Logotipo, nombre de la empresa

Descripción generada automáticamente

|  |
| --- |
| **Actividad 3** |
| APRENDIZAJE NO SUPERVISADO |
| **Caso de estudio:**  Informe Técnico del Proyecto “Customer Personality Analysis”. |

|  |  |
| --- | --- |
| **Profesora** | Gladys María Villegas Rugel |
| **Materia** | APRENDIZAJE AUTOMATICO |
| **Alumnos** | * Bolaños Escandón María Fernanda * Estupiñán Andrade Francisco Javier * Montaño Cárdenas Fernando Xavier |
| **Fecha** | 16/05/2025 |

Contenido

[**1.** **OBJETIVOS** 3](#_Toc198232326)

[**2.** **INTRODUCCIÓN** 3](#_Toc198232327)

[**3.** **DESCRIPCIÓN DEL CONJUNTO DE DATOS** 3](#_Toc198232328)

[3.1 Información Demográfica 3](#_Toc198232329)

[3.2 Comportamiento de Compra 4](#_Toc198232330)

[3.3 Participación en Campañas de Marketing 4](#_Toc198232331)

[3.4 Comportamiento de Compra en Línea y en Tienda 4](#_Toc198232332)

[**4.** **ANÁLISIS TÉCNICO** 4](#_Toc198232333)

[4.1 Preprocesamiento de Datos 4](#_Toc198232334)

[4.2 Aplicación del Algoritmo KMeans 5](#_Toc198232335)

[4.3 Aplicación del Algoritmo DBSCAN 5](#_Toc198232336)

[4.4 Selección del Número Óptimo de Clusters 5](#_Toc198232337)

[4.5 Visualización de los Métodos de Selección 5](#_Toc198232338)

[4.6 Aplicación de Modelos de Clustering y Reducción de Dimensionalidad 6](#_Toc198232339)

[4.7 Visualización Comparativa de Modelos 8](#_Toc198232340)

[**5.** **ANÁLISIS ESTADÍSTICO** 10](#_Toc198232341)

[5.1 Resumen estadístico de las variables numéricas 10](#_Toc198232342)

[**6. RESUMEN DE CARACTERÍSTICAS** 13](#_Toc198232343)

[6.1 Resumen de Características Numéricas por Clúster (K-Means) 13](#_Toc198232344)

[6.2 Resumen de Características Numéricas por Clúster (DBSCAN) 13](#_Toc198232345)

[**7. RESULTADOS Y COMPARACIÓN** 13](#_Toc198232346)

[**8.** **CONCLUSIONES** 14](#_Toc198232347)

[**9.** **RECOMENDACIONES** 15](#_Toc198232348)

[**10.** **ANEXOS** 15](#_Toc198232349)

# **OBJETIVOS**

* Implementar y analizar modelos de aprendizaje no supervisado (K-means, DBSCAN, PCA y t-SNE) para segmentar perfiles del comportamiento de usuario.
* Visualizar resultados, comparar métodos y comunicar conclusiones de forma técnica y visual.

# **INTRODUCCIÓN**

En este informe se presenta el análisis de un conjunto de datos utilizando dos algoritmos de clustering: KMeans y DBSCAN. El objetivo principal es segmentar los datos en diferentes grupos de acuerdo con sus características similares, para poder extraer conclusiones y patrones relevantes. El análisis se realizó en un conjunto de datos con variables numéricas, y se utilizaron técnicas de preprocesamiento para preparar los datos antes de aplicar los algoritmos de clustering.

# **DESCRIPCIÓN DEL CONJUNTO DE DATOS**

El conjunto de datos se obtuvo de la plataforma kaggle, mismo que se describe a continuación.

**Análisis de la Personalidad del Cliente:**

Es un análisis detallado de los clientes ideales de una empresa. Ayuda a la empresa a comprender mejor a sus clientes y facilita la modificación de productos según las necesidades, comportamientos y preocupaciones específicas de los diferentes tipos de clientes.

El análisis de la personalidad del cliente ayuda a una empresa a modificar su producto en función de sus clientes objetivo de diferentes segmentos. Por ejemplo, en lugar de gastar dinero para comercializar un nuevo producto a todos los clientes de la base de datos de la empresa, esta puede analizar qué segmento de clientes es más probable que compre el producto y luego dirigir la estrategia de marketing solo a ese segmento en particular.

El dataset consta de **2,240 registros** y **29 columnas**, organizadas en cuatro categorías principales, los campos relevantes de cada sección se detallan a continuación.

## 3.1 Información Demográfica

* **ID**: Identificador único del cliente.
* **Year\_Birth**: Año de nacimiento del cliente.
* **Education**: Nivel educativo del cliente.
* **Marital\_Status**: Estado civil del cliente.
* **Income**: Ingreso anual del hogar del cliente.
* **Kidhome**: Número de niños en el hogar del cliente.
* **Teenhome**: Número de adolescentes en el hogar del cliente.
* **Dt\_Customer**: Fecha de inscripción del cliente en la empresa.
* **Recency**: Número de días desde la última compra del cliente.
* **Complain**: Indicador de si el cliente presentó una queja en los últimos 2 años (1 = sí, 0 = no).

## 3.2 Comportamiento de Compra

* **MntWines**: Monto gastado en vino en los últimos 2 años.
* **MntFruits**: Monto gastado en frutas en los últimos 2 años.
* **MntMeatProducts**: Monto gastado en productos cárnicos en los últimos 2 años.
* **MntFishProducts**: Monto gastado en productos pesqueros en los últimos 2 años.
* **MntSweetProducts**: Monto gastado en productos dulces en los últimos 2 años.
* **MntGoldProds**: Monto gastado en productos de oro en los últimos 2 años.

## 3.3 Participación en Campañas de Marketing

* **NumDealsPurchases**: Número de compras realizadas con descuento.
* **AcceptedCmp1** a **AcceptedCmp5**: Indicadores de aceptación de las campañas de marketing 1 a 5 (1 = aceptó, 0 = no aceptó).
* **Response**: Indicador de aceptación de la última campaña de marketing (1 = aceptó, 0 = no aceptó).

## 3.4 Comportamiento de Compra en Línea y en Tienda

* **NumWebPurchases**: Número de compras realizadas a través del sitio web de la empresa.
* **NumCatalogPurchases**: Número de compras realizadas utilizando un catálogo.
* **NumStorePurchases**: Número de compras realizadas directamente en tiendas.
* **NumWebVisitsMonth**: Número de visitas al sitio web de la empresa en el último mes.[RStudio Pubs](https://rstudio-pubs-static.s3.amazonaws.com/832210_c452d6fb6d924bf5aedbed80ba5a91fe.html?utm_source=chatgpt.com)

# **ANÁLISIS TÉCNICO**

## 4.1 Preprocesamiento de Datos

Se cargaron las librerías necesarias (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn) y se leyó el archivo marketing\_campaign.csv. En la etapa de preprocesamiento se eliminaron columnas irrelevantes como ID, Dt\_Customer, Z\_CostContact y Z\_Revenue, además de eliminar filas con valores nulos. Las variables categóricas fueron transformadas en variables dummies y los datos fueron escalados con StandardScaler para normalizar las variables numéricas.

Antes de realizar cualquier análisis de clustering, fue necesario aplicar un preprocesamiento adecuado a los datos. Los pasos realizados incluyen:

* **Escalado de los Datos**: Utilizando el StandardScaler, los datos fueron escalados para asegurar que las diferentes magnitudes de las variables no afectaran el rendimiento de los algoritmos de clustering.
* **Análisis de Distribución de Datos**: Se inspeccionaron visualmente las distribuciones de las características para verificar la normalidad y la presencia de posibles outliers.

## 4.2 Aplicación del Algoritmo KMeans

**KMeans**, que segmenta los datos en un número predefinido de clusters. La elección del número de clusters se realizó utilizando el **método del codo** (elbow method), que analiza la variabilidad explicada por diferentes valores de k.

* **Número de Clusters Determinado**: 5 clusters
* **Resumen de Clústeres KMeans**: Después de aplicar el algoritmo, se presentó un resumen de las características de cada clúster calculando la media para cada característica numérica dentro de cada grupo.

## 4.3 Aplicación del Algoritmo DBSCAN

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**. A diferencia de KMeans, DBSCAN no requiere especificar el número de clústeres de antemano, sino que agrupa los puntos de acuerdo con la densidad de puntos cercanos.

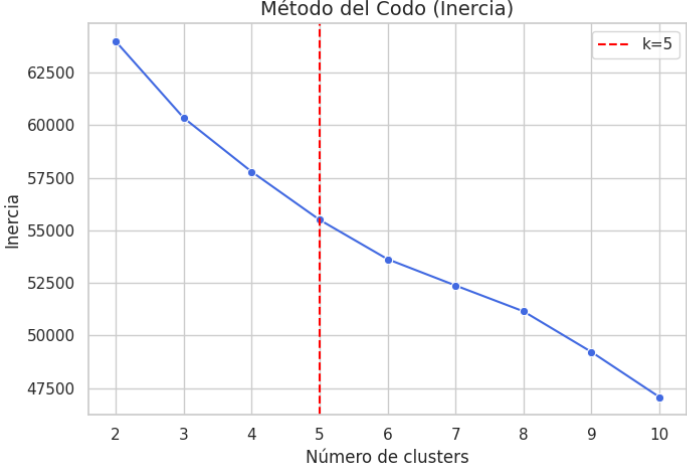
* **Parámetros del Algoritmo**: El valor de eps (distancia máxima entre dos puntos para que sean considerados vecinos) y min\_samples (número mínimo de puntos necesarios para formar un clúster) fueron ajustados utilizando el análisis de la distancia de los vecinos más cercanos.
* **Resumen de Clústeres DBSCAN**: Similar al caso de KMeans, se presentó un resumen con las características promedias por clúster generado por DBSCAN.

## 4.4 Selección del Número Óptimo de Clusters

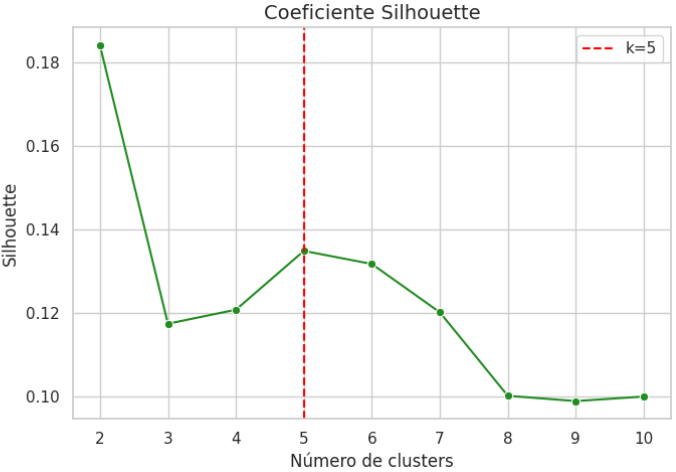
Para determinar el número óptimo de clusters se utilizaron dos métodos: el método del codo y el coeficiente de Silhouette. El método del codo evalúa la inercia (la suma de los errores cuadráticos dentro de los clusters) y mostró un punto de inflexión en K=5, lo que sugiere que añadir más clusters no mejora significativamente la segmentación. Por otro lado, el coeficiente de Silhouette mide qué tan bien está un punto dentro de su cluster en comparación con otros clusters. Aunque su valor máximo se presenta en K=2, el valor en K=5 también fue alto, lo cual valida esta elección desde otro enfoque.

### 4.5 Visualización de los Métodos de Selección

**Figura 1: Método del Codo**



**Figura 2: Coeficiente de Silhouette**

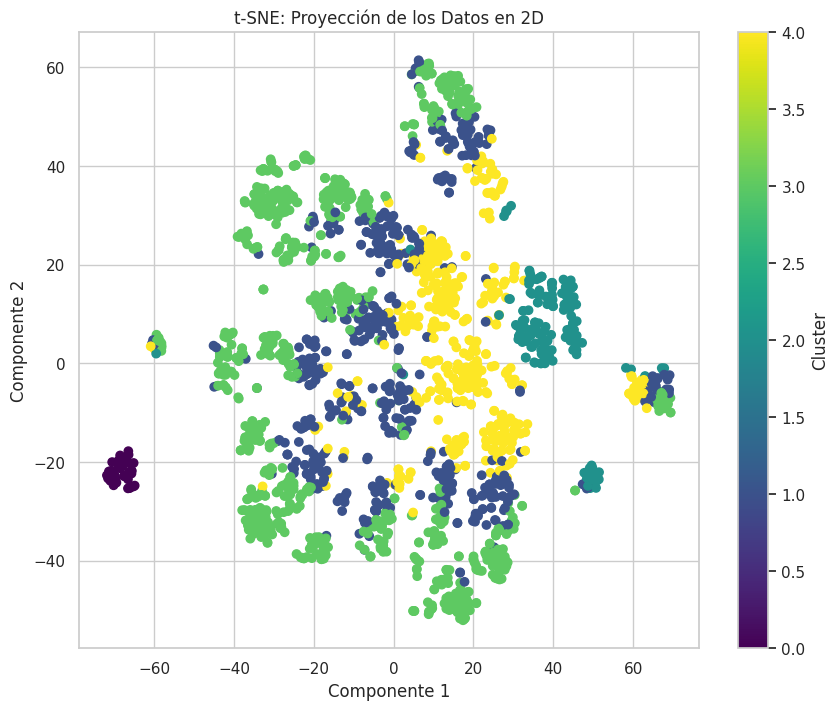


Estas gráficas evidencian que K=5 representa un buen equilibrio entre simplicidad del modelo y segmentación detallada.

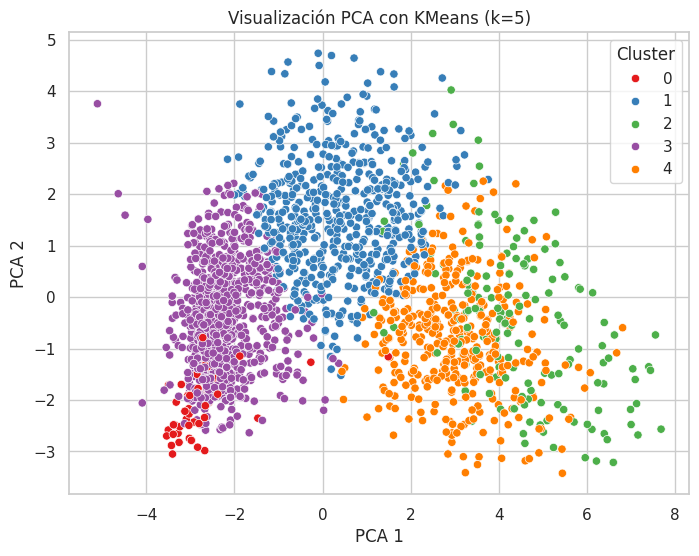
## 4.6 Aplicación de Modelos de Clustering y Reducción de Dimensionalidad

Posteriormente, se procedió a la aplicación de los algoritmos de clustering K-Means y DBSCAN, y se utilizaron técnicas de reducción de dimensionalidad como PCA y t-SNE para visualizar los resultados. Se generaron proyecciones en dos dimensiones para ambos algoritmos, permitiendo evaluar la coherencia visual de los clusters.

**Figura 3: t-SNE: Proyección de los Datos en 2D**



**Figura 4: PCA - (Principal Component Analysis)**

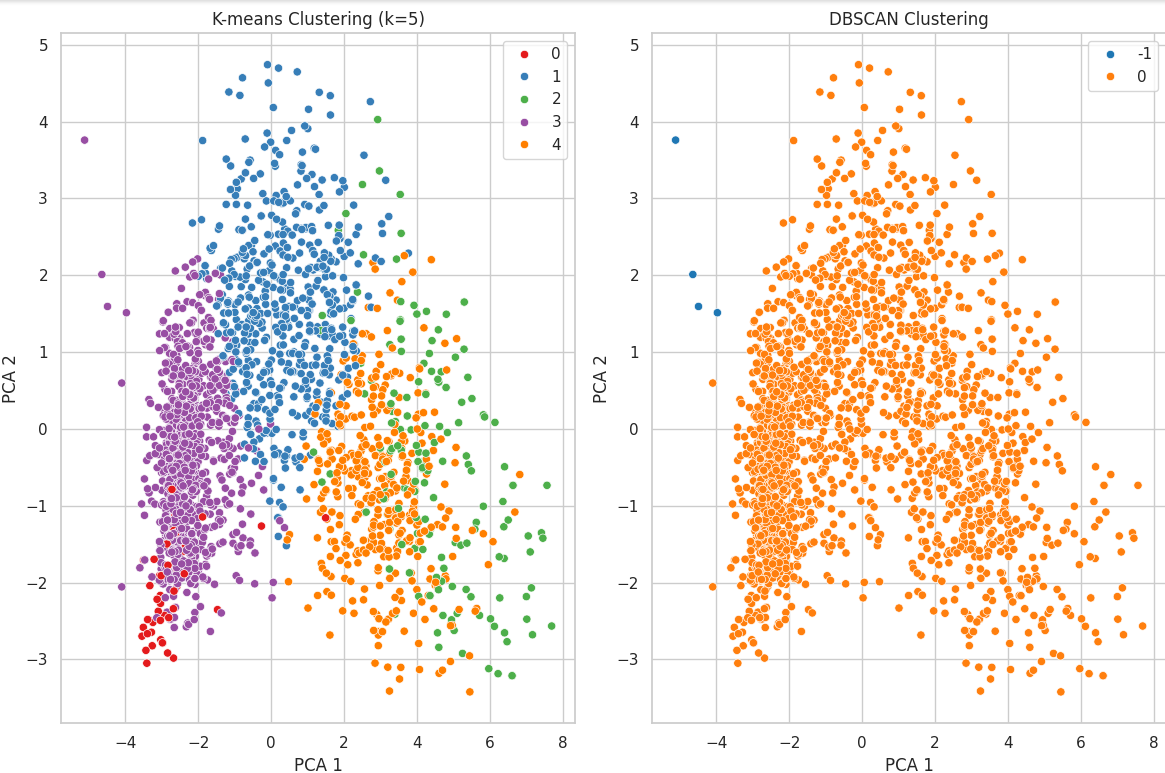


Facilita la reducción de dimensiones a 2 usando PCA, esto para la visualización

## 4.7 Visualización Comparativa de Modelos

A continuación, se presentan las gráficas comparativas entre los algoritmos de clustering K-Means y DBSCAN, empleando tanto PCA como t-SNE como técnicas de reducción de dimensionalidad para visualizar la distribución de los datos.

**Figura 5: Comparativa entre K-Means y DBSCAN con PCA**

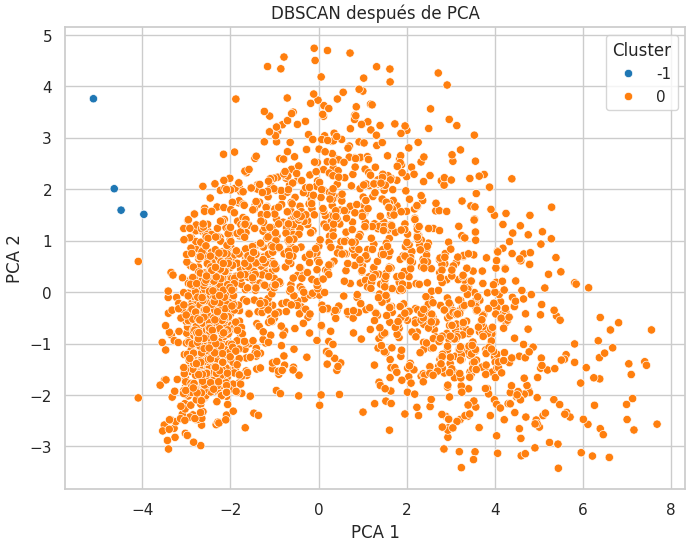


En la parte izquierda de la imagen, K-Means logra separar los datos en cinco agrupaciones bastante diferenciadas.

Se observa una distribución más compacta y organizada, con límites claros entre los grupos. Por otro lado, DBSCAN (parte derecha) identifica un único cluster principal y solo unos pocos puntos como ruido (etiquetados como -1).

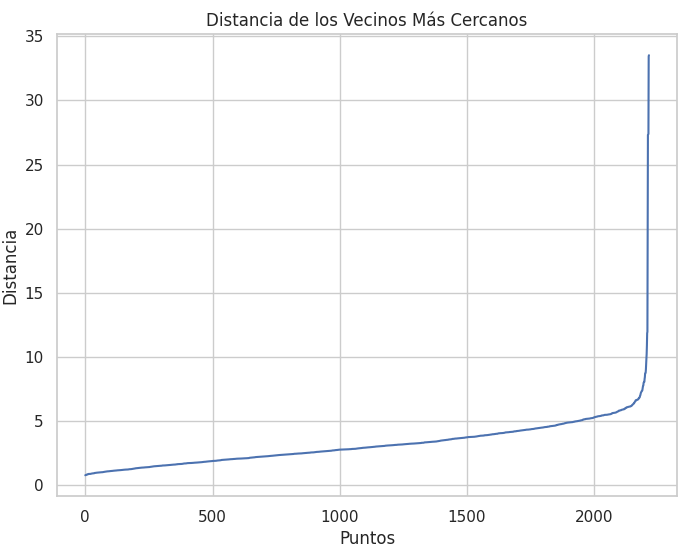
Esto indica que, con la configuración actual de parámetros (*eps* y *min\_samples*), DBSCAN no encuentra estructura suficiente para generar múltiples grupos significativos en este conjunto de datos escalados.

**Figura 6: DBSCAN con PCA (zoom para ver ruido)**



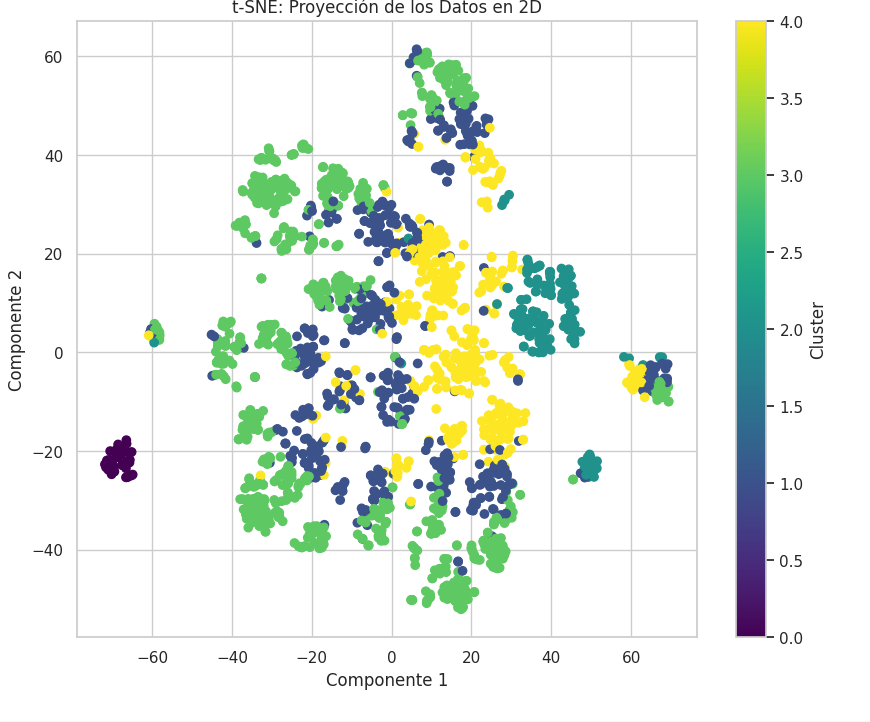
Aquí se confirma que DBSCAN clasificó a casi todos los puntos dentro de un único grupo, con solo algunos considerados como ruido. La falta de subdivisión en varios grupos podría deberse a una configuración de eps demasiado grande o a una baja variabilidad entre los datos tras PCA.

**Figura 7: Curva de Distancia de Vecinos (K-distance plot)**



Este gráfico permite identificar el valor óptimo del parámetro eps para DBSCAN. El punto de mayor curvatura ("rodilla") se encuentra aproximadamente entre 3 y 5 unidades de distancia. Un ajuste fino en este parámetro podría permitir a DBSCAN generar agrupaciones más detalladas.

**Figura 8: K-Means con t-SNE (Visualización No Lineal)**



Las visualizaciones obtenidas muestran que el algoritmo K-Means produce agrupaciones simétricas y bien definidas tanto en PCA como en t-SNE. Por su parte, DBSCAN permite identificar agrupaciones más libres y detecta puntos de ruido (outliers), los cuales aparecen dispersos. Con t-SNE se aprecia una mejor separación entre clusters, lo que permite visualizar estructuras complejas que PCA no logra capturar completamente.

# **ANÁLISIS ESTADÍSTICO**

## 5.1 Resumen estadístico de las variables numéricas

La tabla proporciona estadísticas como la media, desviación estándar, mínimo, máximo, cuartiles (25%, 50%, 75%) para las columnas numéricas.

A continuación, se presenta el análisis estadístico por clúster. En lugar de usar gráficos como mapas de calor o histogramas, se utilizaron tablas que muestran los valores promedio por grupo.

**Tabla 1. Resumen estadístico de las variables numéricas**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | | **min** |  | | --- | --- | | **25%** | **50%** | **75%** | **max** |
| **Year\_Birth** | 2216.0 | 1968.820397 | 11.985554 | 1893.0 | 1959.0 | 1970.0 | 1977.00 | 1996.0 |
| **Income** | 2216.0 | 52247.251354 | 25173.076661 | 1730.0 | 35303.0 | 51381.5 | 68522.00 | 666666.0 |
| **Kidhome** | 2216.0 | 0.441787 | 0.536896 | 0.0 | 0.0 | 0.0 | 1.00 | 2.0 |
| **Teenhome** | 2216.0 | 0.505415 | 0.544181 | 0.0 | 0.0 | 0.0 | 1.00 | 2.0 |
| **Recency** | 2216.0 | 49.012635 | 28.948352 | 0.0 | 24.0 | 49.0 | 74.00 | 99.0 |
| **MntWines** | 2216.0 | 305.091606 | 337.327920 | 0.0 | 24.0 | 174.5 | 505.00 | 1493.0 |
| **MntFruits** | 2216.0 | 26.356047 | 39.793917 | 0.0 | 2.0 | 8.0 | 33.00 | 199.0 |
| **MntMeatProducts** | 2216.0 | 166.995939 | 224.283273 | 0.0 | 16.0 | 68.0 | 232.25 | 1725.0 |
| **MntFishProducts** | 2216.0 | 37.637635 | 54.752082 | 0.0 | 3.0 | 12.0 | 50.00 | 259.0 |
| **MntSweetProducts** | 2216.0 | 27.028881 | 41.072046 | 0.0 | 1.0 | 8.0 | 33.00 | 262.0 |
| **MntGoldProds** | 2216.0 | 43.965253 | 51.815414 | 0.0 | 9.0 | 24.5 | 56.00 | 321.0 |
| **NumDealsPurchases** | 2216.0 | 2.323556 | 1.923716 | 0.0 | 1.0 | 2.0 | 3.00 | 15.0 |
| **NumWebPurchases** | 2216.0 | 4.085289 | 2.740951 | 0.0 | 2.0 | 4.0 | 6.00 | 27.0 |
| **NumCatalogPurchases** | 2216.0 | 2.671029 | 2.926734 | 0.0 | 0.0 | 2.0 | 4.00 | 28.0 |
| **NumStorePurchases** | 2216.0 | 5.800993 | 3.250785 | 0.0 | 3.0 | 5.0 | 8.00 | 13.0 |
| **NumWebVisitsMonth** | 2216.0 | 5.319043 | 2.425359 | 0.0 | 3.0 | 6.0 | 7.00 | 20.0 |
| **AcceptedCmp3** | 2216.0 | 0.073556 | 0.261106 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **AcceptedCmp4** | 2216.0 | 0.074007 | 0.261842 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **AcceptedCmp5** | 2216.0 | 0.073105 | 0.260367 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **AcceptedCmp1** | 2216.0 | 0.064079 | 0.244950 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **AcceptedCmp2** | 2216.0 | 0.013538 | 0.115588 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **Complain** | 2216.0 | 0.009477 | 0.096907 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |
| **Response** | 2216.0 | 0.150271 | 0.357417 | 0.0 | 0.0 | 0.0 | 0.00 | 1.0 |

# **6. RESUMEN DE CARACTERÍSTICAS**

## 6.1 Resumen de Características Numéricas por Clúster (K-Means)

Se presenta una tabla que resume las características numéricas medias para cada clúster generado por K-Means. Incluye variables como ingresos, gasto por tipo de producto y frecuencia de compra en diferentes canales.

## 6.2 Resumen de Características Numéricas por Clúster (DBSCAN)

Del mismo modo, se muestra un resumen de los datos agrupados por DBSCAN, aunque en este caso la mayoría de puntos pertenecen a un único grupo (etiqueta 0), con algunos clasificados como ruido (etiqueta -1).

Estas tablas permiten comparar fácilmente las características dominantes en cada grupo, y son clave para la creación de perfiles de usuario.

# **7. RESULTADOS Y COMPARACIÓN**

* Los clústeres generados por KMeans son generalmente más esféricos y de tamaño similar. En los resúmenes de cada clúster se observó que algunas características numéricas presentan diferencias notables entre los clústeres.
* Los resultados de DBSCAN mostraron una distribución más irregular de los puntos, y también detectó algunos puntos como ruido (que no pertenecen a ningún clúster).
* Comparando los modelos utilizados, se destaca que, aunque K=2 presentó el mayor Silhouette Score, resultó ser una agrupación demasiado general. En cambio, K=5 ofreció un mejor balance entre segmentación detallada y coherencia estructural. DBSCAN mostró un desempeño aceptable, pero fue sensible a la parametrización y generó una cantidad notable de ruido. Las proyecciones con PCA ayudaron a ver una representación lineal de los clusters, mientras que t-SNE permitió capturar relaciones más complejas y no lineales entre los datos.
* **K-Means con K=5** es la mejor alternativa para segmentar este conjunto de clientes por su equilibrio entre precisión, estabilidad y facilidad de interpretación.
* DBSCAN puede ser útil como método exploratorio para detectar ruido o anomalías, especialmente si se ajustan sus parámetros.
* Las visualizaciones con PCA mostraron que K-Means logra una segmentación simétrica y estructurada, mientras que DBSCAN, con los parámetros usados, generó un único grupo principal y pocos puntos de ruido. La curva de distancia de vecinos ayudó a identificar un valor de eps más adecuado para DBSCAN, aunque aún se requiere mayor ajuste para obtener una separación clara de grupos.
* t-SNE reveló con mayor claridad la distribución compleja de los clusters, especialmente para K-Means, donde los cinco grupos se separan con claridad en el espacio no lineal. Esta técnica confirmó la validez estructural de la segmentación más allá de lo que permite visualizar PCA.
* El análisis estadístico por clúster mostró diferencias significativas entre los grupos. Por ejemplo, ciertos clústeres mostraron mayor gasto en productos de lujo como vinos, mientras otros destacaron por una menor actividad en canales digitales. Estas diferencias ofrecen una base sólida para personalizar campañas, mejorar el enfoque de ventas y optimizar recursos.
* Entre los perfiles identificados se encuentran grupos con altos ingresos y consumo de lujo, así como clientes con baja participación digital y escasa respuesta a campañas, lo que permite estrategias de marketing diferenciadas.

# **CONCLUSIONES**

* El análisis integral realizado mediante los algoritmos K-Means y DBSCAN, junto con las técnicas de reducción de dimensionalidad PCA y t-SNE, permitió identificar patrones significativos en el comportamiento de los clientes. A partir del método del codo y el coeficiente de Silhouette, se determinó que **K=5** es el número óptimo de clusters, equilibrando simplicidad y precisión.
* Las visualizaciones con PCA mostraron que K-Means logra una segmentación simétrica y estructurada, mientras que DBSCAN, con los parámetros usados, generó un único grupo principal y pocos puntos de ruido.
* La curva de distancia de vecinos ayudó a identificar un valor de *eps* más adecuado para DBSCAN, aunque aún se requiere mayor ajuste para obtener una separación clara de grupos.
* t-SNE reveló con mayor claridad la distribución compleja de los clusters, especialmente para K-Means, donde los cinco grupos se separan con claridad en el espacio no lineal. Esta técnica confirmó la validez estructural de la segmentación más allá de lo que permite visualizar PCA.
* La variable **Income** (ingreso) fue clave para identificar clientes de alto poder adquisitivo, especialmente en los clusters 0 y 2.
* Variables como **MntWines**, **MntGoldProds** y **NumWebPurchases** revelaron hábitos de consumo de lujo o alto valor, más frecuentes en ciertos clusters. En contraste, clusters con valores bajos en estas variables reflejan perfiles de clientes con menor poder de compra o menor interacción con canales digitales.
* Las variables **Recency** y **AcceptedCmp1-5** también ayudaron a entender la fidelidad y respuesta a campañas, indicando qué grupos han respondido más o menos a las promociones previas.
* **K-Means con K=5** es la mejor alternativa para segmentar este conjunto de clientes por su equilibrio entre precisión, estabilidad y facilidad de interpretación. DBSCAN puede ser útil como método exploratorio para detectar ruido o anomalías, especialmente si se ajustan sus parámetros. Las técnicas de reducción de dimensionalidad fueron fundamentales para validar visualmente la estructura de los datos y facilitar la interpretación de los resultados.
* K-Means con K=5 es el modelo más robusto en este contexto, por su rendimiento computacional, claridad visual y segmentación efectiva. DBSCAN puede considerarse una opción complementaria, especialmente cuando se quiere identificar ruido o patrones menos estructurados. El análisis estadístico de las características de los clusters permitió construir perfiles detallados y útiles para la personalización de campañas de marketing. Las técnicas de reducción de dimensionalidad confirmaron la validez visual de las agrupaciones.
* Una limitación relevante fue la sensibilidad de DBSCAN a la parametrización, lo que dificultó una segmentación clara. Para mejorar su desempeño se recomienda ajustar automáticamente los valores de *eps* y *min\_samples*, o explorar alternativas como HDBSCAN y clustering jerárquico.”
* Otra limitación es la sensibilidad del algoritmo DBSCAN a la selección de parámetros. Un valor inapropiado de eps o min\_samples puede hacer que el algoritmo agrupe la mayoría de los datos en un solo clúster o detecte demasiado ruido, afectando la utilidad de la segmentación. Para abordarlo, se recomienda ajustar los parámetros usando validación visual (como el gráfico de distancia de vecinos).
* Otra limitación presente es la asignación uniforme de peso a todas las variables**.** El uso de escalado estándar asume que todas las variables son igual de importantes, lo que puede no ser cierto en un contexto comercial.

# **RECOMENDACIONES**

Como recomendaciones, se sugiere implementar campañas segmentadas basadas en los perfiles identificados, utilizar los clusters como variables predictoras en modelos supervisados, mantener actualizada la segmentación con nuevos datos y considerar ajustes en DBSCAN o el uso de modelos jerárquicos si se desea explorar una segmentación más fina. Este estudio proporciona una base sólida para decisiones estratégicas basadas en datos y una comprensión profunda del comportamiento de los clientes.

# **ANEXOS**

Data set:

* <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

Repositorio:

* <https://github.com/FranciscoEstAn/Analisis-de-la-personalidad-del-cliente/blob/main/README.md#an%C3%A1lisis-de-gr%C3%A1ficas-y-resultados>

Código:

* <https://colab.research.google.com/drive/1KXlM2nXO0XJ0sURAcPP_kiyRJ36DBqcP?authuser=1#scrollTo=ggi3z2eJ1zTQ>