Aplicación del algoritmo DBSCAN para la identificación de patrones en datos de ventas de café

Autor: José Javier Tovar Pérez
Institución: Universidad Autónoma de Nuevo León
Correo: jjavier.tovar@uanl.edu.mx

Octubre 2025

Resumen

Este trabajo presenta un estudio sobre la aplicación del algoritmo de agrupamiento no supervisado DBSCAN (Density-Based Spatial Clustering of Applications with Noise) a un conjunto de datos de ventas de café. El objetivo fue explorar estructuras subyacentes y patrones de comportamiento temporal sin utilizar etiquetas predefinidas. Los resultados muestran la formación de tres grupos principales distinguibles visualmente mediante reducción de dimensionalidad (PCA), a pesar de que el algoritmo generó un mayor número de clústeres (51) debido a la sensibilidad de sus parámetros. Se discute el comportamiento del modelo, la influencia del parámetro *eps* y la relevancia de DBSCAN en contextos de datos transaccionales.

1. Introducción

El análisis de datos transaccionales en puntos de venta permite descubrir patrones que pueden optimizar estrategias comerciales. En particular, los establecimientos dedicados a la venta de café generan información continua sobre horarios de compra, montos, frecuencia y tipo de transacción. Este tipo de datos puede revelar comportamientos de consumo característicos que resultan útiles para la planeación de inventarios o estrategias de marketing.

El presente trabajo aplica un algoritmo de aprendizaje no supervisado, **DBSCAN**, con el propósito de identificar estructuras latentes en los datos de ventas de café. A diferencia de los métodos basados en centroides como K-Means, DBSCAN agrupa observaciones según la densidad de puntos en el espacio de características, lo que lo hace especialmente adecuado para datos con ruido o distribuciones irregulares.

2. Modelo matemático del algoritmo DBSCAN

DBSCAN (Ester et al., 1996) define los grupos en función de dos parámetros: *eps* (radio de vecindad) y *minPts* (número mínimo de puntos requeridos para formar un clúster).

Sea un conjunto de puntos $D = \{x_1, x_2, \ldots, x_n\}$ en un espacio métrico. Para un punto p, se define su vecindad ε como:

$$N_\varepsilon(p) = \{q \in D \mid dist(p, q) \leq \varepsilon\}$$

Un punto p es considerado un *punto núcleo* si $|N_\varepsilon(p)| \geq minPts$. Los puntos que no cumplen este criterio pero se encuentran dentro de la vecindad de un punto núcleo son denominados *puntos frontera*. Aquellos que no pertenecen a ningún grupo se etiquetan como *ruido*.

El algoritmo forma clústeres expandiendo recursivamente las regiones de densidad alta. Formalmente, dos puntos p y q son *conectados por densidad* si existe una cadena de puntos núcleo que une a ambos dentro del radio ε.

3. Metodología

3.1. Datos

Se trabajó con un conjunto de **3547 registros de ventas de café** entre marzo de 2024 y marzo de 2025. Las variables consideradas fueron:

- **Date:** fecha de la transacción.
- **hour_of_day:** hora del día (entero de 0–23).
- **money:** monto de venta.
- **coffee_ordered:** tipo de café vendido.
- **Time_of_Day, Weekdaysort, Monthsort:** variables categóricas numéricas derivadas.

3.2. Preprocesamiento

Antes del agrupamiento, las variables fueron escaladas con *StandardScaler* para normalizar la influencia de magnitudes. Posteriormente se aplicó una reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) para visualización, conservando dos componentes principales.

3.3. Implementación de DBSCAN

Se utilizó el algoritmo DBSCAN disponible en la librería `scikit-learn`. El parámetro *eps* se determinó mediante el método de la **k-distancia**, que consiste en graficar la distancia media al vecino más cercano para distintos valores de k y localizar el punto de inflexión de la curva.

En este caso, el punto de cambio significativo se observó en **eps = 0.63**, como puede apreciarse en la figura 1. Para valores inferiores, el número de clústeres disminuye abruptamente, mientras que a partir de ese punto la curva muestra un comportamiento exponencial creciente, por lo que se adoptó dicho valor como óptimo.

El parámetro *minPts* se estableció en 5, valor comúnmente recomendado para bases con ruido moderado.

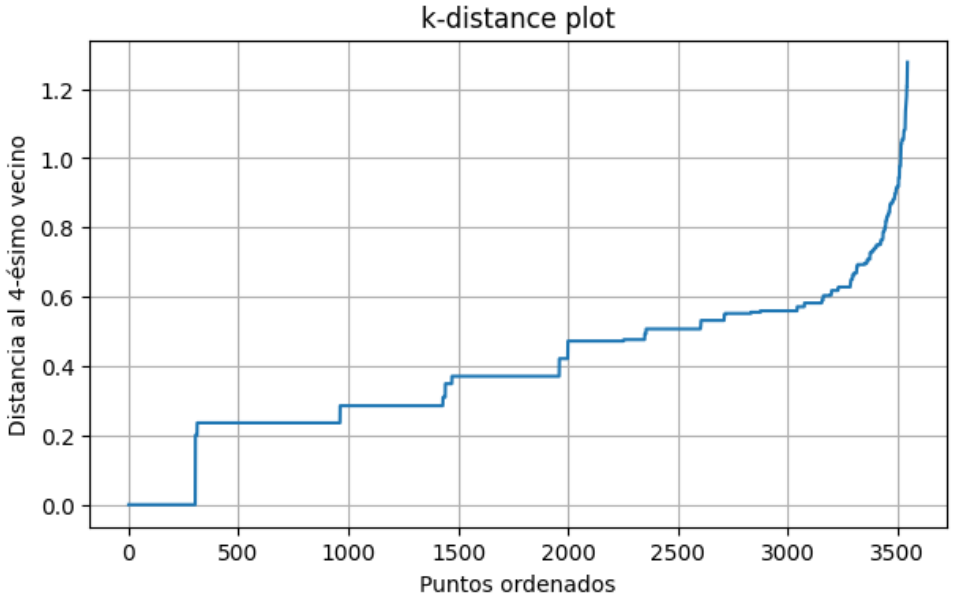

Figura 1: Se representan los puntos ordenados de la k-distancia entre los mismos.

4. Resultados

El algoritmo identificó un total de **51 clústeres** al usar $eps = 0{,}63$, de los cuales tres agrupamientos principales mostraron cohesión visual clara en el plano de componentes principales (PCA), como puede apreciarse en la figura 2.

Los tres grupos principales se distribuyen en forma de rectángulos girados aproximadamente 40° en sentido horario, con ligera separación en la parte superior de los conjuntos, lo que sugiere diferencias de densidad entre segmentos horarios o patrones de consumo diferenciados.

El número elevado de grupos menores (51) refleja la sensibilidad del modelo a pequeñas variaciones en los datos, lo que podría deberse a la presencia de transacciones atípicas o diferencias horarias finas.

5. Discusión

La aplicación de DBSCAN permitió detectar estructuras que no serían capturadas por métodos basados en centroides. El modelo distingue regiones de alta densidad en torno a ciertos horarios o montos de compra, separando transacciones aisladas como ruido. Aunque el número de clústeres formales fue alto, el análisis visual evidencia solo tres agrupaciones significativas, lo que resalta la importancia de la interpretación posterior al modelado.

El método de k-distancia resultó adecuado para determinar el valor de *eps*, pues mostró un cambio abrupto en la pendiente alrededor de 0.63, coherente con la transición entre densidad estable y sobrefragmentación.

DBSCAN demostró ser eficaz para datos transaccionales, donde la forma de los grupos no necesariamente es esférica. Sin embargo, su desempeño depende críticamente de la correcta elección de *eps*, lo que limita su automatización sin análisis gráfico o heurístico.

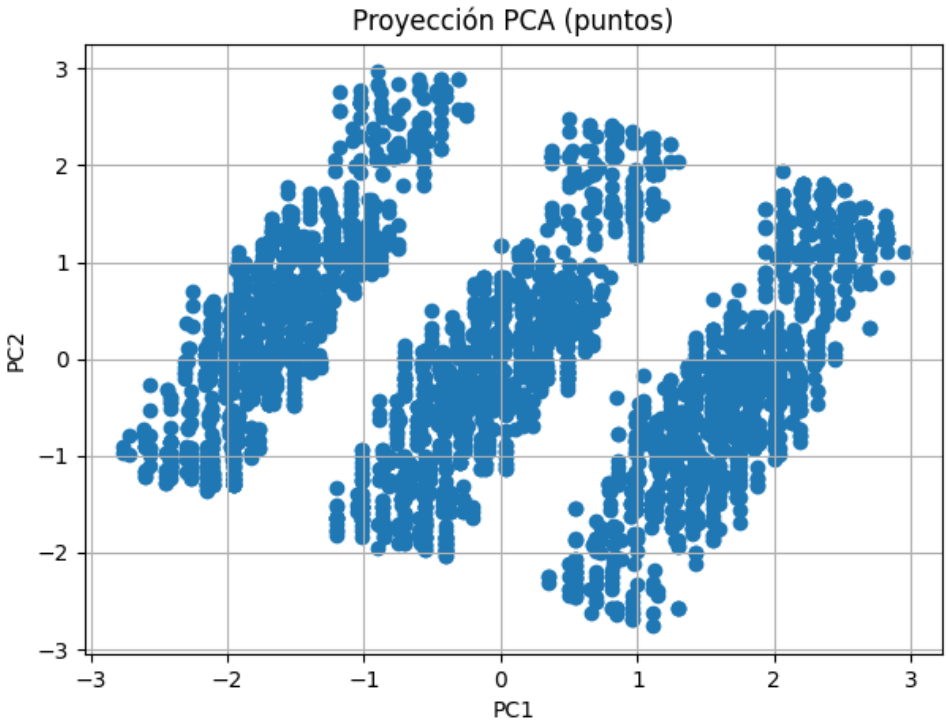

Figura 2: Visualización PCA de los clústeres detectados por DBSCAN. Se aprecian tres conglomerados principales con orientación similar.

6. Metodología supervisada: predicción del monto de venta

6.1. Objetivo supervisado

Con el propósito de complementar el análisis no supervisado, se formuló un modelo supervisado orientado a predecir el monto de venta (`money`) a partir de variables temporales y de transacción. Entre las variables predictoras consideradas se incluyeron el tipo de café solicitado (`coffee_ordered`), la hora del día (`hour_of_day`), el tipo de pago (`cash_type`), y variables derivadas como el periodo del día (`Time_of_Day`), el día de la semana (`Weekdaysort`) y el mes (`Monthsort`).

Este enfoque permite estimar la relación entre los patrones de consumo y el valor de las transacciones, con el objetivo de generar un modelo interpretable que apoye decisiones de pronóstico y planeación de ventas.

6.2. Modelo seleccionado

De entre los algoritmos supervisados revisados (Regresión Lineal, Random Fprest Regressor, Ridge-Lasso, K-NN), se eligió el modelo **Random Forest Regressor** por su capacidad de capturar relaciones no lineales, su robustez frente a ruido y outliers, y su interpretabilidad mediante la importancia de características. El modelo fue entrenado sobre un conjunto de entrenamiento equivalente al 75 % de los datos, reservando el 25 % restante para validación.

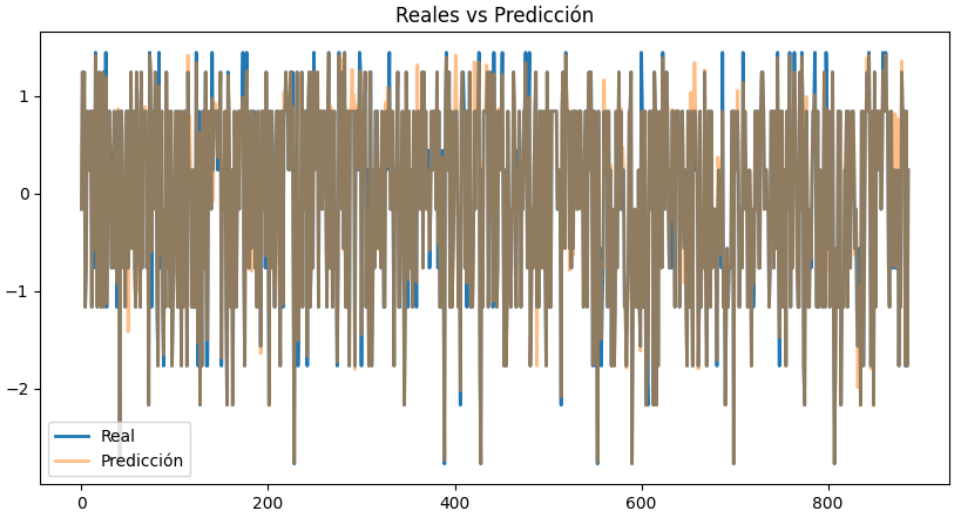

Figura 3: Visualización de las diferencias entre los valores reales y los predichos por el modelo.

6.3. Métricas de evaluación

Para evaluar el desempeño del modelo se emplearon las siguientes métricas:

- **MAE (Mean Absolute Error)**: error medio absoluto, interpretable en unidades monetarias.
- **RMSE (Root Mean Squared Error)**: raíz del error cuadrático medio, sensible a errores grandes.
- **MAPE (Mean Absolute Percentage Error)**: error porcentual medio absoluto, expresado como porcentaje.

Los resultados obtenidos fueron:

$$\text{MAE} = 0{,}050, \quad \text{RMSE} = 0{,}144, \quad \text{MAPE} = 9{,}486\,\%$$

Estos valores indican que el modelo logra un error promedio de alrededor del 9.5 % en la predicción del monto de venta, lo cual se considera adecuado para un modelo con variables discretas y temporales limitadas, dichos resultados pueden apreciarse en la figura 3.

6.4. Importancia de variables

El análisis de importancia de características del modelo mostró los siguientes resultados:

El tipo de café resulta ser el factor determinante del monto de venta, lo cual concuerda con la estructura de precios diferenciados entre variedades (por ejemplo, *americano*, *capuccino*, *moka*). La influencia del mes sugiere cierta estacionalidad en la demanda, mientras que las variables horarias y semanales tienen impacto marginal.

Variable	Importancia	Interpretación
`coffee_ordered`	0.858	Tipo de café comprado (mayor influencia en el precio)
`monthsort`	0.125	Variación estacional de ventas
`hour_of_day`	0.009	Fluctuación por hora (consumo matutino/vespertino)
`weekdaysort`	0.006	Día de la semana (ligera influencia)
`time_of_day`	0.001	Momento general del día (marginal)

Cuadro 1: Importancia relativa de las variables predictoras en el modelo Random Forest.

7. Discusión de resultados

El desempeño del modelo supervisado muestra que es posible estimar el monto de venta con una precisión razonable (*MAPE* debajo del 10 %) utilizando únicamente variables internas del punto de venta. Dado que la mayor parte de la variabilidad está explicada por el tipo de café, el modelo podría ampliarse incorporando nuevas variables (promociones, tamaño de porción, temperatura ambiental, o día festivo) para capturar patrones de consumo más finos.

En comparación con el análisis no supervisado (DBSCAN), donde se identificaron aproximadamente tres grupos visualmente distinguibles, el modelo supervisado confirma la relación dominante entre la composición del pedido y el monto total. Ambas aproximaciones ofrecen perspectivas complementarias: mientras DBSCAN revela segmentaciones naturales en el comportamiento de compra, el Random Forest cuantifica la contribución de cada factor a la variabilidad del ingreso.

Referencias

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[2] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*.

[3] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[4] M. Ester, H. P. Kriegel, J. Sander y X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.

[5] R. Hidayat y H. Kusniyati, "Clustering Analysis in Sales Grouping Using The K-Means Algorithm at Cafe 47°Coffee," *CESS (Journal of Computer Engineering, System and Science)*, vol. 7, no. 2, pp. 420–434, 2022.

[6] S. Ghanvat et al., "Customer Segmentation using Clustering Algorithm in Machine Learning," *Journal of Information Technology and Sciences*, 2023.