Análisis de clustering para la segmentación del mercado: un caso de estudio de una aplicación de una bebida alcohólica en las principales ciudades de Colombia

Cynthia Mariño Santos

Universidad El Bosque

Facultad de Ciencias

Departamento de Matématicas

Programa de Estadística o Matemáticas

Bogotá D.C, Colombia

2023

Análisis de clustering para la segmentación del mercado: un caso de estudio de una aplicación de una bebida alcohólica en las principales ciudades de Colombia

Cynthia Mariño Santos

Tesis como requisito parcial para optar al título de:

Estadístico

Director:

Estadístico Lincoln Ernesto Pérez Pérez

Universidad El Bosque

Facultad de Ciencias

Departamento de Matematicas

Programa de Estadística o Matemáticas

Bogotá D.C, Colombia

2023

Agradecimientos

Me gustaría expresar mi sincera gratitud a todas las personas que contribuyeron a hacer realidad este proyecto.

En primer lugar, mi reconocimiento a mi asesor, Lincoln Perez, cuya dirección, soporte y paciencia han sido cruciales en cada etapa de este trayecto. Su sabiduría y experiencia han sido pilares en el logro de nuestros objetivos.

Un especial agradecimiento a mis colegas de estudio, siempre listos para ayudar y compartir sus valiosas ideas en cada oportunidad.

Extiendo mi gratitud a la Universidad El Bosque por darme la oportunidad de desarrollar este proyecto y por facilitarme los recursos necesarios para su ejecución.

Finalmente, mi más profundo agradecimiento a mi familia y amigos por su apoyo incondicional, motivación y comprensión a lo largo de este proceso. Sin su estímulo y amor, este proyecto no habría sido una realidad.

De corazón, gracias a todos por su invaluable aporte.

Declaración personal del autor

DECLARO QUE

Este documento es una creación propia y original. No he reproducido ni utilizado partes de otras obras sin indicar su fuente de manera explícita y detallada, tanto en el texto como en la bibliografía. Además, no he utilizado datos de terceros sin obtener la autorización correspondiente. Soy plenamente consciente de que el incumplimiento de esta norma puede conllevar sanciones académicas, además de otras posibles acciones legales.

Resumen

Esta tesis propone el uso de técnicas de clustering para la segmentación de mercado en la industria de bebidas alcohólicas en Colombia. Para ello, se plantea el uso de DBSCAN en combinación con técnicas de reducción de dimensionalidad como UMAP, T-SNE y PCA. El estudio busca identificar patrones entre los consumidores de alcohol, generando información que puede ser útil para el desarrollo de estrategias de mercadeo efectivas y personalizadas. Se realiza la sintonización de técnicas de reducción y se comparan los resultados entre los distintos métodos para llegar a una conclusión.

Palabras clave: DBSCAN, UMAP, T-SNE, PCA, clustering.

Abstract

This thesis proposes the use of clustering techniques for market segmentation in the alcoholic beverage industry in Colombia. For this purpose, the use of DBSCAN in combination with dimensionality reduction techniques such as UMAP, T-SNE, and PCA is proposed. The study aims to identify patterns among alcohol consumers, generating information that can be useful for the development of effective and personalized marketing strategies. The tuning of reduction techniques is carried out, and the results from the different methods are compared to reach a conclusion.

Keywords: DBSCAN, UMAP, T-SNE, PCA, clustering.

Índice general

Índice de figuras

Índice de tablas

Lista de Algoritmos

1. Introducción

Históricamente, las segmentaciones de mercado se habían realizado de manera un tanto aleatoria debido a la multitud de características del mercado que se podrían considerar similares o diferentes. Por esta razón, Green, Frank y Robinson [2] proponen el uso de una técnica numérica llamada análisis de conglomerados para hacer coincidir los mercados de prueba potenciales en función de una amplia gama de características que podrían afectar los resultados del marketing de prueba. Este enfoque permitiría la preselección de mercados para minimizar las variaciones no deseadas entre las regiones de prueba.

En la actualidad, existen diversos métodos de clustering aplicados en la segmentación de mercado. Según Reutterer, los dos principales son los métodos basados en modelos y los que se basan en distancias [3]. Asimismo, aunque existen una gran cantidad de trabajos sobre la segmentación de clientes basada en clústeres, pocos de ellos han utilizado algoritmos de agrupamiento basados en densidad. El anterior artículo aplica este método y encontró que este acercamiento proporciona una segmentación del mercado significativa, así como detecta algunos clientes anómalos cuyos hábitos de gasto son diferentes al compararlo con otras técnicas de clustering más comunes [4].

Sin embargo, como menciona Vijendra, las densidades sufren la maldición de la dimensionalidad. Una de las formas tradicionales de abordar el problema de la alta dimensionalidad es realizar una reducción de la dimensionalidad antes del proceso de clustering. Uno de los acercamientos más comunes es la extracción de características como el ACP (Análisis de Componentes Principales), que se usa comúnmente en conjuntos de datos de alta dimensión, pero al ser una técnica lineal, solo tiene en cuenta las dependencias lineales entre variables [5].

Recientemente, se han propuesto métodos alternativos de reducción de dimensionalidad, no lineales, que tienen el potencial de ser superiores. Estos métodos tienen como objetivo crear una representación de menor dimensión del espacio de datos de alta

dimensión mientras se reproducen las relaciones entre los puntos de datos vecinos como UMAP (Uniform Manifold Approximation and Projection) [6] y T-SNE (T-distributed Stochastic Neighbor Embedding) [7].

Teniendo en cuenta lo anterior, en este proyecto, se van a aplicar las técnicas de reducción de dimensionalidad expuestas anteriormente para intentar potenciar el algoritmo de clustering basado en densidad DBSCAN (Density-Based Spatial Clustering of Applications with Noise) en el contexto de segmentación de mercado.

2. Planteamiento del problema

En 2019, el DANE y el Ministerio de Justicia de Colombia realizaron la Encuesta Nacional de Consumo de Sustancias Psicoactivas (ENCSPA) para evaluar el consumo de drogas psicoactivas entre los ciudadanos de 12 a 65 años. El estudio buscó medir la prevalencia del consumo de estas sustancias a lo largo de la vida de los encuestados, así como su uso reciente (en el último año y mes). Los resultados mostraron que un notable 84,0 % de la población en el rango de edad mencionado admitió haber consumido alcohol alguna vez. Las regiones con la mayor prevalencia de consumo de alcohol fueron Boyacá con un 92,9 % y Risaralda con un 92,5 %, mientras que las tasas más bajas se observaron en el Archipiélago de San Andrés y Providencia (54,7 %) y Guainía (65,7 %). En la capital, Bogotá, el 87,0 % de las personas de 12 a 65 años informaron haber bebido alcohol, cifra cercana al 86,1 % registrado en Manizales. Por otro lado, en las áreas metropolitanas de Cali y Medellín, las tasas fueron del 82,8 % y 80,0 % respectivamente. [8]

A partir de lo anterior, se entiende que el consumo de alcohol en Colombia realiza una parte esencial de la cultura del país. Según la Organización Mundial de la Salud (OMS), el consumo nocivo de alcohol es un factor importante en la carga mundial de enfermedades y lesiones, y es responsable de 3 millones de muertes cada año. Además, el consumo de alcohol ha sido asociado con una variedad de problemas de salud, como enfermedades hepáticas, trastornos mentales y problemas sociales [9].

Como se explica en [10], comprender los factores que se asocian con la preferencia y el consumo de alcohol en poblaciones no clínicas permite el desarrollo de productos y oportunidades de innovación para la industria de bebidas alcohólicas. Asimismo, a nivel de mercadeo, agrupar estos factores puede permitir la personalización de sus estrategias de marketing para dirigirse a segmentos de clientes específicos, lo que puede llevar a campañas de marketing más eficientes y exitosas. Estas agrupaciones, también conocidas como segmentaciones, pueden basarse en distintas características de los

consumidores, como datos demográficos, comportamientos y geografía. Sin embargo, el enfoque más efectivo es la segmentación basada en el comportamiento del usuario, ya que tiene en cuenta las actitudes, opiniones, intereses y otros criterios relevantes del consumidor.

Por lo tanto, en este trabajo de investigación, se quiere realizar una segmentación de mercado a partir de los datos recopilados de una encuesta que hace parte de una investigación de mercado de una bebida alcohólica. Los datos empleados en este proyecto se obtuvieron mediante una encuesta de opinión que involucró preguntas realizadas en campo con el fin de realizar una segmentación del mercado de potenciales consumidores de una bebida alcohólica. Los participantes de este estudio se seleccionaron mediante un proceso de muestreo que consideró los resultados de dos investigaciones realizadas por el DANE. En primer lugar, se tomaron en cuenta las proyecciones de población departamental para el período 2018-2050, las cuales se calcularon a partir de los datos del Censo Nacional de Población y Vivienda (CNPV) de 2018. Además, se consideraron los resultados de la Encuesta Nacional de Consumo de Sustancias Psicoactivas (ENCSPA) de 2019, cuyo objetivo principal era obtener información estadística necesaria para estimar el alcance del consumo de sustancias psicoactivas en Colombia dentro de la población de 12 a 65 años [8].

El muestreo se centra en regiones específicas, como Antioquia, Atlántico y Bogotá, entre otras. A partir de aquí, se considera la población masculina en varios grupos de edad, ajustando estos números mediante un factor de 0.380, que representa la prevalencia mensual de consumo de bebidas alcohólicas en hombres. Luego, se consolidan los datos en categorías de edad más amplias, aplicando ajustes únicos a cada región. Se sigue un enfoque similar para la población femenina, pero con un factor de escala de 0.227, basado en la prevalencia mensual de consumo. Además, se exploran divisiones socioeconómicas o estratos, enfocándose en regiones y ajustando los datos utilizando multiplicadores específicos para cada estrato. Aunque hay seis posibles estratos, los cálculos se concentran en los últimos cuatro debido al conocimiento anecdótico que se tiene sobre los consumidores del tipo de bebida alcohólica abordada en el cuestionario. Finalmente, se combinan todos los datos refinados, que incluyen las poblaciones masculina y femenina, junto con las proyecciones de estrato, región, prevalencia de consumo

y edad, para obtener un tamaño de muestra de 671 encuestados.

El objetivo es agrupar a los consumidores por lo tanto, se propone la realización de una segmentación de mercado utilizando la técnica de clustering DBSCAN para identificar grupos de consumidores de la bebida alcohólica en estudio en función de sus características y comportamientos. Se busca identificar patrones y tendencias en los datos, lo que permitirá a la empresa comprender mejor a sus clientes y ajustar su estrategia de marketing para dirigirse de manera efectiva a cada uno de los grupos identificados.

Esta información es especialmente relevante en un contexto en el que la competencia en la industria de bebidas alcohólicas es cada vez más fuerte y diversa, y se busca diferenciarse y ser más efectivo en la comercialización de los productos. La segmentación de mercado puede proporcionar una ventaja competitiva para la empresa al permitirle desarrollar una estrategia de marketing más eficaz y personalizada, lo que a su vez puede mejorar la satisfacción y lealtad de los clientes lo que puede aumentar las ventas.

3. Justificación

Este proyecto tiene como objetivo principal aplicar técnicas de clustering en un conjunto de datos de consumidores de bebidas alcohólicas para identificar patrones de comportamiento y preferencias en los clientes potenciales de una marca en particular. La segmentación de mercado es una herramienta importante para las empresas, ya que les permite centrarse en grupos específicos de consumidores con necesidades y deseos similares, lo que a su vez puede ayudarles a diseñar estrategias de marketing más efectivas y personalizadas.

Además, en un contexto en el que la salud y el bienestar son cada vez más importantes para los consumidores, es fundamental que las empresas de bebidas alcohólicas comprendan los patrones de consumo de los clientes y las posibles consecuencias para la salud. La Organización Mundial de la Salud (OMS) ha señalado que el consumo de alcohol está relacionado con una serie de problemas de salud, incluyendo enfermedades cardiovasculares, cáncer y trastornos mentales y del comportamiento [9]. Por lo tanto, es importante que las empresas en la industria del alcohol se aseguren de que están comercializando sus productos de manera responsable y ética.

En este sentido, el proyecto también busca contribuir a la responsabilidad social empresarial al identificar patrones de consumo de alcohol y segmentar el mercado de manera que las empresas puedan dirigirse a los consumidores de manera responsable y ética. Los resultados de este proyecto podrían ayudar a la empresa a la que pertenece la marca de bebidas alcohólicas a comprender mejor a sus clientes potenciales y diseñar estrategias de marketing más efectivas y responsables. Además, al segmentar el mercado, la empresa podría optimizar sus recursos y reducir costos al enfocarse en grupos específicos de consumidores.

4. Objetivos

4.1. Objetivo general

Evaluar el desempeño de diversas técnicas de reducción de dimensionalidad, como UMAP, T-SNE y PCA, en la mejora de un algoritmo DBSCAN para un escenario de segmentación de mercado, utilizando datos recolectados a través de un cuestionario diseñado para identificar posibles consumidores de una bebida alcohólica.

4.2. Objetivos específicos

1. Analizar las características demográficas, comportamentales e intereses de los participantes del cuestionario para identificar patrones y tendencias en el mercado.
2. Aplicar técnicas de reducción de dimensionalidad a los datos recopilados con el objetivo de elegir la que mejor respalde la segmentación en base a métricas de evaluación.
3. Generar perfiles de clientes a partir de las conclusiones obtenidas del agrupamiento para obtener una comprensión más completa de las características de los posibles clientes.

5. Marco Teórico

5.1. Antecedentes

En la segmentación del mercado cuatro pilares se han vuelto los más destacados en investigaciones [11]: segmentación geográfica (por ejemplo, mercados divididos según la región geográfica, densidad poblacional o clima); segmentación demográfica (por ejemplo, mercados divididos por criterios como edad, género, tamaño y tipo de familia); segmentación psicográfica (por ejemplo, mercados divididos según variables relacionadas con el estilo de vida); y segmentación conductual (por ejemplo, mercados divididos en función de la ocasión de compra, beneficios buscados y estatus del consumidor).

Uno de los primeros usos de clusterización en pruebas de mercado se dio en 1967 pues esto comprendía una tarea importante si se pretendía realizar comparaciones fiables entre mercados. Anteriormente esto se ha llevado a cabo de manera bastante arbitraria, en gran parte debido a la gran cantidad de características del mercado en las que los mercados pueden considerarse similares o diferentes. Pero Paul E. Green, Ronald E. Frank, Patrick J. Robinson [2], propusieron un procedimiento numérico, el análisis de conglomerados, para emparejar mercados de prueba potenciales en función de una amplia variedad de características que podrían afectar los resultados de las pruebas de mercado. De esta manera, los mercados pueden preseleccionarse para reducir la variabilidad no deseada entre las áreas de prueba.

Posteriormente, la aplicación de técnicas de clustering usadas exclusivamente para el caso de segmentación fueron ganando popularidad, pues había quienes afirmaban, como Myers and Tauber [12], que existían segmentos de mercado dentro del campo de la investigación de segmentación que eran agrupaciones naturales de personas claramente definidas. Por esto, hay quienes afirman, como Wedel y Kamakura [13], que desde hace tiempo, el análisis de clusters ha sido el enfoque líder y más favorecido para

la segmentación de mercados, en gran medida debido a su amplia presencia en una variedad de paquetes de software estándar

Asimismo, los autores mencionados anteriormente, también confirman que aunque hay una gran cantidad disponible de métodos de clustering no jerárquicos; K-means es el más conocido y utilizado de esos procedimientos dentro del contexto de segmentación del mercado. Pero como menciona A. S. M. Shahadat Hossain [4], a pesar de la existencia de varias investigaciones relacionadas con el uso de clustering para la segmentación de clientes, no hay tanta exploración del uso de algún algoritmo de clustering basado en la densidad.

Algunos ejemplos de los algoritmos mencionados se encuentran en el trabajo de Zakrzewska y Murlewski [14], donde comparan algoritmos de clustering en casos de alta dimensionalidad con ruido para el caso de segmentación de clientes bancarios, donde se aplican tres algoritmos: DBSCAN basado en densidad, k-means y un algoritmo de clustering en dos fases. Comparaciones entre distintos tipos de clustering también se realizan por parte de Koul y Philip [15] en el campo del comercio electrónico, donde se analizan varios métodos de clustering para el caso de segmentación de clientes como k-medias y algoritmos basados en densidad. Asimismo, A. S. M. Shahadat Hossain [4] confirma que los resultados obtenidos a través de la aplicación de estos dos algoritmos demuestran que cualquiera de ellos se puede utilizar para la segmentación de clientes. Sin embargo, en comparación con k-means, DBSCAN ofrece una característica adicional que puede identificar clientes poco comunes que exhiben comportamientos de gasto distintos y también afirma que esto es altamente efectivo para garantizar la satisfacción del cliente y ganancias óptimas.

Por otro lado, Ospina [16] implementó 3 modelos de clusterización combinados con métodos de reducción de dimensionalidad para la segmentación del mercado de clientes de EPS Sura. Los modelos fueron evaluados utilizando diversos factores como el número de clusters, la existencia de ruido y la agrupación de los datos, entre otros aspectos. Se descubrió que el uso de UMAP para reducir la dimensionalidad de los datos proporciona resultados altamente efectivos cuando se combina con la técnica de segmentación HDBSCAN, en comparación con el uso del método K-Means.

5.2. Segmentación del mercado

El concepto de la segmentación del mercado nace a mediados de la década de los 50's como una estrategia alternativa de mercadeo disponible para planeadores y comerciantes en el contexto de competencia imperfecta. La segmentación de mercado implica considerar un mercado diverso, definido por demandas variadas, como una colección de mercados más pequeños y homogéneos, cada uno con sus propias preferencias de producto entre segmentos significativos del mercado. Esta práctica se origina en las preferencias y deseos de los consumidores o usuarios [17].

De acuerdo a Ernawati, Baharin y Kasmin [18], la segmentación de clientes se puede utilizar para adaptar las estrategias de marketing a cada grupo individualmente. Al hacerlo, las empresas pueden dirigirse de manera efectiva a clientes valiosos y desarrollar actividades de marketing que satisfagan sus necesidades y preferencias específicas.

5.3. Clustering

El clustering, también conocido como análisis de conglomerados, es una tarea general que consiste en agrupar un conjunto de objetos de manera que los objetos con características similares permanezcan en el mismo grupo. Es un aspecto vital de la minería de datos y una técnica comúnmente empleada en varios campos [19].

En *"Data Mining: Concepts and Techniques"* de J. Han, M. Kamber y J. Pei[20], se describe una variedad de algoritmos de clustering empleados en el análisis de datos. Estos incluyen métodos basados en centroides, donde se utilizan puntos centrales para representar los clusters; técnicas basadas en la densidad, que agrupan datos según la concentración de puntos en una región específica; algoritmos jerárquicos, que construyen clusters mediante la creación de jerarquías basadas en la proximidad de los datos; y métodos basados en distribución, que emplean modelos estadísticos para identificar la distribución de los datos. Cada uno de estos enfoques tiene sus propias ventajas y se utiliza según el tipo de datos y el objetivo del análisis [20].

5.3.1. DBSCAN

En 1996, Ester et al. propusieron DBSCAN como el algoritmo pionero para el agrupamiento basado en la densidad. Su propósito era agrupar conjuntos de datos con formas arbitrarias en bases de datos de alta dimensión, tanto espaciales como no espaciales, teniendo en cuenta el ruido [21].

La idea clave de DBSCAN es que para cada objeto de un grupo, el vecindario de un radio dado (Eps) debe contener al menos un número mínimo de objetos ($MinPts$), lo que significa que la cardinalidad del vecindario tiene que exceder algún umbral. El vecindario 'ϵ' de un punto arbitrario 'p' se define como:

$$N_{Eps} = \{q \in D \mid \text{dist}(p, q) < Eps\}$$

Aquí, D es el conjunto de datos. Si los vecindarios 'ϵ' de un punto 'p' contienen al menos un número mínimo de puntos, entonces este punto se llama punto central. El punto central se define como:

$$N_{Eps}(P) > MinPts$$

Eps y $MinPts$ son los parámetros especificados por el usuario, que representan el radio del vecindario 'ϵ' y el número mínimo de puntos en el vecindario de un punto central, respectivamente. Si esta condición no se cumple, este punto no se considera como un punto central.

DBSCAN busca los clústeres revisando el vecindario 'ϵ' de cada objeto en el conjunto de datos. Si el vecindario 'ϵ' de un objeto contiene más de $MinPts$, un nuevo clúster se crea con objeto central 'p'. Posteriormente, el algoritmo reúne de forma iterativa aquellos objetos que se encuentran directamente dentro del alcance de densidad de los objetos centrales, lo cual podría resultar en la integración de un nuevo clúster basado en la densidad. Este proceso concluye una vez que no es posible añadir más objetos a los clústeres existentes [22].

Algorithm 1 Pseudocódigo del algoritmo DBSCAN. Tomado de [1]

{DBSCAN(D, ε, minPts)}
{D is a set of unclassified points}
{ε is the maximum distance}
{minPts is the minimum points to form a cluster}
Initialize cluster id $C = 0$
for each unclassified point $p \in D$ **do**
    $N_\varepsilon(p) = \text{RangeQuery}(p, \varepsilon)$
    if $|N_\varepsilon(p)| \geq \text{minPts}$ **then**
        Set p's cluster id to C
        $\text{ExpandCluster}(p, N_\varepsilon(p), C, \varepsilon, \text{minPts})$
        $C \leftarrow C + 1$
    else
        Label p as noise
    end if
end for

5.3.2. Validación de clústeres

La validación de clústeres es una paso esencial para evaluar los de resultados de la agrupación. Existen tres tipos de procedimientos de validación de clústeres: validación externa de clústeres, validación interna de clústeres y validación relativa de clústeres [23].

La validación interna de clústeres se basa en la información intrínseca de los datos y utiliza la información interna del proceso de clustering para evaluar el procedimiento. Por lo tanto, los índices internos solo utilizan la información sobre el conjunto de datos y los resultados de agrupamiento, mientras que los índices externos también requieren etiquetas de datos independientes. En este caso se usará el índice Calinski-Harabasz como método de validación interna donde se busca comparar la varianza entre grupos con la varianza dentro de cada grupo. Esté índice no tiene límite y un valor más alto de este indica una mejor separación entre clústeres [24].

5.4. Métodos de reducción de dimensionalidad

Los métodos de reducción de la dimensionalidad se pueden categorizar principalmente en dos grupos: lineales y no lineales. Los métodos lineales, como el análisis de componentes principales (ACP), buscan construir nuevas variables colectivas mediante la realización de combinaciones lineales de las variables de entrada. Por el contrario, los métodos no lineales, como el método de incrustación de vecinos estocásticos distribuidos (t-SNE) y el método de aproximación y proyección de variedad uniforme (UMAP), construyen nuevas variables colectivas mediante el mapeo las variables de entrada a una función no lineal [25].

Reflexionando sobre las estrategias de análisis de datos, es fundamental considerar las orientaciones proporcionadas por I. T. Jolliffe en su obra *"Principal Component Analysis for special types of data"* [26]. En en este texto se enfatiza la importancia de seleccionar cuidadosamente el método de reducción de dimensionalidad más adecuado para cada conjunto de datos. Jolliffe explica que la elección debe basarse en las características únicas y los objetivos específicos del análisis de datos en cuestión. Por lo tanto, recomienda evaluar distintas técnicas de reducción de dimensionalidad, comparando sus resultados para determinar cuál es más efectivo en revelar las estructuras subyacentes del conjunto de datos o en simplificar la información sin sacrificar aspectos críticos. Este enfoque comparativo es esencial para facilitar un análisis de datos más eficiente y preciso [26].

5.4.1. UMAP

UMAP, cuyo acrónimo proviene de "Uniform Manifold Approximation and Projection", constituye un algoritmo de aprendizaje automático empleado para la reducción de la dimensionalidad y la visualización de conjuntos de datos que poseen una alta cantidad de dimensiones. Su creación se atribuye a McInnes y Healy en el año 2018 [6]. La técnica subyacente en la reducción de dimensionalidad de UMAP se sustenta en tres suposiciones fundamentales. En primer lugar, se postula que los datos se distribuyen de manera uniforme en una variedad de Riemann. Seguidamente, se asume que la

métrica de Riemann es constante en un contexto local. Por último, se considera que la estructura subyacente, denominada colector, presenta conexiones locales. UMAP sigue un enfoque basado en grafos, donde su principal premisa consiste en la construcción de una representación gráfica UMAP de baja dimensionalidad y ponderada para cada punto de datos en el espacio de alta dimensionalidad. Este proceso busca lograr una correspondencia cercana entre el gráfico ponderado y los datos originales. Los vectores propios de esta representación gráfica en baja dimensión (k-dimensional) se utilizan posteriormente para representar los puntos de datos originales de manera más eficiente y comprensible [6].

El proceso comienza con la construcción de grafos ponderados de k-vecinos, utilizando el algoritmo del vecino más cercano, ya que UMAP emplea este enfoque. UMAP trabaja con un conjunto de datos de entrada denotado como $X = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R}^M$ con el objetivo de encontrar una representación de baja dimensionalidad óptima $\{y_1, \ldots, y_N \mid y_i \in \mathbb{R}^k\}$ tal que $k < M$. En este contexto, se define una métrica $d : X \times X \to \mathbb{R}^+$. Luego, se establece un hiperparámetro $k \ll M$, y para cada punto x_i en el conjunto de datos de entrada, se calculan los k vecinos más cercanos utilizando la métrica d. Se selecciona ρ_i para garantizar que al menos un punto de datos esté conectado a x_i y tenga un peso de borde igual a 1 [16].

$$\rho_i = \min\{d(x_i, x_j) \mid 1 \leq j \leq k, d(x_i, x_j) > 0\}$$

Se establece σ_i como un parámetro de escala de longitud, tal que:

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k$$

Se establece un grafo dirigido ponderado $\overline{G} = (V, E, \omega)$, donde V corresponde al conjunto de vértices del conjunto de datos X, E representa el conjunto de bordes $E = \{(x_i, x_j) \mid 1 \leq h \leq k, 1 \leq i \leq N\}$ y ω es la función de peso de los bordes definida como:

$$\omega(x_i, x_j) = \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$$

UMAP intenta definir un grafo ponderado no dirigido G a partir de graficar $\overline{G}$ por simetrización. Sea A la matriz de adyacencia del grafo $\overline{G}$. Se puede obtener una matriz simétrica

$$\mathbf{B} = \mathbf{A} + \mathbf{A}^T - \mathbf{A} \otimes \mathbf{A}^T$$

Una vez que se ha establecido el grafo de k vecinos ponderados, se avanza hacia la creación de un grafo de menor dimensión en el que UMAP emplea fuerzas de atracción y repulsión para determinar las coordenadas y_i y y_j en el nuevo gráfico, de acuerdo a las especificaciones de UMAP.

$$\frac{-2ab\,\|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2}\omega(x_i, x_j)(y_i, y_j)$$

Donde a y b son valores ajustables, y la intensidad de la fuerza repulsiva se describe mediante:

$$\frac{2b}{(\varepsilon + \|y_i - y_j\|_2^2)(1 + a(y_i - y_j)^{2b})}(1 - \omega(x_i, x_j))(y_i - y_j)$$

Es posible seleccionar un valor ε de pequeña magnitud para prevenir la ocurrencia de un denominador igual a cero. El objetivo principal radica en determinar las coordenadas de baja dimensión óptimas $\{y_i\}_{i=1}^N, y_i \in \mathbb{R}^k$, que buscan minimizar la entropía cruzada de los bordes con respecto a los datos originales en cada punto [16].

5.4.2. T-SNE

t-SNE (t-distributed Stochastic Neighbor Embedding) es un algoritmo de reducción dimensional no lineal que resulta muy adecuado para la reducción de datos de alta dimensión a un espacio de dos o tres dimensiones. Hay dos etapas principales en t-SNE. Primero, construye una distribución de probabilidad sobre pares de datos de modo que a pares de puntos cercanos se les asigna una alta probabilidad, mientras que a pares de puntos más distantes se les asigna una baja probabilidad. En la segunda etapa, t-SNE define una distribución de probabilidad en el espacio de incrustación que es similar a la del espacio dimensional original de alta dimensionalidad, con el objetivo de minimizar la divergencia de Kullback-Leibler (KL) entre ambas [27].

El principal hiperparámetro considerado en t-SNE es la perplejidad, que puede interpretarse como una medida de suavidad del número efectivo de vecinos. El rendimiento de t-SNE es bastante robusto a los cambios en la perplejidad, siendo los valores típicos entre 5 y 50 [7].

5.4.3. ACP

El análisis de componentes principales (ACP) reduce la dimensionalidad de los datos proyectando cada punto sobre algunos componentes principales. Estos representan una versión de menor dimensionalidad de los datos originales mientras preservan la variación de los mismos. Los componentes en ACP son combinaciones lineales de las variables de entrada y son ortogonales entre sí. Dadas dos variables, x y y, su covarianza muestral mide cómo estas dos variables varían conjuntamente de sus medias $\overline{x}$ y $\overline{y}$, basado en n observaciones.

$$\sigma(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

En ACP, se construye una matriz de covarianza C de tamaño $p \times p$, para un conjunto de datos con p variables. Cada elemento C_{ij} representa la covarianza entre dos variables, x_i y x_j, es decir, $C_{ij} = \sigma(x_i, x_j)$. Los vectores propios de C son los componentes del ACP. Los valores propios de C indican la importancia de cada componente en el conjunto de datos; a mayor magnitud del valor propio, mayor es la contribución de su correspondiente componente principal. En general, los vectores propios con los valores propios más altos se seleccionan como componentes principales para formar un espacio 2D o 3D para la proyección de los datos [25].

6. Metodología

La siguiente figura representa los pasos que se seguirán en el desarrollo de este proyecto, abarcando desde la fase inicial de análisis hasta la etapa práctica de interpretación de los resultados. A continuación, se detalla cada uno de estos pasos de manera más exhaustiva.

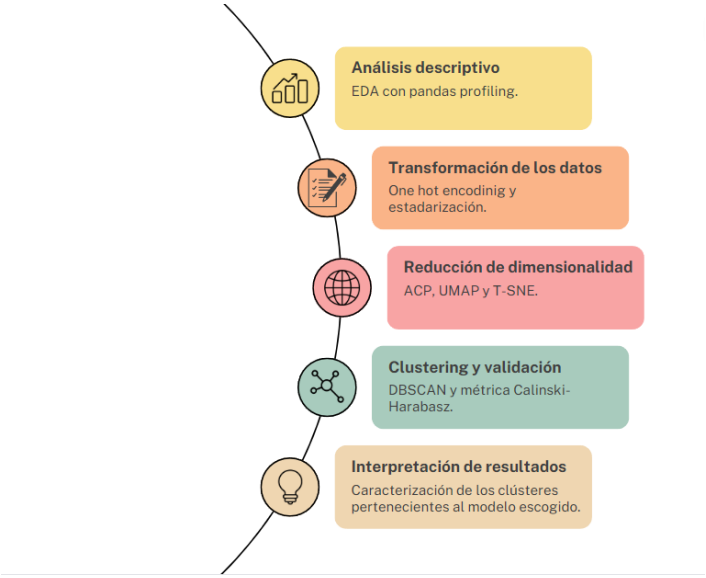


Figura 6.1: Metodología

6.1. Análisis descriptivo

Una vez preparada la información, se lleva a cabo el análisis descriptivo de los resultados obtenidos. Este análisis busca identificar las características principales de los datos, como su tendencia central, variabilidad, distribución y las relaciones existentes entre las distintas variables. El propósito del análisis descriptivo es ofrecer un resumen claro, conciso y comprensible de los datos, que sirva de base para análisis más profundos o para la toma de decisiones informadas.

6.2. Transofrmación de los datos

En la siguiente etapa, se emplea la técnica de codificación one-hot para la transformación de variables categóricas en numéricas. Esto implica la creación de nuevas columnas binarias para cada categoría de la variable categórica original, asignando un valor de 1 si la categoría es relevante y 0 en caso contrario. Este proceso incrementa la cantidad de variables en comparación con el conjunto de datos original.

A continuación, se lleva a cabo una estandarización, también conocida como estandarización de la puntuación z. Este método transforma los datos numéricos de modo que tengan una media de cero y una desviación estándar de uno. Se logra esto restando la media del conjunto de datos a cada dato y dividiendo el resultado por la desviación estándar. Los datos resultantes, por ende, presentan una media de cero y una desviación estándar de uno. Este paso es crucial, ya que normaliza los datos que originalmente se miden en distintas escalas, evitando así que variables con valores numéricos más altos dominen el análisis y conduzcan a conclusiones erróneas o sesgos en los resultados.

6.3. Aplicación técnicas de reducción de dimensionalidad

En esta fase de reducción de la dimensionalidad, el enfoque se centra principalmente en dos técnicas: UMAP y t-SNE. Adicionalmente, se incorporará el Análisis de Componentes Principales (ACP) como un modelo de referencia para comparación con las otras técnicas. Este proceso comienza con la carga del conjunto de datos, previamente preparado en la fase inicial, que contiene los datos de segmentación de los clientes. Posteriormente, se definen y aplican los parámetros específicos para cada una de estas técnicas de reducción de la dimensionalidad.

6.4. Clusterización

Una vez completada la fase de reducción de la dimensionalidad, el proyecto avanza hacia la implementación de modelos con el objetivo de identificar agrupaciones en los datos de los encuestados, empleando para ello el algoritmo DBSCAN. En esta etapa, se realiza un proceso detallado que implica la generación de gráficos basados en distintos conjuntos de parámetros para DBSCAN. Tras una cuidadosa selección y optimización de estos parámetros, se consolidan cinco modelos principales:

- UMAP y DBSCAN
- UMAP y DBSCAN (con hiperparámetros optimizados)
- T-SNE y DBSCAN
- T-SNE y DBSCAN (con hiperparámetros optimizados)
- ACP y DBSCAN

6.5. Validación

La validación de los modelos propuestos se realizará evaluando múltiples características. Primero, se efectuará una comparación interna buscando asegurar que los clústeres sean compactos, estén claramente separados entre sí y mantengan conectividad dentro de cada grupo. Para la selección de parámetros de los modelos mencionados anteriormente, se utilizarán el índice Calinski-Harabasz, el coeficiente de silueta y el índice de Davies-Bouldin. Durante el proceso de selección del modelo, se llevará a cabo un análisis detallado basado en estas métricas predefinidas, complementado con la representación gráfica de los clústeres.

6.6. Interpretación de los Resultados

Una vez seleccionado el modelo más adecuado, se procederá a la caracterización detallada de los clústeres generados por dicho modelo. Este análisis permitirá identificar

las similitudes y diferencias entre los grupos, ofreciendo una visión detallada de las dinámicas y comportamientos dentro de estos clústeres.

7. Resultados

7.1. Análisis descriptivo

Para el desarrollo de este proyecto se lograron obtener 696 registros, obtenidos a partir de las respuestas obtenidas al cuestionario suministrado. Era importante que lo encuestados cumplieran cierto requisito para participar en el análisis, como el hecho de no trabajar en áreas relacionadas a la temática que se iba a abordar o haber participado en investigaciones de mercado, entrevistas o encuestas sobre bebidas alcohólicas recientemente.

Se observa que hay varias variables dentro del cuestionario que tienen respuestas abiertas y aunque estas sean de gran valor analítico, se remueven del análisis pues no son fácilmente cuantificables y dada la cantidad de respuestas distintas categorizar sería una tarea extensa. Por ende, para la realización del proyecto solo se utilizan 51 variables.

En primer lugar, se realiza un análisis descriptivo de las variables incluídas para así tener una visión general del conjunto de datos. Para esto, se hizo uso de la librería de Python, pandas-profiling cuyo objetivo principal es llevar a cabo un análisis exploratorio de datos (EDA). Así se proporcionó un análisis detallado del conjunto de datos y permite exportar este análisis de datos en formato HTML. A continuación se presenta un resumen de las variables demográficas del cuestionario:

La edad de los participantes restantes varió desde un mínimo de 18 años hasta un máximo de 65 años, con una edad media de 37.8 años y una edad mediana de 36 años. La distribución de edades de los encuestados, que se puede visualizar en la figura 7.1, tuvo una desviación estándar de 11.3 años, lo que indica un grado moderado de variabilidad.

En la Figura 7.2, se refleja de manera visual la distribución de los encuestados

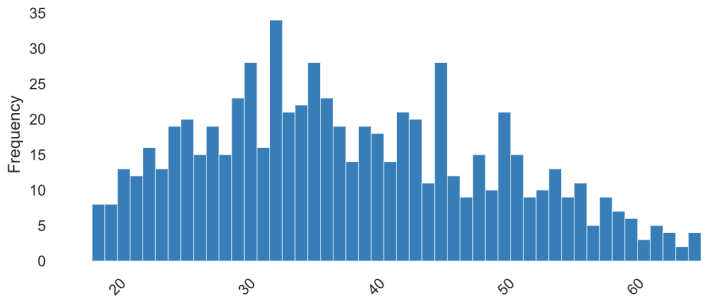


Figura 7.1: Edad

Value	Count	Frequency (%)
3	395	56.8%
4	202	29.0%
5	77	11.1%
6	22	3.2%

Figura 7.2: Estrato

según su estrato socioeconómico. Los resultados indican que un notable 56.8 % pertenece al estrato 3, siendo la categoría más representada en la muestra. Le sigue el estrato 4, al cual corresponde el 29.0 % de los encuestados. Además, el estrato 5 está representado por el 11.1 % de los participantes, mientras que el estrato 6 cuenta con un 3.2 % de representación en la población encuestada.

Value	Count	Frequency (%)
Bogotá	327	47.0%
Medellín	180	25.9%
Cali	122	17.5%
Barranquilla	66	9.5%
zipaquira	1	0.1%

Figura 7.3: Ciudad

La Figura 7.3 muestra el análisis demográfico de la muestra, se observa que de los encuestados, un significativo 47 % correspondía a residentes de Bogotá, siendo la ubicación más representativa en la encuesta con un total de 327 personas. Le sigue Medellín, con un 25.9 %, equivalente a 180 participantes. Cali también estuvo representada, con un 17.5 %, lo que equivale a 122 personas en la muestra. Por otro lado, Barranquilla contribuyó con un 9.5 %, contando con la participación de 66 individuos. Finalmente, solo un individuo, representando el 0.1 % de la muestra, provenía de Zipaquirá.

Las demás variables tenían respuesta likert con declaraciones de mito de la industria, intención de compra, intención de consumo, posicionamiento, entre otros. Asi-

mismo, una variable que incluía percepción de valor, donde los encuestados detallaban cuánto estarían dispuestos a pagar por el producto y otra variable dicotómica para medir la intención de prueba.

Siguiendo la transformación de los datos mencionada en la metodología se realizó una función representada en la Figura 7.4 que con el objetivo de automatizar la determinación de los hiperparámetros de DBSCAN, específicamente eps y min samples, basándose en el punto de codo en el gráfico de distancia-k. Para luego evualuar la calidad del agrupamiento utilizando tres métricas: índice Calinski-Harabasz, índice Davies-Bouldin y coeficiente de silueta.

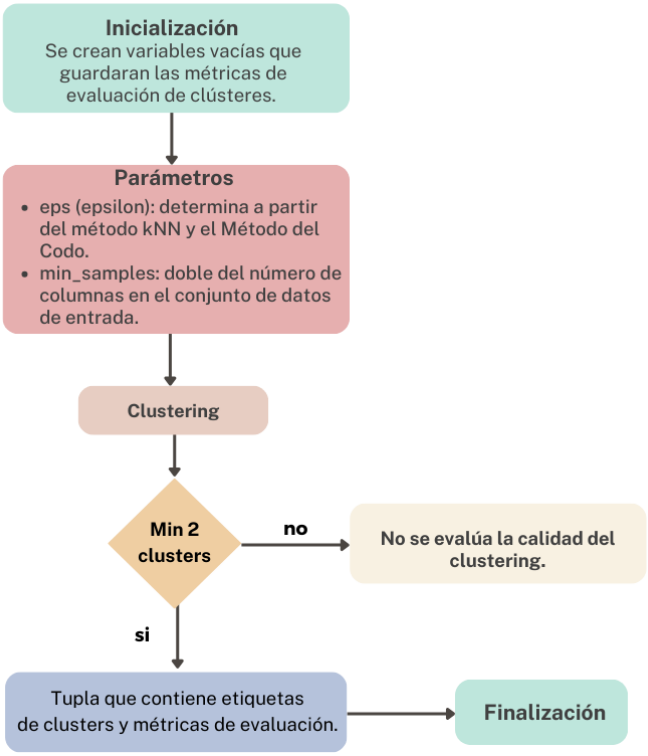
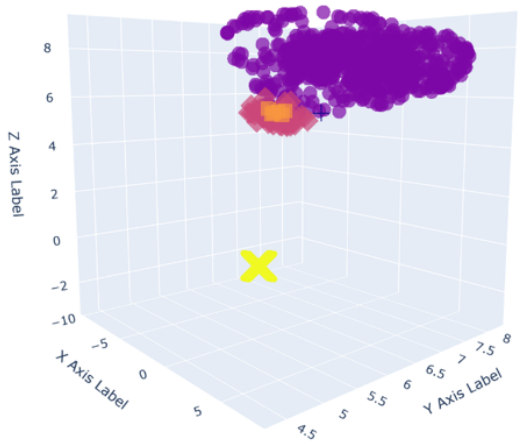


Figura 7.4: Función DBSCAN

Esta función se aplicó en cada uno de los modelos y se utilizó para simplificar el proceso de obtención de métricas de validación interna de clústeres una vez al mismo tiempo que se aplicaban las técnicas de reducción de dimensionalidad.

Para el primer modelo se utilizó UMAP para reducir la dimensionalidad del conjunto de datos y visualizar los grupos de datos en tres dimensiones utilizando los hiperparámetros por defecto de la librería UMAP y luego se aplicó la función de clus-

terización de DBSCAN mencionada anteriormente. Ese modelo está representado en la Figura 7.5 con su respectivo índice de Calinski-Harabasz, coeficiente de silueta e índice Davies-Bouldin.



Características	
Número de clústeres	4
Índice de Calinski-Harabasz	334.7598
Coeficiente de silueta	0.6589
Índice Davies–Bouldin	0.1450

Figura 7.5: Modelo 1: UMAP y DBSCAN sin hiperparámetros optimizados

Los resultados del Modelo 1 revelaron la formación de 4 clusteres dentro del conjunto de datos, proporcionando una estructura discernible. Con un índice Calinski de 334.7598, se indica una buena separación entre los clústeres, lo que sugiere distinción entre los grupos identificados. El índice de silueta, con un valor de 0.6589, refleja la calidad de la asignación de puntos a sus respectivos clústeres. Además, el índice de Davies-Bouldin, con un valor de 0.1450, sugiere una baja dispersión dentro de los clústeres.

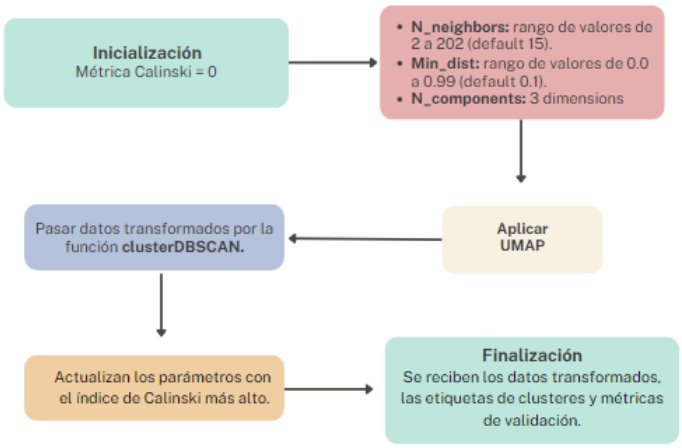
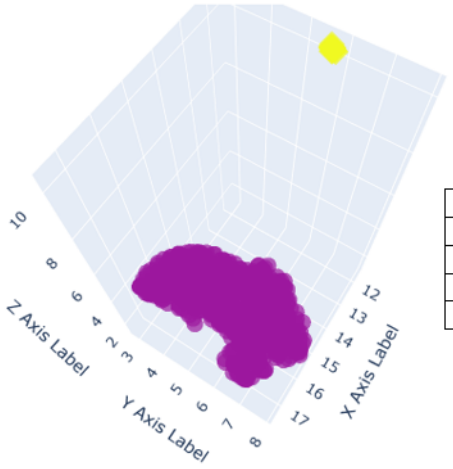


Figura 7.6: Optimización de hiperparámetros UMAP

Posteriormente se realizó una función que hace una búsqueda de los mejores

hiperparámetros para UMAP (n_neighbors y min_dist) mediante un proceso iterativo. Se prueba una variedad de combinaciones y se selecciona aquella que maximiza el índice Calinski-Harabasz entre todas las iteraciones. Luego, la función aplica DBSCAN sobre los datos transformados por UMAP con los parámetros óptimos encontrados. Este proceso está representado gráficamente en la Figura 7.6.



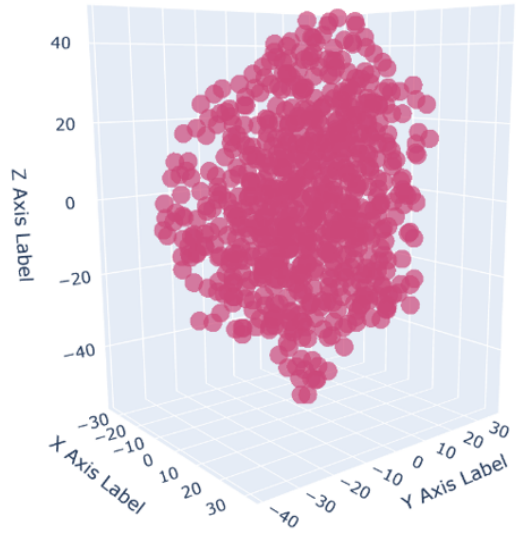
Características	
Número de clústeres	2
Índice de Calinski-Harabasz	281.8079
Coeficiente de silueta	0.2113
Índice Davies–Bouldin	0.7292

Figura 7.7: Modelo 2: UMAP y DBSCAN con hiperparámetros optimizados

Los resultados obtenidos del Modelo 2, representados en la Figura 7.7, revelan la conformación de dos clusters dentro del conjunto de datos, evidenciando una clara pero más simple separación. Con un índice Calinski-Harabasz de 281.8079, se sugiere una distinción notable entre los dos clusters identificados. No obstante, el índice de silueta, con un valor de 0.2113, indica una asignación menos homogénea de puntos a sus respectivos clusters en comparación con agrupamientos más cohesivos. Además, el índice de Davies-Bouldin de 0.7292 sugiere cierta dispersión y una menor compacidad en la configuración de los clusters.

Posteriomente, se aplicó la técnica de t-SNE para reducir la dimensionalidad del conjunto de datos y visualizar los grupos de datos en tres dimensiones utilizando los hiperparámetros por defecto de la librería t-SNE, a los datos transformados se les aplicó el algortimo DBSCAN y representado en la Figura 7.8. Este fue el Modelo 3 con sus respectivas métricas de validación interna.

Los resultados del Modelo 3 revelaron la formación de un solo cluster en el conjunto de datos, según la configuración establecida. Sin embargo, las métricas de validación



Características	
Número de clústeres	1
Índice de Calinski-Harabasz	0
Coeficiente de silueta	0
Índice Davies–Bouldin	0

Figura 7.8: Modelo 3: T-SNE y DBSCAN sin hiperparámetros optimizados

interna, como el índice Calinski, el índice de silueta y el índice de Davies-Bouldin, se registraron como 0. Esto se debe a una condición incorporada en la función DBSCAN creada, que especificaba que solo se calcularan y mostraran estas métricas si se identifican al menos 2 clusters al transformar los datos mediante la técnica de t-SNE. Dado que en esta instancia solo se formó un cluster, las métricas de validación interna no se generaron.

Dados estos resultados se crea una función que realizaba una búsqueda iterativa de los mejores hiperparámetros para t-SNE, la perplejidad, mediante un proceso de evaluación exhaustiva. Se probó una variedad de valores y se seleccionó aquel que maximizara el índice Calinski-Harabasz entre todas las iteraciones. Luego, la función aplica DBSCAN sobre los datos transformados por t-SNE con los parámetros óptimos encontrados y devuelve los datos transformados por t-SNE, las etiquetas de los clústeres, el índice Calinski-Harabasz, el índice Davies-Bouldin y el índice de silueta.

La aplicación de esta función arroja los el Modelo 4 representado en la Figura 7.10.

El Modelo 4 resultó en dos clusters con un índice Calinski de 61.7761, un índice de silueta de 0.4919 y un índice de Davies-Bouldin de 0.3781. En comparación con los modelos previos, este modelo exhibe una menor cantidad de clusters, lo que sugiere una partición más simplificada de los datos. Sin embargo, las métricas de evaluación

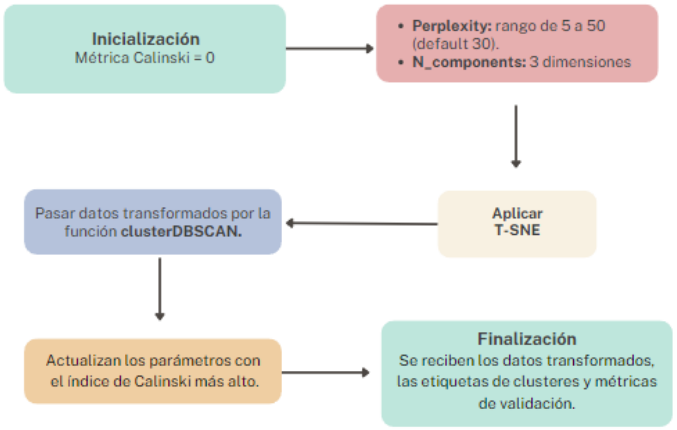


Figura 7.9: Optimización de hiperparámetros T-SNE

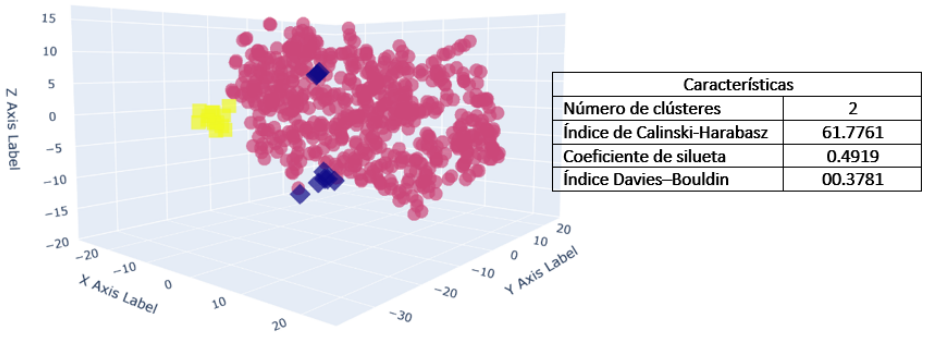
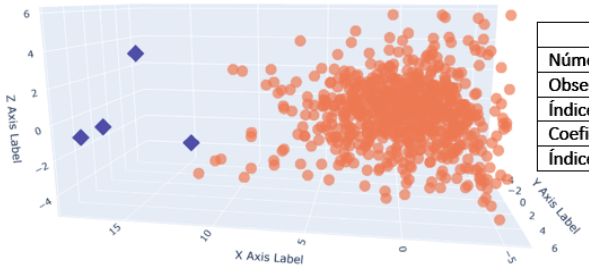


Figura 7.10: Modelo 4: T-SNE y DBSCAN con hiperparámetros optimizados

interna indican que el Modelo 4 logra una buena separación y coherencia dentro de los clusters identificados.

Por último, se utilizó un ACP para reducir la dimensionalidad del conjunto de datos y visualizar los grupos de datos en tres dimensiones utilizando los hiperparámetros por defecto de la librería PCA y luego se aplicó la clusterización DBSCAN, resultando en el Modelo 5 representado en la Figura 7.8 con su respectivo índice de Calinski-Harabasz, coeficiente de silueta e índice Davies-Bouldin.

Los resultados del Modelo 5 revelaron la formación de un único cluster en el conjunto de datos, con un índice Calinski de 61.5740, un índice de silueta de 0.4315 y un índice de Davies-Bouldin de 0.6645. Sin embargo, se identificaron cuatro observaciones clasificadas como ruido, indicando puntos que no se asignaron a ningún cluster específico según la configuración del algoritmo.



Características	
Número de clústeres	1
Observaciones ruido	4
Índice de Calinski-Harabasz	61.5740
Coeficiente de silueta	0.4315
Índice Davies–Bouldin	0.6645

Figura 7.11: Modelo 5: PCA y DBSCAN

Los cinco modelos evaluados muestran variaciones significativas en la estructura y calidad de los agrupamientos. El Modelo 1, con múltiples clusters, exhibió un índice Calinski-Harabasz más alto, indicando una mayor distinción entre clusters. El Modelo 2, también con dos clusters, presentó una estructura clara y bien definida, aunque con métricas ligeramente inferiores en comparación con el Modelo 4. Este último, resaltó por su buena coherencia interna, como evidenciaron el índice de silueta y el índice de Davies-Bouldin. En contraste, el Modelo 3, que resultó en un solo cluster, no generó métricas de validación interna debido a su configuración específica. El Modelo 5, con un cluster y observaciones clasificadas como ruido, muestra una estructura homogénea en la mayoría de los datos pero con algunas observaciones atípicas.

Es importante recalcar que la elección entre estos modelos dependerá de los objetivos del análisis, donde se quiere dar prioridad a la simplicidad y la distinción entre clusters, por esto también se tendrán en cuenta los gráficos de cada uno. Asimismo, se evaluarán dependiendo según las necesidades específicas del proyecto y los objetivos del mismo.

8. Discusión

Se implementaron cinco modelos de clusterización para analizar la segmentación de clientes de una bebida alcohólica a partir de los resultados obtenidos de un cuestionario. Las características detalladas y comparativas de cada uno de estos modelos se encuentran representadas en la Tabla 8.1.

El Modelo 1 muestra el índice Calinski-Harabasz más alto con una puntuación de 334.76, lo que sugiere la presencia de grupos bien separados y cohesivos. El Modelo 2, con una puntuación de 281.81, también presenta características de agrupamiento sólidas. Sin embargo, el Modelo 3, con una puntuación de 0, indica un agrupamiento deficiente, implicando la existencia de un único grupo no diferenciado. Los Modelos 4 y 5, con puntuaciones de 61.78 y 61.57, respectivamente, tienen valores de índice intermedios, lo que sugiere una calidad de agrupamiento moderada. Por ende, los Modelos 1 y 2 muestran el agrupamiento más robusto.

Con respecto a las otras métricas de validación, el Modelo 2 presenta un índice de Davies-Bouldin significativamente más bajo (0.211) en comparación con el Modelo 1 (0.659). Esto sugiere que el Modelo 1 tiene agrupamientos más cohesivos y mejor definidos en comparación con el Modelo 1. Asimismo, el Modelo 2 tiene un puntaje de silueta más alto (0.729) en comparación con el Modelo 1 (0.145), lo que respalda la observación anterior de que el Modelo 1 presenta agrupamientos más sólidos y bien definidos en comparación con el Modelo 2.

Aunque el Modelo 1 exhibe métricas de evaluación superiores en comparación con el Modelo 2, se opta por seleccionar este último debido a los objetivos específicos del análisis. La elección se basa en priorizar la simplicidad y la capacidad de obtener una mejor distinción entre clusters a nivel gráfico. A pesar de que el Modelo 2 pueda presentar métricas ligeramente inferiores, su estructura clara y bien definida permite visualizar de manera más nítida las características distintivas de los grupos identificados. Esta decisión se alinea con los objetivos del trabajo, donde la claridad en la

distinción entre clusters aporta una comprensión más accesible y visualmente evidente de las características particulares de cada grupo.

Entre las características más evidentes del cluster más pequeño identificado en el Modelo 2, seleccionado como el óptimo para el análisis, revelan un perfil homogéneo entre sus 13 observaciones. Todas las integrantes de este grupo son mujeres y se desempeñan laboralmente en empresas bancarias o crediticias. Además, es notable que la gran mayoría de estas mujeres tienen su residencia en Bogotá. Esto respalda la elección del Modelo 2 pues sus características poseían capacidad para ofrecer insights precisos y fácilmente interpretables.

Tabla 8.1

Tabla comparativa de resultados.

Característica	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Número de clústers	4	2	1	2	1
Indice Calinski-Harabazs	334.7598	281.8079	0	61.7761	61.5740
Coeficiente de silueta	0.6589	0.2113	0	0.4919	0.4315
Indice Davies-Bouldin	0.1450	0.7292	0	0.3781	0.6645

9. Conclusiones

En el análisis emprendido para entender la segmentación de clientes de una bebida alcohólica, se llevó a cabo una evaluación minuciosa de cinco modelos distintos. El objetivo principal era identificar el modelo más adecuado para capturar las particularidades de los distintos grupos de clientes. Se logró la implementación y evaluación de cinco modelos de clustering. Estos modelos incluyeron combinaciones de técnicas de reducción de dimensionalidad y algoritmos de clustering. Los modelos implementados fueron: UMAP combinado con DBSCAN, UMAP con DBSCAN aplicando hiperparámetros optimizados, T-SNE con DBSCAN, T-SNE y DBSCAN con hiperparámetros optimizados, y finalmente, ACP combinado con DBSCAN.

La evaluación de estos modelos se basó en índices de calidad como Calinski-Harabasz, coeficiente de silueta y Davies-Bouldin, logrando así el objetivo inicial de implementar y comparar diferentes modelos. A partir de un análisis detallado, se identificó un modelo específico como el más idóneo, distinguiéndose por su eficacia en proporcionar una segmentación clara y útil.

A pesar de que el Modelo 1 presentó métricas superiores, se optó por el Modelo 2 debido a su simplicidad y claridad en la diferenciación entre clusters. Esta decisión estuvo en consonancia con uno de los objetivos específicos del estudio: seleccionar un modelo que, además de ser eficaz en términos de métricas, facilitara una interpretación visual y clara de los datos. Este modelo ofreció percepciones particularmente reveladoras sobre un grupo específico de clientes: mujeres en el sector bancario o crediticio, mayoritariamente localizadas en Bogotá. La elección de este modelo se basó no solo en su precisión analítica, sino también en su capacidad para simplificar y clarificar la comprensión de los datos.

Finalmente, la capacidad del modelo para ofrecer una interpretación visualmente intuitiva y accesible de los datos fue un factor determinante en su selección. Esta metodología de análisis no solo facilitó una segmentación efectiva de los clientes, sino que

también propició una comprensión más profunda de sus características y preferencias.

Bibliografía

[1] A. Gunawan and M. de Berg, "A faster algorithm for dbscan," *Master's thesis*, 2013.

[2] P. E. Green, R. E. Frank, and P. J. Robinson, "Cluster analysis in test market selection," *Management science*, vol. 13, no. 8, pp. B–387, 1967.

[3] T. Reutterer and D. Dan, "Cluster analysis in marketing research," in *Handbook of Market Research*. Cham: Springer International Publishing, 2020, pp. 1–29.

[4] A. S. M. S. Hossain, "Customer segmentation using centroid based and density based clustering algorithms," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, Dec. 2017.

[5] S. Vijendra, "Efficient clustering for high dimensional data: Subspace based clustering and density based clustering," *Information Technology Journal*, vol. 10, pp. 1092–1105, 2011.

[6] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," Feb. 2018.

[7] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[8] G. de Colombia, "Encuesta nacional de consumo de sustancias psicoactivas (encspa). resultados 2019 [national survey on psychoactive substances consumption. results 2019]," 2020.

[9] World Health Organization, "Global status report on alcohol and health 2018," Available from: https://www.who.int/publications/i/item/9789241565639., 2018, accessed: 2023-04-24.

[10] M. Thibodeau and G. J. Pickering, "The role of taste in alcohol preference, consumption and risk behavior," *Critical Reviews in Food Science and Nutrition*, vol. 59, pp. 676–692, 2 2019.

[11] P. Kotler, G. Armstrong, J. Saunders, V. Wong, S. Miquel, E. Bigné, and D. Cámara, *Introducción al marketing*. Pearson Prentice Hall, 2000.

[12] J. H. Myers and E. Tauber, *Market structure analysis*. Marketing Classics Press, 2011.

[13] M. Wedel and W. A. Kamakura, *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media, 2000.

[14] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. IEEE, 2005, pp. 197–202.

[15] S. Koul and T. M. Philip, "Customer segmentation techniques on e-commerce," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2021, pp. 135–138.

[16] A. C. Ospina Pérez, "Análisis de segmentación del cliente empresa de eps sura," 2021.

[17] W. R. Smith, "Product differentiation and market segmentation as alternative marketing strategies," *Journal of marketing*, vol. 21, no. 1, pp. 3–8, 1956.

[18] E. Ernawati, S. Baharin, and F. Kasmin, "A review of data mining methods in rfm-based customer segmentation," in *Journal of Physics: Conference Series*, vol. 1869, no. 1. IOP Publishing, 2021, p. 012085.

[19] J. Hartigan, "Clustering algorithms wiley," *New York*, 1975.

[20] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[21] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[22] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE, 2014, pp. 232–238.

[23] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, pp. 107–145, 2001.

[24] M. I. Oliveira and A. R. Marcal, "Clustering lidar data with k-means and dbscan," 2023.

[25] F. Trozzi, X. Wang, and P. Tao, "Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study," *The Journal of Physical Chemistry B*, vol. 125, no. 19, pp. 5022–5034, 2021.

[26] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002.

[27] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data," *Nature methods*, vol. 16, no. 3, pp. 243–245, 2019.

Anexos

A. Anexo I: Código

Código en Google Colab