

# Aplicación de algoritmos no supervisados y supervisados para la identificación y predicción 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 combinada de técnicas de aprendizaje no supervisado y supervisado en un conjunto de datos de ventas de café. En una primera fase se empleó el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) para descubrir estructuras latentes en los registros transaccionales, mientras que en una segunda fase se implementó un modelo de regresión basado en Random Forest para predecir el monto de venta. Los resultados muestran la existencia de tres grupos principales de comportamiento de compra, y un modelo predictivo con un error medio absoluto del 5 % y un error porcentual medio del 9.5 %. El análisis demuestra el potencial de la minería de datos en entornos comerciales para la segmentación y predicción de patrones de consumo.

## 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 enfoque dual: primero, un algoritmo de aprendizaje no supervisado (**DBSCAN**) con el propósito de identificar estructuras latentes en los datos de ventas; y posteriormente, un modelo de aprendizaje supervisado (**Random Forest Regressor**) para predecir el monto de venta. Esta combinación permite no solo descubrir patrones naturales de comportamiento, sino también cuantificar la influencia de las variables sobre las transacciones.

## 2. Metodología

El proceso metodológico se desarrolló en dos fases: (1) descubrimiento de patrones mediante agrupamiento no supervisado, y (2) predicción del monto de venta a través de

un modelo supervisado. Ambas se fundamentan en literatura consolidada en aprendizaje automático.

## 2.1. Datos y preprocesamiento

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 (0–23).
- **money:** monto de venta.
- **coffee\_ordered:** tipo de café vendido (valor categórico).
- **Time\_of\_Day, Weekdaysort, Monthsort:** variables categóricas numéricas derivadas.

Antes del modelado, las variables numéricas fueron escaladas con *StandardScaler* y las categóricas fueron codificadas numéricamente. Para visualización se empleó reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), reteniendo dos componentes principales.

## 2.2. Fase no supervisada: agrupamiento con DBSCAN

El algoritmo DBSCAN [3] define grupos en función de dos parámetros: *eps* (radio de vecindad) y *minPts* (número mínimo de puntos por clúster). A diferencia de métodos basados en centroides como K-Means, DBSCAN agrupa observaciones por densidad, lo que lo hace adecuado para datos con ruido o distribuciones irregulares.

El valor óptimo de *eps* se determinó mediante el método de la **k-distancia** [4], identificando el punto de inflexión de la curva en **0.63**. Para valores inferiores, el número de clústeres decrece abruptamente, mientras que a partir de dicho punto la gráfica adopta una tendencia exponencial, indicativa de sobrefragmentación. El parámetro *minPts* se fijó en 5.

El modelo detectó **51 clústeres**, aunque la reducción PCA mostró tres conglomerados principales bien definidos (Figura 2), interpretables como patrones de compra diferenciados por horario o tipo de pedido.

## 2.3. Fase supervisada: predicción mediante Random Forest

En la segunda fase se implementó un modelo **Random Forest Regressor** [1] con el objetivo de predecir el monto de venta (**money**) a partir de variables temporales y de transacción (**coffee\_ordered**, **hour\_of\_day**, **cash.type**, **Time\_of\_Day**, **Weekdaysort**, **Monthsort**). El conjunto de datos se dividió en 75 % para entrenamiento y 25 % para validación.

El modelo se seleccionó por su capacidad para capturar relaciones no lineales, su robustez ante ruido y su interpretabilidad mediante la medida de importancia de características. La Figura 3 muestra la comparación entre valores reales y predichos.

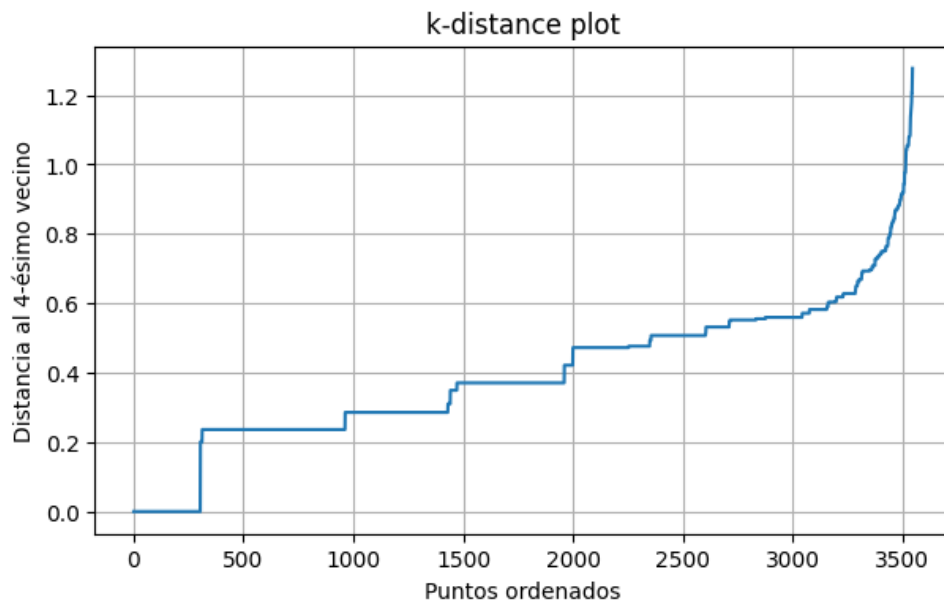


Figura 1: Curva de k-distancia empleada para determinar el valor óptimo de  $\epsilon$ .

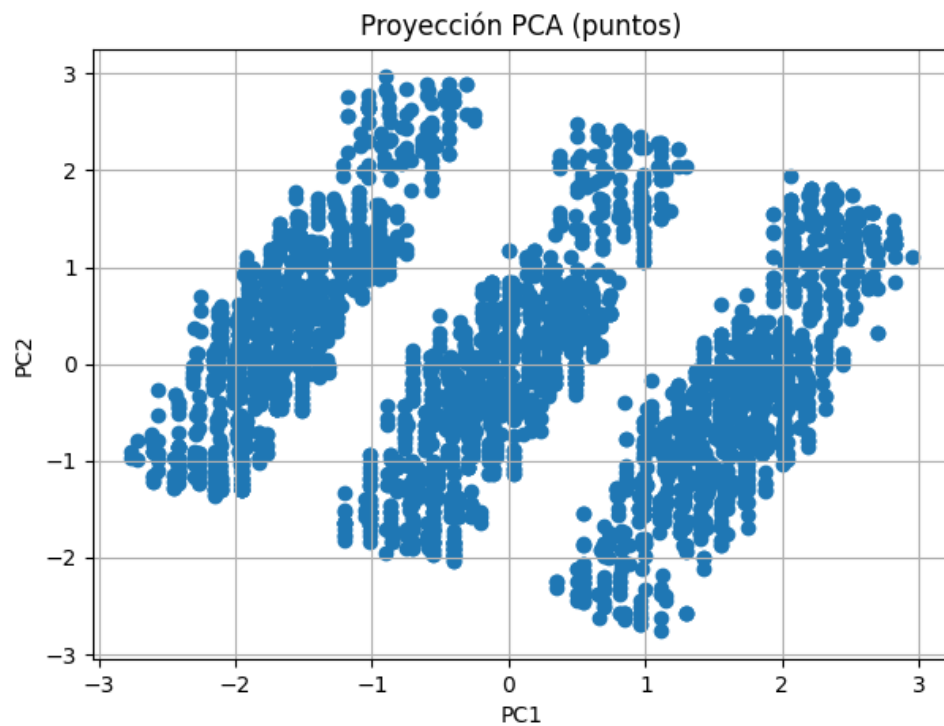


Figura 2: Visualización PCA de los clústeres detectados por DBSCAN. Se aprecian tres grupos principales con orientación similar.

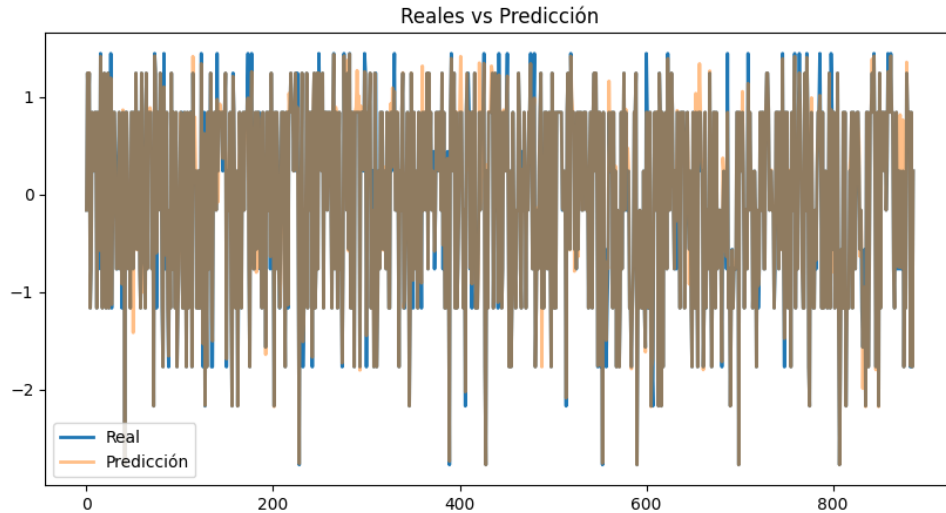


Figura 3: Comparación entre valores reales y predichos del monto de venta mediante Random Forest.

## 2.4. Evaluación del modelo

El desempeño se evaluó mediante las métricas MAE, RMSE y MAPE [2]. Los resultados obtenidos fueron:

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

Estos valores reflejan un error promedio relativo menor al 10 %, lo que indica un desempeño aceptable considerando la naturaleza discreta y temporal de las variables.

## 2.5. Importancia de variables

Variable	Importancia	Interpretación
coffee_ordered	0.858	Tipo de café vendido (mayor influencia)
monthsort	0.125	Variación estacional
hour_of_day	0.009	Fluctuación horaria
weekdaysort	0.006	Día de la semana (impacto leve)
time_of_day	0.001	Momento general del día (marginal)

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

El tipo de café representa el factor determinante del monto de venta, seguido por la estacionalidad mensual. Las variables horarias y semanales aportan menor influencia, lo cual coincide con la segmentación observada en la fase no supervisada.

## 3. Discusión de resultados

La aplicación de DBSCAN permitió identificar tres patrones principales de comportamiento en las ventas, posiblemente asociados con horarios o tipos de consumo. El alto

número de clústeres menores refleja la sensibilidad del algoritmo a la densidad local, pero su visualización mediante PCA confirma una estructura consistente.

El modelo Random Forest mostró un buen desempeño predictivo (MAPE inferior al 10 %), destacando la fuerte influencia del tipo de producto en el monto total. La combinación de ambos enfoques —agrupamiento y predicción— ofrece una visión integral: el primero revela la estructura natural de los datos, y el segundo cuantifica las relaciones internas.

## Referencias

- [1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- [2] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.
- [3] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96*, 226–231.
- [4] Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2), 169–194.
- [5] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.