**Universidad Nacional Mayor de San Marcos**
**Universidad del Perú. Decana de América**
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

# Clustering de clientes de un grupo de e-Marketplaces del Perú

**TRABAJO DE SUFICIENCIA PROFESIONAL**

Para optar el Título Profesional de Licenciada en Estadística

**AUTOR**

Majorie Denisse BILLADONI VILLAVICENCIO

**ASESOR**

Mg. Ricardo Luis POMALAYA VERÁSTEGUI

Lima, Perú

2021

**Referencia bibliográfica**

Billadoni, M. (2021). *Clustering de clientes de un grupo de e-Marketplaces del Perú*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

## Metadatos complementarios

| | |
|---|---|
| **Datos de autor** | |
| Nombres y apellidos | Majorie Denisse Billadoni Villavicencio |
| Tipo de documento de identidad | DNI |
| Número de documento de identidad | 76374318 |
| URL de ORCID | No aplica |
| **Datos de asesor** | |
| Nombres y apellidos | Ricardo Pomalaya Verastegui |
| Tipo de documento de identidad | DNI |
| Número de documento de identidad | 10460674 |
| URL de ORCID | https://orcid.org/0000-0002-3021-6895 |
| **Datos del jurado** | |
| **Presidente del jurado** | |
| Nombres y apellidos | Roger Pedro Norabuena Figueroa |
| Tipo de documento | DNI |
| Número de documento de identidad | 41493243 |
| **Miembro del jurado 1** | |
| Nombres y apellidos | Oscar Antonio Robles Villanueva |
| Tipo de documento | DNI |
| Número de documento de identidad | 32762171 |
| **Datos de investigación** | |
| Línea de investigación | A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad |
| Grupo de investigación | No aplica |
| Agencia de financiamiento | Sin financiamiento |
| Ubicación geográfica de la investigación | Universidad Nacional Mayor de San Marcos<br>País: Perú<br>Departamento: Lima<br>Provincia: Lima<br>Distrito: Lima<br>Latitud: -12.0560257<br>Longitud: -77.0844226 |

| | |
|---|---|
| Año o rango de años en que se realizó la investigación | Julio 2021 – setiembre 2021 |
| URL de disciplinas OCDE | Estadísticas, Probabilidad<br>https://purl.org/pe-repo/ocde/ford#1.01.03 |

# ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL EN LA MODALIDAD VIRTUAL PARA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA

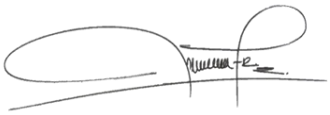
En Lima, siendo las 19:00 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional: Dr. Roger Pedro Norabuena Figueroa (PRESIDENTE), Dr. Oscar Antonio Robles Villanueva (MIEMBRO) y el Mg. Ricardo Pomalaya Verastegui (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**CLUSTERING DE CLIENTES DE UN GRUPO DE E-MARKETPLACES DEL PERÚ**", presentado por la señorita **Bachiller Majorie Denisse Billadoni Villavicencio**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del trabajo de suficiencia, el Presidente invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, la expositora mereció la aprobación de **SOBRESALIENTE,** con un calificativo promedio de **DIECISIETE (17)**.

A continuación, los miembros del Jurado dan manifiesto que la participante **Bachiller Majorie Denisse Billadoni Villavicencio** en vista de haber aprobado la sustentación del Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 19:30 horas se levantó la sesión firmando para constancia la presente Acta.



Dr. Roger Pedro Norabuena Figueroa
PRESIDENTE

Dr. Oscar Antonio Robles Villanueva
MIEMBRO

Mg. Ricardo Pomalaya Verastegui
MIEMBRO ASESOR

La Vicedecana de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, el titulando, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

# RESUMEN

El crecimiento exponencial de los e-maketplaces en el Perú junto a la relevancia de los datos transaccionales que se generan con cada compra, hace que cada vez más profesionales de las áreas comerciales y de marketing estén interesados y se encuentren en la necesidad de descubrir los patrones ocultos tras los comportamientos de compra de los clientes, los cuales son utilizados posteriormente para fijar estrategias que sigan beneficiando al crecimiento de este sector.

El presente trabajo presenta, en esencia, la aplicación del algoritmo de clusterización bietápica sobre un conjunto de datos que reúne las características de compra de clientes de cuatro e-marketplaces del Perú que realizaron transacciones durante los meses de mayo, junio y julio de 2021. El dataset utilizado consistió de un total de 35,881 clientes y se consideró una variable cuantitativa: Recencia, y dos cualitativas: NombreDePago y Marketplace.

Tras la aplicación se encontró una agrupación aceptable de la que resultaron dos clústeres: el primero conformado por el 31% de clientes y el segundo, por el 69% restante. Se describió el perfil del cliente tipo de cada uno de los clústeres y se propusieron estrategias comerciales y de marketing que serán de gran valor para impulsar el crecimiento de las ventas.

Palabras claves: Clúster bietápico, e-commerce, marketplace.

# ABSTRACT

The exponential growth of e-maketplaces in Peru along with the relevance of transactional data generated with purchases, makes professionals in commercial and marketing areas want to get involved and feel in need to discover hidden patterns behind customers purchasing behaviors, which are later used to set strategies that continue to benefit the growth of this sector.

This work presents the application of two-step clustering algorithm on a set of customers' purchasing behaviors of four e-marketplaces in Peru from May to July 2021. The final dataset consisted of 35,881 clients, a quantitative variable: Recency, and two qualitative variables: PaymentMethod and Marketplace.

After the application, a valid grouping of two clusters was found: the first gathered 31% of customers and the second, 69%. The average client's profile was described for each cluster and commercial and marketing strategies were proposed in order to increase sales.



Key words: Two-step cluster, e-commerce, marketplace.

# ÍNDICE

# ÍNDICE DE TABLAS



# ÍNDICE DE FIGURAS

# I. Introducción

Durante el año 2020, el comercio electrónico o e-commerce se ubicó en el primer puesto de entre los sectores económicos con mayor crecimiento en el Perú, obteniendo una mejora del 50% versus el año 2019 y moviendo US$ 6,000 millones al cierre del año. Además, de manera particular, los e-marketplaces percibieron un crecimiento de 295%. (CAPECE, 2021)

Con el crecimiento de tráfico, sesiones y ventas dentro de los e-marketplaces, también aumentó la información disponible sobre los clientes que apuestan por este tipo de comercio, lo que da cabida a profesionales de distintas áreas a explorar más sobre los posibles patrones escondidos que hay detrás del comportamiento de compra de este gran grupo de usuarios. Los resultados obtenidos de dichas indagaciones son de gran valor para los equipos comerciales y de marketing, ya que, de esta manera, las diferentes estrategias de publicidad, como ads en redes sociales, mails y otros, pueden ser personalizadas en función el perfil del cliente, lo que finalmente se reflejará en una mayor tasa de conversión.

Es así que surge la necesidad de un grupo de e-marketplaces, de investigar más a fondo las preferencias y conducta de compra de su público objetivo, de tal forma que sea capaz de implementar una nueva estrategia comercial y de marketing respaldándose en datos. Para ello se utilizó el método de clusterización bietápico, el cual se encarga revelar agrupaciones naturales dentro de un conjunto de datos y que, además, presenta grandes ventajas sobre el resto de algoritmos de clusterización como: la posibilidad de trabajar con variables categóricas y continuas, seleccionar el número de grupos automáticamente y la aplicabilidad sobre grandes bases de datos. (IBM Corporation, 2017)

El informe está dividido en seis capítulos detallados a continuación.

El primer capítulo plasma en términos generales el contexto de la problemática sobre la actividad realizada, además de introducir la metodología estadística a utilizar para el procedimiento.

El segundo capítulo brinda una somera descripción de la institución donde se desarrolló la implementación de la actividad. Se incluyen los objetivos y el periodo de ejercicio de labores dentro de la empresa.

El tercer capítulo se enfoca directamente en la actividad realizada dentro de la institución. Aquí se pueden observar a detalle los objetivos, problemática, finalidad, metodología, resultados y el proceso seguido para cumplir con los objetivos planteados.

El cuarto capítulo se refiere a las conclusiones obtenidas como resultado de los procedimientos listados en el tercer capítulo.

El quinto capítulo plasma algunas recomendaciones.

El sexto capítulo lista la bibliografía utilizada a lo largo del informe.

## II. Información del lugar donde se desarrolló la actividad

- Institución donde se desarrolló la actividad. Corporación que tiene bajo su supervisión a cuatro de los principales e-marketplaces del Perú.
- Periodo de duración de la actividad. Del 04 de enero de 2021 al 31 de julio de 2021.
- Finalidad y objetivos de la entidad. La entidad tiene como finalidad acercar pequeños negocios y compradores en plataformas donde puedan encontrar un gran surtido de productos con el mayor beneficio posible; además, su objetivo principal se centra en posicionarse de manera conjunta como el grupo de e-marketplaces N°1 en el Perú.
- Razón social. No aplica por términos de confidencialidad.
- Dirección postal. No aplica por términos de confidencialidad.
- Correo electrónico del profesional a cargo. No aplica por términos de confidencialidad.

## III. Descripción de la actividad

- Organización de la actividad.

Con base en la metodología CRISP-DM, utilizada ampliamente en trabajos de minería de datos, se listan los pasos seguidos para el cumplimiento de los objetivos planteados:

- Entendimiento del negocio. Como primer paso, fue necesario comprender qué necesitaba la institución, esto recaía en implementar una nueva estrategia comercial y de marketing basada en los datos recopilados sobre las preferencias y/o comportamiento de compra de los clientes.
- Entendimiento de los datos. En segundo lugar, se necesitó trasladar el problema de la realidad hacia un problema de datos, para ello, se necesitó explorar sobre la base general de transacciones de clientes para así obtener un subset del conjunto de datos originales que contenga las variables que consideramos de mayor utilidad para encontrar grupos de clientes con similitudes dentro de ellos y diferencias entre ellos.
- Preparación de los datos. Durante esta etapa, se aplicó el Análisis Exploratorio de Datos para observar el comportamiento de las variables originales; también, fue posible agregar valor al nuevo dataset mediante la ingeniería o creación de nuevas variables. El software utilizado durante esta etapa fue Python a través de la interfaz de Google Colaboratory. Al completar esta etapa, se obtuvo la data final sobre la que se aplicó el modelo.
- Modelamiento. Con el pre procesamiento de los datos listo, se aplicó el algoritmo que mejor se acopló a los objetivos trazados además de adaptarse a las variables seleccionadas. En este caso, la técnica estadística seleccionada fue el análisis de clúster bietápico. El software utilizado durante esta etapa fue IBM SPSS Statistics 21.
- Evaluación. Para evaluar los resultados del modelo, se hizo uso de la medida de silueta de cohesión y separación brindados por el software IBM SPSS Statistics 21. Esta es aceptable cuando es mayor o igual a 0 y da un mejor resultado cuando es superior a 0.2, pues es indicador de la existencia de una separación significativa entre los grupos. (Dietrich y Rundle-Thiele, 2017)

- Finalidad y objetivos de la actividad.

Finalidad.

El estudio tiene como finalidad segmentar a los consumidores según sus características de compra (acotadas por variables), de tal manera, la institución se ve en la posibilidad de ofrecer promociones en base a los perfiles identificados.

Objetivo general.

Obtener clústeres de clientes en base a sus principales características de compra.

Objetivos específicos.

- Identificar el número de clústeres resultantes.
- Describir al cliente tipo de cada clúster en base a sus características de compras.
- Proponer acciones comerciales y de marketing a trabajar sobre cada clúster encontrado.

- Problemática

En algunos casos se ve en una gran base de clientes una gran oportunidad para generar ventas; sin embargo, el envío de promociones y publicidad masiva suele generar el efecto contrario al esperado, pues al no ser estos especializados según sus preferencias puede llegar inclusive a incomodar al cliente. Esto es una gran problemática para los e-marketplaces, ya que, siendo la industria número 1 en crecimiento, se pueden perder muchas oportunidades de conversión si no se tiene cuidado; por lo tanto, el principal problema con el que se cuenta es realizar un análisis que permita conocer qué tipos de clientes tienen los e-marketplaces y cómo se puede tratar a cada uno de ellos en base a sus preferencias.

A continuación, se presenta una serie de trabajos que han utilizado la segmentación de clientes o mercados como base para una mejor toma de decisiones, según sea el contexto.

En su trabajo de investigación, Chirinos y Villalobos (2017) hicieron uso de la segmentación de mercados, mediante el análisis de clúster bietápico aplicado en el software SPSS Statistics v.19, para comprender al target del sector de ventas de accesorios de moda emergente en Venezuela. En primer lugar, se valieron de un conjunto de datos conformado por características demográficas y psicográficas de 384 clientes potenciales o reales, de donde se obtuvieron tres segmentos: cautelosos-dinámicos, conformado por el 49.2% de la población; explorador-experimentador, conformado por el 25.8% y práctico-equilibrado, conformado por el 25% restante; además, se validó que calidad de los segmentos era aceptable a través de la medida de silueta de cohesión y separación, pues resultó mayor a 0.2. Además, se analizó un segundo conjunto de datos que contenía información sobre la interacción con cuentas de accesorios de moda y las características propias de 1236 usuarios de Instagram, con la que se logró identificar a cuatro segmentos: sociables altamente influyentes, conformado por el 0.3% de la población; sociables muy influyentes, conformado por el 21.1%; sociables activos, conformado por el 38.2%; y sociables de poca actividad, por el 40.3%; además, la calidad de los segmentos fue buena, pues su medida de silueta de cohesión y separación resultó mayor a 0.5.

Tang y Vargas (2016), en su trabajo de titulación, aplicaron el análisis de conglomerados en dos etapas con el fin de segmentar a los clientes de una tienda de electrodomésticos y proponer nuevas estrategias específicas de marketing para cada uno de los grupos encontrados; dicho proceso se realizó haciendo uso del software IBM SPSS Statistics v.22 y con una base de datos que contenía información acerca de nueve características, medidas por variables cualitativas y cuantitativas, de 6284 clientes que adquirieron un electrodoméstico de la tienda gracias al otorgamiento de un préstamo crediticio durante el primer trimestre del año 2013. Se lograron identificar tres segmentos de clientes a los que se le asignó un perfil y estrategia de marketing según sus características, estos fueron: los cazaofertas, conformado por el 36.5% del total de clientes, los estacionales, conformado por el 27.9% del total y los estacionales/cautos conformado por el 35.6% del total.

En su trabajo de investigación, Doğan et al. (2018) realizaron una segmentación de clientes de una de las cadenas de retail de deportes más grandes de Turquía basándose en el modelo RFM. Dicha investigación tuvo como finalidad encontrar nuevos clústeres que ayudaron a redefinir el sistema de fidelización por tarjetas con la que se contaba, para esto hicieron uso de un set de datos conformado por los indicadores de Recencia, Frecuencia y Monetario (RFM) de 700032 clientes que realizaron compras de manera presencial o en línea durante el 1 de enero al 31 de diciembre del 2016. Se aplicaron dos metodologías; en primer lugar, la clusterización bietápica, con la que se obtuvieron a los grupos de clientes de Bronce, con indicadores RFM por debajo del promedio; Oro, con un indicador R por encima del promedio e indicadores F y M por debajo y Premium, con RFM por encima del promedio; y en segundo lugar, la clusterización por k-means, de la que resultaron los grupos de clientes Regulares, que englobaron al 92% de los clientes y con indicadores RFM por debajo del promedio; Leales, con indicadores RFM por encima del promedio; Estrella, con indicadores RFM muy por encima del promedio y que representan a menos del 0.015% de los clientes; y Avanzados, con indicadores RFM por encima del promedio pero menores que los del grupo de Leales.

Tavakoli et al. (2018) desarrollaron un artículo de investigación donde se propuso segmentar a clientes utilizando el método de K-Means basado en un modelo R + FM que, en comparación al RFM tradicional, toma en consideración los cambios del negocio, haciéndolo más eficaz. El procedimiento se aplicó en Digikala, el e-commerce más grande de medio oriente, se lograron obtener cuatro segmentos: Activos con valor alto, Medianamente activos con valor monetario alto, Medianamente activos con frecuencia alta y Valor bajo de actividad, sobre los que se construyeron y aplicaron estrategias de marketing para cada segmento. Los resultados de las campañas diferenciadas mostraron que el nuevo modelo de segmentación generó un mayor impacto en los clientes, por lo tanto, se consiguió una mayor efectividad.

Wu et al. (2020) desarrollaron un paper en el que se explican los resultados obtenidos al realizar una segmentación de clientes en una empresa dedicada al comercio electrónico en Beijing, China. Durante la investigación se analizó la data transaccional de la empresa desde noviembre del 2017 hasta abril del 2019, sobre la que se aplicó el algoritmo de k-means fusionado con el modelo RFM. Como resultado, se obtuvieron cuatro clústeres segmentados por sus hábitos de compra, para los que se trabajó una distinta estrategia de CRM con el fin de obtener un mejor nivel de satisfacción por parte de los clientes. Finalmente, se comprobó la efectividad del método aplicado ya que los KPIs de la empresa mejoraron: el número de clientes activos se incrementó en 519, el volumen de compra se incrementó en un 279% y el total de consumo se incrementó en 102%.

- Metodología

Tipo de investigación y diseño

Según Carrasco (2013), la presente se trata de una investigación de tipo aplicada puesto que se busca resolver un problema de la realidad y de diseño no experimental transversal descriptivo debido a que los datos utilizados se acotaron sobre un periodo de tiempo específico y su finalidad fue encontrar segmentos de clientes en base a sus diversas características de compra.

Población y muestra

La población estuvo conformada por los clientes que realizaron por lo menos una compra a través de los e-marketplaces en estudio. Además, debido a que la actividad se centra en descubrir patrones escondidos en base a una cantidad de datos considerablemente grande, se tomó como muestra a 35,881 clientes que efectuaron una o más compras durante los meses de marzo a julio del 2021 en los e-marketplaces seleccionados.

Identificación de variables

El dataset utilizado contiene información transaccional de clientes de cuatro e-marketplaces peruanos que realizaron compras a través de las respectivas plataformas durante los meses de mayo a junio de 2021. En él se presentan sus características de compra definidas en base a tres variables cuantitativas y tres cualitativas.

**Tabla 1**

*Variables utilizadas durante la actividad*

| Tipo de variable | Variable | Descripción |
|---|---|---|
| Continua/ Cuantitativa | Recencia | Número de días transcurridos desde la última compra |
| | Frecuencia | Número de veces que el cliente realizó una compra |
| | Monetario | Desembolso total del cliente en soles |
| Categórica/ Cualitativa | Marketplace | E-Marketplace por el que se realizaron las compras con mayor frecuencia |
| | Zona | Zona más frecuente hacia la que se hizo el envío de pedidos |
| | Medio de pago | Vía más frecuente por la que se realizaron las compras |

- Procedimientos

A continuación, se describen algunos conceptos clave relacionados al estudio.

Análisis clúster

Según Aldás y Uriel (2017), el análisis clúster o de conglomerados es una técnica estadística que busca clasificar un grupo de *n* observaciones y *k* variables en *g* nuevos grupos de observaciones, desconocidos a priori.

Estos nuevos grupos deben cumplir con ser:

- Homogéneos dentro de sí, las observaciones dentro de estos deben ser lo más parecidas posibles según las variables en análisis.
- Heterogéneos entre ellos, cada grupo debe ser lo más distinto posible del otro según las variables en análisis.

Además, la técnica de clusterización es considerada como parte de las técnicas de aprendizaje no supervisado debido a que su objetivo es descubrir patrones ocultos dentro de los datos que le permitan generar nuevas agrupaciones, mas no encontrar una variable respuesta; por este motivo, las variables incluidas en el análisis tienen la misma importancia y los resultados obtenidos no pueden ser considerados precisamente como correctos o incorrectos, pues todo depende del enfoque y objetivos que el analista haya planteado (IBM Corporation, 2020).

Análisis de clúster bietápico

El análisis clúster bietápico o en dos fases es uno de los métodos de análisis de conglomerados, este se diferencia del resto por desarrollarse en dos etapas además de contar con bondades que serán comentadas a lo largo de este apartado.

- Medida de similaridad. Para poder agrupar las observaciones es necesario contar con un indicador que cuantifique qué tanto se parecen cada par de observaciones entre sí (Aldás y Uriel, 2017), en el análisis de clúster bietápico se hace uso de la distancia de log-verosimilitud y se explicará a continuación según IBM (2013).

  La distancia de log-verosimilitud está basada en probabilidades y puede ser utilizada con variables continuas y categóricas. Además, en este caso particular, la distancia entre dos conglomerados, *i* y *j*, está relacionada con la disminución del logaritmo de verosimilitud a medida que se combinan en un solo clúster y está representada de la siguiente manera:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>}$$

  Donde

$$\xi_v = -N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}^2{}_k + \hat{\sigma}^2{}_{vk}) + \sum_{k=1}^{K^B}\hat{E}_{vk}\right)$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}\,log\,\frac{N_{vkl}}{N_v}$$

  Además

  $K^A$: Número total de variables continuas usadas en el procedimiento.

  $K^B$: Número total de variables categóricas usadas en el procedimiento.

  $L_k$: Número de categorías de la *k-ésima* variable categórica.

  $N_v$: Número de observaciones o registros en el clúster *v*.

  $N_{vkl}$: Número de observaciones o registros en el clúster *v* para los que la *k-ésima* variable categórica toma la *l-ésima* categoría.

$\hat{\sigma}^2{}_k$: Varianza estimada de la *k-ésima* variable continua del dataset general.

$\hat{\sigma}^2{}_{vk}$: Varianza estimada de la *k-ésima* variable continua en el clúster *v*.

Nota: Si se ignora $\hat{\sigma}^2{}_k$ en la ecuación $\xi_v$, la distancia entre los clústeres *i* y *j* representaría exactamente el decrecimiento en log-verosimilitud cuando dos clústeres se combinan. El término $\hat{\sigma}^2{}_k$ es agregado para resolver el problema de obtener un logaritmo natural indefinido causado por $\hat{\sigma}^2{}_{vk} = 0$; este caso puede ocurrir, por ejemplo, cuando un clúster está conformado por un solo caso.

Supuestos. Teóricamente, el uso de la distancia log-verosimilitud supone que:

- Todas las variables consideradas en el modelo son independientes.
- Las variables continuas siguen una distribución normal.
- Las variables categóricas siguen una distribución multinomial.

Sin embargo, de manera empírica se ha comprobado que el procedimiento es robusto frente a violaciones de dichos supuestos (IBM Corporation, 2017).

- Algoritmo de agrupación. Después de determinar la proximidad de las observaciones gracias al cálculo de las distancias, es necesario formar los grupos en base a un algoritmo de agrupación, usualmente se elige entre los enfoques jerárquico y no jerárquico.
  El análisis clúster bietápico, tal como su nombre lo indica, se desarrolla en dos etapas y hace uso de un algoritmo distinto en cada una de ellas. Este procedimiento está fundamentando en el algoritmo de clusterización BIRCH, propuesto por Zhang et al. (1996), el cual se basa en una estructura de datos jerárquica llamada CFT (Clustering Feature Tree) y se ejecuta en dos fases.

  Algoritmo de clusterización BIRCH. Según Jiawei et al. (2012), este algoritmo usa las nociones de CF y CFT (características de agrupamiento y árbol de características de agrupamiento, respectivamente) para resumir un clúster y representarlo jerárquicamente, ambas estructuras ayudan al algoritmo a alcanzar rapidez de procesamiento y escalabilidad sobre grandes conjuntos de datos e inclusive de tiempo real.

Para un clúster de *n* casos y *d* dimensiones, el CF representa un vector de tres dimensiones que resume información sobre los casos contenidos en él. Está definido como $CF =< n, LS, SS >$, donde $LS$ es la suma lineal de los *n* casos y $SS$ es la suma de cuadrados de los *n* casos. Además, el CF tiene propiedades aditivas; es decir, para dos clústeres disjuntos, $C_1$ y $C_2$, con $CF_1 =< n_1, LS_1, SS_1 >$ y $CF_2 =< n_2, LS_2, SS_2 >$, el CF para el clúster formado por la unión de $C_1$ y $C_2$ es $CF_1 + CF_2 =< n_1 + n_2, LS_1 + LS_2, SS_1 + SS_2 >$. Adicionalmente, es importante notar que al resumir un clúster mediante el CF evitamos almacenar información detallada sobre los objetos, necesitando solo espacio en la memoria para almacenar los CF, esta es la clave de la eficiente del algoritmo BIRCH.

Por otro lado, el CFT es un árbol balanceado compuesto por un nodo raíz, nodos sin hojas (puede haber varios subniveles) y nodos con hojas, en cada uno de los niveles se almacena un número determinado de CF que está delimitado por los valores de *B*: número máximo de CF por cada nodo sin hojas y *L*: número máximo de CF por cada nodo con hojas, este se construye durante la primera fase, mientras que en la segunda fase se aplica un algoritmo de clusterización sobre los nodos con hojas del CFT. Ambas fases se realizan con una sola corrida de los datos.

Entendidos los conceptos de este algoritmo usado como base para desarrollar el de la clusterización bietápica, se procede a describir el procedimiento de cada una de las etapas de este último según lo expuesto por Dietrich y Rundle-Thiele (2017).

Etapa 1: Pre-clustering. El objetivo de esta etapa es reducir el tamaño de la matriz de distancias y esto se consigue formando agrupaciones de las observaciones originales a través de una estructura modificada del CFT que usa CF formados por el triplete de: número de registros, media de los registros y varianza de los registros; una vez terminado el proceso, el tamaño de la matriz de distancias ya no depende del número de observaciones sino del número de preclústeres obtenidos, el cual es menor en tamaño respecto a la matriz original de datos.

El uso de CF en los CFT hace que se ocupe un espacio reducido de memoria en comparación a la utilización de los datos brutos, esto permite que el algoritmo sea bastante funcional frente a grandes bases de datos.

Etapa 2: Clustering. Esta etapa hace uso del algoritmo de clusterización jerárquico aglomerativo sobre los preclústeres obtenidos en la primera etapa, esto es posible debido a que la cantidad de preclústeres es menor a la cantidad de registros con los que inicialmente se contaban. El uso de un método jerárquico permite explorar un rango de soluciones con diferente número de clústeres, lo que la facilita la autodeterminación del número óptimo de clústeres, para esto se vale del criterio de información bayesiano BIC (IBM, 2013).
El uso de un algoritmo jerárquico permite obtener el número óptimo de clústeres, para esto se vale del BIC o Criterio de Información Bayesiano.

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(N)$$

Donde

$$m_J = J\left\{2K^A + \sum_{k=1}^{K^B}(L_K - 1)\right\}$$

*N*: Número total de observaciones o registros.

Las notaciones de $K^A$, $K^B$ y $L_K$ son las descritas en la sección de medida de similaridad.

Validación. Según Dietrich y Rundle-Thiele (2017), al usar el Criterio de Información Bayesiano, es posible validar los resultados del clúster obtenido mediante la medida de silueta de cohesión y separación, la cual se encarga de medir la relación dentro y entre conglomerados. Una puntuación superior a 0 asegura que la distancia dentro y entre los conglomerados es válida entre las diferentes variables, ya que existen variaciones entre ellas; sin embargo, se obtienen mejores resultados cuando la puntuación está por encima de 0.2.

Clustering de clientes.

Es el proceso de aplicar el análisis y modelo específico de clusterización sobre un conjunto de clientes de cierta empresa con el fin de obtener *g* nuevos grupos en base a características predefinidas.

Los resultados obtenidos suelen usarse como inputs para análisis posteriores, un ejemplo de esto es la creciente aplicación de esta metodología para optimizar las estrategias de mercado. Esto se sustenta en el hecho de que, al identificar las diferencias y similitudes entre los segmentos de clientes, los equipos de marketing pueden diseñar campañas especiales que atraigan a una proporción más amplia de consumidores, por otro lado, también podrían optar por reducir el enfoque a un solo tipo de consumidor y de esta manera mejorar su desempeño en el mercado.

Análisis RFM

Según la documentación de IBM Corporation (2014), el análisis RFM es una técnica que identifica a los clientes con mayor probabilidad de responder ante una nueva oferta, para esto se les asigna puntuaciones a tres indicadores y se agrupan las observaciones con puntajes similares. Los indicadores son:

- R: Recencia. Indica qué tan recientemente se ha hecho la última compra.
  Existe una mayor probabilidad de que los clientes que recientemente han realizado una compra vuelvan a comprar en comparación a clientes que compraron más en el pasado.
- F: Frecuencia. Indica qué tantas veces se realizaron las compras.
  Existe una mayor probabilidad de que los clientes que compraron más veces respondan a la nueva oferta en comparación a los que realizaron menos compras.
- M: Monetario. Indica cuánto fue el gasto total de todas las compras realizadas.
  Existe una mayor probabilidad de que los clientes que gastan más en sus compras respondan a la nueva oferta en comparación a los que han gastado menos.

- Resultados de la actividad

- Entendimiento del problema.
  Las necesidades de la institución se explicaron durante la introducción y planteamiento de objetivos.

- Entendimiento de los datos.

  El paso que prosiguió a la comprensión de las necesidades fue determinar qué variables serían de ayuda para alcanzar los objetivos; para esto, y según la información puesta a disposición, inicialmente se recolectó una total de 76,202 registros transaccionales de los meses mayo, junio y julio de 2021; además, se seleccionaron las variables o características de compra que, por criterio comercial, podrían ser más interesantes de investigar, estas fueron: día en el que se realizó la compra, valor de la compra en soles, nombre del e-marketplace por el que se realizó la compra, distrito hacia donde se realizó el envío de la compra y medio de pago utilizado para efectuar la compra, además del ID de cliente, ID de pedido y Nombre_producto que fueron necesarios para realizar la agrupación en pasos posteriores y verificar completitud de los datos.

- Preparación de los datos.

  El subset de datos y variables obtenido en la etapa anterior fue llevado hacia el entorno de Python, en el que se trabajaron tres subetapas.

  Limpieza de datos. Para poder aplicar cualquier tipo de modelo, siempre es imprescindible contar con completitud de los datos, dicho de otra manera, es necesario imputar o eliminar los registros que cuenten con valores perdidos. Una forma sencilla de identificarlos es obteniendo la cantidad de valores perdidos por variable.

**Tabla 2**

*Número de valores perdidos por variable*

| Variable | Número de valores perdidos |
|---|---|
| ID_cliente | 1 |
| ID_pedido | 0 |
| Nombre_producto | 0 |
| Día_compra | 0 |
| Valor_compra | 0 |
| Nombre_marketplace | 1 |
| Distrito_envío | 10 |
| Medio_pago | 1 |

El primer paso fue eliminar el registro que no contaba con ID_cliente además de otro grupo de registros para los que se cumplía ID_cliente igual a cero, pues este dato no es imputable y es de vital importancia para realizar las agrupaciones, con esto se redujeron a cero los valores perdidos de las variables Nombre_marketplace y Medio_pago, ya que se trataba del mismo registro.

Además, se identificaron los registros que contaban con valores perdidos en el campo Distrito_envío y se observó que la información estaba incompleta por tratarse de compras de prueba; en ocasiones los e-marketplaces suelen realizar tests para verificar que no existan problemas en el journey de compra del cliente y esta se efectúe sin mayor inconveniente. Para solucionar este problema, se eliminaron los registros cuyo Nombre_producto contenga la palabra "prueba".

Al realizar ambas acciones detalladas líneas arriba, se obtuvo un dataset completo sin valores perdidos.

Transformación de variables. Durante esta sección, en primera instancia se buscó transformar las variables que se tenían inicialmente, de forma que representen mejor al dataset obtenido en la etapa de limpieza. Para esto, se crearon tres nuevas variables: Marketplace_pre, MedioDePago_pre y Zona_pre.

La variable Marketplace_pre se obtuvo tras renombrar las categorías de la variable Nombre_marketplace, esta modificación fue necesaria para asegurar la confidencialidad de los datos. Los nuevos nombres son: MP1, MP2, MP3 y MP4.

La variable MedioDePago_pre se obtuvo estandarizando los valores de las categorías iniciales con las que contaba la variable Medio_pago, por ejemplo: Visa y tarjeta_visa fueron renombrados como Tarjeta Visa.

Por último, para obtener la nueva variable Zona_pre se hizo una reagrupación de la variable inicial Distrito_envío siguiendo el esquema mostrado a continuación.

**Tabla 3**

*Esquema de reagrupación de distritos hacia zonas*

| Zona_pre | Distrito_Compra |
|---|---|
| Callao | Callao, Ventanilla, La Punta, La Perla, Carmen de La Legua, Bellavista. |
| Lima Norte | Ancón, Santa Rosa, Puente Piedra, Comas, Carabayllo, Independencia, San Martin de Porres, Los Olivos. |
| Lima Este | San Luis, Ate, Santa Anita, El Agustino, San Juan de Lurigancho, Lurigancho, Cieneguilla, Chaclacayo. |
| Lima Sur | Villa Maria del Triunfo, Villa El Salvador, Santa Maria, San Juan de Miraflores, Chorrillos, Punta Negra, Punta Hermosa, San Bartolo, Pucusana, Pachacamac, Lurín. |
| Lima Céntrica | Lima, Rimac, La Victoria, Surquillo, Surco, Jesús María, Lince, Breña, Pueblo Libre, San Borja, San Miguel, Magdalena, San Isidro, Barranco, Miraflores, La Molina |
| Provincias | Compuesto por el resto de distritos que no sea parte de los listados anteriormente |

Ahora, es importante recordar que en el dataset obtenido hasta este paso, cada registro representa una transacción; sin embargo, el proceso que se desea realizar es una clusterización de clientes, por lo que fue necesario agrupar los registros por ID_cliente. Al realizar este tratamiento, las variables también cambiaron, siendo las definitivas: Recencia, Frecuencia, Monetario, Marketplace, Zona y MedioDePago, además del IDcliente como identificador.

La variable Recencia para cada IDcliente se obtuvo como la diferencia entre la fecha en la que se registró la última compra en cualquiera de los e-marketplaces y la fecha más reciente en la que un cliente determinado realizó una compra.

La variable Frecuencia se obtuvo como la cuenta de ID_pedido para cada IDcliente, esto resume la cantidad de veces que un determinado cliente realizó compras a través de cualquiera de los e-marketplaces.

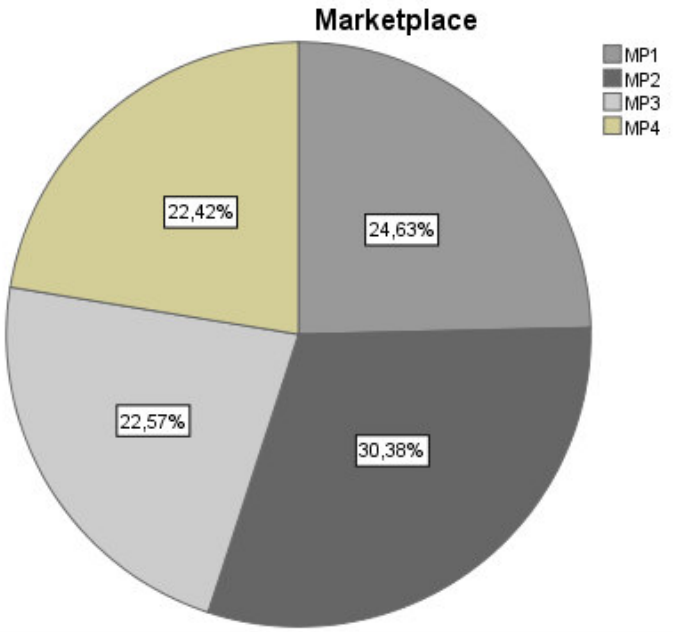
La variable Monetario se obtuvo sumando Valor_compra para cada IDcliente, de tal manera que en cada registro se muestre el total de dinero desembolsado por determinado cliente a lo largo de los meses mayo, junio y julio de 2021.

Por último, las variables Marketplace, Zona y MedioDePago se obtuvieron hallando la moda de Marketplace_pre, Zona_pre y MedioDePago_pre para cada IDcliente respectivamente. Estas representan el e-marketplace por donde se realizaron las compras, zona hacia donde se enviaron las compras y medio de pago utilizado con mayor frecuencia respectivamente.

Análisis Exploratorio de Datos. En esta sección se muestra el comportamiento univariado de las variables en estudio.

**Figura 1**

*Gráfico de sectores de compras por e-marketplace*




De la gráfica se puede notar que la distribución de las compras por Marketplace es bastante homogénea, pues aproximadamente el 25% de ellas se ha dado en cada uno de los e-marketplaces. Según el porcentaje de frecuencias, el Marketplace donde se realizó la mayor parte de las ventas fue MP2 (30.38%), seguido de MP1 (24.63%), MP3 (22.57%) y MP4 (22.42%).

**Tabla 4**

*Compras por zona*

| Zona | Compras | |
|---|---|---|
| | F | % |
| Lima Céntrica | 29,845 | 83.18 |
| Provincias | 2,616 | 7.29 |
| Callao | 1,021 | 2.85 |
| Lima Norte | 959 | 2.67 |
| Lima Este | 816 | 2.27 |
| Lima Sur | 624 | 1.74 |

De la tabla, se observa que la zona hacia la que se enviaron la mayor parte de las compras es Lima Céntrica (83.18%), Provincias ocupa el segundo lugar (7.29%) pero hay que considerar que esta categoría no está realmente desagregada por provincias, eso puede explicar por qué resulta mayor en proporción respecto a las compras enviadas a Callao (2.85%), Lima Norte (2.67%), Lima Este (2.27%) y Lima Sur (1.74%).

**Tabla 5**

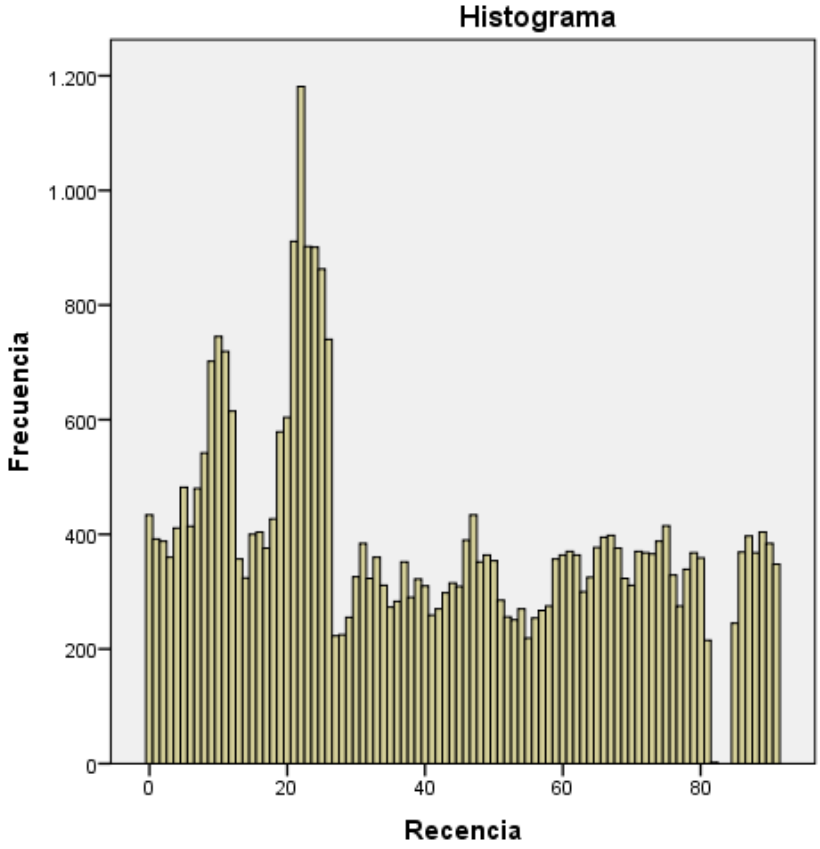
*Compras por medio de pago*

| Medio de pago | Compras | |
|---|---|---|
| | F | % |
| Tarjeta de crédito/débito | 15,408 | 42.94% |
| Visa | 9,825 | 27.38% |
| Tarjeta Oh | 6,684 | 18.63% |
| Mastercard | 1,340 | 3.73% |
| Safetypay | 1,258 | 3.51% |
| American Express | 415 | 1.16% |
| Tarjeta Agora | 325 | 0.91% |
| Diners | 318 | 0.89% |
| Otro | 308 | 0.86% |

El medio de pago más utilizado con el que se realizaron compras fue la tarjeta de débito/crédito (42,94%), seguido de tarjetas Visa (27.38%) y tarjetas Oh (18.63%), estos tres medios de pago representan más del 80% de compras que se realizaron los e-marketplaces.

**Figura 2**

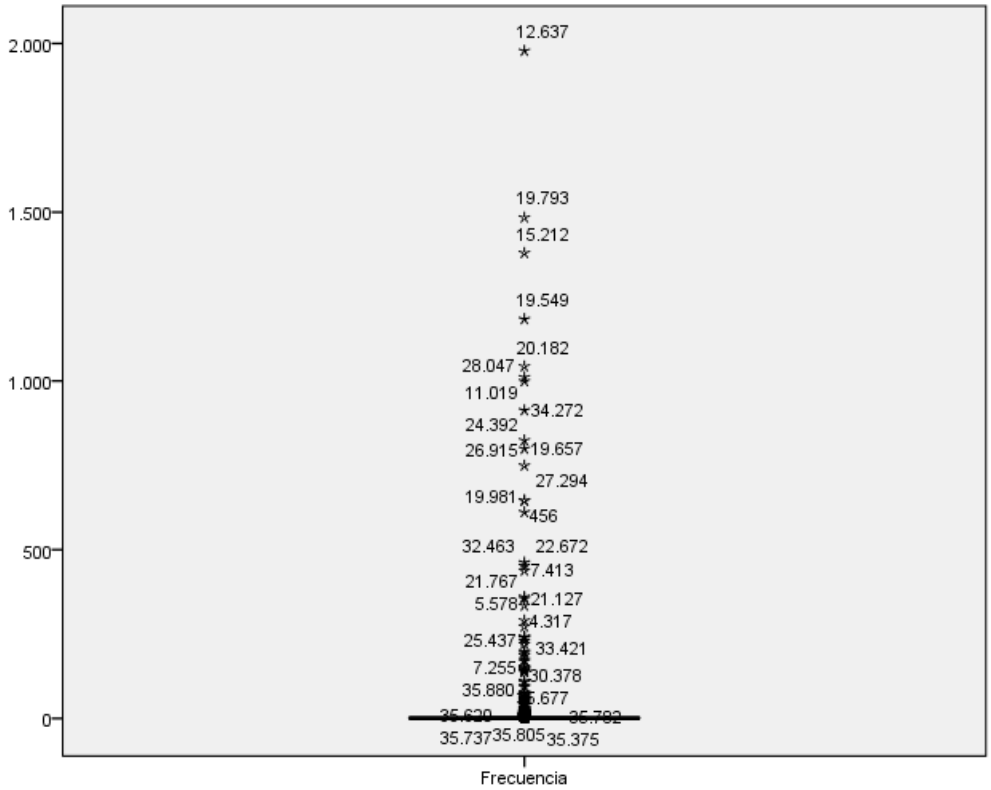
*Histograma de frecuencias de Recencia*




La variable Recencia presenta una distribución asimétrica parecida a la uniforme con una mayor concentración alrededor de los 22 días.

**Figura 3**

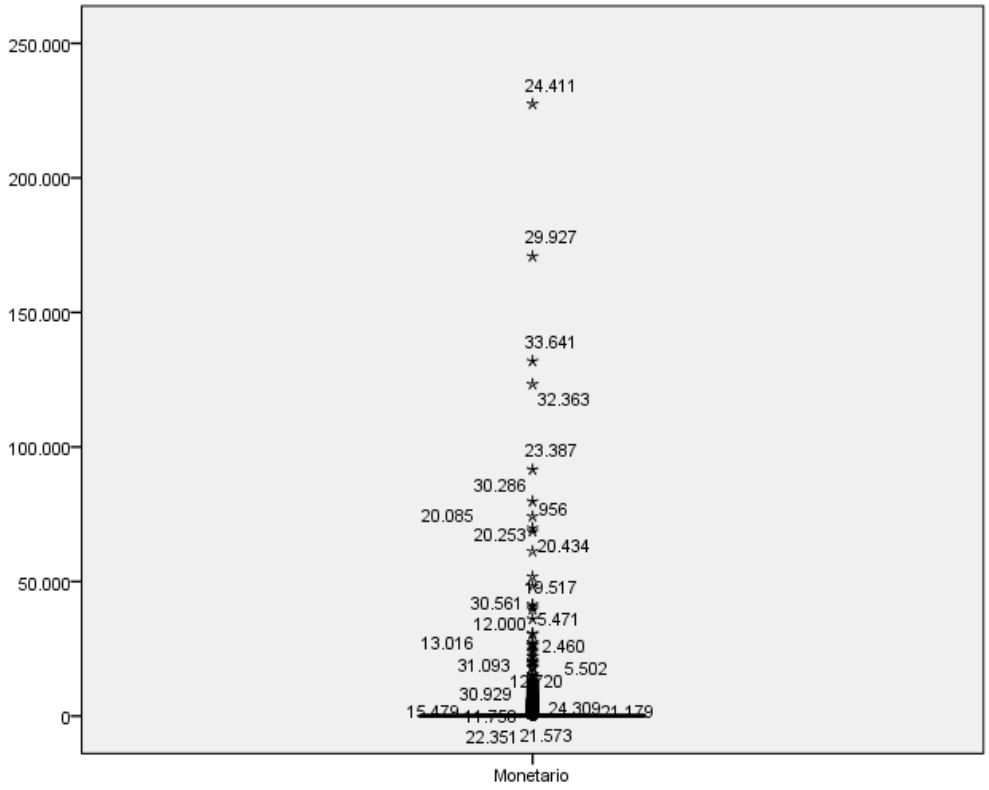
*Diagrama de cajas de la frecuencia de compra*




El diagrama de cajas muestra que existe una gran cantidad de valores atípicos, esto debido a que los e-marketplaces suelen ofrecer precios bastante bajos durante fechas especiales como Cybers, lo que genera un aumento sobre la cantidad de transacciones generadas.

**Figura 4**

*Diagrama de cajas de la variable Monetario*



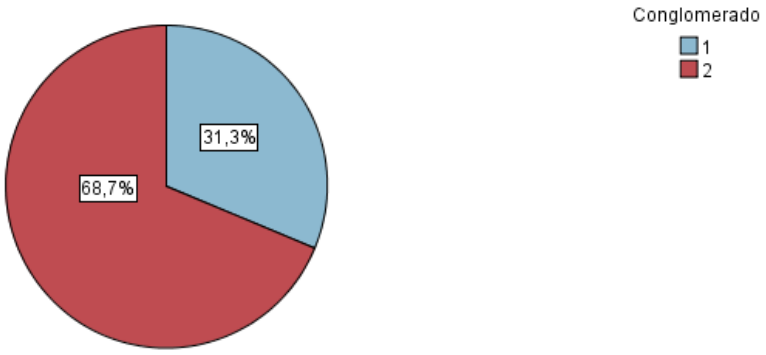

Sobre la variable Monetario se observa un comportamiento similar al descrito en el *Gráfico 3*, no existe normalidad de los datos, esto debido a que la frecuencia de compra está relacionada al gasto total hecho por el cliente.

A pesar de la identificación de valores atípicos e incumplimiento de supuestos durante esta etapa, se decide no alterar la base de datos, ya que la aplicación del método de clusterización bietápico mediante el software IBM SPSS Statistics v.21 realiza una exclusión de valores atípicos de manera automática antes de proporcionar los resultados finales y según lo explicado en el apartado de *Procedimientos*, el algoritmo es robusto frente a violaciones de normalidad e independencia.

- Modelamiento de los datos. Se realizó una primera corrida de los datos y se identificó que las variables Zona, Frecuencia y Monetario no estaban siendo significativas; es decir, que no estaban aportando suficiente información para realizar la clusterización de clientes, por lo que se decidió retirarlas del modelo. Con las tres variables restantes: Recencia, Marketplace y MedioDePago se obtuvieron los siguientes resultados.

**Figura 5**

*Clientes por clúster*

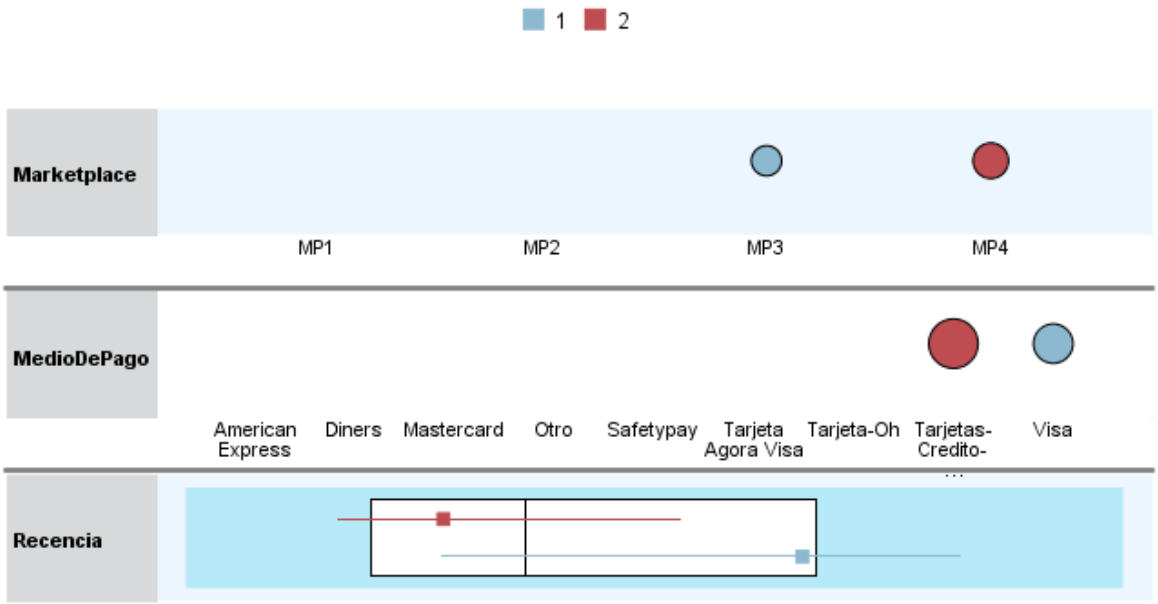



| | |
|---|---|
| **Tamaño de conglomerado más pequeño** | 9802 (31,3%) |
| **Tamaño de conglomerado más grande** | 21554 (68,7%) |
| **Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño** | 2,20 |

Se obtuvieron dos clústeres, el primero estuvo conformado por el 31,3% de los clientes y el segundo por el 68,7% restante.

**Figura 6**

*Comparación de los clústeres*



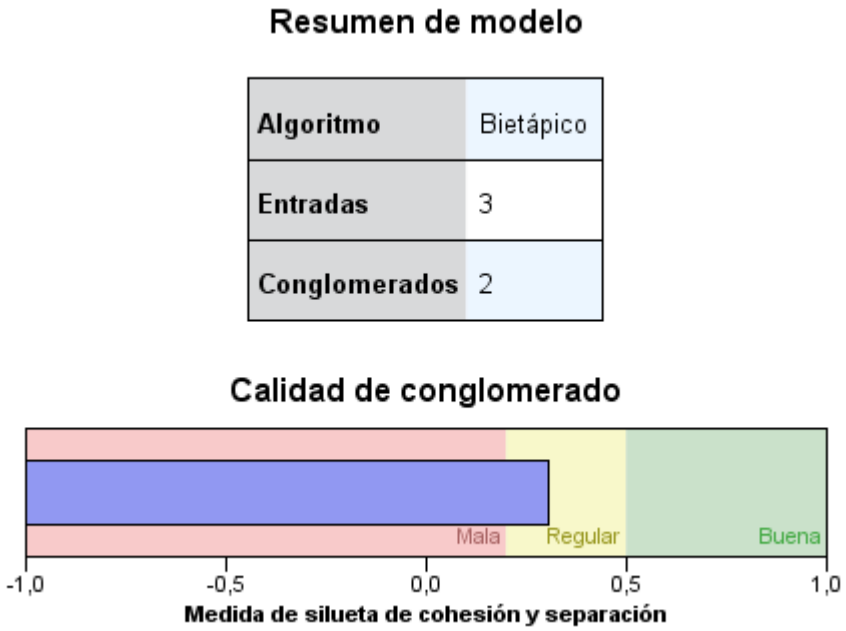

El primer clúster o clúster 1 está conformado en su mayoría por clientes que prefieren realizar sus compras en el e-marketplace N°3 (60%), utilizar la tarjeta Visa como medio de pago (100%) y en promedio, tienden a comprar aproximadamente cada 2 meses (mediana igual a 60 días). Sobre este clúster de clientes se plantea trabajar una campaña de descuento especial con tarjeta Visa por un mínimo de dos compras en el e-marketplace N°3, además de enviar mailings programados en base a productos similares al adquirido en su última compra, de esta manera el cliente será más propenso a generar una recompra en el futuro y podrá acortar su indicador de Recencia.

Por otro lado, la mayor parte de clientes que conforman el segundo clúster prefieren comprar a través del e-marketplace N°4 (63%), pagar haciendo uso de una tarjeta de crédito o débito (71%) y en promedio suelen realizar compras cada 25 días. Debido a que el segundo clúster de clientes espera menos de un mes para realizar una nueva compra, será más fácil que respondan de manera positiva ante las campañas que se le ofrezcan; para este grupo se plantea enviar publicidad a través de redes sociales o mailings de manera que estén al tanto de los nuevos descuentos que se suelen preparar para campañas estacionales a lo largo del año, así se captará su atención y se logrará fidelizar.

- Validación. La medida de silueta de cohesión y separación fue utilizada durante esta sección para comprobar que exista suficiente homogeneidad dentro de los grupos y heterogeneidad entre los grupos.

**Figura 7**

Validación de los clústeres obtenidos




La medida de silueta de cohesión y separación resultó igual a 0.3, por lo tanto, puede ser considerada como aceptable e indica que se cuenta con modelamiento válido.

## IV. Conclusiones

Al aplicar el método de clusterización bietápica sobre un conjunto de datos que reúne las características de compra de los clientes de cuatro e-marketplaces durante los meses de mayo, junio y julio de 2021, se lograron encontrar a dos clústeres de clientes. El primer clúster estuvo conformado por 9,802 clientes (31.3%) y el segundo por 21,554 (68.7%); además, el clustering obtenido se consideró como válido ya que la medida de silueta de cohesión y separación resultó mayor a 0.2

El primer clúster estuvo conformado en su mayoría por clientes que prefieren realizar sus compras en el e-marketplace N°3, utilizan la tarjeta Visa como medio de pago y en promedio, tienden a comprar aproximadamente cada 2 meses, mientras que el segundo contó con clientes que prefieren comprar a través del e-marketplace N°4, hacen uso de una tarjeta de crédito o débito y en promedio esperan menos de 25 días para realizar una nueva compra.

Sobre el primer clúster se propone trabajar una campaña exclusiva de descuento con tarjeta Visa por un mínimo de dos compras en el e-marketplace N°3, además de mantener informado mediante mails a los clientes sobre productos similares a los adquiridos, de esta manera mejorarán los indicadores de Recencia en el clúster. Por otro lado, sobre el segundo clúster se propone enviar publicidad a través de redes sociales o mailings de manera que los clientes puedan estar constantemente informados de las campañas estacionales a lo largo del año así se logrará mantenerlos fidelizados.

## V. Recomendaciones

Se recomienda que el equipo de TI pueda poner a disposición un set de datos que considere las transacciones desde el 2020, además de nuevas variables que no pudieron ser consideradas en el dataset con el que se trabajó la actividad, como la categoría del ítem o producto adquirido por el cliente, de esta manera el modelo podrá aprender mejor sobre las características de compra de los clientes y brindará mejores resultados. Además, se sugiere volver a correr un modelo que considere nuevas variables, por ejemplo, una variable que sume la cantidad de dinero desembolsado por cliente en fechas de Cyber, Cyber Wow, Black Friday, etc. También, considerar transformar las variables Recencia, Frecuencia y Monetario a cualitativas asignándole puntuaciones tal cual se hace en los análisis RFM tradicionales. Ambas recomendaciones influirán sobre número de clústeres a encontrar tras la aplicación del análisis y sobre la descripción del cliente tipo de cada clúster encontrado.

Respecto a las propuestas comerciales y de marketing, se recomienda fijar una fecha de corte para realizar una supervisión de la implementación y determinar si se aplaza su aplicación o se amerita un reajuste, que debe ir de la mano con las mejoras del dataset comentadas líneas arriba.

## VI. Bibliografía


Aldás, J., & Uriel, E. (2017). *Análisis multivariante aplicado con R*.

CAPECE. (2021). REPORTE OFICIAL DE LA INDUSTRIA ECOMMERCE EN PERÚ Impacto del COVID - 19 en el comercio electrónico en Perú y perspectivas al 2021. In *Reporte oficial de la industria del ecommerce en Perú 2020*. https://www.capece.org.pe/wp-content/uploads/2021/03/Observatorio-Ecommerce-Peru-2020-2021.pdf

Carrasco Díaz, S. (2013). *METODOLOGÍA DE LA INVESTIGACIÓN CIENTÍFICA Pautas metodológicas para diseñar y elaborar el proyecto de investigación*.

Chirinos, R., & Villalobos, M. (2017). BIG DATA PARA LA SEGMENTACIÓN DE MERCADOS EN REDES SOCIALES E ACCESORIOS DE MODA EMERGENTE. *MARKETING VISIONARIO UNIVERSIDAD Privada DR. RAFAEL BELLOSO CHACÍN*, *6*, 116–145.

Dietrich, T., & Rundle-Thiele, S. (2017). The importance of segmentation in social marketing strategy. In *Segmentation in Social Marketing: Process, Methods and Application* (pp. 109–113). https://doi.org/10.1007/978-981-10-1835-0_3

Doğan, O., Ayçin, E., & Bulut, Z. A. (2018). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. *International Journal of Contemporary Economics and Administrative Sciences*, *8*(1), 1–19. https://doi.org/10.1007/s10997-018-9447-3

IBM. (2013). IBM SPSS Statistics 22 Algorithms. In *Ibm* (p. 1151). http://library.uvm.edu/services/statistics/SPSS22Manuals/IBM SPSS Statistics Algorithms.pdf

IBM Corporation. (2014). *Análisis RFM*. https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=option-rfm-analysis

IBM Corporation. (2017). *Análisis de clústeres en dos fases*. https://www.ibm.com/docs/es/spss-statistics/25.0.0?topic=features-twostep-cluster-analysis

IBM Corporation. (2020). *Agrupación en clúster de modelos*. https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=nodes-clustering-models

Jiawei, H., Micheline, K., & Jian, P. (2012). Cluster Analysis: Basic Concepts and Methods. In *Data Mining: Concepts and Techniques* (Vol. 3, pp. 443–495). https://doi.org/10.1016/B978-

0-12-381479-1.00010-1

Tang, F., & Vargas, C. (2016). *SEGMENTACIÓN DE CLIENTES DE UNA TIENDA DE ELECTRODOMÉSTICOS UTILIZANDO EL ANÁLISIS DE CONGLOMERADOS*. http://repositorio.lamolina.edu.pe/bitstream/handle/UNALM/2214/E73-T3-T.pdf?sequence=1&isAllowed=y

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. *2018 IEEE 15th International Conference on E-Business Engineering, ICEBE 2018*, 119–126. https://doi.org/10.1109/ICEBE.2018.00027

Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/8884227

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *25*(2), 103–114. https://doi.org/10.1145/235968.233324