



**Universidad
Zaragoza**



**Facultad de Ciencias
Universidad Zaragoza**

UNIVERSIDAD DE ZARAGOZA

MÁSTER EN MODELIZACIÓN E INVESTIGACIÓN MATEMÁTICA,
ESTADÍSTICA Y COMPUTACIÓN

TRABAJO DE FIN DE MÁSTER

**SISTEMA DE CLUSTERIZACIÓN
Y RECOMENDACIONES AUTOMÁTICAS
BASADO EN EL ETIQUETADO DE CLIENTES**

AUTOR: JUAN MANUEL PÉREZ GARCÍA DE CARELLÁN

DIRECTOR: DR. RICARDO LÓPEZ-RUIZ

CODIRECTOR: ALEJANDRO CARBONELL NAVARRO

Septiembre 2024

Año Académico 2023–2024

Quiero expresar mi agradecimiento a Ricardo por su apoyo y colaboración, y en especial a Álex y a Paula, por toda la ayuda y orientación que me han brindado a lo largo de la realización de este trabajo. Su acompañamiento y consejos han sido fundamentales, no solo en este proyecto, sino también en mis primeros pasos dentro del ámbito profesional.

Resumen

En un entorno empresarial cada vez más competitivo, ofrecer recomendaciones precisas y personalizadas a los clientes es fundamental para conseguir el máximo beneficio en las ventas de una empresa, además de optimizar la satisfacción del cliente. En los últimos años, el uso de técnicas de inteligencia artificial ha demostrado ser clave en este proceso, permitiendo a las empresas analizar grandes volúmenes de datos y mejorar la calidad de sus recomendaciones. Estas tecnologías no solo ayudan a comprender mejor los comportamientos de los consumidores, sino que también incrementan la eficiencia de las estrategias de marketing.

Este trabajo de fin de máster se centra en la aplicación de técnicas de aprendizaje no supervisado para la segmentación de los clientes de una empresa, con el objetivo final de mejorar las recomendaciones de productos. Los algoritmos de agrupamiento se aplican a un conjunto de datos que refleja el interés de los clientes en diversas categorías de productos.

Este trabajo contribuye al campo de la inteligencia artificial aplicada al marketing al proporcionar un enfoque sólido y reproducible para la segmentación de clientes, utilizando técnicas de aprendizaje no supervisado. Además, sienta las bases para futuros estudios que busquen mejorar las estrategias de recomendación, explorando nuevas formas de segmentación dinámica y ajustada a cambios en el comportamiento del consumidor a lo largo del tiempo.

Abstract

In an increasingly competitive business environment, providing accurate and personalized recommendations to customers is critical to maximizing a company's sales revenue, as well as optimizing customer satisfaction. In recent years, the use of artificial intelligence techniques has proven to be key in this process, enabling companies to analyze large volumes of data and improve the quality of their recommendations. These technologies not only help to better understand consumer behaviors, but also increase the efficiency of marketing strategies.

This master's thesis focuses on the application of unsupervised learning techniques for the segmentation of a company's customers, with the ultimate goal of improving product recommendations. Clustering algorithms are applied to a dataset reflecting customer interest in various product categories.

This work contributes to the field of artificial intelligence applied to marketing by providing a robust and reproducible approach to customer segmentation using unsupervised learning techniques. It also lays the groundwork for future studies that seek to improve recommendation strategies by exploring new forms of dynamic segmentation adjusted to changes in consumer behavior over time.

Índice general

Resumen	III
Abstract	V
1. Introducción	1
2. Marco teórico	5
2.1. Introducción al aprendizaje no supervisado	5
2.2. Reducción de la dimensionalidad	10
2.2.1. Análisis de Componentes Principales	10
2.3. Métodos de agrupamiento	11
2.3.1. K -medias	12
2.3.2. DBSCAN	13
2.3.3. Jerárquico aglomerativo	14
2.4. Métodos de evaluación	17
2.4.1. Método de la silueta	17
2.4.2. Índice de Dunn	18
2.4.3. Método del codo	18
3. Metodología de resolución	21
3.1. Introducción al problema	21
3.2. Preprocesamiento de los datos: creación de las matrices de conocimiento	23
3.2.1. <i>Knowledge_Items_Category</i>	23
3.2.2. <i>Knowledge_Customers_Category</i>	24
3.3. Clusterización de los clientes	27
3.3.1. Análisis de Componentes Principales	30
3.3.2. K -medias	31
3.3.3. DBSCAN	34
3.3.4. Jerárquico aglomerativo	36
3.4. Estrategia de recomendación	39
3.4.1. Popularidad dentro de los clústeres	39
3.4.2. Proximidad geográfica	40
3.4.3. Filtración colaborativa basada en usuarios	41
3.4.4. Venta cruzada y venta adicional	41
4. Conclusiones y trabajo futuro	43

Bibliografía	47
Anexos	49
Anexo A	49
Anexo B	50

Capítulo 1

Introducción

La inteligencia artificial (IA) se ha vuelto una parte fundamental de nuestra vida cotidiana, integrándose de manera fluida en múltiples aspectos de nuestras rutinas. Desde asistentes de voz que organizan nuestras agendas hasta recomendaciones personalizadas que moldean nuestras experiencias en línea, la IA se ha convertido en una presencia constante [1]. Esta incorporación natural de la IA es un reflejo de los enormes avances en la capacidad computacional y del impulso constante hacia el progreso científico.

Para resolver un problema en un ordenador, necesitamos un algoritmo, que es una secuencia de instrucciones que deben llevarse a cabo para transformar la entrada en salida. Por ejemplo, se puede idear un algoritmo para ordenar. La entrada es un conjunto de números y la salida es su lista ordenada. Para la misma tarea, puede haber varios algoritmos y nos puede interesar encontrar el más eficiente, que requiera el menor número de instrucciones o de memoria, o ambas cosas.

Sin embargo, para algunas tareas no disponemos de un algoritmo. Por ejemplo, para distinguir los correos spam de los legítimos. Sabemos cuál es la entrada: un documento de correo electrónico que, en el caso más sencillo, es un archivo de caracteres. Sabemos que la salida debe ser una respuesta afirmativa o negativa que indique si el mensaje es spam o no. Pero no sabemos cómo transformar la entrada en salida. Lo que puede considerarse spam cambia con el tiempo y de un individuo a otro.

Lo que nos falta en conocimientos lo compensamos con datos. Podemos compilar fácilmente miles de mensajes de ejemplo, algunos de los cuales sabemos que son spam, y lo que queremos es “aprender” lo que constituye el spam a partir de ellos. En otras palabras, nos gustaría que el ordenador (la máquina) extrajera automáticamente el algoritmo (aprendiera) para esta tarea. Este proceso constituye el llamado aprendizaje automático.

Los enfoques del aprendizaje automático se dividen tradicionalmente en tres grandes categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Estas categorías proporcionan un marco para comprender los distintos enfoques de resolución de problemas de aprendizaje y orientar el desarrollo de algoritmos y técnicas en el campo del aprendizaje automático.

En el aprendizaje supervisado, los modelos se entrenan con datos etiquetados, es decir, con ejemplos de entrada y salida conocidas, lo que permite al modelo aprender a predecir resultados específicos basándose en patrones previos. Por otro lado, el aprendizaje no supervisado trabaja con datos no etiquetados, buscando identificar patrones o estructuras ocultas sin la necesidad de ejemplos predefinidos.

En el caso del aprendizaje no supervisado, que es el que nos interesa en este trabajo, el objetivo principal es descubrir agrupaciones o relaciones dentro de los datos, lo que es útil en situaciones donde no se dispone de etiquetas o categorías previas. Las técnicas de reducción de la dimensionalidad y los distintos métodos de agrupamiento permiten extraer información relevante y segmentar los datos en grupos significativos, ayudando a comprender mejor su estructura subyacente sin intervención directa.

La segmentación de clientes ha sido un campo de estudio esencial en marketing y análisis de datos desde hace décadas. Los primeros métodos de segmentación de clientes se basaban principalmente en criterios demográficos o geográficos. Sin embargo, con la disponibilidad masiva de datos y los avances en inteligencia artificial y minería de datos, hoy en día es posible realizar segmentaciones mucho más precisas basadas en patrones de comportamiento de los clientes.

El proceso de recomendar productos a los clientes es crucial para mejorar la experiencia de compra y aumentar las ventas de una empresa. Para poder ofrecer recomendaciones personalizadas y efectivas, es necesario primero entender las diferencias entre los clientes, para así poder segmentarlos según sus intereses y comportamientos de compra. El aprendizaje no supervisado, abordado en la asignatura de minería de datos, cobra una especial relevancia en este contexto. Esta segmentación es clave para diseñar estrategias de recomendación más precisas y efectivas, alineando mejor los productos ofrecidos con las necesidades y preferencias de cada grupo de clientes.

El objetivo principal de este trabajo de fin de máster es identificar patrones o tendencias en los datos de compra de los clientes que permitan segmentarlos de manera efectiva en grupos con características comunes. Esto servirá como base para generar recomendaciones personalizadas. Para ello, se emplean métodos de agrupamiento y reducción de dimensionalidad para analizar los datos sin etiquetas previas, con el fin de descubrir agrupaciones que puedan dar lugar a estrategias de marketing más efectivas.

En cuanto a la metodología empleada, el primer paso ha sido recopilar los datos de facturación de la empresa, que contienen información sobre los clientes y sus compras, además de los datos que contienen información de los productos que se venden. Seguidamente, se ha limpiado y transformado la información, con el fin de construir una matriz que refleje los intereses de los clientes en las diferentes categorías de productos. Sobre esta matriz se han empleado varios algoritmos de agrupamiento para identificar grupos de clientes con comportamientos similares. Finalmente, se ha evaluado la calidad de los resultados de la segmentación mediante varios indicadores, lo que ha permitido determinar qué método produce las mejores agrupaciones. Todo el proceso de análisis se ha realizado utilizando el lenguaje de programación Python.

El trabajo se estructura en varios capítulos, cada uno enfocado en una parte clave del análisis. En el Capítulo 2, se presentan los conceptos matemáticos esenciales para abordar el problema de la segmentación de clientes. Se inicia con una introducción al aprendizaje no supervisado, explicando cómo identificar estructuras agrupables dentro de un conjunto de datos. A continuación, se profundiza en la reducción de dimensionalidad mediante el Análisis de Componentes Principales, técnica crucial para manejar conjuntos de datos con un gran número de variables. El capítulo concluye con la exposición de tres métodos distintos de agrupamiento (K -medias, clustering jerárquico y DBSCAN) y varios indicadores (coeficiente de silueta e índice de Dunn) para evaluar y comparar la calidad de los clústeres generados.

En el Capítulo 3, se aplican los métodos desarrollados previamente a un conjunto de datos reales de clientes de una empresa, con el fin de segmentarlos en grupos y diseñar estrategias de recomendación personalizadas basadas en los resultados obtenidos.

Por último, el Capítulo 4 está dedicado a la discusión y análisis de los resultados. Se destacan las fortalezas del proceso de segmentación, así como las limitaciones y desafíos encontrados durante su implementación. Además, se presentan sugerencias de mejora y posibles variantes del problema a considerar en estudios futuros.

Capítulo 2

Marco teórico

Este capítulo está dedicado a establecer las bases matemáticas que sustentan este trabajo. Se explorarán una de las principales técnicas de reducción de dimensionalidad, así como varios métodos de agrupamiento y criterios de evaluación para los clústeres generados.

En un problema típico de agrupamiento, se parte de un conjunto de n observaciones, cada una descrita por un conjunto de p variables y representada mediante un vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$. Este enfoque permite caracterizar cada observación a través de sus atributos, lo que facilita la identificación de patrones y la agrupación de datos similares de manera eficiente.

2.1. Introducción al aprendizaje no supervisado

Los métodos exploratorios básicos suelen ser de gran utilidad para comprender la complejidad de las relaciones entre múltiples variables [2]. Una técnica clave en este proceso es buscar patrones de agrupamiento “naturales” en los datos. Estos grupos pueden ayudar a evaluar de forma preliminar la dimensionalidad de los datos, detectar posibles valores atípicos y generar hipótesis interesantes sobre cómo se relacionan las variables entre sí.

Agrupar, o hacer clustering, difiere de los métodos de clasificación. La clasificación se refiere a un número conocido de grupos, y el objetivo es asignar nuevas observaciones a uno de estos grupos. El análisis de clústeres es una técnica más primitiva en el sentido de que no se hacen suposiciones sobre el número de grupos o la estructura del grupo. La agrupación se realiza sobre la base de similitudes o distancias. Las entradas requeridas son medidas de similitud o datos de los cuales se pueden calcular similitudes.

Para simplificar una estructura de grupos a partir de un conjunto de datos complejo, es necesario utilizar una medida de “proximidad” o “similitud”. Cuando se realizan agrupaciones, la proximidad suele indicarse por algún tipo de distancia, d , entre dos vectores que definen dos observaciones, $\mathbf{x}$ e $\mathbf{y}$. Estas distancias son funciones de las observaciones que cumplen ciertas propiedades:

- No negativas: $d(\mathbf{x}, \mathbf{y}) \geq 0$ y $d(\mathbf{x}, \mathbf{x}) = 0$
- Simétricas: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

- Cumplen la desigualdad triangular: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$

Las similitudes, s , son funciones que cumplen que:

- $s(\mathbf{x}, \mathbf{y}) \leq s_0$
- $s(\mathbf{x}, \mathbf{x}) = s_0$
- $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$

La elección de una medida de similitud a menudo conlleva una considerable subjetividad. Es crucial tener en cuenta la naturaleza de las variables (discretas, continuas, binarias), las escalas de medición y el conocimiento del tema. En la mayoría de las aplicaciones prácticas del análisis de clústeres, el investigador sabe lo suficiente sobre el problema para distinguir entre agrupaciones “buenas” y “malas”.

Para un conjunto de variables puramente cuantitativo, se emplea la distancia de Minkowski, la cual es una generalización de varias otras métricas de distancia:

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^p |x_i - y_i|^m \right]^{\frac{1}{m}}, \quad (2.1.1)$$

donde $m > 0$ se elige en función del peso que se desee dar a las diferencias más grandes y más pequeñas. Para $m = 1$, se denomina distancia de Manhattan (o “city-block”) y para $m = 2$ se convierte en la distancia euclídea, $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. Para $m \rightarrow +\infty$ la Ecuación (2.1.1) se denomina distancia de Chebychev, la cual es equivalente a $d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$, dando la máxima importancia a la mayor diferencia entre las componentes de las observaciones. La Figura 2.1.1 ilustra las distintas distancias de una manera visual.

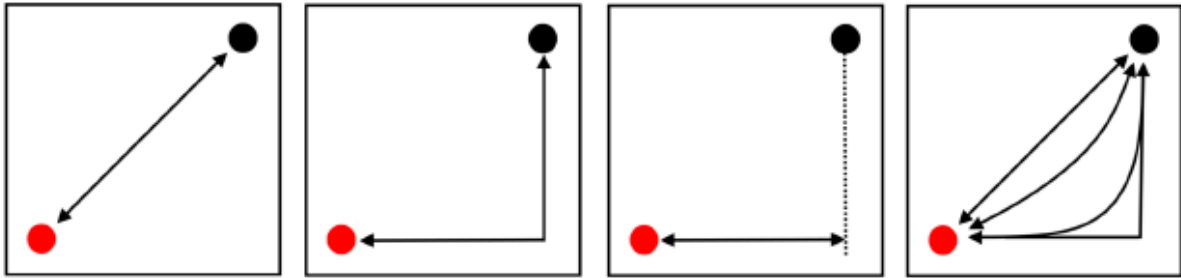


Figura 2.1.1: **Diferentes ejemplos de distancias. De izquierda a derecha: euclídea, Manhattan, Chebyshev y Minkowski para varios valores de m . Fuente: [3].**

La distancia euclídea, conocida como norma L_2 de un vector, es la más común de las métricas de Minkowski, y representa la distancia más corta para ir de un punto a otro. Es la elección de distancia empleada a lo largo del trabajo. Por otro lado, la distancia Manhattan o norma L_1 se emplea cuando se quieren calcular distancias entre puntos situados en una cuadrícula.

Cuando los ítems no pueden representarse mediante medidas p -dimensionales significativas, los pares de elementos se comparan a menudo en función de la presencia o ausencia de determinadas características. Los elementos similares tienen más características en común que los

no similares. La presencia o ausencia de una característica puede describirse matemáticamente introduciendo una variable binaria x_{ij} , la cual toma el valor 1 si la característica j está presente en el ítem i y el valor 0 si la característica está ausente. Dados dos ítems, i y k , se cumple:

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{si } x_{ij} = x_{kj} = 1 \text{ o } x_{ij} = x_{kj} = 0 \\ 1 & \text{si } x_{ij} \neq x_{kj} \end{cases} \quad (2.1.2)$$

y se le conoce como la distancia de Hamming, la cual proporciona un recuento del número de discordancias. Una distancia grande corresponde a muchos desajustes.

A menudo, las variables medidas en los individuos de estudio presentan una naturaleza diversa, incluyendo características numéricas continuas, categóricas y binarias. Para abordar el desafío de calcular distancias en conjuntos de datos de este tipo, se puede emplear la técnica de *one-hot encoding*, que transforma las variables categóricas en variables binarias, con la consiguiente pérdida de información. Otra alternativa consiste en “inventar” una matriz de distancias o similitudes, aprovechando el conocimiento que el científico de datos tiene sobre el experimento. Finalmente, se puede optar por la distancia de Gower [4], que ofrece una media ponderada de la contribución de cada variable a la distancia. En el contexto de nuestro trabajo, todas las variables son numéricas continuas, lo que significa que no tendremos que enfrentar este problema.

Muchas de las distancias son sensibles a las diferentes escalas en las que se miden las variables. Esta sensibilidad puede llevar a que ciertas magnitudes “dominen” la medida de distancia, haciendo que las variables con mayor rango de valores tengan una influencia desproporcionada en el cálculo de la distancia. Para mitigar este efecto y asegurar que todas las variables contribuyan de manera equitativa, se emplean técnicas de normalización o estandarización de datos.

Una técnica común es la estandarización, que consiste en tipificar las variables. Esto se logra restando la media y dividiendo por la desviación típica de cada variable, de modo que los nuevos valores resultantes tengan una media nula y una desviación típica igual a 1. Dada una variable aleatoria X con media μ y desviación típica σ , la variable aleatoria tipificada Z viene entonces dada por:

$$Z = \frac{X - \mu}{\sigma} \quad (2.1.3)$$

La transformación mostrada en la Ecuación (2.1.3) no solo iguala las escalas de las variables, sino que también facilita la comparación entre ellas, eliminando cualquier sesgo derivado de las diferencias en magnitud original.

Otra técnica utilizada es el escalado de las variables. Este método incluye dividir cada valor por su desviación típica, o bien restar el valor mínimo y dividir por el rango de la variable, obteniendo así valores que se sitúan entre 0 y 1. Este tipo de normalización es útil cuando se busca mantener la interpretación de los datos en términos de proporciones relativas.

No solo es importante estudiar cada variable por separado, sino también las relaciones entre sí. Para ello, se emplea la matriz de correlaciones, la cual mide cómo de relacionadas linealmente están las variables dos a dos. La medida más conocida para medir la correlación entre dos variables aleatorias X e Y con valores esperados μ_X y μ_Y y desviaciones típicas σ_X y σ_Y ,

respectivamente, es el coeficiente de correlación de Pearson $\rho_{X,Y}$ definido como:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \quad (2.1.4)$$

donde corr y cov se refieren a correlación y covarianza, respectivamente, y E es el operador valor esperado. El valor del coeficiente de correlación varía entre -1 y +1, correspondiendo estos valores extremos a una relación lineal inversa y directa perfecta, respectivamente, y los valores cercanos a 0 indican una menor relación lineal. Cabe mencionar que si X e Y son independientes, entonces $\rho_{X,Y} = 0$, pero lo recíproco no es cierto porque el coeficiente de correlación sólo detecta dependencias lineales entre dos variables.

Además, las técnicas de agrupamiento son particularmente sensibles a la presencia de datos atípicos (outliers) y datos faltantes. Los valores extremos pueden distorsionar los resultados del análisis, afectando negativamente la formación de grupos y la interpretación de los mismos. De igual manera, los datos faltantes suponen una pérdida de información crucial para el cálculo de medidas de distancia o similitud y pueden llevar a conclusiones incorrectas o a la exclusión de casos relevantes del análisis.

Por ello, es esencial llevar a cabo una depuración previa de los datos, identificando y tratando los outliers y datos faltantes antes de proceder con el agrupamiento. Esta depuración puede implicar la eliminación o transformación de los datos atípicos, y la imputación de valores ausentes sustituyéndolos por estimaciones razonables.

Una herramienta muy útil para llevar a cabo un estudio de la distribución de los datos de una variable determinada es mediante los diagramas de caja, los cuales representan gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, se muestran a simple vista la mediana y los cuartiles de los datos, y también pueden representarse sus valores atípicos.

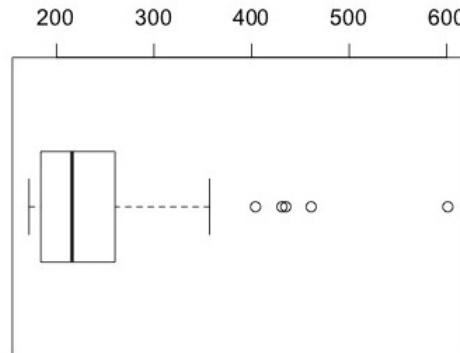


Figura 2.1.2: **Ejemplo de diagrama de caja. Fuente: [5].**

En la Figura 2.1.2., los extremos de la caja marcan el primer y tercer cuartil, Q_1 y Q_3 , y la línea gruesa dentro de la caja denota la mediana de los datos. Los extremos de los llamados bigotes, las líneas que se extienden desde la caja, se corresponden con los valores máximo y mínimo de la serie o hasta 1.5 veces el rango intercuartílico, definido como $RIC = Q_3 - Q_1$. Los puntos que caen más allá de los bigotes son los datos atípicos.

Finalmente, antes de aplicar una técnica de agrupamiento a un conjunto de datos, es crucial examinar si tiene sentido organizarlos en grupos, es decir, si existe alguna estructura de agrupamiento inherente en los datos.

Una de las maneras de comprobarlo es mediante el estadístico de Hopkins [6], el cual es una medida utilizada para evaluar la tendencia de un conjunto de datos a formar clústeres. Se parte de un conjunto X de n datos, y se toma una muestra aleatoria $\tilde{X}$ de $m \ll n$ puntos muestreados sin reemplazamiento de X . Además, se genera un conjunto Y de m datos aleatorios distribuidos uniformemente en el mismo intervalo definido por X . Con esto, se calcula la media de las u_i , las distancias mínimas (dada alguna métrica adecuada) desde cada $y_i \in Y$ a su vecino más cercano $x_j \in X$. También se calcula la media de las ω_i , las distancias mínimas desde cada $\tilde{x}_i \in \tilde{X} \subset X$ a su vecino más cercano $x_j \in X$, $\tilde{x}_i \neq x_j$. El estadístico de Hopkins H se define de la siguiente manera:

$$H = \frac{\sum_{i=1}^m u_i^p}{\sum_{i=1}^m u_i^p + \sum_{i=1}^m \omega_i^p}, \quad (2.1.5)$$

donde un valor de H cercano a 0.5 indica que los datos son aleatorios, y un valor cercano a 1 (superior a 0.7 se suele considerar suficiente) indica que el conjunto de datos es significativamente diferente de datos aleatorios uniformes y que contiene estructuras agrupables.

Otra manera de examinar la posible estructura de grupos subyacente es mediante el gráfico VAT [7] (siglas en inglés de *Visual Assessment of cluster Tendency*), el cual permite evaluar visualmente si los datos muestran indicios de algún tipo de agrupación. Para crear el gráfico VAT, se calcula una matriz de distancias (dada alguna métrica adecuada) entre todos los pares de observaciones y se reordena de forma que las observaciones similares estén situadas cerca unas de otras. Con esto, se representa gráficamente la matriz de distancias ordenada, empleando un gradiente de color para el valor de las distancias, y, si existen agrupaciones subyacentes en los datos, se forma un patrón de bloques cuadrados, tal y como se puede observar en la Figura 2.1.3.

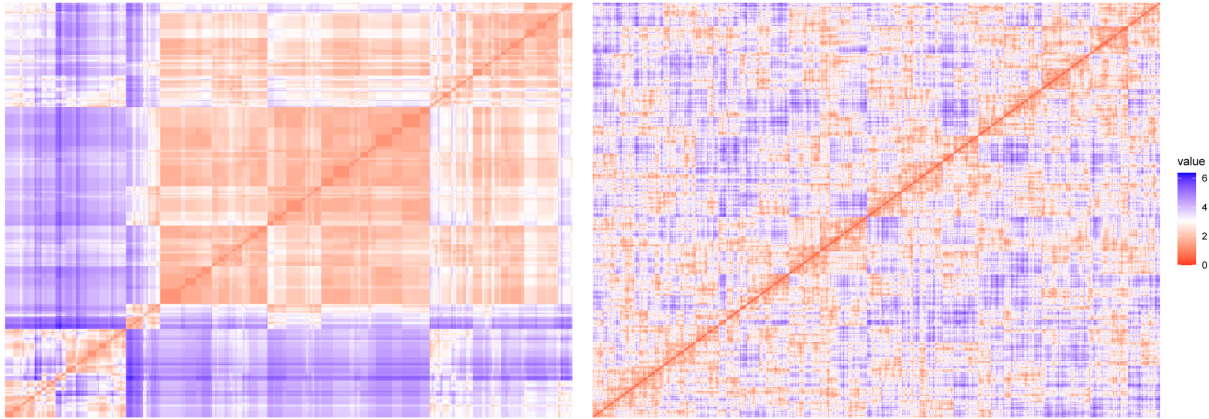


Figura 2.1.3: Comparación entre los gráficos VAT de unos datos con estructuras agrupadas (izquierda) y de unos datos aleatorios (derecha). Fuente: [8].

La capacidad de discernir patrones y agrupaciones es fundamental en el aprendizaje no supervisado, especialmente cuando se trabaja con conjuntos de datos de alta dimensionalidad. Mientras que la intuición humana puede identificar agrupaciones visibles en representaciones de baja dimensión, el desafío incrementa significativamente a medida que la cantidad de dimensiones aumenta. Esta complejidad hace esencial el uso de técnicas especializadas que faciliten la comprensión y el análisis de los datos. En este contexto, se vuelve imprescindible contar con métodos efectivos para reducir la dimensionalidad, preservando al mismo tiempo las relaciones estructurales clave entre los datos. Un procedimiento típico que permite representar observaciones de alta dimensionalidad en un menor número de dimensiones (pudiendo llegar a representarlos de manera visual en dos o tres dimensiones), de manera que los objetos similares estén, en cierto sentido, cercanos entre sí, se considera en la Sección 2.2.

2.2. Reducción de la dimensionalidad

La reducción de la dimensionalidad es un aspecto crucial en el análisis de datos, ya que simplifica modelos complejos sin perder la información esencial, facilitando así la interpretación y el análisis de grandes volúmenes de datos. En esta sección, se exploran uno de los métodos más destacados para este propósito: el Análisis de Componentes Principales (ACP). Este enfoque no solo permite visualizar los datos de manera más clara, sino que también mejora el rendimiento de los algoritmos de aprendizaje automático al eliminar el ruido y la redundancia en los conjuntos de datos, optimizando así los resultados obtenidos.

2.2.1. Análisis de Componentes Principales

El Análisis de Componentes Principales o ACP (*Principal Component Analysis* o PCA en inglés) [9] se utiliza para describir un conjunto de datos en términos de nuevas variables (llamadas componentes) no correlacionadas. Es una poderosa herramienta que permite visualizar y simplificar datos de alta dimensión preservando la información esencial. El ACP implementa una transformación lineal que establece un nuevo sistema de coordenadas para el conjunto original de datos. En este nuevo sistema, el eje que captura la mayor varianza de los datos se convierte en la primera componente principal. Asimismo, el eje que capta la segunda mayor varianza se convierte en la segunda componente principal, y este proceso se repite para las subsiguientes componentes principales. De esta manera, el ACP reordena la varianza de los datos de manera descendente a lo largo de los nuevos ejes coordenados.

Existen dos formas básicas de aplicar el ACP:

1. Método basado en la matriz de correlación, empleado cuando los datos no son dimensionalmente homogéneos o el orden de magnitud de las variables aleatorias no es el mismo.
2. Método basado en la matriz de covarianzas, usado cuando los datos son dimensionalmente homogéneos.

En lo que sigue, se explicará el método basado en la matriz de covarianzas, debido a la naturaleza del conjunto de datos que se empleará en el Capítulo 3.

Dada una muestra de n individuos para los cuales se tienen registros de p variables aleatorias (X_j) para cada uno, el objetivo es transformar el conjunto de datos inicial X en otro conjunto de datos Y de dimensión $n \times m$, con $m < p$, de manera que se pierda la menor información posible, utilizando para ello la matriz de covarianzas. Además, se tiene que cumplir que $m \leq \min\{n, p\}$.

El primer paso es estandarizar o tipificar las variables, empleando para ello la ecuación 2.1.3 y consiguiendo que las nuevas variables tengan media nula y desviación estándar igual a 1. El ACP se basa en la descomposición en vectores propios de la matriz de covarianzas definida como:

$$\text{cov}(X) = \frac{X^T X}{n - 1} \quad (2.2.1)$$

Con esto, se calculan los vectores propios $\mathbf{v}_\alpha$ y sus valores propios asociados λ_α resolviendo la ecuación:

$$\text{cov}(X)\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha, \quad (2.2.2)$$

lo cual equivale a resolver la ecuación característica $\det(\text{cov}(X) - \lambda I_p) = 0$, donde I_p se corresponde con la matriz identidad de dimensión $p \times p$.

Los autovalores λ_α indican cuánta varianza del conjunto de datos captura cada autovector de la matriz de covarianzas. Además se cumple que:

$$\sum_{\alpha=1}^p \lambda_\alpha = 1 \quad (2.2.3)$$

Finalmente, el conjunto de datos inicial X queda expresado en función de las m componentes principales de la siguiente manera:

$$X = \sum_{\alpha=1}^m \mathbf{t}_\alpha \mathbf{v}_\alpha^T + E, \quad (2.2.4)$$

donde los vectores $\mathbf{t}_\alpha$ se corresponden con las proyecciones de X sobre las direcciones definidas por los autovectores, $\mathbf{t}_\alpha = X\mathbf{v}_\alpha$, y la matriz E da cuenta del error que se produce al tomar un número de componentes principales menor que el número inicial de variables ($m < p$).

2.3. Métodos de agrupamiento

En la siguiente sección se analizarán diferentes técnicas de agrupamiento, esenciales para el análisis y segmentación de datos en diversas disciplinas. Los métodos de agrupamiento tienen como objetivo principal dividir un conjunto de datos en grupos homogéneos o clústeres, donde los elementos dentro de cada grupo son más similares entre sí que con los de otros grupos. Este proceso ayuda a descubrir patrones ocultos y estructuras subyacentes en los datos, facilitando así una mejor comprensión y toma de decisiones.

Los métodos de agrupamiento se clasifican en tres categorías principales según la estrategia utilizada para obtener los grupos: métodos jerárquicos, métodos de partición y modelos de mixturas. Los métodos jerárquicos pueden ser aglomerativos, uniendo sucesivamente las observaciones

más cercanas, o divisivos, comenzando con todos los datos en un único grupo y dividiéndolos sucesivamente, y el número de grupos se elige según las agrupaciones resultantes. Los métodos de partición dividen el conjunto de datos en un número predefinido de grupos, minimizando la distancia entre las observaciones del mismo grupo y su centro. Finalmente, los modelos de mixturas describen cada grupo mediante una distribución de probabilidad, comúnmente normal, y estiman los parámetros maximizando la verosimilitud.

2.3.1. K -medias

K -medias [10] es uno de los métodos de agrupamiento más utilizados en el análisis de datos. Es un método de partición cuyo objetivo es dividir el conjunto de puntos que definen los datos en K clústeres, donde $K \leq n$ es un número entero predefinido especificado por el investigador.

El algoritmo construye una partición de las observaciones en K conjuntos $\mathbf{C} = \{C_1, C_2, \dots, C_K\}$ con el objetivo de minimizar la suma de las distancias dentro de cada grupo:

$$\arg \min_{\mathbf{C}} \sum_{i=1}^K \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \boldsymbol{\mu}_i), \quad (2.3.1)$$

donde $\boldsymbol{\mu}_i$ se corresponde con el centroide del clúster i , el cual es la media de los puntos dentro de C_i .

El algoritmo más comúnmente empleado es el que emplea una técnica de refinamiento iterativo. Para ello, definiendo K centroides iniciales $\{\boldsymbol{\mu}_i^{(1)}\}_{i=1}^K$ (los cuales se suelen asignar aleatoriamente de entre las n observaciones), se alterna entre dos pasos:

1. **Asignación:** cada observación en la iteración t se asigna al grupo definido por el centroide que tiene a una menor distancia.

$$C_i^{(t)} = \{\mathbf{x}_l : d(\mathbf{x}_l, \boldsymbol{\mu}_i^{(t)}) \leq d(\mathbf{x}_l, \boldsymbol{\mu}_j^{(t)}) \forall 1 \leq j \leq K\} \quad (2.3.2)$$

2. **Actualización:** se calculan los nuevos centroides a partir de las asignaciones previamente hechas, calculando la media de los puntos de cada clúster.

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{\mathbf{x}_j \in C_i^{(t)}} \mathbf{x}_j \quad (2.3.3)$$

Se obtiene el resultado cuando el algoritmo ha convergido (las asignaciones no cambian) o se alcanza un número máximo de iteraciones. Además, la métrica que emplea el algoritmo K -medias para calcular las distancias es la euclídea, por tanto $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^p |x_i - y_i|^2} \equiv \|\mathbf{x} - \mathbf{y}\|$.

Como el algoritmo descrito es un algoritmo heurístico, el resultado obtenido puede no ser el óptimo global de la función objetivo definida en (2.3.1), y además este puede depender de la inicialización de los centroides. La convergencia a óptimos locales puede acarrear malos resultados. Como el algoritmo suele ser muy rápido (en la práctica, la mayoría de las veces el tiempo de ejecución es polinomial), es común ejecutar varias veces con diferentes condiciones de partida.

Por otro lado, la correcta elección del parámetro K es esencial para el buen funcionamiento del método sobre el conjunto de datos a estudiar. En futuros apartados se desarrollaran métodos que permiten hallar el valor de K idóneo para los datos a tratar.

Una limitación clave del K -medias es su modelo de agrupamiento. El concepto se basa en grupos esféricos que son separables de una forma en que el valor de la media converge hacia el centro del grupo. Se espera que los grupos tengan igual tamaño, por lo que la asignación al grupo más cercano es la asignación correcta.

Como con cualquier otro algoritmo de agrupamiento, el resultado de K -medias depende del conjunto de datos para satisfacer las necesidades del algoritmo, trabajando bien en algunos conjuntos de datos y fallando en otros.

2.3.2. DBSCAN

El agrupamiento espacial basado en densidad de aplicaciones con ruido o DBSCAN [10] (en inglés *Density-Based Spatial Clustering of Applications with Noise*) se clasifica dentro de los métodos de agrupamiento basados en densidad, los cuales se suelen comparar con los métodos de partición debido a que también dividen el conjunto de datos en distintos clústeres. El principio fundamental de DBSCAN es que los clústeres son regiones densas de puntos en el espacio de características, mientras que las áreas escasamente pobladas son consideradas como ruido.

Los puntos que definen las observaciones se clasifican en tres tipos: puntos núcleo (*core points*), puntos frontera (*border points*) y puntos de ruido (*noise points*). DBSCAN utiliza dos parámetros fundamentales para clasificar cada punto: ε , que define el radio de la vecindad alrededor de un punto, y $MinPts$, que establece el número mínimo de puntos necesarios a una distancia menor o igual que ε de un punto dado para que sea considerado un núcleo.

Un punto núcleo tiene al menos $MinPts$ puntos dentro de un radio ε , un punto frontera no es un núcleo pero está dentro del radio ε de un núcleo, y un punto de ruido no cumple con ninguna de estas condiciones.

El algoritmo DBSCAN comienza identificando los puntos núcleo. Para cada punto del conjunto de datos, se calcula cuántos puntos están dentro de un radio ε . Si este número es al menos $MinPts$, el punto se considera un núcleo. A partir de cada punto núcleo no visitado, se crea un nuevo clúster. Se añaden todos los puntos dentro de su radio ε al clúster, y para cada uno de estos puntos, si también es un núcleo, se repite el proceso de añadir sus vecinos al clúster. Este proceso continúa hasta que no se puedan añadir más puntos al clúster. Los puntos que no son núcleos pero están dentro del radio ε de algún núcleo se etiquetan como puntos frontera o de borde, mientras que los puntos que no cumplen con los criterios anteriores se etiquetan como puntos de ruido. La Figura 2.3.1 muestra un ejemplo de esta clasificación junto con la aplicación del método sobre un conjunto de datos.

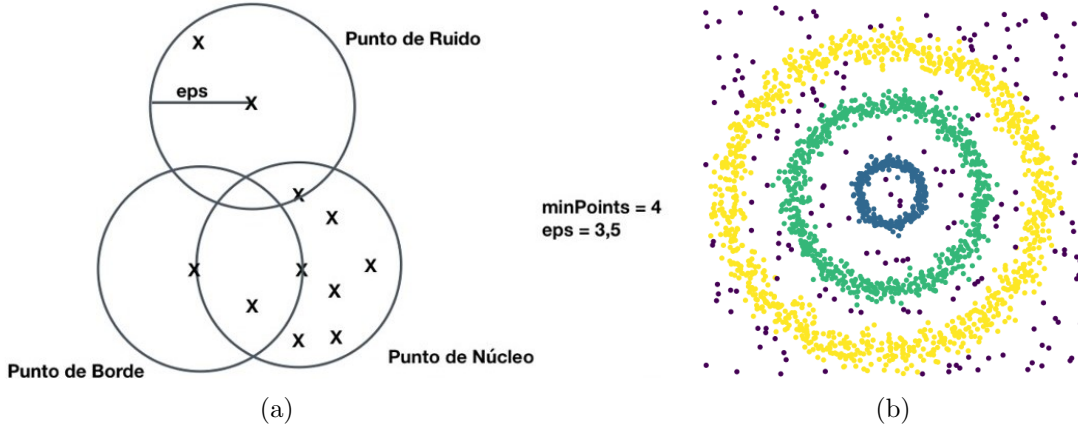


Figura 2.3.1: (a) Ejemplo de clasificación de tres puntos empleando el método de agrupamiento DBSCAN con $\varepsilon = 3.5$, $MinPts = 4$. Fuente: [11]. (b) Ejemplo de aplicación del método DBSCAN sobre un conjunto de puntos que definen tres circunferencias concéntricas y ruido. Fuente: [12].

DBSCAN tiene varias ventajas y desventajas. Entre sus ventajas, destaca su capacidad para encontrar clústeres con formas geométricas arbitrarias y su robustez al manejar ruido, ya que puede identificar puntos de ruido que no pertenecen a ningún cluster. Además, no requiere que se especifique el número de clústeres de antemano, a diferencia del método K -medias previamente explicado.

Por otro lado, elegir los valores adecuados para ε y $MinPts$ puede ser difícil y a menudo requiere de prueba y error. Además, no es un método enteramente determinista, ya que los puntos borde que son alcanzables desde más de un clúster pueden etiquetarse en cualquiera de estos. Por último, DBSCAN no funciona bien en conjuntos de datos con densidades muy variables, ya que no puede adaptarse a diferentes densidades de clústeres.

2.3.3. Jerárquico aglomerativo

En el agrupamiento jerárquico aglomerativo [10], cada observación comienza en un grupo formado por ella misma, y los pares de grupos se van uniendo conforme se sube en la jerarquía, hasta llegar a un grupo formado por la totalidad de las observaciones. El número de grupos se elige en función de las agrupaciones resultantes. Se ha optado por explicar el algoritmo jerárquico aglomerativo en lugar del divisivo debido a su menor complejidad computacional, su facilidad de implementación y su mayor popularidad en la literatura y herramientas de software.

El orden para realizar las uniones entre grupos en el agrupamiento jerárquico aglomerativo se elige a partir de una medida de similitud. En la mayoría de métodos de agrupamiento jerárquico, y en el caso que nos atañe, esta medida se consigue mediante el uso de una métrica apropiada y un criterio de enlace, el cual especifica la similitud de los conjuntos como una función de las distancias dos a dos entre observaciones en los conjuntos.

La elección de la métrica apropiada influenciará la forma de los grupos. Las métricas comúnmente usadas son las explicadas en la Sección 2.1.

Algunos criterios de enlace frecuentemente empleados entre dos conjuntos de observaciones A y B , dada una métrica d , se definen a continuación, y se muestran de manera visual en la Figura 2.3.2

- **Completo:** $\max\{d(\mathbf{a}, \mathbf{b}) : \mathbf{a} \in A, \mathbf{b} \in B\}$
- **Simple:** $\min\{d(\mathbf{a}, \mathbf{b}) : \mathbf{a} \in A, \mathbf{b} \in B\}$
- **Promedio:** $\frac{1}{|A||B|} \sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} d(\mathbf{a}, \mathbf{b})$
- **Método de Ward:** consiste en minimizar, a cada paso del algoritmo, la pérdida de información al unir dos grupos. Esta pérdida de información W se mide como la suma de las distancias (euclídeas) al cuadrado de cada observación $\mathbf{x}_{ij}$ a la media de su grupo $\boldsymbol{\mu}_j$:

$$W = \sum_{j=1}^K \sum_{i=1}^{n_j} \|\mathbf{x}_{ij} - \boldsymbol{\mu}_j\|^2, \quad (2.3.4)$$

donde K denota el número de grupos y n_j ($j = 1, \dots, K$) el tamaño de cada uno. De tal manera que al inicio del algoritmo, $W = 0$ (cada grupo contiene un único elemento) y a cada paso se realiza una unión que busca minimizar W .

El enlace completo tiende a formar grupos esféricos con diámetros similares, mientras que el enlace simple a menudo no distingue adecuadamente entre grupos que no están claramente separados, lo que resulta en clústeres alargados. Por otro lado, el enlace promedio es menos sensible a los valores extremos (outliers) y tiende a unir clústeres con pequeñas varianzas, al igual que el método de Ward, que genera grupos con formas y tamaños similares. No existe una regla fija para elegir el método de agrupamiento; es recomendable probar varios y optar por el que produzca la agrupación más coherente.

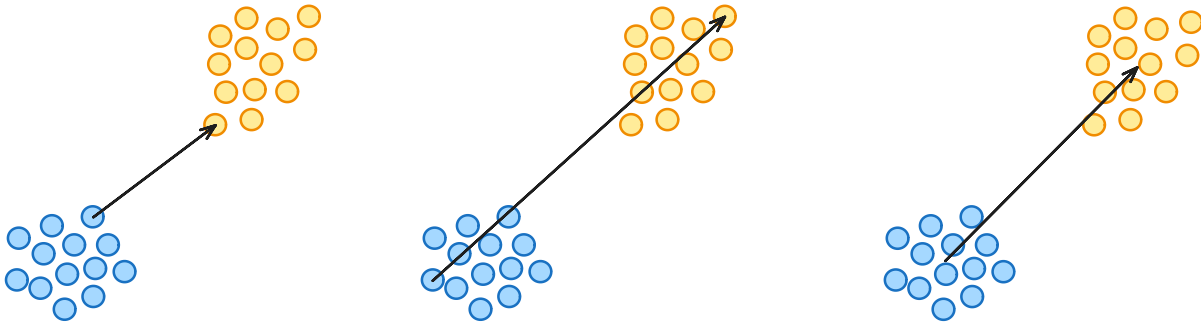


Figura 2.3.2: Representación visual de tres criterios de enlace. De izquierda a derecha: enlace simple, enlace completo y enlace promedio.

El resultado final se muestra mediante un dendrograma que muestra las sucesivas uniones de los clústeres en cada paso, así como las distancias en las que han ocurrido (en el eje de ordenadas). Cuanto más abajo en el dendrograma se realiza la unión, más parecidos son los

elementos que se han unido. El dendrograma proporciona una representación visual intuitiva de las relaciones jerárquicas entre las observaciones, lo cual puede ser útil para la interpretación del conjunto de datos de estudio. La Figura 2.3.3 muestra un ejemplo de un dendrograma.

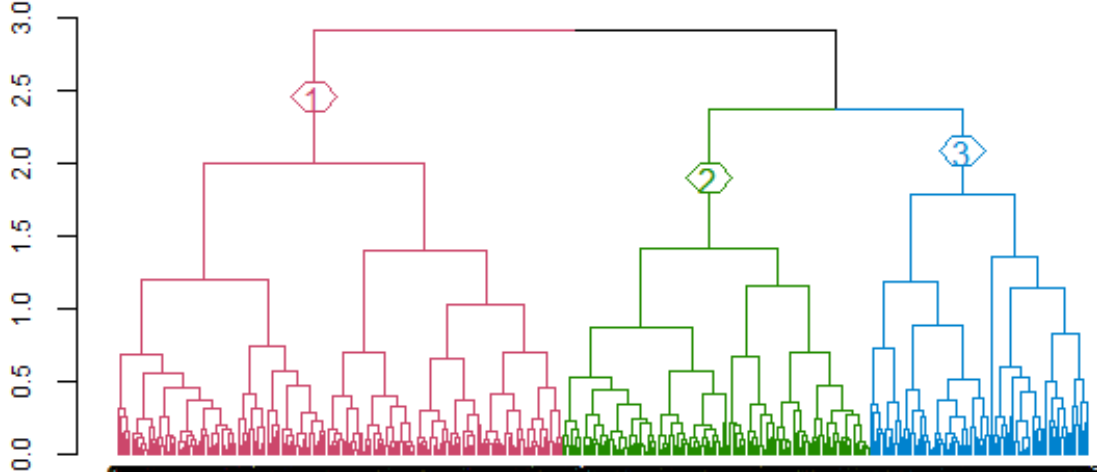


Figura 2.3.3: **Ejemplo de dendrograma, separación de los datos en 3 grupos distintos. El eje vertical mide la distancia de cada unión. Fuente: [13].**

Se comienza con n grupos formados por cada una de las observaciones y una matriz (simétrica de dimensión $n \times n$) de distancias o similitudes entre ellas. El algoritmo que permite realizar el agrupamiento jerárquico aglomerativo repite los siguientes dos pasos hasta que todos los elementos están agrupados en un solo clúster:

1. Se eligen los dos elementos más próximos en la matriz de distancias y se forma un clúster con ellos.
2. Se calculan las distancias del nuevo grupo al resto de los elementos teniendo en cuenta la métrica y el criterio de enlace escogidos y se reconstruye la matriz de distancias.

Una ventaja significativa del agrupamiento jerárquico aglomerativo, al igual que en el DBSCAN, es que no requiere especificar el número de clústeres K de antemano, a diferencia del algoritmo K -medias. Además, este método de agrupamiento permite observar la estructura jerárquica de los datos y decidir el número óptimo de clústeres examinando el dendrograma.

Por otro lado, el agrupamiento jerárquico aglomerativo también tiene sus deficiencias y limitaciones, como pueden ser la alta complejidad computacional y la sensibilidad a las métricas y criterios de enlace.

En cuanto a la complejidad computacional, la necesidad de calcular y actualizar una matriz de distancias de tamaño inicial $n \times n$ puede ser muy costosa en términos de tiempo y memoria, especialmente para conjuntos de datos grandes. Esto lo hace menos escalable comparado con otros métodos de agrupamiento como K -medias.

En lo que se refiere a la sensibilidad a las métricas y criterios de enlace, la forma y tamaño de los clústeres resultantes dependen en gran medida de la elección de la métrica de distancia y del criterio de enlace. Diferentes combinaciones pueden producir resultados muy distintos, y no siempre es evidente cuál es la mejor elección para un conjunto de datos dado. Aunque sí es cierto que la capacidad de definir diferentes criterios de enlace proporciona flexibilidad para adaptar el algoritmo a las características específicas del conjunto de datos y a los objetivos del análisis.

2.4. Métodos de evaluación

La validación de la agrupación obtenida sigue siendo un problema no supervisado, salvo que se disponga de al menos un subconjunto de datos correctamente clasificados de antemano. Para llevar a cabo la evaluación de los resultados, se emplean mayoritariamente el método de la silueta y el índice de Dunn.

Además, para métodos de agrupamiento de partición, los cuales requieren un número predefinido de grupos en los que dividir el conjunto de datos, se suele emplear el método del codo para evaluar cuál es el número óptimo de clústeres.

2.4.1. Método de la silueta

El coeficiente de silueta [10] es una medida utilizada para evaluar la calidad de un agrupamiento. Este índice considera tanto la cohesión dentro del clúster como la separación entre clústeres. Dada una observación $\mathbf{x}$ perteneciente al clúster C_i , el coeficiente de silueta $s(\mathbf{x})$ se define de la siguiente manera:

$$s(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max(a(\mathbf{x}), b(\mathbf{x}))}, \quad (2.4.1)$$

donde

$$a(\mathbf{x}) = \frac{1}{|C_i| - 1} \sum_{\mathbf{y} \in C_i, \mathbf{y} \neq \mathbf{x}} d(\mathbf{x}, \mathbf{y}) \quad (2.4.2)$$

define la distancia media entre el punto $\mathbf{x}$ el resto de puntos de su mismo clúster C_i , y donde

$$b(\mathbf{x}) = \min_{\mathbf{y} \neq \mathbf{x}} \left(\frac{1}{|C_j|} \sum_{\mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}) \right) \quad (2.4.3)$$

se corresponde con la distancia media entre la observación $\mathbf{x}$ y todas aquellas observaciones pertenecientes al clúster más cercano a C_i .

El coeficiente de silueta de una agrupación dada, S , viene dado por la media aritmética de los coeficientes de silueta de cada una de las n observaciones, el cual proporciona una medida de la calidad global del agrupamiento.

$$S = \frac{1}{n} \sum_{\mathbf{x}} s(\mathbf{x}) \quad (2.4.4)$$

De la anterior ecuación se tiene que $-1 \leq S \leq 1$. Para $S \rightarrow 1$, se cumple que $b(\mathbf{x}) \gg a(\mathbf{x})$, $\forall \mathbf{x}$, lo cual indica que la distancia entre clústeres es mucho mayor que el tamaño de los grupos, por tanto señalando una buena agrupación. En el otro extremo, para $S \rightarrow -1$, se cumple que $b(\mathbf{x}) \ll a(\mathbf{x})$, $\forall \mathbf{x}$, indicando una mala agrupación. Por tanto, el método de la silueta se empleará para comparar modelos, prefiriendo aquellos que aporten mayores valores de S .

2.4.2. Índice de Dunn

El índice de Dunn [14] se calcula como el cociente entre la separación mínima inter-grupo, calculada como el mínimo de las distancias de cada observación a la del resto de los grupos, y la separación máxima intra-grupo, calculada como la máxima distancia entre las observaciones de cada grupo. Dados K clústeres, el índice de Dunn D viene dado por:

$$D = \frac{\min_{1 \leq i < j \leq K} d(i, j)}{\max_{1 \leq l \leq K} d'(l)}, \quad (2.4.5)$$

donde $d(i, j)$ representa la distancia entre los grupos i y j , y $d'(l)$ mide la distancia dentro del grupo l .

El criterio del índice de Dunn busca grupos con alta semejanza dentro de un mismo grupo y baja semejanza entre grupos distintos, consiguiendo identificar clústeres densos y bien separados. En la práctica, se prefieren aquellos grupos que dan índices de Dunn más altos.

2.4.3. Método del codo

El método del codo proporciona una forma intuitiva y sencilla de determinar el número adecuado de clústeres para un conjunto de datos. Esta técnica se basa en la relación entre el número de clústeres y una métrica de evaluación de la calidad del agrupamiento, generalmente la suma de las distancias al cuadrado de cada punto al centro su clúster, también conocido como inercia. Esta expresión se puede ver en la Ecuación (2.3.4), ya que es la misma métrica empleada en el método de Ward.

El método del codo consiste en representar gráficamente la inercia en función del número de clústeres, variando este número entre 1 y N , donde N es un valor especificado por el investigador. En nuestro trabajo $N = 10$. En esta gráfica, se espera observar un cambio notable en la evolución de la inercia, de modo que la curva tome la forma de un brazo doblado, similar a un codo. El punto donde ocurre este cambio brusco en la inercia señala el número óptimo de clústeres $K \leq N$ para el conjunto de datos analizado. Más específicamente, K se identifica en el valor del eje de abscisas donde la pendiente de la curva de inercia disminuye de manera significativa. Este descenso abrupto en la tasa de mejora sugiere que añadir más clústeres no aporta una reducción considerable en la dispersión dentro de estos, lo que indica que a partir de ese punto, agregar más clústeres no mejora sustancialmente la cohesión del agrupamiento.

Un ejemplo de esta gráfica se muestra en la Figura 2.4.1, donde se aprecia claramente el punto de inflexión en $K = 3$, marcando el “codo” característico de este método.

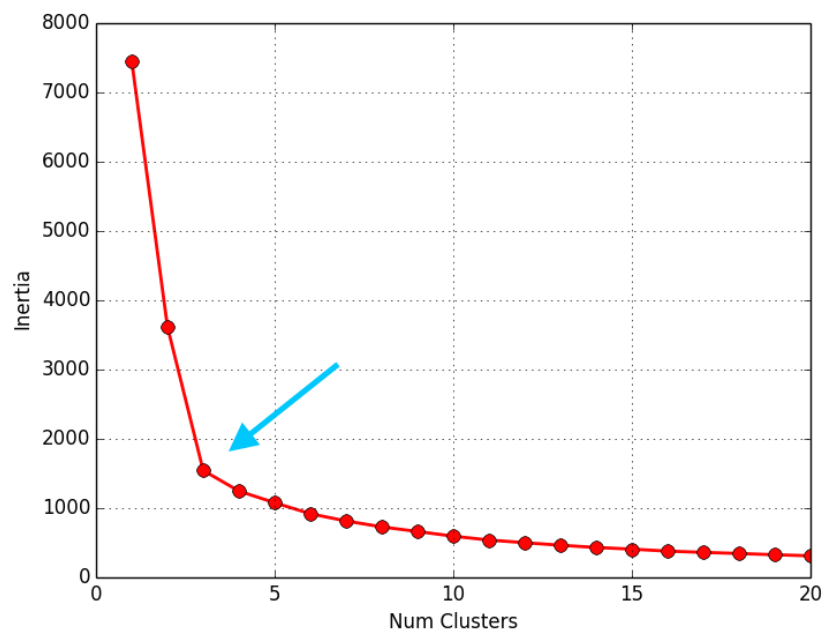


Figura 2.4.1: Representación de la inercia frente al número de clústeres para un conjunto de datos, indicando con una flecha dónde se encuentra el codo. Fuente: [15].

Capítulo 3

Metodología de resolución

En el capítulo anterior se han presentado las diversas técnicas necesarias para llevar a cabo el análisis de datos requerido en este trabajo. A continuación, en el próximo capítulo se detallará el preprocesamiento y la transformación de unos datos en información útil, con el fin de agrupar a los clientes de una empresa en distintos clústeres. El objetivo principal es establecer estrategias de recomendación más especializadas, basadas en los resultados obtenidos de dicha agrupación.

A lo largo de este trabajo se ha empleado el lenguaje de programación Python, utilizando diversas librerías especializadas que han facilitado el análisis de datos y la implementación de técnicas de agrupamiento. Por un lado, *pandas* y *numpy* se han usado para la manipulación y procesamiento eficiente de los datos. Para la visualización, se ha recurrido principalmente a *matplotlib* y *seaborn*, que han permitido generar gráficos claros y representativos del análisis realizado. Asimismo, las técnicas de clusterización y análisis de datos se han implementado mediante las librerías *scikit-learn* y *scipy*. Todo el código se puede encontrar en el repositorio de GitHub mostrado en el [Anexo A](#).

3.1. Introducción al problema

El conjunto de datos empleado para la realización del trabajo ha sido proporcionado por la empresa Fersa, una multinacional española con más de 50 años de experiencia en el diseño, desarrollo, fabricación y distribución de soluciones completas de rodamientos para el sector de la automoción a nivel mundial [16].

El estudio se lleva a cabo bajo la iniciativa *SARA* (Sistema de Aprendizaje para Recomendaciones Automáticas). *SARA* es un proyecto en desarrollo por parte del departamento de Inteligencia Artificial del área de I+D de Fersa, diseñado para asistir a los comerciales en la tarea de ofrecer nuevos productos a sus clientes. El núcleo del proyecto reside en comprender las dinámicas y características de los clientes, con el fin de categorizarlos de manera que se facilite su agrupación por intereses o características similares, y explorar las conexiones que permitan una mejor comprensión de sus necesidades.

El propósito de *SARA* es elaborar recomendaciones personalizadas para los clientes utilizando los registros de compras de estos (base de datos de facturas) y detalles sobre los productos

ofrecidos (base de datos de aplicaciones). A partir de estos datos, se pretende generar una propuesta que anticipa el interés de los clientes en productos que aún no han adquirido, mejorando así la precisión y relevancia de las sugerencias comerciales. A continuación, se detallan los dos conjuntos de datos de entrada que sustentan el proceso de generación de recomendaciones.

1. **FACTURAS:** contiene el histórico de facturas desde 2019 hasta 2024, con los siguientes atributos:

- ITEMID: código del producto que se ha comprado.
- CUSTOMERID: código del cliente que ha realizado la compra.
- INVOICEID: código de la factura.
- INVOICEDATE: fecha de la compra.
- QTY: cantidad de unidades que se han comprado del producto ITEMID.
- UNITPRICE: precio unitario del producto ITEMID que se ha comprado.
- AMOUNT: el valor de QTY multiplicado por el valor de UNITPRICE. Es el precio de la compra en unidades de la moneda empleada para el pago.
- CURRENCYCODE: código de la moneda empleada para el pago.
- EXCHANGERATE_EUR: tipo de cambio a euros.
- AMOUNT_EUR: el valor de AMOUNT multiplicado por el valor de EXCHANGERATE_EUR. Es el precio de la compra en euros.
- COUNTRY: país del cliente que ha realizado la compra.
- CONTINENT: continente del cliente que ha realizado la compra.
- LINEOFBUSINESS: línea de negocio del cliente.

2. **APLICACIONES:** incluye información relevante de los productos que se venden, de los cuales se conocen los siguientes atributos:

- ITEMID: código del producto que se vende.
- MODEL: nombre concreto de la aplicación donde va montado el producto.
- APPLICATION: tipo de aplicación en la que el rodamiento va montado (eje, diferencial, caja de cambios). Hay 5 aplicaciones distintas.
- BRAND: marca donde se monta el producto. Hay 57 marcas distintas.
- SECTOR: sector comercial donde se vende el producto.
- VEHICLE: tipo de vehículo donde se monta el producto.

Cabe mencionar que no todos aquellos productos que salen en las facturas también salen en las aplicaciones, y viceversa. En el primer caso, un producto que salga en facturas y no en aplicaciones se corresponde con un cliente que ha comprado un producto el cual no se ha registrado o no se conoce su información en una aplicación. En el segundo caso, un producto que salga en aplicaciones y no en facturas se corresponde con uno del cual se conoce su información, pero sin embargo no se tienen registros de compras.

Por otro lado, en el conjunto de datos de aplicaciones, el único valor del atributo SECTOR es ‘Commercial’, y en el atributo VEHICLE el único valor es ‘Truck’. Esto se debe a que Fersa en España está especializado en rodamientos de camión.

3.2. Preprocesamiento de los datos: creación de las matrices de conocimiento

Una vez introducidos los conjuntos de datos a tratar, el primer objetivo de cara a la creación de *SARA* es construir una matriz de conocimiento que vincule a cada cliente, listado en las facturas, con su interés por las diversas categorías de los productos incluidos en el conjunto de datos de aplicaciones. Las categorías se corresponden con los distintos atributos del conjunto de datos de aplicaciones. Este interés se cuantificará en una escala de 0 a 1 y servirá como métrica fundamental para llevar a cabo la clusterización y determinar las distancias y similitudes entre los clientes. Por tanto, el primer paso es describir cómo se construye la matriz de conocimiento, la cual se llama *Knowledge_Customers_Category*.

3.2.1. *Knowledge_Items_Category*

El primer paso para crear la matriz de conocimiento de los clientes en las distintas marcas, sectores, tipos de vehículo y aplicaciones donde se montan los rodamientos que compran es la realización de la matriz *Knowledge_Items_Category*. Esta, como su nombre indica, muestra el conocimiento de cada producto sobre las distintas categorías del conjunto de datos de aplicaciones.

La manera de crearla es sencilla: se parte de una matriz llena de ceros, de l filas (número de productos de los cuales se tiene información) y p columnas, tantas como distintas categorías haya para los productos (en nuestro ejemplo concreto $p = 64$: 1 columna para el sector comercial, 1 para el tipo de vehículo, 5 para la aplicación donde se monta y 57 para las distintas marcas).

Con esto, para cada producto denominado por Item_i , $i = 1, \dots, l$, se escriben unos en las columnas que se corresponden con las categorías que ese producto tiene en los datos de aplicaciones. Finalmente, se normaliza mediante el método unidad, dividiendo cada valor de cada fila por la suma de celdas con valor 1 para cada subgrupo de categorías. Esta suma, para cada fila $i = 1, \dots, l$, se denota por $s_{x,i}$, con $x = a$ para el subgrupo de categorías referente a la aplicación donde el producto va montado, y $x = b$ se refiere a las marcas que emplean dicho producto. En el caso de las categorías de sector y de vehículo, todas las filas están llenas de unos, ya que, en el conjunto de datos de aplicaciones, solo se muestran registros de productos que van montados en camión (truck), y pertenecen al sector comercial.

En el Cuadro 3.2.1 se muestra un ejemplo que ilustra cómo se estructura y organiza la matriz de conocimiento previamente descrita.

ITEMID	SECTOR (Commercial)	VEHICLE (Truck)	APP. 1	...	APP. 5	BRAND 1	...	BRAND 57
Item ₁	1	1	0	...	$1/s_{a,1}$	0	...	$1/s_{b,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	...	$\vdots$
Item _l	1	1	0	...	0	$1/s_{b,l}$	...	$1/s_{b,l}$

Cuadro 3.2.1: **Ejemplo de la matriz de conocimiento de los productos sobre las distintas categorías, *Knowledge_Items_Category*.**

De esta forma, los ítems en la matriz *Knowledge_Items_Category* que presentan valores no nulos en numerosas celdas se corresponden con productos generales que se integran en diversas aplicaciones y marcas. En cambio, aquellos ítems con una alta cantidad de valores nulos y solo unos pocos no nulos representan productos muy específicos. Por este motivo, estos últimos reciben una puntuación más alta en términos de conocimiento especializado.

3.2.2. *Knowledge_Customers_Category*

Una vez obtenida la matriz que ofrece información de los distintos productos en las distintas categorías, se pretende obtener una matriz con las mismas columnas, pero en la cual las filas se correspondan con los clientes, denotados por $Cust_i$, $i = 1, \dots, n$, donde n es el número de clientes de los cuales se tiene información de compras en los datos de facturas.

El contenido de la matriz buscada estará moldeado por el momento en que se realicen las compras y la cantidad de producto adquirido. Así, esta matriz capturará de manera más completa y detallada la dinámica de compra, reflejando no solo el qué, sino también el cuándo y el cuánto.

Los datos de las facturas abarcan un periodo desde 2019 hasta 2024. Para la creación de la matriz objetivo, el proceso de aprendizaje se realiza de manera anual, otorgando mayor relevancia a la información más reciente. Esto se debe a que las compras realizadas en años anteriores pueden no ser tan influyentes para predecir los intereses actuales de los clientes, en comparación con las transacciones más cercanas en el tiempo.

Para capturar de manera más fiel las preferencias de los clientes, es fundamental que el interés en una categoría crezca si, en el transcurso de un año, el cliente adquiere algún producto que pertenezca a esa categoría, y disminuya en caso contrario. Además, este crecimiento debe ser proporcional a la cantidad de producto adquirido. Con este fin, se define Q_i , la evolución del interés en una categoría por cada unidad de producto comprada en un mismo año:

$$Q_i = (1 - \alpha)Q_{i-1} + \alpha, \quad (3.2.1)$$

donde $0 < \alpha < 1$ representa la tasa de aprendizaje (learning rate). Para valores de α cercanos a 0, el interés en las categorías crece de manera pausada, requiriéndose una cantidad considerable de compras para alcanzar intereses cercanos a 1, que es el máximo que puede alcanzar el interés en cada categoría. Por el contrario, cuando α se aproxima a 1, el interés se incrementa rápidamente, reflejando con inmediatez la afinidad del cliente por una categoría tras cada compra.

Fijado un valor para α y un interés inicial Q_0 , se despeja la siguiente relación para Q_m , el interés en una categoría tras la m -ésima compra en un mismo año:

$$\begin{aligned}
Q_1 &= (1 - \alpha)Q_0 + \alpha \\
Q_2 &= (1 - \alpha)Q_1 + \alpha \\
&= (1 - \alpha)^2 Q_0 + (1 - \alpha)\alpha + \alpha \\
&\vdots \\
Q_m &= (1 - \alpha)Q_{m-1} + \alpha \\
&= (1 - \alpha)^m Q_0 + \sum_{k=0}^{m-1} (1 - \alpha)^k \cdot \alpha \\
&= (1 - \alpha)^m Q_0 + \alpha \cdot \sum_{k=0}^{m-1} (1 - \alpha)^k
\end{aligned} \tag{3.2.2}$$

Del anterior resultado se deduce, empleando la fórmula de la suma de los $n + 1$ primeros términos de la serie geométrica, para $|r| < 1$:

$$\sum_{k=0}^n r^k = \frac{1 - r^{n+1}}{1 - r}, \tag{3.2.3}$$

el siguiente resultado para Q_m :

$$\begin{aligned}
Q_m &= (1 - \alpha)^m Q_0 + \alpha \cdot \frac{1 - (1 - \alpha)^m}{1 - (1 - \alpha)} \\
Q_m &= (1 - \alpha)^m Q_0 + 1 - (1 - \alpha)^m \Rightarrow \boxed{Q_m = (1 - \alpha)^m (Q_0 - 1) + 1}
\end{aligned} \tag{3.2.4}$$

Tras el desarrollo anterior, estamos en condiciones de explicar cómo se obtiene la matriz de intereses de los clientes en las categorías, *Knowledge_Customers_Category*.

En primer lugar, se construye, para cada año k considerado ($k = 2019, \dots, 2024$), una matriz $\mathbf{A}^k$ de n filas (una para cada cliente) y l columnas (una para cada ítem), que contenga en cada celda $\mathbf{A}_{ij}^k$ la cantidad de producto j que ha comprado el cliente i en el año k .

Además, como la matriz *Knowledge_Items_Category* contiene valores x en el intervalo $[0, 1]$, se transforma atendiendo a la siguiente norma:

- Si $x = 1$, entonces se modifica a $x = 2 \cdot \beta$.
- Si $x < 1$, entonces se modifica a $x = 0.05 \cdot \beta \cdot x$,

con $0 < \beta < 0,5$. El valor de este parámetro depende directamente de la frecuencia con la que se ejecuta el proceso de aprendizaje, siendo mayor cuanto menor es dicha frecuencia. Para ajustar el valor de β , es necesario realizar el entrenamiento que se describe posteriormente, buscando un valor óptimo que evite alcanzar el máximo interés en cualquiera de las categorías (un valor

máximo de 1). De esta forma, se previene el sobreajuste. En este trabajo, el valor seleccionado para la frecuencia actual de un año es $\beta = 0.02$.

Atendiendo a la transformación anteriormente descrita, se otorga mayor valor a los ítems muy específicos, mientras que se reduce la importancia de aquellos que son menos específicos. El nuevo valor de $x \in [0, 1)$ se corresponde con la tasa de aprendizaje descrita en la Ecuación (3.2.1).

Tras crear cada una de las matrices $\mathbf{A}^k$ y transformar la matriz que contiene la tasa de aprendizaje x para las categorías de cada ítem, el proceso de aprendizaje para cada cliente en un año determinado sigue los siguientes pasos:

1. De la matriz $\mathbf{A}^k$ correspondiente, se selecciona la fila que contiene la información de compra del cliente en estudio, extrayendo únicamente aquellos elementos no nulos que reflejan la cantidad de producto comprado ese año.
2. Se transforma el conocimiento de los ítems que ha comprado el cliente ese año en conocimiento sobre el propio cliente. Para ello, se itera sobre cada uno de los ítems seleccionados en el paso 1. Para cada una de las categorías, se emplea la fórmula definida anteriormente en la Ecuación (3.2.4) para calcular el nuevo interés en dicha categoría, Q_m :

$$Q_m = (1 - x)^m(Q_0 - 1) + 1, \quad (3.2.5)$$

donde x denota el valor transformado de la matriz *Knowledge_Items_Category* (tasa de aprendizaje), m se corresponde con la cantidad total de ítems que ha comprado el cliente ese año con interés no nulo en la categoría de estudio, y Q_0 denota el interés inicial del cliente en esa categoría. $Q_0 = 0$ para el primer año ($k = 2019$), y para años posteriores Q_0 se corresponde con el valor del interés final del cliente (devaluado por un factor Δ) obtenido en el año anterior para la categoría de estudio.

3. En caso de que no se esté tratando con el último año ($k = 2024$), se devalúa la información obtenida en el paso anterior multiplicando la matriz por un factor $\Delta = 0.2$. De esta manera, la información más antigua tiene menor importancia que la más reciente.

Este proceso se aplica de manera general para todos los años, clientes y categorías consideradas, construyendo así la matriz final deseada, *Knowledge_Customers_Category*, cuyo esquema general se muestra en el Cuadro 3.2.2. Esta matriz será la que muestre el interés de cada cliente en las distintas categorías y la que se utilice para llevar a cabo el proceso de clustering.

CUSTOMERID	SECTOR (Commercial)	VEHICLE (Truck)	APP. 1	...	APP. 5	BR. 1	...	BR. 57
Cust ₁				...			...	
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮
Cust _n				...			...	

Cuadro 3.2.2: Ejemplo de la matriz de conocimiento de los clientes sobre las distintas categorías, *Knowledge_Customers_Category*.

3.3. Clusterización de los clientes

Después de explicar cómo obtener *Knowledge_Customers_Category*, la matriz que asocia a cada cliente con varios intereses en las distintas categorías de productos, estamos listos para iniciar el proceso de clusterización partiendo de dicha matriz.

Contamos con un conjunto de $n = 1837$ clientes distintos, descritos mediante $p = 64$ variables que indican el nivel de interés de cada cliente en varias categorías: diferentes sectores, tipos de vehículos, cinco aplicaciones específicas (eje delantero, eje trasero, eje de remolque, caja de cambios y diferencial) y 57 marcas. Este interés se mide en una escala de 0 a 1, siendo 0 indicativo de no tener interés y 1 el máximo posible valor del interés.

Además, se reduce en primer lugar el número de filas de 1837 a 1542. Estas 295 filas eliminadas se corresponden con clientes representados por vectores nulos, es decir, clientes que salen en las facturas pero que compran productos de los cuales no se tiene información en los datos de aplicaciones. Por tanto, se parte de un conjunto de $n = 1542$ datos, cada uno descrito por un conjunto de $p = 64$ variables y representado como un vector $\mathbf{x} = (x_1, x_2, \dots, x_{64})$.

El objetivo de esta sección es emplear los algoritmos de agrupamiento descritos en la Sección 2.3 para posteriormente compararlos y extraer conclusiones a partir de los resultados obtenidos. Antes de pasar a realizar el proceso de clusterización sobre el conjunto de datos obtenido, se estudian los datos con el objetivo de ver de antemano si hace falta hacer alguna transformación de las variables y si hay estructuras agrupables, tal y como se explica en la Sección 2.1.

En primer lugar, se toma la decisión de no realizar ningún tipo de escalado sobre la matriz de datos. Esto se debe a que el proceso de construcción de dicha matriz es el mismo para todas las variables, luego todas ellas contribuyen de manera equitativa.

Por otro lado, se examina si existe alguna estructura de agrupamiento inherente en los datos, empleando para ello el estadístico de Hopkins definido en la Ecuación (2.1.5). El valor obtenido es $H = 1$, lo cual indica que el conjunto de datos de estudio es significativamente diferente de datos aleatorios uniformes y que contiene estructuras agrupadas, luego tiene sentido llevar a cabo el proceso de clusterización.

Finalmente, se examina la posible estructura de agrupamiento mediante el gráfico VAT mostrado en la Figura 3.3.1. La distancia escogida para su construcción es la euclídea porque refleja directamente la distancia geométrica entre los puntos en el espacio de características, lo que facilita la interpretación de la proximidad entre las observaciones.

En dicha gráfica, se puede observar un grupo compacto de aproximadamente 1000 clientes que están muy cercanos entre sí, lo que sugiere que podrían formar un grupo bien definido. Además, se puede notar que el resto de los clientes están más dispersos y alejados de este grupo principal.

A continuación, se procede al análisis de las variables que integran el conjunto de datos utilizando la matriz de correlación lineal y la visualización mediante diagramas de caja. Dado que el conjunto de variables es extenso ($p = 64$), la representación completa de la matriz y de todos los

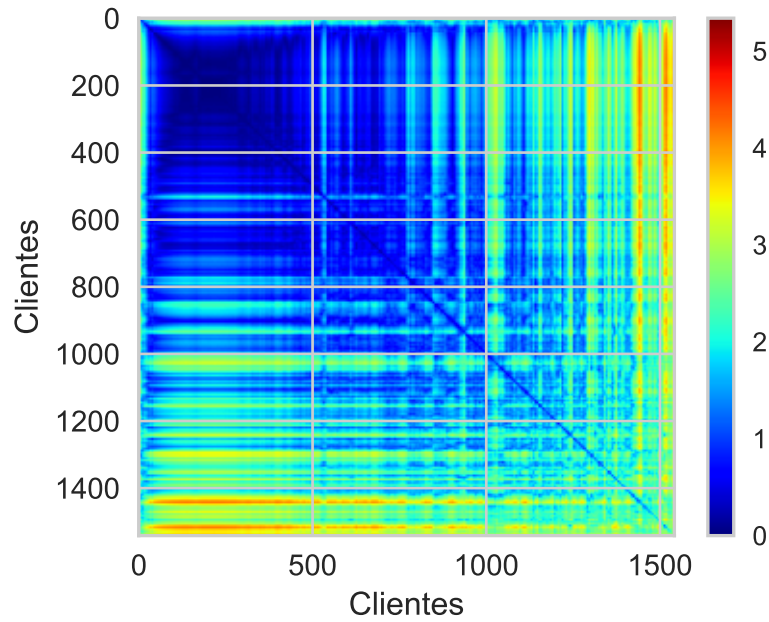


Figura 3.3.1: **Gráfico VAT de los clientes. Los distintos colores denotan la distancia euclídea.**

diagramas de caja resulta compleja y consume un espacio innecesario. Por ello, se ha decidido mostrar estas visualizaciones solo para un subconjunto reducido de variables. La visualización de todas las variables y de otras gráficas se encuentra disponible en el [Anexo B](#).

La Figura 3.3.2 muestra la matriz de correlaciones entre las variables ‘Commercial sector’, ‘Truck vehicle’, las cinco aplicaciones ‘App’, y dos marcas ‘Brand’ escogidas arbitrariamente. Es directo observar cómo la correlación lineal entre las dos primeras variables es 1, lo cual era de esperar ya que, como se ha explicado previamente, todos los productos del conjunto de datos de aplicaciones son productos de camión y pertenecientes al sector comercial. Es por ello que, en la Figura 3.3.3, no se muestra el diagrama de caja de ‘Truck vehicle’, por ser el mismo que el de ‘Commercial sector’.

Además, se observa que todas las correlaciones son positivas y, en muchos casos, bastante altas, superando un valor de 0.6. En particular, ‘Commercial sector’, ‘App_01’ y ‘App_05’ están altamente correlacionadas entre sí, mientras que ‘App_02’, ‘App_03’ y ‘App_04’ forman otro grupo con una alta correlación entre sus miembros. En contraste, las marcas presentan correlaciones generalmente más bajas con el resto de las variables.

Por otro lado, la Figura 3.3.3 revela que ‘Commercial sector’, ‘App_01’ y ‘App_05’ son las variables con mayor dispersión y valores más elevados de interés, en comparación con ‘App_02’, ‘App_03’ y ‘App_04’, que muestran menor dispersión y menor interés en general por parte de los clientes. Finalmente, ‘Brand_01’ y ‘Brand_29’ reflejan una tendencia general hacia un interés menor en las marcas, con la menor dispersión entre todas las variables, lo cual es esperable dada la amplia variedad de marcas disponibles.

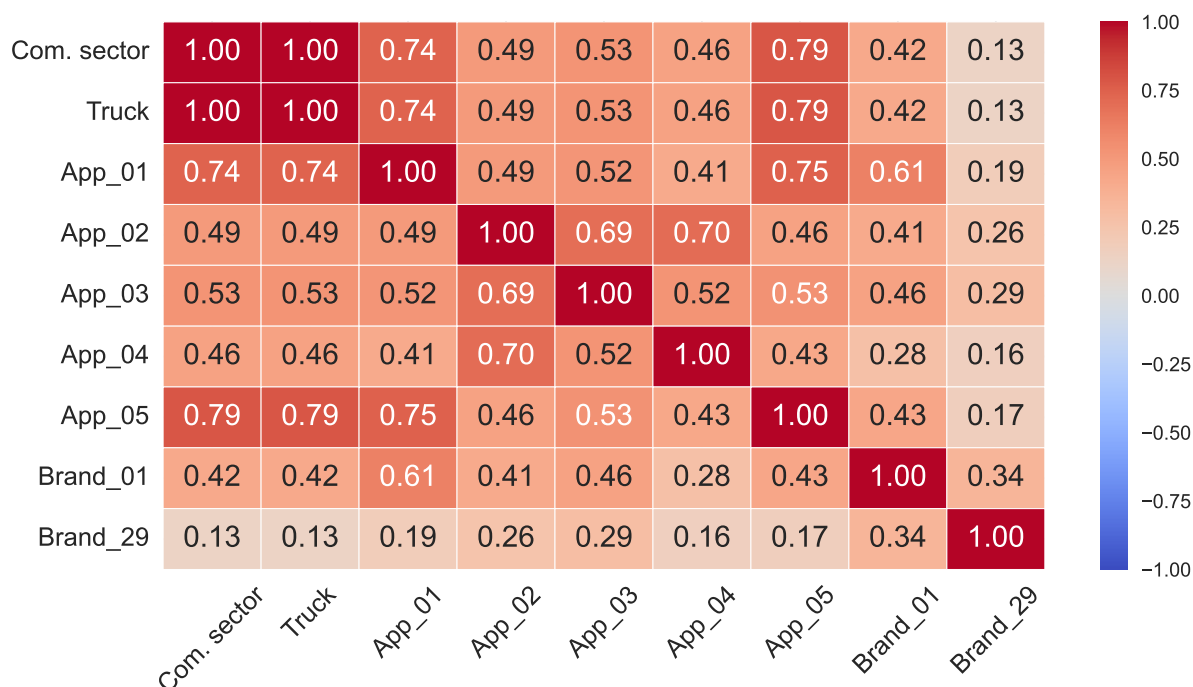


Figura 3.3.2: Matriz de correlaciones entre una selección de algunas variables.

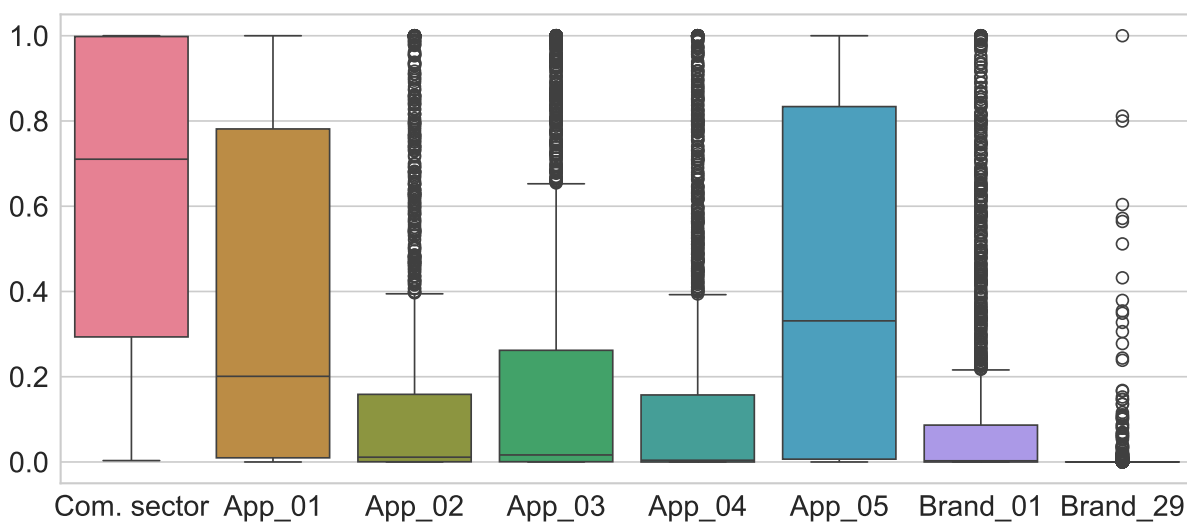


Figura 3.3.3: Selección de diagramas de caja de algunas variables.

Es importante señalar que la única técnica de reducción de dimensionalidad aplicada al conjunto de datos es el análisis de componentes principales (ACP) realizado en el siguiente apartado, en lugar de una selección tradicional de variables. La razón detrás de esta elección es que el problema en estudio no presenta limitaciones en cuanto a costes computacionales o tiempos de ejecución elevados, lo que permite utilizar la totalidad de los datos.

3.3.1. Análisis de Componentes Principales

El objetivo ahora es describir el conjunto de datos completo utilizando nuevas variables no correlacionadas, conocidas como componentes principales, para visualizar y simplificar los datos de alta dimensión de los que se parte, conservando la mayor cantidad de información posible.

La primera componente principal (CP1) define el eje que captura la mayor varianza en los datos, seguida de la segunda componente principal (CP2). Este proceso se repite para las siguientes componentes principales, reordenando la varianza de manera descendente a lo largo de los nuevos ejes coordenados. En el Cuadro 3.3.1 se muestra el porcentaje de varianza explicada por cada una de las siete primeras componentes principales, así como el porcentaje de varianza acumulada.

Componente principal	PC1	PC2	PC3	PC4	PC5	PC6	PC7
% varianza explicada	51.37	7.74	6.40	5.10	4.11	3.27	2.39
% varianza acumulada	51.37	59.11	65.51	70.61	74.72	77.99	80.38

Cuadro 3.3.1: Porcentaje de la varianza explicada por las siete primeras componentes principales, y porcentaje de varianza acumulada.

Se puede observar que las dos primeras componentes principales logran capturar casi el 60 % de la información original. Nos centraremos en estas dos componentes, ya que nos permitirán representar gráficamente los puntos que corresponden a los clientes en un plano bidimensional.

En los Cuadros 3.3.2 y 3.3.3 se presentan los coeficientes que denotan en mayor medida, tanto de manera positiva como negativa, las direcciones de la primera y segunda componente principal en el espacio de variables original, respectivamente. Estos coeficientes indican cuánto contribuye cada variable original a cada componente principal.

Variable	App_05	App_01	Com. sector	Truck	Brand_13	Brand_12	Brand_27
Coefficiente	0.363	0.340	0.305	0.305	0.293	0.270	0.256

Cuadro 3.3.2: Variables que más contribuyen a la CP1 y sus coeficientes asociados.

En el caso de la primera componente principal, todos los coeficientes son positivos, lo que significa que las variables originales están alineadas en la misma dirección en esta componente. Los puntos que representen clientes con valores altos en la primera componente principal (eje de abscisas) corresponderán a clientes con un alto interés en las categorías ‘App_05’, ‘App_01’, ‘Commercial sector’, ‘Truck vehicle’, ‘Brand_13’, ‘Brand_12’ y ‘Brand_27’.

Variable	Com. sector	Truck	App_01	App_05	App_04	App_03	App_02
Coefficiente	0.359	0.359	0.261	0.255	-0.290	-0.318	-0.324

Cuadro 3.3.3: Variables que más contribuyen a la CP2 y sus coeficientes asociados.

Por otro lado, para los puntos proyectados en el plano de las dos primeras componentes principales, aquellos con un valor grande positivo en la segunda componente principal (eje de ordenadas) corresponderán a clientes con un alto interés en las categorías ‘Commercial sector’, ‘Truck vehicle’, ‘App_01’ y ‘App_05’. En cambio, los clientes con un valor grande negativo en la segunda componente principal tendrán un alto interés en las categorías ‘App_02’, ‘App_03’ y ‘App_04’.

Finalmente, se transforman los datos originales a un nuevo espacio definido por las dos primeras componentes principales, proyectando las observaciones (tras estandarizarlas) en los nuevos ejes, tal y como se observa en la Figura 3.3.4.

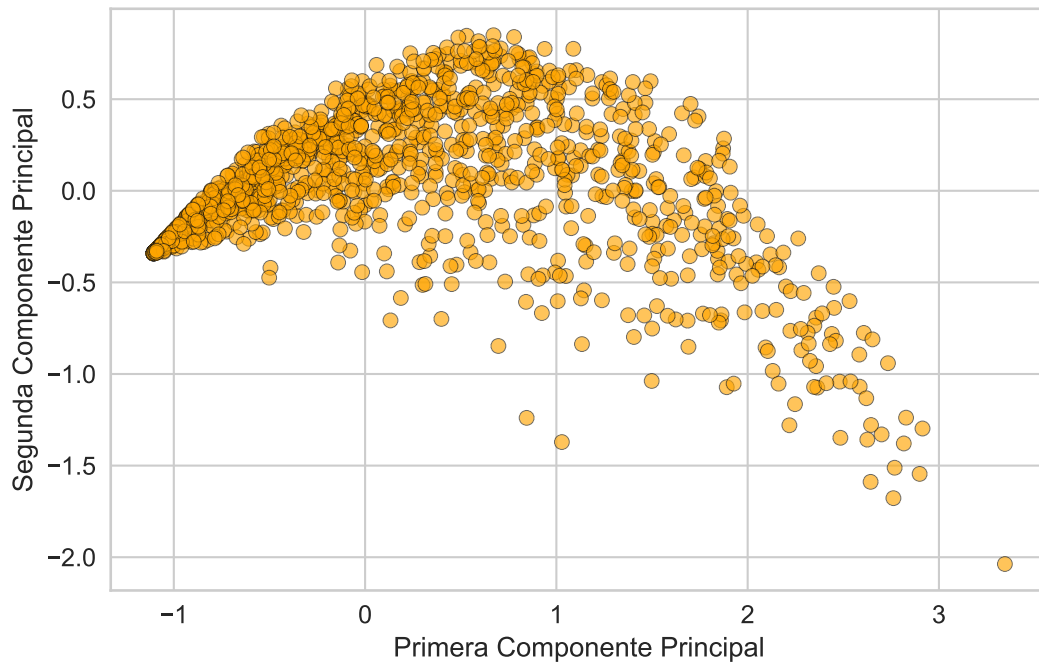


Figura 3.3.4: Visualización de los puntos que definen el interés de los clientes, proyectados sobre las dos primeras componentes principales.

Tras analizar el conjunto de datos almacenado en *Knowledge_Customers_Category* y su representación sobre un número reducido de componentes principales, ahora podemos aplicar los métodos de agrupamiento descritos en la Sección 2.3.

3.3.2. K -medias

El único parámetro necesario para aplicar el algoritmo de agrupamiento K -medias es K , el número de clústeres en los que se debe dividir el conjunto de datos. Para determinar este valor, se utilizan los métodos descritos en la sección 2.4: el método del codo, el método de la silueta y el índice de Dunn. Los resultados obtenidos se muestran en la Figura 3.3.5.

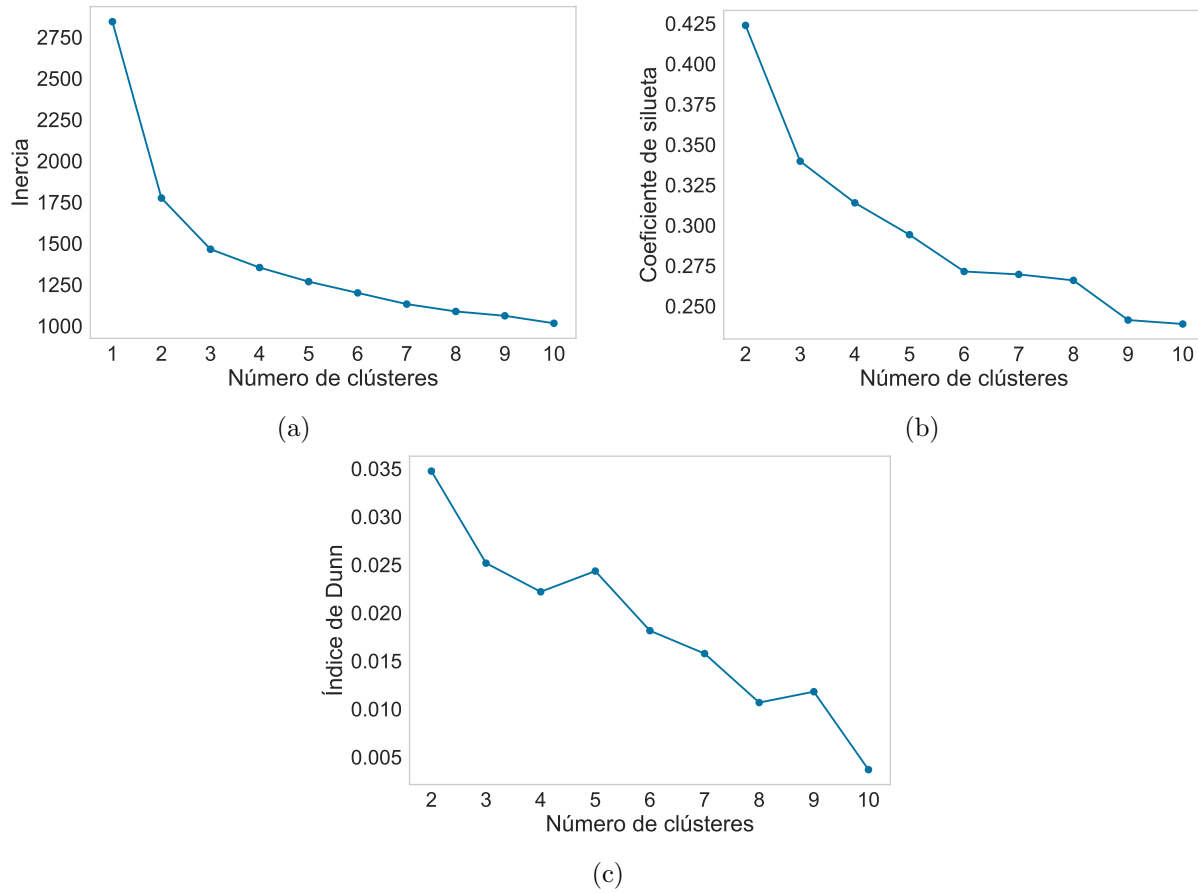


Figura 3.3.5: Estudio del número de clústeres óptimo para un agrupamiento K -medias. (a) Método del codo. (b) Método de la silueta. (c) Método del índice de Dunn.

El método del codo sugiere que el número óptimo de clústeres es $K = 2$ o $K = 3$. Para resolver esta incertidumbre, se consultan los dos métodos adicionales en los que la métrica (coeficiente de silueta o índice de Dunn) alcanza su valor máximo para $K = 2$. Es por ello por lo que se decide tomar 2 como el número de clústeres en los que dividir el conjunto de datos original.

Tras aplicar el algoritmo K -medias, se obtiene la distribución de los datos en los grupos denominados '0' y '1', como se muestra en la Figura 3.3.6, para un subconjunto de las variables originales. Se observa un grupo '1' con interés medio mayor en las distintas categorías que el grupo '0'. Además, por lo general, los datos del clúster '1' tienen una mayor varianza que los del grupo '0'.

Finalmente, los puntos ya agrupados proyectados sobre las componentes principales se muestran en la Figura 3.3.7. En dicho gráfico se observa claramente una frontera entre los dos grupos, con una mayor densidad de puntos (menor varianza) en el grupo '0', formado por 1015 puntos, concentrados en valores cercanos a 0 en ambas componentes principales, indicando un interés medio bajo en las categorías en general, en comparación con el grupo '1', formado por 527 puntos.

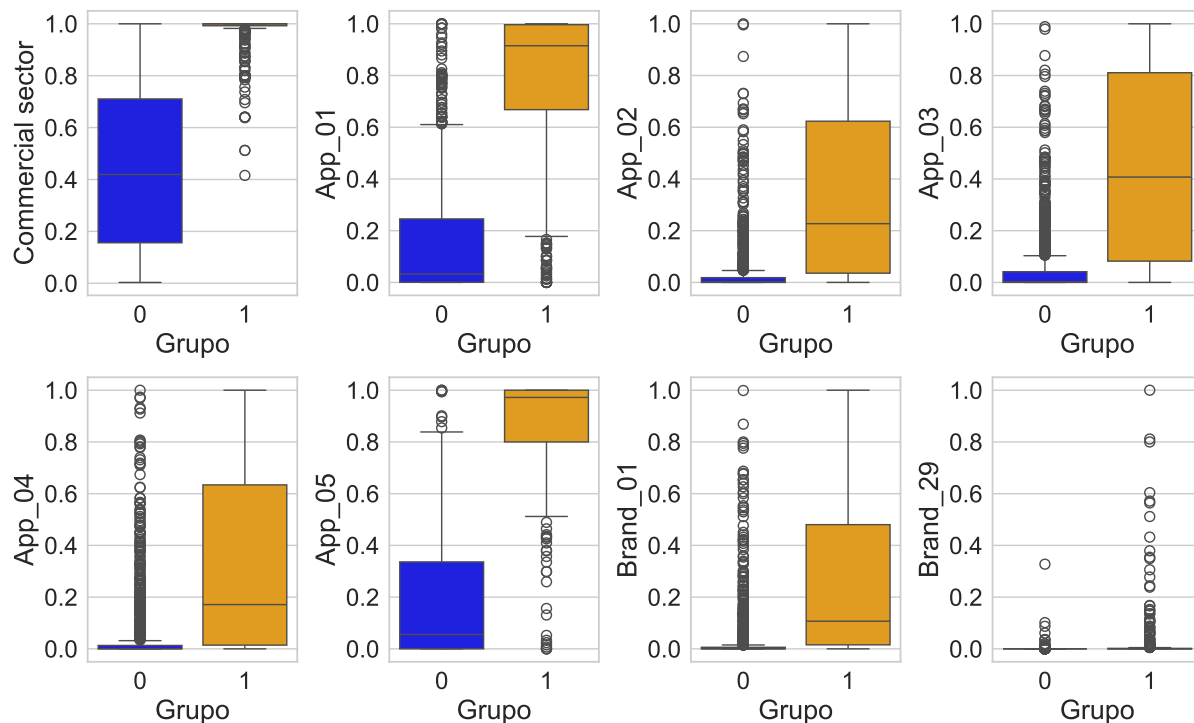


Figura 3.3.6: Selección de diagramas de caja de algunas variables, separados por grupos según el método *K-medias*.

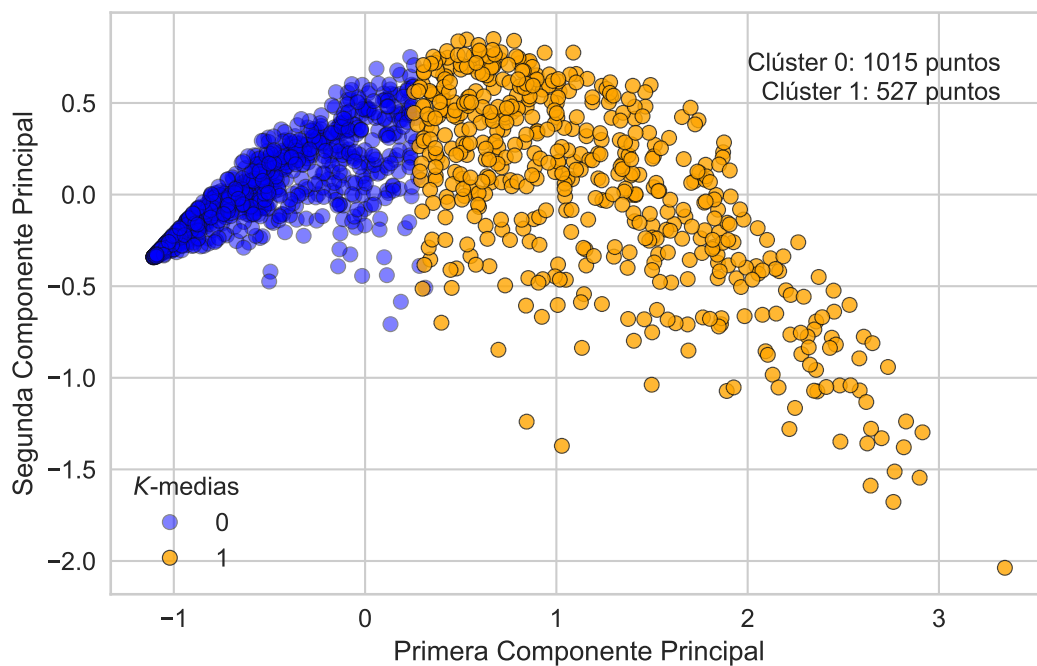


Figura 3.3.7: Visualización de los grupos creados por el algoritmo *K-medias*, proyectados sobre las dos primeras componentes principales.

3.3.3. DBSCAN

El algoritmo DBSCAN requiere de un tuneado de los parámetros ε , el radio de la vecindad alrededor de un punto, y $MinPts$, el número mínimo de puntos necesarios que un dato ha de tener a una distancia menor que ε para que sea considerado núcleo.

La selección de dichos parámetros se lleva a cabo de manera simultánea. Para ello, se toma un rango de valores para $MinPts$. Seguidamente, se realiza un proceso análogo al método del codo sobre una gráfica de los k -vecinos más cercanos, donde $k = MinPts$. Es decir, para un valor seleccionado de $MinPts$, se trazan las distancias al $MinPts$ -ésimo vecino más cercano de cada punto, ordenadas de manera ascendente, y se observa dónde ocurre un cambio abrupto en la pendiente de la curva, conocido como el “codo”. Este punto se corresponde con el valor óptimo de ε para dicho valor de $MinPts$, ya que indica la distancia a partir de la cual los puntos comienzan a agruparse de manera efectiva, separándose del ruido.

Por otro lado, para obtener el valor óptimo de $MinPts$, se utiliza el índice de Dunn, que mide la calidad de la agrupación. Se calcula este índice para cada valor dentro del rango de $MinPts$ estudiado, y se selecciona el valor de $MinPts$ que maximice dicho índice, asegurando la mejor separación entre grupos y la mayor compacidad dentro de cada grupo.

El rango escogido es $45 \leq MinPts \leq 345$, con saltos de 10 en 10. El motivo es que este rango permite explorar un conjunto lo suficientemente amplio de valores de $MinPts$, asegurando que se capturen tanto configuraciones de agrupamiento más estrictas como aquellas más flexibles. La Figura 3.3.8 muestra el máximo índice de Dunn para $MinPts = 65$, y su correspondiente gráfica de distancia al 65-ésimo vecino más cercano, para el cálculo de ε .

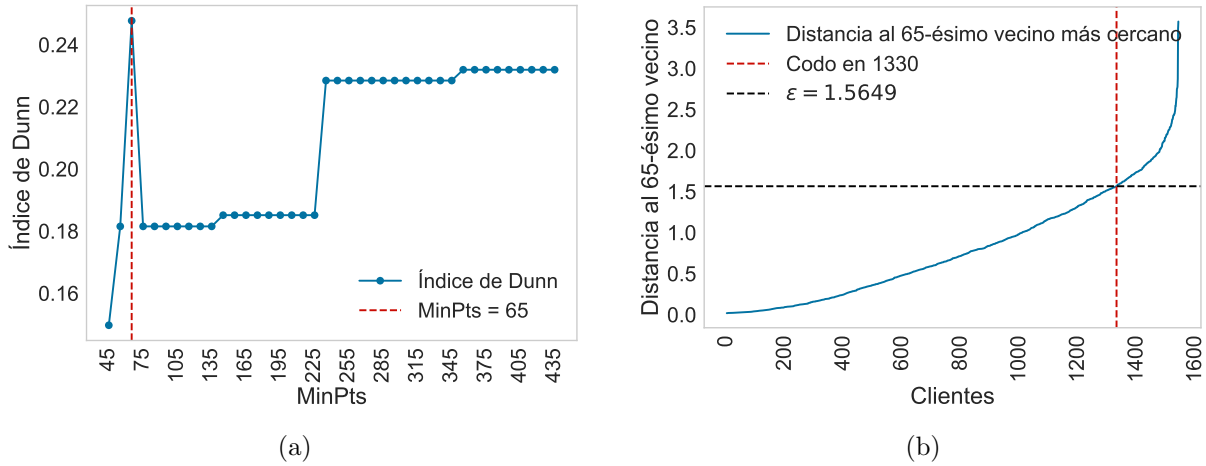


Figura 3.3.8: Estudio de los parámetros $MinPts$ y ε para un agrupamiento DBSCAN. (a) Método del índice de Dunn para determinar $MinPts$. (b) Método del codo para determinar ε .

Atendiendo a las gráficas anteriores, los valores óptimos de los parámetros para la aplicación del método de agrupamiento DBSCAN sobre el conjunto de datos de estudio son $MinPts = 65$ y $\varepsilon = 1,565$.

Cabe mencionar que, aunque no se emplea el método del coeficiente de silueta en el estudio de los parámetros óptimos, el valor obtenido para la agrupación realizada con el método DBSCAN y parámetros $MinPts = 65$ y $\varepsilon = 1,565$ es 0.493. El motivo de no emplearlo como criterio de selección radica en que, al aumentar considerablemente $MinPts$, el número de clientes asignados a un grupo pequeño se reduce cada vez más. En estos casos, el coeficiente de silueta puede seguir aumentando, pero esto no tiene mucho sentido, ya que la división en grupos pierde relevancia al quedar prácticamente todos los datos en un solo clúster grande.

La aplicación del algoritmo da como resultado un grupo, denotado como clúster ‘0’, y otro grupo denotado como clúster ‘-1’, que se corresponde con el ruido, es decir, aquellos puntos que no son núcleo ni se encuentran dentro de un radio ε de algún núcleo. La distribución de los datos en los dos grupos para un subconjunto de las variables originales se muestra en la Figura 3.3.9.

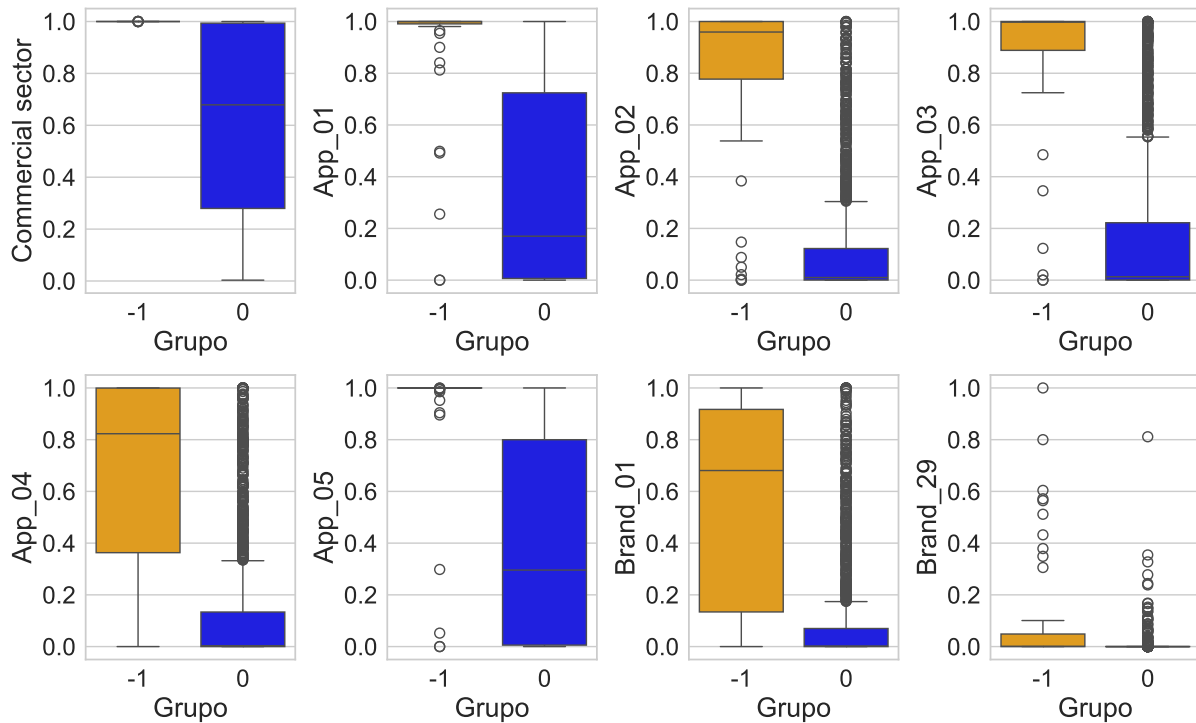


Figura 3.3.9: Selección de diagramas de caja de algunas variables, separados por grupos según el método DBSCAN.

Se observa de nuevo, al igual que sucedía con el algoritmo K -medias, que el grupo ‘0’, formado por la gran mayoría de los clientes, tiene un menor interés medio por las categorías que el grupo ‘-1’ de los puntos que conforman el ruido. Es decir, hay un subconjunto pequeño de los clientes con un interés medio “más alto de lo normal” que denota el grupo ‘0’. También se vuelve a dar que los valores de interés del grupo ‘0’ en las distintas marcas no tiene apenas varianza, y el interés del grupo ‘-1’ en el sector comercial es prácticamente 1 para todos los clientes pertenecientes a dicho clúster.

Al proyectar los datos de clientes ya agrupados sobre las dos primeras componentes principales, se obtiene el resultado mostrado en la Figura 3.3.10. En este caso, el clúster ‘0’ está formado por 1490 puntos, casi la totalidad de los clientes, y el clúster ‘-1’ por 52 clientes únicamente. Como cabía esperar, los clientes del clúster ‘-1’ poseen valores muy altos de la primera componente principal en su mayoría, y también valores extremos (positivos y negativos) de la segunda componente principal.

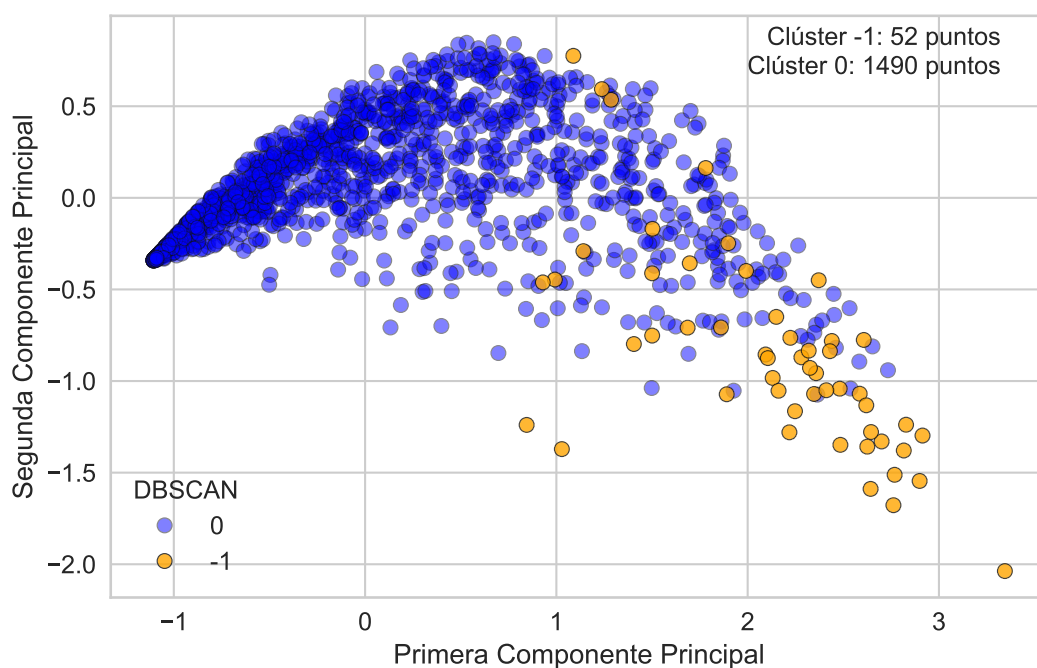


Figura 3.3.10: Visualización de los grupos creados por el algoritmo DBSCAN, proyectados sobre las dos primeras componentes principales.

3.3.4. Jerárquico aglomerativo

La manera de proceder con la aplicación del algoritmo jerárquico aglomerativo es similar a la empleada en el método K -medias. El primer paso es decidir el criterio de enlace que usar al unir los sucesivos clústeres. De entre los vistos en el apartado 2.3.2, se muestran en la Figura 3.3.11 sus dendrogramas respectivos.

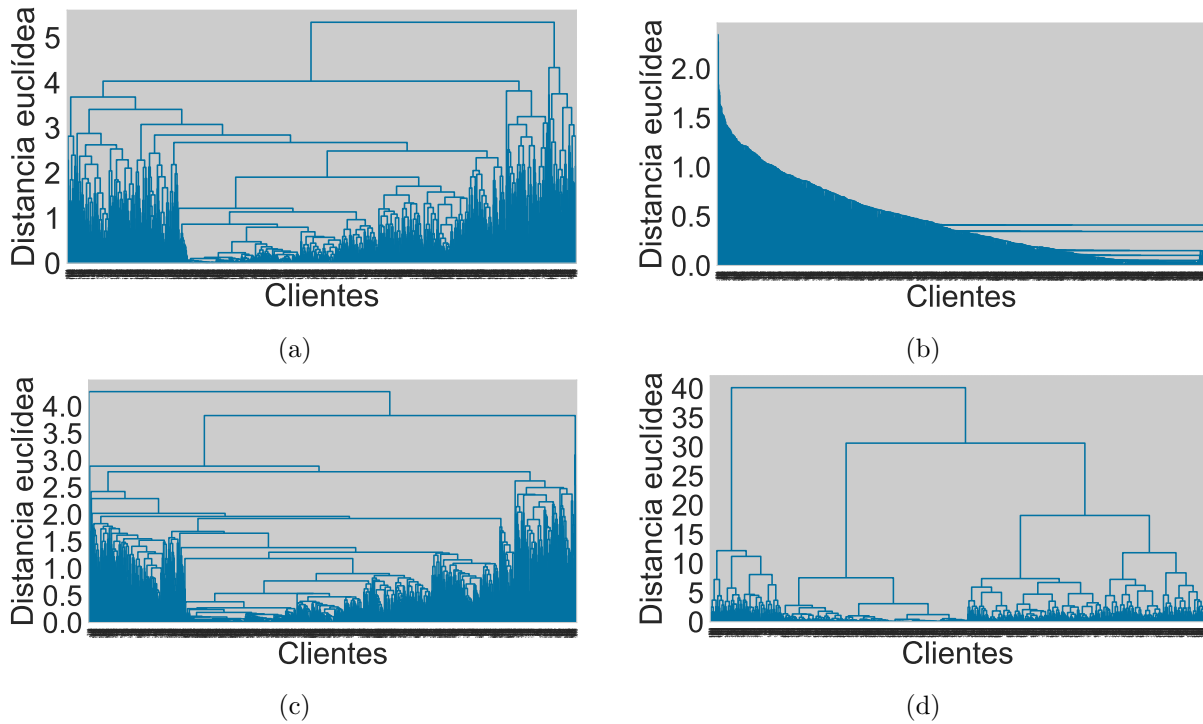


Figura 3.3.11: **Dendrogramas de los datos para los distintos criterios de enlace.** (a) Completo. (b) Simple. (c) Promedio. (d) Método de Ward.

Basándose en los dendrogramas previos, se selecciona el método de Ward como el criterio de enlace para la agrupación. De acuerdo con estos dendrogramas, los datos parecen ajustarse a una partición en $K = 2$ o $K = 3$ clústeres. La elección final se determina mediante los métodos de evaluación empleados hasta ahora, como el coeficiente de silueta y el índice de Dunn, cuyos resultados se presentan en la Figura 3.3.12.

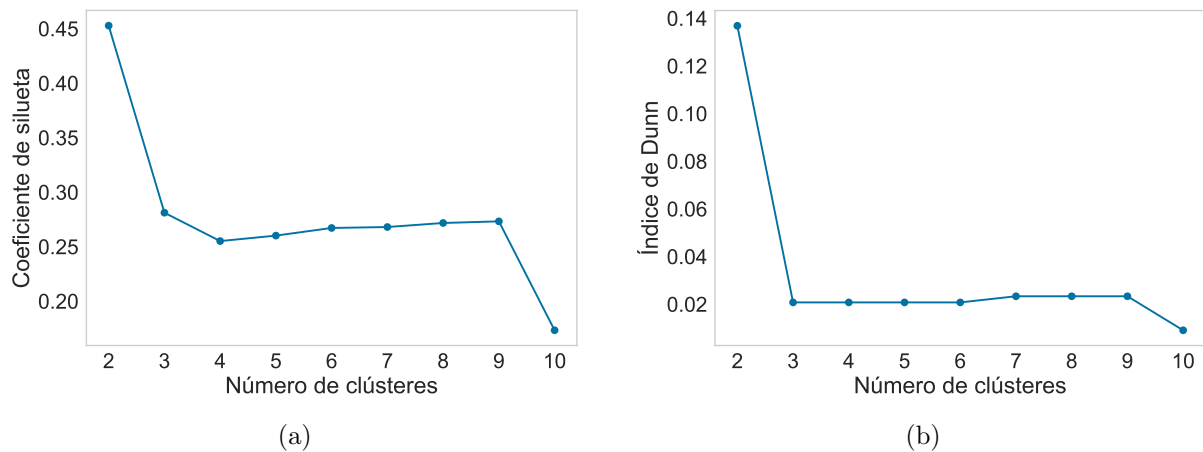


Figura 3.3.12: **Estudio del número de clústeres óptimo para un agrupamiento jerárquico aglomerativo.** (a) Método de la silueta. (b) Método del índice de Dunn.

En vista de los anteriores resultados, se escoge $K = 2$ como el número óptimo de clústeres, y se procede a realizar el proceso jerárquico aglomerativo sobre el conjunto de datos en el espacio de variables original. La distribución de los datos en los dos grupos para un subconjunto de las variables originales se muestra en la Figura 3.3.13.

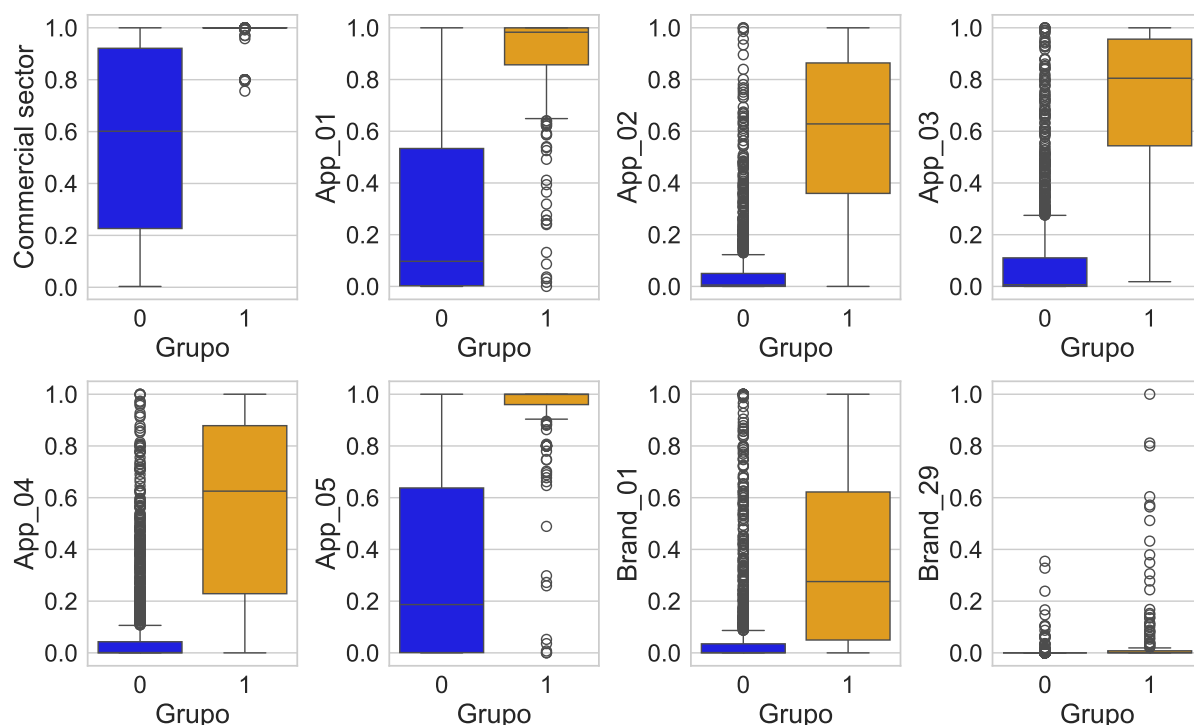


Figura 3.3.13: Selección de diagramas de caja de algunas variables, separados por grupos según el método jerárquico aglomerativo.

Una vez más, el grupo '0' está formado por la mayoría de los clientes, los cuales muestran un interés medio menor en todas las categorías, con un interés prácticamente nulo en la mayoría de las marcas. Además, su varianza vuelve a ser menor en la mayoría de las variables, comparada con la varianza del grupo '1'. El grupo '1' vuelve a estar formado por clientes con un interés prácticamente igual a 1 en el sector comercial.

Finalmente, se proyectan los puntos que definen a los clientes divididos en los dos clústeres sobre el plano definido por las dos componentes principales, y se muestra el resultado en la Figura 3.3.14. El clúster '0' está formado por 1318 individuos, y el clúster '1' por los 224 restantes, los cuales tienen valores grandes positivos de la primera componente principal y valores extremos negativos de la segunda componente principal.

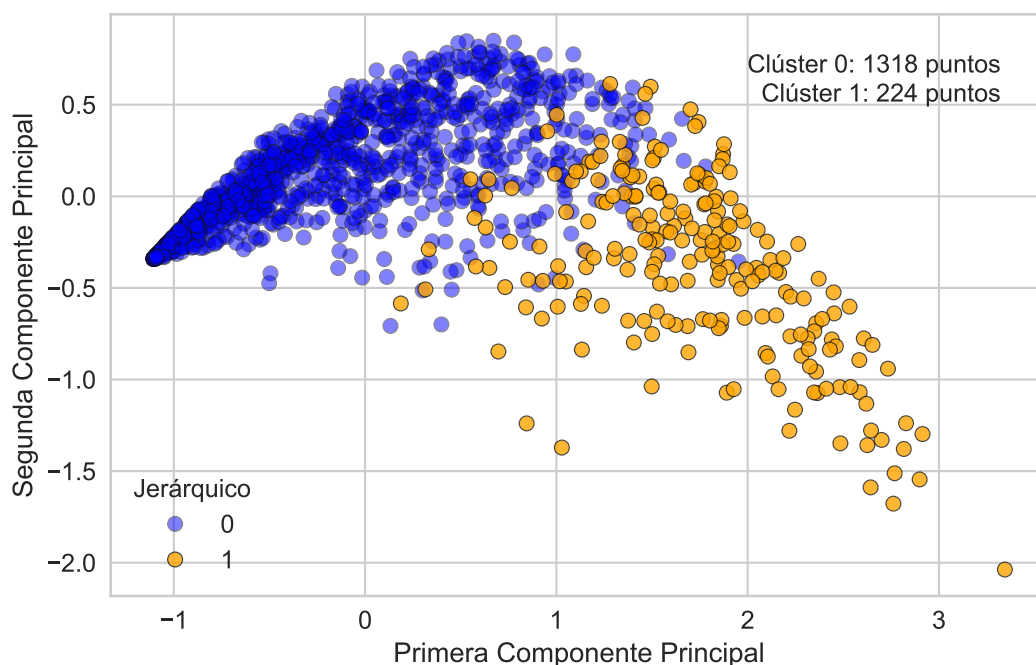


Figura 3.3.14: Visualización de los grupos creados por el algoritmo jerárquico aglomerativo, proyectados sobre las dos primeras componentes principales.

3.4. Estrategia de recomendación

La sección anterior nos ha permitido analizar los intereses de los distintos clientes registrados en el conjunto de datos de facturas. Mediante tres algoritmos de agrupamiento, los clientes han sido clasificados en dos grupos: uno compuesto por la mayoría de los clientes y otro por un grupo reducido. En todos los casos, este último grupo destaca por presentar un mayor interés medio en las diversas categorías de productos.

En la siguiente sección, se expondrán algunas ideas sobre las estrategias que un sistema de recomendación automática como *SARA* podría seguir para realizar recomendaciones basadas en los resultados obtenidos.

3.4.1. Popularidad dentro de los clústeres

Tras identificar dos grupos de clientes, uno con mayor interés general y otro con menor interés, resulta natural aplicar estrategias de recomendación diferenciadas para cada segmento.

Por un lado, para el grupo con alto interés, se podrían recomendar productos que sean tendencia o especialmente populares dentro de las categorías de su mayor interés. Dado que este grupo es más reducido, es posible realizar un análisis más detallado de cada cliente, lo que podría

traducirse en mayores beneficios. Esta estrategia parte de la suposición de que estos clientes tienden a probar nuevos productos o a realizar compras con mayor frecuencia.

Por otro lado, para el clúster con menor interés general, es recomendable optar por una estrategia de recomendación más conservadora. Aquí sería preferible sugerir productos con buena aceptación general y que se utilicen en una amplia variedad de aplicaciones. En lugar de proponer productos innovadores o de nicho, se recomendarían productos bien valorados por clientes con hábitos de compra similares.

3.4.2. Proximidad geográfica

Esta estrategia está dirigida a clientes nuevos de los que únicamente se conoce su localización geográfica (por ejemplo, mediante su dirección IP). Se basa en realizar una recomendación inicial apoyada en los intereses de los clientes de la misma región. Esta aproximación es especialmente útil para productos cuya demanda varía según la ubicación.

En la Figura 3.4.1 se resaltan en rojo los clientes de tres zonas geográficas distintas, sobre un fondo de puntos grises que representan al resto de los clientes, proyectados en las dos primeras componentes principales.

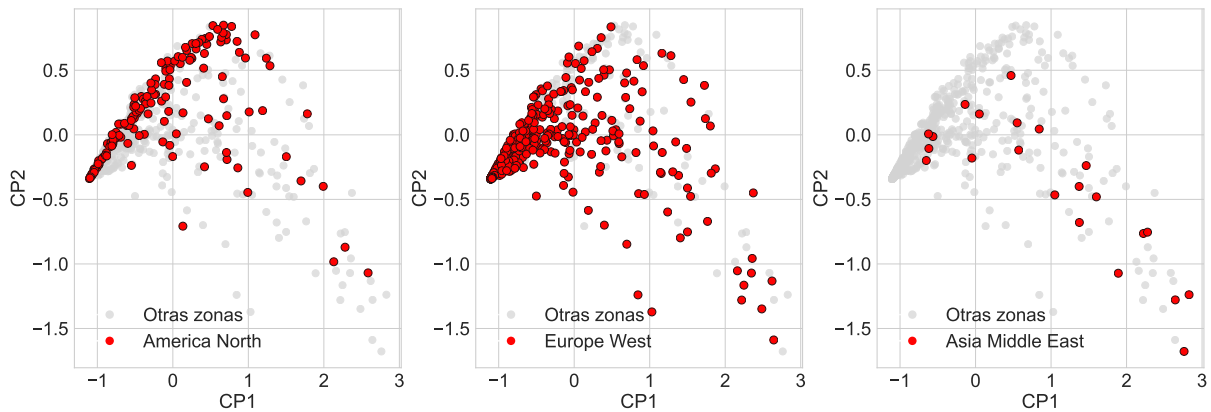


Figura 3.4.1: Visualización de los clientes sobre las dos primeras componentes principales, para varias zonas geográficas.

Atendiendo a los gráficos anteriores, se observa que hay distintos intereses en función de las regiones, por lo que tiene sentido hacer una recomendación basada en la zona geográfica. Por ejemplo, en el gráfico correspondiente a América del Norte, a un nuevo cliente de dicha región sin información previa se le recomendarían productos de las categorías que más interesan a los clientes de esa misma zona.

3.4.3. Filtración colaborativa basada en usuarios

La idea clave detrás de la filtración colaborativa basada en usuarios es que los clientes que han mostrado comportamientos o intereses similares en el pasado probablemente estarán interesados en las mismas cosas en el futuro. Este enfoque busca recomendar productos a un usuario basándose en los comportamientos de otros usuarios con gustos similares.

Aplicado a nuestro problema, la filtración colaborativa se aplica directamente sobre *Knowledge_Customers_Category*, la matriz que indica el interés de los clientes en las distintas categorías de los productos. El proceso de recomendación a un cliente dado consistiría únicamente en sugerirle aquellos productos que sus clientes más cercanos hayan comprado y que el primero no haya adquirido. Dichos clientes más cercanos son aquellos que se encuentran a una menor distancia del cliente de estudio. En el código accesible en el [Anexo A](#) se muestra una implementación de una función que devuelve los N clientes más cercanos a uno dado.

3.4.4. Venta cruzada y venta adicional

La venta cruzada, o *cross-selling*, se refiere a la recomendación de productos complementarios o adicionales a los que el cliente ya ha adquirido. La venta adicional, o *up-selling*, es una estrategia que consiste en ofrecer al cliente una versión mejorada o más completa del producto que planea adquirir. A diferencia de la venta cruzada, la venta adicional no busca añadir productos complementarios, sino proponer una alternativa que incluya el producto deseado junto con otros elementos que mejoran o facilitan su uso. Ambas estrategias de recomendación tienen el objetivo de mejorar la experiencia de compra y aumentar el valor total de la transacción. Aunque estas estrategias no se basan directamente en el análisis de datos realizado en este trabajo, los métodos de venta cruzada y venta adicional están implementados en el sistema *SARA*.

El funcionamiento de la venta cruzada se apoya en un estudio de los productos que suelen comprarse juntos. En el contexto de una empresa como Fersa, un ejemplo típico sería recomendar grasas o lubricantes especializados a un cliente que ha comprado un tipo específico de rodamiento, ya que estos productos suelen adquirirse de forma conjunta. Por otro lado, un ejemplo claro de venta adicional sería recomendar a un cliente que está por adquirir un rodamiento específico, un buje completo ya montado con todos los componentes necesarios. De este modo, el cliente no solo obtiene el rodamiento que necesita, sino también otros productos esenciales para un montaje más eficiente, lo que mejora tanto la operatividad como el valor percibido de la compra.

Capítulo 4

Conclusiones y trabajo futuro

El trabajo realizado se centra en la transformación de unos datos de facturas y la información sobre distintos productos en un conjunto de datos que describe el interés de 1542 clientes en 64 categorías a las que pertenecen dichos productos. A partir de esta transformación, se lleva a cabo un análisis detallado del conjunto de datos creado, con el objetivo principal de identificar patrones que permitan agrupar a los clientes y, en última instancia, generar recomendaciones especializadas para los distintos clústeres resultantes.

Para el proceso de clusterización, se han utilizado los tres métodos descritos en la Sección 2.3. Todos los métodos coinciden en dividir a los clientes en dos grupos: uno conformado por la mayoría de los clientes, que muestran un menor interés, y otro formado por un pequeño número de clientes que presentan un mayor interés en las categorías de los productos.

La calidad de los agrupamientos realizados se puede evaluar mediante los métodos presentados en la Sección 2.4. El Cuadro 4.0.1 muestra los valores obtenidos del coeficiente de silueta y del índice de Dunn para los tres métodos de agrupamiento, junto con el número de clientes que pertenecen al grupo de mayor interés.

	Coef. de silueta	Índ. de Dunn	Clientes alto interés
<i>K</i>-medias	0.424	0.035	527
DBSCAN	0.493	0.248	52
Jerárquico aglomerativo	0.453	0.137	224

Cuadro 4.0.1: Comparativa de los valores del coeficiente de silueta, del índice de Dunn y del número de clientes en el grupo de alto interés obtenidos en los distintos métodos de agrupamiento empleados.

Un mayor valor del índice de Dunn indica una mejor calidad de los clústeres, ya que refleja una mayor densidad dentro de los grupos y una mejor separación entre ellos. De manera similar, un coeficiente de silueta más alto señala una mayor cohesión interna en los clústeres y una mayor distancia entre ellos. Basándonos en estos dos criterios, el método DBSCAN es el que produce la mejor agrupación, seguido del método jerárquico aglomerativo, mientras que el método *K*-medias obtiene los resultados más débiles.

Este patrón parece estar inversamente relacionado con el número de clientes en el grupo menos poblado y de mayor interés. Esto se debe a la naturaleza de los índices utilizados: tanto el coeficiente de silueta (definido en la Ecuación (2.4.4)) como el índice de Dunn (definido en la Ecuación (2.4.5)) tienden a mejorar cuando hay menos puntos dentro de un clúster pequeño, ya que esto reduce las distancias internas y maximiza la separación entre grupos, lo que incrementa los valores de ambos indicadores.

Además, al analizar la proyección de los datos sobre el plano de las dos primeras componentes principales (Figura 3.3.4), se observa una gran aglomeración de puntos, con algunos individuos alejados del clúster principal. DBSCAN es particularmente eficaz en este tipo de escenarios, donde existe un grupo dominante y un pequeño número de puntos que se desvían significativamente. El algoritmo detecta estos puntos atípicos, clasificándolos como ruido y destacando a los clientes con mayor interés, que son aquellos que se salen de lo habitual.

Finalmente, el reducido número de clientes en el clúster de alto interés también facilita un análisis más detallado e individualizado de estos clientes, lo que abre la puerta a estrategias de recomendación más personalizadas, como se explica en la Sección 3.4.

En cuanto a las limitaciones y posibles variantes del problema a explorar en el futuro, la creación de la matriz de datos *Knowledge_Customers_Category* presenta un inconveniente: los clientes que realizan compras de grandes volúmenes de productos, pero de forma esporádica, tienden a aparecer con un interés reducido en las distintas categorías si ha pasado un tiempo considerable desde su última transacción. Esto puede llevar a una subestimación de su verdadero interés en los productos, ya que el patrón de compra no refleja una falta de interés, sino simplemente una menor frecuencia en las adquisiciones. Para mitigar este problema, sería recomendable implementar un enfoque que no penalice a estos clientes por su falta de regularidad, y que en su lugar considere su historial de compras en su conjunto. Esto permitiría ajustar las recomendaciones en función de su comportamiento a largo plazo, ofreciendo una representación más precisa de sus intereses.

El entrenamiento de la matriz de intereses se realiza con una frecuencia de un año. Ejecutar el entrenamiento en intervalos muy cortos no permitiría acumular suficientes compras entre sesiones, lo que reduciría la posibilidad de detectar cambios significativos en los intereses de los clientes y en las recomendaciones. Por otro lado, si el entrenamiento se realiza con demasiada poca frecuencia, se corre el riesgo de no captar los intereses reales de los clientes en las distintas categorías. Esto podría llevar a recomendar repetidamente productos ya sugeridos anteriormente a los mismos clientes. Por lo tanto, es fundamental realizar un análisis exhaustivo para determinar el intervalo óptimo entre entrenamientos, de modo que el sistema refleje adecuadamente las preferencias actuales sin redundancias en las recomendaciones.

Para mejorar la precisión y la efectividad del sistema de recomendación a lo largo del tiempo, una opción relevante sería integrar tanto la retroalimentación explícita como implícita de los usuarios. El feedback explícito, como las valoraciones y reseñas, ofrece información directa sobre el grado de satisfacción del cliente con las recomendaciones. Por su parte, el feedback implícito (clics, tiempo de permanencia en la página o compras realizadas) aporta información valiosa sobre las preferencias del usuario sin requerir una interacción directa. Al incorporar ambos tipos

de retroalimentación en los modelos de recomendación, se lograría una mayor personalización de las sugerencias, optimizando así la experiencia del usuario y aumentando la relevancia de los productos recomendados.

Además de la retroalimentación por parte de los usuarios, es crucial considerar la opinión del equipo comercial de la empresa sobre la calidad de las recomendaciones. Dado su profundo conocimiento cualitativo de los clientes, el equipo comercial puede aportar una perspectiva que va más allá de los datos cuantitativos, permitiéndoles ajustar manualmente los intereses o preferencias de los clientes en función de características subjetivas que el sistema automatizado podría no captar. Al incorporar esta retroalimentación cualitativa, se puede lograr que el sistema de recomendación sea más preciso y refleje mejor las necesidades reales de cada cliente, proporcionando un valor añadido significativo.

Por último, la implementación de técnicas avanzadas como el clustering dinámico permitiría que los grupos de clientes evolucionen de forma continua a medida que se recopilan nuevos datos. Esto facilitaría que el sistema se ajuste de manera dinámica a los cambios en los comportamientos de compra y las preferencias del mercado.

En conclusión, el trabajo realizado ha permitido no solo llevar a cabo un análisis detallado de los datos de los clientes, sino también sentar las bases para futuras mejoras en la estrategia de recomendación.

Bibliografía

- [1] Juan Manuel Pérez García de Carellán. “Quantum Machine Learning for Disease Diagnosis”. Dirigido por Dr. David Zueco Lainez y Dr. Rafael del Hoyo Alonso. Trabajo de Fin de Grado. Universidad de Zaragoza, 2023.
- [2] Richard A. Johnson y Dean W. Wichern. *Applied multivariate statistical analysis*. 5. ed. Upper Saddle River, NJ: Prentice Hall, 2002. XVIII, 767. ISBN: 0130925535.
- [3] Tuhin Mukherjee. *Different Types of Distances Used in Machine Learning, Explained*. 2021. URL: <https://tuhinmukherjee74.medium.com/different-types-of-distances-used-in-machine-learning-explained-550e2979752c>.
- [4] Brian S. Everitt et al. *Cluster Analysis*. 5th. Pág. 54. Wiley, 2011. URL: https://cicerocq.wordpress.com/wp-content/uploads/2019/05/cluster-analysis_5ed_everitt.pdf.
- [5] Rosana Ferrero. *Diagramas de cajas (boxplot), cuartiles y percentiles*. <https://rosanaFerrero.blogspot.com/2011/09/diagramas-de-cajas-boxplot-cuantiles.html>. 2011.
- [6] Anil K. Jain y Richard C. Dubes. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988, págs. 218-219.
- [7] James Bezdek y Richard J. Hathaway. “VAT: A tool for visual assessment of (cluster) tendency”. En: vol. 3. Feb. de 2002, págs. 2225-2230. ISBN: 0-7803-7278-6. DOI: 10.1109/IJCNN.2002.1007487.
- [8] Virginia Ahedo et al. “Let’s go fishing: A quantitative analysis of subsistence choices with a special focus on mixed economies among small-scale societies”. En: *PLOS ONE* 16.8 (ago. de 2021), págs. 1-47. DOI: 10.1371/journal.pone.0254539. URL: <https://doi.org/10.1371/journal.pone.0254539>.
- [9] Jonathon Shlens. “A Tutorial on Principal Component Analysis”. En: *CoRR* abs/1404.1100 (2014). arXiv: 1404.1100. URL: <http://arxiv.org/abs/1404.1100>.
- [10] Jiawei Han, Micheline Kamber y Jian Pei. *Data Mining: Concepts and Techniques*. 3rd. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2011.
- [11] Ligdi Gonzalez. *DBSCAN: Teoría*. 2020. URL: <https://aprendeia.com/dbscan-teoria/>.
- [12] Ander Fernández. *DBSCAN en Python: Algoritmo de Clustering Explicado*. URL: <https://anderfernandez.com/blog/dbscan-python/>.
- [13] Enrique J. De La Hoz, Efraín De La Hoz y Tomás Herrera. “Metodología de Aprendizaje Automático para la Clasificación y Predicción de Usuarios en Ambientes Virtuales de Educación”. En: *Información tecnológica* 30 (feb. de 2019), págs. 247-254. DOI: 10.4067/S0718-07642019000100247.

- [14] J. C. Dunn†. “Well-Separated Clusters and Optimal Fuzzy Partitions”. En: *Journal of Cybernetics* 4.1 (1974), págs. 95-104. eprint: <https://doi.org/10.1080/01969727408546059>. URL: <https://doi.org/10.1080/01969727408546059>.
- [15] Ricardo Moya. *Selección del número óptimo de Clusters*. 2016. URL: <https://jarroba.com/seleccion-del-numero-optimo-clusters/>.
- [16] *Fersa Bearings — Fabricante europeo de rodamientos*. <https://www.fersa.com>.

Anexos

Anexo A

Tanto el código como los archivos empleados en la realización de este trabajo se pueden obtener del siguiente [repositorio de GitHub](#).

Anexo B

Este anexo contiene una serie de imágenes de gran tamaño que complementan el análisis presentado en el cuerpo principal del trabajo.

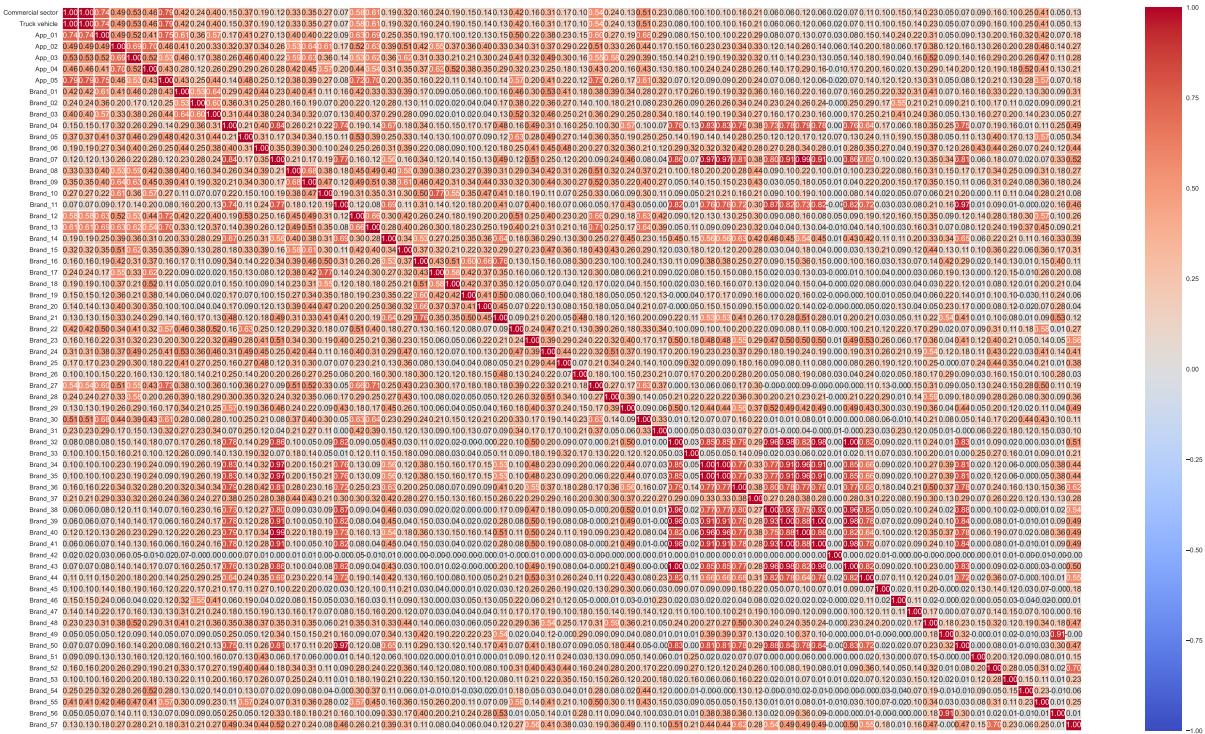


Figura 4.0.1: Matriz de correlaciones entre todas las variables.

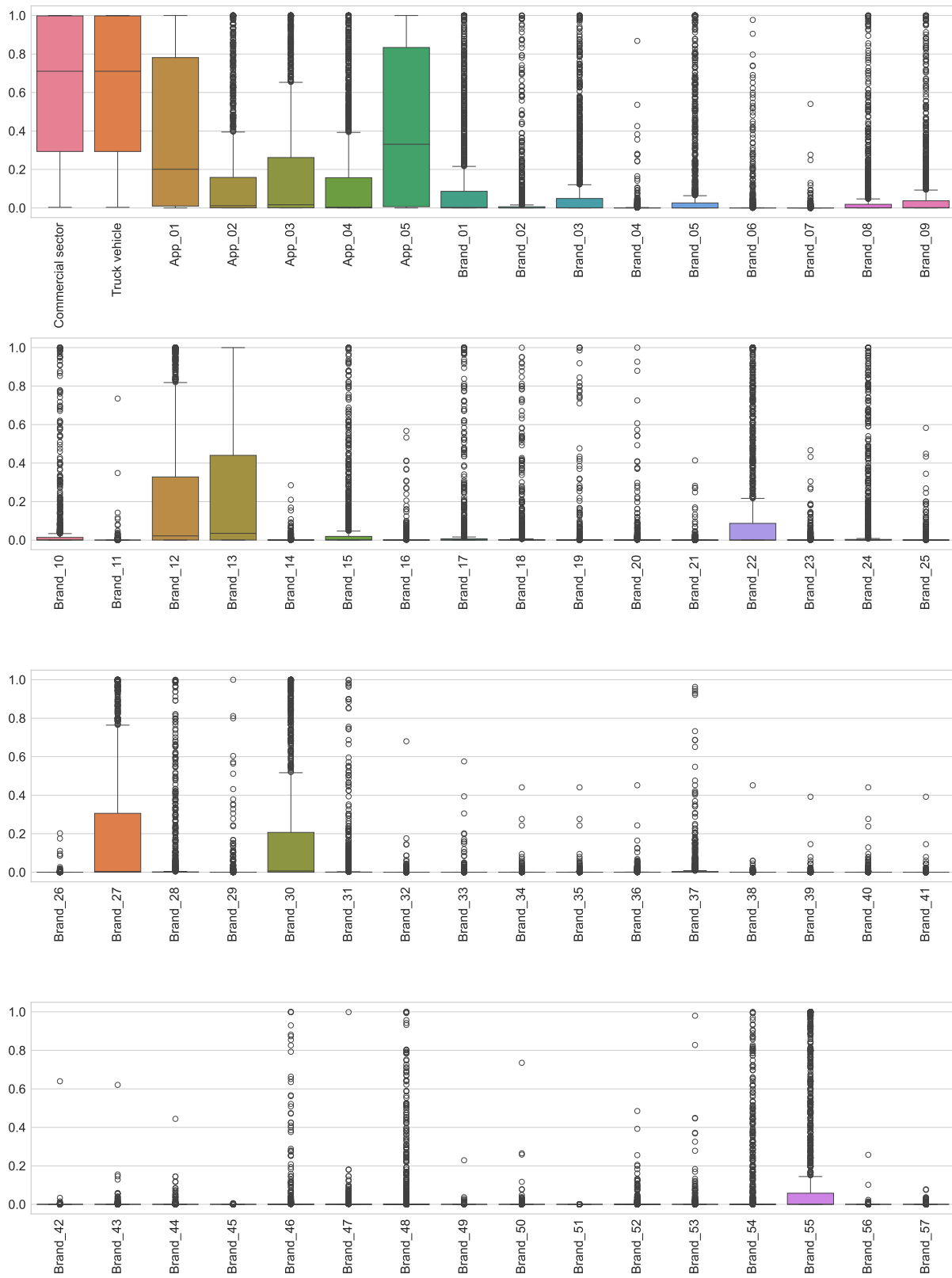


Figura 4.0.2: Diagramas de caja de todas las variables.

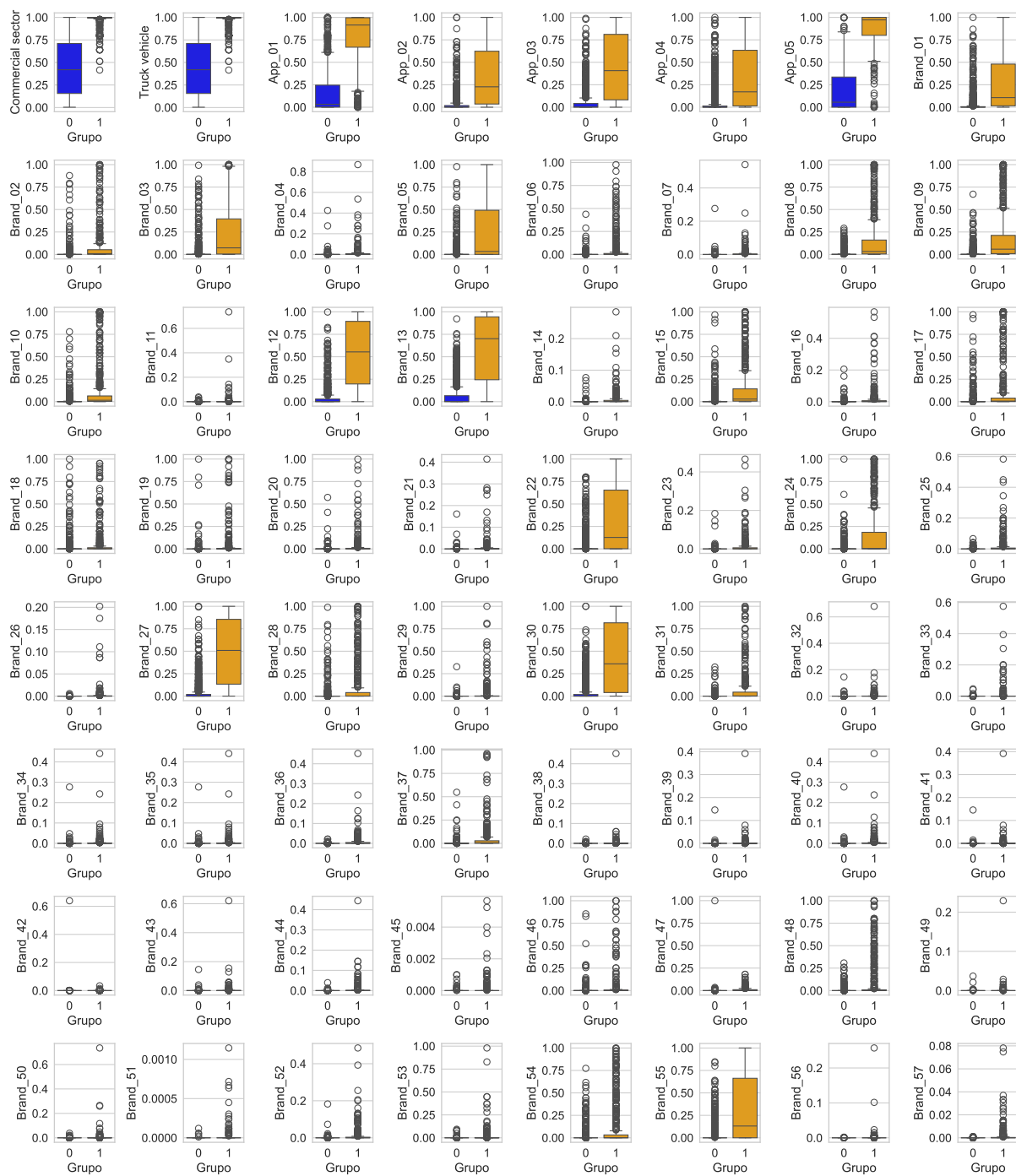


Figura 4.0.3: Diagramas de caja de todas las variables, separados por grupos según el método *K-medias*.

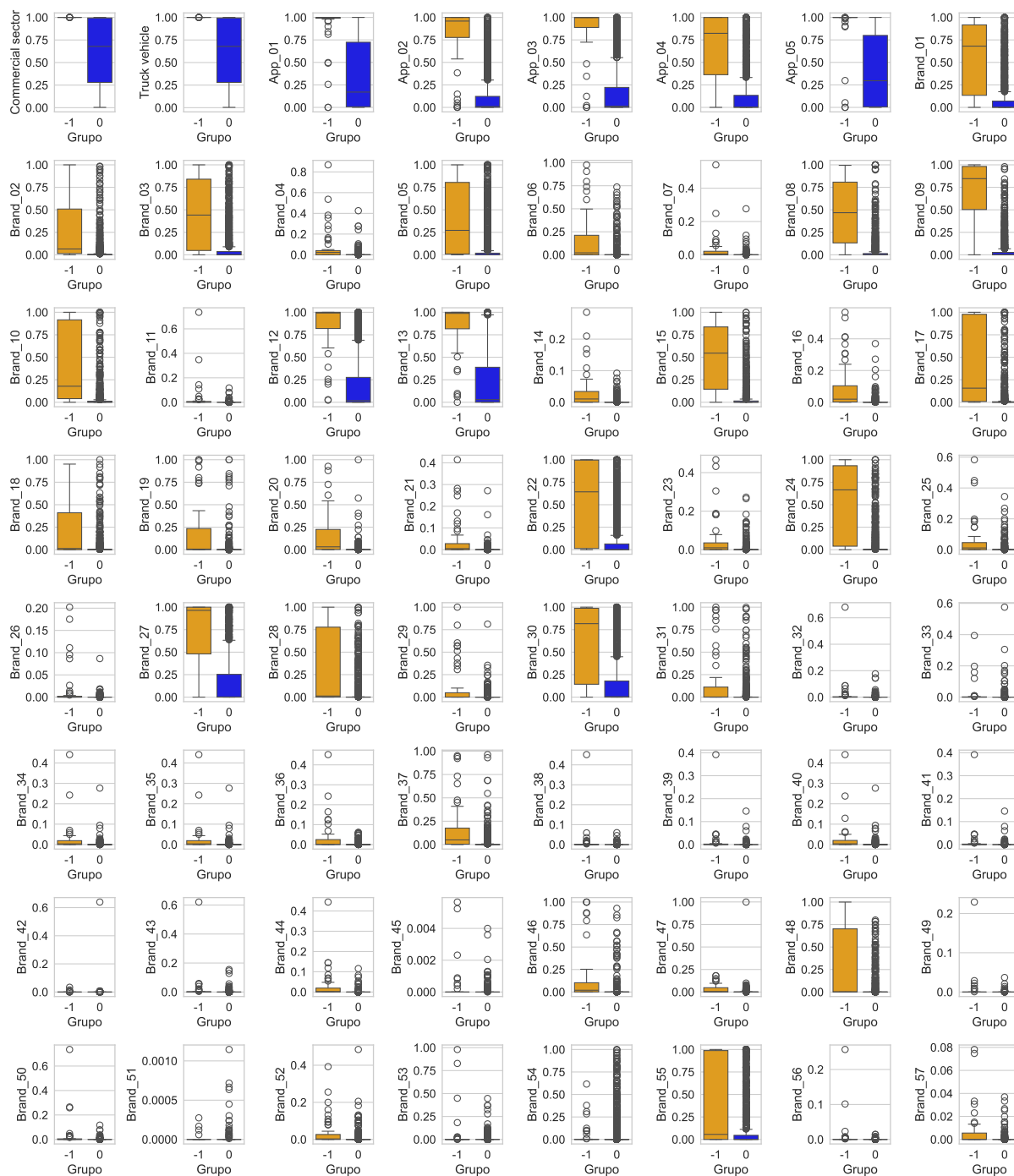


Figura 4.0.4: Diagramas de caja de todas las variables, separados por grupos según el método DBSCAN.

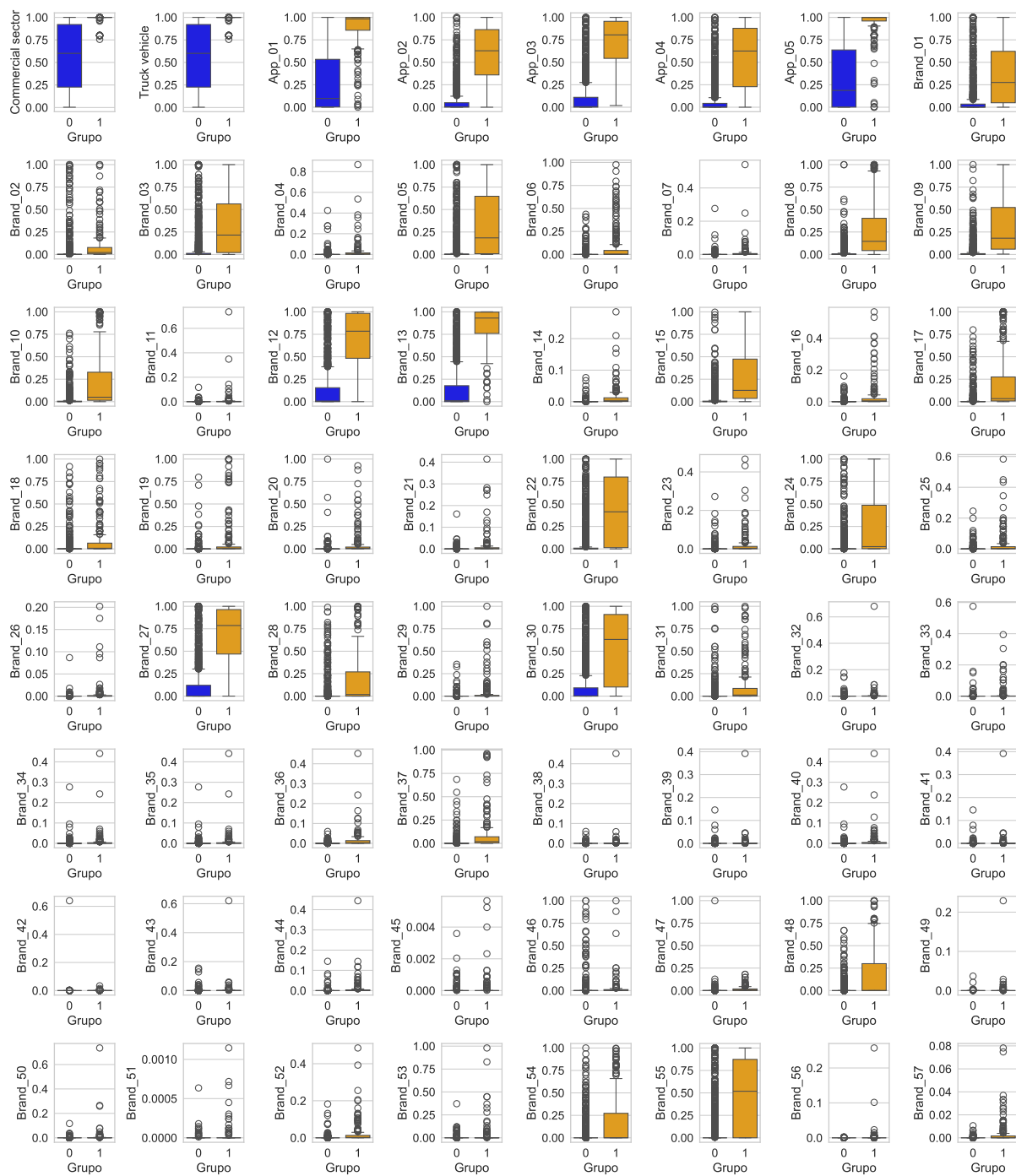


Figura 4.0.5: Diagramas de caja de todas las variables, separados por grupos según el método jerárquico aglomerativo.