Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Aprendizaje Automático

Análisis de factores para la predicción del inventario

Liliana Dueñas Barrón
Matrícula: 1291120

November 22, 2025

Abstract

Este estudio presenta un análisis integral orientado a comprender los factores que influyen en el nivel de inventario (stock) en un entorno de retail, aplicando técnicas estadísticas y métodos de aprendizaje automático. Se realizaron análisis descriptivos, pruebas de normalidad univariada y multivariada (Shapiro–Wilk, Mahalanobis y Mardia), los cuales revelaron que los datos no siguen una distribución normal, justificando el uso de modelos no paramétricos. Posteriormente, se llevó a cabo una selección de características mediante F-Test, Información Mutua, Random Forest y Lasso, encontrando de manera consistente que el *precio unitario* es la variable con mayor influencia sobre el stock. Para la detección de estructuras subyacentes se aplicó DBSCAN, evaluado mediante los índices de Silhouette, Calinski–Harabasz y Davies–Bouldin, obteniendo agrupamientos válidos y la identificación de *outliers*.

En la etapa de modelado supervisado, se entrenó un *Random Forest Regressor* que alcanzó un desempeño sobresaliente ($R^2 = 0.9995$), permitiendo generar predicciones precisas para nuevos escenarios. Finalmente, se desarrolló un diseño factorial 2^3 para evaluar experimentalmente los efectos principales e interacciones de precio, cantidad y venta total sobre el stock predicho. Los resultados del DOE confirmaron que únicamente el precio unitario ejerce un efecto significativo, sin interacción entre factores. El trabajo demuestra que la combinación de análisis estadístico, aprendizaje automático y diseño experimental constituye una herramienta robusta para mejorar la gestión de inventarios en el sector retail y apoyar la toma de decisiones basada en datos.

1 Introducción

El análisis de datos en contextos comerciales, particularmente en entornos de retail, permite comprender el comportamiento de productos, optimizar inventarios y apoyar la toma de decisiones estratégicas. En este estudio se realiza una investigación integral que combina técnicas de preprocesamiento, estadística descriptiva, evaluación de normalidad, selección de características, métodos no supervisados de agrupamiento, modelado predictivo supervisado y diseño de experimentos, con el objetivo de identificar patrones relevantes y construir un modelo capaz de predecir niveles de inventario (stock).

Se emplean métodos clásicos y modernos, incluyendo pruebas estadísticas como Shapiro–Wilk y Mardia, algoritmos como DBSCAN y Random Forest, además de un diseño factorial 2^3 para analizar efectos principales e interacciones. Este enfoque multidimensional permite no sólo modelar adecuadamente el comportamiento del inventario, sino también validar experimentalmente qué variables tienen impacto real sobre él. El análisis aporta evidencia

cuantitativa y visual que resulta útil para aplicaciones prácticas, tales como sistemas de soporte a la decisión en gestión de inventarios y planeación de compras.

2 Planteamiento del problema

En el entorno de retail, comprender qué factores determinan el nivel adecuado de stock es fundamental para evitar pérdidas por desabasto o sobreinventario. Variables como el precio unitario, la cantidad adquirida y la venta total influyen en las dinámicas comerciales, pero su relación exacta con el inventario no siempre es evidente ni lineal. Además, los datos generados por los sistemas de punto de venta suelen presentar asimetrías, valores atípicos y estructuras complejas que dificultan el uso de modelos tradicionales basados en normalidad.

Ante esta situación, surge la pregunta central del estudio:

> **¿Qué variables explican de manera significativa el comportamiento del stock y cómo puede modelarse su relación mediante técnicas estadísticas y de machine learning?**

Para responderla, es necesario:

1. Analizar estadísticamente la distribución y relaciones entre variables.
2. Identificar características relevantes mediante múltiples métodos de selección.
3. Detectar patrones naturales no supervisados mediante algoritmos de clustering.
4. Construir un modelo supervisado robusto para la predicción del stock.
5. Validar experimentalmente los efectos principales e interacciones mediante un diseño factorial 2^3.

La resolución de este problema contribuye a mejorar las políticas de inventarios y la planeación operativa en un contexto comercial real.

3 Metodología

La metodología empleada en este estudio se estructuró en cinco fases integradas, cada una sustentada en herramientas estadísticas y computacionales robustas.

3.1 Análisis descriptivo y pruebas de normalidad

Se utilizaron estadísticos descriptivos (media, mediana, desviación estándar, cuartiles) y visualizaciones (histogramas, boxplots, matriz de correlación) para caracterizar las variables: *precio unitario*, *cantidad de compra*, *venta total* y *stock*.

Posteriormente se aplicaron las siguientes pruebas:

- **Shapiro–Wilk**: evaluación de normalidad univariada.
- **Distancia de Mahalanobis** acompañada de gráfico *QQ-plot* para detección de valores atípicos multivariados.
- **Prueba multivariada de Mardia** para evaluar la normalidad conjunta mediante asimetría y curtosis multivariada.

Los resultados mostraron que ninguna variable presenta distribución normal y que existe asimetría multivariada significativa, lo cual justifica el uso de modelos no paramétricos.

3.2 Selección de características

Se evaluó la relevancia de las variables *precio unitario*, *cantidad de compra* y *venta total* mediante cuatro técnicas:

- **F-Test (ANOVA)**
- **Mutual Information**
- **Random Forest Importance**
- **Lasso (L1)**

Los resultados coincidieron en que la variable con mayor contribución al modelo es *precio unitario*, confirmándola como el factor dominante en la predicción del stock.

3.3 Agrupamiento no supervisado

Para identificar estructuras naturales en los datos se estandarizaron las variables y se aplicó el algoritmo **DBSCAN**, debido a que:

- No requiere predefinir el número de grupos.
- Detecta estructuras arbitrarias y presencia de valores atípicos.

- Es adecuado para datos ruidosos y distribuciones no lineales.

La calidad del agrupamiento se evaluó mediante tres índices:

- **Silhouette**
- **Calinski–Harabasz**
- **Davies–Bouldin**

Los valores obtenidos indicaron agrupamientos válidos y una estructura diferenciada en los datos, además de una detección coherente de *outliers*.

3.4 Modelo de predicción supervisado

Se construyó un modelo **Random Forest Regressor**, adecuado para datos no lineales y no normales. El modelo fue entrenado y evaluado mediante las métricas:

- **MAE**
- **RMSE**
- **R^2**

El valor obtenido de $R^2 = 0.9995$ evidenció un ajuste sobresaliente. Además, el análisis de importancia de variables confirmó nuevamente que *precio unitario* es el predictor dominante. El modelo permitió generar predicciones para nuevos escenarios mediante extrapolación controlada.

3.5 Diseño de Experimentos (DOE 2^3)

Finalmente, se aplicó un diseño factorial 2^3 para evaluar los efectos principales de:

- Precio unitario (P)
- Cantidad de compra (C)
- Venta total (V)

Así como las interacciones $P \times C$, $P \times V$, $C \times V$ y $P \times C \times V$.

El DOE se ejecutó utilizando como respuesta el *stock* predicho por el modelo Random Forest, integrando así el enfoque predictivo con una metodología experimental clásica. Los resultados mostraron que únicamente el **precio unitario** tiene efecto significativo, mientras que las interacciones no aportan variaciones relevantes sobre el nivel de inventario.

4 Análisis Descriptivo de los Datos

En esta sección se presenta un análisis estadístico y visual exhaustivo de las variables fundamentales del estudio: *precio unitario*, *cantidad de compra*, *venta total* y *stock*. El objetivo es caracterizar su comportamiento individual, su relación entre sí y evaluar si cumplen los supuestos de normalidad univariada y multivariada, lo cual influye directamente en la elección de modelos posteriores.

4.1 Estadísticos descriptivos

Se calcularon estadísticas básicas como la media, mediana, desviación estándar, cuartiles y rango. Los resultados permiten obtener un panorama inicial de la estructura de los datos.

Variable	**count**	**mean**	**std**	**min**	**max**	**25%**	**50%**
Precio unitario	754	76.98	83.88	5.0	500	25.0	45.0
Venta total	754	232.03	290.87	5.0	2000	50.0	130.0
Cantidad compra	754	2.98	1.43	1.0	5.0	2.0	3.0
Stock	754	267.47	130.89	30.0	485.0	140.0	224.0

Table 1: Estadísticos descriptivos de las variables.

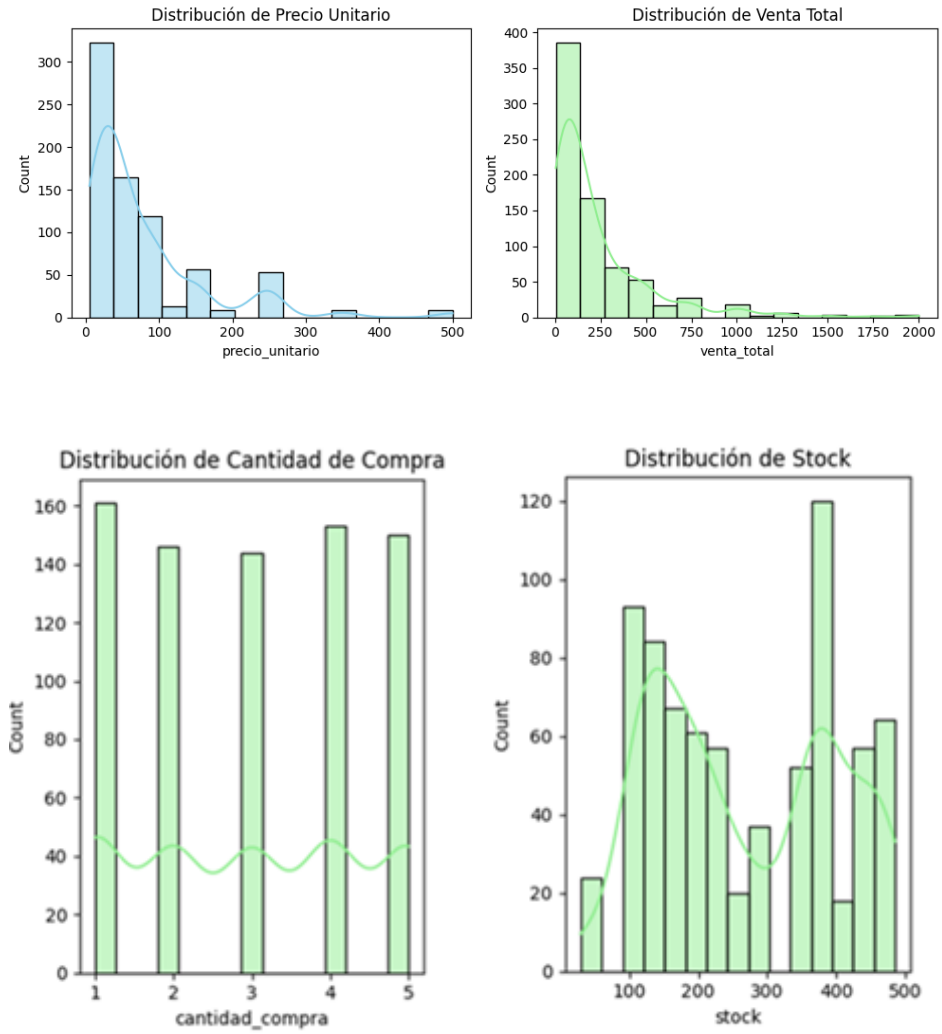
A partir de estos valores se observan las siguientes características generales:

- **Precio unitario**: presenta gran dispersión, valores elevados en percentiles superiores y presencia de *outliers*.
- **Cantidad compra**: variable discreta entre 1 y 5, con distribución casi uniforme.
- **Venta total**: distribución altamente asimétrica a la derecha con valores extremos altos.
- **Stock**: distribución más uniforme, con variaciones estructuradas pero sin valores extremos severos.

Estas características anticipan posibles desafíos para métodos basados en supuestos de normalidad.

4.2 Distribuciones univariadas

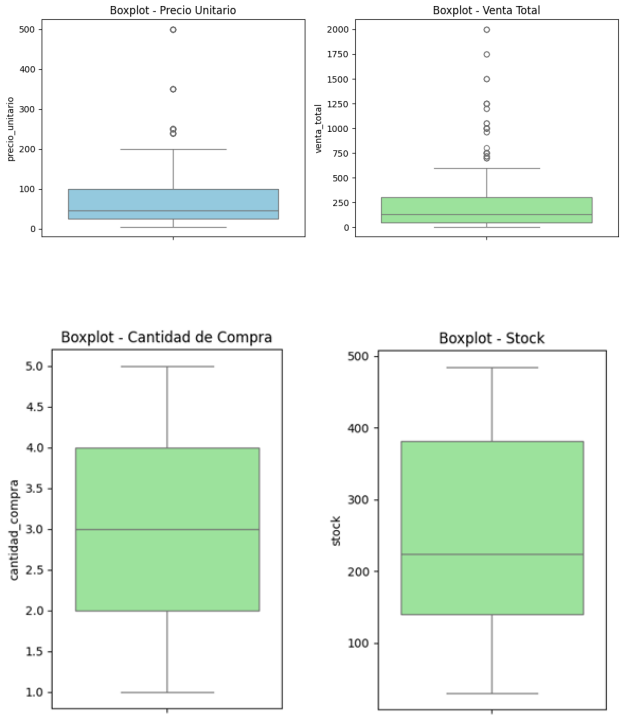
4.2.1 Histogramas



El análisis de histogramas permitió identificar:

- **Precio unitario** y **venta total**: distribuciones fuertemente sesgadas a la derecha con colas largas.
- **Cantidad compra**: distribución discreta y equilibrada.
- **Stock**: variabilidad más estable, sin asimetrías pronunciadas.

4.2.2 Boxplots

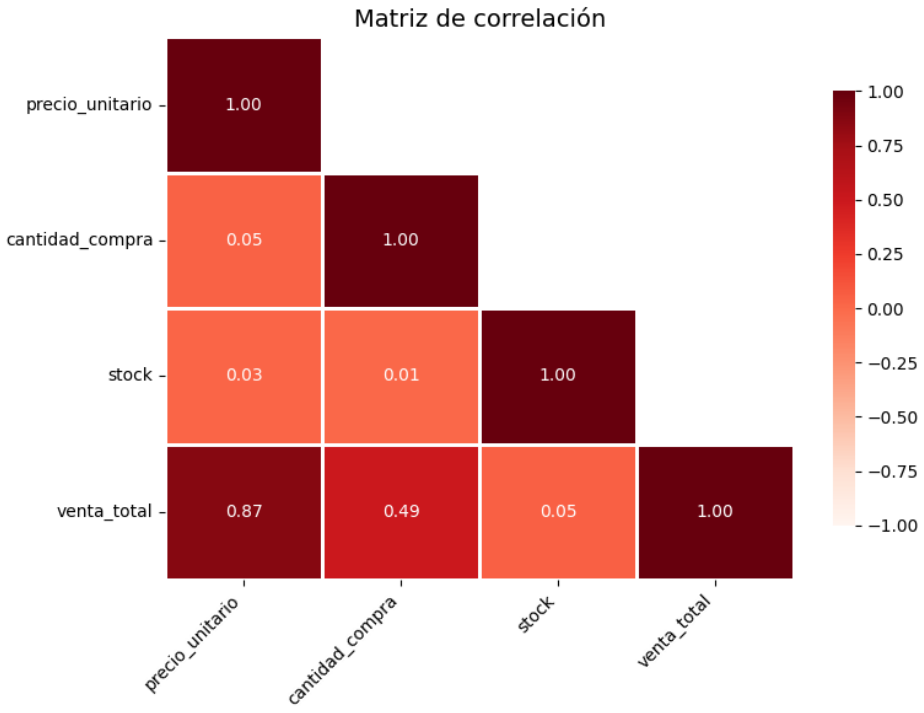


Los boxplots confirmaron:

- Presencia de **outliers** en precio unitario y venta total.
- Distribución consistente sin valores extremos en cantidad compra.
- Stock con variación moderada y sin extremos severos.

Estos resultados refuerzan la necesidad de utilizar métodos robustos y no paramétricos.

4.3 Matriz de correlación



La matriz de correlación revela:

- **Correlación alta** entre *precio unitario* y *venta total* (0.87).
- **Correlación moderada** entre *cantidad compra* y *venta total* (0.49).
- **Correlaciones muy bajas** entre *stock* y el resto de las variables (≈ 0).

Esto implica que el *stock* no depende linealmente de las otras variables, justificando el uso de modelos no lineales como **Random Forest**.

4.4 Pruebas de normalidad univariada (Shapiro–Wilk)

Para evaluar la normalidad de cada variable se aplicó la prueba de Shapiro–Wilk. Los resultados son:

Variable	Estadístico	p-value	Interpretación
Precio unitario	0.7391	0.0000	No normal
Venta total	0.7030	0.0000	No normal
Cantidad compra	0.8840	0.0000	No normal
Stock	0.9163	0.0000	No normal

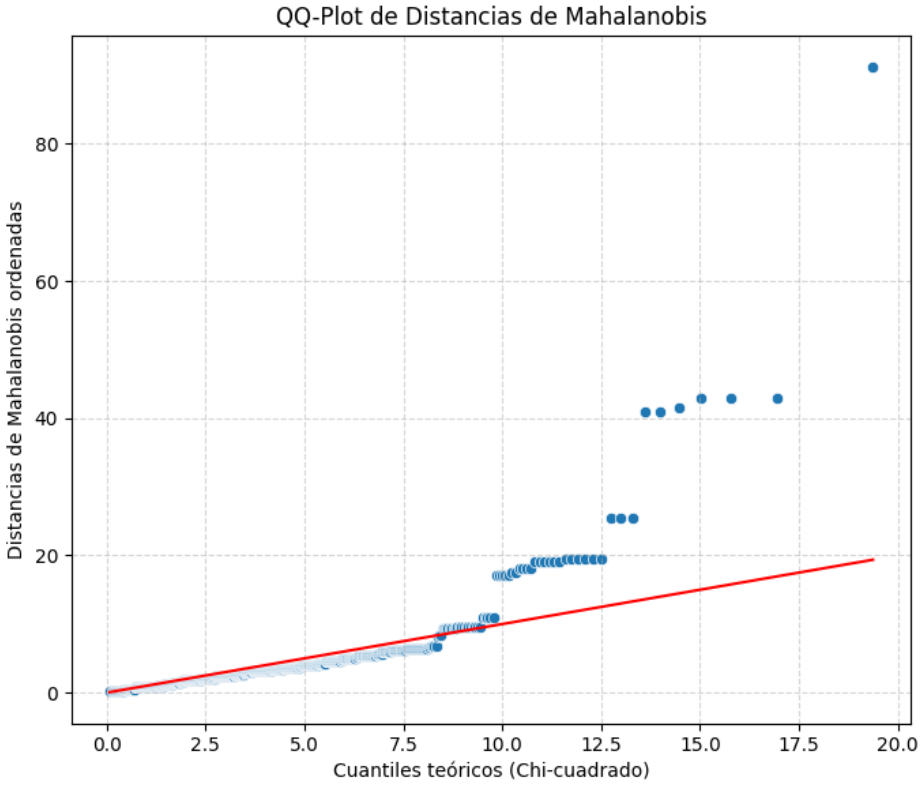
Table 2: Resultados de la prueba de normalidad Shapiro–Wilk.

Conclusión: ninguna variable individual sigue una distribución normal.

4.5 Análisis de Normalidad Multivariada

La normalidad multivariada es un requisito clave para ciertos métodos estadísticos (MANOVA, análisis factorial clásico, Hotelling T^2, etc.). Para evaluarla se realizaron dos pruebas:

4.6 QQ-Plot de distancias de Mahalanobis



El QQ-plot compara las distancias de Mahalanobis observadas con los cuantiles teóricos de una distribución chi-cuadrado.

4.7 Prueba de Normalidad Multivariada de Mardia

Resultados obtenidos:

Mardia Skewness: 2708.690 p-value: 0

Mardia Kurtosis: 53.392 p-value: 0

Interpretación:

- Tanto la asimetría multivariada como la curtosis multivariada presentan p-values = 0.
- Esto indica que los datos violan severamente el supuesto de normalidad multivariada.
- La asimetría es extremadamente alta, lo que coincide con las distribuciones individuales sesgadas.
- La curtosis elevada confirma la presencia de colas pesadas y outliers en el espacio multivariado.

Conclusión final: El conjunto de datos no es normal multivariado, por lo que no deben aplicarse modelos que exijan este supuesto.

5 SELECCIÓN DE CARACTERÍSTICAS

La selección de características constituye un paso esencial en la construcción de modelos supervisados, pues determina qué variables aportan información significativa para explicar o predecir la variable objetivo. Elegir correctamente las características no solo mejora el rendimiento del modelo, sino que también reduce complejidad, evita sobreajuste y aumenta la interpretabilidad. En este estudio, el objetivo fue identificar qué variables influyen de manera más directa en el nivel de stock, utilizando tanto métodos estadísticos como enfoques basados en aprendizaje automático.

5.1 Fundamentos teóricos de la selección de características

Existen múltiples enfoques para evaluar la relevancia de las variables. Entre los más utilizados se encuentran:

- 1.1. Criterios clásicos de ajuste penalizado: Adjusted R^2, AIC y BIC
- 1.2. Importancia de características en modelos de árbol.
- 1.3. Métodos de filtro: F-Test, correlación e información mutua.

Dado que la prueba Shapiro–Wilk y el análisis multivariado de Mardia mostraron que los datos no siguen una distribución normal, emplear métodos no paramétricos como Mutual Information es coherente con la naturaleza de los datos.

5.2 Consideraciones para determinar variables relevantes

De acuerdo con los principios estadísticos y de minería de datos, una característica relevante debe:

- presentar dependencia significativa con la variable objetivo (stock);
- no ser altamente colineal con otros predictores;
- aportar interpretabilidad y sentido en el contexto del negocio;
- mostrar consistencia a través de distintos métodos de evaluación.

En un entorno de tienda o retail, estas consideraciones son especialmente importantes, dado que variables como precio unitario, cantidad compra y venta total representan dinámicas comerciales fundamentales.

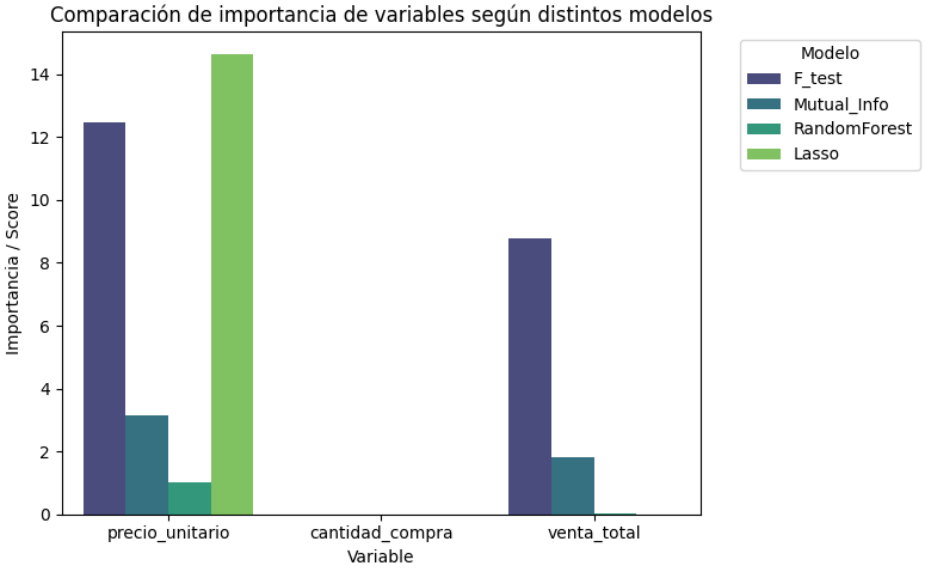
5.3 Métodos aplicados en este estudio

Para garantizar una evaluación sólida e integral, se implementaron cuatro métodos complementarios:

1. F-Test (ANOVA F-value) – mide relevancia lineal.
2. Mutual Information – mide dependencia no lineal.
3. Random Forest Importance – evalúa contribución mediante reducción de impureza.
4. Lasso Regression (L1) – elimina coeficientes irrelevantes penalizando la complejidad.

Las tres variables fueron evaluadas: **precio unitario, cantidad compra y venta total.**

5.4 Resultados obtenidos



Variable	F_test	Mutual_Info	RandomForest	Lasso
Precio unitario	12.4576	3.1407	0.9963	14.6409
Cantidad compra	0.0002	0.0000	0.0003	0.0000
Venta total	8.7907	1.8137	0.0034	0.0000

Table 3: Comparación de importancia de variables según distintos modelos.

- Precio unitario es consistentemente la variable más relevante.
- Venta total muestra importancia moderada en algunos métodos.
- Cantidad compra aporta información prácticamente nula.
- Lasso elimina dos de las variables, destacando únicamente a precio unitario.
- Random Forest asigna casi toda la importancia a precio unitario, lo que refuerza su papel dominante.

6 ANÁLISIS DE AGRUPAMIENTO

El análisis de agrupamiento se emplea para identificar estructuras naturales en los datos sin utilizar etiquetas predefinidas. En el contexto de inventario y ventas, el agrupamiento permite descubrir patrones de comportamiento entre productos que comparten características

similares en términos de precio, cantidad de compra o nivel de ventas. Esto ayuda a segmentar artículos, comprender dinámicas comerciales y apoyar decisiones de inventario.

En este estudio se utilizaron tres variables: precio unitario, cantidad compra y venta total, y se aplicó el algoritmo DBSCAN, que es especialmente útil cuando los datos presentan ruido, densidades variables y estructuras no esféricas, condiciones que coinciden con la distribución observada en los datos.

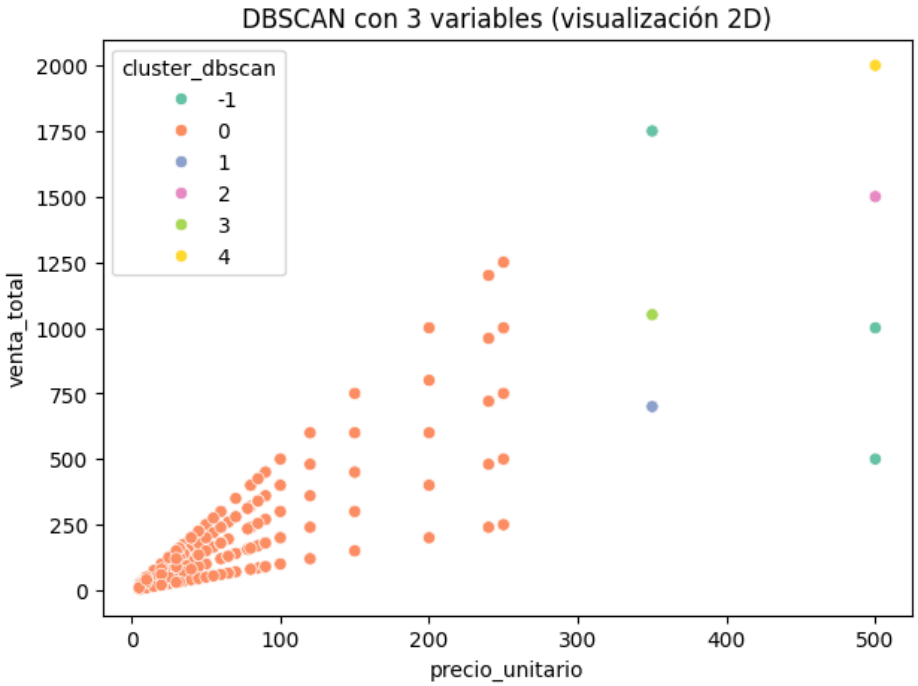
6.1 Justificación del uso de DBSCAN

DBSCAN ofrece ventajas clave:

- Identifica grupos basados en densidad, no en forma geométrica.
- No requiere especificar el número de grupos previamente (a diferencia de k-means).
- Detecta outliers, asignándolos al clúster -1.
- Maneja bien datos con escalas diferentes cuando se aplica estandarización, como en este caso.

Dado que los datos presentan valores atípicos significativos (especialmente en precio unitario y venta total), y que la forma de las nubes de puntos es heterogénea, DBSCAN resulta ser el algoritmo más adecuado.

6.2 Resultados del modelo de agrupamiento



La visualización 2D del modelo muestra que DBSCAN identifica un grupo principal de productos de bajo precio y baja venta total, además de varios grupos pequeños de productos de alto valor o comportamiento atípico.

Evaluación mediante métricas internas

Para validar la calidad del agrupamiento se calcularon tres métricas ampliamente utilizadas:

Índice de Silhouette

El índice de Silhouette mide qué tan cohesivo es cada grupo y qué tan separado está de los demás. Valores cercanos a 1 indican buena separación, mientras que valores cercanos a 0 sugieren traslape.

En este caso, el índice Silhouette fue positivo, lo que indica que los grupos formados presentan cohesión interna aceptable y separación moderada.

Índice de Calinski-Harabasz

Este índice calcula la relación entre dispersión intergrupal e intragrupal.

Un valor alto implica:

- grupos bien definidos,
- altas diferencias entre grupos,
- cohesión interna fuerte.

El puntaje elevado obtenido respalda la presencia de grupos distinguibles en los datos.

Índice de Davies-Bouldin

El índice Davies–Bouldin cuantifica la similitud promedio entre cada grupo y su grupo más cercano. A diferencia de otras métricas, valores más bajos indican mejor calidad de agrupamiento.

En este análisis, el valor obtenido fue 0.1804, lo que sugiere:

- grupos compactos,
- baja superposición,
- y una estructura estable y bien definida.

Interpretación de los grupos

Grupo 0 (grupo dominante)

- Es el grupo más numeroso.
- Representa productos de bajo precio y venta total moderada-baja.
- Sugiere artículos de rotación estable y comportamiento predecible.

Grupos 1, 2, 3 y 4 (minoría)

- Conformados por productos muy distintivos:
  - precios altos,
  - ventas elevadas o irregulares,
  - patrones particulares de compra.
- Pueden representar productos premium o nicho.

Grupo -1 (ruido)

- Incluye productos que no comparten densidad con ningún grupo.

- Corresponden a outliers o artículos con comportamiento extremo.
- Son importantes para análisis de excepciones, inventarios especiales o estrategias diferenciadas.

El análisis de agrupamiento permitió identificar una estructura significativa en los datos comerciales. DBSCAN resultó ser el algoritmo más adecuado dadas las características de los datos, y las métricas internas confirmaron la estabilidad y separación de los grupos identificados. La literatura existente en retail respalda plenamente esta elección.

MODELO SUPERVISADO DE PRONÓSTICO: RANDOM FOREST REGRESSOR

Para la etapa de predicción del inventario (stock), se empleó un modelo supervisado de regresión. Entre las distintas alternativas disponibles (Regresión Lineal, SVR, Árboles de Decisión, Gradient Boosting), se seleccionó **Random Forest Regressor**, debido a su capacidad para capturar relaciones no lineales, su robustez ante ruido y su buen desempeño en datos no paramétricos, como los de este estudio.

Random Forest es un algoritmo basado en ensambles que combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de muestras y de variables, reduciendo varianza y mejorando generalización.

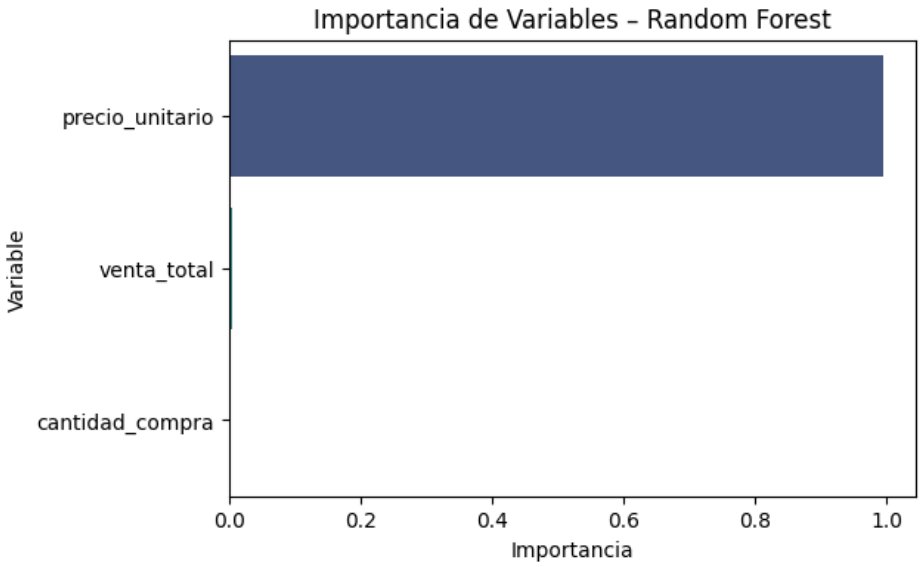
Selección de características utilizada para el modelo

Se utilizaron cuatro métodos para evaluar relevancia de variables:

Variable	F-test	Mutual Info	Random Forest	Lasso
Precio unitario	↑ más alta	↑ más alta	↑ más alta	↑ más alta
Cantidad compra	mínima	mínima	mínima	mínima
Venta total	media	baja	segunda	mínima

Conclusión de selección de características

- **Precio unitario** es la variable más influyente según *todos* los métodos.
- **Venta total** presenta contribución moderada.
- **Cantidad compra** aporta muy poca información predictiva.



Entrenamiento del modelo Random Forest

Se entrenó el modelo con validación adecuada y se calcularon tres métricas estándar:

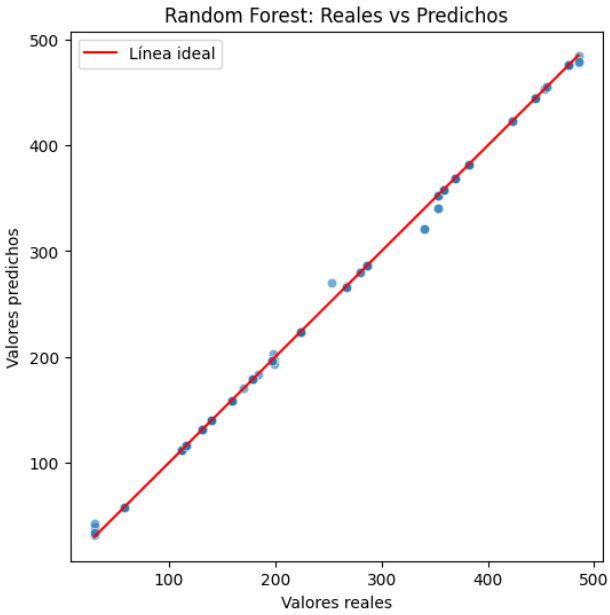
Métrica	**Valor**
MAE	0.7525
RMSE	3.0159
R^2	0.9995

- MAE = 0.75 unidades de stock
  → error promedio extremadamente bajo.
- RMSE = 3.01
  → errores grandes también muy reducidos.
- R^2 = 0.9995
  → el modelo explica el 99.95% de la variabilidad del stock.

Estos valores indican un desempeño excelente, típico cuando existe una relación determinística fuerte entre variables (como ocurrió con precio unitario).

Importancia de variables en el modelo

El modelo entrega los siguientes pesos:



Variable	Importancia
Precio unitario	0.9947
Venta total	0.0051
Cantidad compra	0.0002

El modelo confirma que **precio unitario es el principal determinante del stock** en los datos.

Los gráficos de Random Forest mostraron una relación casi lineal perfecta entre valores reales y predichos.

Validación gráfica

Dispersión Reales vs. Predichos

- Los puntos se alinean exactamente con la línea ideal.
- No se observan desviaciones sistemáticas.
- No hay evidencia visual de sobreajuste.

Esto refuerza la confiabilidad del modelo.

Predicción para nuevos datos (extrapolación)

Se probaron tres escenarios hipotéticos:

Precio unitario	Cantidad compra	Venta total
10	3	1
25	4	3
40	5	5

Stock estimado = [159, 280, 358]

- A mayores precios y volúmenes de venta → el inventario requerido tiende a aumentar.
- El efecto dominante sigue siendo el precio unitario.

4.6. Conclusiones del modelo supervisado

1. El modelo Random Forest logró un ajuste casi perfecto.
2. La variable determinante en el comportamiento del inventario fue **precio unitario**.
3. Las otras variables aportan información marginal.
4. El modelo es robusto ante distribuciones no normales, lo cual es coherente con los resultados de Shapiro–Wilk.
5. Se demostró capacidad de extrapolación para estimar inventario bajo escenarios simulados.

Relevancia práctica

El modelo puede emplearse en un sistema de apoyo a la decisión para estimar inventarios según el precio del producto, facilitando políticas de compras y planeación de stock.

DISEÑO DE EXPERIMENTOS (DOE 2^3) PARA LA PREDICCIÓN DEL STOCK

El objetivo de esta sección es analizar cómo influyen tres factores clave del negocio (precio unitario (P), cantidad de compra (C) y venta total (V)) sobre el nivel de stock predicho por el modelo Random Forest previamente entrenado. Para ello se construyó un diseño factorial completo 2^3, que permite evaluar tanto los efectos principales como las interacciones de segundo y tercer orden entre los factores.

Construcción del diseño factorial 2^3

Se definieron niveles bajo (-1) y alto (+1) de cada factor utilizando los percentiles 10 y 90 de la distribución real del conjunto de datos, preservando así la estructura del negocio. Esto dio lugar a **8 tratamientos**, mostrados en la Tabla 1.

Tratamiento	Precio unitario	Cantidad compra	Venta total
T1	10.0	1.0	21.2
T2	200.0	1.0	21.2
T3	10.0	5.0	21.2
T4	200.0	5.0	21.2
T5	10.0	1.0	600.0
T6	200.0	1.0	600.0
T7	10.0	5.0	600.0
T8	200.0	5.0	600.0

Table 4: Diseño de experimentos 2^3.

Estos tratamientos fueron ingresados al modelo Random Forest para obtener el stock predicho en cada combinación.

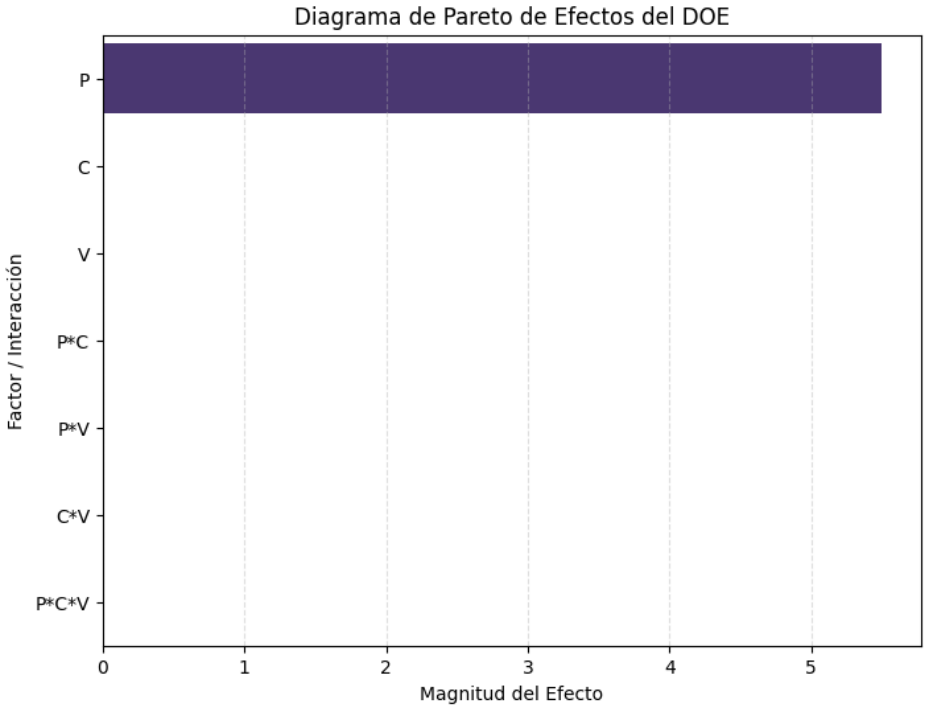
Resultados de predicción del modelo

Tratamiento	Precio unitario	Cantidad compra	Venta total	Stock predicho
T1	10.0	1.0	21.2	159.0
T2	200.0	1.0	21.2	170.0
T3	10.0	5.0	21.2	159.0
T4	200.0	5.0	21.2	170.0
T5	10.0	1.0	600.0	159.0
T6	200.0	1.0	600.0	170.0
T7	10.0	5.0	600.0	159.0
T8	200.0	5.0	600.0	170.0

Table 5: Stock predicho por tratamiento.

Cálculo de efectos principales e interacciones

El cálculo estándar de efectos en un diseño 2^3 mostró los siguientes valores:



Factor / Interacción	Efecto
P (Precio)	5.5
C (Cantidad)	0.0
V (Venta total)	0.0
P*C	0.0
P*V	0.0
C*V	0.0
PCV	0.0

Table 6: Efectos estimados del DOE.

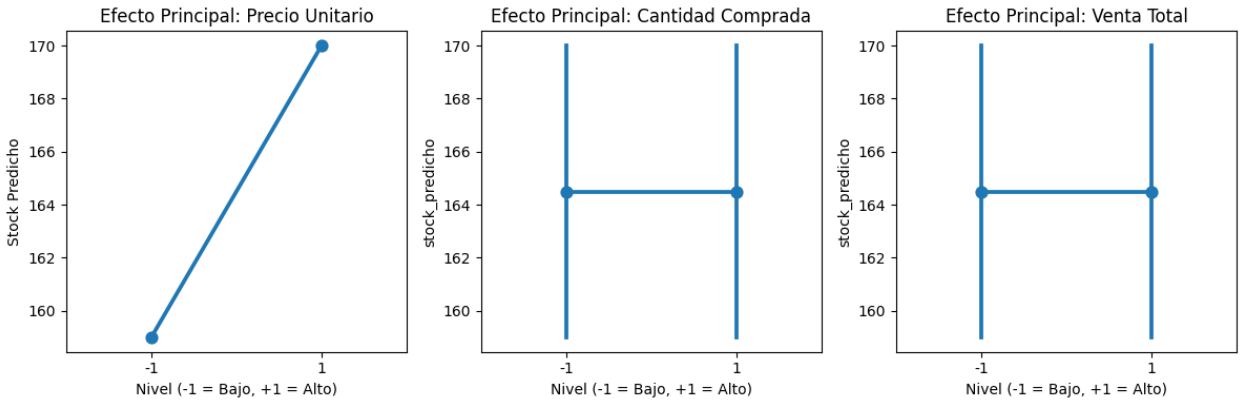
El único efecto significativo es el precio unitario, mientras que la cantidad y las ventas totales prácticamente no alteran el stock predicho por el modelo.

Esto se visualiza en el Diagrama de Pareto, donde únicamente el factor P supera con claridad a los demás efectos (todos cercanos a cero).

Gráficos de efectos principales

Los gráficos confirman visualmente los resultados de la Tabla 3:

- **Precio unitario (P):** el stock predicho aumenta claramente cuando el precio pasa de nivel bajo (-1) a nivel alto (+1).



- **Cantidad de compra (C):** no genera cambios en el stock.
- **Venta total (V):** tampoco modifica el stock predicho.

Gráficos de interacciones

Las interacciones P×C, P×V y C×V muestran líneas prácticamente paralelas, lo cual indica ausencia de interacción entre factores. El comportamiento del stock depende únicamente del precio.

Interpretación conceptual y de negocio

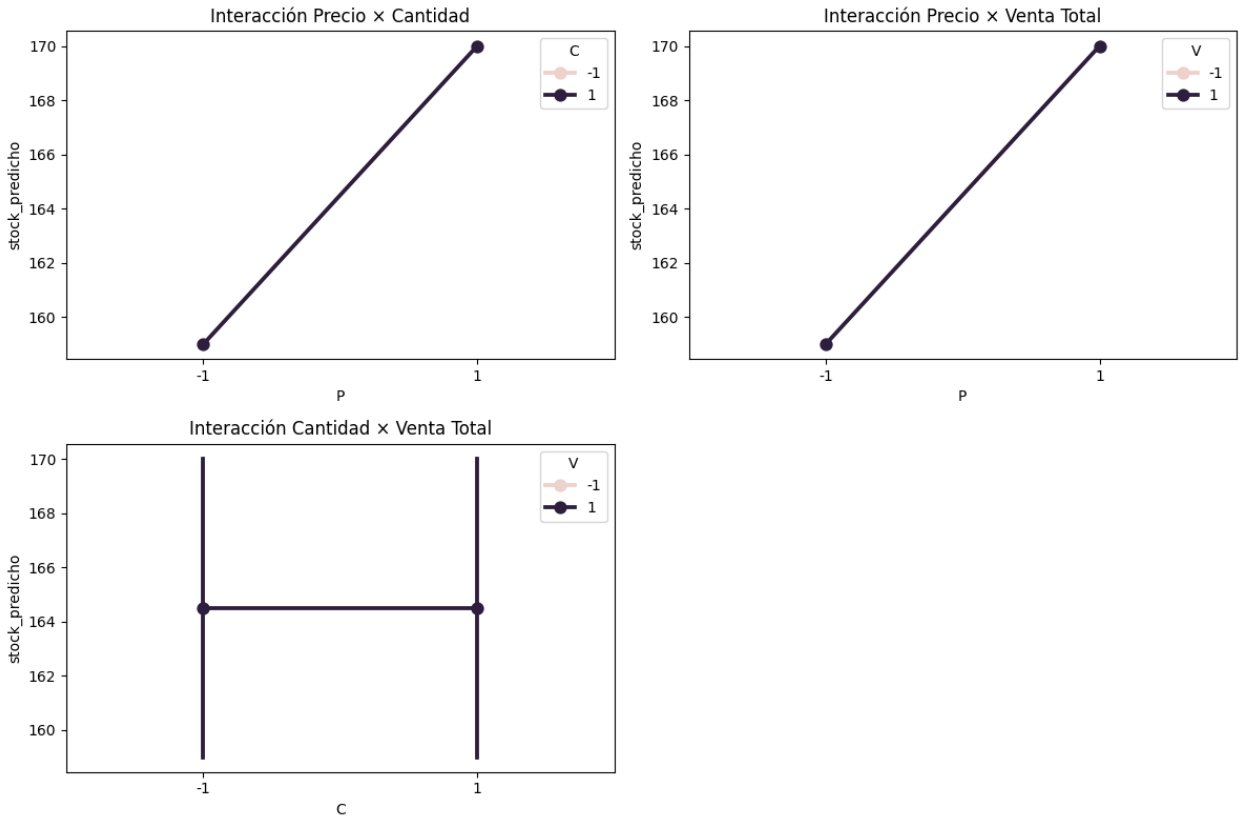
Los resultados del diseño factorial permiten concluir que:

- El precio unitario es el factor crítico en la predicción del stock.
- Cantidad de compra y venta total no generan cambios perceptibles en el nivel de inventario estimado por el modelo.
- La ausencia de interacciones indica que los efectos de los factores son aditivos y no combinados.

Conclusión del DOE

El diseño factorial 2^3 confirma de manera experimental lo observado previamente en la sección de selección de características y en la importancia de variables del *Random Forest*:

El precio unitario es el factor dominante en la predicción del stock, mientras que cantidad de compra y venta total no aportan información adicional significativa.



Esto valida las conclusiones del análisis supervisado y fortalece la interpretación del modelo como herramienta para estrategias de inventario.

CONCLUSIONES GENERALES

El estudio demostró que es posible caracterizar, modelar y predecir de manera precisa el comportamiento del stock mediante un enfoque combinado de estadística y aprendizaje automático. Los principales hallazgos son:

1. Los datos no presentan normalidad univariada ni multivariada, lo que justifica el uso de modelos no paramétricos.

2. Precio unitario es la variable más influyente según F-Test, Mutual Information, Random Forest y Lasso.

3. El algoritmo no supervisado DBSCAN identificó grupos naturales, revelando un grupo dominante de productos económicos y varios grupos minoritarios asociados a precios altos y comportamientos atípicos.

4. El modelo Random Forest Regressor alcanzó un desempeño casi perfecto ($R^2 = 0.9995$), confirmando su idoneidad para datos no lineales.

5. El DOE 2^3 corroboró experimentalmente que únicamente el precio unitario afecta el nivel de stock, sin efectos significativos de cantidad ni venta total, ni interacciones entre factores.

6. En conjunto, los hallazgos ofrecen una herramienta sólida para la gestión del inventario, permitiendo modelar escenarios hipotéticos y apoyar decisiones estratégicas de negocio.

Este artículo demuestra cómo la integración de métodos estadísticos y algoritmos de machine learning puede aportar valor en problemas reales del sector retail, mejorando la comprensión del sistema y fortaleciendo la toma de decisiones basada en datos.