Tarea 5

Sonia Torres Ibarra

12 de octubre de 2025

1. Introducción

En el análisis de datos, los algoritmos no supervisados juegan un papel fundamental al permitir descubrir estructuras subyacentes sin necesidad de etiquetas o clases predefinidas. Estos métodos son ampliamente utilizados en contextos donde se busca identificar patrones, segmentar observaciones o reducir la complejidad de los datos, como en la exploración de clientes, análisis de comportamiento o clasificación automática de información.

Entre las técnicas más comunes se encuentran los algoritmos de clustering, que agrupan elementos con características similares. Sin embargo, la elección del algoritmo adecuado depende tanto de la naturaleza de los datos como de la forma en que se desea representar la similitud entre observaciones. En este trabajo se investigó una alternativa adicional a los métodos vistos en clase, considerando el modelo DBSCAN, el cual ofrece ventajas frente a algoritmos clásicos como k-means, especialmente cuando los datos presentan ruido, densidades variables o formas de agrupamiento no esféricas.

Asimismo, se abordan métricas de evaluación para determinar el número óptimo de grupos, tales como los índices de Silhouette Score y Davies-Bouldin, que permiten cuantificar la calidad de la segmentación obtenida. La combinación de estos enfoques busca no solo obtener una agrupación más representativa, sino también justificarla matemáticamente a partir de medidas objetivas de desempeño.

Finalmente, se presenta la aplicación práctica de un algoritmo no supervisado a un conjunto de datos reales, acompañada del análisis de resultados, la discusión de las métricas empleadas y la referencia a trabajos científicos previos relacionados con la metodología implementada.

2. Descripción de los datos

La Comisión Nacional Bancaria y de Valores (CNBV), en colaboración con el Instituto Nacional de Estadística y Geografía (INEGI), realizaron la Encuesta Nacional de Inclusión Financiera (ENIF) 2024. Su objetivo es generar información estadística e indicadores oficiales a nivel nacional que permitan a las autoridades financieras hacer diagnósticos, diseñar políticas públicas y establecer metas en materia de inclusión y educación financieras. Asimismo, incorporar cambios y actualizaciones para dar atención a nuevos requerimientos de información y consideraciones en la Política Nacional de Inclusión Financiera (PNIF).

3. Antecedentes

Investigaciones relacionadas:

El autor Cando [1] menciona que "DBSCAN, se distingue por su capacidad para identificar agrupaciones de formas irregulares y destacar puntos de datos aislados como outliers, lo que lo convierte en una herramienta valiosa para discernir comportamientos financieros altamente atípicos entre los socios".

4. Metodología

El DBSCAN es un algoritmo no supervisado muy conocido en materia de Clustering. Fue presentado en 1996 por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiawei Xu.[3]

El DBSCAN es un algoritmo sencillo que define los clústeres mediante la estimación de la densidad local. Se puede dividir en 4 etapas:

- Para cada observación miramos el número de puntos a una distancia máxima ϵ de ella. Esta zona se denomina ϵ-vecindad de la observación.
- Si una observación tiene al menos un cierto número de vecinos, incluida ella misma, se considera una observación central. En este caso, se ha detectado una observación de alta densidad.
- Todas las observaciones en la vecindad de una observación central pertenecen al mismo clúster. Puede haber observaciones centrales cercanas entre sí. Por lo tanto, de un paso a otro, se obtiene una larga secuencia de observaciones centrales que constituyen un único clúster.
- Cualquier observación que no sea una observación central y que no tenga ninguna observación central en su vecindad se considera una anomalía.

DBSCAN por lo general utiliza la distancia euclidiana

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_i - q_i)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

En cada observación, para contar el número de vecinos a como máximo una distancia ϵ, calculamos la distancia euclidiana entre el vecino y la observación y comprobamos si es inferior a ϵ.

DBSCAN utiliza dos parámetros principales [2]:

- eps (épsilon): La distancia máxima entre dos puntos para que se consideren vecinos.
- min samples: El número mínimo de puntos necesarios para formar una región densa.

Por lo tanto, es necesario definir 2 datos antes de utilizar DBSCAN:

- ¿Qué distancia ϵ hay que determinar para cada observación la ϵ-vecindad?
- ¿Cuál es el número mínimo de vecinos necesario para considerar una observación como una observación central?

En este trabajo, para seleccionar los parámetros óptimos del algoritmo DBSCAN , se utilizaron métricas internas de validación de clustering, específicamente el Silhouette Score y el Davies-Bouldin Index.

El Silhouette Score [5] mide la cohesión interna de un cluster y la separación entre clusters. Para cada punto i , se calcula la distancia promedio a los demás puntos de su propio cluster (a_i) y la distancia promedio al cluster más cercano (b_i), dando como resultado un valor:
$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$

con rango [1,1], donde valores cercanos a 1 indican clusters bien definidos y separados, valores cercanos a 0 indican puntos en los límites de los clusters, y valores negativos sugieren posible asignación incorrecta [6].

El coeficiente de la Silhouette (s_i) se define para cada clúster. Cuando se dispone de más de dos clústeres se calcula el valor promedio de la Silhouette para disponer de una medida de la calidad de los clústeres.

Por otro lado, el Davies-Bouldin Index (DBI) [4] se basa en relacionar la dispersión dentro de los clústeres (intra-clúster) y la separación entre clústeres (inter-clúster). Por un lado, la dispersión intra-clúster mide la separación de los puntos dentro de cada clúster. Una dispersión intra-clúster baja indica que los puntos dentro de un grupo están muy cercanos entre sí, algo que es deseable en un buen clustering. Por otro lado, la dispersión inter-clúster mide la separación entre los grupos. Una dispersión inter-clúster alta indica que los grupos están muy alejados entre sí, lo que también es deseable en un buen clustering. El índice de Davies-Bouldin se construye como el cociente de ambos valores. Por lo que cuando los clústeres están separados y son compactos el valor de este índice se minimiza.:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}}$$

donde S_i es la dispersión promedio dentro del cluster i y M_{ij} la distancia entre los centroides de los clusters i y j. Valores bajos de DBI indican clusters más compactos y mejor separados.

Para aplicar estas métricas, se probó DBSCAN con distintas combinaciones de eps y min samples, calculando Silhouette y Davies-Bouldin para los clusters detectados (excluyendo outliers). La combinación que maximizó el Silhouette Score y minimizó el Davies-Bouldin Index fue seleccionada como óptima, asegurando así una segmentación más representativa y coherente de los datos.

5. Resultados

En la figura 1 (p.4) se muestran los resultados de aplicar DBSCAN sobre los datos, evaluando distintos valores de los parámetros eps y min samples. Se incluyeron métricas de validación como Silhouette Score y Davies-Bouldin Index para determinar la calidad de los clusters obtenidos y los parámetros a utilizar.

```
=== Resultados métricas DBSCAN ===
     eps  min_samples  n_clusters  n_outliers  silhouette  davies_bouldin
2   0.75            8           3        5900        0.72            0.37
0   0.75            3          76        5633        0.72            0.33
6   1.00            8           4        5892        0.69            0.44
1   0.75            5          13        5842        0.68            0.43
3   0.75           10           2        5909        0.68            0.47
7   1.00           10           2        5909        0.68            0.47
4   1.00            3          96        5558        0.67            0.42
5   1.00            5          16        5825        0.65            0.48
11  1.50           10           6        5777        0.47            0.77
8   1.50            3         142        5049        0.38            0.81
10  1.50            8          16        5666        0.37            0.93
9   1.50            5          46        5429        0.35            0.89
14  2.00            8          20        5177        0.12            1.35
15  2.00           10          18        5281        0.11            1.35
13  2.00            5          55        4797        0.08            1.15
12  2.00            3         132        4410        0.08            1.03

Mejores parámetros -> eps: 0.75, min_samples: 8
```

Figura 1: Resultados para pruebas DBI y Silhouette

El valor más alto de Silhouette (0.72) se obtiene para eps=0.75 y min samples=3 o min samples=8. Esto sugiere que los clusters detectados bajo estos parámetros presentan una buena separación relativa entre ellos.

Para el índice de Davies-Bouldin, los valores más bajos (0.33–0.37) se observan también en los parámetros de eps=0.75 y min samples bajos, corroborando la cohesión detectada por Silhouette.

Cuadro 1: Clusters obtenidos

Cluster	Casos
-1	5900
0	20
1	15
2	9

Considerando ambas métricas, los parámetros eps=0.75 y min samples=8 fueron seleccionados como los más adecuados, ya que logran un balance entre número razonable de clusters y alta calidad de agrupamiento.

DBSCAN, con eps=0.75 y min samples=8, detecta tres grupos principales de usuarios y destaca la heterogeneidad mediante la gran cantidad de outliers.

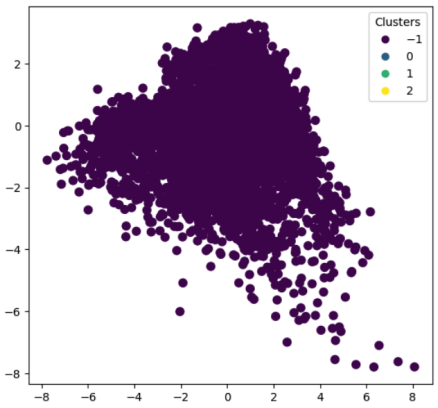

Figura 2: Gráfico DBSCAN con los parámetros seleccionados.

6. Conclusiones y discusión

Esta tarea permitió explorar la aplicación del algoritmo DBSCAN, como una alternativa para la detección de patrones y agrupamientos en aprendizaje no supervisado. A diferencia de otros métodos más rígidos como k-means, DBSCAN no requiere definir previamente el número de grupos.

Para mejorar la calidad del agrupamiento y reducir el uso de recursos computacionales, se aplicó una reducción de variables mediante PCA (Análisis de Componentes Principales), que no se menciono en los resultados ya que no era parte del fin de esta actividad. Posteriormente, se realizó una búsqueda de los parámetros eps y min samples.

La búsqueda de parámetros se realizo con el índice de Silhouette y el índice de Davies-Bouldin. Estos cálculos permitieron analizar la relación interna de los

clústeres y la separación entre ellos, proporcionando una visión más completa del desempeño del modelo. Los valores reflejaron una estructura de clústeres bien definida y con baja superposición entre grupos. Sin embargo, se identificó una cantidad considerable de outliers, lo que sugiere la existencia de observaciones atípicas o de baja densidad en el conjunto de datos.

Los hallazgos confirman que DBSCAN es un buen algoritmo para la identificación de patrones en datos complejos, siempre que se realice una selección cuidadosa de sus parámetros y se complementen los resultados con métricas de validación adecuadas.

Referencias

[1] J. G. G. CANDO. Facultad ingeniería maestría en sistemas de información mención ciencia de datos.

[2] DataCamp. Guía del algoritmo de agrupación dbscan, 2025. URL https://www.datacamp.com/es/tutorial/dbscan-clustering-algorithm.

[3] DataScientest. Machine learning clustering: el algoritmo dbscan, 2022. URL https://datascientest.com/es/machine-learning-clustering-dbscan.

[4] D. Rodríguez. El índice de davies-bouldinen para estimar los clústeres en k-means e implementación en python., 2023. URL https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-c

[5] D. Rodríguez. Número óptimo de clústeres con silhouette e implementación en python, 2023. URL https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-impl

[6] Scikit-Learn. Clustering. URL https://scikit-learn.org/stable/modules/clustering.html.