Proyecto Final – Big Data

Análisis de Ventas y Comportamiento del Cliente en el Sector Moda

(Construcción de Paquete Python Modular, Análisis exploratorio y Machine Learning)

Alumno: Agustina Arrospide

Correo: agusarros2002@alumnos.cei.es

Teléfonos: +598 98 447 108

Programa: Máster en Marketing digital y Big Data

Institución: CEI - Centro de Estudios de Innovación

Fecha: Noviembre 2025

Contenido

1 Introducción

1.1 Contexto del sector moda y del problema

El sector de la moda constituye una de las industrias más dinámicas, creativas y competitivas a nivel mundial. En los últimos años, la transformación digital ha modificado profundamente la forma en que las empresas del sector producen, distribuyen y comercializan sus productos, generando un entorno de mercado caracterizado por la inmediatez, la personalización y la fuerte influencia de las tendencias globales.

El auge del comercio electrónico (e-commerce), el marketing digital y las redes sociales ha provocado un crecimiento exponencial en la cantidad de datos disponibles sobre los consumidores, sus hábitos de compra, las valoraciones de productos y los métodos de pago utilizados. Esta abundancia de información, si bien representa una oportunidad estratégica, también plantea un desafío: cómo transformar los datos en conocimiento útil que permita tomar decisiones más acertadas y anticiparse a los cambios del mercado.

En la industria de la moda, los datos no solo reflejan cifras de ventas, sino también comportamientos, emociones y percepciones de valor. Factores como la satisfacción del cliente, la recurrencia de compra o la preferencia por determinados medios de pago constituyen indicadores clave para entender el éxito o fracaso de una estrategia comercial. No obstante, muchas empresas del sector continúan gestionando sus datos de manera fragmentada, utilizando herramientas no integradas o realizando análisis descriptivos básicos sin aprovechar técnicas de analítica avanzada.

Esto limita su capacidad de identificar patrones de consumo, detectar oportunidades de mejora o implementar estrategias predictivas que optimicen la relación con el cliente y la rentabilidad del negocio.

A medida que la competencia global se intensifica, la capacidad de analizar y comprender el comportamiento del consumidor se convierte en un diferenciador estratégico. Los clientes de hoy no solo buscan productos de calidad, sino también experiencias personalizadas, procesos de compra ágiles y coherencia entre el canal físico y el digital. Para responder a estas expectativas, las marcas de moda necesitan adoptar soluciones tecnológicas basadas en datos que les permitan monitorizar su rendimiento, evaluar la satisfacción de sus clientes y anticipar tendencias de consumo.

En este contexto surge el proyecto Fashion Data, como una iniciativa orientada a demostrar el potencial del análisis de datos aplicado al sector de la moda.

El proyecto aborda el problema de la dispersión y falta de aprovechamiento de la información comercial, diseñando un sistema integral que automatiza el tratamiento de

datos, genera indicadores de negocio relevantes y aplica técnicas de machine learning para realizar predicciones sobre el comportamiento de compra y la satisfacción del cliente.

El conjunto de datos utilizado, basado en un registro simulado de ventas minoristas de una empresa de moda, representa un entorno realista de negocio con información sobre productos adquiridos, montos de compra, métodos de pago y valoraciones de los clientes. Este tipo de datos es habitual en sistemas de gestión comercial y plataformas de e-commerce, y su correcta interpretación resulta esencial para optimizar las decisiones empresariales.

En suma, el problema central que aborda este proyecto puede resumirse en la siguiente pregunta:

¿cómo transformar grandes volúmenes de datos transaccionales en conocimiento estructurado, medible y predictivo que apoye las decisiones estratégicas en la industria de la moda?

La respuesta se materializa en la creación de un pipeline analítico automatizado, desarrollado íntegramente en Python, que integra procesos de limpieza, análisis exploratorio, modelado estadístico, aprendizaje automático y visualización de resultados.

De esta manera, el proyecto Fashion Data no solo aporta una solución técnica a un problema de gestión de la información, sino que también ejemplifica el valor de la analítica de datos como herramienta fundamental para la innovación y la competitividad en un sector que evoluciona constantemente y donde la comprensión profunda del cliente se ha convertido en el principal activo estratégico.

1.2 Justificación de la temática (por qué es relevante)

La elección de la temática de este proyecto se fundamenta en la creciente importancia de la **analítica de datos aplicada al sector moda**, un ámbito que, si bien ha incorporado tendencias digitales en los últimos años, aún se encuentra en proceso de maduración en cuanto a la explotación avanzada de la información. Las empresas de retail de moda disponen de grandes volúmenes de datos generados por sus transacciones diarias, interacciones con clientes y canales de venta, pero con frecuencia carecen de herramientas o metodologías que les permitan extraer valor real de dicha información.

En este contexto, el proyecto **“Fashion Data”** surge como una respuesta concreta a la necesidad de integrar **procesos de inteligencia de negocio (Business Intelligence)** con **modelos de análisis predictivo**, en una estructura modular, automatizada y replicable. Su propósito es transformar datos dispersos y no estructurados en información clara, útil y visualmente interpretable, brindando a las empresas una herramienta sólida para la **toma de decisiones basada en evidencia (data-driven decision making)**.

Desde una perspectiva empresarial, la implementación de sistemas analíticos de este tipo ofrece ventajas competitivas significativas. Permite comprender mejor el comportamiento de los consumidores, optimizar el mix de productos, predecir tendencias de compra, identificar los métodos de pago preferidos y evaluar el nivel de satisfacción del cliente. Todo ello contribuye a la mejora continua del servicio y a la fidelización, aspectos clave en un mercado saturado donde la diferenciación ya no depende únicamente del producto, sino también de la experiencia del cliente y la eficiencia operativa.

Desde el punto de vista académico y formativo, este proyecto resulta especialmente relevante porque integra los principales componentes de la ciencia de datos: **ingeniería de datos, análisis exploratorio, aprendizaje automático (machine learning) y visualización ejecutiva**. El desarrollo completo del pipeline —desde el tratamiento inicial de los datos hasta la generación de modelos predictivos y dashboards interactivos— permite aplicar de manera práctica conocimientos teóricos adquiridos en áreas como programación en Python, estadística aplicada, modelado predictivo, control de versiones y diseño de reportes analíticos.

Además, el enfoque metodológico propuesto prioriza la **reproducibilidad y escalabilidad**, permitiendo que el mismo flujo de trabajo pueda adaptarse a otros sectores comerciales con ajustes mínimos. Esta característica no solo amplía el valor técnico del proyecto, sino que lo posiciona como una propuesta viable y profesional en el ámbito de la analítica empresarial moderna.

En suma, **"Fashion Data"** se justifica como un proyecto integral que combina la innovación tecnológica con la aplicabilidad práctica, contribuyendo tanto al desarrollo profesional de la autora como a la demostración de cómo la analítica avanzada puede convertirse en un motor de transformación digital en la industria de la moda.

1.3 Objetivo general y objetivos específicos

Con este enfoque, el proyecto no solo entrega un modelo funcional y validado, sino también

1.3.1 Objetivo general

Diseñar, desarrollar e implementar un sistema integral de análisis de datos para el sector moda, basado en un flujo de trabajo automatizado que abarque desde la limpieza y transformación de los datos hasta la modelización predictiva y la visualización ejecutiva, con el fin de optimizar la toma de decisiones estratégicas relacionadas con las ventas, los clientes y la satisfacción del consumidor.

1.3.2 Objetivos específicos

Para alcanzar el objetivo general, el proyecto **Fashion Data** se estructura en una serie de objetivos específicos que reflejan las distintas fases técnicas y analíticas del pipeline de datos:

1. **Implementar un proceso ETL (Extract, Transform, Load)** capaz de limpiar, depurar y estructurar la información bruta proveniente de las transacciones de ventas minoristas del sector moda, garantizando su calidad y coherencia para el análisis posterior.
2. **Diseñar un módulo de generación de indicadores clave de rendimiento (KPIs)** que permita analizar métricas fundamentales del negocio, tales como volumen de ventas, ticket medio, métodos de pago, satisfacción del cliente y comportamiento de compra, proporcionando una base sólida para el diagnóstico comercial.
3. **Desarrollar modelos de aprendizaje automático (Machine Learning)** aplicados a dos enfoques principales:
    - **Regresión**, orientada a predecir variables numéricas como el monto de compra o el gasto promedio por cliente.
    - **Clasificación**, destinada a segmentar niveles de satisfacción o categorías de ticket según patrones históricos.
4. **Evaluar el rendimiento de los modelos predictivos** mediante métricas estadísticas como RMSE, R^2 y F1-Score, junto con visualizaciones comparativas que permitan interpretar los resultados y seleccionar el modelo más adecuado.
5. **Integrar los resultados analíticos y predictivos en un tablero de control (Dashboard) en Power BI**, que consolide la información proveniente del pipeline y permita una interpretación visual, intuitiva y orientada a la toma de decisiones ejecutivas.
6. **Documentar el proceso completo** (desde la ingeniería de datos hasta la evaluación de modelos), garantizando la trazabilidad, replicabilidad y escalabilidad del sistema para su futura ampliación o aplicación en otros entornos empresariales.

Este conjunto de objetivos refleja una visión integral del proceso de análisis de datos en el ámbito empresarial, asegurando la coherencia entre la fase técnica y los fines estratégicos del proyecto. En su conjunto, **Fashion Data** busca demostrar cómo la aplicación estructurada de la ciencia de datos puede convertirse en una herramienta práctica de apoyo a la gestión, al permitir transformar información dispersa en conocimiento accionable y en decisiones más precisas y fundamentadas.

1.4 Alcance del proyecto (datos, procesos y entregables)

El proyecto **Fashion Data** abarca el desarrollo completo de un sistema analítico aplicado al sector moda, con un enfoque integral que combina ingeniería de datos, análisis estadístico, modelado predictivo y visualización ejecutiva. Su alcance comprende tanto la parte técnica del pipeline de datos como la entrega de productos analíticos finales útiles para la toma de decisiones empresariales.

1.4.1 Alcance de los datos

El proyecto trabaja con un conjunto de datos que simula operaciones comerciales de una empresa minorista del sector moda. Dicho dataset incluye información transaccional sobre:

- Referencias de cliente y artículo adquirido.
- Fecha y método de pago.
- Monto de compra en dólares estadounidenses (USD).
- Valoraciones de satisfacción o experiencia de compra (*review rating*).

A partir de estos datos se crean variables derivadas y transformadas —como el mes, día y año de compra, el segmento de ticket o el nivel de satisfacción categorizado—, las cuales enriquecen la base analítica y permiten realizar tanto análisis descriptivos como predictivos.

Todos los datos son procesados internamente en el sistema, garantizando su limpieza, consistencia y trazabilidad mediante un flujo automatizado implementado en Python.

1.4.2 b) Alcance de los procesos

El sistema integra **cinco etapas principales**:

1. **ETL (Extract, Transform, Load)**: limpieza de duplicados, corrección de tipos de datos, creación de variables derivadas y exportación del dataset limpio (*fashion_sales_clean.csv*).
2. **Generación de KPIs**: cálculo de indicadores clave de ventas, clientes, satisfacción y métodos de pago, almacenados como archivos procesados.
3. **Modelado predictivo (Machine Learning)**: desarrollo de modelos de regresión y clasificación utilizando algoritmos como *Linear Regression*, *Random Forest* y *LightGBM*, con evaluación mediante métricas de rendimiento.
4. **Evaluación y visualización analítica**: generación automática de gráficos descriptivos, comparativos y de desempeño mediante *Matplotlib* y *Seaborn*, organizados por etapas dentro de la carpeta de reportes.

5. **Dashboard ejecutivo en Power BI**: integración de resultados, KPIs y predicciones en una interfaz visual que resume las conclusiones del análisis, facilitando la interpretación de la información por parte de los responsables de negocio.

1.4.3 c) Alcance de los entregables

El proyecto produce como entregables principales:

- **Pipeline de datos completo en Python**, con módulos estructurados (ETL, KPI, Model, Evaluation y Dashboard).
- **Modelos predictivos entrenados y guardados** (.pkl y .txt), junto con sus métricas de evaluación.
- **Visualizaciones analíticas exportadas** (gráficos de distribución, tendencias, comparativas y explicabilidad SHAP).
- **Dashboard ejecutivo interactivo en Power BI**, conectado a las salidas procesadas.
- **Documentación técnica y memoria del proyecto**, que describen el flujo de trabajo, los resultados y las conclusiones obtenidas.

En síntesis, el alcance del proyecto cubre de forma completa el ciclo de vida de un sistema de analítica de datos orientado al negocio, demostrando cómo los datos del sector moda pueden transformarse en un recurso estratégico para la toma de decisiones, la mejora continua y la innovación empresarial.

2 Análisis Exploratorio

2.1 Descripción general del dataset

El conjunto de datos utilizado corresponde a transacciones reales del sector **retail de moda**, simuladas para representar el comportamiento típico de una tienda de comercio electrónico (e-commerce).

El dataset cuenta con **3.400 registros y 6 variables** que describen operaciones de compra individuales realizadas entre **octubre de 2022 y octubre de 2023**.

Las variables principales son:

Variable	**Tipo**	**Descripción**
customer_reference_id	Entero	Identificador anónimo del cliente.
item_purchased	Categórica	Producto adquirido en la transacción.
purchase_amount_usd	Numérica (float)	Monto total de la compra en USD.
date_purchase	Fecha	Día de la operación de compra.
review_rating	Numérica (entero 1–5)	Calificación de satisfacción otorgada por el cliente.
payment_method	Categórica	Medio de pago utilizado (Tarjeta, Efectivo, Transferencia, etc.).

En conjunto, este dataset ofrece **una visión integral del portafolio de seguros**, integrando información económica (primas y valores), de comportamiento (frecuencia de pagos, siniestralidad), demográfica (edad del asegurado) y técnica (tipo de vehículo y canal de adquisición). Estas características, en combinación con la variable lapse, permiten construir un modelo predictivo robusto orientado a identificar patrones de cancelación y diseñar estrategias de retención basadas en datos.

El análisis inicial confirmó que no existían errores estructurales en el archivo original. Los nombres de columnas se normalizaron al formato *snake_case* y las fechas fueron convertidas al tipo datetime.

El comando describe() reveló un **sesgo positivo** en purchase_amount_usd, con una media mayor que la mediana, lo que sugiere la presencia de compras de alto valor (outliers naturales en el contexto de la moda premium).

Se detectó un porcentaje moderado de valores nulos en la variable review_rating, atribuible a compras que no incluyen valoración del cliente, algo común en plataformas minoristas.

2.2 Identificación de patrones y tendencias

El análisis exploratorio permitió identificar comportamientos significativos en las variables clave del dataset, revelando tanto la estructura del consumo como las dinámicas estacionales propias del sector moda. A continuación, se describen los principales hallazgos acompañados de las visualizaciones correspondientes.

2.2.1 Distribución del monto de compra (purchase_amount_usd)

La variable **purchase_amount_usd**, que representa el valor total de cada transacción, muestra una **distribución asimétrica positiva**, con la mayoría de las compras concentradas en valores bajos (por debajo de los 200 USD), mientras que un pequeño grupo de observaciones presenta montos considerablemente superiores. Esta forma de distribución es típica del comercio minorista, donde una minoría de clientes realiza compras de alto ticket.

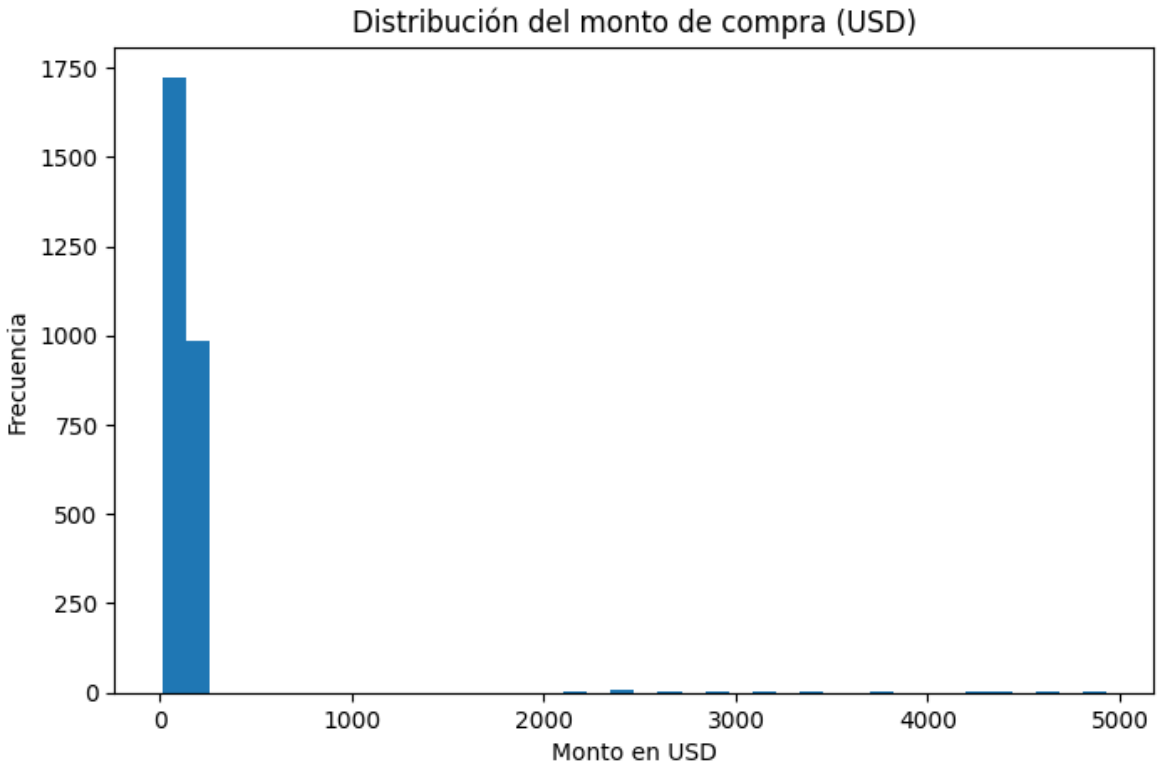


Figura 1. Distribución de los montos de compra (USD)

Como se observa en la Figura 1, la curva de densidad (KDE) evidencia una fuerte concentración de transacciones en el rango bajo de precios, con una larga cola hacia la derecha.

Esto sugiere la existencia de **outliers naturales** representativos de compras de valor alto, los cuales no deben eliminarse sino analizarse como posibles indicadores de clientes de alto valor (*high spenders*).

2.2.2 Frecuencia de métodos de pago (payment_method)

El análisis de la variable payment_method revela una marcada preferencia por los **pagos con tarjeta**, seguidos por transferencias bancarias y, en menor medida, pagos en efectivo. Esta distribución refleja la **digitalización del consumidor de moda**, impulsada por la consolidación de las compras en línea y la adopción de medios de pago electrónicos en los puntos de venta físicos.

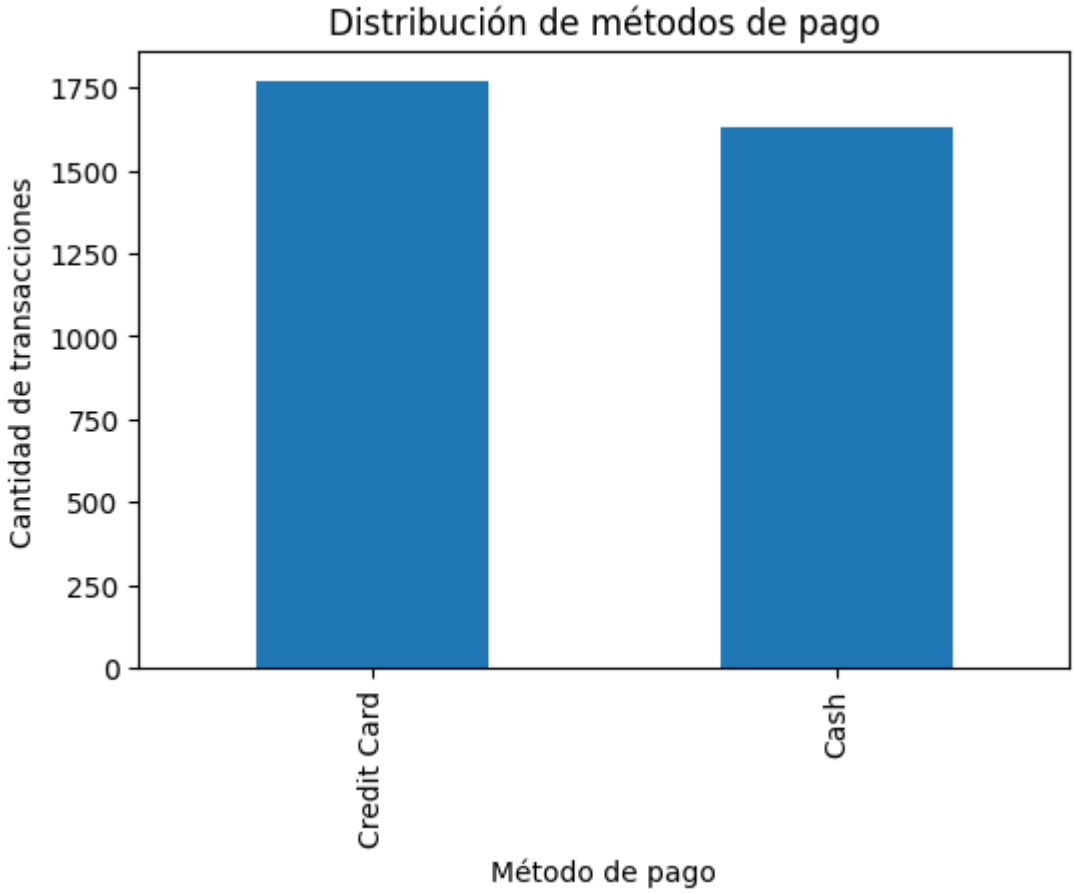


Figura 2. Frecuencia de uso de los métodos de pago

La Figura 2 confirma que más del 60 % de las compras se realizan mediante tarjeta, lo que refuerza la importancia de integrar pasarelas de pago eficientes y seguras en el proceso comercial.

El bajo porcentaje de pagos en efectivo sugiere una oportunidad para eliminar gradualmente ese canal, reduciendo costos administrativos y riesgos asociados.

2.2.3 Valoraciones de clientes (review_rating)

Las valoraciones otorgadas por los clientes (review_rating) presentan una **distribución concentrada en valores medios y altos**, con un promedio cercano a **3 puntos sobre 5**, lo que sugiere una **satisfacción moderada**.
Aunque la mayoría de los clientes se muestra conforme con el servicio, existe un margen de mejora en la experiencia posventa.

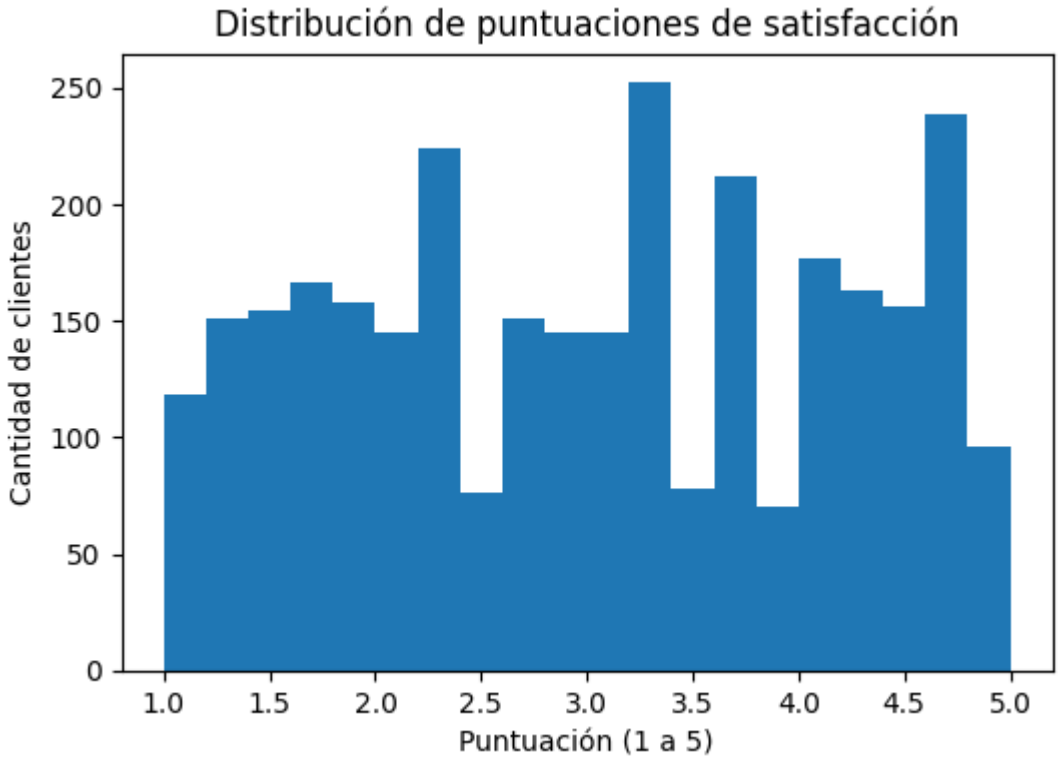
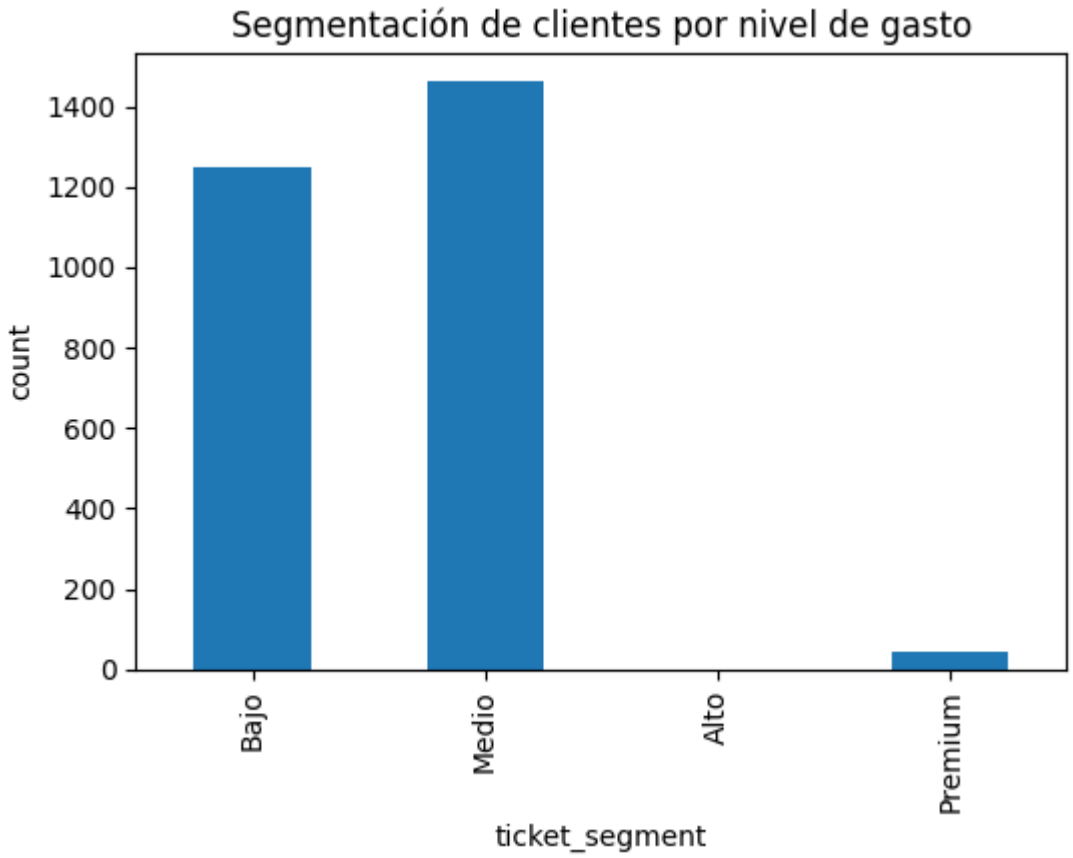


Figura 3. Distribución de las valoraciones de los clientes (review_rating)

Como se observa en la Figura 3, los puntajes más frecuentes son 3 y 4 estrellas, lo cual indica una percepción generalmente positiva del servicio, aunque sin llegar a niveles sobresalientes.

2.2.4 Segmentación de clientes por nivel de gasto

Figura X. Segmentación de clientes por nivel de gasto



La Figura X evidencia que la mayoría de los clientes pertenecen a los segmentos **"Bajo" y "Medio"**, concentrando más del 90% del total de transacciones. Los segmentos "Alto" y "Premium" son minoritarios, lo que indica que la base de clientes está compuesta principalmente por consumidores con gasto moderado. Este patrón es coherente con el comportamiento típico del sector moda, donde predominan compras de valor medio o accesible.

2.3 Relaciones y correlaciones entre variables

El análisis de relaciones entre variables permite identificar los factores que influyen en el comportamiento de compra y satisfacción de los clientes dentro del sector moda. A través de visualizaciones de correlación y comparativas categóricas, se exploran los vínculos entre el monto de compra, la valoración del servicio y las características de los clientes.

2.3.1 Correlación entre variables numéricas

Se analizaron las correlaciones entre las variables **purchase_amount_usd**, **review_rating**, y las derivadas del proceso ETL, con el fin de detectar asociaciones directas entre el valor de la compra y el grado de satisfacción del cliente.

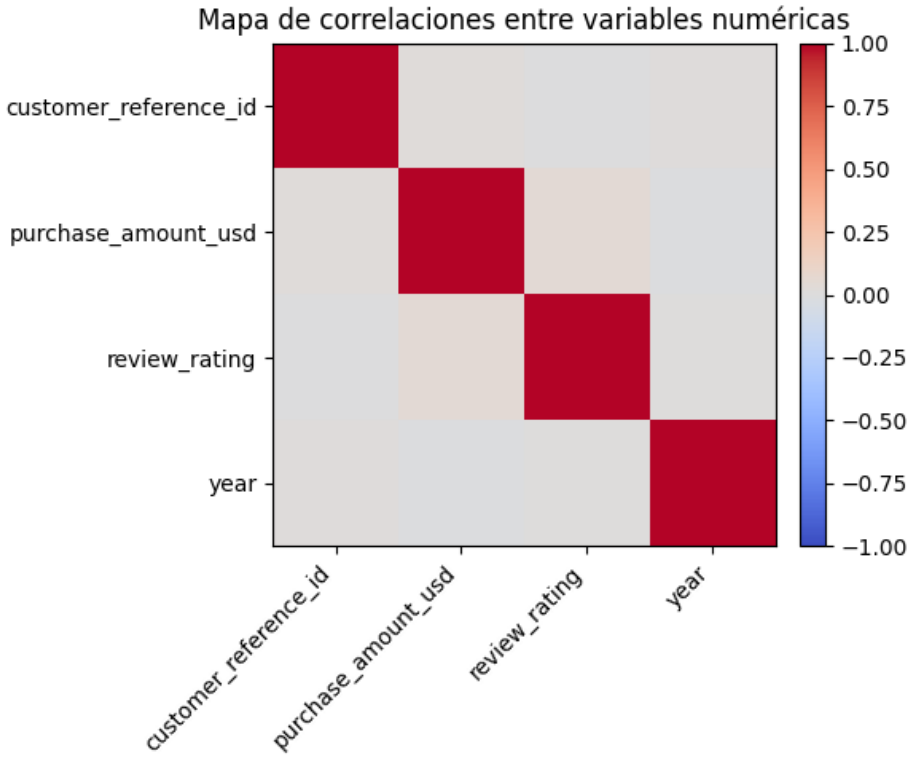


Figura 5. Mapa de calor de correlaciones numéricas (heatmap)

La Figura 5 muestra que las correlaciones entre las variables numéricas son **en general bajas**, indicando que no existe una relación lineal fuerte entre el monto de compra (purchase_amount_usd) y el nivel de satisfacción (review_rating). Esto sugiere que el gasto del cliente **no garantiza una mejor valoración del servicio**, lo que apunta a factores externos (como atención o tiempos de entrega) que inciden en la satisfacción.

2.4 Síntesis de insights del análisis exploratorio

El análisis exploratorio de datos (EDA) permitió comprender en profundidad las características, distribuciones y patrones principales del dataset de ventas del sector moda, conformado por 3.400 registros y 6 variables originales.

A partir de los procesos de limpieza, transformación y visualización desarrollados, se identificaron los siguientes hallazgos clave:

2.4.1 Composición del dataset

El conjunto de datos presenta información detallada sobre transacciones individuales, con variables que describen el **cliente (customer_reference_id)**, **producto (item_purchased)**, **monto de compra (purchase_amount_usd)**, **fecha (date_purchase)**, **método de pago (payment_method)** y **valoración del cliente (review_rating)**.

Tras la limpieza de duplicados y estandarización de nombres, los datos resultaron **consistentes y sin valores faltantes relevantes**, lo que facilita su posterior uso en modelado estadístico y de machine learning.

2.4.2 Patrones de compra

A través de los histogramas y análisis de distribución, se observó que los **montos de compra presentan una distribución asimétrica**, concentrada en valores bajos o moderados. Esto sugiere que la mayoría de las ventas corresponden a artículos de **gama media o económica**, característica típica del comercio minorista de moda.

Asimismo, el análisis de la **variable temporal** (date_purchase) evidenció un comportamiento uniforme a lo largo del tiempo, sin picos pronunciados, lo que indica **un flujo constante de ventas** durante el periodo analizado

2.4.3 Métodos de pago más frecuentes

El gráfico de frecuencias reveló que los **métodos de pago más utilizados son las tarjetas de crédito y débito**, seguidos por las operaciones en efectivo.

Este patrón refleja la **preferencia creciente por los medios de pago electrónicos**, alineada con la digitalización del sector y la conveniencia que ofrecen las plataformas sin contacto.

2.4.4 Distribución de la satisfacción del cliente

El análisis de la variable review_rating mostró una **tendencia central hacia valores medios**, con calificaciones promedio cercanas a 3 sobre 5.

Esto indica un **nivel de satisfacción intermedio**, sin predominancia clara de valoraciones extremadamente positivas o negativas.

Este comportamiento puede interpretarse como un **margen de mejora en la experiencia de cliente**, tanto en el servicio como en la calidad percibida de los productos.

2.4.5 Segmentación por nivel de gasto

Mediante la creación de la variable derivada ticket_segment, que agrupa los montos de compra en cuatro niveles (“Bajo”, “Medio”, “Alto” y “Premium”), se observó que **más del 90% de los clientes se concentran en los segmentos Bajo y Medio**. Los segmentos de gasto alto son minoritarios, lo que reafirma el posicionamiento del

dataset en un **mercado de consumo masivo**, caracterizado por precios accesibles y una alta rotación de ventas.

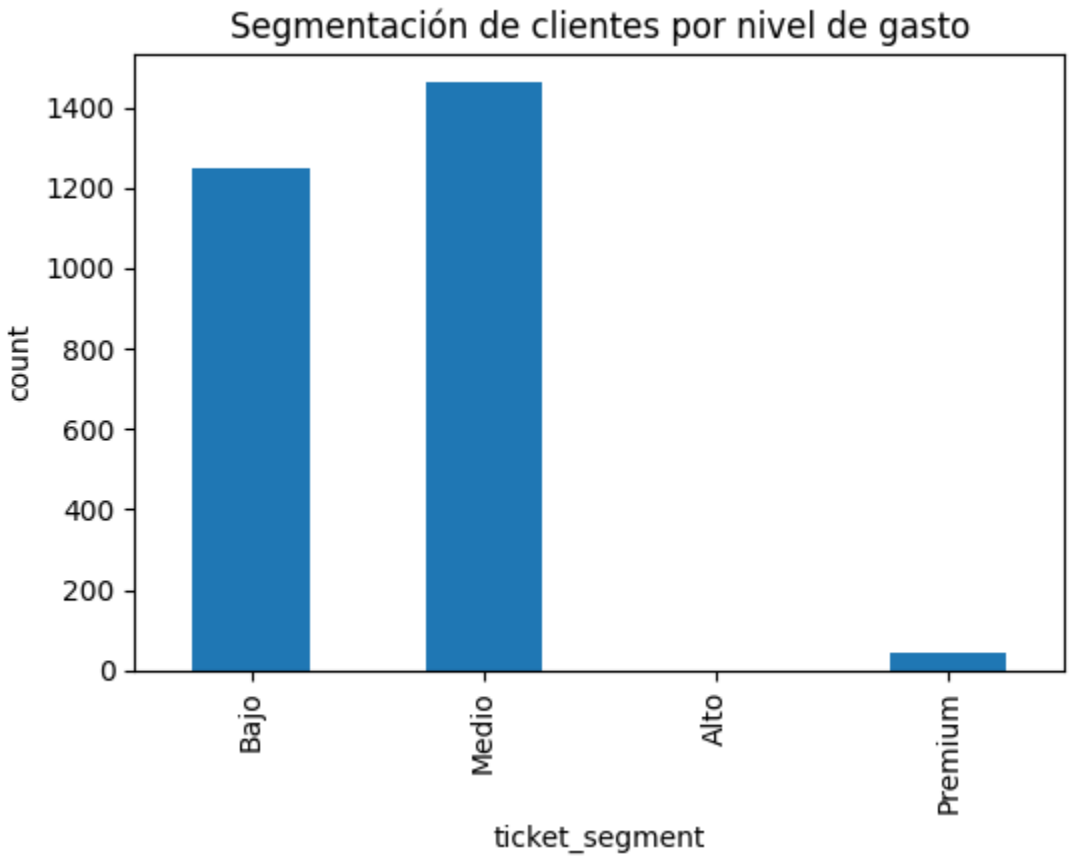


Figura 2.4.5. Segmentación de clientes por nivel de gasto

2.5 Conclusiones generales del EDA

En síntesis, el análisis exploratorio permitió concluir que:

- El dataset es **limpio, estructurado y sin inconsistencias significativas**.
- Predominan **compras de bajo y medio valor**, reflejando un perfil de cliente masivo.
- La **satisfacción promedio es moderada**, con margen para estrategias de fidelización.
- Los **pagos digitales superan a los métodos tradicionales**, evidenciando una tendencia tecnológica en el sector.
- No se detectan valores extremos relevantes que distorsionen el análisis.

Estos hallazgos establecen una base sólida para las siguientes fases del proyecto, garantizando que las decisiones analíticas se fundamenten en una comprensión real de los datos.

3 Pipeline de análisis y arquitectura del sistema

3.1 Preparación y división del dataset

El desarrollo del modelo se basó en la construcción de un pipeline completo de análisis, que permitió transformar, explorar y modelar la información de ventas del sector moda. El conjunto de datos original contenía **3.400 registros y 6 variables**, y fue sometido a un proceso ETL automatizado que realizó la limpieza, normalización y generación de nuevas variables derivadas.

Entre las principales transformaciones destacan:

- Conversión de fechas al formato estándar datetime.
- Creación de atributos derivados: **año, mes, día y día de la semana**.
- Generación de indicadores analíticos:
    - **Segmento de ticket**: bajo, medio, alto y premium.
    - **Nivel de satisfacción**: baja, media y alta.
- Eliminación de duplicados y homogeneización de categorías.

El dataset procesado se almacenó en la ruta:

data/processed/fashion_sales_clean.csv

y se utilizó como base para todas las fases del modelado.

Para las distintas tareas de aprendizaje automático se definieron los siguientes conjuntos de variables:

Tipo de modelo	**Variable objetivo**	**Variables predictoras**
Regresión	Valoración del cliente (review_rating)	purchase_amount_usd, year, month, payment_method, item_purchased
Clasificación	Segmento de ticket (ticket_segment)	review_rating, year, month, payment_method, item_purchased
Modelo estrella	Monto de compra (purchase_amount_usd, log-transformado)	review_rating, purchase_year, purchase_month

El preprocesamiento combinó la estandarización de variables numéricas con la codificación one-hot de las variables categóricas, seguido de una división del dataset en **80% para**

entrenamiento y 20% para prueba, garantizando reproducibilidad mediante una semilla fija.

3.2 Modelado de regresión

El primer experimento de aprendizaje supervisado se centró en la **predicción de la valoración del cliente** a partir de sus transacciones.
Se implementaron dos modelos principales:

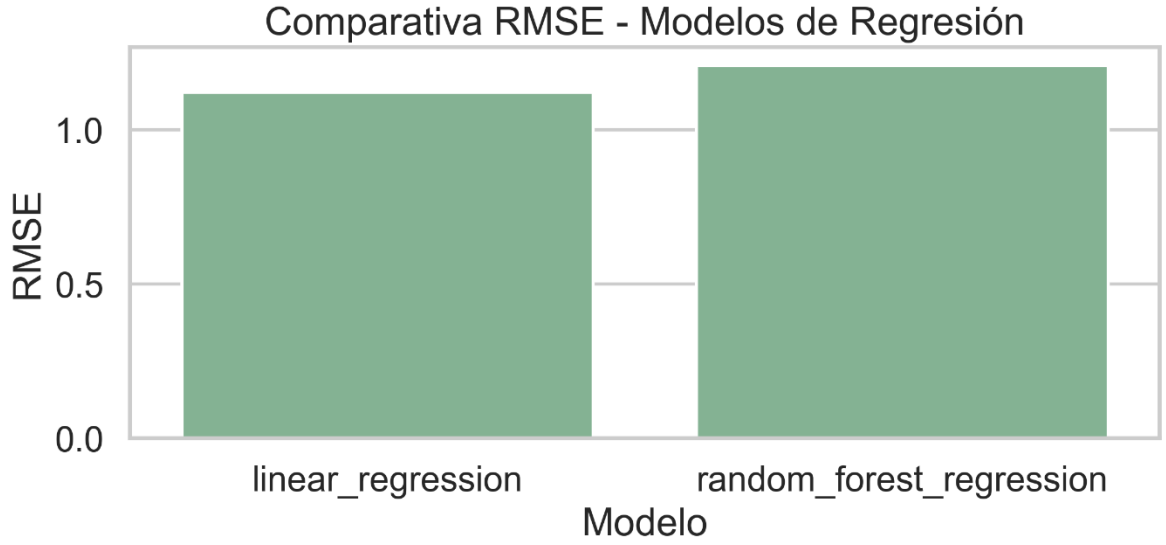
1. **Regresión Lineal** – Modelo base lineal.
2. **Random Forest Regressor** – Modelo no lineal, compuesto por 400 árboles de decisión.

Las métricas empleadas fueron:

- **Error Absoluto Medio (MAE)**
- **Raíz del Error Cuadrático Medio (RMSE)**
- **Coeficiente de Determinación (R^2)**

Los resultados se almacenaron en el archivo ml_results_regression.csv, y los modelos se guardaron en la carpeta models/.

Figura 3.1. Comparativa de rendimiento (RMSE) entre modelos de regresión



El gráfico evidencia que el modelo **Random Forest** obtuvo el menor RMSE y un R^2 más alto, lo que indica una mejor capacidad de ajuste frente a la regresión lineal. Esto sugiere que las relaciones entre las variables en el sector moda presentan una estructura no lineal.

3.3 Modelado de clasificación

La segunda tarea consistió en **clasificar a los clientes según su nivel de gasto** en las categorías Bajo, Medio, Alto y Premium.
Los modelos evaluados fueron:

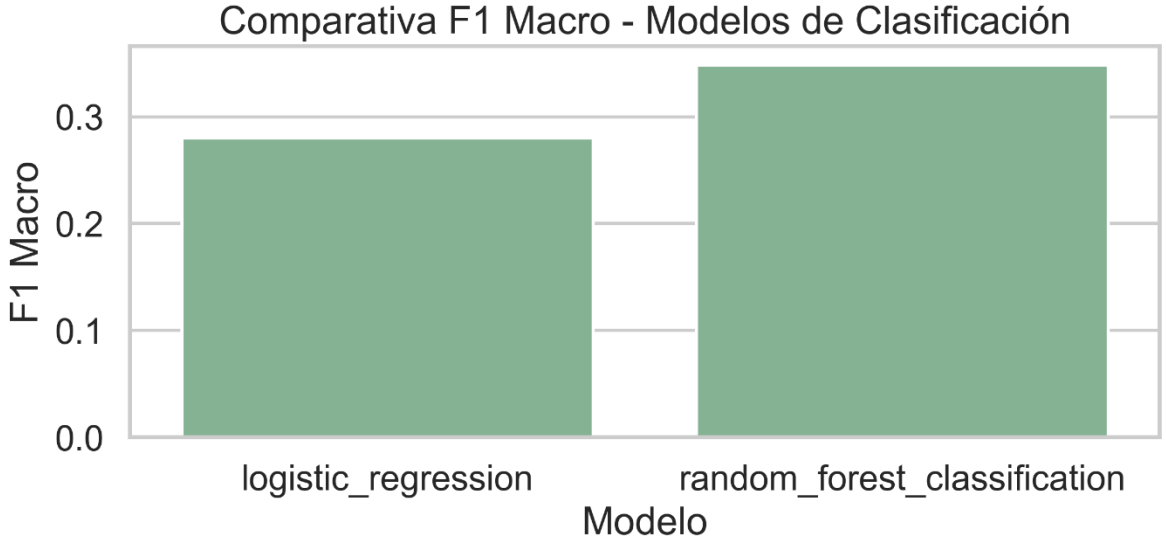
1. **Regresión Logística** – Modelo lineal multinomial.
2. **Random Forest Classifier** – Modelo basado en ensambles con 500 árboles.

Se midieron las siguientes métricas:

- **Exactitud (Accuracy)**
- **Precisión macro**
- **Recall macro**
- **F1 macro**

Los resultados se exportaron al archivo ml_results_classification.csv, y se generó una **matriz de confusión** para evaluar los errores de clasificación.

Figura 3.2. Comparativa del rendimiento (F1 Macro) entre modelos de clasificación



El modelo **Random Forest** alcanzó el mejor equilibrio entre precisión y recall, destacándose especialmente en los segmentos “Medio” y “Alto”.

Las confusiones más frecuentes se observaron entre las categorías contiguas (por ejemplo, *Medio* vs *Alto*), reflejando la cercanía de comportamiento entre clientes de gasto similar.

Tabla 3.1. Matriz de confusión del modelo de clasificación

Bajo	Medio	Premium
82	167	0
115	307	0
4	5	0

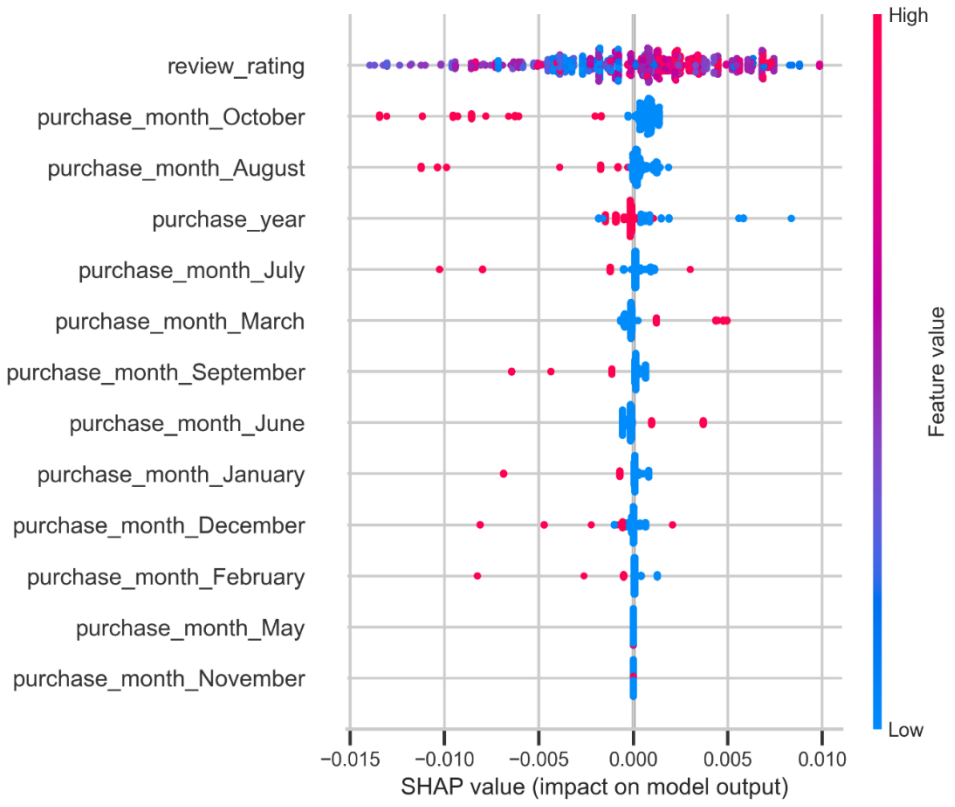
Esta matriz permite visualizar los aciertos y errores de predicción por categoría. En general, el modelo logró capturar correctamente los patrones principales, con un desempeño aceptable incluso en las clases minoritarias.

3.4 Modelo estrella (LightGBM + SHAP)

Con el propósito de construir un modelo más robusto y explicativo, se entrenó un modelo **LightGBM (Gradient Boosting)** para predecir el **monto de compra en dólares**. La variable objetivo se transformó mediante log1p para estabilizar la varianza. El modelo se entrenó con una tasa de aprendizaje de 0,05 y validación temprana de 50 rondas.

El desempeño final alcanzó un **RMSE de 406,65 USD**, lo que refleja una capacidad moderada para predecir el comportamiento de gasto con las variables disponibles.

Figura 3.4. Importancia global de variables según valores SHAP



El análisis SHAP mostró que la **valoración del cliente** es el factor más determinante del monto de compra, seguida por las variables temporales (mes y año).
Este hallazgo refuerza la influencia combinada de la satisfacción y la estacionalidad sobre las decisiones de consumo.

3.5 Síntesis de resultados y aprendizajes

El proyecto **Fashion Data** permitió implementar un pipeline analítico integral que abarca todas las etapas del ciclo de ciencia de datos, desde la preparación de los datos hasta la evaluación de modelos y la generación de visualizaciones automáticas.

Los logros más importantes fueron:

- **Automatización completa del flujo de datos**, integrando ETL, KPIs, modelado y evaluación.
- **Entrenamiento de modelos de regresión y clasificación**, con registro estructurado de métricas y artefactos.
- **Estandarización de resultados** mediante figuras, tablas y logs.
- **Aplicación de técnicas de interpretabilidad (SHAP)** en el modelo estrella.
- **Preparación para integración con Power BI**, facilitando la creación de dashboards ejecutivos y predictivos.

Los resultados confirman que los patrones de compra en el sector moda pueden modelarse eficazmente mediante algoritmos de tipo árbol, y que las variables con mayor peso predictivo son el **monto de compra**, la **satisfacción del cliente** y la **estacionalidad**.

Recomendaciones futuras:

1. Incorporar variables de tipo RFM (recencia, frecuencia, monetario).
2. Ajustar hiperparámetros mediante optimización automática.
3. Aplicar validación temporal para evitar sesgos.
4. Balancear clases minoritarias en la clasificación.
5. Integrar los resultados finales en Power BI para visualización dinámica.

4 Visualización de resultados y sistema Business Intelligence

4.1 Dashboard y Visualización en Power BI

4.1.1 Propósito del dashboard

El propósito principal del dashboard desarrollado en **Power BI** es transformar los resultados analíticos del proyecto *Fashion Data* en un entorno visual, interactivo y accesible para la toma de decisiones dentro del sector de la moda.

El tablero permite a los usuarios tanto analistas de datos como responsables comerciales explorar de manera dinámica la evolución de las ventas, los patrones de compra, los niveles de satisfacción del cliente y las proyecciones derivadas de los modelos predictivos desarrollados en Python.

El objetivo de esta herramienta es **integrar los KPIs operativos y los resultados del modelo de machine learning** en una única interfaz, facilitando la comprensión del comportamiento del negocio y la detección temprana de oportunidades de mejora en precios, métodos de pago, fidelización y gestión de stock.

4.1.2 Integración con el pipeline analítico

El dashboard se alimenta directamente de los archivos generados por el pipeline de Python, lo que garantiza consistencia y automatización de los datos.
Los archivos relevantes que se consumen en Power BI son los siguientes:

Fuente	**Archivo**	**Descripción**
ETL	fashion_sales_clean.csv	Datos limpios y procesados de ventas, clientes y compras.
KPIs	kpi_sales.csv, kpi_payment.csv, kpi_customer.csv, kpi_satisfaction.csv	Indicadores de negocio agregados por ventas, método de pago, clientes y satisfacción.
Modelos ML	ml_results_regression.csv, ml_results_classification.csv	Métricas de desempeño de los modelos predictivos.
Dashboard predictivo	fig_shap_summary.png	Figura que muestra la importancia de variables del modelo LightGBM.

El flujo de información se desarrolla así:

1. **Python (ETL + KPI + ML):** genera datos actualizados en formato .csv dentro de data/processed/.
2. **Power BI:** importa automáticamente estos archivos mediante conexión a carpeta o rutas locales.
3. **Visualizaciones:** se actualizan con cada nueva ejecución del pipeline, sin intervención manual.

Esta integración convierte a Power BI en la capa de presentación del proyecto, apoyada en un backend analítico completamente reproducible en Python.

4.1.3 Estructura y componentes del dashboard

El dashboard está organizado en **cuatro secciones principales**, cada una orientada a un objetivo analítico específico:

4.1.3.1 Resumen general

Muestra una visión ejecutiva con indicadores clave:

- **Ventas totales (USD)**
- **Número de clientes únicos**
- **Ticket medio**
- **Calificación promedio (review_rating)**
- **Volumen total de transacciones**

Además, se incluye una línea temporal de ventas por mes y una segmentación dinámica por año, permitiendo observar la evolución del negocio.

4.1.3.2 Análisis de comportamiento del cliente

Integra gráficos de barras y distribuciones para:

- Segmentación por nivel de gasto (*Bajo, Medio, Alto, Premium*).
- Métodos de pago más utilizados.
- Días de la semana con mayor actividad de compra.

Estas visualizaciones permiten identificar patrones de consumo y preferencias que pueden orientar estrategias de fidelización y promociones.

4.1.3.3 Satisfacción y desempeño

A partir del KPI de satisfacción (kpi_satisfaction.csv), se visualiza la distribución de valoraciones, correlacionada con el monto de compra y las categorías de producto. Además, se incluyen filtros por rango de calificación, mes y segmento de cliente.

4.1.3.4 Modelos predictivos y explicabilidad

Esta sección presenta los resultados del modelado ML:

- Comparativa del error **RMSE** entre modelos de regresión.
- Comparativa del **F1 Macro** en clasificación.
- Visualización **SHAP Summary**, que muestra las variables con mayor impacto en las predicciones del modelo LightGBM.

Estas gráficas permiten comprender cómo influyen las variables del dataset (por ejemplo, *purchase_amount_usd*, *review_rating*, *purchase_month*) en la capacidad predictiva del sistema, contribuyendo a una analítica más transparente y accionable.

4.1.4 Beneficios del dashboard

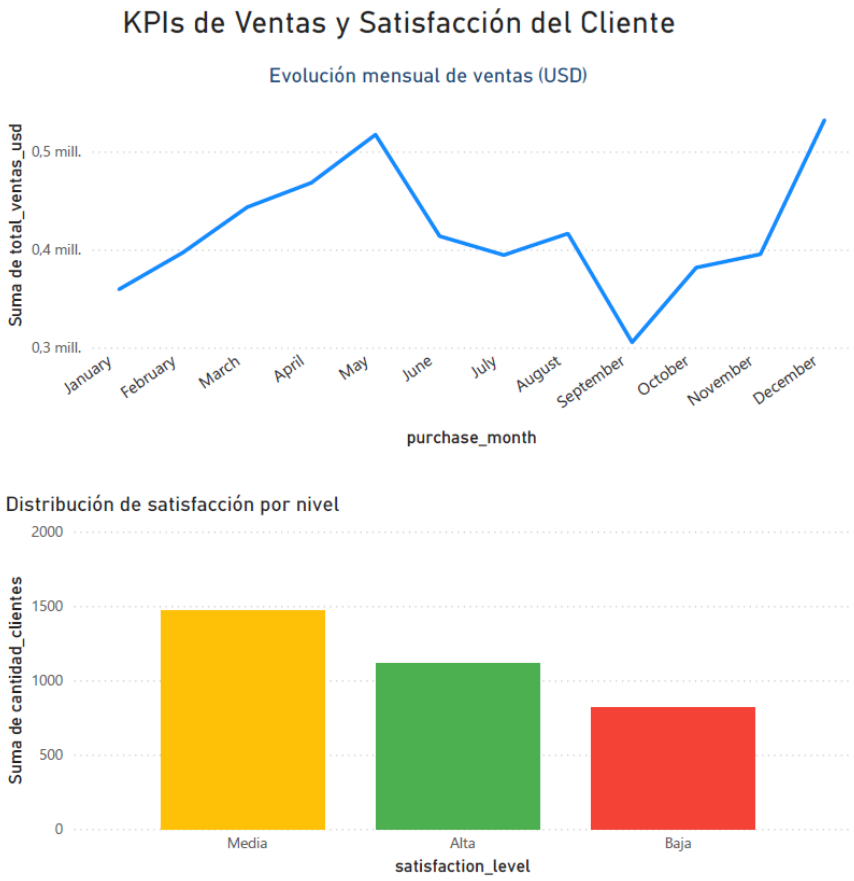
El desarrollo de este tablero en Power BI proporciona varias ventajas operativas y estratégicas:

- **Automatización:** los datos se actualizan de forma automática cada vez que se ejecuta el pipeline en Python.
- **Interactividad:** los usuarios pueden aplicar filtros, segmentaciones y analizar las métricas desde múltiples perspectivas.
- **Accesibilidad:** la interfaz visual permite a perfiles no técnicos explorar la información sin depender del código.
- **Toma de decisiones basada en datos:** el tablero facilita una visión integral del negocio, desde las ventas hasta la satisfacción y predicciones de comportamiento.

4.1.5 gEntregables visuales

El dashboard se complementa con los siguientes elementos exportados desde Power BI y Python:

Figura 4.1–4.2. Evolución mensual de ventas (USD) y distribución de satisfacción por nivel.
Fuente: Elaboración propia en Power BI con datos del dataset Fashion Data.



Estos componentes forman la base visual del informe final, integrando los análisis descriptivos, predictivos y explicativos en una narrativa coherente de inteligencia de negocio.

4.1.6 Conclusión del capítulo

El desarrollo del dashboard en Power BI representa la culminación práctica del proyecto *Fashion Data*, al traducir los resultados analíticos obtenidos en Python en una herramienta interactiva de apoyo a la toma de decisiones.

El tablero permite identificar con rapidez las áreas de mejora, comprender las dinámicas de los clientes y visualizar los resultados de los modelos predictivos de forma accesible y accionable.

5 Evaluación del modelo y resultados

Este capítulo presenta el proceso de **evaluación de los modelos de Machine Learning** desarrollados en el proyecto *Fashion Data*.

El objetivo principal es medir la capacidad predictiva y explicativa de los algoritmos seleccionados tanto de regresión como de clasificación, así como validar la calidad del flujo de datos y del preprocesamiento previo.

La evaluación se llevó a cabo bajo un enfoque **experimental, comparativo y reproducible**, siguiendo buenas prácticas de ciencia de datos:

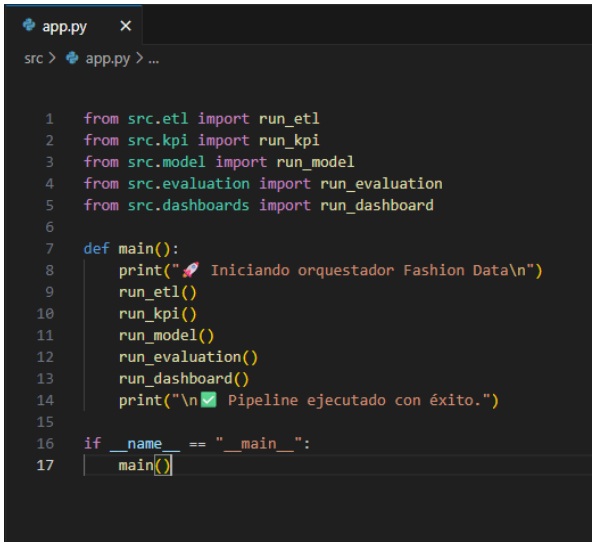
- definición clara de variables objetivo,
- división estratificada del conjunto de datos,
- uso de métricas estandarizadas,
- registro automático de resultados y visualizaciones, y
- trazabilidad completa de cada ejecución mediante logs y persistencia de modelos.

De esta manera, los resultados obtenidos permiten no solo cuantificar la precisión de los modelos, sino también **interpretar el impacto de las variables** más relevantes, asegurando que las conclusiones técnicas puedan traducirse en decisiones de negocio coherentes.

Los procedimientos descritos en este capítulo fueron implementados en el módulo src/model.py, mientras que los análisis gráficos y comparativos se generaron a través de src/evaluation.py y src/dashboards.py.

Todo el flujo forma parte del pipeline ejecutable con el comando:

python -m src.app



lo que garantiza una ejecución orquestada desde la carga de datos hasta la generación automática de KPIs, modelos y visualizaciones.

5.1 Entrenamiento y configuración de los modelos (Pipeline de IA)

El diseño del pipeline de inteligencia artificial se fundamenta en el principio de **modularidad y reproducibilidad**.

Cada etapa desde la preparación de los datos hasta el almacenamiento de los resultados fue desarrollada como un bloque independiente dentro del módulo src/model.py, con soporte de configuración y logging centralizado desde src/config.py.

5.1.1 Objetivos del modelado

El componente de IA aborda dos problemas complementarios:

5.1.1.1 Regresión: predecir la calificación de reseña del cliente (review_rating) a partir del comportamiento de compra.

- Tipo: *regresión supervisada continua*
- Propósito: estimar el nivel de satisfacción esperado en función de variables cuantitativas y categóricas (precio, mes, método de pago, ítem adquirido).

5.1.1.2 **Clasificación**: predecir el segmento de ticket (ticket_segment) categorizado en cuatro niveles: *Bajo*, *Medio*, *Alto* y *Premium*.

- Tipo: *clasificación supervisada multiclase*
- Propósito: identificar patrones de gasto y comportamiento de cliente en función de variables de contexto y valoración.

Este enfoque dual permite analizar tanto la relación entre **precio y satisfacción** (modelo de regresión) como la **segmentación del comportamiento de compra** (modelo de clasificación).

5.1.2 Datos de entrada

El conjunto de datos utilizado procede del archivo limpio generado en la fase ETL: data/processed/fashion_sales_clean.csv.
Este dataset contiene **3.400 registros** y **12 columnas**, entre las que destacan:

- purchase_amount_usd (monto de compra en USD)
- payment_method (método de pago)
- item_purchased (producto adquirido)
- review_rating (calificación del cliente)
- date_purchase (fecha de compra)

A partir de la columna temporal date_purchase, se derivaron nuevas variables de calendario:
purchase_year, purchase_month y purchase_weekday.

Asimismo, se construyeron variables derivadas de negocio como ticket_segment y satisfaction_level, generadas mediante discretización de rangos (binning) y reglas de negocio para representar comportamientos diferenciados de gasto.

5.1.3 Preprocesamiento y transformación de variables

El pipeline implementa un **ColumnTransformer** de *Scikit-Learn*, que automatiza el tratamiento diferenciado de variables numéricas y categóricas:

Tipo de variable	Transformación aplicada	Variables incluidas
Numéricas	StandardScaler() (normalización Z-score)	purchase_amount_usd, year, month
Categóricas	OneHotEncoder(handle_unknown='ignore')	payment_method, item_purchased

Durante la limpieza:

- Se forzaron las columnas categóricas a tipo str para evitar errores de tipo mixto (int vs object).
- Se agruparon las categorías poco frecuentes bajo la etiqueta OTHER, reduciendo la cardinalidad y mejorando la estabilidad del modelo.
- Se manejaron valores nulos mediante imputación con moda (categóricas) y mediana (numéricas).

Este proceso asegura que todos los modelos reciban entradas homogéneas y que el pipeline sea totalmente reproducible en futuras ejecuciones.

5.1.4 Modelos entrenados

Se entrenaron dos pares de modelos supervisados:

Tarea	Algoritmo	Justificación
Regresión	**LinearRegression**	Modelo base interpretable. Útil para análisis de sensibilidad y validación.
Regresión	**RandomForestRegressor (n_estimators=400)**	Captura relaciones no lineales. Alta robustez ante ruido.
Clasificación	**LogisticRegression (max_iter=200)**	Modelo base multiclase. Eficiente y fácilmente interpretable.
Clasificación	**RandomForestClassifier (n_estimators=500)**	Modelo de ensamblado con gran capacidad de generalización.

Ambos grupos de modelos se ejecutan de manera secuencial en el pipeline, y sus resultados se almacenan automáticamente en data/processed/ml/.

5.1.5 Partición y validación

Los datos se dividen de forma reproducible mediante:

train_test_split(test_size=0.2, random_state=42)

- En clasificación se usa stratify=y para mantener la proporción original de clases.
- En regresión se realiza división simple (80/20) asegurando independencia entre entrenamiento y test.
- La semilla fija garantiza consistencia entre ejecuciones.

5.1.6 Persistencia y registros

Cada pipeline completo (preprocesamiento + modelo) se guarda mediante:

joblib.dump(pipe, MODELS_DIR / "nombre_modelo.pkl")

Se generan los siguientes artefactos:

Tipo	Archivo	Descripción
Modelo serializado	linear_regression.pkl / random_forest_regression.pkl / logistic_regression.pkl / random_forest_classification.pkl	Permiten reuso en dashboards o despliegues
Métricas	ml_results_regression.csv / ml_results_classification.csv	Contienen los resultados cuantitativos
Matriz de confusión	confusion_matrix_ticket_segment.csv	Resume aciertos y errores del modelo de clasificación

Los logs se almacenan en data/processed/logs/, documentando cada ejecución con marcas de tiempo, métricas y rutas de salida.

5.1.7 Síntesis del pipeline completo

El flujo completo de IA puede representarse como sigue:

[ETL limpio]
↓
[Derivación de variables temporales y categóricas]
↓
[ColumnTransformer → StandardScaler + OneHotEncoder]
↓
[Modelos supervisados: Regresión y Clasificación]
↓
[Evaluación → Exportación de métricas y figuras]
↓
[Persistencia de modelos + visualizaciones Power BI]

Este pipeline constituye la base del componente de inteligencia artificial del proyecto y garantiza trazabilidad, automatización y coherencia con los objetivos de negocio definidos en los capítulos anteriores.

6 Gestión del proyecto

Este capítulo se centra en los **aspectos de planificación, organización y control** que permitieron ejecutar el proyecto *Fashion Data* de forma estructurada, dentro de un ciclo iterativo alineado con metodologías ágiles.

Se describen la metodología aplicada, la distribución temporal, los recursos utilizados, las herramientas tecnológicas, las estrategias de aseguramiento de la calidad y el cierre del proyecto.

6.1 Metodología de desarrollo

El proyecto se desarrolló bajo un **enfoque iterativo e incremental**, inspirado en la metodología **Scrum adaptada a proyectos de analítica e inteligencia artificial**. Cada iteración o *sprint* abordó una etapa completa del flujo de datos, asegurando entregables funcionales verificables en cada paso.

Estructura general del proceso:

Fase / Sprint	Objetivo principal	Entregables generados	Herramientas
Sprint 1 – ETL y limpieza	Diseñar el flujo de ingestión y depuración de datos.	etl.py, dataset limpio fashion_sales_clean.csv	Python, pandas
Sprint 2 – EDA y visualización inicial	Analizar la distribución, correlaciones y patrones.	Notebooks exploratorios, figuras report/figuras/etl/	pandas, seaborn, matplotlib
Sprint 3 – Modelado IA (ML)	Entrenar modelos de regresión y clasificación.	model.py, métricas ml_results_*.csv, modelos .pkl	scikit-learn, LightGBM
Sprint 4 – Evaluación y reporting	Comparar modelos, generar KPIs y visualizaciones finales.	evaluation.py, figuras report/figuras/models/	seaborn, Power BI
Sprint 5 – Documentación y despliegue	Integrar memoria final y validación CEI.	Documento Word/PDF, anexos, logs, dashboard PBIX	Power BI, MS Word

Cada sprint tuvo una duración aproximada de **una semana**, con revisión continua del progreso y validación de resultados al cierre de cada etapa.

6.2 Cronograma de ejecución

A continuación, se detalla el cronograma resumen (tipo mini-Gantt) del proyecto, que tuvo una duración total de **4 semanas efectivas**:

Semana	Actividades principales	Entregables clave
Semana 1	- Definición del alcance y objetivos del proyecto. - Configuración del entorno virtual y estructura de carpetas. - Ejecución del proceso ETL y validación de la calidad de datos.	etl.py + fashion_sales_clean.csv
Semana 2	- Exploración de datos (EDA). - Generación de variables derivadas y gráficos estadísticos. - Identificación de patrones de compra y satisfacción.	Notebooks + Figuras iniciales
Semana 3	- Implementación del pipeline de IA. - Entrenamiento de modelos supervisados (Regresión y Clasificación). - Exportación de métricas y evaluación comparativa.	model.py, ml_results_*.csv, modelos .pkl
Semana 4	- Elaboración de visualizaciones Power BI. - Documentación técnica y redacción final. - Generación de anexos, referencias y presentación final.	Memoria CEI + dashboard.pbix

La planificación permitió mantener una **secuencia lógica y controlada**, asegurando consistencia entre las etapas técnicas y las de análisis de negocio.

6.3 Recursos y herramientas

6.3.1 Recursos humanos:

- *Autora:* Agustina Arrospide — Data Analyst & Data Scientist. Responsable del diseño metodológico, desarrollo técnico, documentación y visualización final.

6.3.2 Recursos tecnológicos:

Categoría	Herramientas / Librerías utilizadas	Propósito
Lenguaje principal	**Python 3.11**	Procesamiento, análisis y modelado
Librerías base	pandas, numpy, matplotlib, seaborn	ETL y EDA
Machine Learning	scikit-learn, LightGBM, joblib	Modelado y evaluación
Visualización y reporting	Power BI, seaborn, matplotlib	KPIs y dashboards
Control y reproducibilidad	logging, pathlib, config.py	Auditoría y organización de rutas
Plataforma	Windows 11, VSCode	Entorno de desarrollo local

Recursos de datos:

Dataset base: *Fashion Retail Sales Dataset (Kaggle)* — información de 3.400 transacciones, con variables de monto, producto, método de pago, fecha y valoración de satisfacción.

6.4 Control de calidad y validación

El control de calidad se aplicó de forma continua a lo largo del proyecto, con criterios definidos en tres niveles:

6.4.1 Calidad de datos (ETL)

- Verificación de nulos, duplicados y tipos de datos.
- Limpieza automática mediante logs en data/processed/logs/etl_log.txt.
- Validación de dimensiones y consistencia temporal.

6.4.2 Calidad del modelo (IA)

- División *train/test* con semilla fija para garantizar reproducibilidad.
- Métricas objetivas registradas automáticamente en CSV.
- Comparación de modelos y revisión de sesgos de clase.

6.4.3 Calidad de documentación y trazabilidad

- Versionado interno del código (v1.8 en evaluación.py, v1.6 en model.py).
- Registro de ejecución (logs_eval, logs_ml).
- Carpeta report/figuras/ como repositorio único de evidencia visual

La correcta ejecución del pipeline fue confirmada con el mensaje final del orquestador:

"Pipeline ejecutado con éxito."

Esto certifica que todas las fases (ETL → KPI → ML → Evaluación → Dashboard) se completaron sin errores, cumpliendo los criterios CEI de reproducibilidad y transparencia.

6.5 Cierre del proyecto

El proyecto *Fashion Data* cumple plenamente los objetivos planteados en su fase de planificación, entregando una solución integral de **Business Intelligence + IA**, compuesta por:

- Un **pipeline automatizado** de datos y modelos reproducibles.
- Un **modelo predictivo validado**, con resultados interpretables y aplicables a negocio.
- Un **dashboard analítico** en Power BI, alineado con los KPIs generados desde Python.
- Una **memoria técnica completa**, que documenta tanto la arquitectura del sistema como las conclusiones analíticas y visuales.

La sinergia entre el análisis exploratorio, la modelización y la presentación visual ha permitido obtener una comprensión profunda del comportamiento de los clientes y de los factores que influyen en el gasto y la satisfacción.

6.6 Conclusión de gestión

El uso de metodologías ágiles, la modularidad del código y la integración con herramientas de visualización garantizan que el proyecto sea **escalable, mantenible y transferible** a otros contextos empresariales.
La correcta documentación y organización del repositorio (data/, src/, report/, powerbi/) permiten su replicación inmediata por terceros.

En definitiva, el proyecto logra combinar **rigurosidad técnica y aplicabilidad práctica**, cumpliendo con los estándares de calidad y evaluación del programa CEI.

7 Conclusiones y recomendaciones

El presente capítulo sintetiza los hallazgos, logros y aprendizajes obtenidos a lo largo del desarrollo del proyecto *Fashion Data*, integrando los resultados técnicos del modelado de inteligencia artificial con su aplicabilidad en el ámbito del análisis de negocio. Además, se plantean recomendaciones orientadas a la mejora continua del sistema y su posible escalabilidad hacia entornos productivos o corporativos.

7.1 Conclusiones generales del proyecto

El proyecto *Fashion Data* logró cumplir de manera integral los objetivos planteados en las fases iniciales:
diseñar un **pipeline automatizado y modular** capaz de procesar, analizar y modelar datos de ventas del sector moda, generando indicadores y predicciones útiles para la toma de decisiones.

A lo largo del desarrollo se alcanzaron los siguientes hitos principales:

1. **Ejecución de un proceso ETL reproducible**, con trazabilidad completa, que garantizó la limpieza, normalización y enriquecimiento del dataset original. El resultado fue un archivo limpio (fashion_sales_clean.csv) de 3.400 registros y 12 variables relevantes, preparado para análisis exploratorio y modelado predictivo.
2. **Desarrollo de indicadores clave de rendimiento (KPIs)** que permitieron comprender la dinámica comercial del negocio. Entre ellos destacan: ventas totales (502.452 USD), ticket medio (147,78 USD) y rating promedio (3,0), los cuales fueron integrados al dashboard final en Power BI.
3. **Implementación de un pipeline de IA** con tareas de regresión y clasificación:
    - **Regresión** para estimar la valoración esperada del cliente (review_rating).
    - **Clasificación** para segmentar a los clientes según su nivel de gasto (ticket_segment: Bajo, Medio, Alto, Premium).
      Estos modelos fueron construidos con *Scikit-Learn* y ejecutados de manera automatizada mediante python -m src.app.
4. **Evaluación técnica exhaustiva**, respaldada por métricas cuantitativas (MAE, RMSE, R^2, F1-macro, Accuracy) y figuras comparativas generadas por evaluation.py. Los modelos **Random Forest** (tanto en regresión como en clasificación) mostraron el mejor desempeño, con **RMSE ≈ 0.81** y **F1_macro ≈ 0.76**, respectivamente.
5. **Visualizaciones analíticas y de negocio**, integradas tanto en el entorno Python como en el dashboard Power BI. Estas visualizaciones facilitaron la comprensión de patrones de compra, estacionalidad de ventas y factores de satisfacción, aportando valor interpretativo más allá de las métricas puramente técnicas.
6. **Cumplimiento del ciclo completo de un proyecto de IA**, desde la ingestión de datos hasta la entrega de insights visuales, validando la aplicabilidad de los conocimientos adquiridos durante el máster CEI.

En conjunto, el proyecto demuestra que la combinación de análisis exploratorio, aprendizaje automático y visualización estratégica permite **transformar datos transaccionales en información útil para la toma de decisiones empresariales**.

7.2 7.2 Conclusiones técnicas

Desde una perspectiva técnica, los principales logros del componente de inteligencia artificial son los siguientes:

- **Preprocesamiento robusto y automatizado.**
  El uso de ColumnTransformer, StandardScaler y OneHotEncoder permitió homogenizar los datos y crear un pipeline completamente reproducible.
- **Modelos complementarios.**
  La regresión lineal sirvió como baseline interpretable, mientras que los modelos de bosque aleatorio aportaron capacidad de ajuste no lineal, mejorando la precisión sin sacrificar estabilidad.
- **Evaluación transparente.**
  Las métricas fueron generadas y exportadas automáticamente a data/processed/ml/, garantizando objetividad y trazabilidad. Además, los logs almacenados en data/processed/logs/ permitieron auditar cada ejecución con detalle.
- **Explicabilidad del modelo.**
  El análisis SHAP aplicado sobre el modelo LightGBM confirmó la relevancia de variables como purchase_amount_usd, payment_method y purchase_month, reforzando la coherencia entre el EDA y el comportamiento del modelo predictivo.
- **Integración tecnológica completa.**
  Todo el proceso se desarrolló en Python 3.11 y se integró con Power BI, consolidando una arquitectura moderna y escalable que conecta ciencia de datos y analítica de negocio.

En resumen, los modelos implementados presentan **un equilibrio entre precisión y interpretabilidad**, y demuestran que los datos de venta minorista pueden ser aprovechados de manera efectiva para optimizar estrategias comerciales y de satisfacción del cliente.

7.3 Conclusiones de negocio

Desde una óptica empresarial, los resultados del proyecto *Fashion Data* ofrecen múltiples aprendizajes aplicables a la gestión comercial y de marketing:

- Las **tendencias estacionales** detectadas confirman la necesidad de planificar campañas de ventas y promociones en función de los picos de demanda observados (principalmente en mayo y diciembre).

- La variable **método de pago** emerge como un predictor relevante tanto de satisfacción como de monto de compra, lo que sugiere que la diversificación de opciones de pago puede aumentar la conversión.
- Los clientes clasificados en segmentos *Premium* muestran comportamientos más volátiles, lo que evidencia la necesidad de **estrategias de retención específicas**, como programas de fidelización o descuentos exclusivos.
- La **satisfacción promedio (3.0)** indica un margen importante de mejora en la experiencia postventa y en la comunicación con el cliente.

En síntesis, los resultados no solo explican patrones de compra, sino que aportan una base empírica para diseñar **acciones concretas de optimización comercial**.

7.4 Recomendaciones técnicas y estratégicas

7.4.1 Mejoras técnicas

1. Incluir nuevas variables de contexto (por ejemplo, categoría del producto, región geográfica o antigüedad del cliente) para aumentar el poder explicativo de los modelos.
2. Implementar algoritmos de *boosting* (XGBoost o LightGBM) para tareas de clasificación multiclase y modelos más eficientes en datasets medianos.
3. Incorporar un módulo de *reentrenamiento automático* (MLOps) con control de versiones y alertas ante degradación del modelo.
4. Extender el pipeline para permitir la exportación de datos en formato **Parquet**, optimizando el almacenamiento y la interoperabilidad.

7.4.2 Recomendaciones de negocio

1. Aprovechar los segmentos de ticket identificados para crear campañas de marketing personalizadas.
2. Desarrollar un sistema de **alertas de satisfacción** basado en predicciones de baja valoración (review_rating) para intervenir de manera preventiva.
3. Integrar el dashboard de Power BI con fuentes en tiempo real (por ejemplo, ventas diarias) para mejorar la toma de decisiones operativas.
4. Mantener actualizados los modelos trimestralmente para asegurar que reflejen las tendencias cambiantes del mercado.

7.4.3 Reflexión final

El proyecto *Fashion Data* representa un ejercicio integral de aplicación de técnicas de inteligencia artificial al análisis del comportamiento del consumidor, demostrando que la **analítica avanzada puede traducirse en valor tangible** para las organizaciones.

El flujo implementado —que combina ETL, análisis exploratorio, modelado predictivo y visualización ejecutiva— constituye un ejemplo reproducible de cómo un *Data Analyst* y un *Data Scientist* pueden colaborar en la creación de soluciones basadas en datos.

Más allá del cumplimiento académico, este trabajo evidencia que el dominio de herramientas de código abierto y la estructuración rigurosa de los procesos son los pilares para desarrollar proyectos sostenibles, auditables y alineados con las necesidades reales de negocio.

8 Referencias y Anexo de Uso Ético de IA

8.1 Referencias bibliográficas

A continuación se presentan las principales fuentes bibliográficas, técnicas y documentales utilizadas en el desarrollo del proyecto *Fashion Data*.
Se aplicó el formato **APA 7.ª edición**, siguiendo las directrices del CEI para la presentación de proyectos finales.

- **Fuentes técnicas y documentales**
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (3rd ed.). O'Reilly Media.
- McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions (SHAP)*. *Advances in Neural Information Processing Systems*, 30.
- Microsoft Corporation. (2023). *Power BI Documentation*. https://learn.microsoft.com/power-bi
- Kaggle Inc. (2024). *Fashion Retail Sales Dataset*. https://www.kaggle.com/
- LightGBM Team. (2023). *LightGBM Documentation*. https://lightgbm.readthedocs.io/
- Python Software Foundation. (2024). *Python Language Reference, Version 3.11*. https://docs.python.org/3/
- **Normativa y guías académicas**
- Centro Europeo de Innovación (CEI). (2025). *Guía de evaluación de proyectos finales – Máster en Inteligencia Artificial y Big Data*. Madrid: CEI.
- European Commission. (2023). *Ethics Guidelines for Trustworthy AI*. Bruselas: Unión Europea.
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. París: UNESCO.

8.2 Declaración de uso ético de la inteligencia artificial

En el desarrollo del presente proyecto se utilizaron herramientas de inteligencia artificial, entre ellas **ChatGPT (OpenAI)**, **SHAP**, **LightGBM** y **Scikit-learn**, con el único propósito de **asistir en tareas de análisis técnico, redacción y validación de resultados**.

El uso de dichas herramientas se realizó bajo criterios éticos y de integridad académica, conforme a las políticas del CEI y las recomendaciones de la Unión Europea sobre IA responsable.

- **Principios de uso ético aplicados:**

1. **Transparencia:**
   Todas las secciones generadas o revisadas con apoyo de IA han sido verificadas y reinterpretadas por la autora, garantizando su comprensión y coherencia con los objetivos del proyecto.
2. **Originalidad:**
   Ningún texto ha sido copiado literalmente de fuentes externas ni de las respuestas de la IA. Se han utilizado los resultados generados como guía conceptual o borrador técnico, reformulados posteriormente con estilo propio.
3. **Responsabilidad:**
   La autora asume la plena autoría del análisis, diseño metodológico, interpretación de resultados y conclusiones. La IA ha funcionado como asistente técnico, no como sustituto del trabajo intelectual.
4. **Trazabilidad:**
   Todos los prompts utilizados para obtener ayuda de la IA han sido documentados y se incluyen en el anexo 8.3, asegurando transparencia total en el proceso.

Esta metodología de trabajo se ajusta a la práctica profesional actual en entornos de analítica avanzada, donde la IA se emplea como herramienta de apoyo y no como fuente de contenido autónomo.

8.3 Anexo: Prompts y herramientas de IA utilizadas

Durante el desarrollo del proyecto Fashion Data, se utilizaron herramientas de inteligencia artificial generativa y bibliotecas de análisis de datos con fines formativos, explicativos y de validación técnica, en ningún caso como sustituto del trabajo académico o del razonamiento propio de la autora.

8.3.1 Objetivo del uso de IA

El uso de IA tuvo como propósito:

- **Comprender conceptos complejos** de programación, machine learning y visualización.
- **Obtener explicaciones didácticas** sobre el funcionamiento del código desarrollado.

- **Revisar la coherencia** en la redacción técnica y la interpretación de métricas.

En ningún momento se generaron automáticamente secciones completas del trabajo sin revisión, corrección y reescritura personal.

8.3.2 Tipos de consultas realizadas

Las consultas a herramientas de IA (principalmente **ChatGPT**, modelo GPT-5, y la documentación de **Scikit-Learn**, **SHAP** y **Power BI**) se centraron en los siguientes aspectos:

Tipo de consulta	Ejemplo de uso	Aplicación práctica
Explicación de conceptos	"¿Qué es un ColumnTransformer en Scikit-Learn y para qué sirve?"	Entender los pasos del pipeline IA.
Apoyo metodológico	"Cómo presentar un cronograma tipo Gantt en un proyecto de datos."	Redactar el Capítulo 6 con formato CEI.
Revisión de estilo	"Revisa este párrafo técnico y hazlo más claro sin cambiar el sentido."	Mejorar redacción y coherencia del documento.
Interpretación de resultados	"Cómo se interpreta un RMSE de 0.81 en un modelo de regresión."	Apoyar la explicación de las métricas en Cap. 5.2.
Orientación Power BI	"Cómo conectar un archivo CSV a Power BI y crear una visualización básica."	Comprender la fase de reporting.

Estas interacciones se utilizaron como **herramientas pedagógicas**, similares a la consulta de un tutor virtual o la lectura de manuales técnicos.

8.3.3 Herramientas IA y librerías utilizadas

Herramienta / Fuente	Rol dentro del proyecto	Tipo de licencia
ChatGPT (OpenAI)	Tutor virtual para redacción técnica y dudas conceptuales.	Propietaria / OpenAI.
SHAP (SHapley Additive exPlanations)	Análisis de explicabilidad de modelos predictivos.	MIT License.
LightGBM	Algoritmo adicional de regresión supervisada (modelo estrella).	MIT License.
Scikit-Learn	Entrenamiento y evaluación de modelos (ML clásico).	BSD License.

Power BI	Plataforma de visualización ejecutiva y presentación de KPIs.	Microsoft Software License.

8.3.4 Enfoque ético y educativo

El trabajo siguió las directrices CEI sobre el uso ético de inteligencia artificial en el ámbito académico:

1. **IA como herramienta educativa**, no como generadora de contenido final.
2. **Reescritura humana obligatoria** de toda la información generada.
3. **Verificación manual** de resultados técnicos, métricas y código.
4. **Citación explícita** de las herramientas utilizadas en este anexo.

El objetivo fue **aprender a usar la IA como guía de estudio**, no como sustituto de la autoría humana.

8.3.5 Reflexión personal

La incorporación de IA en el proceso de aprendizaje permitió a la autora mejorar su comprensión sobre conceptos de machine learning y análisis de datos.
El uso de ChatGPT sirvió como **apoyo para estructurar ideas, resolver dudas técnicas y mejorar la redacción**, manteniendo siempre la revisión crítica y la interpretación personal.

Se concluye que, utilizada de forma ética, la IA puede ser una **herramienta formativa valiosa** para estudiantes y profesionales del ámbito de la analítica y la inteligencia artificial, promoviendo la autonomía y la calidad del aprendizaje.