 3: Identificar los grupos de productos con mayor **monto** de ventas.

El grupo de productos puede ser la agrupación de marcas, categorías, tipos, etc.

Procedimiento:

Seleccionar ponderación:

En base al objetivo seleccionado se ponderan las variables RFM. Para ello se signar el peso correspondiente a cada uno de ellos siendo el más alto el de mayor relevancia según el objetivo. Tal como se muestra en la Tabla 4:

Tabla 4

Lista de valores RFM según el objetivo

Objetivo	Peso(R)	Peso(F)	Peso(M)
Objetivo 1	2	1	1
Objetivo 2	1	2	1
Objetivo 3	1	1	2

Fuente: Elaboración propia

Salida:

Variable RFM ponderada:

Si por ejemplo se selecciona el objetivo 2 de la tabla 4, las variables RFM tienen la ponderación que se indica en su respectiva fila**.**

Paso 2. Selección de datos

En este paso, se debe seleccionar los atributos de la base de datos para obtener las características RFM de los clientes. La Tabla 5 muestra las entradas, herramientas y técnicas, así como la salida.

Tabla 5

Paso 2: Selección de datos

2. Selección de datos		
Entrada	Procedimiento	Salida
- Período de análisis - Tipos de documentos de venta - Atributos de la base de datos	- Seleccionar atributos	- Registro de venta

Fuente: Elaboración propia

Entradas:

Período de Análisis:

Es el período comercial en el cual se va a realizar el análisis para la segmentación de clientes, basado en un rango de fechas.

Tipos de documentos de venta:

Son los tipos de documentos comerciales como boleta, factura o algún otro comprobante de pago.

Atributos de la base de Datos:

Son los campos de la base de datos que servirán para el análisis RFM de los clientes y los grupos de productos. (Ver Tabla 6)

Tabla 6

Ejemplo de lista de campos seleccionados

ID	Atributo	Tipo
1	Código de cliente	Entero
2	Fecha de venta	Fecha
3	Monto de venta	Decimal
4	Número de documento	Texto
5	Tipo de documento	Entero
6	Grupo de Producto	Texto

Fuente: Elaboración propia

Procedimiento

Selección de atributos:

En base a los atributos de entrada se transforman a las variables RFM mediante una consulta de selección a la base de datos.

Salida

Registro de venta

Considerar los atributos de la base de datos necesarios para el RFM y también para los grupos de productos. (Ver Tabla 7)

Tabla 7

Ejemplo de registro de venta

Fecha de Venta	Tipo de Documento	Número de Documento	Grupo de Productos	Id Cliente	Monto de Venta
01/01/2019	BOL	001	Grupo 1	C001	1,000.00
01/01/2019	FAC	002	Grupo 2	C002	2,000.00
01/01/2019	BOL	003	Grupo 3	C003	4,000.00

Fuente: Elaboración propia

Paso 3. Limpieza y pre-procesado

Para la limpieza y el pre-procesado se analizan los registros de ventas obtenidos en el paso 2. (Ver Tabla 8)

Tabla 8

Paso3: Limpieza y pre-procesado

3. Limpieza y pre-procesado		
Entrada	Procedimiento	Salida
- Registro de venta	- Limpiar datos	- Registro de venta depurado

Fuente: Elaboración propia

Entrada:

Registro de venta

Proviene del paso 2

Procedimiento

Limpiar datos:

Con todos los registros ventas obtenidos, se deben eliminar del análisis los que algunos parámetros que no se consideren necesarios, tal como se muestra en la Tabla 9:

Tabla 9

Ejemplo de limpieza y preprocesado de datos

Problema	Solución
Clientes vinculados a la misma empresa (Retail)	Eliminar registros
Artículos promocionales con facturación a costo 0.	Eliminar registros
Valores nulos o vacíos generados por el sistema.	Eliminar registros

Fuente: Elaboración propia

Salida

Registro de venta depurado.

Se obtiene un nuevo de registro de ventas con los filtros aplicados en el procesamiento.

Paso 4. Transformación

En este paso, obtenemos el análisis RFM de los clientes con los datos provenientes del paso 3, para ello se utiliza el enfoque de cuartiles con la finalidad de minimizar las diferencias (ver Tabla 10).

Tabla 10

Paso 4: Transformación

4. Transformación		
Entrada	Procedimiento	Salida
- Registro de venta depurado.	- Transformar datos - Enfoque de cuartiles	- Lista de clientes con puntuación RFM. - Matriz de clientes por grupo de productos

Fuente: Elaboración propia

Entrada:

Registro de venta depurado.

Proviene del paso 3.

Procedimiento

Transformar de datos

Transformar los datos provenientes del paso 3 para que puedan ser utilizados en el análisis RFM, es conveniente utilizar el enfoque de cuartiles como una forma de discretización de datos.

Adicionalmente, utilizaremos el registro de ventas depurado obtenido en el paso 3, para crea una matriz de clientes por segmento de producto para determinar si un cliente compra o no un grupo determinado de productos, mediante una asignación booleana Verdadero (V) o Falso (F).

Salida

Lista de Clientes con puntación RFM

Aplicando el proceso obtenemos a cada cliente con su respectiva puntuación (1-4) por variable (ver Tabla 11).

Tabla 11

Ejemplo de lista de clientes con puntuación RFM

ID	R	F	M
C001	4	4	4
C002	3	3	4
C003	2	2	3
C004	1	2	1

Fuente: Elaboración propia

Matriz de clientes por grupo productos

Se elabora una matriz en donde listamos los grupos de productos por cliente (ver Tabla 12).

Tabla 12

Ejemplo de matriz de clientes por grupo de productos

ID	Grupo 1	Grupo 2	Grupo 3
C001	V	F	F
C002	F	V	V
C003	V	V	F
C004	V	V	F

Fuente: Elaboración propia

Paso 5. Segmentación

En este paso es donde se genera la obtención de conocimiento aplicando una técnica de minería de datos en base a los datos obtenidos en el paso 4, con la finalidad de obtener segmentos de clientes. En la Tabla 13, se muestran las entradas, procedimientos y salidas de este paso.

Tabla 13

Paso 5: Segmentación

5. Segmentación		
Entrada	Procedimientos	Salida
- Lista de clientes con puntuación RFM - Variables RFM ponderadas - Número de segmentos -Técnica de segmentación - Herramienta para minería de datos	- Segmentar - Ponderación de segmentos -Definir segmentos (A-Z) - Distancia Euclidiana	- Definición de segmentos. - Lista de clientes segmentados

Fuente: Elaboración propia

Entrada:

Lista de clientes con puntuación RFM

Proviene del paso 4

Variables RFM ponderadas

Proviene del paso 1

Número óptimo de clústeres

Se puede utilizar algún método para indicar el número óptimo de clústeres como el método del codo, método de la silueta, entre otros, la organización puede elegir el número de grupos de acuerdo a su necesidad.

Herramienta para minería de datos

Software libre o de pago que permita la aplicación de la técnica de clustering. Weka, R, Sciki-learn, etc.

Técnica de segmentación

Es alguno de los algoritmos de segmentación para agrupar a los clientes, K-means, K-medoides, KNN, etc.

Procedimiento:

Segmentar

Tenemos las siguientes consideraciones:

- Seleccionar una técnica de clustering (definido como entrada)
- Indicar el número óptimo de clústeres (definido como entrada)
- Aplicar la técnica de clustering. (con la herramienta de minería de datos)
- Obtener los medoides o centroides de cada clúster.

Ponderación de segmentos

Luego de obtener los segmentos, utilizamos los medoides o centroides para categorizar a los clientes; para ello, usamos los pesos asignados en el paso 1 y obtenemos el valor ponderado por cada segmento, basado en la siguiente fórmula:

$$P = \frac{R \times Peso(R) + F \times Peso(F) + M \times Peso(M)}{4}$$

Donde:

P = Valor de Ponderación del segmento

R= Centroide o Medoide del segmento en Recencia.

F= Centroide o Medoide del segmento en Frecuencia.

M=Centroide o Medoide del segmento en Monto.

Posteriormente a cada valor ponderado asignarle un grupo A-Z en orden ascendente.

En la Tabla 14, se muestra un ejemplo de segmentos categorizados de acuerdo a su ponderación:

Tabla 14

Ejemplo de segmentos categorizados.

Clúster	R(Peso1)	F(Peso1)	**M(Peso2)**	Ponderado	Segmento
Clúster01	3.5	3.6	3.5	3.525	A
Clúster02	2.4	2.3	2.3	2.325	B
Clúster03	1.3	1.4	1.4	1.375	C

Fuente: Elaboración propia

Definir los segmentos (A-Z)

Luego de ponderar los segmentos se debe definir a cada letra en relación al objetivo.

- **Por ejemplo**, si el objetivo seleccionado es el número 3, indica que el segmento A se trata de clientes que tienen un mayor monto de compras, y que el segmento C se trata de clientes con menor monto de ventas.

Distancia euclidiana

Para categorizar a cada cliente de la organización se debe emplear la fórmula de distancia euclidiana, para aproximar a los clientes al segmento (A-Z) más cercano:

$$D_{(i,j)} = \sqrt{(X1_i - X1_j)^2 + (X2_i - X2_j)^2 + \ldots + (XZ_i - XZ_j)^2}$$

Por ejemplo, en la tabla 15 tenemos los centroides de los segmentos A (R =3.5, F=3.6, M=3.5), B (R=2.4, F=2.3, M=2.3) y C (R=1.3, F=1.4, M=1.4)

Tabla 15

Ejemplo de aproximación con distancia euclidiana

ID	R	F	M	Distancia a los centroides			Más cercano	Grupo
				A	B	C		
C001	4	4	4	**0.81**	2.89	4.56	**0.81**	A
C002	3	3	4	**0.93**	1.93	3.49	**0.93**	A
C003	2	2	3	2.25	**0.86**	1.85	**0.86**	B
C004	1	2	1	3.88	1.93	**0.78**	**0.78**	C

Fuente: Elaboración propia

Salidas

Lista de clientes segmentados

Es la lista de todos los clientes con su respectivo grupo o segmento. (Ver Tabla 16)

Tabla 16

Ejemplo de lista de clientes segmentados

ID	Grupo
C001	A
C002	A
C003	B
C004	C
C005	B
C006	C

Fuente: Elaboración propia

Definición de segmentos

Es la lista de los segmentos con la definición propia de la organización. (Ver Tabla 17)

Tabla 17

Ejemplos de definición de segmentos

Segmento	Definición
A	Cliente Preferencial. Son los clientes con mayor monto de compra.
B	Cliente Intermedio. Son los clientes con un regular monto de compra.
C	Cliente Potencial. Son los clientes con un monto de compra menor.

Fuente: Elaboración propia

Paso 6. Matriz Comparativa

En este paso, utilizamos los segmentos de clientes y los grupos de productos para elaborar una matriz comparativa. (Ver Tabla 18).

Tabla 18

Paso 6: Matriz Comparativa

6. Matriz Comparativa		
Entrada	Procedimiento	Salida
- Lista de clientes segmentados - Matriz de clientes por grupo de productos	- Elaboración de matriz	- Matriz comparativa

Fuente: Elaboración propia

Entradas:

Lista de clientes segmentados

Proviene del paso 5.

Matriz de clientes por grupo de productos

Proviene del paso 4.

Procesamiento:

Elaboración de una matriz.

Elaborar una matriz, agrupada por segmento de cliente, que contenga el número total de clientes que compró un grupo de producto determinado, tal como se muestra en la Tabla 19. (Suma de verdaderos por segmento de clientes de la Tabla 12).

Tabla 19

Ejemplo de matriz grupo productos por segmento de cliente

Grupo de Producto	A	B	C
Grupo 1	120	10	50
Grupo 2	70	100	10
Grupo 3	30	40	150

Fuente: Elaboración propia

Aplicar el enfoque de cuartiles a la matriz de productos por segmento de cliente para puntuar los grupos de productos en rangos de 1-4. (Ver Tabla 20)

Tabla 20

Ejemplo de matriz grupo de productos por segmento en cuartiles

Grupo de Producto	A	B	C
Grupo 1	4	1	2
Grupo 2	3	4	1
Grupo 3	1	2	4

Fuente: Elaboración propia

Finalmente, agrupamos los grupos de productos por cuartil y segmento de cliente. (Ver Tabla 21)

Tabla 21

Ejemplo de matriz comparativa

(Puntuación)	A	B	C
4	Grupo 1	Grupo 2	Grupo 3
3	Grupo 2	Grupo 4	Grupo 4
2	Grupo 4	Grupo 3	Grupo 1
1	Grupo 3	Grupo 1	Grupo 2

Fuente: Elaboración propia

Salida

Matriz Comparativa

Se obtiene la matriz comparativa del grupo de producto y segmento de cliente.

Paso 7. Interpretación

En este paso se definen las estrategias según el objetivo. (Ver Tabla 22)

Tabla 22

Paso 7: Interpretación.

7. Interpretación		
Entrada	Procedimiento	Salida
- Matriz comparativa -Definición de segmentos. -Objetivo de Estudio	- Interpretar	- Estrategia

Fuente: Elaboración propia

Entrada

Matriz comparativa.

Proviene del paso 6

Definición de segmentos.

Proviene del paso 5

Objetivo de Estudio

Proviene del paso 1

Procedimiento

Interpretar

Realizamos el análisis de la tabla comparativa y definición de segmento con el objetivo seleccionado.

Por ejemplo:

El objetivo 3: Identificar los grupos de productos con mayor **monto** de ventas.

Interpretación 1: Los clientes que pertenecen al segmento A, tienen un mayor monto de compra en el grupo de productos del cuartil 4

Interpretación 2: Los clientes que pertenecen al segmento C, tienen un menor monto de compra en el grupo de productos del cuartil 1.

Salida

Estrategias:

Según la interpretación 1, la organización puede implementar estrategias de fidelidad al segmento de clientes preferenciales.

Según la interpretación 2, la organización puede implementar estrategias para incrementar la venta de grupos de productos identificados para los clientes potenciales.

2.2. Métodos

2.2.1. Tipo y nivel de investigación

La presente investigación fue de tipo aplicada porque está orientada a mejorar el funcionamiento del proceso de segmentación con la aplicación de los avances de la ciencia y tecnología. Y es de nivel descriptivo, debido a que se obtuvieron los datos de los clientes en un periodo específico (ene 2019 – nov 2019), describiendo las agrupaciones que hay con sus respectivas características comerciales.

2.2.2. Diseño de investigación

La presente investigación se realizó mediante el diseño descriptivo, conocido también como "no experimental" de estudio transversal, ya que se recolectarán datos en un periodo de tiempo específico.

Es descriptiva porque según Bernal (2010), menciona que la investigación descriptiva sirve para seleccionar características principales de objeto de estudio y la descripción detallada de las categorías, clases o partes del mismo.

2.2.3. Población y muestra

2.2.3.1. Población y muestra

"La población se define como la totalidad del fenómeno a estudiar donde las unidades de población poseen una característica común la cual se estudia y da origen a los datos de la investigación". (Tamayo y Tamayo, 1997, p. 114)

Y la muestra se conceptualiza como: "El grupo de individuos que se toma de la población, para estudiar un fenómeno estadístico". (Tamayo y Tamayo, 1997, p. 38)

Siendo que el diseño de la investigación es de tipo descriptivo propositivo y que el objeto de estudio de la presente es "segmentación de clientes", el mismo no cuenta con una población y muestra.

CAPÍTULO III
RESULTADOS Y DISCUSIÓN

La presente investigación buscó desarrollar un modelo de segmentación clientes basado en técnicas de minería de datos, para ello se utilizó la metodología KDD, además se aplicó el modelo diseñado en la empresa Suministros del Oriente SA:

3.1 Desarrollo de caso práctico

Para aplicar el modelo se tomó como de estudio a la organización Suministros del Oriente, empresa dedicada a la distribución de productos para el mercado de vehículos en todas sus categorías. La empresa realiza aproximadamente 4,000 transacciones al mes en sus 5 sucursales en Tarapoto, Pucallpa, Iquitos, Jaen y Tocache.

La organización tiene la necesidad de clasificar a sus clientes, ya que en la actualidad no cuenta con una segmentación definida y por tal motivo, la aplicación de estrategias comerciales con los grupos de productos que oferta (categoría de producto) sean iguales para todos sus clientes, tanto en precio como en cantidad. El modelo permitirá identificar el segmento de sus "mejores clientes" con los "grupos de productos más vendidos", que ayudará en el proceso de toma de decisiones en el área comercial.

A continuación, se detallan los pasos aplicados por el método propuesto:

PASO 1: DETERMINACIÓN DE OBJETIVO

Establecemos el objetivo de la segmentación basada en la característica del análisis RFM que se desee estudiar y seleccionar la variable predeterminante.

Objetivo: Segmentar a los clientes por la **frecuencia de compras** para las categorías de productos. (Objetivo 2)

Por lo cual se realiza la ponderación como se muestra en la Tabla 23.

Tabla 23

Variables RFM ponderadas

Objetivo	Peso(R)	Peso(F)	Peso(M)
Objetivo 2	1	2	1

Fuente: Elaboración propia

PASO 2: SELECCIÓN DE DATOS

- Definimos el período comercial a evaluar.

Período de Análisis: ENE2019 – NOV2019

- Especificamos los tipos de documentos de venta que se consideren necesarios en el análisis.

Tipos de documentos de venta: Boleta y Factura

- Seleccionamos los campos de la base de datos que servirán para realizar el análisis RFM y los grupos de productos. (Ver Tabla 24)

Tabla 24

Lista de campos de seleccionados

Atributo	Descripción	Tipo
Idcliente	Código único del cliente	Entero
FechaVenta	Fecha de venta del documento	Fecha
CategoríaProducto	Nombre de la categoría al que pertenece el producto	Texto
MontoVenta	Contiene el importe de por Categoría	Decimal
NroDocumento	Contiene el número de documento asignado a esa transacción	Texto
TipoDocumento	Contiene el código único de tipo de documento asignado a la transacción	Entero

Fuente: Elaboración propia

PASO 3: LIMPIEZA Y PRE-PROCESADO

Eliminamos los datos atípicos para que el contenido sea correctamente procesado por el software informático.

Tabla 25

Limpieza en atributos seleccionados

Problema	**Solución**
El atributo Categ*oríaProducto* contiene grupos que no estaban vinculadas a productos de implicancia comercial y que están direccionados a campañas promocionales.	Se eliminaron registros
El atributo *GrupoProducto* contiene categorías vacías	Se eliminaron registros
El atributo *CodDocumento* contiene tipos de documentos que no corresponden a los del análisis (boletas y facturas)	Se eliminaron registros
El atributo *IdCliente* contiene clientes que forman parte del grupo comerciales y que no son parte del análisis respectivo.	Se eliminaron registros
El atributo *FechaVenta* se obtiene con el formato: "dd-MMM"	Cambiamos formato: "dd/mm/aaaa"

Fuente: Elaboración propia

PASO 4: TRANSFORMACIÓN

-Transformamos los atributos del registro de ventas depurado al modelo RFM (Recencia, Frecuencia y Monto).

a) Recencia: Intervalo entre la última fecha de transacción y el final del período de análisis. (días)

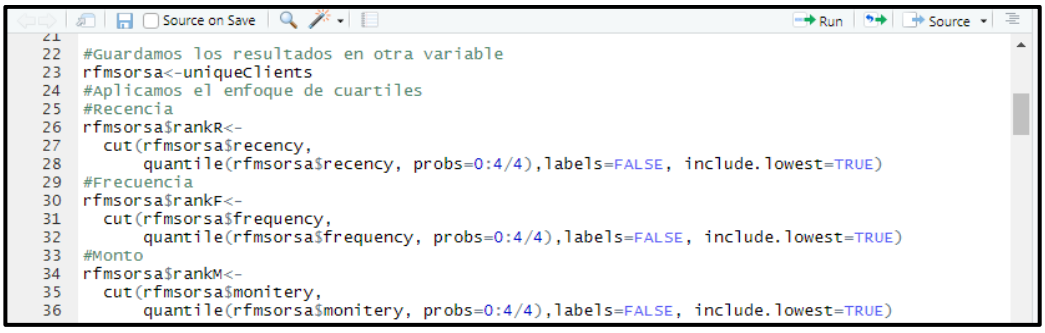
b) Frecuencia: Cantidad de transacciones por documento del período de análisis. (entero)

c) Monto: Importe total de la suma del período de análisis. (decimal)

```
# Cargamos los datos en el data.frame itemsSold para procesarlos
itemsSold <- data.frame( read.table("datasorsarfm.csv", header = TRUE, sep=";", dec=".") )
# Nos aseguramos de que la columna orderdate es de tipo Date
itemsSold$orderdate <- as.Date( itemsSold$orderdate, format="%d/%m/%Y")
# Obtenemos los usuarios únicos de la empresa
uniqueClients <- with( itemsSold, data.frame( iduser = sort(unique(iduser)) ) )
# Añadimos la columna Recencia con los días transcurridos desde la última compra
uniqueClients <- cbind(uniqueClients, recency =aggregate(round(
  as.numeric(difftime(Sys.Date(),itemsSold$orderdate, units="days")) ),
  list(itemsSold$iduser), min )$x)
# Añadimos la columna Frecuencia con el número de compras realizadas por factura
uniqueClients <- cbind(uniqueClients,frequency = with( itemsSold, as.numeric(
  by(doc, iduser, function(x) length(unique(x)) ) ) ))
# Añadimos la columna Monto con el valor total de las compras
uniqueClients <- cbind(uniqueClients,monitery = with( itemsSold, as.numeric(
  by(unitprice,iduser,sum) )))
```

Figura 16. Código en R – Carga de data (Elaboración Propia).

-Aplicamos el enfoque de cuartiles para producir las variables en rangos de 1-4 de los valores RFM obtenidos de todos los clientes.



```
#Guardamos los resultados en otra variable
rfmsorsa<-uniqueClients
#Aplicamos el enfoque de cuartiles
#Recencia
rfmsorsa$rankR<-
  cut(rfmsorsa$recency,
      quantile(rfmsorsa$recency, probs=0:4/4),labels=FALSE, include.lowest=TRUE)
#Frecuencia
rfmsorsa$rankF<-
  cut(rfmsorsa$frequency,
      quantile(rfmsorsa$frequency, probs=0:4/4),labels=FALSE, include.lowest=TRUE)
#Monto
rfmsorsa$rankM<-
  cut(rfmsorsa$monitery,
      quantile(rfmsorsa$monitery, probs=0:4/4),labels=FALSE, include.lowest=TRUE)
```

Figura 17. Código en R – Enfoque de cuartiles (Elaboración Propia).

-Utilizamos el mismo registro para crear la matriz de clientes por grupo de producto.

```
#Matriz de clientes -categorías
require("dplyr")
bkitemsSold<-itemsSold

skusorsa<-bkitemsSold %>% group_by(bkitemsSold$iduser,
                                   bkitemsSold$sku,
                                   bkitemsSold$skuorder)%>%
                                   dplyr::summarize(cnt=n()) %>% as.data.frame()

names(skusorsa)<-c("iduser","sku","doc","cnt")
#Contabilizamos las compras por usuario
catsorsa<-xtabs(~iduser+sku,data=skusorsa)
#Convertimos el resultado en una matriz
tab_productos<-as.data.frame.matrix(catsorsa)
#Añadimos una columna con los códigos de cliente
tab_productos$idcliente<-rownames(tab_productos)
```

Figura 18. Código en R - Matriz de clientes-categorías

	ACEITE	ADITIVOS	CALCIO	CERA LIQUIDA	CINTA	DOT3	DOT4	FILTROS	FRENOS
1	F	F	F	F	F	F	F	F	F
2	F	F	F	F	V	F	F	F	F
3	F	F	F	F	F	F	F	F	F
4	F	F	V	F	V	F	F	V	F
5	F	F	F	F	F	F	F	V	F
6	F	F	F	F	F	F	F	F	F
7	F	F	F	F	F	F	F	F	F
8	F	F	V	F	F	F	F	V	F
9	F	F	F	F	V	F	F	F	F
10	F	F	F	F	F	F	F	F	F
11	F	F	V	F	F	F	F	F	F
12	F	F	F	F	F	F	F	F	F
13	F	F	F	F	F	F	F	F	F
14	F	F	F	F	F	F	F	F	F
15	F	F	V	F	F	F	F	F	F
16	F	F	F	F	F	F	F	F	F

Showing 1 to 17 of 2,525 entries, 25 total columns

Figura 19. Vista en R - Matriz de clientes – categoría

PASO 5: SEGMENTACIÓN

- Se utilizaron los métodos del codo y silueta para definir el número óptimo de segmentos, así como la técnica de segmentación (Ver Anexo 3). Luego de realizar las respectivas comparaciones se obtuvo el valor de 5 para el número óptimo de segmentos y la mejor técnica K-means.

Realizamos la segmentación en el software R, considerado las siguientes entradas:

Registro de ventas depurado.

Número óptimo de clústeres: 5.

Técnica de segmentación: K-means.

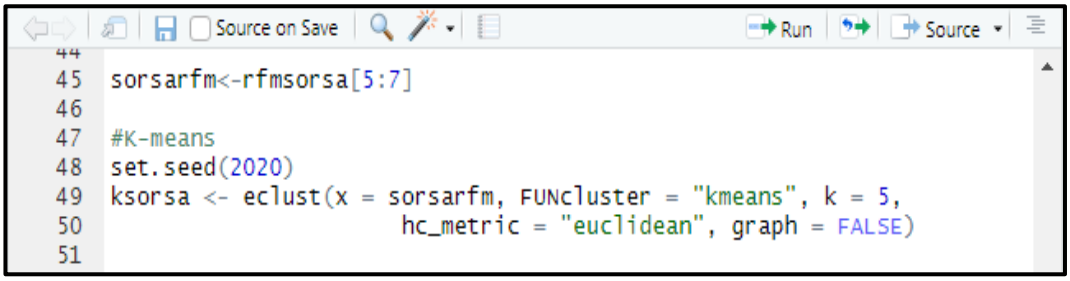


Figura 20. Código en R – Kmeans con 5 segmentos (Elaboración Propia)

- Luego de obtener los clústeres, utilizar los centroides para segmentar a los clientes; para ello, usamos los pesos asignados en el paso 1 (ver Tabla 24) y obtenemos el valor ponderado por cada clúster (En la Tabla 26, se detallan los centroides obtenidos de la aplicación de la técnica de segmentación K-means con 5 clúster)

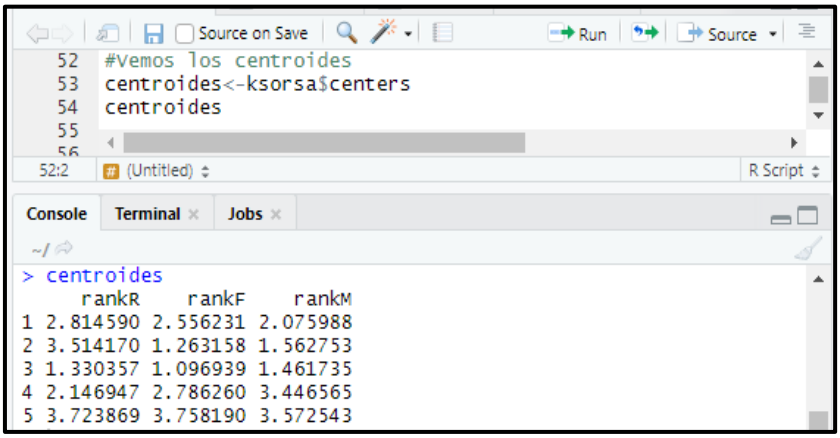


Figura 21. Código en R – Centroides (Elaboración Propia)

Tabla 26

Centroides de cada variable por clúster

Centroides	R	F	M
Cluster 1	2.81	2.56	2.08
Cluster 2	3.51	1.26	1.56
Cluster 3	1.33	1.10	1.46
Cluster 4	2.15	2.79	3.45
Cluster 5	3.72	3.76	3.57

Fuente: Elaboración propia

$$P = \frac{R \times Peso(R) + F \times Peso(F) + M \times Peso(M)}{4}$$

- Luego de asignar los pesos respectivos a cada variable, asignamos la tipificación A-Z, en forma descendiente según el valor ponderado, tal como se muestra en la Tabla 27.

Tabla 27

Categorización alfabética de acuerdo al ponderado

Centroides	R	F	M	Ponderado	Grupo
Clúster 1	2.81	2.56	2.08	2.50	C
Clúster 2	3.51	1.26	1.56	1.90	D
Clúster 3	1.33	1.10	1.46	1.25	E
Clúster 4	2.15	2.79	3.45	2.80	B
Clúster 5	3.72	3.76	3.57	3.70	A

Fuente: Autor

- Asignamos a cada cliente, de acuerdo a su aproximación a los centroides de cada clúster encontrado, utilizando la fórmula de distancia euclidiana y le asignamos la tipificación correspondiente (En la Tabla 28Tabla , se muestra un fragmento del proceso de asignación de cada cliente de acuerdo a su aproximación a los centroides definidos en la Tabla 27, por medio de la distancia euclidiana)

Fórmula euclidiana:

$$D_{(i,j)} = \sqrt{(X1_i - X1_j)^2 + (X2_i - X2_j)^2 + \ldots + (XZ_i - XZ_j)^2}$$

Tabla 28

Fragmento de resultado del cálculo de la distancia euclidiana por clúster

ID	R	F	M	Distancia a los centroides					Más cercano	Grupo
				A	B	C	D	E		
C005	4	4	4	**0.56**	2.28	2.68	3.70	4.69	0.56	A
C007	2	4	2	2.35	1.89	**1.66**	3.16	3.03	1.66	C
C015	1	1	1	4.65	3.24	2.62	2.59	**0.58**	0.58	E
C020	4	3	1	2.70	3.08	**1.66**	1.89	3.31	1.66	C
C027	1	3	1	3.82	2.71	2.16	3.11	**1.99**	1.99	E
C037	4	3	3	**0.99**	1.92	1.57	2.31	3.62	0.99	A
C044	3	3	3	1.19	**0.99**	1.04	2.31	2.96	0.99	B

Fuente: Elaboración propia

```
56  #Distancia euclidiana
57  sorsaeu<-rfmsorsa[,c(1,5,6,7)]
58
59  #A<-5 /B<-4 /C<-1 /D<-2 /E<-3
60
61  sorsaeu$A<-sqrt((sorsaeu$rankR - centroides[5,1])^2
62                  +(sorsaeu$rankF - centroides[5,2])^2
63                  +(sorsaeu$rankM - centroides[5,3])^2)
64
65  sorsaeu$B<-sqrt((sorsaeu$rankR - centroides[4,1])^2
66                  +(sorsaeu$rankF - centroides[4,2])^2
67                  +(sorsaeu$rankM - centroides[4,3])^2)
68
69  sorsaeu$C<-sqrt((sorsaeu$rankR - centroides[1,1])^2
70                  +(sorsaeu$rankF - centroides[1,2])^2
71                  +(sorsaeu$rankM - centroides[1,3])^2)
72
73  sorsaeu$D<-sqrt((sorsaeu$rankR - centroides[2,1])^2
74                  +(sorsaeu$rankF - centroides[2,2])^2
75                  +(sorsaeu$rankM - centroides[2,3])^2)
76
77  sorsaeu$E<-sqrt((sorsaeu$rankR - centroides[3,1])^2
78                  +(sorsaeu$rankF - centroides[3,2])^2
79                  +(sorsaeu$rankM - centroides[3,3])^2)
80
81  #Valor próximo
82  vp<-sorsaeu[,5:9]
83  valor_p<-apply(vp, 1, function(x) {which.min(x)[1]})
84  sorsaeu$vp<-valor_p
85  sorsaeu$vp[sorsaeu$vp==1]<-"C"
86  sorsaeu$vp[sorsaeu$vp==2]<-"D"
87  sorsaeu$vp[sorsaeu$vp==3]<-"E"
88  sorsaeu$vp[sorsaeu$vp==4]<-"B"
89  sorsaeu$vp[sorsaeu$vp==5]<-"A"
90
```

Figura 22. Código en R – Determinación de segmentos (Elaboración propia)

sorsaeu

Filter

	iduser	rankR	rankF	rankM	A	B	C	D	E	vp
1	4	1	2	1	4.1386800	2.8141376	2.1817146	2.6796788	1.0667026	E
2	5	4	4	4	0.5634180	2.2832572	2.6816952	3.6968241	4.6901544	A
3	6	3	4	2	1.7479558	2.0720560	1.4576076	2.8188403	3.3919318	C
4	7	2	4	2	2.3458660	1.8940193	1.6594576	3.1581958	3.0275263	C
5	8	4	4	4	0.5634180	2.2832572	2.6816952	3.6968241	4.6901544	A
6	9	1	3	2	3.2353061	1.8584191	1.8696098	3.0868846	2.0051203	B
7	10	1	1	1	4.6524244	3.2391186	2.6215149	2.5897865	0.5759617	E
8	11	3	4	2	1.7479558	2.0720560	1.4576076	2.8188403	3.3919318	C
9	13	4	4	1	2.5985952	3.3004019	2.1557867	2.8360230	3.9708886	C
10	15	1	1	1	4.6524244	3.2391186	2.6215149	2.5897865	0.5759617	E
11	17	1	1	1	4.6524244	3.2391186	2.6215149	2.5897865	0.5759617	E
12	18	1	2	2	3.6032744	2.0065494	1.8994480	2.6561574	1.1019914	E
13	20	4	3	1	2.6961228	3.0765519	1.6612882	1.8892703	3.3108660	C
14	22	1	1	1	4.6524244	3.2391186	2.6215149	2.5897865	0.5759617	E

Figura 23. Vista en R – Asignación de clientes

PASO 6: MATRIZ COMPARATIVA

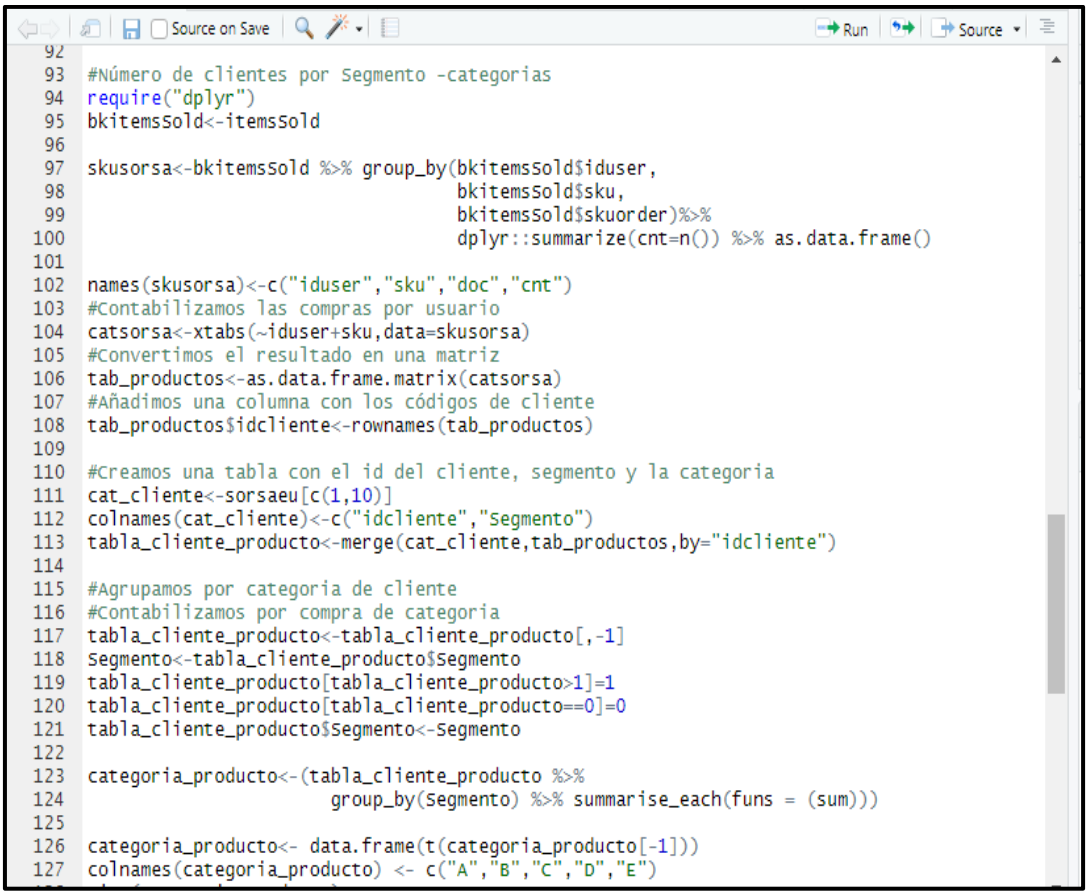
- Elaboramos una matriz, donde se contabilice el número de clientes por segmento, que compra una categoría de producto (Para nuestro caso tenemos 25 categorías, en la Tabla 29, mostramos un ejemplo de los resultados del número de clientes por segmento que compra un determinado grupo de producto)

Tabla 29

Número de clientes por segmento – categoría de producto

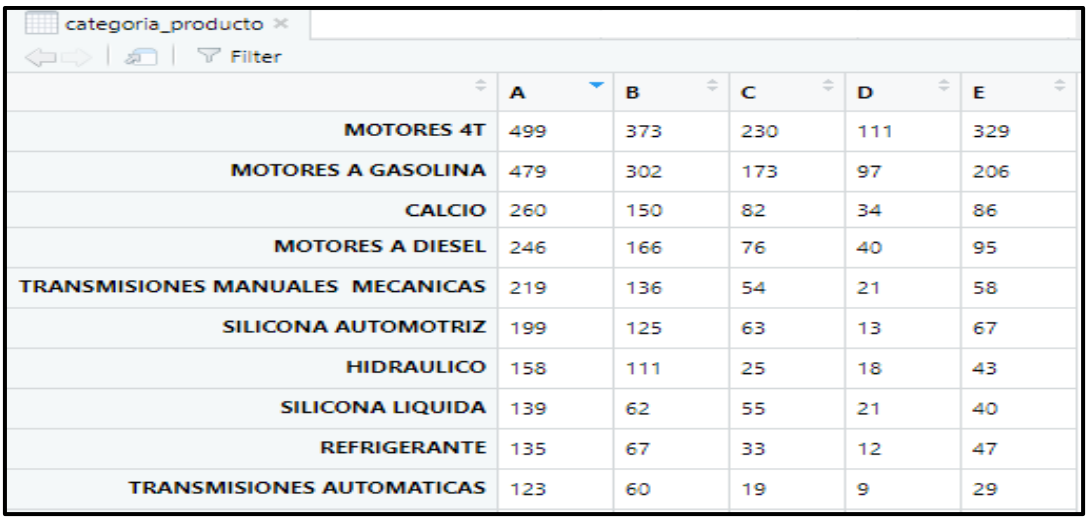
Categoría producto	A	B	C	D	E
Motores 4t	499	373	230	111	329
Motores a gasolina	479	302	173	97	206
Calcio	260	150	82	34	86
Motores a diesel	246	166	76	40	95
Transmisiones manuales mecanicas	219	136	54	21	5
Silicona automotriz	199	125	63	13	67
Hidraulico	158	111	25	18	43
Silicona liquida	139	62	55	21	40
Refrigerante	135	67	33	12	47
Transmisiones automáticas	123	60	19	9	29
Pintura en spray	106	50	44	13	27
Cinta	91	24	21	11	13
Dot3	80	56	30	9	21
Lithium	75	40	10	8	14
Limpiadores y desengrasantes	70	47	24	6	21
Filtros	55	30	16	13	38
Motores 2t	48	27	12	9	24
Tractores agrícolas	28	15	9	4	10
Aditivos	12	5	2	1	3
Sodio	7	4	5	1	3
Cera liquida	3	0	2	0	0
Frenos	3	3	2	0	0
Petroleo	3	1	0	1	1
Aceite	2	0	0	0	0
Dot4	0	2	0	0	0

Fuente: Elaboración propia



```
#Número de clientes por Segmento -categorias
require("dplyr")
bkitemsSold<-itemsSold

skusorsa<-bkitemsSold %>% group_by(bkitemsSold$iduser,
                                   bkitemsSold$sku,
                                   bkitemsSold$skuorder)%>%
                                   dplyr::summarize(cnt=n()) %>% as.data.frame()

names(skusorsa)<-c("iduser","sku","doc","cnt")
#Contabilizamos las compras por usuario
catsorsa<-xtabs(~iduser+sku,data=skusorsa)
#Convertimos el resultado en una matriz
tab_productos<-as.data.frame.matrix(catsorsa)
#Añadimos una columna con los códigos de cliente
tab_productos$idcliente<-rownames(tab_productos)

#Creamos una tabla con el id del cliente, segmento y la categoria
cat_cliente<-sorsaeu[c(1,10)]
colnames(cat_cliente)<-c("idcliente","Segmento")
tabla_cliente_producto<-merge(cat_cliente,tab_productos,by="idcliente")

#Agrupamos por categoria de cliente
#Contabilizamos por compra de categoria
tabla_cliente_producto<-tabla_cliente_producto[,-1]
Segmento<-tabla_cliente_producto$Segmento
tabla_cliente_producto[tabla_cliente_producto>1]=1
tabla_cliente_producto[tabla_cliente_producto==0]=0
tabla_cliente_producto$Segmento<-Segmento

categoria_producto<-(tabla_cliente_producto %>%
                       group_by(Segmento) %>% summarise_each(funs = (sum)))

categoria_producto<- data.frame(t(categoria_producto[-1]))
colnames(categoria_producto) <- c("A","B","C","D","E")
```

Figura 24. Código en R – Matriz categoría por segmento (Elaboración propia)



categoria_producto

Filter

	A	B	C	D	E
MOTORES 4T	499	373	230	111	329
MOTORES A GASOLINA	479	302	173	97	206
CALCIO	260	150	82	34	86
MOTORES A DIESEL	246	166	76	40	95
TRANSMISIONES MANUALES MECANICAS	219	136	54	21	58
SILICONA AUTOMOTRIZ	199	125	63	13	67
HIDRAULICO	158	111	25	18	43
SILICONA LIQUIDA	139	62	55	21	40
REFRIGERANTE	135	67	33	12	47
TRANSMISIONES AUTOMATICAS	123	60	19	9	29

Figura 25. Vista de Tabla en R - Tabla compras de segmento por categoría

- Aplicamos el enfoque de cuartiles para puntuar las categorías de productos en rangos de 1-4, para cada categoría de cliente. (En la Tabla 30, se muestra un ejemplo de la aplicación del enfoque de cuartiles aplicada a la categoría de producto)

Tabla 30

Puntuación de cuartiles por categoría de producto-segmento

CATEGORÍA PRODUCTO	A	B	C	D	E
Motores 4t	4	4	4	4	4
Motores a gasolina	4	4	4	4	4
Calcio	4	4	4	4	4
Motores a diesel	4	4	4	4	4
Transmisiones manuales mecánicas	4	4	3	4	4
Silicona automotriz	4	4	4	3	4
Hidraulico	3	3	3	3	3
Silicona liquida	3	3	4	4	3
Refrigerante	3	3	3	3	3
Transmisiones automáticas	3	3	2	2	3
Pintura en spray	3	3	3	3	3
Cinta	3	2	2	3	2
Dot3	2	3	3	2	2
Lithium	2	2	2	2	2
Limpiadores y desengrasantes	2	2	3	2	2
Filtros	2	2	2	3	3
Motores 2t	2	2	2	2	2
Tractores agrícolas	2	2	2	2	2
Aditivos	1	1	1	1	1
Sodio	1	1	1	1	1
Cera liquida	1	1	1	1	1
Frenos	1	1	1	1	1
Petroleo	1	1	1	1	1
Aceite	1	1	1	1	1
Dot4	1	1	1	1	1

Fuente: Elaboración propia

```
#ordenamos la tabla para el gráfico final
ult_fila<-nrow(categoria_producto)
a<- cbind(rownames(categoria_producto),rep("A",ult_fila),categoria_producto$rnkA)
b<- cbind(rownames(categoria_producto),rep("B",ult_fila),categoria_producto$rnkB)
c<- cbind(rownames(categoria_producto),rep("C",ult_fila),categoria_producto$rnkC)
d<- cbind(rownames(categoria_producto),rep("D",ult_fila),categoria_producto$rnkD)
e<- cbind(rownames(categoria_producto),rep("E",ult_fila),categoria_producto$rnkE)
tabla_final<-rbind(a,b,c,d,e)

colnames(tabla_final)<-c("Producto","Segmento","RnkP")
tabla_final<-as.data.frame(tabla_final)
tabla_final$RnkP<-as.numeric(tabla_final$RnkP)

ggplot(tabla_final,
       aes(x=Producto,
           y=RnkP,
           group = Segmento,
           colour=Segmento))+
  geom_line()+
  geom_point( size=2, shape=5, fill="white") +
  theme_minimal()
```

Figura 26. Código en R – Gráfico lineal de segmentos (Elaboración Propia).

3.2 Resultados y Discusión

a) Visualización de Datos

En la siguiente figura se muestra la distribución de los segmentos visualizados en un gráfico en R.

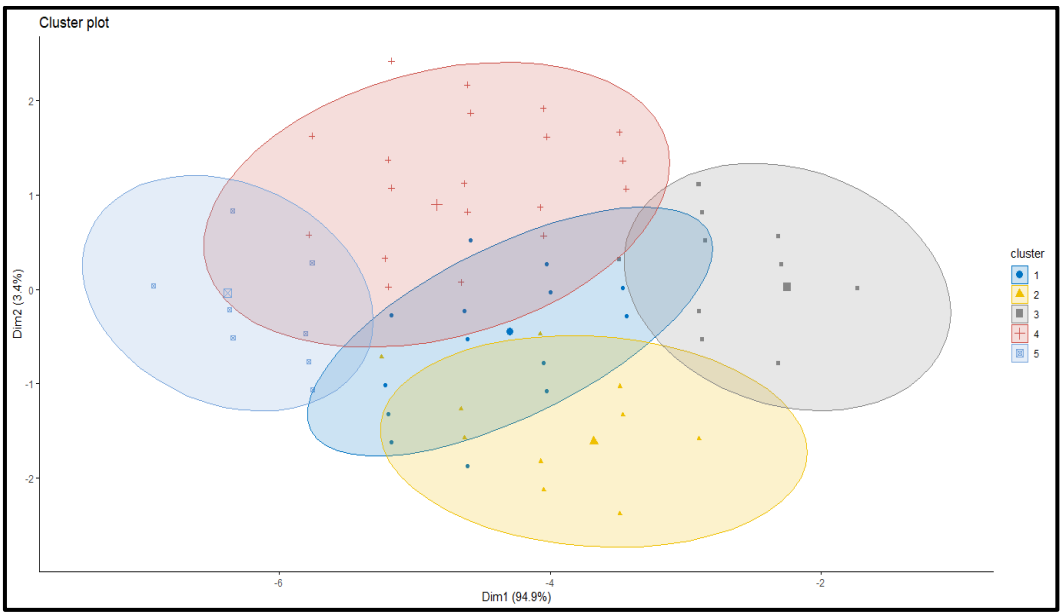


Figura 27. Gráfico en R- Visualización de 5 clústeres por K-means

b) Matriz de resultados e interpretaciones

Obtenemos la matriz final (Tabla) y realizamos algunas interpretaciones con los segmentos encontrados:

Tabla 31. *Matriz final comparativa de los segmentos A, B, C, D, E.*

	A	B	C	D	E
4	Motores 4t	Motores 4t	Motores 4t	Motores 4t	Motores 4t
	Motores a gasoline	Motores a gasolina	Motores a gasoline	Motores a gasoline	Motores a gasoline
	Calcio	Calcio	Calcio	Calcio	Calcio
	Motores a diesel	Motores a diesel	Motores a diesel	Motores a diesel	Motores a diesel
	Transmisiones manuales mecanicas	Transmisiones manuales mecanicas	Silicona liquida	Transmisiones manuales mecanicas	Transmisiones manuales mecanicas
	Silicona automotriz	Silicona automotriz	Silicona automotriz	Silicona liquida	Silicona automotriz
3	Hidraulico	Hidraulico	Hidraulico	Hidraulico	Hidraulico
	Silicona liquida	Silicona liquida	Transmisiones manuales mecanicas	Silicona automotriz	Silicona liquida
	Refrigerante	Refrigerante	Refrigerante	Refrigerante	Refrigerante
	Transmisiones automaticas	Transmisiones automaticas	Limpiadores y desengrasantes	Filtros	Transmisiones automaticas
	Pintura en spray	Pintura en spray	Pintura en spray	Pintura en spray	Pintura en spray
	Cinta	Dot3	Dot3	Cinta	Filtros
2	Dot3	Cinta	Cinta	Dot3	Dot3
	Lithium	Lithium	Lithium	Lithium	Lithium
	Limpiadores y desengrasantes	Limpiadores y desengrasantes	Transmisiones Automaticas	Limpiadores y desengrasantes	Limpiadores y desengrasantes
	Filtros	Filtros	Filtros	Transmisiones automaticas	Cinta
	Motores 2t	Motores 2t	Motores 2t	Motores 2t	Motores 2t
	Tractores agricolas	Tractores agricolas	Tractores agricolas	Tractores agricolas	Tractores agricolas
1	Aditivos	Aditivos	Aditivos	Aditivos	Aditivos
	Sodio	Sodio	Sodio	Sodio	Sodio
	Cera liquida	Cera liquida	Cera liquida	Cera liquida	Cera liquida
	Frenos	Frenos	Frenos	Frenos	Frenos
	Petroleo	Petroleo	Petroleo	Petroleo	Petroleo
	Aceite	Aceite	Aceite	Aceite	Aceite
	Dot4	Dot4	Dot4	Dot4	Dot4

Fuente: Elaboración propia

Interpretación 1: De la Matriz comparativa, podemos deducir que las categorías de productos que se encuentran en el cuartil 4 y el segmento A, tienen una mayor frecuencia de ventas:

-MOTORES 4T

-MOTORES A GASOLINA

-CALCIO

-MOTORES A DIESEL

-TRANSMISIONES MANUALES MECANICAS

-SILICONA AUTOMOTRIZ

Estrategias:

- Se pueden utilizar las categorías de productos de este cuartil para fomentar una mejora en la frecuencia, monto o recencia de ventas de los cuartiles 1 o 2 del mismo segmento A.

- Al ser el grupo con mayor frecuencia de ventas, se pueden utilizar para fidelizar al segmento A de clientes.

- Se pueden utilizar los productos como portafolio sugerido en el trabajo con los clientes del segmento D y E. Así como en la captación de clientes nuevos.

Interpretación 2: De la Matriz comparativa, podemos deducir que los grupos de productos que se encuentran en el cuartil 1 y el segmento E, tienen una menor frecuencia de ventas:

- ADITIVOS

- SODIO

- CERA LIQUIDA

- FRENOS

- PETROLEO

- ACEITE

- DOT4

Estrategias:

- Se pueden utilizar los grupos de productos de este cuartil para crear promociones con los del cuartil 4 para mejorar en la frecuencia, monto o recencia de ventas del mismo segmento E.

- Al ser el grupo con menor frecuencia de ventas, se deben analizar las posibles causas y soluciones.

- Se pueden utilizar vínculos comerciales con clientes estratégicos para fomentar la venta de estos productos.

-Como complemento a las estrategias o a la ejecución del método se muestran los siguientes gráficos para análisis comparativo entre segmentos.

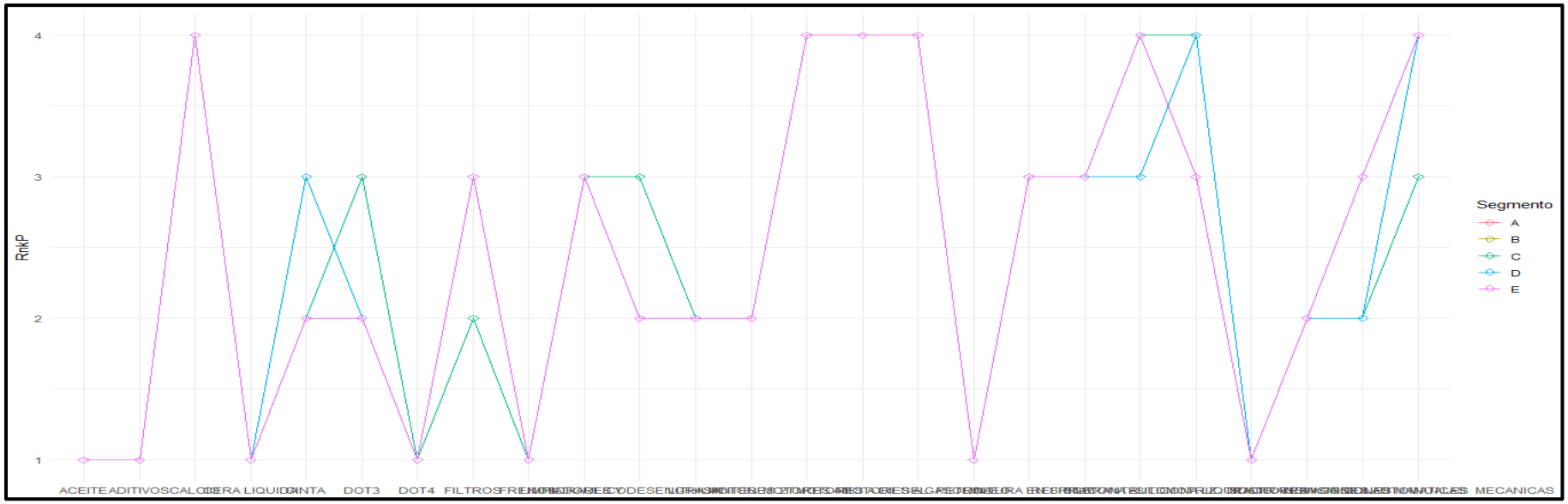


Figura 28. Gráfico en R – Diagrama de líneas de los segmentos - categoría de productos.

Interpretación: De la Figura 28, podemos observar que existen diferencias en cada uno de los 05 segmentoss de clientes generados por la técnica de Segmentación K means , respecto a la frecuencia de compra de las diversas categorías que ofrece la empresa Suministros del Oriente SA.

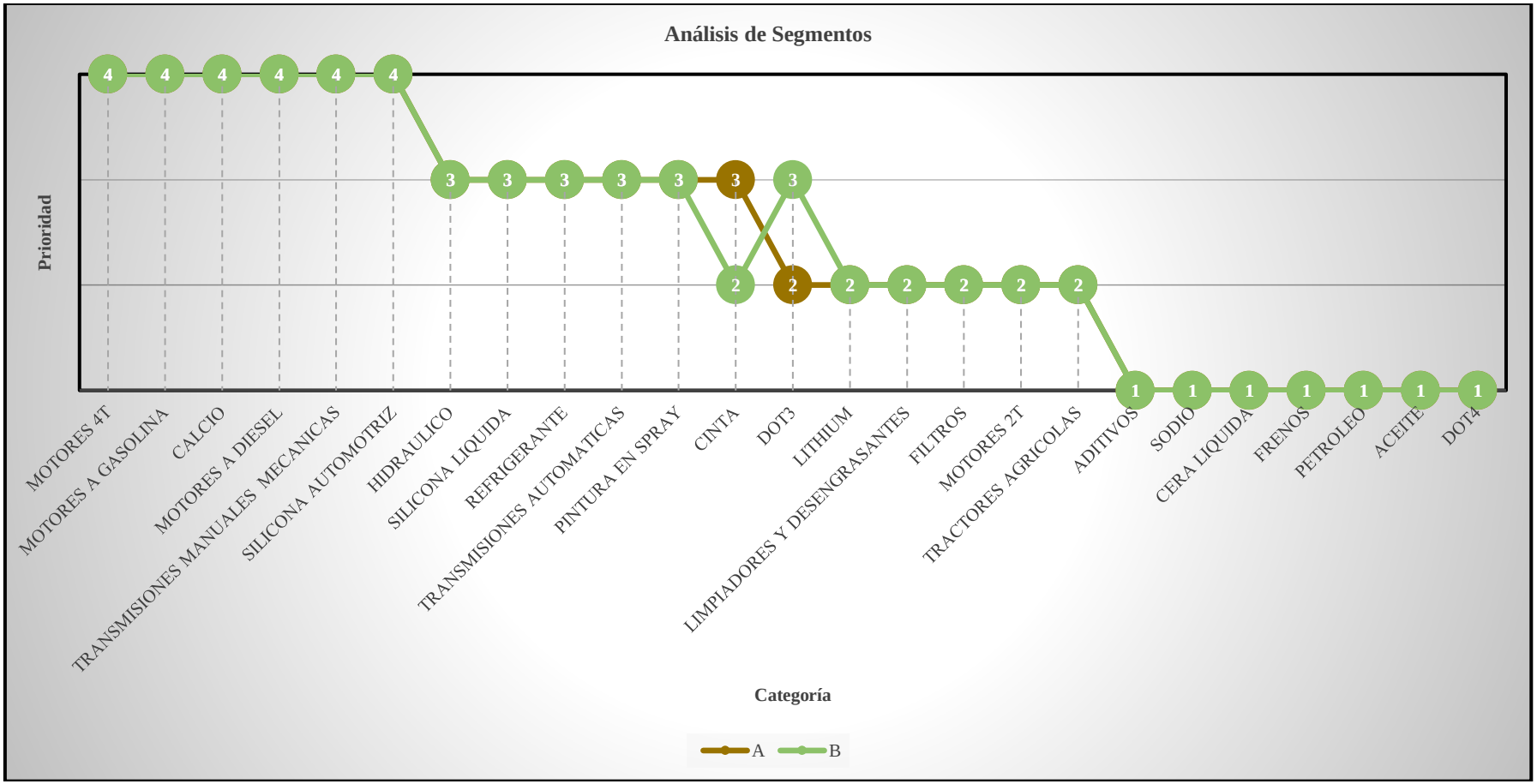


Figura 29. Segmento A vs B (Elaboración propia)

Interpretación: En la Figura 29, podemos ver que los segmenteos A y B tienen comportamientos de compra similares en las categorías de productos con puntuación 4 y 1, solo existe una pequeña variación en 02 de las categorías.

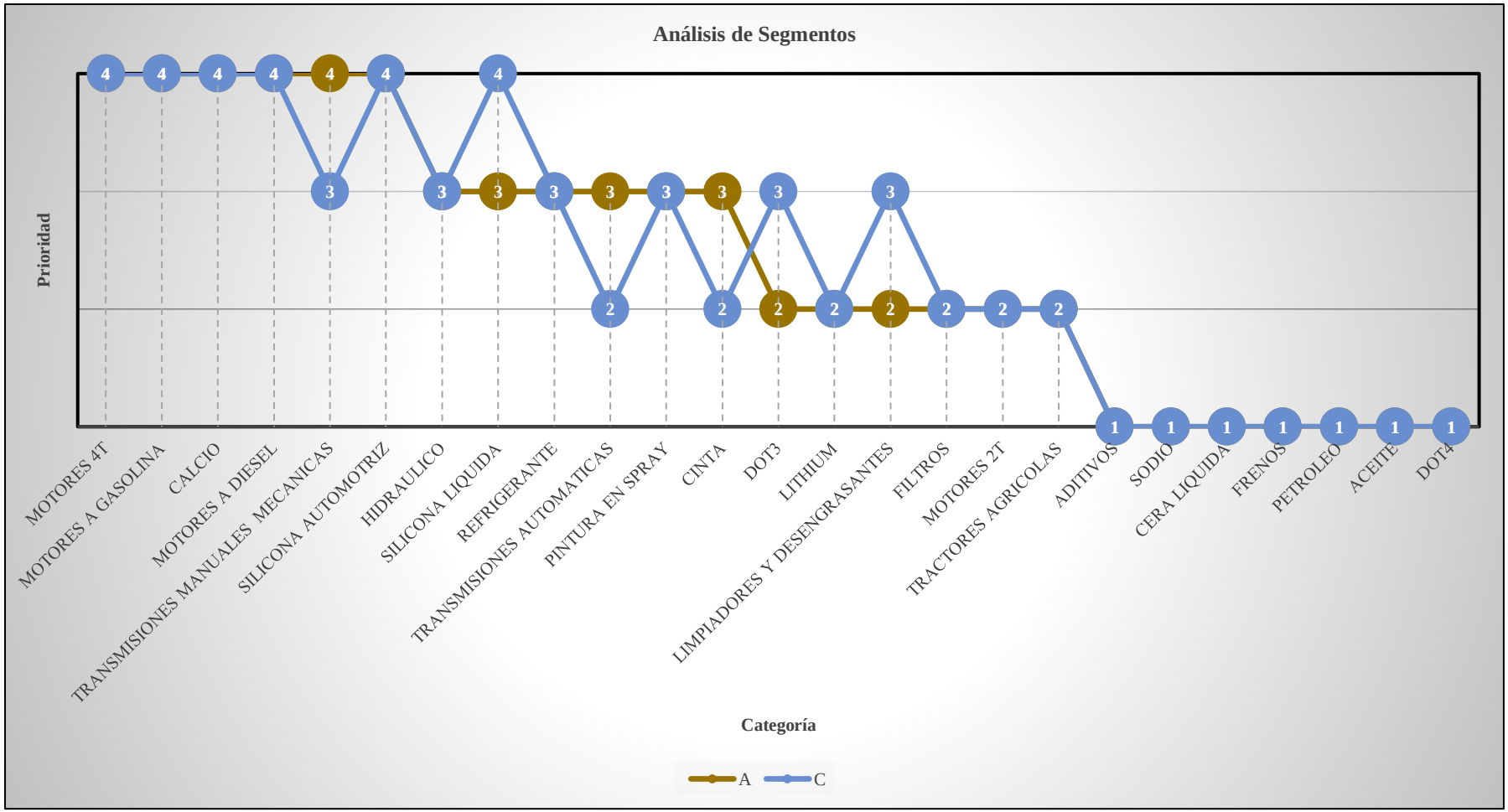


Figura 30. Segmento A vs C (Elaboración propia)

Interpretación: La Figura 30, nos muestra que los segmenteos A y C tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 06 variaciones demostrando que tienen una conducta de compra diferenciada.

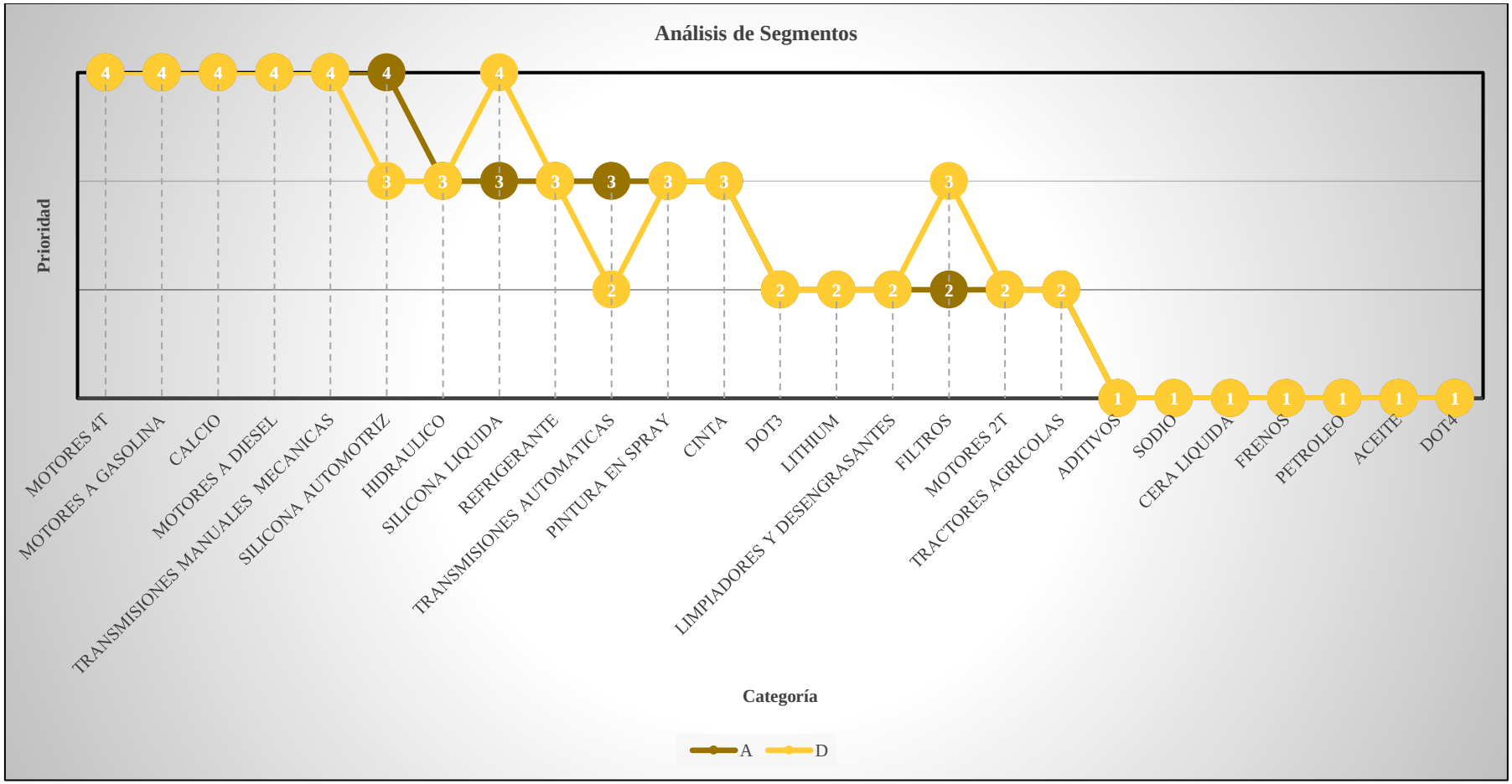


Figura 31. Segmento A vs D (Elaboración propia)

Interpretación: En la Figura 31, nos muestra que los segmentos A y D tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 04 variaciones demostrando que tienen una conducta de compra diferenciada.

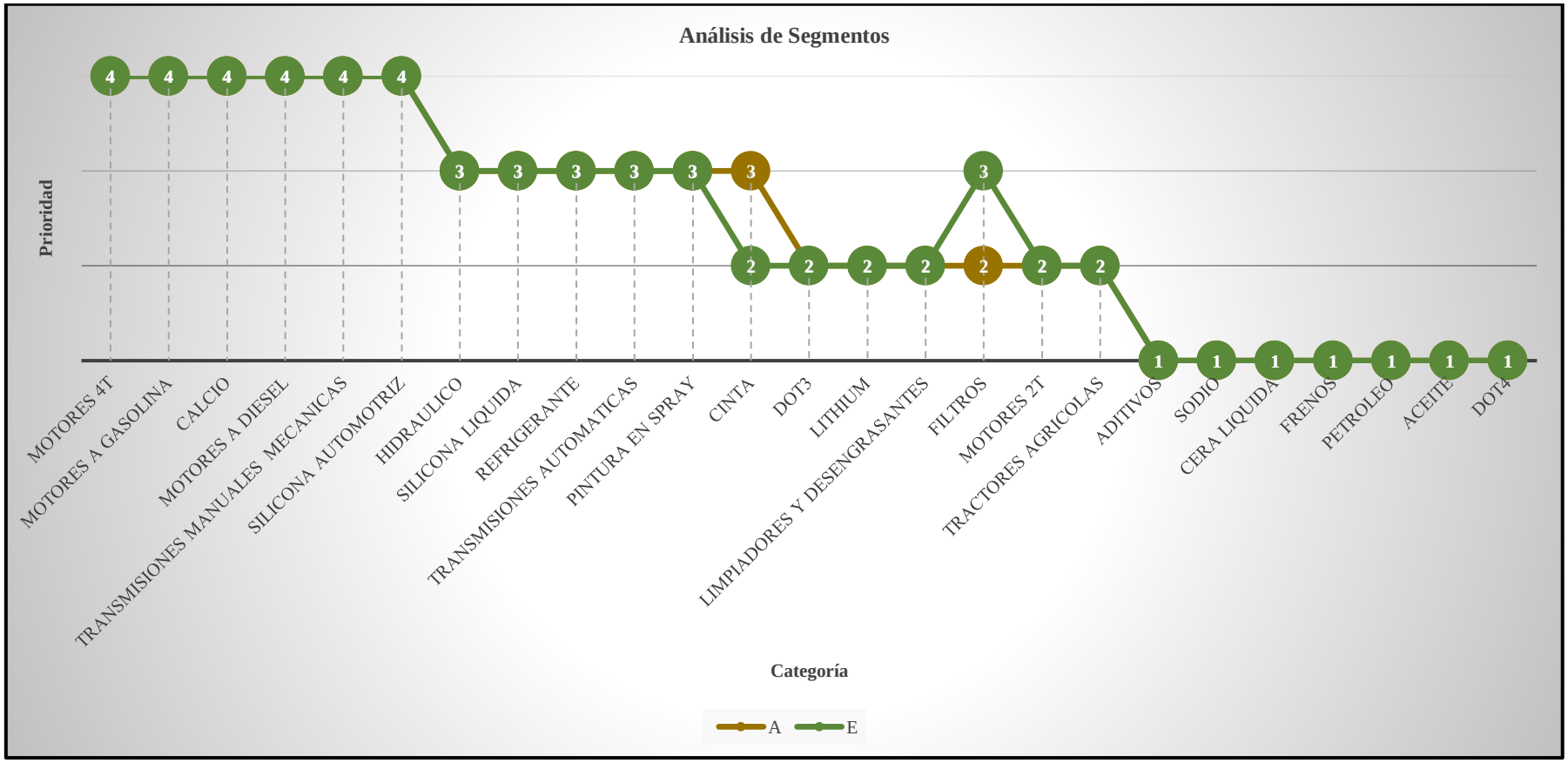


Figura 32. Segmento A vs E (Elaboración propia)

Interpretación: La Figura 32,***Figura 32*** nos muestra que los segmentos A y E tienen comportamientos de compra similares en las categorías de productos con puntuación 4 y 1, solo existe una pequeña variación en 02 de las categorías.

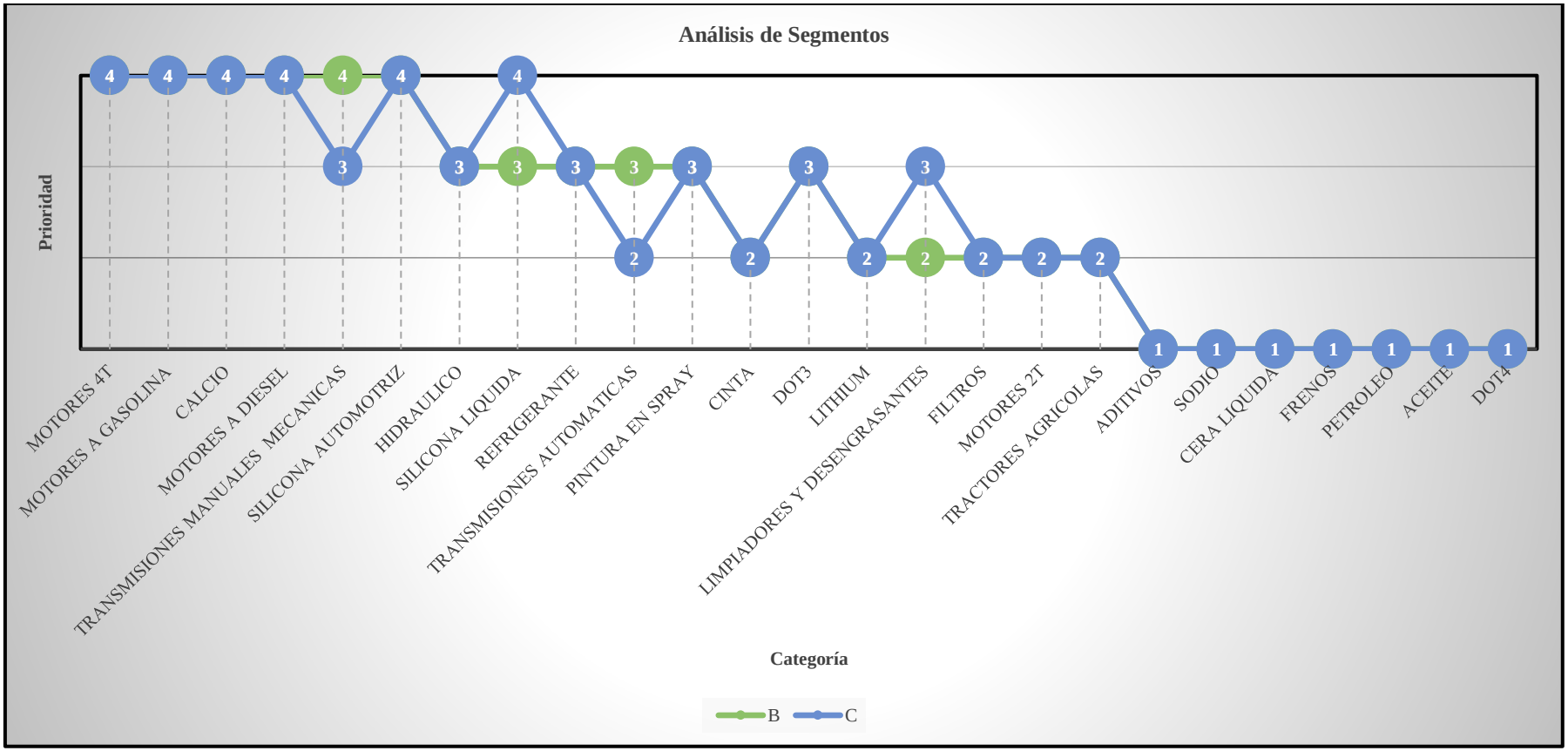


Figura 33. Segmento B vs C (Elaboración propia)

Interpretación: La Figura 33, nos muestra que los segmentos B y C tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 04 variaciones demostrando que tienen una conducta de compra diferenciada.

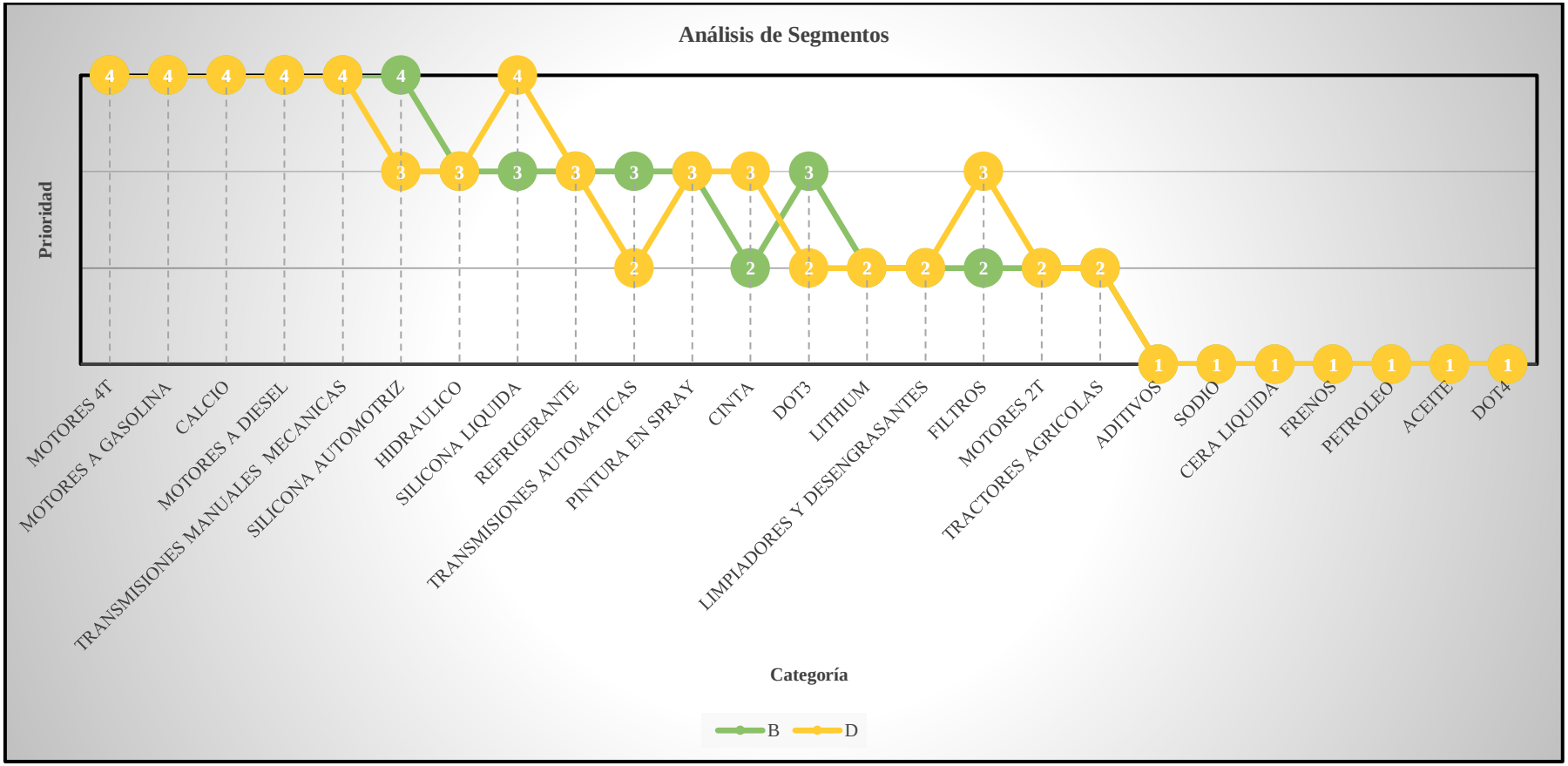


Figura 34. Segmento B vs D (Elaboración propia)

Interpretación: La Figura 34, nos muestra que los segmentos B y D tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 06 variaciones demostrando que tienen una conducta de compra diferenciada.

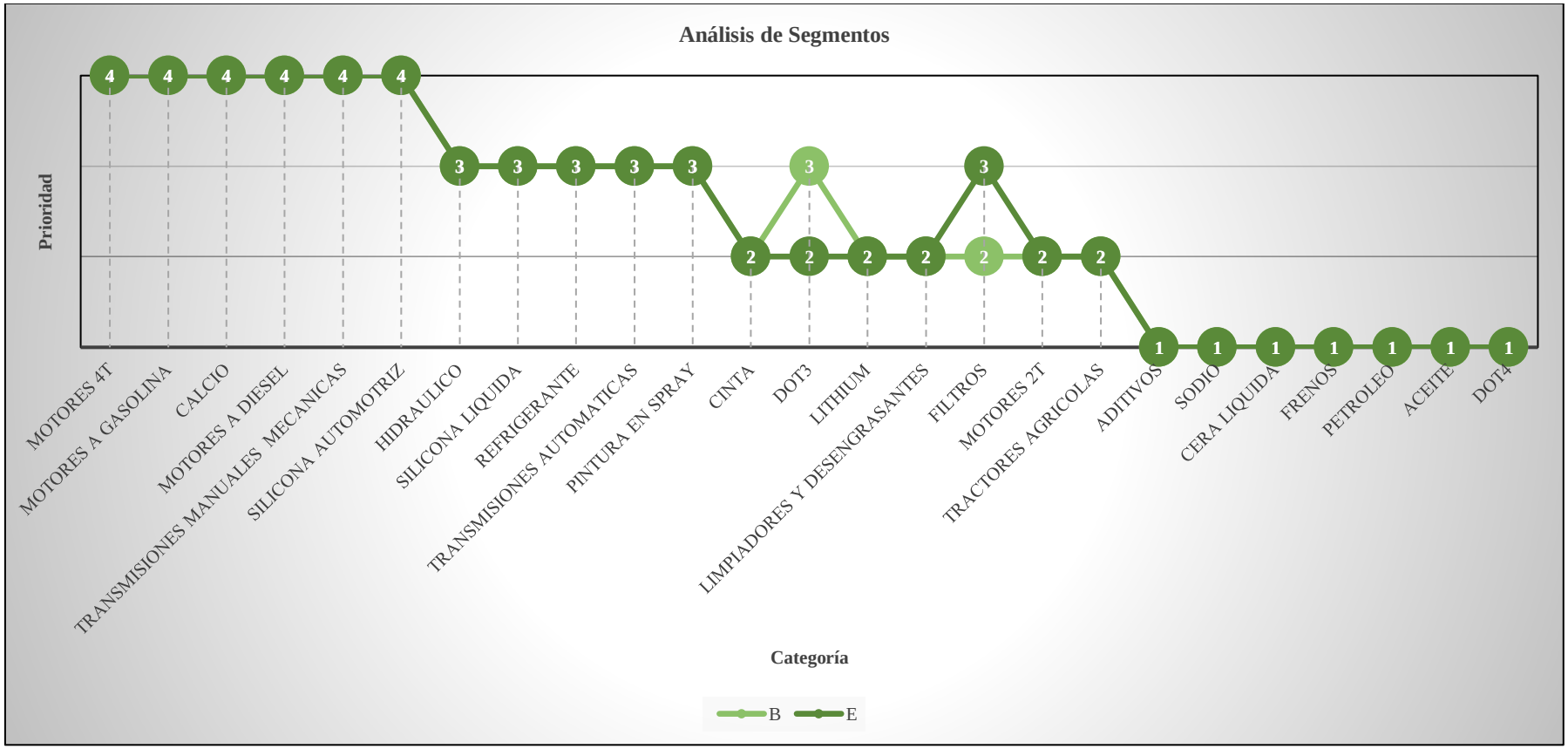


Figura 35. Segmento B vs E (Elaboración propia)

Interpretación: La Figura 35, nos muestra que los segmentos B y E tienen comportamientos de compra similares en las categorías de productos con puntuación 4 y 1, solo existe una pequeña variación en 02 de las categorías.

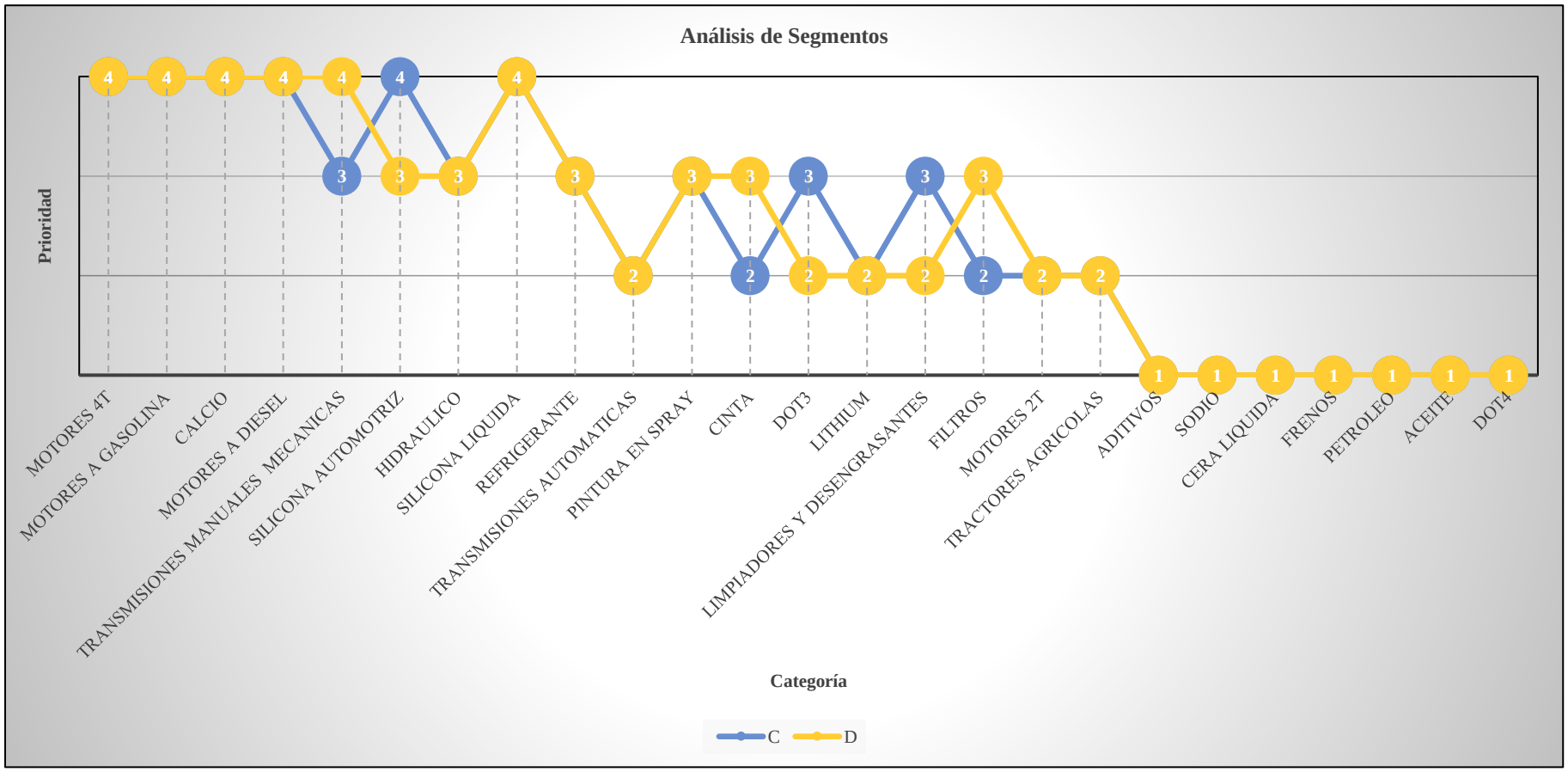


Figura 36. Segmento C vs D (Elaboración propia)

Interpretación: La Figura 36, nos muestra que los segmentos C y D tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 06 variaciones demostrando que tienen una conducta de compra diferenciada.

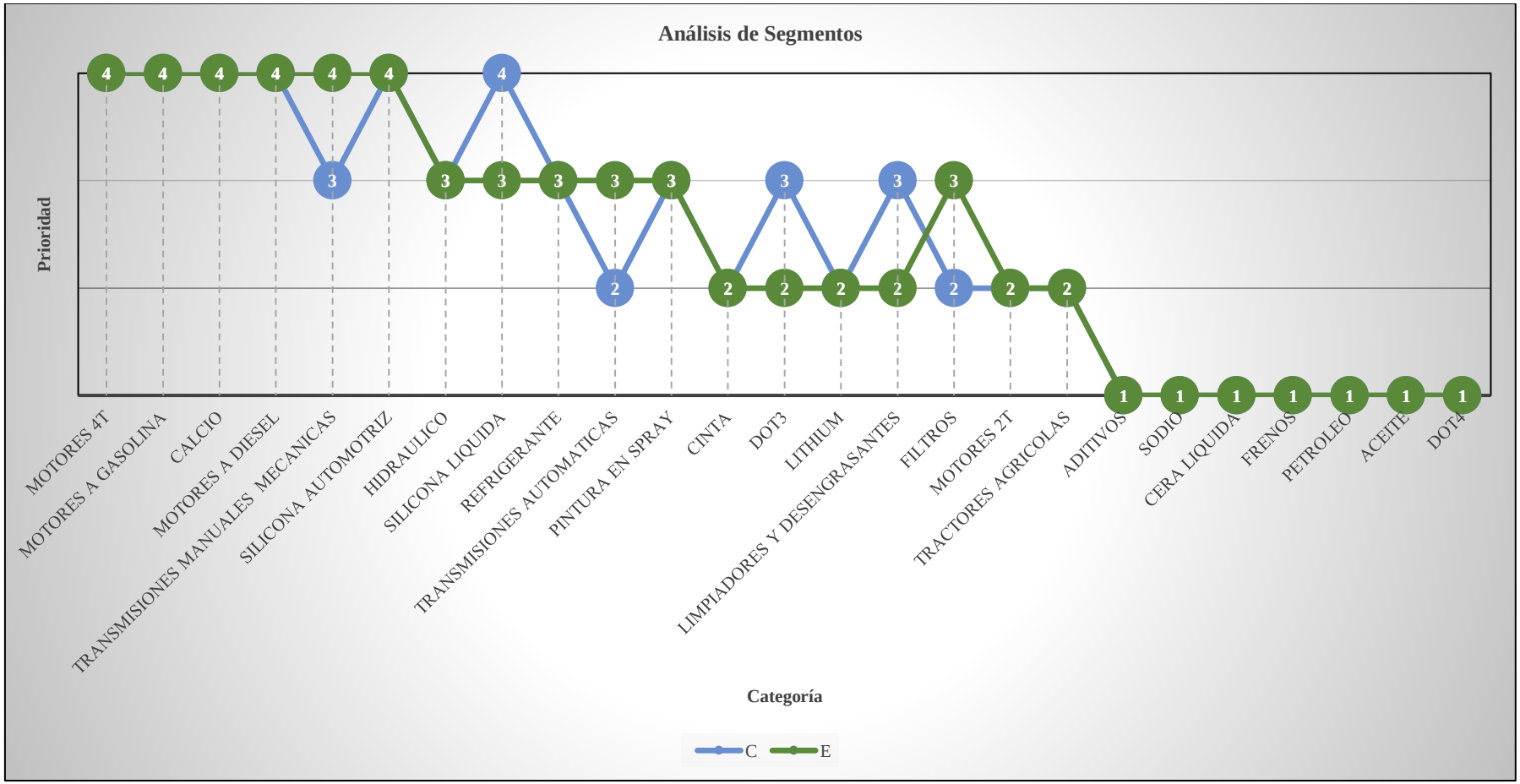


Figura 37. Segmento C vs E (Elaboración propia)

Interpretación: La Figura 37 nos muestra que los segmentos C y E tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 06 variaciones demostrando que tienen una conducta de compra diferenciada.

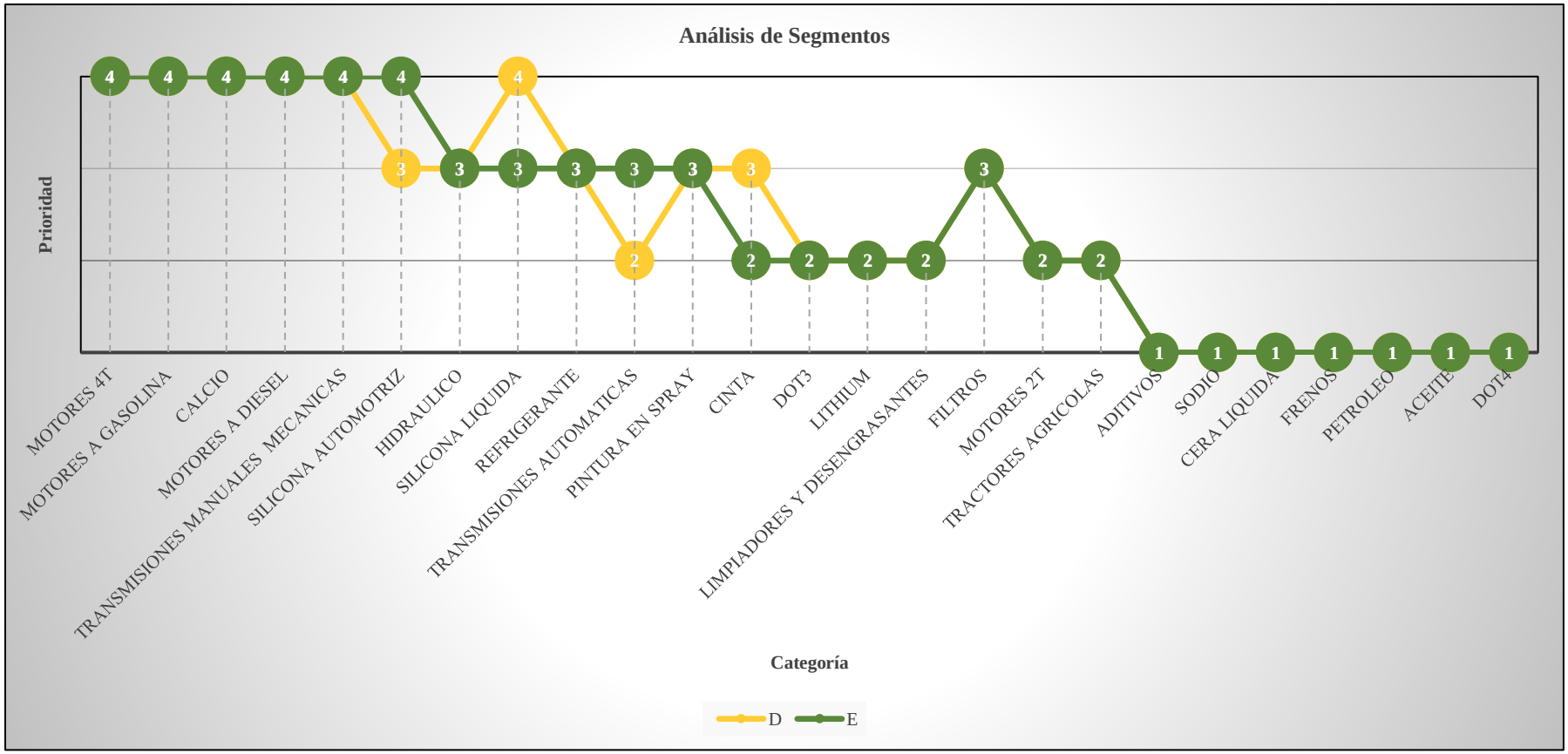


Figura 38. Segmento D vs E (Elaboración propia)

Interpretación: La Figura 38 nos muestra que los segmentos D y E tienen comportamientos de compra similares solo en la categoría de productos con puntuación 1, existen 04 variaciones demostrando que tienen una conducta de compra diferenciada.

3.2 Discusión de resultados

El modelo presentado para la segmentación de clientes en la empresa Suministros del Oriente S.A.C, ha sido elaborado teniendo en cuenta las variables RFM que intervienen en el proceso de segmentación de clientes y siguiendo la metodología KDD, en este caso se realizó un modelo de segmentación.

- De los resultados obtenidos en el punto anterior, podemos indicar que el modelo permite segmentar a los clientes, de acuerdo al caso de estudio, en 3 segmentos utilizando las variables RFM que también se utilizaron en las siguientes investigaciones: Durson y caber (2016), Cuadros et al. (2017), Batchiar (2018) y Aryuni et al. (2018).

- El modelo propuesto permite identificar las características comerciales de los clientes de la empresa suministros del oriente, que servirán como apoyo en la toma de decisiones. Esta afirmación tiene concordancia con las investigaciones de Durson y caber (2016), Cuadros et al. (2017), Qadadeh y Abadallah (2018), Batchiar (2018), Aryuni et al. (2018), Flores et al. (2019), Laura et al. (2016), Reyes (2018), De la Cruz (2017).

- El modelo propuesto también contempla el uso de enfoque de cuartiles y la agrupación por categoría de productos complementando la información brindada por las investigaciones citadas en la presente tesis, generando un nuevo análisis para la toma de decisiones.

CONCLUSIONES

1. La presente investigación permitió diseñar un modelo basado en técnicas de minería de datos que permite la segmentación de clientes en la empresa Suministros del Oriente.

2. La aplicación del modelo logró identificar los segmentos de clientes de la empresa, de acuerdo a sus características comerciales de recencia, frecuencia y monto, así como sus características de hábitos de compra de acuerdo a los grupos de productos.

3. El presente trabajo de investigación propone de manera satisfactoria, un modelo de segmentación de clientes basado en KDD y técnicas de minerías de datos. Tomando en cuenta el análisis de las investigaciones previas, además de la información y los datos proporcionados por la empresa Suministros del Oriente SA.

4. La aplicación del modelo en el caso de estudio permitió recomendar estrategias comerciarles de acuerdo a los segmentos encontrados.

5. El modelo propuesto permite utilizar diferentes agrupaciones de productos de acuerdo a la necesidad de la organización que lo aplique. Estas agrupaciones pueden ser, por ejemplo: marca, categoría, clasificación o tipo de uso de los productos.

6. La herramienta R permitió llevar a cabo todo el proceso de KDD en todos los pasos propuestos por el modelo, ya que contiene una gran cantidad de librerías complementarías (funciones estadísticas, algoritmos de minería, gráficas, etc.). Por lo tanto, proporciona una mejor experiencia al momento de utilizar una sola herramienta en toda la investigación.

RECOMENDACIONES

Al finalizar el presente trabajo de investigación es necesario dejar las siguientes recomendaciones que son importantes para la entidad o demás organizaciones que tengan acceso a esta información.

1. Es importante definir el objetivo que se pretende analizar con la utilización de este modelo, ya que en base a ello se podrán seleccionar los datos complementarios para que los resultados finales, sean importantes en la toma de decisiones de la organización.

2. En el paso de limpieza y preprocesamiento de datos se debe realizar una exhaustiva revisión de los datos, antes de ser sometido al siguiente paso, ya que esto tiene un alto índice de influencia en la confiabilidad de los resultados finales de la aplicación del modelo.

3. El modelo permite a las organizaciones definir el número de segmentos de clientes que desee encontrar, de no tener un conocimiento exacto del mercado, se sugiere aplicar algunos métodos como lo son el de la silueta, el del codo, la suma de cuadrados, entre otros. Esto permite tener una idea básica sobre los segmentos que proporciona la propia data almacenada.

4. Existen diversos algoritmos de segmentación, una forma recomendable de definir cuál técnica es la mejor, es utilizar el método de la silueta promedio.

5. El modelo de segmentación de clientes proporciona una vista sobre el nivel de consumo de los grupos de productos en cada segmento, es importante que la organización defina estrategias comerciales adecuadas para cada uno de ellos, ya que, según las revisiones realizadas, es importante dar prioridad a los clientes de alto valor teniendo en cuenta que cuesta más adquirir nuevos clientes que mantener satisfechos a los actuales.

.

REFERENCIAS BIBLIOGRÁFICAS

Alaminos Chica, A., Francés García, F. J., Penalva Verdú, C., & Santacreu Fernández, Ó. A. (2015). *Análisis multivariante para las Ciencias Sociales I.* Cuenca: Pydlos ediciones.

Albrecht, T. D., & Savio, T. D. (2015). A (I)LEGALIDADE DO SISTEMA SCORING: (im)possibilidade de concessão de dano moral diante da decisão do Superior Tribunal de Justiça. *Contribuciones a las Ciencias Sociales*(2015-12).

Aldas Manzano, J., & Uriel, J. E. (2017). *Análisis multivariante aplicado con R.* Madrid: Ediciones Parainfo, S.A.

Al-Hagery, M. A., Alfaiz, A. S., Alorini, F. S., & Althunayan, M. S. (2015). Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community. *IJCER, 4*(6), 118-125.

Alvarez Horn, H. I. (2018). *Mercadotecnia al alcance de todos.* Ciudad de México: Grp México.

Álvarez López, A. C., & Sánchez López, D. (2015). La formación del contador público de la Universidad de Antioquia en tecnologías de información y comunicación. *Tesis de Grado*. Antioquia, Colombia: Universidad de Antioquia.

Amador Garcia, M., Baltazar Martínez , M., Rodríguez Camacho, M., & Ruiz Perales, C. (2015). DETERMINACIÓN DE PERFILES DELICTIVOS EN EL ESTADO DE JALISCO UTILIZANDO WEKA A TRAVÉS DE MINERÍA DE DATOS. *TECTZAPIC*(2).

Arroyo López, P. E., & Borja Medina, J. C. (2018). *Análisis multivariante para la inteligencia de mercados.* Monterrey: Editorial Digital- Tecnológico de Monterrey.

Aryuni, M., Madyatmadja, E. D., & Miranda, E. (2018). Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. *In 2018 International Conference on Information Management and Technology (ICIMTech)*, 412-416.

Bachtiar, F. A. (2018). Customer Segmentation Using Two-Step Mining Method Based on RFM Model. *International Conference on Sustainable Information Engineering and Technology (SIET)*, 10-15.

Bastos Boubeta, A. I. (2007). *Fidelización Del Cliente.* Ideaspropias Editorial SL.

Benitez, M. Á., & Arias, Á. (2015). *Curso de Introducción a la Administración de Bases de Datos.* IT Campus Academy.

Bernal Torres, C. A. (2010). *Metodología de la investigación: Administración, Economía, Humanidades y Ciencia Social.* Ciudad de México: Pearson Educación de México.

Bernal, S. (2017). Inteligencia de Mercados.

Bokan Garay, A., Patiño Escarcina, R., & Túpac Valdivia, Y. (2011). Validacion de Clusters usando IEKA y SL-SOM . *Proceedings del X Congreso de la Sociedad Peruana de Computación, CSPC2011(Pucallpa, Perú)(Alex Cuadros-Vargas, ed.), Sociedad Peruana de Computación*, 161-170.

Cabeza, R. (2016). Localización de datos de contactos personales utilizando técnicas de minería web y redes sociales. *Investigación e Innovación en ingenierías, 4*(1).

Castillo Rojas, W., Medina Quispe, F., & Vega Damke, J. (2017). Esquema de Visualización para Modelos de Clústeres en Minería de Datos. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*(21), 67-80. doi:https://doi.org/10.17013/risti.21.67-84

Christy, A., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences.* doi:https://doi.org/10.1016/j.jksuci.2018.09.004

Cuadros, Á. L., Gonzales, C. C., & Jiménez, P. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41-51. doi:https://doi.org/10.14483/22487638.12957

De la Cruz Gutierrez, K. D. (2017). Segmentación de clientes con Inteligencia Analítica para personalizar las Ventas de los Servicios de las Agencias Turísticas. *Tesis de Maestría*. Lima, Perú: Universidad Peruana Unión.

Desgraupes, B. (Noviembre de 2017). *R-project.* Obtenido de https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives, 18*, 153-160. doi:https://doi.org/10.1016/j.tmp.2016.03.001

Escobar Terán , H. E., Alcivar, M., & Puris, A. (2016). Aplicaciones de minería de datos en marketing. *Revista Publicando, 3*(8), 203-512.

Estupiñan Castellanos, A. R., Carvajal, L., & Ochoa, A. (2016). Estudio de la precipitación media mensual a partir de los datos de precipitación del Centro de Climatología de Precipitación Global (GPCC). *XXVII Congreso Latinoamericano de Hidráulica, At Lima, Perú*.

Etzel, M. J., Staton, W., & Walker, B. (2004). *Fundamentos de Marketing.* México: McGrawHill .

Flores Azuela, J. I., Ochoa Hernández, M. L., & Ayup González, J. (2019). Segmentación del comprador online en México: un estudio con base en la frecuencia de compra electrónica. *CIENCIA ergo-sum, 26*(2). doi:https://doi.org/10.30878/ces.v26n2a1

Giraldo Mejía, J. C. (2019). Aplicación de la técnica regresión logística de la minería de datos en el proceso de descubrimiento de conocimiento (KDD) en bases de datos operativas o transaccionales. *Perspectiv@ s, 14*(13), 51-55.

Gončarovs, P. (2018). Using Data Analytics for Customers Segmentation: Experimental Study at a Financial Institution. *In 2018 59th International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, 1-5. doi:https://doi.org/10.1109/ITMS.2018.8552951

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio , M. (2010). *Metodología de la investígación (Quinta Ed.).* México: McGraw-Hill.

Hernandez, H., & Francisco, E. (2017). Aplicativo web para localizar datos de contactos personales utilizando técnicas de minería web y redes sociales. *Revista Investigación y Desarrollo en TIC, 5*(2), 1-4.

Hernández, R. (2006). *Metodología de la Investigación.* Retrieved from http://sistemas.unicesar.edu.co/documentossistemas/sampieri.pdf.

Joyanes Aguilar, L. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfaomega Grupo Editor.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning.* (Vol. 1): STHDA.

Kerlinger, F., & Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales.* México: McGraw-Hill.

Laura Ochoa, L., Rosas Paredes, K., & Esquicha Tejada, J. (2017). Global Partnerships for Development and Engineering Education: Proceedings of the 15th LACCEI International Multi-Conference for Engineering, Education and Technology, July 19-21, 2017, Boca Raton, FL, United States. (pág. 115). Latin American and Caribbean Consortium of Engineering Institutions. doi:http://dx.doi.org/10.18687/LACCEI2017.1.1.115

Leiva-Valdebenito, S., & Torres-Avilés, F. (2010). Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo. *Revista Colombiana de Estadística, 33(2)*, 321-339.

Marmol Sinclair, P., & Ojeda García, C. (2016). *Marketing turístico 2.* Ediciones Paraninfo, SA.

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

Moya Garcia, R. (2016). *MACHINE LEARNING(en python), con ejemplos.*

Muñoz, M. (2015). Conceptualización Del Neuromarketing: Su Relación Con El Mix De Marketing Y El Comportamiento Del Consumidor (Conceptualisation of Neuromarketing: Its Relationship with the Mix of Marketing and Consumer Behavior). *RAN-Revista Academia & Negocios, 1*(2).

Murillo, D., & Saavedra, D. (2017). Implementación de algoritmo en el Lenguaje R para extraer los datos de los Perfiles en Google Scholar utilizando la técnica web Scraping de Minería de datos. *Congreso Compdes*.

Ñaupas Caraza, C. M. (2016). Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias. Lima, Perú: Tesina.

Nuñez Cardenas, F. J., Hernandez Reyes, J. F., Felipe Redondo, A. M., & Tomas Mariano, V. T. (2019). Aplicación del algoritmo k-means como técnica de minería de datos para determinar el nivel de autoestima en los alumnos universitarios mediante la escala de rosenberg. *Ciencia Huasteca Boletín Científico de la Escuela Superior de Huejutla, 7*(14), 9-17. doi:https://doi.org/10.29057/esh.v7i14.4428

Peña Cuellar, R., Ortiz Sandoval, J. D., & Espitia Cuchango, H. E. (2015). Análisis del efecto-día en el mercado accionario colombiano empleando mapas autoorganizados. *ITECKNE: Innovación e Investigación en Ingeniería, 12*(1), 84-94.

Prieto, N., Casanova, A., Marqués, F., Llorens, M., Galiano, I., Gómez, J. A., . . . Piris, J. (2016). *Empezar a programar usando JAVA.* Valencia: Editorial UPV.

Qadadeh, W., & Abdallah, S. (2018). Customers Segmentation in the Insurance Company (TIC) Dataset. *Procedia computer science, 144*, 277-290. doi:https://doi.org/10.1016/j.procs.2018.10.529

Ramos, J., Alturas, B., & Moro, S. (2017). Business intelligence num organismo público-avaliação de um data mart financeiro. In 12th Iberian Conference on Information Systems and Technologies. *CISTI 2017*, 2274-2279. doi:https://doi.org/10.23919/CISTI.2017.7975672

Ramos, S. (2016). *Data Warehouse, data marts y modelos dimensionales. Un pilar fundamental para la toma de decisiones.* Albatera: SolidQ.

Reyes Vargas, H. R. (2018). Segmentación de clientes y posicionamiento de la marca de materiales de construcción en el distrito de Victor Larco,2017. *Tesis de Pregrado*. Trujillo, Perú: Universidad Nacional de Trujillo.

Rodríguez León, C., & García Lorenzo, M. M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. *Revista Universidad y Sociedad, 8*(4), 43-53.

Rojas Gutiérrez, E. A., & Sebastián Aguilar, J. (2017). Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en bogotá, Colombia. *Trabajo de Investigación*. Bogotá, Colombia.

Ruiz Chavez, Z. (2019). Técnicas de aprendizaje automático aplicadas al procesamiento de Información demográfica. *Tesis Doctororal*. Quito, Ecuador.

Ruiz Larrocha, E. (2017). *Nuevas tendencias en los sistemas de información.* Editorial Centro de Estudios Ramon Areces SA.

Srivastava, R. (2016). Identification of customer clusters using rfm model: A case of diverse purchaser classification. *International Journal of Business Analytics and Intelligence 4.2*, 45-50.

Tang Tong, M. M. (2015). La inteligencia de mercado en las empresas exportadoras e importadoras peruanas. *Revista Oidles, 18*, 71-97.

Vicente Cestero, E., & Mateos Caballero, A. (2018). *Data science y redes complejas: Métodos y aplicaciones.* Editorial Centro de Estudios Ramon Areces SA.

Westwood, J. (2016). *Preparar un plan de marketing.* Profit Editorial.

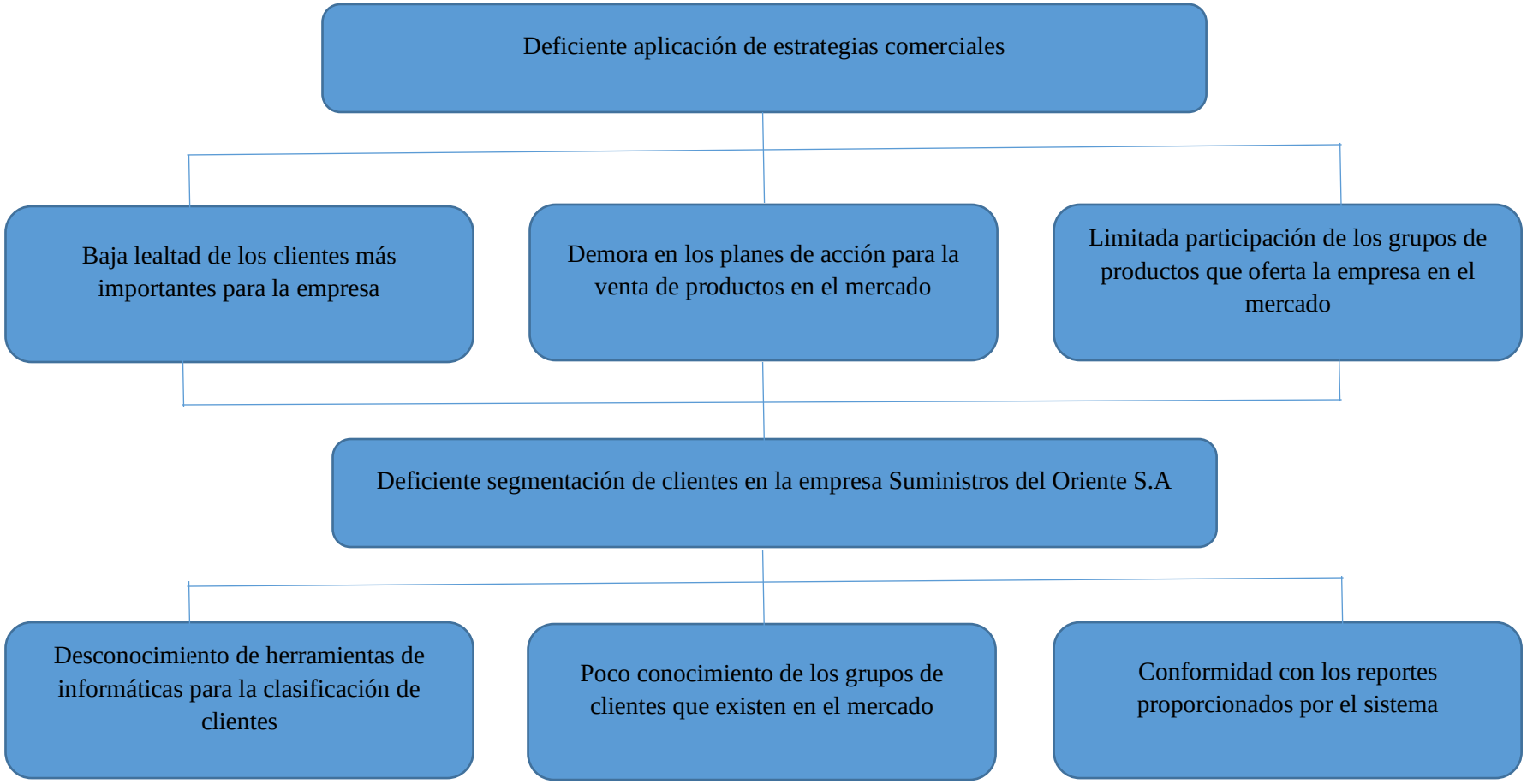
ANEXOS

Anexo 1. Matriz de Consistencia

PROBLEMA	OBJETIVO	HIPÓTESIS	DISEÑO DEL ESTUDIO	POBLACIÓN Y MUESTRA	VARIABLES	DIMENSIONES	INDICADORES
¿Cómo segmentar a los clientes de Suministros del oriente SA?	**Objetivo general** Mejorar el proceso de segmentación de clientes aplicando técnicas de minería de datos para la empresa distribuidora Suministros del Oriente SA. **Objetivos específicos** **OE1:** Identificar las características de los hábitos de compra de los clientes **OE2:** Proponer un modelo basado en KDD y técnicas de minería de datos. **OE3:** Recomendar estrategias comerciales de segmentación de clientes	**Hi:** El modelo basado en técnicas de minería de datos segmentará a los clientes de la empresa distribuidora Suministros del oriente SA	El diseño de investigación que se empleará es el diseño DESCRIPTIVO.	Siendo que el diseño de la investigación es de tipo descriptivo propositivo y que el objeto de estudio de la presente es "segmentación de clientes", el mismo no cuenta con un universo y muestra, ya que los instrumentos diseñados para la recolección de datos para la realización de la prueba de hipótesis basado en el juicio de expertos, será aplicado a los expertos que evaluarán en qué medida la misma mejorará la propuesta.	Técnicas de minería de datos	Clustering o Agrupamiento	Número de clústeres
					Segmentación de clientes	Clientes	Número de Clientes
						Segmentos	Número de Segmentos
						Características Comerciales	Número de Características Comerciales

Fuente: Elaboración propia

Anexo 2. Árbol de Causa - Efecto



Anexo 3. Comparación entre las técnicas de segmentación

1. Utilizaremos dos técnicas de segmentación k-medias(k-means) y k-medoides (clara)
2. Encontramos el número óptimo de clústeres (utilizamos el método del codo y el método de la brecha).

```
#Determinamos el número de clúster óptimos por el método de la brecha
#y el método del codo

library(cluster)
library(factoextra)
library(ggpubr)

#Encontramos el número óptimo de clúster con el método de la Brecha
set.seed(2020)

km_codo<-fviz_nbclust(x = sorsarfm, kmeans, dist(x= sorsarfm, method = "euclidean"),
                      method = c("wss"))+ggtitle("Método del Codo - Kmeans")
km_gap<-fviz_nbclust(x = sorsarfm, kmeans, dist(x= sorsarfm, method = "euclidean"),
                     method = c("gap_stat"))+ggtitle("Estadística de Brecha - Kmeans")

clara_codo<-fviz_nbclust(x = sorsarfm, clara, dist(x= sorsarfm, method = "euclidean"),
                         method = c("wss"))+ggtitle("Método del Codo - Clara")
clara_gap<-fviz_nbclust(x = sorsarfm, clara, dist(x= sorsarfm, method = "euclidean"),
                        method = c("gap_stat"))+ggtitle("Estadística de Brecha - Clara")

ggarrange(km_codo, km_gap, clara_codo, clara_gap)
```

Figura 39. Código en R –Número óptimo de clústeres. (Elaboración Propia).

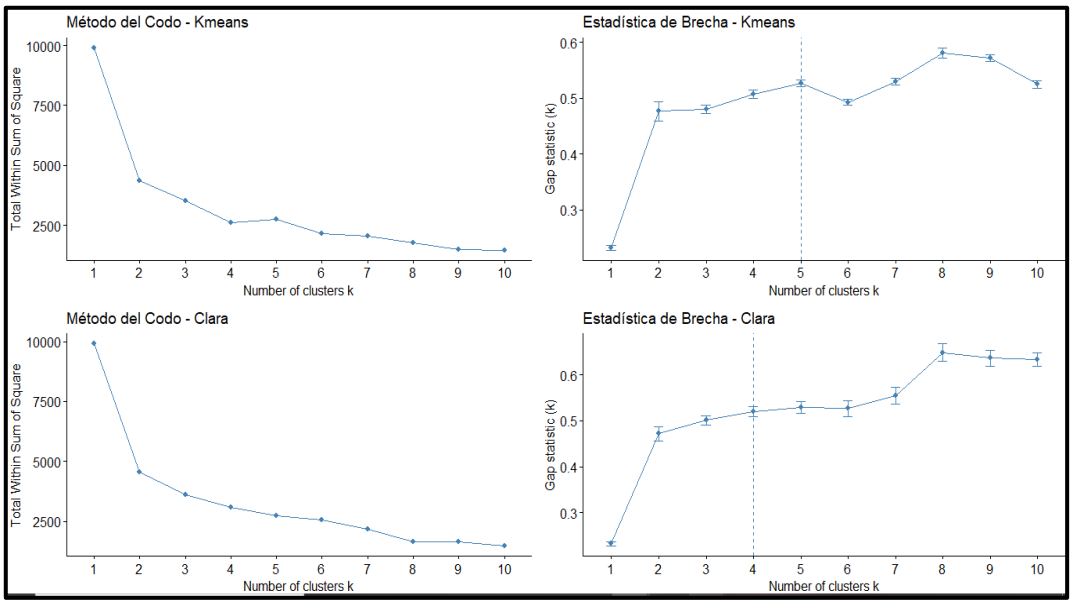


Figura 40. Gráfico en R - Método del codo vs Estadística de la brecha

Del gráfico anterior, definimos que en el método del codo para K-means el número óptimo de segmentos es > 2, coincidiendo con la estadística de brecha que precisa 5 como mejor medida. Para Clara al igual que K-means en el método del codo el número óptimo es >2, y coincide con 4 en la estadística de la brecha.

3. Aplicamos el coeficiente de la silueta para determinar la técnica de segmentación y el número de clústeres más óptimo tal como se muestra en la Figura 41.

```
#Utilizamos el coeficiente de Silueta para comparar las ténicas
library(NbClust)
set.seed(2020)
km_clusters <- eclust(x = sorsarfm, FUNcluster = "kmeans", k = 5,
                      hc_metric = "euclidean", graph = FALSE)
v1<-fviz_silhouette(sil.obj = km_clusters, print.summary = TRUE, palette = "jco",
                    ggtheme = theme_classic())
clara_clusters <- eclust(x = sorsarfm, FUNcluster = "clara", k = 4,
                         hc_metric = "euclidean", graph = FALSE)
v2<-fviz_silhouette(sil.obj = clara_clusters, print.summary = TRUE, palette = "jco",
                    ggtheme = theme_classic())

ggarrange(v1, v2)
```

```
> km_clusters <- eclust(x = sorsarfm, FUNcluster = "kmeans", k = 5,
+                       hc_metric = "euclidean", graph = FALSE)
> v1<-fviz_silhouette(sil.obj = km_clusters, print.summary = TRUE, palette = "jco",
+                     ggtheme = theme_classic())
  cluster size ave.sil.width
1       1  329          0.26
2       2  247          0.38
3       3  784          0.53
4       4  524          0.19
5       5  641          0.49
> clara_clusters <- eclust(x = sorsarfm, FUNcluster = "clara", k = 4,
+                          hc_metric = "euclidean", graph = FALSE)
> v2<-fviz_silhouette(sil.obj = clara_clusters, print.summary = TRUE, palette = "jco",
+                     ggtheme = theme_classic())
  cluster size ave.sil.width
1       1  600          0.65
2       2  475          0.58
3       3  767          0.09
4       4  683          0.05
```

Figura 41. Resultados en R (Elaboración Propia).

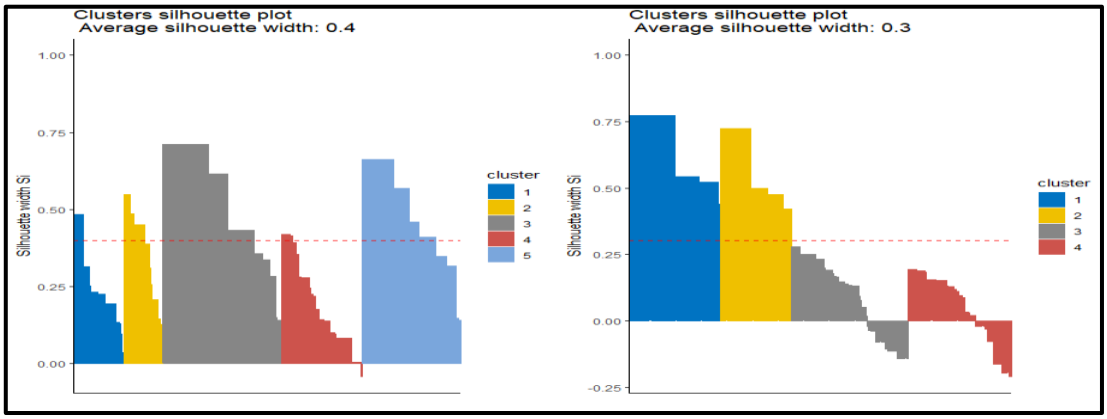


Figura 42. Gráfico en R – K-means vs Clara

Resultado: Del gráfico anterior, se define que la técnica de segmentación k-medias(k-means) con 5 clústeres, permite tiene un promedio de silueta de 0.4; es mejor que k medoides (clara) de 4 clústeres, con un promedio de silueta de 0.3.