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

Caracterización de clientes de la marca TGI Fridays, mediante los algoritmos K-Means y K-Medoids

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Yeny LARICO SONCCO

ASESOR

Mg. Hugo Marino RODRIGUEZ ORELLANA

Lima, Perú

2021

Referencia bibliográfica

Larico, Y. (2021). *Caracterización de clientes de la marca TGI Fridays, mediante los algoritmos K-Means y K-Medoids*. [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	Yeny Larico Soncco
Tipo de documento de identidad	DNI
Número de documento de identidad	47936331
URL de ORCID	https://orcid.org/0000-0003-2364-6368
Datos de asesor	
Nombres y apellidos	Hugo Marino Rodriguez Orellana
Tipo de documento de identidad	DNI
Número de documento de identidad	40162362
URL de ORCID	https://orcid.org/0000-0002-7958-8550
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Helfer Joel Molina Quiñones
Tipo de documento	DNI
Número de documento de identidad	40014631
Miembro del jurado 1	
Nombres y apellidos	Antonio Bravo Quiroz
Tipo de documento	DNI
Número de documento de identidad	10130035
Datos de investigación	
Línea de investigación	A.3.2.1. Análisis Multivariante

Grupo de investigación	No Aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Torre 2 - Av. Circunvalación del Club Golf Los Incas N°134, Torre 2, piso 4, oficina 402 (Patio Panorama) País: Perú Departamento: Lima Provincia: Lima Distrito: Surco Latitud: -12.082644 Longitud: -76.968806
Año o rango de años en que se realizó la investigación	Julio 2021 – setiembre 2021
URL de disciplinas OCDE	Estadísticas, Probabilidad 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 18:00 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional: Dr. Helfer Joel Molina Quiñones (PRESIDENTE), Mg. Carlos Alberto Jaimes Velasquez (MIEMBRO) y el Mg. Hugo Marino Rodriguez Orellana (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**CARACTERIZACIÓN DE CLIENTES DE LA MARCA TGI FRIDAYS, MEDIANTE LOS ALGORITMOS K-MEANS Y K-MEDOIDS**", presentado por la señorita **Bachiller Yeny Larico Soncco**, 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 **BUENO,** con un calificativo promedio de **DIECISEIS (16)**.

A continuación, los miembros del Jurado dan manifiesto que la participante **Bachiller Yeny Larico Soncco** 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 18:30 horas se levantó la sesión firmando para constancia la presente Acta.

Dr. Helfer Joel Molina Quiñones
PRESIDENTE

Mg. Carlos Alberto Jaimes Velasquez
MIEMBRO

Mg. Hugo Marino Rodriguez Orellana
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.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADÍSTICA

INFORME DE EVALUACIÓN DE ORIGINALIDAD

El Director de la Escuela Profesional de Estadística, Dr. Roger Pedro Norabuena Figueroa, informa lo siguiente:

1. Operador del programa informático de similitudes: Dr. Roger Pedro Norabuena Figueroa
2. Documento evaluado: CARACTERIZACIÓN DE CLIENTES DE LA MARCA TGI FRIDAYS, MEDIANTE LOS ALGORITMOS K-MEANS Y K-MEDOIDS
1. Autor de la tesis: YENY LARICO SONCCO
2. Fecha de recepción de la tesis: 13/10/2022
3. Fecha de aplicación del programa informático de similitudes: 13/10/2022
   - Software utilizado: Turnitin
4. Configuración del programa detector de similitudes:
   - Excluye textos entrecomillados
   - Excluye bibliografía
   - Excluye cadenas menores a 40 palabras
5. Porcentaje de similitudes según programa detector de similitudes: diez por ciento (10%)
6. Fuentes originales de las similitudes encontradas:
   Fuentes de internet: 8 %
   Publicaciones: 0 %
7. Calificación de originalidad:
   - Documento cumple criterios de originalidad, sin observaciones

Lima, 13 de octubre del 2022

UNMSM
Firmado digitalmente por NORABUENA FIGUEROA Roger Pedro FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 14.10.2022 08:34:16 -05:00

Dr. Roger Pedro Norabuena Figueroa
Director

Resumen

En el siguiente trabajo se caracterizó a los clientes de la marca TGI Fridays, para poder tener éxito en las estrategias de Marketing.

Para esta categorización primero se tuvo que realizar la limpieza de datos, luego se seleccionaron las variables para el análisis, las cuales fueron edad, ganancia y cantidad de productos. Luego se aplicó el análisis jerárquico, el cual se realizó para poder obtener el número óptimo de clúster. Luego se aplicaron los análisis no jerárquicos de conglomerados K-means y K-meodids. Esto nos brindó los grupos de clientes con las características más homogéneas.

Seguido se realizó la evaluación de cada agrupamiento con el método de la silueta. Finalmente, se obtuvo como resultado tres grupos de clientes. El primer grupo se define por la alta demanda de productos y alta ganancia, el segundo grupo se define por ser los más jóvenes y con poca ganancia y el tercer grupo se define por la demanda regular de productos y mayor edad.

Con esta clasificación tendríamos más conocimiento de ellos, y en consecuencia elegir cuales serían las estrategias adecuadas para la publicidad.

Palabras clave: cluster, kmeans, kmedoids, CRISP-DM, algoritmo no supervisado

Abstract

In the following work, the clients of the TGI Fridays brand were characterized, in order to be able to be successful in Marketing strategies.

For this categorization, first the data cleaning had to be carried out, then the variables for the analysis were selected, which were age, profit and quantity of products. Then the hierarchical analysis was applied, which was carried out in order to obtain the optimal cluster number. Then the non-hierarchical K-means and K-meodids cluster analyzes were applied. This gave us the customer groups with the most homogeneous characteristics.

The evaluation of each grouping was then carried out with the silhouette method.

Finally, three groups of clients were obtained as a result. The first group is defined by the high demand for products and high profit, the second group is defined by being the youngest and with little profit and the third group is defined by the regular demand for products and older age.

With this classification we would have more knowledge of them, and consequently choose which would be the appropriate strategies for advertising.

Keywords: cluster, kmeans, kmedoids, CRISP-DM, unsupervised algorithm

1 Tabla de Contenido

Lista de Tablas

Lista de Figuras

1. Introducción

La empresa Franquicias Alimentarias S.A, propietaria de la marca TGI Fridays en Perú, es una empresa ubicada en el rubro gastronómico. La primera tienda se inauguró en Lima, en 1997 en el Ovalo Gutiérrez. Y gracias a la enorme acogida se expandió a otros distritos de Lima y en provincias como Arequipa y Trujillo.

Hoy en día, por la actual coyuntura debido a la pandemia generada por el COVID-19, ha llegado a cerrar una tienda, y por el Decreto Supremo N°183 de 2020-PCM por el cual se aprueba la "Reanudación de Actividades" conforme a la estrategia elaborada por el Grupo de Trabajo Multisectorial conformado mediante la Resolución Ministerial, las tiendas actuales cuentan con un aforo del 50%. Esto ha ocasionado un bajo performance en ventas de algunas tiendas. El negocio de venta se rige mediante dos modalidades; modalidad de atención en salón y modalidad de delivery.

Por ello en el área de Business Intelligence, se planteó desarrollar estrategias de negocio mediante la caracterización de los clientes de la marca TGI Fridays. Esto permitirá al área de Marketing entender a los clientes y construir estrategias diferenciadas en base a los resultados, tales como campañas, promociones, entre otros.

En el capítulo II, se brinda información sobre la institución donde se desarrollará esta actividad, como razón social, dirección postal y el periodo. Asimismo, se detalla la finalidad y objetivos de la empresa y área.

En el capítulo III, se presenta la organización de la empresa, así como la finalidad, objetivos y problemática que aborda el trabajo. Además, se expondrá la metodología, procedimientos y resultados de esta actividad.

Para el capítulo IV, se exponen las conclusiones del trabajo. Finalmente, en el capítulo V, se brindan las recomendaciones necesarias.

2. Información del Lugar Donde se Desarrolló la Actividad

2.1. Datos de la Empresa

Franquicias Alimentarias S.A. | T.G.I Fridays. Es una empresa que se encarga de brindar experiencia y hacer que los clientes sientan que dentro de las cuatro paredes siempre sea viernes, ya que este es nuestro propósito. Asimismo, el objetivo del área de BI es generar conocimientos del negocio y de nuestros huéspedes para garantizar valor agregado.

2.1.1. Misión

Siempre ha sido y será superar las expectativas de los clientes, en cuanto a servicio, calidad de comida y ambiente.

2.1.2. Visión

Es ser considerado el concepto preferido de comida casual en Perú.

2.1.3. Razón Social

Marcas Alimentarias S.A.

RUC: 20298674611

2.1.4. Dirección

Av. Circunvalacion del Golf L Nro. 134 Int. 402, Santiago de Surco.

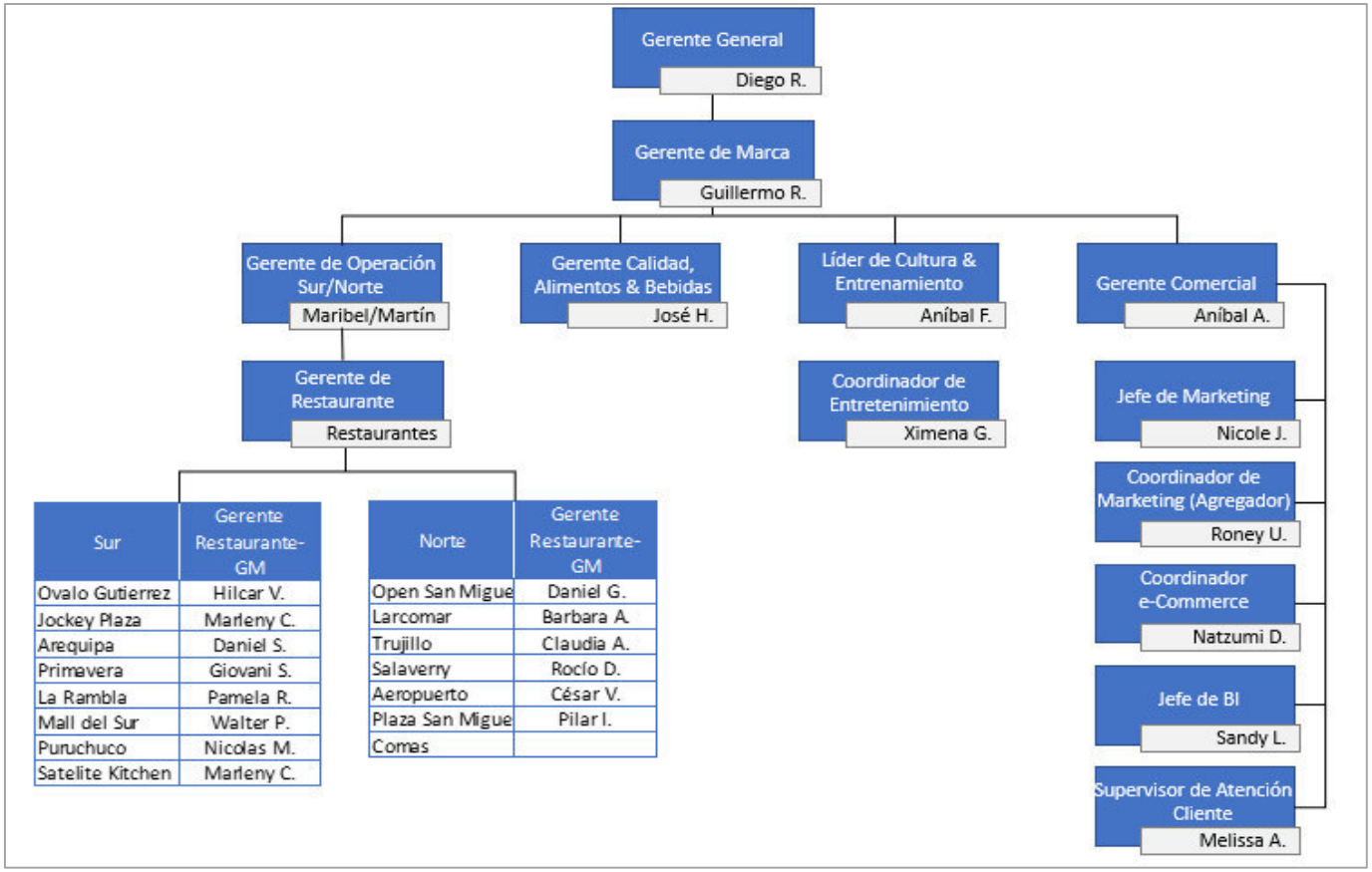
2.1.5. Correo Electrónico del Profesional a Cargo

ylarico@fridaysperu.com

2.1.6. Organigrama de la Empresa

Figura 1.

Organigrama de la Empresa de la marca TGI Fridays.



Sur	Gerente Restaurante-GM
Ovalo Gutierrez	Hilcar V.
Jockey Plaza	Marleny C.
Arequipa	Daniel S.
Primavera	Giovani S.
La Rambla	Pamela R.
Mall del Sur	Walter P.
Puruchuco	Nicolas M.
Satelite Kitchen	Marleny C.

Norte	Gerente Restaurante-GM
Open San Migue	Daniel G.
Larcomar	Barbara A.
Trujillo	Claudia A.
Salaverry	Rocío D.
Aeropuerto	César V.
Plaza San Migue	Pilar I.
Comas	

Fuente: Marcas Alimentarias S.A. | T.G.I Friday's

3. Descripción De La Actividad

Se ha seguido la metodología CRISP-DM

1) Comprensión del negocio
   - Realizar el plan de la actividad
2) Comprensión de los datos
   - Esquema de BD
   - Análisis exploratorio de datos
3) Preparación de los datos
   - Selección de datos
   - Limpieza de datos
   - Estructurar los datos
4) Modelado
   - Seleccionar técnica de modelado
   - Construir el modelo
   - Evaluación del modelo
5) Evaluación
   - Evaluar los resultados
   - Revisión del proceso
6) Implementación
   - Plan de implementación
   - Informe final
   - Revisión de la actividad

3.1. Finalidad de la actividad

Este documento tiene como finalidad conocer el perfil de los clientes que consumen en las diferentes tiendas de la marca TGI Fridays. Esto para poder aplicar estrategias en base a los segmentos obtenidos.

3.1.1. Objetivos del TSP

Objetivo general

Determinar la caracterización de clientes de la marca TGI FRIDAYS, mediante los algoritmos K-Means y K-Mediods.

Objetivos específicos

- Identificar el número de conglomerados aplicando el algoritmo clúster jerárquico a los clientes de la marca TGI Fridays.
- Aplicar y evaluar el algoritmo K-Means a los clientes de la marca TGI Fridays.
- Aplicar y evaluar el algoritmo K-Medoids a los clientes de la marca TGI Fridays.
- Identificar las características de los segmentos de los clientes de la marca TGI Fridays.

3.2. Problemática

En la empresa no se contaba con un análisis que permita conocer el perfil de los clientes, para poder incrementar la fidelización y demanda mediante estrategias que se puedan aplicar a nivel de marketing. Asimismo, el análisis de conglomerados jerárquico y no jerárquico nos permitirá explorar la existencia de los perfiles de los clientes de la marca TGI Fridays.

3.2.1. Antecedentes

Elguera (2018), realizó un estudio acerca de la segmentación de clientes, mediante el algoritmo de partición alrededor de medoides. Esta investigación tuvo como objetivo la aplicación del algoritmo PAM y así poder agrupar a los clientes de un casino. El tipo de investigación fue no experimental con diseño descriptivo dispuesto a una muestra de 2798 clientes, donde concluyó que el agrupamiento primeramente por el método de la silueta que el número adecuado de clúster es de tres para el algoritmo PAM. Finalmente aplicando el logaritmo de segmentación para los clientes del casino nos da una proporción de 49.4%, 11.3% y 39.4% para el primer, segundo y tercer grupo respectivamente.

Cuadros et al. (2017), realizó un estudio sobre la aplicación de la técnica multivariante no paramétrica K-means, realizando de manera previa un análisis RFM. La cual tuvo como objetivo detectar aquellos clientes más valiosos. El tipo de investigación en este trabajo fue no experimental con diseño transversal, donde tras la depuración, la base de datos contó con 304 clientes, los cuales generaron durante 8 meses 5962 transacciones. Finalmente, con los cinco grupos establecidos, se calculó el puntaje de cada grupo donde los grupos 1 y 4 obtuvieron los mayores puntajes y el grupo 3 tuvo el puntaje mínimo.

Azuela et al. (2019), realizó un estudio sobre la segmentación de los compradores online. Este trabajo tuvo como objetivo la segmentación de los compradores online con base a su periodicidad de consumo por internet. El tipo de investigación fue no experimental con diseño descriptivo. La muestra trabajada fue de 1495 personas que afirmaron haber realizado compras por internet. Para ese trabajo se aplicó el análisis multivariante clúster no jerárquico K-means, donde obtuvo como resultado que el primer grupo con 441

clientes presentó que sus consumos eran una vez al mes, el cual es clasificado como frecuente, ya que fija que el promedio de navegación por internet es de 6.6 horas por día. Por otro lado, el segundo grupo con 567 clientes presentó que sus consumos eran dos veces por año, el cual es clasificado como ocasional, ya que fija que el promedio de navegación por internet es de 5.3 horas por día, y el tercer grupo con 446 clientes, presentó que sus consumos eran una vez al año, el cual es clasificado como esporádico, ya que fija que el promedio de navegación por internet es de 5.6 horas por día. Concluyendo que los compradores frecuentes y ocasionales compran aproximadamente $250, y en el proceso van complementando el tiempo personal y finalizando Compra y vende principalmente a través de teléfonos móviles.

3.3. Definiciones

3.3.1. Metodología CRISP-DM

Según Rodríguez (2015), este es un proceso de modelo para proyectos de minería de datos en general, comprende una serie de seis fases:

- Comprensión del negocio: esto tiene como objetivo comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial y convertir este conocimiento en un problema de minería de datos.
- Comprensión de los datos: esto pretende familiarizarse con los datos, identificar problemas de calidad y obtener los primeros conocimientos sobre los datos.
- Preparación de los datos: es la fase donde se realiza la selección de variables, transformación de datos, limpieza de datos y cualquier otra tarea que se considere necesaria.
- Modelado: es la fase donde se seleccionan las técnicas de modelado, construcción y evaluación del modelo.
- Evaluación: antes de proceder a la implementación final del modelo, es importante realizar un análisis más completo. evaluación, revisando los pasos ejecutados para su construcción, y cerciorarse de que cumpla adecuadamente con los objetivos de negocios.
- Implementación: esta es la fase de explotación del proyecto.

3.3.2. Minería de datos.

Esta definición fue desarrollada en la década de los 80 por desarrolladores de base de datos. Minería de datos puede definirse como el proceso de descubrimiento de nuevas e importantes relaciones explorando grandes volúmenes de datos, según Rodríguez (2015)

3.3.2.1 Métodos no supervisados. Es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones. Los algoritmos más comunes encontramos al Clúster Jerárquico, K-means, K-medoids, etc.

3.3.3. Análisis de conglomerados o Clustering

Según Aldas y Uriel (2017), es una técnica multivariante que se utiliza para clasificar diferentes observaciones en grupos. Donde dentro de cada grupo (clúster) sea lo más homogéneo en cuanto a las variables usadas, y que entre grupos sean diferentes (pp. 77).

3.3.3.1. Método de Agrupación.

- **Método Jerárquico.**

Según Kassambara (2017), se caracterizan porque en cada fase del algoritmo solamente un objeto puede cambiar de grupo, estos grupos estarán anidados en las fases anteriores. En caso de que un objeto sea asignado a determinado grupo ya no cambiará a otro grupo. La construcción de su estructura tiene la forma de un árbol. Además, no requiere conocer previamente el número de grupos.

- **Método No Jerárquico.**

Se caracteriza porque ya cuenta de manera a priori con el número de grupos fijos.

La agrupación jerárquica se puede subdividir en dos tipos:

- Método jerárquico aglomerativo, en el que cada observación se considera inicialmente como un racimo propio. Luego, los clústeres más similares se fusionan sucesivamente hasta que haya un solo clúster.

- Método jerárquico divisivo, la cual es una inversa del método jerárquico aglomerativo, comienza con la raíz, en que todos los objetos están incluidos a un grupo. Entonces en cada paso el algoritmo divide sucesivamente al grupo más heterogéneo hasta que el cada clúster quede con un elemento cada uno.

3.3.3.2. Distancia. Según Mardia et al. (1995), sean P y Q dos puntos los cuales representan las medidas x y y en dos objetos. Una función de valores reales $d(P, Q)$ es denominada función de distancia si cumple las siguientes propiedades:

a. Simetría: $d(P, Q) = d(Q, P)$
b. No negativo: $d(P, Q) \geq 0$
c. $d(P, P) = 0$

Para muchas funciones de distancia se sostiene también las siguientes propiedades:

d. $d(P, Q) = 0 \; si \; y \; sólo \; si \; P = Q$
e. Desigualdad triangular: $d(P, Q) \leq d(P, R) + d(R, Q)$

3.3.3.3. Medidas de similaridad para variables métricas. Según Aldas y Uriel (2017), cuando las variables son de tipo cuantitativas, ya sea de escala de razón o intervalar se puede utilizar las siguientes medidas de similaridad:

- **Distancia Euclidiana.**

$$D_{ij} = \sqrt{\sum_{p=1}^{k} \left(x_{ip} - x_{j\dot{p}}\right)^2}$$

- **Distancia Ecludiana al cuadrado.**

$$D_{ij} = \sum_{p=1}^{k} \left(x_{ip} - x_{j\dot{p}}\right)^2$$

- **Distancia de Minkowski.**

$$D_{ij} = \left[\sum_{p=1}^{k} \left|x_{ip} - x_{jp}\right|^{n}\right]^{1/n}$$

- **Distancia Manhattan.**

$$D_{ij} = \sum_{p=1}^{k} \left|x_{ip} - x_{jp}\right|^{n}$$

Donde las notaciones son similares en cada ecuación:

- i y j son observaciones de n posibles.

- x_{ip} y $x_{j\dot{p}}$ son valores k variables existentes en dichas observaciones.

3.3.3.4. Métodos de Enlace

● **Método del vecino más cercano.**

Según Kassambara (2017), se define como el valor mínimo de todas las distancias por pares entre los elementos del conglomerado A y los elementos del conglomerado B. Este método se expresa de la siguiente manera según Díaz y Morales (2012):

$$d_{AB} = min\,\{d_{ij}) \qquad i \in A\; j \in B$$

● **Método del vecino más lejano**

Se define como el valor máximo de todas las distancias por pares entre los elementos del conglomerado A y los elementos del conglomerado B. Tiende a producir conglomerados más compactos.

$$d_{AB} = máx\,\{d_{ij}) \qquad i \in A\; j \in B$$

- **Método de la vinculación de promedio**

La distancia entre dos conglomerados se define como la distancia media entre los elementos del conglomerado A y los elementos del conglomerado B.

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$

- **Método del centroide.**

La distancia entre dos grupos se define como la distancia entre el centroide del conglomerado A (un vector medio de variables de longitud p) y el centroide del conglomerado B.

- **Método Ward.**

Minimiza la varianza total dentro del conglomerado. En cada paso, se fusiona el par de grupos con una distancia mínima entre grupos.

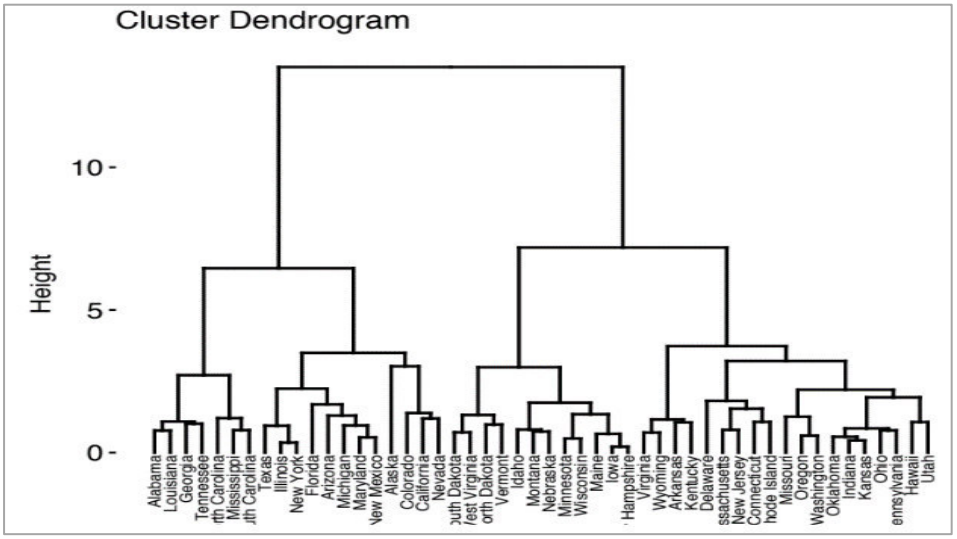
$$SCW = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij} \right)^2 \right)$$

3.3.4. Dendrograma

El dendrograma es una jerarquía multinivel donde los grupos de un nivel se unen para formar los grupos en los siguientes niveles. Esto permite decidir el nivel en que cortar el árbol para generar grupos adecuados de objetos de datos.

Figura 2.

Dendrograma



Nota. El dendrograma muestra que cada hoja corresponde a un objeto. Tomado de *Multivariate Analysis I: Practical Guide To Cluster Analysis in R* (p.72), por A. Kassambara, 2017, STHDA.

3.3.5. Verificación de la Calidad del Dendograma.

Una forma de medir qué tan bien el árbol de clúster es calcular la correlación entre las distancias cofenéticas y los datos de distancia originales. Para verificar que la agrupación se haya realizado correctamente, la alineación de objetos del árbol debe tener un alto nivel de correlación con la distancia entre los objetos en la matriz de distancias original.

Cuanto más cerca de 1 esté el valor del coeficiente de correlación, con mayor precisión refleja la solución de agrupación en clústeres de sus datos. Se considera que los valores superiores a 0,75 son buenos.

3.3.6. Métodos para evaluar la tendencia del clúster o agrupamiento.

● **Método Estadístico.**

La estadística de Hopkins se utiliza para estimar la tendencia de agrupamiento de un conjunto de datos midiendo la probabilidad de que un conjunto de datos dado se genera mediante una distribución de datos uniforme. En otras palabras, prueba la aleatoriedad espacial de los datos.

Por ejemplo, sea D un conjunto de datos reales. La estadística de Hopkins se puede calcular como seguir:

a) Seleccionar una muestra de n puntos $(p_1, \dots, p_n)$ de D.

b) Para cada $p_i \in D$ encontrar su vecino más cercano p_j y calcular

$$dist(p_i, p_j) = x_i$$

c) Generar un conjunto de datos simulados de una distribución uniforme con n puntos $(q_1, \dots, q_n)$ y con la misma dispersión que los datos originales D.

d) Para cada q_i encontrar su vecino más cercano q_j en D y calcular

$$y_i = dist(q_i, q_j)$$

e) Hallar la estadística de Hopkins (H).

La fórmula se define como sigue:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

La hipótesis nula y alternativa está definidas como lo siguiente:

H_0: D está distribuido de forma uniforme (es decir: no hay agrupaciones significativas)

H_1: D no está distribuido de forma uniforme (es decir: el conjunto de datos contiene agrupaciones significativas)

Según (Hopkins, 1954, como se citó en Lawson y Jurs, 1990), se puede realizar la prueba estadística de Hopkins de forma iterativa, utilizando 0.5 como límite para rechazar la hipótesis alternativa. Es decir, si H <0.5, entonces es poco probable que D tenga conglomerados estadísticamente significativos.

En otras palabras, si el valor de la estadística de Hopkins está cerca de 1, entonces podemos rechazar la hipótesis nula y concluir que el conjunto de datos D es significativamente agrupable.

● **Métodos Visuales.**

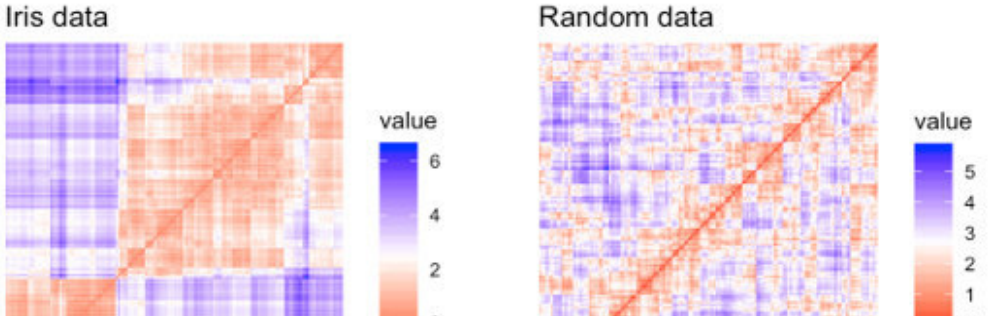
Se utiliza un gráfico para mostrar los resultados de la matriz de disimilaridad para el conjunto de datos reales y los simulados.

Es apropiado para muestras conformadas por conjunto de datos no muy grandes.

1. Calcule la matriz de disimilitud (DM) entre los objetos en el conjunto de datos usando la medida de distancia euclidiana.
2. Reordena el DM para que los objetos similares estén cerca unos de otros. Este proceso crea una matriz de disimilitud ordenada (ODM).
3. El ODM se muestra como una imagen de disimilitud ordenada (ODI), que es el resultado visual del IVA.

Figura 3.

Matriz de disimilaridad para la base de datos Iris y para una base de datos aleatoria (Random data).



Nota. En el gráfico el color rojo indica alta similaridad (es decir baja disimilaridad) y el color azul indica baja similaridad (es decir alta disimilaridad). Tomado de *Multivariate Analysis I: Practical Guide To Cluster Analysis in R* (p.126), por A. Kassambara, 2017, STHDA.

3.3.7. Determinación del número óptimo de clústeres o conglomerados.

Según Kassambara (2017), de acuerdo al método utilizado para medir similitudes y los parámetros utilizados para la partición, se obtendrá el número óptimo de clústeres, aunque resalta que este número óptimo es de alguna manera subjetiva.

Estos métodos incluyen métodos directos y métodos de prueba estadísticos que a continuación se muestran:

1. Métodos directos: consiste en optimizar el criterio de determinación de número de clústeres. Estos métodos se denominan métodos de codo y silueta.

2. Métodos de prueba estadística: se basa en contrastar evidencia contra hipótesis nula. Un ejemplo es el estadístico GAP

- **Método del codo**

La SC dentro de los clústeres debe ser mínima, mide la compacidad de los clústeres.

$$SC = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

Para construir el gráfico:

1. Calcular el algoritmo de agrupamiento para diferentes valores de k; generalmente para valores entre 1 y 10.
2. Para cada valor de k calcular la SC dentro del agrupamiento.
3. Graficar la curva de los valores de la SC dentro de los agrupamientos y los valores de k respectivos.
4. La ubicación de un codo en la curva generalmente se considera como un Indicador del número apropiado de grupos.

- **Método Silueta Promedio**

Mide la calidad de la agrupación, indica que tan bien se encuentra cada individuo dentro del grupo.

Un promedio alto indica buen agrupamiento.

El algoritmo es el siguiente:

1. Calcular el algoritmo de agrupamiento para diferentes valores de k, generalmente de 1 a 10.
2. Para cada k, calcular la silueta promedio de las observaciones.

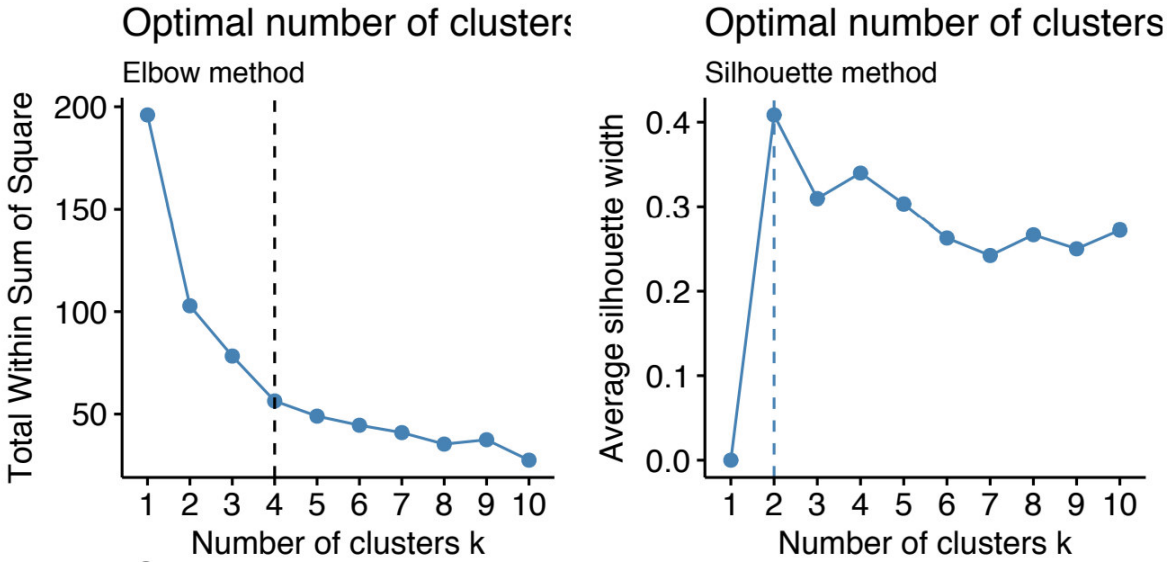
$$S(i) = \frac{b(i) - a(i)}{máx\{a(i) - b(i)\}}$$

a(i) = disimilaridad promedio entre la i-ésima observación y las otras del agrupamiento.

3. Graficar el valor de la silueta promedio para cada valor de k.
4. La ubicación del máximo valor de la silueta promedio, se considera como el número de agrupamientos adecuado.

Figura 4.

Gráficos para la identificación del número óptimo de clústeres, según el Método del Codo y Método de la Silueta.



Nota. En el gráfico el método del codo sugiere que el número óptimo de clúster sería de 4 y el método de la Silueta sugiere que el número óptimo de clúster es de 2. Tomado de *Multivariate Analysis I: Practical Guide To Cluster Analysis in R* (p.134), por A. Kassambara, 2017, STHDA.

Asimismo, se cuentan con otros métodos para encontrar un número óptimo de clústeres como los que se mencionan en Aldas y Uriel (2017), sin embargo, no se presentarán en este trabajo.

3.3.8. Validación de Conglomerados.

El término validación de conglomerados se utiliza para la evaluación de bondad de los grupos resultantes. Esto es primordial para prever encontrar patrones en un dato aleatorio.

Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado.

Generalmente, las estadísticas de validación de la agrupación en clústeres se pueden clasificar en 3 clases:

1. Validación interna de clúster, emplea información interna del desarrollo de agrupamiento para examinar la bondad de una estructura de clúster.

2. Validación externa de clúster, consiste en contrastar los resultados de un análisis de conglomerados con un resultado conocido de otro análisis de clúster, como las etiquetas proporcionadas por cada análisis clúster (por ejemplo: “Grupo 1”, “Grupo 2”, etc).

3. Validación relativa del clúster, evalúa la estructura del clúster variando diferentes resultados de parámetros para el mismo algoritmo. Por lo general, se usa para determinar la cantidad óptima de clústeres.

- **Coeficiente de la Silueta.**

El análisis de la silueta calcula si un objeto está bien agrupado y permite estimar la distancia promedio entre los grupos. El gráfico de silueta muestra una medida de qué tan cerca está cada punto en un grupo de puntos en los grupos vecinos.

Para cada observación i, el ancho de la silueta s_i se calcula de la siguiente manera:

1. Para cada objeto i, calcule la disimilaridad promedio a_i entre i y todos los otros puntos del clúster al cual pertenece i.

2. Para otro clúster C, al cual i no pertenece, calcular la disimilaridad promedio $d(i, C)$ de i a todas las observaciones de C. El más pequeño de estas $d(i, C)$, disimilaridades, está definido como $b_i = min_C d(i, C)$. El valor de b_i puede ser visto como la disimilaridad entre i y el clúster vecino. Es decir, el más cercano al cual no pertenece.
3. Finalmente, el ancho de la silueta de la observación i es definida por la fórmula:

$$S_i = \frac{(b_i - a_i)}{max\ (a_i; b_i)}$$

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

- Las observaciones con un S_i de larga anchura (cercano a 1), están bien agrupados.
- Un S_i cercano a 0 indica que la observación se encuentra entre dos clúster.
- Las observaciones con un S_i negativo probablemente estén ubicadas en el grupo equivocado.

3.3.9. Agrupación K-Means

El agrupamiento en clústeres de K-medias (Según MacQueen 1967, como se citó en Kassambara, 2017) es el método no supervisado más utilizado para particionar un conjunto de datos dado en un conjunto de k grupos (es decir, k conglomerados), donde k representa el número de grupos predefinidos por el analista.

El algoritmo K-means se puede resumir de la siguiente manera:

1. Especifique el número de clústeres (K) que se crearán (por el analista)

2. Seleccionar aleatoriamente k objetos del conjunto de datos como centroides del conglomerado inicial.

3. Asignar a cada observación a su centroide más cercano, basado en el Euclidean distancia entre el objeto y el centroide.

4. Para cada uno de los k conglomerados, actualice el centroide del conglomerado calculando los nuevos valores medios de todos los puntos de datos del conglomerado.

El centoide de un grupo Kth es un vector de longitud p que contiene las medias de todas las variables para las observaciones en el k-ésimo grupo; p es el número de variables.

5. Minimizar iterativamente el total dentro de la suma del cuadrado. Es decir, repita los pasos 3 y 4 hasta que las asignaciones del clúster dejen de cambiar o el número máximo de se alcanza iteraciones.

3.3.10. Agrupación K-Medoids

El algoritmo k-medoids (Según Kaufman y Rousseeuw 1990, como se citó en Kassambara, 2017) es un enfoque de agrupamiento relacionado con el agrupamiento de k-means para particionar un conjunto de datos en k grupos o clústeres. En la agrupación de k-medoides, cada grupo está representado por uno de los puntos de datos del grupo. Estos puntos son medoides de racimo nombrados.

El término medoide se refiere a un objeto dentro de un grupo para el cual la disimilitud promedio entre él y todos los demás, los miembros del clúster son mínimos. Corresponde al punto más céntrico del grupo.

El método de agrupamiento de k-medoids más común es el algoritmo PAM (Partición Alrededor de Medoides).

El algoritmo PAM se basa en la búsqueda de k objetos representativos o medoides entre las observaciones del conjunto de datos.

El algoritmo PAM procede los siguientes pasos:

1. Seleccione k objetos para convertirlos en medoides, o en caso de que se proporcionen estos objetos utilícelos como medoides.
2. Calcule la matriz de disimilitud si no se proporcionó.
3. Asignar cada objeto a su medoide más cercano.
4. Para cada búsqueda de conglomerado, si alguno de los objetos del conglomerado disminuye el coeficiente de disimilitud promedio, si lo hace, seleccione la entidad que disminuye más este coeficiente como el medoide para este conglomerado.

5. Si al menos un medoide ha cambiado, vaya a (3), de lo contrario finalice el algoritmo.

3.4. Metodología

3.4.1. Metodología de Investigación

- **Enfoque:** Cuantitativo
- **Diseño:** No experimental
- **Nivel:** Explicativo
- **Población.**

La población está conformada por 1002 clientes que consumieron al menos una vez en la marca TGI Fridays, en el periodo de abril a junio del 2021.

- **Muestra.**

Para la presente investigación se consideró toda la población, esto es debido a la disponibilidad del total de clientes que consumieron al menos un producto de la marca TGI Fridays. Asimismo, se requería el análisis en toda la población dentro de los requerimientos del área de trabajo.

3.5. Métodos.

3.5.1. Método CRISP-DM

- Entendimiento del Negocio.

El primer paso es comprender las necesidades del negocio de la empresa. Comprender sus expectativas y comprender las capacidades de su negocio para futuros eventos mensuales.

Una de estas es conocer el perfil del cliente. Esto ayudaría mucho al área de Marketing para que puedan realizar campañas, dependiendo de las características del cliente.

- Comprensión de los datos.

Esta etapa consiste en el análisis de la base de datos recolectada. A continuación se encuentra el análisis exploratorio de datos.

Tabla 1

Tabla de estado de variables en la base de datos de muestra proporcionada por la marca Fridays

Variable	Cant. Ceros	Porc. Ceros	Cant. Vacíos	Porc. Vacíos	Cant. Infinitos	Porc. Infinitos	Tipo de Variable	Cant. De Valores Únicos
ID_Tienda	0	0	0	0	0	0	integer	15
MicrosBsnzDate	0	0	0	0	0	0	character	179
MicrosChkNum	0	0	0	0	0	0	integer	941
CustomerID	0	0	0	0	0	0	integer	956
ventaNeta	0	0	0	0	0	0	numeric	619
tickets	0	0	0	0	0	0	integer	1
CostoTotal	0	0	0	0	0	0	integer	96
cantidad	0	0	0	0	0	0	integer	20
Ganancia	5	0.5	0	0	0	0	integer	159
EDAD	0	0	0	0	0	0	integer	53

Nota: Valores Infinitos mencionados en las columnas 6 y 7 hacen referencia hacia aquellos posibles valores que tiendan al infinito.

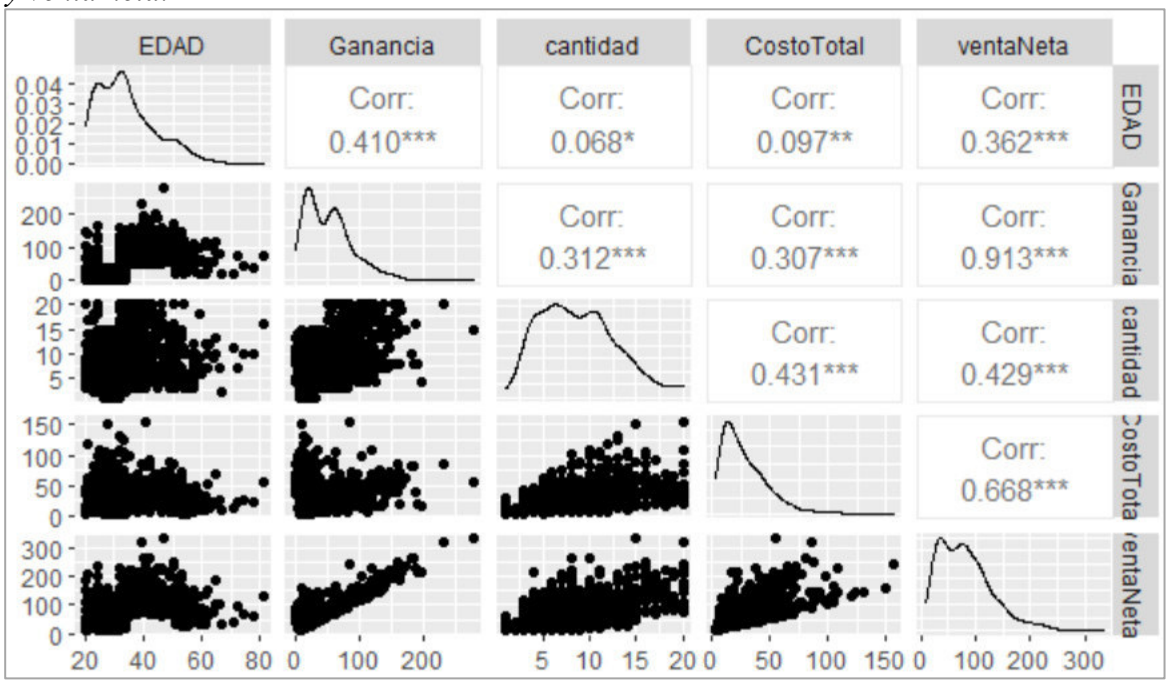
Podemos apreciar en la tabla 1, que ninguna de las variables de estudio cuentan con información faltante para este data set. Se tiene que los 1002 clientes se encuentran distribuidos en 15 restaurantes ubicados en zonas diferentes. El campo de tickets tiene registrado un único valor para todos los clientes, por esta razón se retira del análisis.

Asimismo, para el análisis de segmentación se ha tomado en cuenta las variables edad, ganancia, cantidad, costo total y venta neta.

A continuación, se muestra el análisis de correlación entre las variables edad, ganancia, cantidad, costo total y venta neta.

Figura 5.

Matriz de correlación y dispersión de la edad, ganancia, cantidad adquirida, costo total y venta neta.



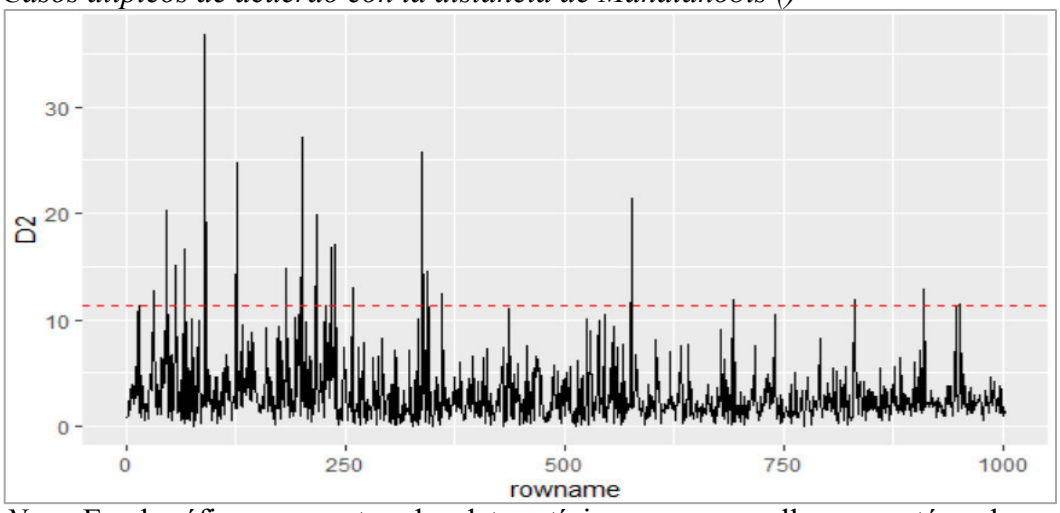
Nota. En el gráfico, el nivel de correlación de Pearson se muestra en la parte superior de la matriz bajo el nombre de Corr.

En la figura 5, podemos notar que las variables ganancia, edad y cantidad presentan un bajo nivel de correlación entre sí. Permitiendo así considerarlas para el análisis de segmentación.

A continuación, se analizará los casos atípicos mediante la distancia de Mahalanobis.

Figura 6.

Casos atípicos de acuerdo con la distancia de Mahalanobis ()



Nota. En el gráfico se muestran los datos atípicos como aquellos que están sobre la línea roja.

Para identificar los casos atípicos se calcula la distancia de Mahalanobis. Dado que, la distancia de Mahalanobis se distribuye como una chi-cuadrado con tantos grados de libertad como variables implicadas bajo la hipótesis nula de que el i-ésimo caso no es atípico, se calcula que el punto crítico, con un nivel de significancia del 99%, dando como resultado 11.3. Esto nos permite trazar una línea horizontal mostrada en la figura 6 para identificar aquellos casos que sobrepasen este punto, teniendo así 21 casos atípicos. En la siguiente etapa del método CRISP – DM aplicado se removerán estos casos atípicos que afectan al análisis de segmentación.

- Preparación de los datos

En esta etapa se procede a seleccionar las variables que cuentan con menor nivel de correlación (edad, ganancia, cantidad). Asimismo, se remueven los casos atípicos identificados en la etapa anterior.

Tabla 2

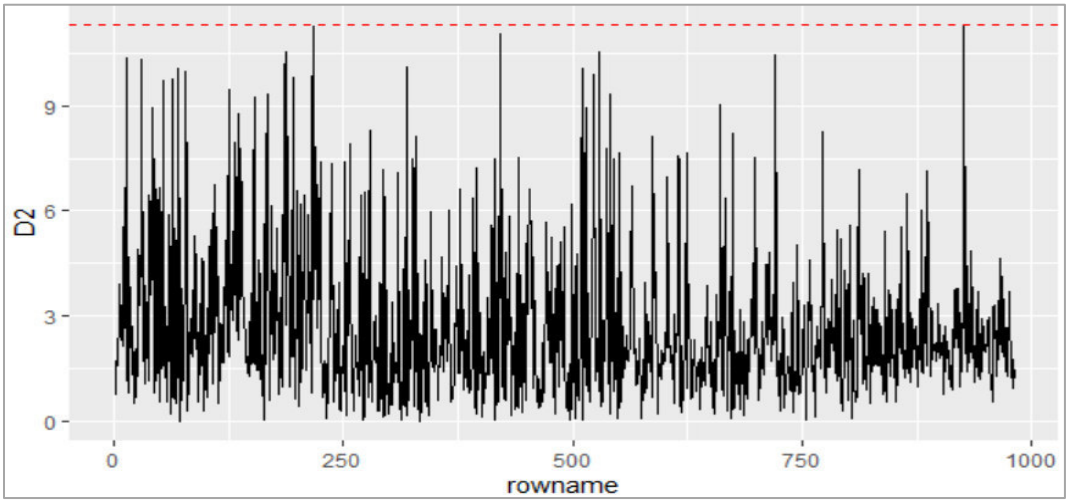
Descripción de variables

Variable	Definición	Tipo de Variable	Escala
Edad	Es la edad de cada cliente	Cuantitativa Discreta	Razón
Ganancia	Indicador económico que representa la ganancia real considerando el costo de venta	Cuantitativa Continua	Razón
Cantidad	Número de productos consumidos por el cliente	Cuantitativa Continua	Razón

Fuente. Elaboración propia

Figura 7.

Gráfico de las distancias de Mahalanobis () sin casos atípicos



Nota. En el gráfico se muestran que los datos atípicos que fueron identificados en la figura 6, fueron removidos.

Para realizar el análisis de segmentación se debe verificar si los datos tienen tendencia a poder ser agrupables mediante el estadístico de Hopkings, previamente estandarizando los datos.

Tabla 3

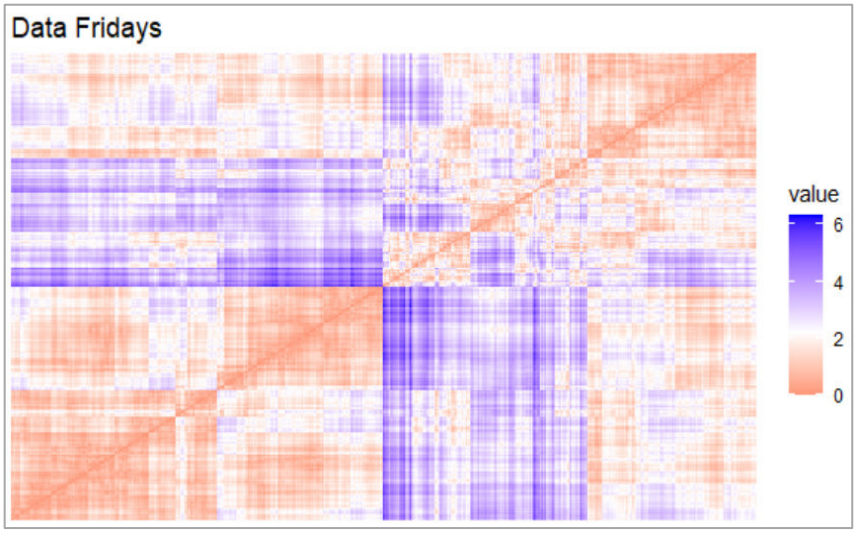
Estadístico de Hopkins

Valor	Hipótesis
0.25	No es agrupable

Fuente. Elaboración propia

Figura 8.

Gráfico de matriz de disimilaridad ordenada

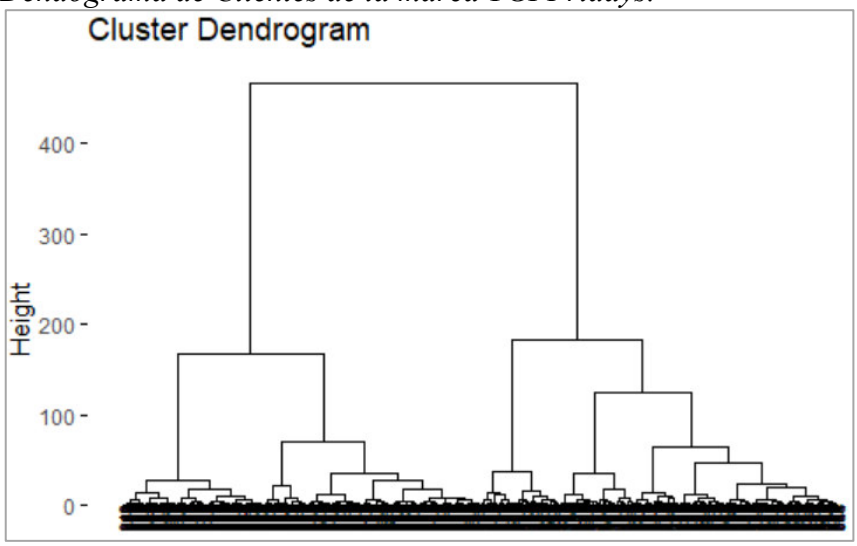


Fuente. Elaboración propia

Se recomienda tomar en cuenta que los resultados obtenidos de la segmentación no son del todo perfectos acorde al marco teórico mostrado. Además, tomar este análisis como una base para poder mejorar y proponer a futuro una estrategia de ventas en la marca TGI Fridays.

Figura 9.

Dendograma de Clientes de la marca TGI Fridays.



Fuente. Elaboración propia

En la figura 9, nos muestra las distancias entre las observaciones representadas por las ramificaciones del dendrograma. Cuanto mayor es la altura de la fusión, menos similares son las observaciones. Esta altura se conoce como la distancia Cophenetic entre dos observaciones, la cual es de 0.55. (A. Kassambara)

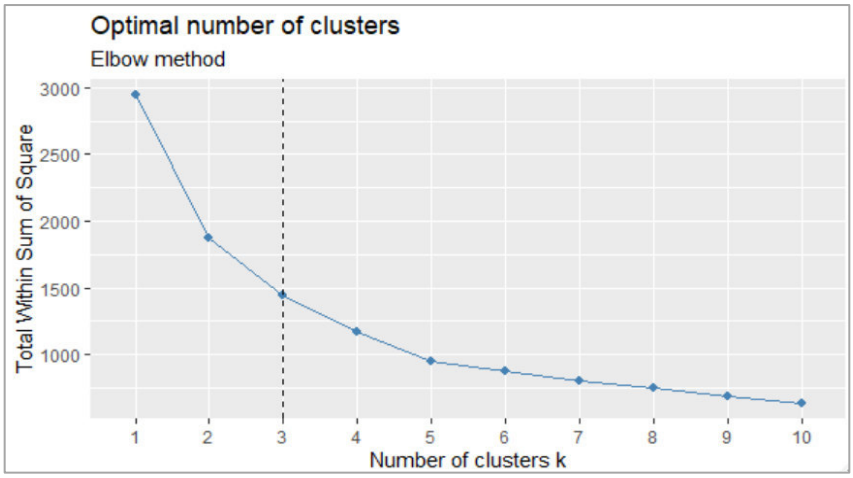
- Modelado

- Aplicación de Análisis Jerárquico de conglomerados.

En esta etapa determinamos el número óptimo de clústeres mediante el método de codo y silueta.

Figura 10.

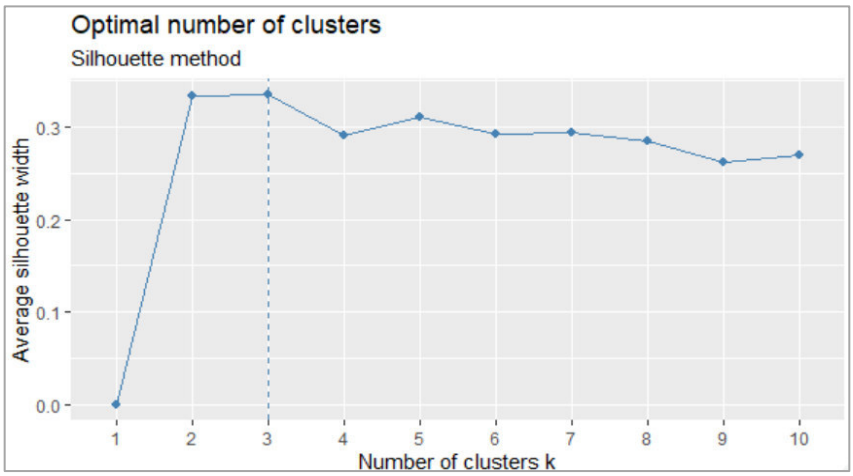
Gráfico del Método del Codo



Fuente. Elaboración propia

Figura 11.

Gráfico del Método de la Silueta



Fuente. Elaboración propia

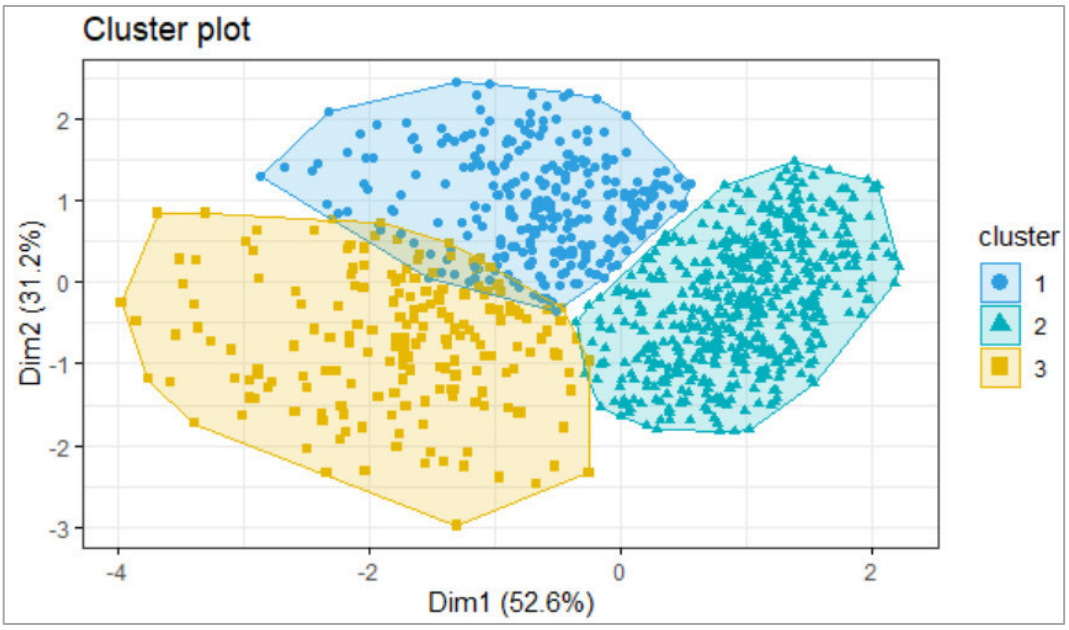
Según estas las figuras 10 y 11, es posible definir k = 3 como el número óptimo de grupos en los datos. Por lo tanto se procedió a ejecutar el algoritmo de clúster jerárquico, y de acuerdo con Kassambara (2017), se utilizó el método Ward y la distancia euclidiana para obtener los segmentos.

- Aplicación de Análisis de conglomerados no Jerárquico: K-means

El clúster jerárquico permitió identificar tres segmentos, estos segmentos se utilizaron como información apriori para poder aplicar un algoritmo de segmentación no jerárquico K - Means con un número de segmentos k = 3. A continuación se presenta el análisis no jerárquico de conglomerados.

Figura 12.

Mapa Perceptual de la Caracterización de Clientes de la marca TGI Fridays del agrupamiento K-means

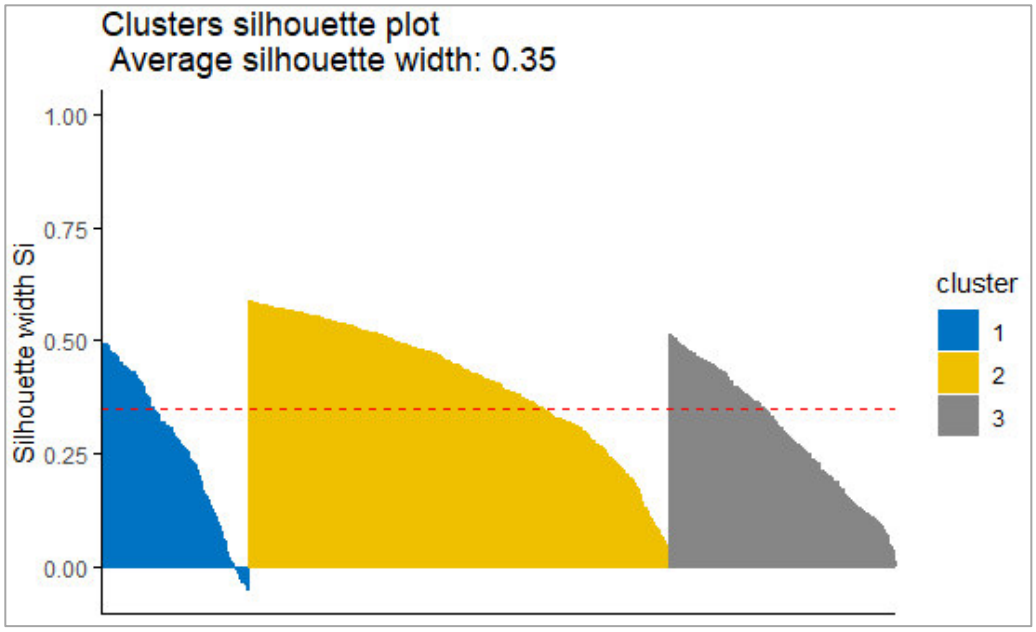


Fuente. Elaboración propia

Se observa que en la figura 12, el mapa perceptual que muestra la formación de tres segmentos, sin embargo, se logra notar que existen traslapes en el clúster 1 y 3, lo cual nos indica que el algoritmo a pesar de lograr agrupar estos clientes, no lo hace de forma tan discriminante.

Figura 13.

Método de validación Coeficiente de Silueta al Agrupamiento K-means



Fuente. Elaboración propia

En la figura 13, se obtiene información de los clústeres por separado como la cantidad de observaciones y el ancho promedio de silueta del agrupamiento que se muestra en la tabla 4.

Podemos observar que solo el clúster 2 tiene una estructura sólida con valor de silueta 0.41, mientras que el clúster 1 y 3 tienen un valor de 0.26 y 0.30 respectivamente.

Tabla 4

Validación Interna del Agrupamiento K-means

clúster	size	ave.sil.width
1	182	0.26
2	519	0.41
3	280	0.30

Fuente. Elaboración propia

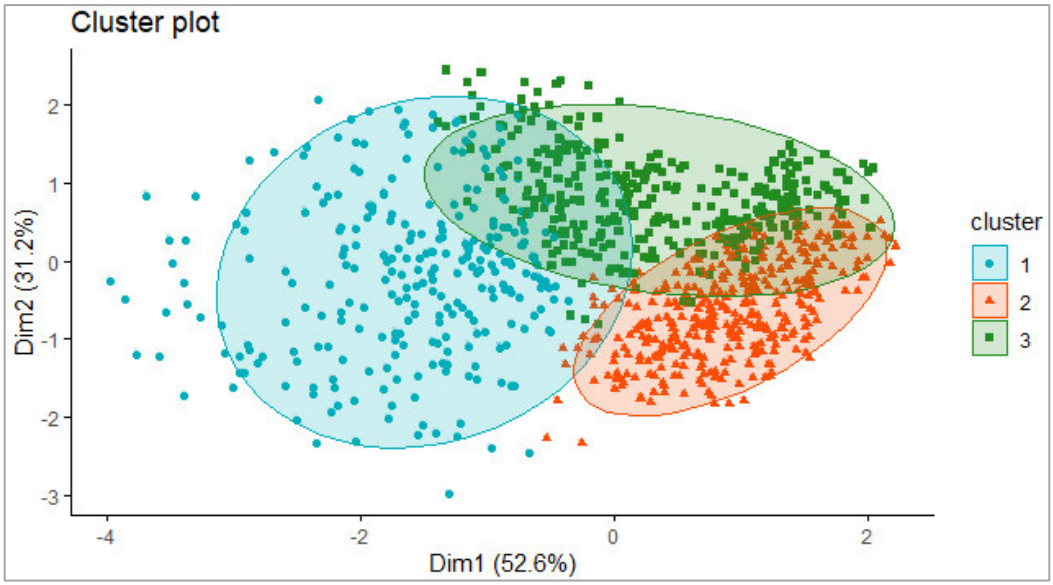
Se puede observar que varias muestras, en el conglomerado 1, tienen un coeficiente de silueta negativo. Lo que indica que esas observaciones podrían estar mal clasificadas.

- Aplicación de Análisis de conglomerados no Jerárquico: K-Medoids

El agrupamiento jerárquico permitió la identificación de tres segmentos, que se utilizan como información a priori para poder aplicar el algoritmo de segmentación no jerárquica K-Medoids, con un número de segmentos k = 3. A continuación se presenta el análisis no jerárquico de conglomerados.

Figura 14.

Mapa Perceptual de la Caracterización de Clientes de la marca TGI Fridays del agrupamiento K-medoids

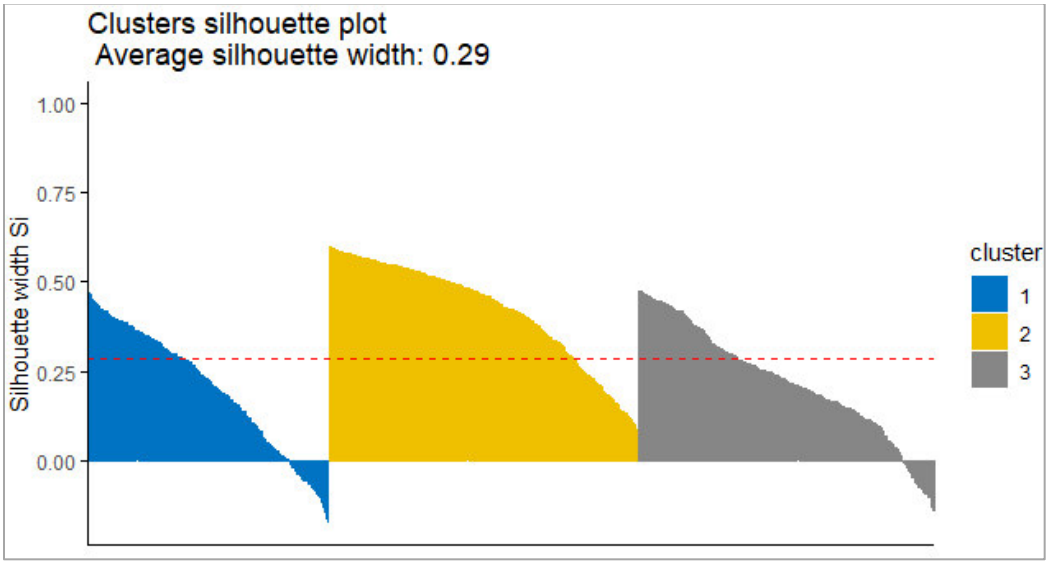


Fuente. Elaboración propia

Se observa que en la figura 14, el mapa perceptual que muestra la formación de tres segmentos, sin embargo, se logra notar que existen traslapes entre todos los clústeres, esto quiere decir que la segmentación no logra realizar una buena clasificación de los grupos.

Figura 15.

Método de validación Coeficiente de Silueta al Agrupamiento K-medoids



Fuente. Elaboración propia

En la figura 15, se obtiene información de los clústeres por separado como la cantidad de observaciones y el ancho promedio de silueta del agrupamiento que se muestra en la tabla 5.

Podemos observar que solo el clúster 2 tiene una estructura sólida con valor de silueta 0.42, mientras que el clúster 1 y 3 tienen un valor de 0.20 y 0.22 respectivamente.

Tabla 5

Validación Interna del Agrupamiento K-medoids

cluster	size	ave.sil.width
1	279	0.20
2	360	0.42
3	342	0.22

Fuente. Elaboración propia

Se puede observar que varias muestras, en el conglomerado 1 y 3, tienen un coeficiente de silueta negativo. Lo que indica que esas observaciones podrían estar mal clasificadas.

• Evaluación

Anteriormente se realizó la validación a cada algoritmo de agrupamiento, mediante el coeficiente de la silueta.

Comparando los resultados se descarta que el mejor algoritmo para la caracterización de clientes es el K-means con un K=3 y un ancho promedio de Silueta de 0.35, mientras que el algoritmo K-medoids tuvo un ancho promedio de silueta de 0.29.

En la siguiente tabla se procederá a identificar el perfil de los clientes.

Tabla 6

Promedio de Edad, Ganancia y Cantidad, según Segmentos

Clúster	Edad	Ganancia	Cantidad
1	39	97.4	14
2	27	27.4	8
3	43	65.7	6

Fuente. Elaboración propia

En la tabla 6, se puede observar el promedio de las variables que se utilizaron para el análisis de segmentación, teniendo así:

Clúster 1 (Alta demanda y alta ganancia):
Representa el 26% del total de la muestra (182 clientes), y se caracteriza por clientes con edad promedio regular (39 años), los cuales generan una alta ganancia (S/ 97.4 en promedio) y ordenan una alta cantidad de productos (14 en promedio).

Clúster 2 (Jóvenes y poca ganancia):
Representa el 41% del total de la muestra (519 clientes), y se caracteriza por clientes con edad promedio baja (27 años), los cuales generan una poca ganancia (S/ 27.4 en promedio) y ordenan una regular cantidad de productos (8 en promedio).

Clúster 3 (Demanda regular y mayor edad):

Representa el 30% del total de la muestra (280 clientes), y se caracteriza por clientes con edad promedio alta (43 años), los cuales generan una alta ganancia (S/ 65.7 en promedio) y ordenan una baja cantidad de productos (6 en promedio).

- Implementación

Una vez obtenido el perfilamiento de los clústeres, se procedió a presentar la implementación que se considera ejecutar el algoritmo en la base de datos de la marca TGI Fridays que cuenten con la información necesaria que se ha mostrado en el análisis (edad, ganancia, cantidad). Se presentó los resultados al equipo de Marketing, para que de esta forma puedan tomar decisiones en base a los perfiles de cada segmento con el objetivo de aumentar la fidelización de los clientes.

4. Conclusiones

Al concluir el presente trabajo correspondiente a la segmentación de clientes de la marca TGI Fridays se ha obtenido los siguientes resultados:

- Para el análisis de conglomerados jerárquico, se aplicaron los métodos de codo y silueta para obtener el número óptimo de clúster los cuales resultaron en cada método ser tres y con este resultado se procedió a aplicar los algoritmos K-means y K-medoids.
- Se consideró al algoritmo de agrupamiento K-means sobre el K-medoids, debido a que cuando se evaluó cada algoritmo de manera individual y se obtuvo que el ancho promedio de Silueta de K-means es 0.35, mientras que ancho promedio de Silueta de K-medoids es 0.29.
- Para el análisis de conglomerado K-means se consideró que sea tres el número óptimo de clúster, debido al análisis previo y se obtuvo que el porcentaje del clúster 1, 2, y 3 son 26%, 41% y 30% respectivamente.
- Con respecto a la caracterización de los clientes de la marca TGI Fridays fueron obtenidos mediante el algoritmo K-means y esto resulto lo siguiente:
    - Clúster 1 (Alta demanda y alta ganancia): está conformado por clientes con edad promedio regular de 39 años, los cuales generan alta ganancia que es de S/97.4 soles en promedio y consumen en promedio 14 productos.
    - Clúster 2 (Jóvenes y poca ganancia): está conformado por clientes con edad promedio baja de 27 años, los cuales generan una baja ganancia promedio de S/ 27.4 soles y consumen en promedio 8 productos.
    - Clúster 3 (Demanda regular y mayor edad): está conformado por clientes con edad promedio alta de 43 años, los cuales generan una ganancia promedio regular de S/ 65.7 soles y consumen en promedio 6 productos.

5. Recomendaciones

- Emplear otros métodos de conglomerados para poder observar el performance de los algoritmos.
- Poner en práctica estrategias de marketing basados en las características de cada clúster encontrado con el análisis de conglomerados K-means.
- Tomar en cuenta que los resultados obtenidos de la segmentación para poder mejorar y proponer a futuro una estrategia de ventas en la cadena Fridays.

6. Referencias Bibliográficas

Aldas, J., y Uriel, E. (2017). *Análisis Multivariante aplicado a R.* España: Paraninfo.

Azuela, J., Ochoa, M., y Ayup, J. (2019). Segmentación del comprador online en México: un estudio con base en la frecuencia de compra electrónica. *Científica Multidisciplinaria de Prospectiva, 26*(2), 2-14. doi:10.30878/ces.v26n2a1

Cuadros, A., Caicedo, C., y Jiménez, P. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41-51. doi:10.14483/22487638.12957

Elguera, R. (2018). *Segmentación de Clientes de un Casino Utilizando el Algoritmo Partición Alrededor de Medoides.* Lima: Universidad Nacional Agraria de la Molina.

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

Lawson, R., y Peter, J. (1990). New Index for Clustering Tendency and Its Application to Chemical Problems. *Journal of Chemical Information and Modeling, 30*(1), 36-41. doi:10.1021/ci00065a010

Mardia, K., Kent, J., y Bibby, J. (1995). *Multivariate Analysis.* London: Academic Press.

Rodriguez, E. (2015). *Unsupervised Learning With R. Birmingham.* United Kingdom: Packt Publishing.

Peña, D (2002). *Análisis de Datos Multivariantes.* Madrid: McGRAW-HILL

7. Anexos e Ilustraciones

- **Distancia de Mahalanobis.**

  Según Peña (2002), se define la distancia de Mahalanobis entre un punto y su vector de medias por

$$D_{s(x_i;x_j)} = \sqrt{(x_i - x_j)' S (x_i - x_j)}$$

- **Herramientas.**

  Para la ejecución del trabajo se utilizó el programa R versión 4.1, ya que cuenta con los paquetes necesarios como la biblioteca, “clusterder”, “factoextra”, “cluster”, “dendextend” entre otros.