Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

TENDENCIAS DE CONSUMO EN MODA: UN ENFOQUE PREDICTIVO DESDE LA INTELIGENCIA ARTIFICIAL

JULIETH TATIANA HERNANDEZ INFANTE

Dirigido por

Dr. CARLOS WOLFRAM ROZAS

CURSO 2024-2025

TÍTULO: TENDENCIAS DE CONSUMO EN MODA: UN ENFOQUE PREDICTIVO DESDE LA INTELIGENCIA ARTIFICIAL

AUTOR: JULIETH HERNANDEZ INFANTE

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)]

DIRECTOR/ES DEL PROYECTO: Dr. CARLOS WOLFRAM ROZAS

FECHA: SEPTIEMBRE 2025

RESUMEN

El presente trabajo se centra en el diseño y desarrollo de un modelo de aprendizaje automático supervisado para predecir el éxito comercial de productos en el sector de la moda, tomando como base un conjunto de datos transaccionales con variables socioculturales. La problemática abordada responde a la necesidad del sector por anticiparse a las preferencias del consumidor y optimizar la toma de decisiones estratégicas en un entorno altamente competitivo y dinámico.

Mediante el uso de herramientas de ciencia de datos y técnicas de minería de datos, se llevó a cabo un proceso completo que incluyó la limpieza y transformación del dataset, la selección de variables relevantes y la implementación de múltiples algoritmos de clasificación supervisada, tales como KNN, Random Forest, AdaBoost, entre otros. El modelo KNN obtuvo el mejor rendimiento en términos de precisión, exactitud y recall.

Los resultados del proyecto evidencian la viabilidad de predecir con alta efectividad qué productos tienen mayores probabilidades de éxito, lo cual permite a las empresas del sector optimizar inventarios, ajustar campañas promocionales y priorizar productos con alto potencial. Además, se identificaron variables clave como el precio, el descuento y la cantidad, que influyen significativamente en la probabilidad de éxito comercial.

Este trabajo contribuye a integrar inteligencia artificial y análisis predictivo en la industria de la moda, y sienta las bases para futuras ampliaciones mediante datos no estructurados, modelos más complejos y aplicaciones en entornos reales.

Palabras clave: aprendizaje automático, predicción de tendencias, industria de la moda, clasificación supervisada.

ABSTRACT

This project focuses on the design and development of a supervised machine learning model to predict the commercial success of fashion products, using transactional datasets enriched with sociocultural variables. The problem addressed arises from the fashion industry's growing need to anticipate consumer preferences and optimize strategic decision-making in a highly dynamic and competitive market.

Through data science tools and data mining techniques, the project involved a full pipeline: data cleaning and transformation, selection of relevant variables, and the implementation of several supervised classification algorithms, including K-Nearest Neighbors (KNN), Random Forest, AdaBoost, and others. Among these, the KNN model demonstrated the highest performance in terms of precision, accuracy, and recall.

The results confirm the feasibility of effectively predicting which products are more likely to succeed commercially, enabling companies to optimize inventory management, adjust marketing strategies, and prioritize high-potential items. Key influencing variables such as unit price, discount, and quantity were identified as significant predictors of success.

This work contributes to the integration of artificial intelligence and predictive analytics in the fashion industry and establishes a foundation for future extensions, such as the incorporation of unstructured data, the use of deep learning models, or the deployment of real-time applications in business environments.

Keywords: machine learning, trend prediction, fashion industry, supervised classification.

AGRADECIMIENTOS

En primer lugar, deseo expresar mi agradecimiento a Dios, por brindarme la claridad, disciplina y perseverancia necesarias para culminar con éxito este trabajo.

Agradezco de manera especial a mi familia, cuyo respaldo constante fue fundamental durante todo el proceso académico. Su apoyo emocional y motivación resultaron clave para mantener la constancia en las etapas más exigentes del proyecto.

Extiendo mi reconocimiento a mi tutor académico, por su orientación técnica, disponibilidad y aportes críticos, los cuales contribuyeron significativamente al rigor metodológico y a la calidad del presente trabajo.

Asimismo, agradezco a los docentes del programa por proporcionar una base sólida en ciencia de datos y aprendizaje automático, que me permitió abordar este proyecto con una perspectiva analítica y estructurada.

Finalmente, valoró el intercambio académico y el apoyo colaborativo de mis compañeros, quienes enriquecieron el proceso mediante discusiones constructivas, recomendaciones y resolución conjunta de problemas técnicos.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Julieth Hernandez Infante
Título del proyecto:	Tendencias de consumo en moda: un enfoque predictivo desde la inteligencia artificial
Directores del proyecto:	Dr. Carlos Wolfram Rozas
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	El objetivo del presente trabajo es diseñar un modelo de aprendizaje automático basado en algoritmos de clasificación supervisada, que permita predecir el éxito de las tendencias de consumo en moda a partir de variables culturales y sociales, con el fin de optimizar la toma de decisiones estratégicas y maximizar los ingresos en el sector.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El presente proyecto se enmarca en la transformación digital de la industria de la moda, un sector altamente influenciado por dinámicas socioculturales y tendencias volátiles. En este entorno cambiante, la capacidad de anticipar qué productos serán exitosos es un desafío estratégico clave. Con el crecimiento del análisis de datos y la inteligencia artificial, surge una oportunidad valiosa para aplicar técnicas de aprendizaje automático que mejoren la planificación, la producción y la comercialización en este sector.

1.2 Planteamiento del problema

Este trabajo busca dar respuesta a la pregunta: ¿Cómo puede la inteligencia artificial, a partir de variables culturales y sociales, predecir el éxito de tendencias emergentes en el consumo de moda y contribuir a la toma de decisiones estratégicas en el sector? A través del análisis de datos transaccionales y el uso de modelos supervisados, se plantea construir un sistema predictivo que brinde valor real a las empresas del sector moda.

1.3 Objetivos del proyecto

Este Trabajo de Fin de Máster tiene como objetivo desarrollar un modelo predictivo basado en algoritmos de aprendizaje automático supervisado, capaz de anticipar el éxito de las tendencias de consumo en moda a partir de variables culturales y sociales. El proyecto busca, además, identificar los factores socioculturales más influyentes en las decisiones de compra, comparar el desempeño de distintos algoritmos de clasificación y generar recomendaciones estratégicas que contribuyan a la toma de decisiones informadas y a la maximización de ingresos en el sector.

1.4 Resultados obtenidos

Se entrenaron y compararon seis algoritmos de clasificación, siendo el modelo K-Nearest Neighbors (KNN) el de mejor desempeño (recall del 99% y exactitud del 98%). Además, se identificaron variables clave como el precio unitario, el descuento y la cantidad, que influyen directamente en el éxito comercial de un producto. Estos

hallazgos permiten generar recomendaciones para la optimización de inventario, campañas de marketing y alineación con preferencias socioculturales.

1.5 Estructura de la memoria

El documento se organiza en seis secciones que recorren, de forma secuencial, desde el planteamiento inicial del problema hasta las conclusiones y perspectivas posteriores.

El primer capítulo contextualiza el estudio dentro de los cambios culturales y sociales que inciden en el consumo de moda, resaltando la necesidad de anticipar estas dinámicas mediante inteligencia artificial. Se destacan tanto los aportes académicos en el ámbito de la ciencia de datos como la utilidad práctica para la industria, y se plantean los objetivos que orientan la investigación.

El segundo capítulo corresponde al estado del arte, donde se examinan investigaciones previas relacionadas con el uso de la inteligencia artificial en la moda. Este repaso permite reconocer los principales avances alcanzados, al tiempo que deja en evidencia vacíos que justifican el presente trabajo.

En el tercer capítulo se detalla la metodología, la cual incluye la caracterización del conjunto de datos empleado, las etapas de limpieza y normalización de variables, y la aplicación de diferentes algoritmos de clasificación supervisada junto con sus criterios de validación.

El cuarto capítulo se centra en la implementación de los modelos predictivos, describiendo las fases del análisis y el proceso seguido en la construcción de los algoritmos. Se detallan los pasos de entrenamiento y validación, junto con la evaluación de métricas que permitieron contrastar el rendimiento de las distintas técnicas aplicadas. Asimismo, se presentan los resultados obtenidos a lo largo de este proceso, los cuales sirvieron de base para determinar el modelo más adecuado en el contexto de la predicción de tendencias de consumo en moda.

El quinto capítulo está dedicado a la discusión de resultados, donde se presentan e interpretan los principales hallazgos en relación con los objetivos planteados y se analizan sus implicaciones tanto en el ámbito académico como en la práctica

empresarial. Además, se señalan las limitaciones identificadas durante el proceso, lo que permite dimensionar adecuadamente el alcance del estudio.

Por último, el sexto capítulo presenta las conclusiones generales del trabajo y los aprendizajes alcanzados durante su desarrollo. Además, expone recomendaciones y posibles líneas futuras de investigación orientadas a profundizar en el uso de la inteligencia artificial para la predicción de tendencias en el sector moda.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

En un contexto donde la hiperconectividad y la transformación acelerada de las dinámicas culturales son predominantes, prever las tendencias de consumo en la industria de la moda representa un desafío relevante tanto para el ámbito académico como empresarial. Esta preocupación ha motivado un notable incremento en los estudios que abordan el uso de la inteligencia artificial (IA), el análisis de grandes volúmenes de datos y la influencia de las redes sociales para identificar patrones emergentes.

A continuación, se presenta una revisión crítica del conocimiento existente sobre predicción de tendencias de moda, agrupado en cinco líneas principales: (1) inteligencia artificial aplicada a la moda, (2) influencia de redes sociales en la configuración de tendencias, (3) modelos predictivos con variables socioculturales, (4) aplicaciones de Deep Learning y Generative AI, (5) casos empresariales reales y (6) Limitaciones actuales y líneas emergentes. Cada sección identifica avances significativos y limitaciones aún no resueltas, lo que justifica la necesidad de propuestas más integradoras y aplicables al contexto actual.

2.1 Estado del arte

En este marco, se expone a continuación una revisión crítica del conocimiento existente, con el propósito de señalar los principales aportes, limitaciones y áreas con potencial para futuras investigaciones.

1. **Inteligencia artificial aplicada a la predicción de tendencias de moda**

En los últimos veinte años, la inteligencia artificial ha ido adquiriendo un papel cada vez más relevante en la industria de la moda, transformando significativamente la forma en que las empresas diseñan, producen y comercializan sus productos.El estudio de Suvarna y Balakrishna (2024), publicado en la revista científica internacional Fashion and Textiles, respalda que la combinación de modelos de aprendizaje profundo, a través de técnicas de ensamblado y transferencia de conocimiento, permite optimizar los sistemas de recomendación visual en el ámbito de la moda, favoreciendo una mayor personalización y precisión en la identificación de productos dentro de catálogos extensos.

Su implementación ha sido especialmente destacada en áreas como la predicción de la demanda, donde los algoritmos permiten anticipar las preferencias de los consumidores y ajustar la producción para evitar el sobrestock o la escasez de productos.

Asimismo, los sistemas de recomendación personalizados se han convertido en una herramienta clave para mejorar la experiencia del cliente, ya que analizan patrones de comportamiento y datos históricos para ofrecer sugerencias precisas y adaptadas a los gustos individuales. Esta evolución tecnológica no solo ha optimizado los procesos internos, sino que también ha redefinido la relación entre las marcas y los consumidores, fomentando una industria más ágil, eficiente y centrada en el usuario.

En esta línea, diversas investigaciones han propuesto soluciones innovadoras que refuerzan el papel de la inteligencia artificial en la gestión de la cadena de suministro. (Kaaij, 2020) plantea un modelo basado en algoritmos de aprendizaje automático que permite anticipar la demanda en el sector minorista de moda, demostrando que la incorporación de datos históricos de ventas junto con factores estacionales puede contribuir significativamente a una gestión de inventario más eficiente.

Sin embargo, el uso de estas tecnologías ha comenzado a expandirse hacia terrenos menos explorados, como la identificación y predicción de tendencias estéticas y de estilo. En este contexto, el estudio de (Dadoun,2017) resulta especialmente relevante, ya que investiga el potencial del aprendizaje supervisado y no supervisado, aplicado a bases de datos visuales y a contenidos generados en plataformas digitales, destacando el valor del análisis automatizado de imágenes para detectar patrones emergentes en las preferencias de los consumidores.

En la misma línea, (Grammenos,2020) presenta una propuesta innovadora que combina herramientas de visión por computadora con técnicas de procesamiento del lenguaje natural, con el objetivo de detectar microtendencias emergentes de manera inmediata. Este enfoque interdisciplinario permite identificar indicios sutiles dentro de grandes cantidades de datos no estructurados, como opiniones de usuarios, descripciones de artículos o imágenes de moda urbana, y convertirlos en información útil para la toma de decisiones. La capacidad de interpretar y reaccionar rápidamente a

estos cambios incipientes se traduce en una ventaja competitiva significativa en un sector donde la rapidez y la adaptación son esenciales.

Por otro lado, (Liao,2020) introduce un modelo de redes neuronales mejorado mediante la incorporación de conocimiento contextual, conocido como Knowledge-Enhanced Neural Networks. Esta arquitectura híbrida permite al sistema integrar información sociocultural relevante como referencias históricas, simbología cultural o contextos regionales, enriqueciendo significativamente su capacidad predictiva. Gracias a este tipo de avances, la previsión de moda comienza a acercarse a un entendimiento más profundo y matizado del comportamiento estético de las audiencias, permitiendo una toma de decisiones más sensible, informada y proactiva.

Complementando este enfoque basado en el conocimiento contextual, (Choi,2024) analiza cómo los influencers, los hashtags y la dimensión temporal actúan como catalizadores clave en la identificación y propagación de nuevas tendencias dentro del entorno digital. A través de un análisis de imágenes de moda provenientes de múltiples fuentes incluyendo redes sociales, pasarelas y plataformas de retail, el estudio demuestra que las publicaciones de influencers no sólo reflejan, sino que también anticipan movimientos estéticos emergentes.

Asimismo, el uso estratégico de hashtags permite rastrear temáticas populares en evolución, mientras que la dimensión temporal ofrece una comprensión más clara del ritmo con que las tendencias se adoptan y se difunden. Esta visión integrada aporta una perspectiva dinámica a la predicción de moda, al incorporar tanto el contenido visual como el contexto socio temporal en el que se desarrolla.

En conjunto, estas propuestas evidencian un cambio profundo en la forma en que la industria de la moda enfrenta la predicción de tendencias. La combinación de herramientas tecnológicas avanzadas con una lectura contextual de factores sociales y culturales da lugar a sistemas de análisis más precisos y proactivos. Este desarrollo responde a la creciente necesidad de las marcas de adaptarse a un mercado dinámico, donde la capacidad de identificar las señales adecuadas en el momento justo resulta crucial para mantenerse competitivas. En este escenario, la inteligencia

artificial se consolida como un recurso estratégico esencial para optimizar los procesos creativos, productivos y comunicativos, permitiendo respuestas más ágiles, personalizadas y culturalmente alineadas.

2. Influencia de las redes sociales en la configuración de tendencias

Las redes sociales han evolucionado de simples canales de interacción a potentes herramientas de generación, propagación y validación de tendencias. En el ámbito de la moda, su influencia es particularmente notable debido a la rapidez con la que logran posicionar estilos, prendas o estéticas específicas entre millones de usuarios.

Esta capacidad de viralización inmediata no sólo transforma los ciclos tradicionales del consumo, sino que también configura nuevamente la forma en que las marcas interpretan y responden a los deseos del público. Además, plataformas como Instagram, TikTok o Pinterest actúan como espacios de experimentación visual, donde tanto creadores como consumidores participan activamente en la construcción de lo que será considerado “tendencia”.

En este contexto de construcción visual colectiva, una investigación reciente publicada en el *International Journal of Consumer Studies* analiza la forma en que la Generación Z responde a los anuncios de moda generados mediante inteligencia artificial. El estudio concluye que aspectos como la autenticidad, la credibilidad del mensaje y la identificación cultural tienen un peso significativo en la manera en que estos contenidos son recibidos por el público joven. Esto evidencia que el éxito de las tendencias en redes sociales no se basa únicamente en su apariencia visual, sino también en la capacidad del contenido para conectar emocional y simbólicamente con los usuarios (Lou, X., & Copeland, L, 2025).

El estudio de Majeed (2020) muestra cómo las plataformas digitales funcionan como espacios donde se genera inspiración colectiva, permitiendo seguir la evolución de las modas a través de la frecuencia con la que se mencionan y cómo cambian visualmente ciertas prendas. Por ejemplo, su investigación sobre la gorra de béisbol explica cómo este accesorio específico puede pasar de ser presentado en las pasarelas a convertirse en un elemento popular en el consumo masivo, gracias a su difusión en redes sociales como Instagram y TikTok.

En esta misma línea, el modelo desarrollado por (Qian,2021) examina las complejas interacciones que se generan entre usuarios, publicaciones y categorías estéticas en entornos digitales, a través de un sistema avanzado de representación semántica. Esta herramienta no solo permite identificar patrones latentes en el comportamiento en línea, sino también revelar relaciones ocultas entre fenómenos de moda emergentes que podrían pasar desapercibidos con enfoques más superficiales.

La propuesta enfatiza que el análisis del contenido explícito como imágenes, descripciones o etiquetas resulta insuficiente para comprender plenamente la dinámica de viralización. Por ello, incorpora dimensiones adicionales como los vínculos sociales entre usuarios, la evolución temporal de las publicaciones y el contexto cultural o geográfico en el que se producen las interacciones. Al integrar estos factores, el modelo ofrece una comprensión más rica y matizada de cómo se configuran y difunden las tendencias en plataformas digitales, permitiendo incluso anticipar futuras direcciones en el ámbito de la moda y el consumo cultural.

Estas investigaciones destacan que, al analizar redes sociales, no basta con aplicar técnicas de minería de texto o imagen de forma aislada. Si bien estas herramientas permiten extraer información relevante a nivel superficial como etiquetas, palabras clave, estilos visuales o descripciones, resultan insuficientes para comprender los procesos complejos que subyacen a la formación y propagación de tendencias.

Es necesario incorporar estructuras relacionales y datos temporales que revelen cómo interactúan los distintos actores dentro del ecosistema digital. Esto implica considerar las redes de conexiones entre usuarios, los patrones de interacción (como comentarios, compartidos o menciones), así como la evolución cronológica de los contenidos, ya que muchas modas no surgen de manera repentina, sino que se gestan de forma progresiva a través de ciclos de retroalimentación social.

Además, integrar estos elementos permite modelar con mayor precisión los mecanismos de influencia, la difusión de ideas estéticas y la emergencia de comunidades de gusto, ofreciendo una visión más holística y dinámica del entorno digital. Así, el análisis deja de ser una simple recopilación de datos estáticos y se convierte en una herramienta para comprender procesos culturales en constante transformación

3. Modelos predictivos basados en variables socioculturales

Un ámbito aún poco abordado, pero con alto potencial, es la inclusión de variables socioculturales en los modelos de predicción de tendencias de moda. La mayoría de las investigaciones analizadas tienden a enfocarse en datos visuales, antecedentes históricos o interacciones en redes sociales como imágenes, etiquetas o dinámicas de usuario, dejando en segundo plano aspectos clave como los contextos culturales, las condiciones económicas o las características demográficas de las audiencias.

No obstante, estos factores socioculturales pueden ser determinantes para comprender cómo y por qué ciertas tendencias se consolidan o fracasan en distintos entornos. Elementos como la clase social, el género, el nivel educativo, los valores culturales o los movimientos migratorios inciden directamente en la forma en que diversos grupos sociales adoptan o reinterpretan determinados estilos estéticos.

La integración de estas variables enriquecería los modelos existentes, permitiendo una predicción más precisa y contextualizada, al reconocer que la moda no solo responde a patrones visuales o tecnológicos, sino que está profundamente arraigada en construcciones sociales e identitarias. Este enfoque más amplio y transversal abriría nuevas oportunidades de análisis en el cruce entre ciencia de datos, estudios culturales y ciencias sociales, proporcionando una visión más profunda y completa del fenómeno.

No obstante, investigaciones más recientes han empezado a incorporar estas variables socioculturales en sus modelos analíticos. Un ejemplo destacado es el sistema Neo-Fashion, desarrollado por Li Zhao (2020), que propone una metodología innovadora al combinar datos provenientes de desfiles de moda con un análisis contextual de estilos vinculados a regiones geográficas específicas, estaciones del año y perfiles demográficos de consumidores.

Este enfoque permite reconocer que las tendencias no se difunden de manera uniforme a nivel global, sino que su adopción está mediada por códigos culturales, climas locales, hábitos de consumo y significados sociales particulares. En lugar de asumir una lógica universal en la propagación de la moda, el modelo de (Li Zhao, 2020) evidencia cómo determinados estilos pueden resonar más fuertemente en

ciertos entornos que en otros, revelando patrones de localización y segmentación del gusto.

Otro aporte relevante proviene del trabajo de (Tang et al. 2020), quienes desarrollan un marco conceptual basado en inteligencia artificial orientado a la predicción de tendencias en moda, incorporando datos estructurados de carácter socioeconómico y demográfico. Este enfoque es validado a través de un estudio de caso que demuestra su aplicabilidad en contextos reales.

Aunque se encuentra aún en una etapa inicial de desarrollo, esta propuesta representa un avance significativo hacia una visión más holística del análisis de tendencias, al combinar el rigor de los modelos computacionales con la riqueza interpretativa de las ciencias sociales. Al integrar variables como nivel de ingresos, densidad poblacional, edad o educación, se logra captar con mayor precisión cómo distintos segmentos sociales responden a las dinámicas del cambio estético y del consumo.

Comprender las dinámicas actuales de la moda exige trascender los enfoques centrados únicamente en lo visual, lo algorítmico o lo puramente cuantitativo. Las tendencias no emergen en un vacío, sino que se configuran en un entramado complejo de valores culturales, significados sociales y condiciones económicas específicas. Por ello, es fundamental adoptar marcos interpretativos que consideren la diversidad de contextos en los que la moda se produce, circula y se transforma.

En este sentido, los modelos predictivos más innovadores no se limitan al análisis de datos visuales o de comportamiento en redes, sino que integran variables socioeconómicas, demográficas y culturales que permiten una lectura más matizada del fenómeno. Esta visión amplia se materializa en enfoques híbridos que combinan herramientas avanzadas de inteligencia artificial con conceptos procedentes de disciplinas como la sociología, la antropología o la economía cultural.

La articulación entre estos campos no solo enriquece la capacidad explicativa de los modelos, sino que también posibilita una mejor adaptación a contextos locales y globales, considerando las distintas formas en que el gusto, el consumo y la identidad se manifiestan. Al generar predicciones más sensibles a la heterogeneidad social y a las dinámicas multiculturales, estos sistemas contribuyen no solo a una comprensión

más profunda del comportamiento del consumidor, sino también al diseño de estrategias más inclusivas, éticas y culturalmente informadas dentro de la industria de la moda.

4. Aplicación de Deep Learning y Generative AI en la industria de la moda

El avance del deep learning y la inteligencia artificial generativa (Generative AI) ha transformado profundamente el panorama analítico y creativo en la industria de la moda. Estas tecnologías han superado los límites tradicionales del diseño, ofreciendo soluciones innovadoras para tareas como la creación automatizada de imágenes, la elaboración de propuestas estilísticas originales y la anticipación de tendencias. Gracias a su capacidad para aprender de grandes volúmenes de datos visuales y semánticos, los sistemas de IA pueden generar resultados altamente precisos y personalizados, abriendo nuevas posibilidades tanto para diseñadores como para marcas.

En esta línea (Faith , 2023) presenta una revisión detallada del impacto de los modelos generativos especialmente las Redes Generativas Antagónicas (GANs) en el desarrollo de colecciones de moda. Su análisis destaca cómo estas herramientas son capaces de producir representaciones visuales de prendas, outfits y estilos que aún no han sido fabricados, lo que resulta de gran utilidad para procesos de ideación, experimentación y validación anticipada. Al permitir la simulación de escenarios estéticos futuros, estas tecnologías se convierten en instrumentos clave para predecir la dirección que podrían tomar las próximas temporadas, minimizando riesgos y acelerando la toma de decisiones estratégicas.

Además de facilitar pruebas de concepto sin necesidad de producción física, estos modelos también fomentan una creatividad más abierta e iterativa. Los diseñadores pueden explorar múltiples variaciones de un mismo concepto en cuestión de segundos, lo que optimiza tiempos de desarrollo y permite una mayor adaptación a las demandas cambiantes del mercado. Así, la IA generativa no sólo redefine los procesos técnicos del diseño, sino que también impulsa una nueva relación entre creatividad, tecnología y estrategia en el ámbito de la moda contemporánea.

Paralelamente, estudios como el de (Chen, 2020) evidencian que la aplicación de redes neuronales profundas puede potenciar significativamente la capacidad de los sistemas automatizados para reconocer patrones visuales sutiles en grandes cantidades de datos, como imágenes provenientes de desfiles de moda o contenidos compartidos en redes sociales. Esta capacidad es especialmente valiosa para la detección temprana de microtendencias, es decir, señales débiles de estilo que podrían evolucionar en fenómenos más amplios dentro del mercado. La identificación oportuna de estos indicios permite a las marcas anticiparse a los cambios en las preferencias del consumidor y responder con mayor agilidad.

Sin embargo, a pesar de sus ventajas, estos enfoques tecnológicos también conllevan una serie de desafíos éticos, creativos y operativos. Entre ellos se encuentran la generación automatizada de diseños que podrían infringir derechos de autor, la potencial pérdida de originalidad debido a la homogeneización algorítmica de los estilos, y la dependencia creciente de modelos que, aunque eficientes, carecen muchas veces de una comprensión semántica profunda del contexto cultural en el que operan.

Por ello, la implementación efectiva de estas tecnologías en entornos reales no puede entenderse como un proceso meramente técnico. Requiere una articulación interdisciplinaria que incluya no solo a expertos en inteligencia artificial, sino también a diseñadores, especialistas en moda, y analistas culturales. Esta colaboración es clave para garantizar que las soluciones desarrolladas no solo sean técnicamente viables, sino también culturalmente significativas, éticamente responsables y estéticamente relevantes. Integrar la dimensión humana en el proceso algorítmico es, por tanto, esencial para que la innovación en la moda no pierda su carácter expresivo, creativo y social.

5. Aplicaciones empresariales de IA en la moda: casos y modelos organizativos

Más allá del ámbito estrictamente académico, la inteligencia artificial se ha consolidado como un recurso estratégico dentro de la industria textil, especialmente entre empresas líderes que buscan reforzar su competitividad a través de la digitalización integral de sus procesos. Un ejemplo paradigmático es el caso de Zara, documentado

por (Gómez, 2024), que demuestra cómo la incorporación de tecnologías basadas en IA puede transformar profundamente la cadena de valor en el sector moda.

Zara ha implementado un sistema de análisis predictivo que integra múltiples fuentes de información, como históricos de ventas, actividad en redes sociales y patrones de comportamiento del consumidor, para optimizar decisiones clave en áreas como el diseño de productos, la planificación de inventario y la logística de distribución. Esta arquitectura de datos facilita una lectura en tiempo real de la demanda y de las preferencias emergentes, permitiendo una respuesta mucho más ágil y ajustada al mercado. Entre los beneficios tangibles se destacan una gestión más eficiente del aprovisionamiento, una reducción significativa en los ciclos de desarrollo y distribución, y una oferta altamente personalizada que responde con precisión a las dinámicas de consumo actuales.

En una línea complementaria, el informe técnico de (Neo-Fashion, 2020) propone una arquitectura de sistema inteligente orientada a la predicción de tendencias, basada en la fusión de datos estructurados (como estadísticas de ventas y metadatos de producto) y no estructurados (como imágenes de desfiles, publicaciones en redes sociales o archivos visuales). Esta aproximación multimodal permite a las marcas no solo monitorizar el presente, sino también anticipar escenarios estilísticos futuros con mayor nivel de precisión, lo que se traduce en una ventaja competitiva clave en un entorno cada vez más acelerado, saturado de información y volátil.

Ambos casos ilustran una transformación profunda en el modo en que se concibe y gestiona el negocio de la moda. La inteligencia artificial ha dejado de ser una herramienta exclusivamente analítica para convertirse en un componente central del modelo organizativo. Desde la planificación estratégica hasta la experiencia del cliente, la IA está redefiniendo no solo los procesos operativos, sino también las lógicas creativas, las estrategias de posicionamiento y la relación emocional con los consumidores.

Su impacto trasciende la eficiencia, habilita nuevas formas de innovación creativa, contribuye a la sostenibilidad mediante una gestión más inteligente del inventario y reduce el sobrestock, al tiempo que potencia experiencias de consumo personalizadas a través de recomendaciones automatizadas y comunicación adaptativa. Asimismo, la

IA permite acortar los ciclos de diseño y producción, facilitando procesos de Innovación colaborativa entre marcas y consumidores, lo que resulta clave para responder a las demandas de inmediatez del mercado actual

Mientras que en el fast fashion la IA se orienta al rendimiento operativo, marcas como Gucci han encontrado en la inteligencia artificial generativa una vía para explorar nuevas dimensiones estéticas y culturales. En lugar de limitarse al uso funcional de esta tecnología, la firma italiana la ha incorporado como herramienta expresiva y narrativa, capaz de enriquecer su identidad visual y su conexión emocional con el público

Un caso emblemático es la instalación de murales digitales interactivos en su tienda de Chengdu, China, donde la IA se emplea para reinterpretar elementos visuales inspirados en el Renacimiento italiano. Esta intervención, analizada por (Rossi,2023), representa una propuesta en la que la inteligencia artificial no sustituye la creatividad humana, sino que la amplifica, generando nuevas formas simbólicas de interacción entre marca, espacio y cultura. Así, Gucci reafirma su posición como referente en la convergencia entre moda, arte y tecnología, y aporta una experiencia de compra emocionalmente rica y culturalmente significativa.

En contraste, Stitch Fix representa una aplicación más funcional y orientada al servicio personalizado mediante algoritmos de recomendación. Esta empresa estadounidense ha construido su modelo de negocio sobre la base del análisis de datos personales como tallas, historial de compras y devoluciones junto con variables subjetivas, como preferencias estéticas, estilo de vida o contexto de uso. Esta integración le permite generar propuestas de vestuario altamente individualizadas.

Más recientemente, Stitch Fix ha comenzado a incorporar modelos predictivos capaces de anticipar con meses de antelación las preferencias de sus usuarios, lo que optimiza la planificación de colecciones y permite una adaptación proactiva a las tendencias del mercado (Vogue Business, 2023). Su enfoque demuestra cómo la IA puede actuar como puente entre el comportamiento del consumidor y la toma de decisiones estratégicas, con impactos directos en fidelización, eficiencia comercial y personalización masiva.

No obstante, esta transformación implica también retos significativos. La adopción efectiva de estas tecnologías requiere no solo inversiones sostenidas en infraestructura tecnológica como plataformas de big data, computación en la nube o herramientas de machine learning, sino también talento interdisciplinario capaz de traducir datos complejos en decisiones accionables. A ello se suma la necesidad de un cambio organizacional profundo, en el que la cultura de decisión basada en datos conviva armónicamente con los procesos creativos, sin desplazar la intuición, la sensibilidad artística ni el conocimiento tácito que caracterizan al sector moda.

En este contexto, las empresas que logren articular eficazmente tecnología, visión de negocio y sensibilidad cultural serán las que lideren la próxima etapa de evolución del sector. No se trata solo de aumentar la eficiencia, sino de crear propuestas de valor relevantes, sostenibles y culturalmente significativas en un entorno global hiperconectado. Además, este nuevo paradigma abre interrogantes para marcas emergentes, que deberán encontrar formas creativas de incorporar estas herramientas sin comprometer su identidad ni su autenticidad, en un mercado donde la innovación tecnológica y la autenticidad simbólica coexisten como expectativas del consumidor contemporáneo.

6. Limitaciones actuales y líneas emergentes

A pesar de los avances recientes en la aplicación de inteligencia artificial para la predicción de tendencias en el sector moda, el campo sigue enfrentando limitaciones estructurales que restringen su evolución y utilidad práctica. Una de las principales debilidades es la fragmentación metodológica existente entre los distintos enfoques de investigación.

Mientras algunos trabajos se centran exclusivamente en el análisis de imágenes por ejemplo, mediante reconocimiento visual de estilos o colores predominantes, otros se enfocan en el procesamiento de lenguaje natural aplicado a textos provenientes de medios digitales o redes sociales. A su vez, hay estudios que exploran únicamente el comportamiento en plataformas como Instagram o TikTok. Sin embargo, son escasos los intentos por integrar de manera coherente y multidimensional estas fuentes de datos heterogéneas, lo cual limita la capacidad de construir modelos realmente representativos del complejo ecosistema de la moda contemporánea.

Además, muchos de los modelos existentes se caracterizan por adoptar un enfoque descriptivo o correlacional, centrado en observar asociaciones entre variables sin llegar a ofrecer una capacidad predictiva robusta ni una comprensión profunda de las dinámicas que explican la evolución de las tendencias. Esta limitación metodológica reduce la utilidad práctica de los modelos en contextos donde se requiere anticipar con precisión comportamientos futuros del mercado, como en la planificación de colecciones, estrategias de diseño o campañas de marketing.

Otro aspecto crítico es la ausencia de marcos teóricos sólidos que provengan de disciplinas como la sociología, la antropología o los estudios culturales. La mayoría de los modelos no incorporan explícitamente teorías que expliquen por qué ciertas estéticas, discursos o prácticas se convierten en tendencias dominantes en determinados contextos sociales o culturales.

Esta carencia teórica impide interpretar los datos más allá de su superficie y limita el potencial explicativo de las herramientas basadas en IA. Integrar este tipo de enfoques permitiría avanzar hacia modelos más completos, capaces no solo de identificar patrones, sino también de ofrecer una lectura crítica de los factores simbólicos, ideológicos y estructurales que moldean el consumo de moda.

Finalmente, la cuestión de la transparencia y aplicabilidad de los algoritmos sigue siendo una barrera significativa para su adopción en el entorno empresarial. En muchos casos, los modelos de inteligencia artificial son percibidos como “cajas negras”, lo que genera desconfianza entre los tomadores de decisiones, especialmente en sectores como el diseño, el marketing o la gestión de inventarios, donde es imprescindible justificar con claridad las decisiones estratégicas.

La falta de interpretabilidad no solo limita la confianza de las empresas en las herramientas tecnológicas, sino que también restringe su capacidad para adaptarse a contextos específicos o para ser utilizadas en procesos colaborativos con diseñadores, creativos y otros actores clave del sector (Faith, 2023).

En este contexto, se abre una oportunidad estratégica para el desarrollo de propuestas metodológicas más integradoras y transparentes, que combinen el rigor computacional con una comprensión crítica y contextualizada del fenómeno de la moda. Este tipo de

enfoques híbridos podría aportar un valor diferencial tanto a nivel académico como en su aplicación práctica dentro de la industria.

En línea con esta necesidad de enfoques más integradores, resulta pertinente realizar una comparación sistemática de los métodos predictivos propuestos en los estudios revisados. La siguiente tabla sintetiza los principales algoritmos utilizados en la predicción de tendencias de moda, especificando el tipo de datos empleados, las métricas de evaluación aplicadas y los aportes o limitaciones observados en cada caso. Esta comparación no solo permite visualizar el panorama actual de soluciones técnicas en el campo, sino que también facilita la identificación de oportunidades para mejorar la precisión, contextualización y aplicabilidad de los modelos predictivos futuros.

Tabla 1

Comparación entre autores

Autor(es)	Año	Algoritmos utilizados	Tipo de datos	Métricas de evaluación	Comentarios clave
Suvarna & Balakrishna	2024	Deep ensemble classifier + Transfer Learning	Imágenes de productos de moda	Precision, Recall, F1-Score	Mejora la personalización en recomendaciones visuales usando aprendizaje profundo
Lou & Copeland	2025	No aplica (estudio cualitativo)	Percepciones de consumidor (entrevistas)	No aplica	La autenticidad y conexión cultural influyen más que el atractivo visual
Kaaij	2020	Random Forest, Regresión logística	Ventas históricas + temporal	Accuracy, MAE	Buen rendimiento para demanda; no integra factores culturales
Dadoun	2017	K-Means, SVM, Redes Neuronales	Imágenes, plataformas sociales	No especificado	Predicción visual útil, sin validación cuantitativa clara
Grammenos	2020	CNN + NLP	Imágenes + textos	Precision, Recall	Detecta microtendencias con buena sensibilidad

Liao	2020	Redes Neuronales con conocimiento contextual	Datos semánticos y culturales	F1-Score	Alta precisión contextual, pero complejidad elevada
Choi	2024	Modelos híbridos de redes + análisis temporal	Imágenes, hashtags, tiempo	Accuracy	Integración eficaz de dinámica temporal-social
Qian	2021	Embeddings semánticos + clasificación	Texto + relaciones entre usuarios	No especificado	Capta relaciones complejas y latentes; limitado a texto
Li Zhao	2020	Regresión logística + Análisis contextual	Desfiles + ubicación y perfil	Accuracy	Alta capacidad de segmentación local
Tang et al.	2020	Árboles de decisión, Random Forest, MLP	Datos socioeconómicos	Precision, Recall, F1-Score	Combinación de IA + variables sociales con buenos resultados
Faith	2023	GANs (Redes Generativas Antagónicas)	Imágenes de moda	No aplica (generación creativa)	Generación visual, sin foco predictivo directo
Chen	2020	Convolutional Neural Networks (CNN)	Imágenes pasarelas/redes	Accuracy	Reconocimiento visual altamente eficaz
Gómez-Romero	2024	Sistemas de recomendación + modelos de IA	Multifuente estructurado	AUC, Accuracy	Orientado a operación más que a predicción de tendencias

Nota. Elaboración propia

Observaciones

- Árboles de decisión, Random Forest y SVM siguen siendo métodos comunes, con buen balance entre precisión y explicabilidad.
- Redes neuronales profundas (CNN, MLP) ofrecen mayor precisión, especialmente en imágenes, pero requieren gran volumen de datos y computación.
- Modelos híbridos (como los de Liao o Choi) tienen el mayor potencial para integrar contexto cultural, semántica y dinámica social, aunque son más difíciles de implementar.
- GANs y métodos generativos no buscan predecir, sino producir contenido visual; útiles para diseño, menos para análisis estratégico.

2.2 Contexto y justificación

La moda, entendida como un fenómeno cultural, económico y social, desempeña un papel central en la configuración de identidades y dinámicas colectivas. No se limita únicamente al diseño o la estética, sino que refleja los valores, intereses y comportamientos de una sociedad en un momento determinado. En el actual contexto de globalización, digitalización y transformación constante de los hábitos de consumo, la industria de la moda experimenta una evolución acelerada en sus procesos de creación, producción, distribución y comunicación (Gonzáles Cueva, 2024).

El auge del modelo de negocio del fast fashion ha intensificado este dinamismo, promoviendo la producción masiva de prendas de bajo costo y rápida rotación. Este sistema, centrado en la inmediatez y el deseo constante de novedad, obliga a las marcas a renovar colecciones con una frecuencia cada vez mayor, estimulando un consumo impulsivo, de corta duración y muchas veces insostenible. En este escenario, la capacidad de anticipar tendencias con rapidez y precisión se ha convertido en un factor crítico de competitividad.

Las redes sociales, en particular plataformas como TikTok, Instagram y Pinterest, han redefinido los mecanismos de influencia en la moda. Hoy en día, las tendencias no se dictan únicamente desde las pasarelas o revistas especializadas, sino que surgen de la viralidad, de los influencers y de fenómenos culturales espontáneos que se propagan globalmente en cuestión de horas (Branch, 2024). Este fenómeno ha democratizado el acceso a la moda y ha permitido que tanto marcas consolidadas como diseñadores emergentes se posicionen ante audiencias amplias sin intermediarios tradicionales. Sin embargo, también ha generado ciclos de vida más breves para las tendencias y ha intensificado la necesidad de adaptación constante por parte de las empresas.

Esta realidad obliga a replantear las estrategias de producción y comercialización en el sector, priorizando la agilidad, la flexibilidad y la capacidad de leer con precisión los cambios en el comportamiento del consumidor. Tradicionalmente, el análisis de tendencias se ha basado en la observación cualitativa, la intuición de expertos o los informes de agencias especializadas. Si bien estas fuentes siguen siendo valiosas, resultan cada vez más insuficientes para abordar la complejidad del entorno actual.

Por ello, se hace necesario explorar enfoques innovadores basados en el análisis de datos y la inteligencia artificial, que permitan integrar información sociocultural como edad, género, estilos de vida, valores, intereses culturales o hábitos digitales en modelos predictivos capaces de anticipar comportamientos de consumo. El uso de herramientas tecnológicas avanzadas, como el aprendizaje automático, permite detectar patrones emergentes que no son evidentes a simple vista y proporciona una base sólida para la toma de decisiones estratégicas.

Este proyecto se desarrolla, por tanto, en un contexto de cambio acelerado, donde el cruce entre análisis sociocultural, tecnología y moda se convierte en un terreno propiciol para la innovación. Su aporte no se limita al ámbito empresarial, sino que también ofrece una contribución relevante al campo científico-técnico, al proponer un enfoque interdisciplinario que combina sociología, inteligencia artificial y estudios de consumo. Además, responde a una necesidad real del sector: optimizar la toma de decisiones en un mercado volátil, competitivo y digitalmente mediatizado (McKinsey & Company, 2023).

2.3 Planteamiento del problema

En un entorno altamente competitivo como el de la industria de la moda, la capacidad para anticipar gustos, valores y preferencias del consumidor se ha convertido en una prioridad estratégica. La rápida evolución de las tendencias, impulsada por fenómenos culturales globales y amplificada por redes sociales como TikTok o Instagram, obliga a las marcas a adaptarse constantemente a nuevas demandas. Sin embargo, muchas decisiones empresariales siguen basándose en intuiciones, en informes generalistas o en metodologías que no logran capturar la complejidad del comportamiento actual del consumidor.

Este desfase entre la abundancia de datos disponibles y su aplicación efectiva en la toma de decisiones puede traducirse en consecuencias significativas: pérdida de competitividad, desarrollo de colecciones desconectadas de los intereses del público, acumulación de inventario no vendido, campañas de marketing ineficientes o falta de conexión con las narrativas culturales que movilizan a los nuevos consumidores (González, 2023). En particular, la carencia de herramientas capaces de identificar y

predecir tendencias emergentes genera una vulnerabilidad estructural en un mercado donde la agilidad y la personalización son claves.

Aunque industrias como el retail, la banca o la salud ya han adoptado con éxito tecnologías de machine learning y análisis predictivo, el sector moda aún se encuentra en una fase temprana de adopción. Existen avances, como sistemas de recomendación o análisis de sentimientos, pero son escasas las investigaciones que integran variables socioculturales en modelos predictivos para anticipar el éxito de tendencias de consumo. Esta brecha revela una oportunidad de innovación metodológica que combine el análisis cultural con enfoques computacionales.

En respuesta a este vacío, el presente proyecto plantea la aplicación de modelos de inteligencia artificial específicamente algoritmos de clasificación supervisada para predecir el éxito comercial de productos de moda a partir de variables culturales y sociales. Parte del supuesto de que estos factores tienen un peso decisivo en las decisiones de compra y que su análisis sistemático puede traducirse en ventajas competitivas para marcas y diseñadores.

Este trabajo se enmarca en una línea de investigación científico-técnica con un fuerte componente de innovación aplicada. Si bien no surge de un caso empresarial concreto, su aplicabilidad es directa para diseñadores, marcas, analistas de datos y agentes involucrados en la toma de decisiones estratégicas del sector. Su propósito es tender un puente entre el análisis cualitativo tradicional de tendencias y las capacidades cuantitativas del aprendizaje automático, proponiendo un enfoque más preciso, escalable y alineado con las dinámicas del consumo digital y globalizado.

Capítulo 3. OBJETIVOS

3.1 Objetivo general

El objetivo general del presente trabajo consiste en diseñar un modelo de aprendizaje automático supervisado que permita predecir el éxito de las tendencias de consumo en moda, a partir de variables socioculturales extraídas de datos estructurados. Este modelo busca apoyar la toma de decisiones estratégicas en el sector de la moda, proporcionando predicciones basadas en datos que pueden facilitar la identificación anticipada de productos con alto potencial de éxito comercial, contribuyendo así a una planificación más eficaz y potencialmente a una mejora en los ingresos.

3.2 Objetivos específicos

1. Identificar las variables socioculturales más relevantes que influyen en el comportamiento de los consumidores de moda y que contribuyen al éxito de las tendencias, con base en el análisis de un conjunto de datos transaccionales.

2. Desarrollar todas las fases del proceso de minería de datos, incluyendo:
   - Comprensión del negocio.
   - Comprensión de los datos.
   - Preparación de los datos.
   - Modelado predictivo.
   - Evaluación del modelo.
   - Generación de recomendaciones.

3. Implementar y comparar distintos algoritmos de clasificación supervisada, tales como árboles de decisión, random forest, KNN, Adaboost, regresión logística y descenso de gradiente estocástico, con el objetivo de seleccionar el modelo con mejor desempeño en la predicción del éxito de las tendencias.

4. Evaluar el desempeño de los modelos predictivos mediante métricas de clasificación como precisión, exactitud, recall y F1-score, garantizando así la validez y fiabilidad de las predicciones.

5. Analizar e interpretar los resultados obtenidos, con el fin de formular recomendaciones estratégicas orientadas a la maximización de ingresos, optimización de inventarios, mejora de las campañas de promoción y alineación con las preferencias socioculturales del público objetivo.

3.3 Beneficios del proyecto

El presente proyecto ofrece una serie de beneficios tanto a nivel empresarial como académico. Desde una perspectiva estratégica, el desarrollo de un modelo predictivo basado en aprendizaje automático permitirá a las empresas del sector moda anticiparse con mayor precisión a las tendencias de consumo, reduciendo los riesgos asociados a decisiones basadas en intuiciones o interpretaciones subjetivas. Esto se traduce en una mejor planificación de inventario, reducción de costes operativos y optimización de las campañas de marketing, al alinear la oferta con las preferencias reales del consumidor.

Asimismo, el uso de variables socioculturales como la edad, género, ubicación e intereses culturales permitirá una segmentación más precisa y personalizada, fortaleciendo la conexión entre las marcas y sus públicos objetivos. Esto no solo mejora la experiencia del consumidor, sino que también favorece la fidelización y la ventaja competitiva en un mercado altamente dinámico y digitalizado.

En el ámbito académico y de investigación, el proyecto aporta valor al combinar herramientas de inteligencia artificial con enfoques socioculturales, uniendo dos disciplinas que hasta ahora han estado relativamente desconectadas en el estudio de la moda. De esta forma, se amplía el conocimiento sobre cómo las variables culturales y sociales inciden en el consumo estético, y se promueve una visión más holística, crítica y contextualizada en la aplicación de modelos predictivos.

Además, el trabajo ofrece una metodología replicable para otros sectores interesados en entender el comportamiento del consumidor a través de datos, lo cual refuerza su aplicabilidad en distintos contextos de análisis de tendencias, más allá de la industria de la moda.

Capítulo 4. DESARROLLO DEL PROYECTO

En este capítulo se expone el proceso de construcción del modelo predictivo, abarcando desde la fase de planificación hasta la obtención de los resultados finales. Se explican las etapas metodológicas seguidas, junto con los recursos utilizados, el presupuesto asignado y el análisis de viabilidad. Finalmente, se presentan los resultados del entrenamiento y la validación de los algoritmos, que constituyen el sustento para la discusión y las conclusiones del trabajo.

4.1 Planificación del proyecto

El presente Trabajo de Fin de Máster se desarrolló en un período de 4 meses, organizando el tiempo de forma intensiva en fases consecutivas que permitieron abarcar la investigación, el análisis técnico, el desarrollo del modelo predictivo y la redacción del informe. A continuación, se describe el cronograma de trabajo, con las principales actividades realizadas y su esfuerzo estimado:

Tabla 2

Cronograma de actividades del proyecto

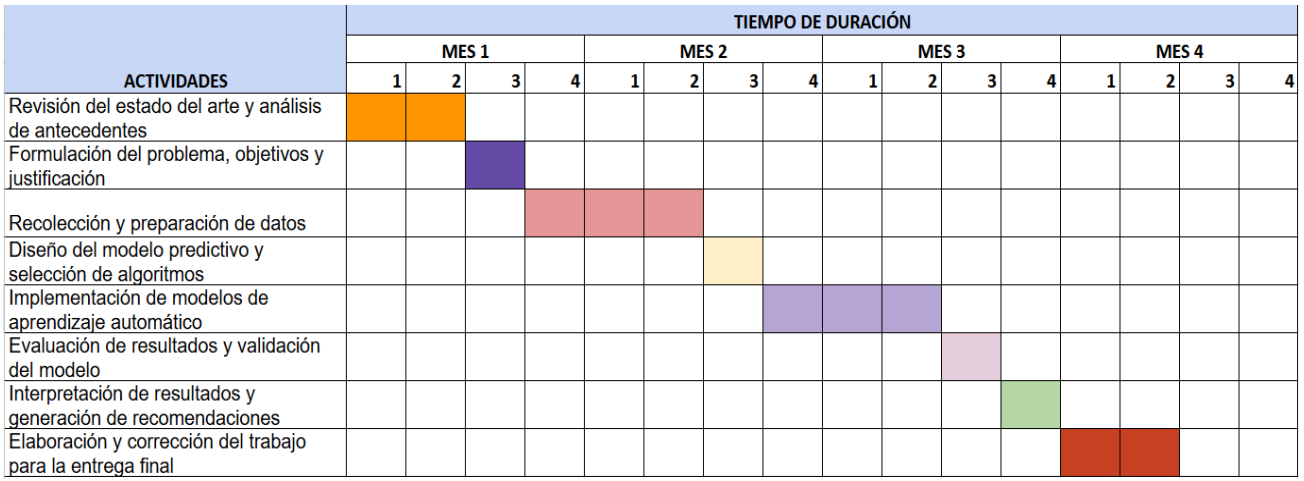
Actividad	Duración estimada	Descripción
Revisión del estado del arte y análisis de antecedentes	2 semanas	Estudio de literatura científica y técnica relacionada con IA, predicción de tendencias y variables socioculturales.
Formulación del problema, objetivos y justificación	1 semana	Redacción de la motivación, planteamiento del problema y definición clara de los objetivos del proyecto.
Recolección y preparación de datos	3 semanas	Revisión del dataset disponible, limpieza de datos, codificación de variables socioculturales y análisis exploratorio.
Diseño del modelo predictivo y selección de algoritmos	1 semana	Selección de técnicas de clasificación (KNN, Random Forest, Regresión logística, etc.) y definición del enfoque.

Implementación de modelos de aprendizaje automático	3 semanas	Programación, entrenamiento y ajuste de los algoritmos seleccionados.
Evaluación de resultados y validación del modelo	1 semana	Aplicación de métricas de desempeño (precisión, recall, F1-score, exactitud) y análisis comparativo entre modelos.
Interpretación de resultados y generación de recomendaciones	1 semana	Análisis de impacto empresarial y formulación de propuestas para el sector moda.
Elaboración y corrección del trabajo para la entrega final	2 semanas	Elaboración del documento final, organización de capítulos, inclusión de visualizaciones y corrección de estilo.

Nota. Elaboración propia

Figura 1

Cronograma de actividades del proyecto



Nota. Elaboración propia

4.2 Descripción de la solución, metodologías y herramientas empleadas

En el desarrollo de este trabajo de investigación, la metodología ocupa un papel fundamental, ya que proporciona la estructura necesaria para alcanzar de manera ordenada y rigurosa los objetivos planteados. Dado que el propósito de este estudio es

diseñar un modelo de aprendizaje automático que permita predecir el éxito de tendencias de consumo en moda, es imprescindible adoptar un enfoque metodológico que integre tanto la comprensión del contexto de negocio como el tratamiento y análisis de grandes volúmenes de datos.

El comportamiento de los consumidores en el sector de la moda está fuertemente influenciado por factores sociales, culturales, económicos y temporales, lo que convierte a este ámbito en un entorno altamente dinámico y complejo. Por esta razón, la metodología seleccionada debe ser capaz de capturar y procesar estos patrones para generar modelos predictivos con aplicabilidad práctica.

En este proyecto se adopta la metodología **CRISP-DM (Cross Industry Standard Process for Data Mining)**, ampliamente reconocida en el ámbito de la ciencia de datos por su enfoque sistemático, flexible y adaptable a diferentes sectores. Esta metodología es especialmente adecuada para proyectos donde es necesario recorrer un ciclo completo desde la identificación del problema de negocio hasta la generación de recomendaciones basadas en modelos predictivos.

El modelo CRISP-DM estructura el proceso en seis fases principales:

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado predictivo
5. Evaluación del modelo
6. Generación de recomendaciones estratégicas

Cada una de estas fases será detallada en el presente capítulo, explicando las actividades realizadas, las herramientas empleadas y las decisiones metodológicas tomadas para garantizar la calidad del proceso..

Además, la metodología contempla la aplicación de algoritmos de clasificación supervisada tales como árboles de decisión, random forest, K-nearest neighbors (KNN), Adaboost, regresión logística y descenso de gradiente estocástico (SGD), que serán entrenados y evaluados para determinar su capacidad de predecir el éxito de las tendencias de consumo. Los modelos serán evaluados utilizando métricas de

desempeño estándar como la precisión, la exactitud, la sensibilidad y el F1-score, con el objetivo de seleccionar la solución más adecuada para el contexto del negocio.

A través de este enfoque, se espera no solo desarrollar modelos predictivos eficientes, sino también generar insumos estratégicos que contribuyan a maximizar los ingresos, optimizar la gestión de inventarios y mejorar las decisiones comerciales en la industria de la moda.

FASES METODOLÓGICAS

4.2.1 Comprensión del negocio

El presente proyecto parte de una necesidad ampliamente contextualizada en los capítulos previos: la urgencia del sector de la moda por adaptarse a un entorno dinámico, marcado por el comportamiento digital y los cambios culturales en el consumo. En este contexto, se define como objetivo general el diseño de un modelo predictivo, basado en técnicas de aprendizaje automático supervisado, que permita anticipar el éxito de tendencias de moda a partir de variables socioculturales.

Este modelo busca contribuir a la toma de decisiones estratégicas en empresas del sector, a través de la identificación temprana de patrones de consumo emergente. En particular, se pretende optimizar procesos relacionados con el diseño de colecciones, gestión de inventario y estrategias de marketing, alineándose con los intereses y perfiles de consumidores en entornos culturalmente diversos.

En cuanto al alcance, el proyecto se enfoca en la aplicación de algoritmos de clasificación sobre datos estructurados, provenientes de fuentes transaccionales y sociodemográficas, con el fin de evaluar su capacidad para anticipar la adopción de nuevas tendencias. El resultado esperado es un sistema predictivo que no solo tenga buen rendimiento, sino que también ofrezca interpretabilidad para su integración en procesos de toma de decisiones en entornos reales de negocio.

4.2.2 Comprensión de los datos

Esta fase se centra en recolectar, explorar y familiarizarse con los datos disponibles, con el objetivo de evaluar su calidad, estructura y relevancia para el análisis. Durante

esta etapa, se realiza un análisis preliminar que incluye la identificación de posibles inconsistencias, valores faltantes, outliers o errores, así como una exploración inicial de patrones o relaciones significativas. Este conocimiento es esencial para tomar decisiones informadas en las etapas siguientes, especialmente en la preparación y modelado de los datos.

Comprender adecuadamente los datos permite al equipo anticipar desafíos técnicos, descubrir oportunidades ocultas y asegurar que los resultados obtenidos estén alineados con los objetivos del negocio.

Dataset seleccionado

Este dataset ha sido seleccionado por contener variables clave relacionadas con el comportamiento de los consumidores, sus preferencias culturales y los contextos sociales de compra, lo que permite abordar el problema de predicción desde una perspectiva integral y contextualizada

Nombre: Global Fashion Retail Sales

Fuente: Kaggle

Archivo Utilizado: El análisis se centra en el archivo transactions.csv, que contiene información detallada de las transacciones individuales realizadas en una empresa multinacional del sector de la moda.

Justificación

El dataset Global Fashion Retail Sales, disponible en la plataforma Kaggle, ha sido seleccionado para este trabajo por múltiples razones que lo convierten en una fuente de datos adecuada, pertinente y estratégica para abordar el problema de investigación planteado.

En primer lugar, este dataset proporciona un volumen considerable de datos transaccionales (aproximadamente 443.000 registros) que reflejan de manera detallada el comportamiento de compra de los consumidores en el sector de la moda a nivel global. La diversidad y cantidad de datos permiten garantizar la robustez

estadística y la viabilidad de aplicar modelos de aprendizaje automático con capacidad predictiva real.

En segundo lugar, el archivo transactions.csv contiene variables clave que permiten abordar el estudio desde una perspectiva sociocultural y comercial. Entre estas variables se encuentran las tallas, los colores, las ubicaciones de las tiendas, los métodos de pago y las monedas utilizadas, lo que permite considerar factores culturales y regionales en la predicción del éxito de las tendencias de consumo. La posibilidad de segmentar los datos por estos criterios ofrece un valor diferencial al análisis, permitiendo construir modelos ajustados a las preferencias específicas de los consumidores en diferentes contextos geográficos y culturales.

En tercer lugar, este dataset proporciona información temporal precisa (fecha y hora de cada transacción), lo que facilita el análisis de patrones estacionales y la identificación de tendencias a lo largo del tiempo. Este aspecto es especialmente relevante en la industria de la moda, donde las preferencias de los consumidores varían según temporadas, eventos sociales y ciclos de moda.

Adicionalmente, el dataset incluye tanto ventas como devoluciones, lo que permite construir un indicador integral de éxito comercial, considerando no solo los productos que se venden, sino también aquellos que son devueltos por los clientes. Esto proporciona una visión más realista del comportamiento de consumo, dado que el éxito de una tendencia no solo se mide por la cantidad vendida, sino también por la aceptación y satisfacción del consumidor final.

Otra razón que respalda la selección de este dataset es su estructura limpia y organizada, que facilita la integración de los datos y permite realizar procesos de limpieza, transformación y análisis de manera eficiente. La variedad de variables disponibles también posibilita la aplicación de diferentes técnicas de ingeniería de variables y la construcción de características derivadas que pueden mejorar el rendimiento de los modelos predictivos.

Por último, el dataset está alineado con los objetivos estratégicos de este trabajo, ya que permite construir un modelo predictivo que anticipe el éxito de las tendencias de

consumo en moda con el fin de maximizar los ingresos, optimizar la gestión de inventarios y mejorar las estrategias comerciales.

En conclusión, la selección del Dataset Global Fashion Retail Sales como fuente de datos es adecuada tanto desde el punto de vista técnico como desde la perspectiva de negocio, ya que proporciona la información necesaria para responder a las preguntas de investigación planteadas y aporta valor al desarrollo de soluciones predictivas aplicables al sector de la moda.

Descripción:

Este conjunto de datos contiene información detallada sobre las operaciones comerciales de una empresa minorista multinacional que tiene 35 tiendas en 7 países y adicional a eso transan en diferentes monedas. Cada registro corresponde a un ítem individual dentro de una factura, abarcando tanto ventas como devoluciones. Estos datos simulan las transacciones realizadas durante un periodo de dos años.

A continuaciòn se relaciona una tabla con la descripción de las 19 variables que se encuentran en el dataset seleccionado:

Tabla 3

Descripción de las variables del dataset

Variable	Descripción Ampliada
Invoice ID	Identificador único asignado a cada transacción, que puede corresponder a una venta (INV) o a una devolución (RET). Este campo permite agrupar todas las partidas de una misma factura y rastrear de manera precisa las operaciones realizadas. Incluye información codificada como país y tienda. Ejemplo: INV-US-001-00001233.
Line	Número secuencial que representa la posición del producto dentro de la factura. Una misma factura puede incluir varias líneas si el cliente adquirió múltiples productos o variantes. Esta variable permite desglosar cada elemento individual de la transacción.

Customer ID	Identificador único y anonimizado que representa a cada cliente. Esta variable permite rastrear el historial de compras por cliente, analizar comportamientos de consumo y segmentar a los clientes según sus patrones de compra.
Product ID	Código único que identifica a cada producto disponible en el inventario. Es la base para asociar los productos vendidos con sus atributos específicos como talla, color y categoría.
Size	Talla del producto vendido. Puede tomar valores como S, M, L, XL, entre otros. Esta variable es clave para analizar las preferencias del consumidor y detectar tendencias asociadas a tallas específicas. Es un factor cultural relevante, ya que las preferencias de talla pueden variar según la región.
Color	Color del producto adquirido. Esta variable permite analizar la aceptación de las diferentes variantes de color en las tendencias de moda y es un indicador de preferencia visual y cultural.
Unit Price	Precio unitario del producto antes de aplicar cualquier descuento. Esta variable refleja el valor base del producto y es esencial para el cálculo de márgenes, análisis de precios y comparación de modelos de consumo.
Quantity	Número de unidades adquiridas por el cliente en la línea correspondiente. Permite medir la cantidad demandada por transacción y calcular la importancia comercial de cada producto.
Date	Fecha y hora exacta en que se realizó la transacción. Es una variable fundamental para realizar análisis temporales, identificar tendencias estacionales y segmentar por periodos (semanas, meses, temporadas de moda).
Discount	Porcentaje de descuento aplicado a la línea de la factura. Se expresa como un valor decimal (por ejemplo, 0.30 corresponde a un 30% de descuento). Es importante para evaluar el impacto de las estrategias de precio y promoción en las decisiones de compra.
Line Total	Valor total pagado por el cliente en la línea de la factura después de aplicar el descuento. Se calcula como: Unit Price × Quantity × (1 - Discount). Este campo permite analizar el gasto por producto y es base para calcular el total de la factura.

Store ID	Identificador único de la tienda física donde se realizó la transacción. Es relevante para el análisis geográfico de las ventas, evaluación del rendimiento por punto de venta y detección de diferencias culturales entre regiones.
Employee ID	Identificador del empleado que gestionó la transacción. Puede ser utilizado para análisis de desempeño individual, eficiencia en las ventas o trazabilidad del proceso de compra.
Currency	Código de moneda internacional (ISO 4217) con el que se realizó la transacción. Permite trabajar con operaciones en diferentes países y facilita el análisis financiero multimoneda. Ejemplos: USD (dólar estadounidense), EUR (euro).
Currency Symbol	Símbolo de la moneda usada en la transacción, correspondiente a la variable Currency. Ejemplos: $, €, £. Se utiliza como información adicional en la representación visual de los datos.
SKU	Código SKU (Stock Keeping Unit), una referencia única que combina el Product ID, la talla y el color. Este campo facilita la gestión del inventario y permite identificar variantes específicas del producto. Ejemplo: FESH81-M-PINK (producto 81, talla M, color rosa).
Transaction Type	Indica si la transacción corresponde a una venta (Sale) o una devolución (Return). Esta variable es crítica para calcular correctamente el volumen de ventas netas y para evaluar la aceptación real de los productos.
Payment Method	Medio de pago utilizado por el cliente, como tarjeta de crédito, efectivo, billeteras digitales, etc. Esta información permite analizar las preferencias de pago por región o por perfil de cliente.
Invoice Total	Monto total de la factura, es decir, la suma de todos los Line Total asociados a un mismo Invoice ID. Este valor es repetido en cada línea de la misma factura y representa el valor final pagado por el cliente en la transacción. Es la base para el análisis financiero global.

Nota. Elaboración propia

Este nivel de detalle permite construir indicadores de éxito, analizar preferencias de consumo y estudiar el comportamiento de compra en diferentes segmentos sociales y culturales.

Descripción estadística y exploratoria

Variables numéricas

A continuación, se presenta el análisis descriptivo de las principales variables numéricas del archivo transactions.csv del dataset Global Fashion Retail Sales. Este análisis permite comprender la estructura de los datos, identificar patrones y detectar posibles valores atípicos que pueden influir en el proceso de modelado.

Tabla 4

Análisis descriptivo de las variables numéricas del dataset

Variable	Mínimo	25%	Mediana	75%	Máximo	Media	Desviación Estándar
Line	1	1	1	2	7	1.58	1.15
Customer ID	1	19,03	37,74	63,024	1,642,526	58,979	91,732
Product ID	1	5,8	8,672	12,723	17,94	8,892	4,242
Unit Price	4.00	31.50	44.50	63.50	152.50	50.75	26.82
Quantity	1	1	1	1	3	1.10	0.39
Discount	0.00	0.00	0.00	0.20	0.60	0.11	0.19
Line Total	-436.50	24.50	38.50	59.00	450.00	44.06	42.07
Store ID	1	1	1	1	2	01.05	0.22
Employee ID	5	7	9	12	24	9.61	3.67
Invoice Total	-715.00	32.00	61.50	130.50	966.50	93.90	115.95

Nota. Elaboración propia con base al dataset Global Fashion Retail Sales

Interpretación Detallada

- **Line:**
  La mayoría de las facturas tienen una sola línea de producto (mediana = 1), aunque algunas alcanzan hasta siete productos por factura. Esto indica que los clientes generalmente realizan compras pequeñas.

- **Customer ID:**
  Hay un rango muy amplio de clientes. La dispersión elevada muestra que el dataset recoge una amplia base de consumidores con muchos identificadores únicos, lo que da diversidad al análisis.

- **Product ID:**
  Los identificadores de productos están distribuidos de manera uniforme, lo que permite trabajar con una amplia variedad de artículos para la predicción.

- **Unit Price:**

  El precio de los productos oscila entre 4 y 152.5 unidades monetarias. El precio promedio es de aproximadamente 50.75 unidades, con una ligera dispersión, lo que indica una mezcla de productos económicos y de mayor valor.

- **Quantity:**
  La mayoría de las transacciones corresponden a la compra de una sola unidad por producto. Esto es típico en el comercio minorista de moda y es relevante para analizar el comportamiento de consumo individual.

- **Discount:**
  El 50% de las transacciones no tienen descuento aplicado. Cuando hay descuentos, suelen situarse en el rango de 20% a 60%, lo que refleja políticas promocionales específica

- **Line Total:**
  Se identificaron valores negativos, que corresponden a devoluciones. La mayoría de las ventas tienen un valor total de aproximadamente 38.50 unidades monetarias por producto. Este campo es clave para identificar compras efectivas versus devoluciones.

- **Store ID:**
  Solo hay dos tiendas registradas en el dataset, lo que permitirá segmentar y comparar los comportamientos de compra por tienda.

- **Employee ID:**
  La variación de empleados que procesan las ventas es amplia. Esto puede ser útil para rastrear desempeño o distribución de trabajo.

- **Invoice Total:**
  Existen facturas con valores negativos, lo que indica transacciones de devolución completas. El valor promedio de una factura es de 93.90 unidades monetarias, aunque existe una alta dispersión (algunas facturas superan las 900 unidades).

Luego de validar es necesario tener en cuenta los siguientes puntos:

- Se debe realizar un tratamiento especial para valores negativos en Line Total e Invoice Total, ya que indican devoluciones. Se pueden etiquetar como "No exitosas" para los modelos de clasificación.
- La mayoría de los clientes compran pocas unidades por transacción, lo que refleja un comportamiento típico en retail de moda.
- La dispersión de precios y montos finales muestra que el dataset incluye productos de diferentes gamas de valor.
- Las promociones y descuentos tienen un impacto visible en las transacciones y deben ser consideradas como posibles variables predictoras clave.

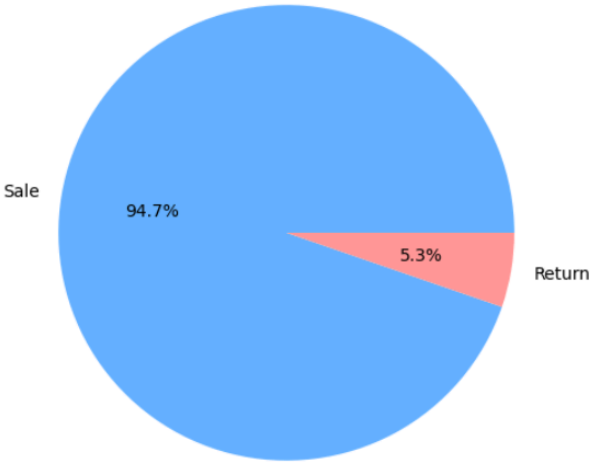
Variables categóricas

El análisis exploratorio de las variables categóricas del dataset transactions.csv permite identificar los patrones de comportamiento de los consumidores, las preferencias de compra y la distribución de las transacciones. A continuación, se describen las principales variables categóricas, sus frecuencias y porcentajes.

1. **Transaction Type (Tipo de transacción**)

Figura 2

Distribución de tipo de transacción



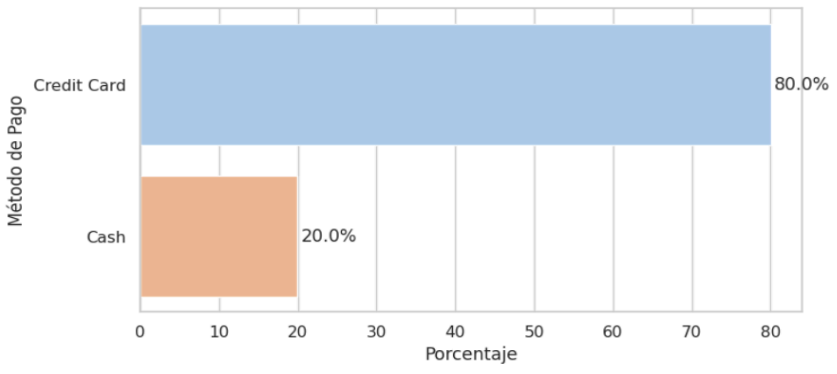
Nota. Elaboración propia

- La gran mayoría de las transacciones son ventas efectivas (94.7%).
- Las devoluciones representan una pequeña proporción (5.3%), pero son relevantes para analizar la satisfacción del cliente y pueden influir en la predicción de tendencias exitosas.

2. **Payment Method (Método de pago)**

Figura 3

Distribución de métodos de pago



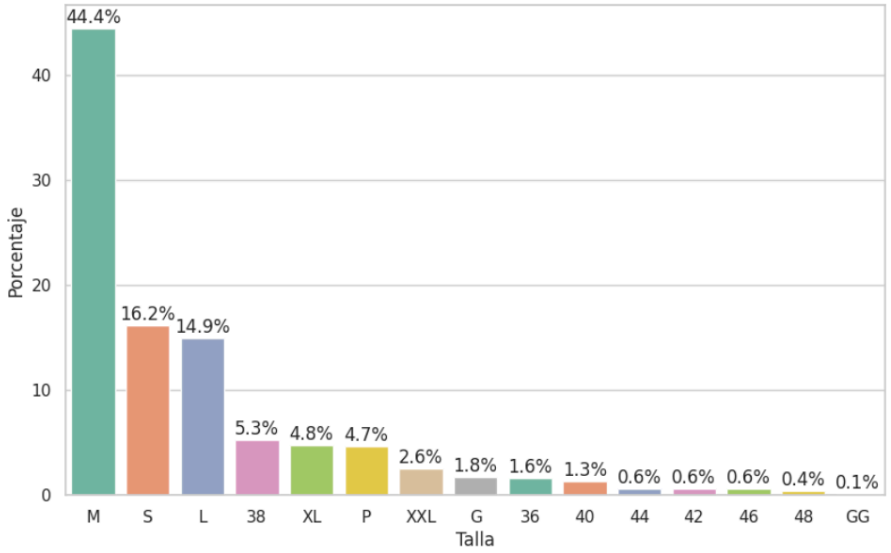
Nota. Elaboración propia

- El pago con tarjeta de crédito es el método predominante en las transacciones.
- El efectivo aún representa una parte importante de las compras, lo que indica la necesidad de atender diferentes perfiles de clientes.

3. **Size (Talla del producto)**

Figura 4

Distribución de métodos de pago



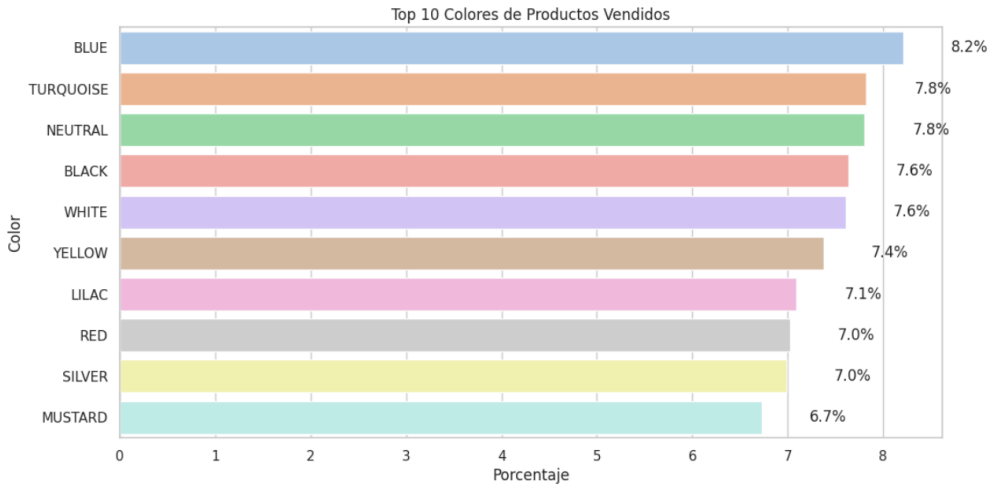
Nota. Elaboración propia

- La talla M es la más vendida, seguida por S y L.
- Las tallas numéricas como 38, 36, y 40 también tienen presencia, posiblemente por productos como calzado.
- El análisis de tallas permite identificar tendencias de consumo asociadas a perfiles demográficos y preferencias regionales

4. **Color (Color del producto)**

Figura 5

Top 10 colores de productos vendidos



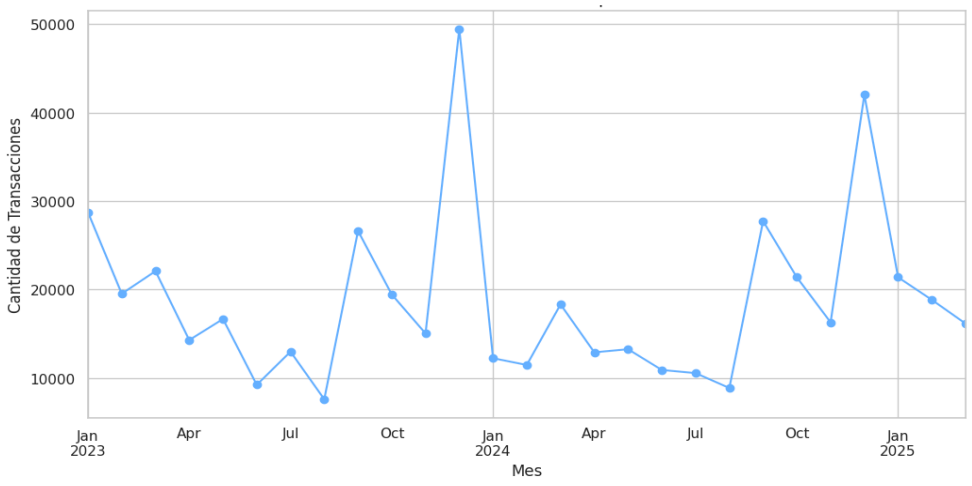
Nota. Elaboración propia

- Los colores azul, turquesa, neutro, negro y blanco son los más demandados.
- Este análisis proporciona información clave sobre las preferencias estéticas del consumidor.

5. Date (Fecha de transacción)

Figura 6

Número de transacciones por mes



Nota. Elaboración propia

- Hay picos de transacciones muy fuertes alrededor de diciembre y enero de cada año. Esto sugiere patrones estacionales relacionados probablemente con compras de fin de año.
- Hay meses con menor actividad (julio, agosto) donde las transacciones son notablemente bajas.
- Los ciclos parecen repetirse, lo que puede ser clave para predecir tendencias de consumo estacional.

SKU (Código único del producto)

Existen 56,578 productos únicos (SKU), lo que muestra una amplia diversidad de productos vendidos. Esta variedad de productos permite trabajar con diferentes segmentos de moda y profundizar en el análisis por categoría.

Currency (Moneda)

Todas las transacciones están registradas en dólares estadounidenses (USD), lo que asegura homogeneidad en el análisis monetario.

Invoice ID (Identificador de factura)

Hay 356,205 facturas únicas. Algunas facturas agrupan varias líneas de productos (ventas múltiples).

Este campo es útil para agregar información a nivel de compra total, especialmente para calcular el valor del ticket promedio y clasificar la compra como exitosa o no

Conclusión del Análisis Categórico:

- La distribución de las categorías muestra patrones consistentes con la industria de la moda.
- La variable Transaction Type puede utilizarse como variable objetivo (éxito = venta, no éxito = devolución)
- Variables como Payment Method, Size, Color y SKU ofrecen información valiosa sobre preferencias de consumo

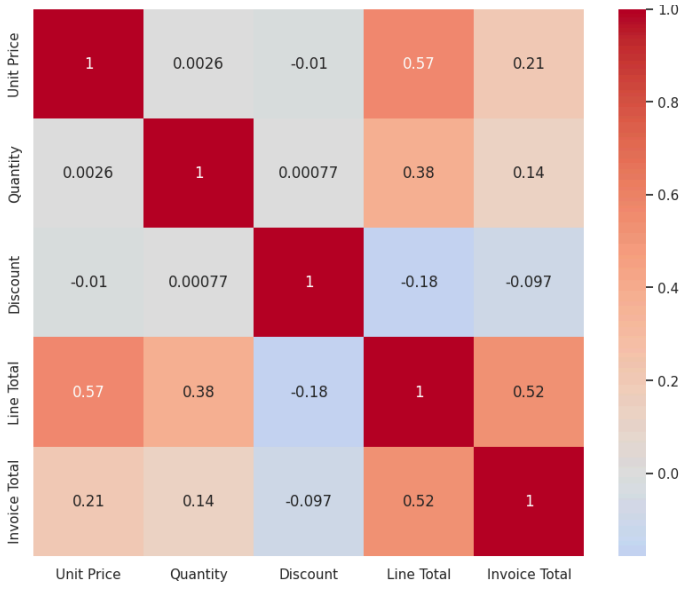
- La fecha puede ser clave para estudiar la estacionalidad y eventos comerciales importantes.

La descripción estadística exploratoria permitió identificar que la mayoría de las transacciones se concentran en tallas medianas, colores neutros y métodos de pago electrónicos, lo que proporciona evidencia preliminar sobre las preferencias actuales de los consumidores en el sector de la moda. Asimismo, se detectaron valores atípicos asociados a compras de alto valor, que deberán ser considerados cuidadosamente durante la fase de modelado para evitar sesgos en las predicciones.

Análisis de correlación

Figura 7

Matriz de correlación entre variables numéricas



Nota. Elaboración propia

Para entender este gráfico se debe tener claro qué significan los colores

- **1** (color rojo oscuro) = correlación perfecta positiva.
- **0** (gris) = no hay correlación.
- **-1** (color azul) = correlación perfecta negativa.

La matriz de correlación obtenida proporciona una visión integral de la relación existente entre las variables numéricas del conjunto de datos. Este análisis es fundamental para identificar posibles dependencias o asociaciones que puedan influir en el comportamiento de las transacciones de consumo en moda.

En primer lugar, se destaca una correlación positiva moderada entre el precio unitario (Unit Price) y el total por línea (Line Total) con un coeficiente de 0.57, lo que indica que a medida que el precio de los productos aumenta, también se incrementa el valor total pagado por cada línea de la factura. Esta relación es lógica desde el punto de vista comercial, ya que productos con precios más altos tienden a generar transacciones de mayor valor.

Asimismo, la cantidad de productos adquiridos (Quantity) presenta una correlación positiva con el total por línea (0.38) y con el total de la factura (0.14). Esto sugiere que el volumen de compra también contribuye de manera relevante al monto total de la transacción, aunque en menor medida que el precio unitario.

Por su parte, el descuento aplicado (Discount) muestra correlaciones negativas con el total por línea (-0.18) y con el total de la factura (-0.097), lo cual es consistente con la naturaleza de esta variable, ya que mayores descuentos generan una disminución en el valor total de la compra. Sin embargo, la baja magnitud de estas correlaciones indica que los descuentos, aunque influyen, no son el factor determinante en el valor total de las facturas dentro de este dataset.

Finalmente, la correlación entre el total por línea y el total de la factura es de 0.52, lo que confirma que el valor final de cada transacción depende significativamente de la suma de los valores individuales de cada línea. Las demás correlaciones observadas son bajas o cercanas a cero, lo que sugiere que no existe una relación lineal fuerte entre el resto de las variables analizadas.

Este análisis proporciona información valiosa para la selección de las variables más relevantes en la construcción del modelo predictivo y permite entender cuáles son los factores que más impactan en el comportamiento de las ventas y en la adopción de tendencias de consumo.

4.2.3 Preparación de los datos

Esta fase tiene como objetivo transformar los datos brutos en un conjunto limpio, estructurado y adecuado para el análisis predictivo. Esta etapa es fundamental, ya que la calidad de los datos influye directamente en la efectividad y fiabilidad de los modelos que se construirán posteriormente.

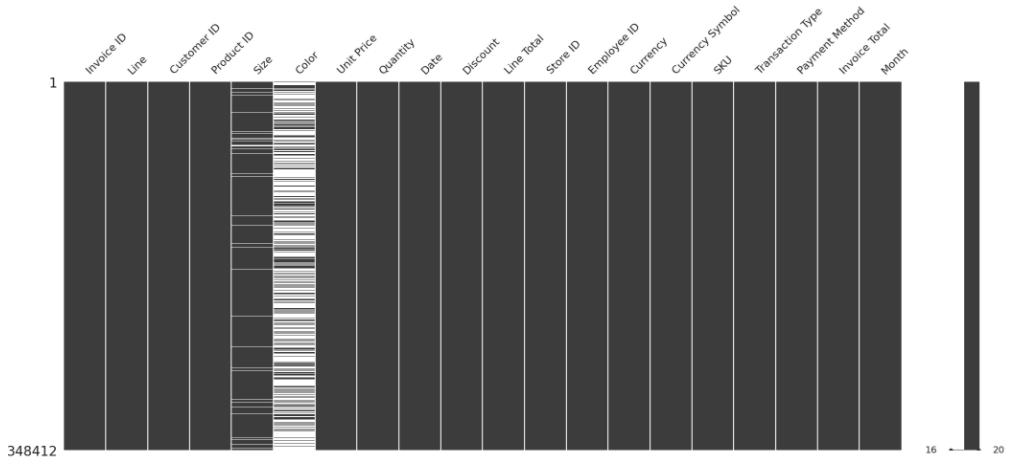
Una adecuada preparación de los datos garantiza no solo la robustez del modelo, sino también la validez de las conclusiones extraídas en función de los objetivos definidos en fases anteriores.

Análisis de Datos Faltantes

Por medio de la siguiente gráfica llamada "Matriz de Ausencia", se va a observar si hay variables con valores faltantes.

Figura 8

Datos faltantes en el dataset



Nota. Elaboración propia

Como se evidencia en la gráfica en donde las líneas blancas son los valores faltantes y las en donde se visualizan más son en las variables size y color, ahora se va a especificar exactamente cuántos son para tener más claridad en los datos

Tabla 5

Valores faltantes por variable en el dataset

Variable	Valores Faltantes	Porcentaje (%)
Color	234097	67.189706
Size	22201	6.372054
Transaction Type	1	0.000287
Payment Method	1	0.000287
Invoice Total	1	0.000287

Nota. Elaboración propia con base al dataset Global Fashion Retail Sales

El análisis de datos faltantes permitió identificar que las variables Color, Size, Transaction Type, Payment Method e Invoice Total presentan valores ausentes. De estas, la variable Color destaca por tener un 67,19% de datos faltantes, lo cual representa una pérdida significativa de información. La variable Size presenta un 6,37% de datos faltantes, mientras que las variables Transaction Type, Payment Method e Invoice Total tienen ausencias mínimas (menores al 0,001 %).

Estas observaciones son coherentes con los resultados visualizados en el heatmap de datos faltantes, donde sólo las variables con ausencias son representadas. Cabe resaltar que las variables completas no aparecen en este tipo de visualización, dado que el enfoque del heatmap es exclusivamente el análisis de las relaciones de ausencia de datos.

Para el tratamiento de estos datos faltantes, se adoptaron las siguientes decisiones:

- La variable **Color** fue descartada del análisis por su alto nivel de ausencia.
- Los valores faltantes en **Size** fueron imputados mediante la moda (valor más frecuente) para conservar la mayor cantidad de registros posibles.
- Los registros con valores ausentes en **Transaction Type, Payment Method e Invoice Total** fueron eliminados debido a su mínima incidencia y para asegurar la consistencia del dataset.

Normalización de las variables

En el presente trabajo, la normalización de las variables numéricas es un paso fundamental debido a las diferencias significativas en la escala y magnitud de los datos incluidos en el conjunto seleccionado. Variables como Unit Price, Quantity,

Discount, Line Total e Invoice Total presentan rangos muy dispares, lo que puede generar sesgos en la visualización gráfica y en el rendimiento de los modelos predictivos basados en algoritmos de clasificación supervisada.

La normalización es especialmente necesaria para este estudio porque algunos de los algoritmos seleccionados, como K-Nearest Neighbors (KNN), Descenso de Gradiente Estocástico (SGD), Regresión Logística con regularización, y Adaboost, son sensibles a las escalas de las variables. Estos métodos calculan distancias, pesos o probabilidades que pueden verse distorsionados si una variable con un rango amplio domina la función de coste o las métricas de distancia.

Además, la normalización facilita la comparación visual entre variables cuando se representan en gráficos conjuntos, como los diagramas de violín múltiples, evitando que variables con menor magnitud queden ocultas o desproporcionadamente pequeñas respecto a las de mayor escala.

Por tanto, la aplicación de la normalización garantiza:

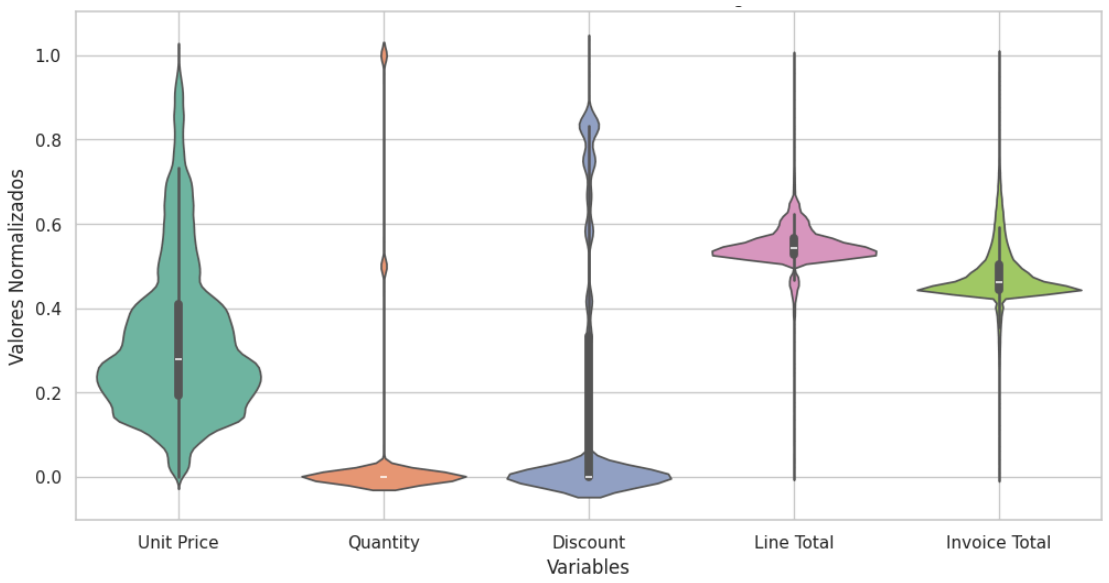
- Una contribución equilibrada de todas las variables al proceso de modelado.
- Una interpretación más clara y objetiva de los gráficos comparativos.
- Un mejor rendimiento y estabilidad de los algoritmos supervisados empleados en este estudio.

Este preprocesamiento es indispensable para cumplir con los estándares metodológicos y asegurar la fiabilidad de los resultados obtenidos en la predicción del éxito de tendencias de consumo en moda.

Detección de Outliers con Boxplots

Figura 9

Distribución de variables normalizadas con diagrama de violín



Nota. Elaboración propia

Unit Price: La distribución muestra una gran concentración de valores en la parte baja de la escala normalizada, lo que indica que la mayoría de los productos tienen precios relativamente bajos en comparación con el máximo. Se observa una ligera asimetría hacia valores intermedios, sugiriendo algunos productos de precio medio, pero con menor frecuencia.

Quantity: La distribución es extremadamente sesgada hacia valores bajos, lo que refleja que en la mayoría de las transacciones se compran pocas unidades. Sin embargo, hay algunos outliers que representan compras de grandes cantidades, aunque son muy poco frecuentes.

Discount:Se presenta una distribución con un claro pico cerca de cero, indicando que la mayoría de los productos vendidos no tienen descuento. No obstante, se observan algunos valores más alejados, lo que representa transacciones con descuentos significativos pero poco comunes.

Line Total:La distribución es relativamente simétrica con una mayor concentración alrededor de valores intermedios. Esto sugiere que el total pagado por línea de factura se mantiene dentro de un rango moderado para la mayoría de las compras.

Invoice Total: La mayor parte de las transacciones totales por factura están concentradas en valores bajos a medios, con algunos valores atípicos que representan facturas con montos significativamente altos.

Este diagrama de violín muestra que, en todas las variables analizadas existen indicios de outliers.

- Estos outliers son especialmente evidentes en Unit Price, Quantity, Line Total e Invoice Total, donde se presentan valores extremos alejados de la mayor concentración de datos.

- La presencia de estos outliers es relevante para el modelado, ya que podrían influir en el rendimiento y la estabilidad de los algoritmos de clasificación supervisada.

Como ya se evidenció anteriormente en el gráfico si se encuentran datos atípicos pero no proporcionan en una cuantificación precisa de los valores fuera de rango. Por esta razón, se hace necesario aplicar un método estadístico más riguroso y específico, como el cálculo del Rango Intercuartílico (IQR).

Este enfoque permite definir límites con base en criterios matemáticos para cada variable numérica, con el fin de identificar rangos válidos y distinguir de forma objetiva entre datos regulares y atípicos, calcula con exactitud el porcentaje de valores atípicos en cada variable, aportando una visión más precisa que la obtenida mediante visualizaciones exploratorias e Identificar las variables con mayor presencia de outliers y establecer un criterio claro para decidir su tratamiento, ya sea mediante eliminación, ajuste o transformación.

Este método no solo mejora la profundidad y solidez del análisis exploratorio, sino que también garantiza que los modelos predictivos que se desarrollarán posteriormente se construyan sobre datos limpios, representativos y confiables.

Tabla 6

Identificación de outliers mediante el método IQR

Variable	Q1	Q3	IQR	Límite Inferior	Límite Superior	Cantidad de Outliers	Porcentaje de Outliers (%)
Unit Price	32.0	63.5	31.5	-15.25	110.75	12440	3.570.486
Quantity	1.0	1.0	0.0	1.00	1.00	23221	6.664.811
Discount	0.0	0.2	0.2	-0.30	0.50	1036	297.349
Line Total	25.0	59.5	34.5	-26.75	111.25	32052	9.199.454
Invoice Total	33.0	132.0	99.0	-115.50	280.50	34271	9.836.343

Nota. Elaboración propia con base al dataset Global Fashion Retail Sales

Unit Price (Precio Unitario):

- Los precios deberían estar entre **-15.25 y 110.75**.
- Hay **12,440 outliers**, lo que representa un **3.57%** del total de registros, lo que indica una baja proporción de datos atípicos.
- Valores negativos no tienen sentido en precios, se debe validar.

Quantity (Cantidad):

- Los valores aceptados son **exactamente 1.0** (rango estrecho).
- Hay **23,221 outliers (6.66%)** con cantidades superiores a 1, puede indicar ventas al por mayor o registros especiales.

Discount (Descuento):

- El rango esperado es de **-0.30 a 0.50**.
- Solo **1,036 outliers (0.29%)** → bajo impacto.
- Valores negativos en descuento podrían ser errores.

Line Total (Total de Línea):

- El valor debería estar entre **-26.75 y 111.25**.
- Tiene **32,052 outliers (9.19%)**.

- Valores negativos no son lógicos, se debe revisar si son devoluciones mal registradas.

Invoice Total (Total de Factura):

- El rango esperado es entre **-115.50 y 280.50**.
- Tiene **34,271 outliers (9.83%)**.
- Facturas con totales negativos parecen errores, probablemente necesitan limpieza

El análisis estadístico basado en el rango intercuartílico (IQR) permitió identificar la presencia de outliers en las principales variables numéricas del dataset. Se observó que variables como Unit Price, Line Total e Invoice Total contienen un porcentaje de valores atípicos que oscila entre el 3% y el 10%, mientras que en la variable Quantity se evidenció una alta concentración en el valor de 1, considerando como outliers a las cantidades mayores. Estos resultados sugieren la necesidad de revisar el impacto que estos datos atípicos pueden tener en la modelación posterior.

Transformaciones para los outliers:

- En las variables **Unit Price y Discount**, los valores atípicos serán eliminados debido a su baja incidencia y su potencial para distorsionar el modelo.
- En el caso de **Quantity**, tras revisar la naturaleza de la variable y los registros específicos, se concluyó que estos valores representan escenarios completamente válidos dentro del contexto del negocio minorista de moda.

  En una tienda de este tipo, es razonable que los clientes puedan adquirir múltiples prendas en una sola transacción, especialmente durante eventos promocionales, temporadas de descuentos o compras grupales. Por tanto, estos valores no son errores ni desviaciones atípicas, sino que reflejan comportamientos reales de compra. Debido a esto, se decidió conservar estos registros sin aplicar correcciones ni transformaciones**.**

- En la variable **Invoice Total** se tienen valores negativos, estos registros fueron evaluados de manera detallada para determinar si correspondían a errores, comportamientos atípicos o escenarios legítimos del negocio.

Tras cruzar la variable Invoice Total con la columna Transaction Type, se comprobó que todos los registros con valores negativos corresponden a transacciones de tipo 'Return' (devolución). Esto valida que estos datos son representaciones fieles de las devoluciones realizadas por los clientes y no son inconsistencias ni registros erróneos.

En consecuencia, se decidió mantener estos registros en el dataset, ya que reflejan operaciones reales dentro del proceso comercial de la empresa. Además, incluir estas devoluciones es fundamental para capturar de forma precisa las tendencias de consumo y los patrones completos de comportamiento de los clientes, considerando tanto las compras como las devoluciones.

Por tanto, no se aplicó ninguna corrección ni eliminación sobre estos registros, asegurando así la integridad del análisis y la fidelidad de los resultados del modelo predictivo.

- Por último se validó la variable **Line total,** esta representa el importe total pagado por el cliente en cada línea de factura, después de aplicar descuentos. Su cálculo responde a la fórmula estándar:

$$\text{Line Total} = \text{Unit Price} \times \text{Quantity} \times (1 - \text{Discount})$$

Se verificó la coherencia de esta variable reproduciendo manualmente el cálculo en una nueva columna y comparándola con los valores originales. Para las transacciones de tipo venta (Sale), los resultados fueron consistentes, evidenciando que Line Total es una variable confiable para su uso como medida del éxito comercial de cada ítem.

Sin embargo, en las transacciones de tipo devolución (Return), se observó que los valores de Line Total aparecían como negativos, lo cual refleja el reembolso asociado a la devolución. Dado que el presente trabajo se enfoca en predecir el éxito de tendencias, se decidió filtrar únicamente las ventas reales, excluyendo las devoluciones del conjunto de datos. Esta decisión permite obtener una representación más precisa del comportamiento de consumo y evita sesgos

derivados de operaciones inversas que no responden a la lógica de adopción de tendencias

4.2.4 Modelado predictivo

Especificación del modelo

En esta etapa del proyecto se procede a la implementación del modelo de predicción, cuyo objetivo es anticipar el éxito comercial de una prenda a partir de atributos transaccionales y socioculturales presentes en los datos. Para ello, se ha formulado el problema como una tarea de clasificación supervisada binaria.

La variable objetivo (y) definida es Trend_Successful, una transformación binaria de la variable continua Line Total, la cual representa el importe pagado por el cliente en una transacción específica (calculado como Unit Price × Quantity × (1 - Discount)). Esta nueva variable toma el valor 1 cuando el valor de Line Total es igual o superior al percentil 75 del conjunto de datos, y 0 en caso contrario. Esta segmentación permite distinguir entre productos que han tenido un comportamiento comercial destacado frente al resto.

Las variables exógenas (X) seleccionadas como predictoras del éxito incluyen:

- Atributos económicos y comerciales:
  - Unit Price, Quantity, Discount.
- Atributos de producto:
  - Size, SKU.
- Atributos de contexto y comportamiento:
  - Store ID, Payment Method, Currency.

Este conjunto de variables ha sido seleccionado con base en su capacidad explicativa y disponibilidad dentro del dataset, así como en los referentes del estado del arte. Aquellas variables con alta proporción de valores nulos o que no aportaban valor predictivo directo (como Color, Employee ID o Transaction Type) fueron descartadas durante el proceso de limpieza y depuración.

Preprocesamiento previo al modelado

Durante el preprocesamiento, se realizaron diversas tareas de limpieza para garantizar la calidad del dataset. En primer lugar, se eliminó la variable Color debido a su alta proporción de valores nulos, lo cual comprometía su utilidad como predictor. La variable Size, por su parte, fue completada mediante imputación con la moda, ya que se trataba de una variable categórica relevante con pocos valores faltantes. Asimismo, se eliminaron los registros que presentaban valores nulos en campos críticos como Transaction Type, Payment Method e Invoice Total, ya que estos datos eran indispensables para caracterizar correctamente cada transacción.

Adicionalmente, se aplicó una técnica de detección de valores atípicos sobre las variables Unit Price y Discount, eliminando los registros fuera del rango percentílico 1-99 para reducir el impacto de valores extremos en el entrenamiento de los modelos. También se excluyeron las variables Invoice ID, Customer ID y Employee ID del análisis, dado que no aportan valor predictivo directo y su inclusión podría introducir ruido o sesgos en los algoritmos.

Finalmente, antes de proceder con el modelado, se realizó la normalización de las variables numéricas mediante escalado estándar (StandardScaler), con el objetivo de mejorar el rendimiento de los algoritmos basados en distancias (como KNN o regresión logística) y asegurar la comparabilidad entre atributos de diferente magnitud.

Algoritmos seleccionados

Basado en los hallazgos del estado del arte, se han seleccionado los siguientes algoritmos de clasificación para su implementación y comparación

- Árboles de decisión (Decision Tree Classifier): Modelo interpretable que divide recursivamente el espacio de variables.
- Random Forest: Ensamble de árboles que mejora la generalización y reduce el sobreajuste.
- K-Nearest Neighbors (KNN): Modelo basado en la similitud entre observaciones.
- AdaBoost: Método de boosting que combina clasificadores débiles para generar uno fuerte.

- Regresión logística: Algoritmo lineal ideal como baseline para problemas binarios.
- Descenso de gradiente estocástico (SGDClassifier): Variante optimizada para grandes volúmenes de datos.

Tabla 7

Comparación entre los principales algoritmos de clasificación

Modelo	Ventajas	Desventajas
Árbol de decisión	Interpretabilidad, fácil visualización	Riesgo de sobreajuste
Random Forest	Precisión, robustez, manejo de no linealidad	Menor interpretabilidad
KNN	Intuitivo, no paramétrico	Sensible a escala y ruido
AdaBoost	Alta precisión, combina errores	Requiere parametrización cuidadosa
Regresión logística	Simplicidad, buena base lineal	No captura relaciones no lineales complejas
SGDClassifier	Eficiente en grandes datasets	Sensible al preprocesado y tuning

Nota. Elaboración propia

Cada uno de estos modelos presenta características particulares en términos de precisión, complejidad y capacidad de generalización, por lo que su rendimiento será evaluado empíricamente para determinar cuál resulta más adecuado en este contexto

Validación del modelo

Cada modelo se entrena utilizando una partición del dataset con un 70% de datos para entrenamiento y 30% para prueba, aplicando train_test_split() con una semilla fija (random_state=42) para garantizar reproducibilidad.

Además, se utiliza validación cruzada (k-fold cross-validation) con k=5 para comprobar la estabilidad de cada modelo en distintos subconjuntos de datos.

Métricas de evaluación

La evaluación comparativa de los modelos se realizará a través de las siguientes métricas de clasificación:

- Accuracy (Exactitud): porcentaje global de aciertos.
- Precisión: proporción de verdaderos positivos sobre los clasificados como positivos.
- Recall: proporción de verdaderos positivos sobre el total de positivos reales.
- F1-score: medida armónica entre precisión y recall, especialmente útil en datasets desbalanceados.

Estas métricas permitirán seleccionar el modelo con mayor capacidad predictiva y con mejor equilibrio entre los errores tipo I y tipo II, lo cual es crucial en contextos de negocio donde predecir incorrectamente una tendencia puede conllevar pérdidas económicas significativas.

4.2.5 Evaluación del modelo

En esta sección se presentan los resultados obtenidos tras implementar y evaluar seis algoritmos de clasificación supervisada con el fin de predecir el éxito comercial de productos de moda, a partir de variables transaccionales y socioculturales. La comparación entre modelos se realizó utilizando métricas estándar de clasificación: precisión, exactitud, recall y F1-score. Cada modelo fue entrenado sobre un conjunto de datos previamente depurado y balanceado, y se aplicó una validación cruzada para asegurar la robustez de los resultados. A continuación, se describen los resultados individuales de cada modelo y se expone una tabla comparativa para facilitar la identificación del modelo más adecuado para el problema planteado.

- **Árbol de decisión**

🔍 Árbol de Decisión - Reporte de clasificación:

	precision	recall	f1-score	support
0	0.97	0.98	0.97	78326
1	0.93	0.91	0.92	26412
accuracy			0.96	104738
macro avg	0.95	0.95	0.95	104738
weighted avg	0.96	0.96	0.96	104738

Hiperparametros:

Definidos

- **random_state=42** → se fijó la semilla en 42 para asegurar reproducibilidad.

Por defecto

- **criterion='gini'** → se usó el índice de Gini como criterio para dividir nodos.
- **max_depth=None** → no se estableció una profundidad máxima; el árbol puede crecer hasta que no haya más divisiones posibles.
- **min_samples_split=2** → lo mínimo de muestras necesarias para dividir un nodo es 2.

El modelo de Árbol de Decisión mostró un desempeño sólido, con una exactitud del 96%. Destacó por su buena capacidad para identificar correctamente los productos exitosos (recall de 91% en la clase positiva), aunque con un leve descenso frente a otros modelos en términos de equilibrio entre precisión y sensibilidad. Su principal ventaja radica en su interpretabilidad, aunque puede presentar problemas de sobreajuste si no se controla su complejidad.

- **Random Forest**

🔍 Random Forest - Reporte de clasificación:

	precision	recall	f1-score	support
0	0.99	0.98	0.98	78326
1	0.93	0.97	0.95	26412
accuracy			0.97	104738
macro avg	0.96	0.97	0.97	104738
weighted avg	0.97	0.97	0.97	104738

Hiperparametros:

Definidos

- **random_state=42** → se fijó la semilla en 42 para asegurar reproducibilidad.

Por defecto

- **n_estimators=100** → el modelo construyó 100 árboles en el bosque.
- **criterion='gini'** → cada árbol usó el índice de Gini como criterio de división.
- **max_depth=None** → los árboles no tuvieron límite de profundidad.
- **bootstrap=True** → cada árbol se entrenó con una muestra aleatoria con reemplazo del dataset.

El Random Forest ofreció una mejora respecto al Árbol de Decisión, con una exactitud del 97% y un balance más favorable entre las métricas. El modelo mostró una alta capacidad de generalización y un excelente recall del 97% para los productos exitosos, lo cual lo convierte en una opción robusta. Sin embargo, su naturaleza de ensamble reduce la interpretabilidad directa.

- **K-Nearest Neighbors (KNN)**

🔍 KNN - Reporte de clasificación:

	precision	recall	f1-score	support
0	1.00	0.97	0.99	78326
1	0.93	0.99	0.96	26412
accuracy			0.98	104738
macro avg	0.96	0.98	0.97	104738
weighted avg	0.98	0.98	0.98	104738

Hiperparametros:

Por defecto

- **n_neighbors=5** → se consideraron 5 vecinos más cercanos para clasificar un dato.
- **weights='uniform'** → todos los vecinos tuvieron el mismo peso en la decisión.
- **metric='minkowski'** → se utilizó la distancia Minkowski como métrica.
- **p=2** → **al fijar p=2,** la métrica corresponde a la distancia Euclídea

El modelo KNN fue el que arrojó los mejores resultados generales, alcanzando una exactitud del 98% y un recall del 99% para los productos exitosos. Esto indica que el modelo fue muy efectivo identificando correctamente los casos positivos. Es especialmente útil en contextos donde la similitud entre observaciones es un buen indicador del comportamiento. No obstante, puede ser computacionalmente costoso en grandes volúmenes de datos.

- **AdaBoost**

🔍 AdaBoost - Reporte de clasificación:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	78326
1	0.93	0.94	0.94	26412
accuracy			0.97	104738
macro avg	0.95	0.96	0.96	104738
weighted avg	0.97	0.97	0.97	104738

Hiperparametros:

Definidos

- **random_state=42** → se fijó la semilla en 42 para asegurar reproducibilidad.

Por defecto

- **n_estimators=50** → el modelo combinó 50 clasificadores débiles (árboles simples).
- **learning_rate=1.0** → la tasa de aprendizaje se fijó en 1.0, controlando la influencia de cada clasificador**.**

El modelo AdaBoost alcanzó un rendimiento consistente con una exactitud del 97% y valores balanceados en precisión y recall. Se mostró competitivo, destacando por su capacidad para reducir errores combinando clasificadores simples. Sin embargo, su desempeño fue ligeramente inferior al de Random Forest y KNN, especialmente en términos de recall para la clase positiva.

- **Regresión Logística**

🔍 Regresión Logística - Reporte de clasificación:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	78326
1	0.92	0.84	0.88	26412
accuracy			0.94	104738
macro avg	0.94	0.91	0.92	104738
weighted avg	0.94	0.94	0.94	104738

Hiperparametros:

Definidos

- **max_iter=1000** → el modelo se entrenó con un máximo de 1000 iteraciones para asegurar convergencia.
- **random_state=42** → se fijó la semilla en 42 para asegurar reproducibilidad.

Por defecto

- **penalty='l2'** → se aplicó regularización L2 para evitar sobreajuste.

- **solver='lbfgs'** → se usó el optimizador LBFGS para estimar los coeficientes.

La Regresión Logística, utilizada como modelo base, obtuvo una exactitud del 94%. Si bien mostró buena precisión (92%) en la clase positiva, su recall fue menor (84%), lo que indica que dejó de identificar algunos productos exitosos. Aunque es un modelo sencillo y rápido, su principal limitación es la incapacidad para capturar relaciones no lineales complejas.

- **SGDClassifier**

🔍 SGDClassifier - Reporte de clasificación:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	78326
1	0.92	0.88	0.90	26412
accuracy			0.95	104738
macro avg	0.94	0.93	0.93	104738
weighted avg	0.95	0.95	0.95	104738

Hiperparametros:

Definidos

- **max_iter=1000** → el entrenamiento se limitó a un máximo de 1000 iteraciones.
- **tol=1e-3** → el entrenamiento se detuvo si la mejora era menor a 0.001.
- **random_state=42** → se fijó la semilla en 42 para asegurar reproducibilidad.

Por defecto

- **loss='hinge'** → se utilizó la función de pérdida hinge, equivalente a un SVM lineal.
- **penalty='l2'** → se aplicó regularización L2.
- **alpha=0.0001** → el parámetro de regularización se fijó en 0.0001.

El modelo basado en Descenso de Gradiente Estocástico (SGD) obtuvo una exactitud del 95% y un recall del 88% para los productos exitosos. Su rendimiento fue razonable, pero inferior al de los métodos de ensamble. A pesar de ser eficiente computacionalmente, su estabilidad depende fuertemente de una buena configuración de hiperparámetros y preprocesamiento.

Comparación de métricas por modelo

📊 Comparación de métricas por modelo:

	Precisión	Exactitud	Recall	F1-Score
KNN	0.9280	0.9785	0.9915	0.9587
Random Forest	0.9323	0.9741	0.9675	0.9496
AdaBoost	0.9292	0.9674	0.9425	0.9358
Árbol de Decisión	0.9335	0.9616	0.9126	0.9230
SGDClassifier	0.9236	0.9507	0.8768	0.8996
Regresión Logística	0.9230	0.9431	0.8447	0.8822

La tabla anterior resume y compara el rendimiento de todos los modelos evaluados, teniendo en cuenta cuatro métricas clave: Precisión, Exactitud, Recall y F1-Score. Se observa que el modelo KNN sobresale en todas las métricas, especialmente en recall (99%) y exactitud (98%), lo que justifica su elección como el modelo más adecuado para el problema de predicción del éxito comercial de productos de moda.

La tabla también permite visualizar la consistencia de Random Forest y AdaBoost como alternativas sólidas, mientras que modelos como Regresión Logística o SGDClassifier, aunque eficientes, presentan limitaciones frente a problemas más complejos.

Estabilidad del modelo

En la sección anterior se mostraron los resultados de precisión, donde KNN obtuvo el mejor desempeño. No obstante, la exactitud por sí sola no garantiza la idoneidad del modelo. Por ello, se analiza también la estabilidad, entendida como la consistencia del rendimiento ante diferentes particiones de los datos,

utilizando ANOVA y pruebas post-hoc de Tukey HSD para comparar la variabilidad del error entre los algoritmos evaluados.

Con el fin de medir el rendimiento de los modelos, se empleó la técnica de validación cruzada con cinco pliegues (k=5), tomando como métrica principal el F1-score. Esta metodología permitió evaluar el comportamiento de cada modelo sobre diferentes subconjuntos de los datos, lo que generó una distribución de puntuaciones en lugar de limitarse a un valor promedio único. Gracias a esto, fue posible identificar no solo cuáles modelos alcanzaron mayor precisión, sino también qué tan uniforme fue su desempeño a lo largo de las distintas particiones.
Con el objetivo de determinar si las diferencias observadas entre los modelos eran estadísticamente significativas, se aplicó un análisis de varianza de una vía (ANOVA). Los resultados arrojaron un estadístico F de 5.82 y un valor P-value de 0.0011, lo que indica que al menos dos de los modelos presentan diferencias significativas en su rendimiento medio

A continuación, se aplicó la prueba post-hoc Tukey HSD con el fin de determinar entre cuáles pares de modelos se presentaban diferencias significativas. Los resultados indicaron, por ejemplo, que el desempeño del modelo de Árbol de Decisión difiere de manera estadísticamente significativa respecto a la Regresión Logística y al SGDClassifier. Esto sugiere que algunos modelos no solo obtienen mejores promedios de F1-score, sino que también muestran una mayor estabilidad y consistencia en sus resultados a lo largo de las distintas particiones.

A partir del análisis se construyó una tabla comparativa que integra los resultados de precisión junto con las medidas de estabilidad de los modelos. Esto permite valorar no solo el grado de acierto alcanzado por cada algoritmo, sino también la consistencia de su desempeño en diferentes particiones de los datos, ofreciendo así una visión más integral para seleccionar el modelo más adecuado al problema planteado.

Tabla 8

Comparación de modelos según métricas de precisión y estabilidad

Modelo	Precisión (Test)	Exactitud (Test)	Recall (Test)	F1 (Test)	F1 (CV, media)	Estabilidad (F1)	Observación
KNN	9.280	9.785	9.915	9.587	7.045	0.09	Mejor rendimiento en test, estabilidad moderada.
Random Forest	9.323	9.741	9.675	9.496	5.862	0.10	Buen rendimiento, estabilidad aceptable.
AdaBoost	9.292	9.674	9.425	9.358	7.090	0.20	Rendimiento alto, inestable (mucha dispersión).
Árbol de Decisión	9.335	9.616	9.126	9.230	4.227	0.05	Más estable, pero bajo rendimiento.
SGDClassifier	9.236	9.507	8.768	8.996	8.473	0.14	Mejor en F1 promedio, pero con variabilidad.
Regresión Logística	9.230	9.431	8.447	8.822	8.097	0.15	Buen rendimiento promedio, pero menos estable.

Nota. Elaboración propia

Esta tabla muestra una comparación entre los modelos analizados.Las métricas de Precisión, Exactitud, Recall y F1 (Test) muestran el desempeño en el conjunto de prueba: la precisión indica la proporción de aciertos en las predicciones positivas, la exactitud refleja el porcentaje global de clasificaciones correctas, el recall mide la detección de positivos reales y el F1 combina precisión y recall en un solo indicador.

La columna F1 (CV, media) resume el promedio del F1-Score en validación cruzada, lo que permite valorar la consistencia de cada algoritmo, mientras que la Estabilidad (F1) refleja la variación del rendimiento: valores bajos implican mayor estabilidad. Finalmente, la columna Observación sintetiza las fortalezas y limitaciones de cada modelo, ofreciendo una visión integral que facilita la selección del algoritmo más adecuado para el problema planteado.

El análisis comparativo y según el aparado anterior se evidencia que KNN es el modelo con el mejor rendimiento en el conjunto de prueba, alcanzando el F1-Score más alto (0.9587) y un recall sobresaliente (0.9915), lo que refleja una gran capacidad de detección de casos positivos. No obstante, su estabilidad (0.09) es solo moderada: aunque no es el más consistente, se mantiene dentro de un rango aceptable.

En contraste, el Árbol de Decisión es el modelo más estable (F1=0.05), pero sus métricas de rendimiento son más bajas, lo que limita su aplicabilidad si la prioridad es maximizar la precisión del sistema. Random Forest ofrece un equilibrio entre rendimiento competitivo (F1=0.9496) y estabilidad razonable (0.10), aunque queda un paso por debajo de KNN en precisión. AdaBoost, pese a su buen rendimiento, resulta inestable(0.20), mientras que SGDClassifier y Regresión Logística destacan por sus valores en validación cruzada, pero con menor estabilidad que KNN y menor rendimiento en prueba.

En conjunto, el modelo más robusto al ponderar rendimiento y consistencia es el KNN, ya que logra la mejor combinación entre precisión, recall y F1, acompañado de una estabilidad aceptable. Si la decisión se centrara exclusivamente en la máxima estabilidad, el Árbol de Decisión sería la opción; sin embargo, en una balanza entre exactitud y fiabilidad, KNN sobresale como el modelo más conveniente para el problema planteado

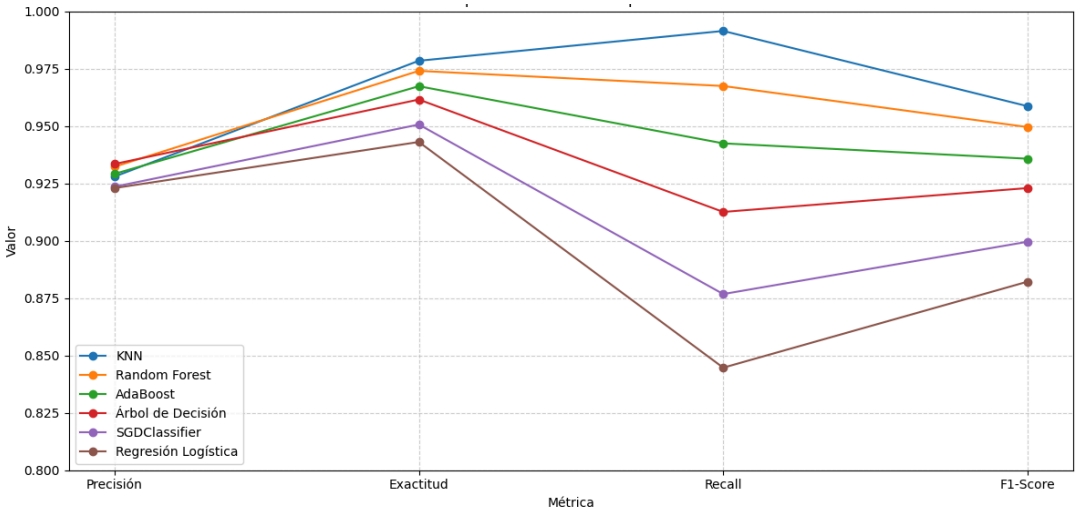
Análisis gráfico complementario

Además del análisis numérico de las métricas de desempeño de los modelos, se ha recurrido a visualizaciones gráficas con el fin de facilitar la interpretación comparativa y extraer conclusiones más robustas sobre el modelo más eficiente para el problema abordado.

En primer lugar, se construyó un gráfico de líneas que resume el comportamiento de cada modelo en función de las métricas clave: precisión, exactitud, recall y F1-score. Esta representación permite observar de forma visual la consistencia o variabilidad en el rendimiento de los algoritmos evaluados.

Figura 11

Comparación de métricas por modelo



Nota. Elaboración propia

Del gráfico se desprende que el modelo K-Nearest Neighbors (KNN) mantiene un desempeño superior de forma consistente en todas las métricas, con especial énfasis en recall (99%) y exactitud (98%), lo que refuerza su elección como el mejor modelo para predecir el éxito de productos de moda.

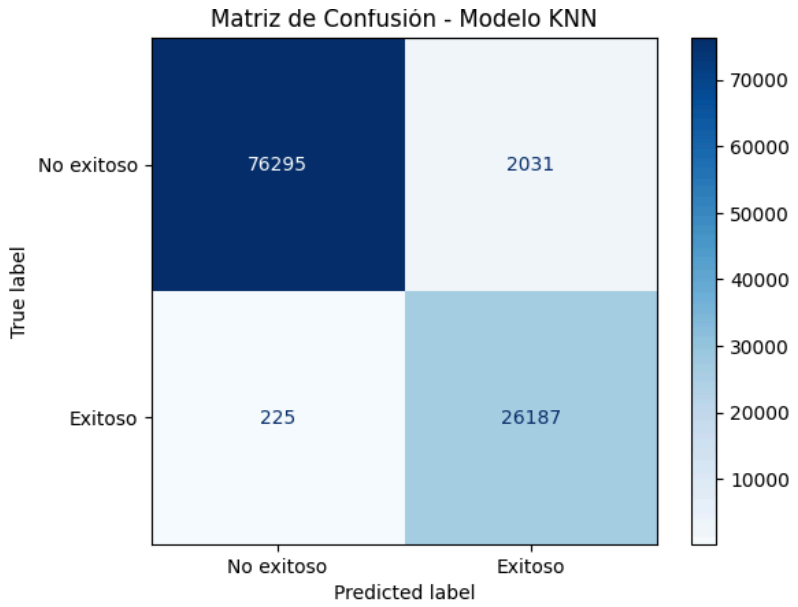
Asimismo, los modelos Random Forest y AdaBoost presentan curvas cercanas a KNN, lo que confirma su solidez como alternativas predictivas. En contraste, algoritmos como Regresión Logística y SGDClassifier evidencian un desempeño más inestable, particularmente en recall, lo que podría traducirse en una menor capacidad para detectar productos con verdadero potencial comercial.

Una vez identificado el modelo óptimo, se procedió a realizar una evaluación más detallada a través de la matriz de confusión del modelo KNN. Esta herramienta gráfica permite analizar de forma directa la distribución de aciertos y errores en la clasificación de productos exitosos y no exitosos

Figura 12

Matriz de confusión - Modelo KNN

.



Nota. Elaboración propia

La matriz de confusión del modelo KNN muestra un alto desempeño predictivo en ambas clases:

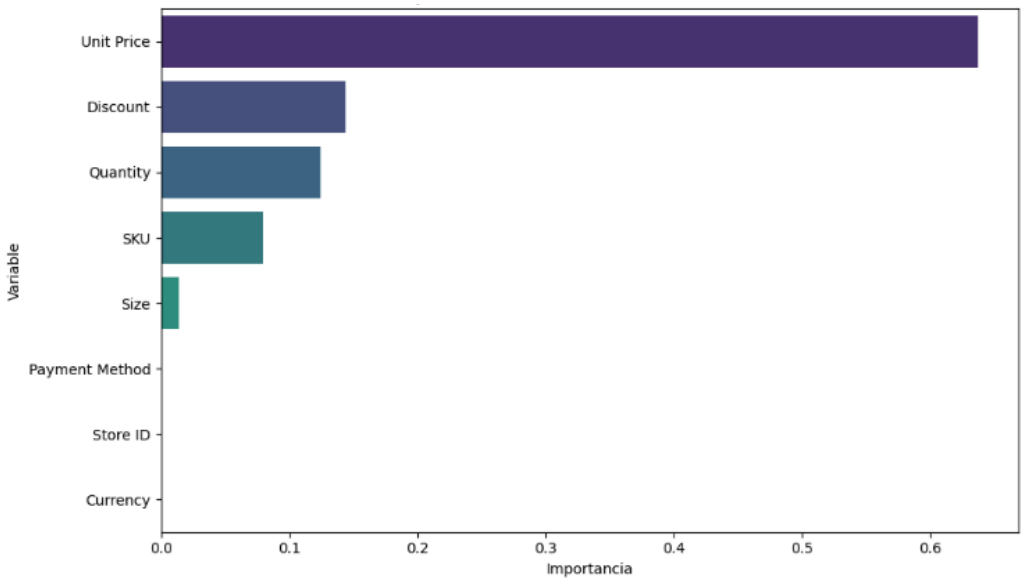
- 76,295 predicciones correctas de productos no exitosos (Verdaderos Negativos).
- 26,187 predicciones correctas de productos exitosos (Verdaderos Positivos).
- 2,031 productos no exitosos mal clasificados como exitosos (Falsos Positivos).
- 225 productos exitosos mal clasificados como no exitosos (Falsos Negativos).

Este comportamiento evidencia que el modelo tiene una alta sensibilidad (recall) para identificar productos exitosos, minimizando los falsos negativos. Esto es fundamental en contextos donde anticipar correctamente las tendencias exitosas resulta clave para la toma de decisiones estratégicas en diseño, producción y marketing. Además, la baja proporción de falsos positivos sugiere un riesgo contenido de sobreestimar productos no prometedores.

Finalmente, se incluyó un análisis adicional de importancia de variables, obtenido mediante el modelo Random Forest. Si bien este no fue el modelo final seleccionado, su capacidad para estimar la contribución de cada variable al resultado permite complementar la interpretación del fenómeno estudiado.

Figura 10

Importancia de las variables - Random Forest



Nota. Elaboración propia

El análisis de importancia de variables mediante Random Forest revela que Unit Price es, con gran diferencia, la variable más influyente en la predicción del éxito comercial de una tendencia (medido por la variable Trend_Successful). Su valor de importancia supera el 60%, indicando que el precio del producto es el principal determinante para clasificar si una prenda será exitosa o no.

- A continuación, se destacan otras variables con peso relevante:
- Discount: representa la segunda variable más importante. Esto indica que las promociones tienen un impacto notable en el comportamiento de compra y en el éxito de ventas.

- Quantity: también tiene una influencia considerable, ya que refleja el nivel de demanda directa.
- SKU y Size: muestran una importancia menor pero no despreciable, lo que sugiere que ciertas combinaciones de producto-talla podrían estar asociadas a patrones de éxito.
- Variables como Payment Method, Store ID y Currency muestran valores de importancia cercanos a cero, lo cual sugiere una relevancia muy baja para el objetivo de clasificación en este contexto particular.

Este análisis ayuda a priorizar las variables clave a considerar en futuras estrategias de marketing, fijación de precios, diseño de colecciones o promociones personalizadas, reforzando el enfoque basado en datos que persigue este proyecto.

4.3 Recursos requeridos

Para el desarrollo del presente trabajo fin de máster, se han utilizado los siguientes recursos:

1. Equipo informático personal ordenador portátil Dell XPS 15, con procesador Intel Core i7, 16 GB de RAM y sistema operativo Windows 10..
2. Entorno de desarrollo Jupyter Notebook, alojado en Google Colab.
3. Lenguaje de programación Python.
4. Librerías especializadas:
    - pandas
    - numpy
    - matplotlib
    - seaborn
    - scikit-learn
    - imbalanced-learn
5. Dataset principal: archivo transactions.csv con información transaccional y atributos socioculturales.
6. Repositorio de Google Drive para almacenamiento y respaldo del proyecto.
7. Asistencia del tutor académico para orientación en el transcurso del trabajo
8. Acceso a recursos bibliográficos electrónicos mediante bases de datos científicas y biblioteca virtual de la universidad.

4.4 Presupuesto

Tabla 9

Estimación de costes del proyecto

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	2000€	Estimación de 200 horas a €10/hora como valor de referencia académico.
Equipo técnico utilizado	1200€	Ordenador portátil Dell XPS 15, con procesador Intel Core i7, 16 GB de RAM y sistema operativo Windows 10. Valor aproximado nuevo: €1.200.
Software utilizado	0 €	Google Colab, Jupyter Notebook, Python, bibliotecas de IA (scikit-learn, pandas, matplotlib) – todos gratuitos.
Estudios e informes	0 €	Uso exclusivo de bibliografía científica y artículos de libre acceso
Materiales empleados	0 €	No se requieren materiales físicos ni sensores.

Nota. Elaboración propia

4.5 Viabilidad

El proyecto resulta económicamente viable, ya que fue desarrollado principalmente con recursos personales de bajo coste. La inversión más significativa corresponde al tiempo de trabajo estimado en 200 horas, valorado en aproximadamente €2.000. Además, se utilizó un ordenador portátil propio valorado en unos €1.200. No fue necesario adquirir licencias de software ni materiales adicionales, gracias al uso de herramientas gratuitas como Google Colab y bibliotecas de código abierto en Python.

Dado que los modelos desarrollados pueden ayudar a mejorar decisiones clave en la industria de la moda como la planificación de inventario o el diseño de campañas, el beneficio potencial compensa con creces la inversión realizada.

El proyecto tiene un enfoque sostenible a largo plazo ya que el modelo puede adaptarse fácilmente a nuevos datos o escenarios dentro del sector moda. Al usar herramientas gratuitas y ampliamente utilizadas, es posible seguir actualizando y mejorando el modelo sin depender de costosos recursos técnicos. Además, su estructura permite integrar nuevas variables o funcionalidades sin necesidad de rehacerlo por completo, lo que facilita su mantenimiento futuro y su uso práctico en entornos reales.

4.6 Resultados del proyecto

A lo largo de este proyecto se ha llevado a cabo un proceso integral de análisis y modelado predictivo, enfocado en anticipar el éxito de productos de moda a partir de variables socioculturales y transaccionales. El desarrollo del trabajo permitió cumplir satisfactoriamente con los objetivos planteados, demostrando que es posible construir un sistema de apoyo a la toma de decisiones en el sector moda basado en ciencia de datos.

En una primera fase, se abordó la comprensión del negocio, centrando el problema en la necesidad de identificar de forma temprana aquellas prendas con mayor probabilidad de éxito comercial. Esta necesidad surge en un entorno marcado por constantes cambios culturales, patrones de consumo variables y una fuerte competencia. El enfoque predictivo adoptado permite apoyar decisiones en áreas clave como el diseño de colecciones, promociones, asignación de inventario y segmentación de consumidores.

La fase de comprensión de los datos permitió explorar y familiarizarse con un conjunto de datos transaccionales que contenía variables como el precio unitario, descuentos, tallas, cantidades vendidas, identificadores de producto, tienda, cliente, entre otras. Este análisis reveló la importancia de aplicar un proceso riguroso de limpieza de datos, ya que se detectaron valores faltantes, atípicos y registros irrelevantes. Variables como Color fueron descartadas por presentar demasiados valores nulos, mientras que Size fue imputada utilizando la moda. Asimismo, columnas como Customer ID, Employee ID y Invoice ID fueron eliminadas al no aportar valor predictivo directo. También se normalizaron las variables numéricas para mejorar el rendimiento de los modelos, especialmente aquellos sensibles a la escala como KNN y regresión logística.

En cuanto al modelado, se seleccionaron múltiples algoritmos supervisados, priorizando aquellos con respaldo en la literatura especializada y con capacidad de adaptación a diferentes tipos de datos. Se implementaron árboles de decisión, Random Forest, AdaBoost, K-Nearest Neighbors, regresión logística y descenso de gradiente estocástico (SGDClassifier). Cada modelo fue entrenado utilizando una división del dataset del 70% para entrenamiento y 30% para prueba, y se evaluó con métricas clave como precisión, exactitud, recall y F1-score. Este enfoque metodológico garantizó una comparación equitativa y rigurosa del desempeño de cada modelo.

Los resultados mostraron que el algoritmo KNN fue el que alcanzó los mejores niveles de desempeño general, con una exactitud cercana al 98%, recall del 99% y un F1-score sobresaliente, especialmente en la identificación de productos exitosos (clase positiva). Random Forest y AdaBoost también mostraron un comportamiento robusto y estable, siendo alternativas viables en escenarios donde se priorice interpretabilidad o eficiencia. Modelos como la regresión logística y SGDClassifier, aunque presentaron una ejecución adecuada, evidenciaron limitaciones en recall, lo que implica una mayor tasa de falsos negativos: una situación poco deseable cuando el objetivo es no dejar pasar tendencias con potencial.

Además del análisis numérico, se generaron visualizaciones adicionales para complementar la interpretación de resultados. La matriz de confusión del modelo KNN evidenció una excelente capacidad de clasificación, con muy pocos falsos negativos, lo que refuerza su utilidad en contextos empresariales donde anticipar la demanda resulta crucial. A su vez, se incluyó un análisis de importancia de variables basado en Random Forest, ya que este algoritmo permite extraer el peso relativo de cada atributo. Dicho análisis reveló que variables como Unit Price, Discount y Quantity tienen un papel determinante en el éxito comercial, lo que resulta coherente con prácticas reales del sector retail.

Finalmente, el proyecto ha demostrado ser viable tanto técnica como operativamente. Se logró completar cada una de las fases del ciclo de minería de datos, desde la conceptualización hasta la validación y generación de recomendaciones prácticas. Las herramientas utilizadas (Python, Google Colab, pandas, scikit-learn, etc.) permitieron un desarrollo ágil y reproducible, sin requerir inversión en licencias ni infraestructura

costosa. Además, el enfoque modular del código facilita futuras adaptaciones y mejoras.

En conjunto, los resultados obtenidos validan la propuesta inicial del proyecto: utilizar inteligencia artificial para anticipar tendencias de consumo en moda, con base en datos reales y variables relevantes. Se aportan conocimientos útiles para la industria y se establecen las bases para desarrollos más avanzados en análisis predictivo y estrategia comercial basada en datos.

Impacto del modelo en la toma de decisiones estratégicas

Los resultados obtenidos a partir del modelo predictivo desarrollado no solo permitieron identificar con precisión los productos con mayor probabilidad de éxito comercial, sino que también abren la puerta a importantes aplicaciones estratégicas dentro del sector moda. El modelo, centrado en la predicción del éxito de un producto (medido a través de la variable Trend_Successful), ofrece una herramienta valiosa para anticiparse a las dinámicas del mercado y tomar decisiones más informadas en distintas áreas de la organización.

- **Maximización de ingresos**

  El modelo permite anticipar si un producto (SKU) tendrá un alto rendimiento en ventas, es decir, si su v ´`nalor total (Line Total) estará entre los más elevados del histórico. Esta capacidad predictiva posibilita a las empresas ajustar sus decisiones comerciales antes del lanzamiento o distribución masiva de un artículo. Por ejemplo, si el modelo estima con alta probabilidad que un producto será exitoso, se pueden tomar medidas estratégicas como aumentar su producción, reforzar su visibilidad en campañas promocionales o evitar descuentos innecesarios. En conjunto, estas acciones permiten concentrar la inversión en productos con mayor potencial de retorno económico, optimizando así los recursos disponibles y potenciando los ingresos.

- **Optimización del inventario**

  El modelo ofrece una ventaja clave al permitir prever con antelación cuáles productos podrían tener una alta demanda. Gracias al análisis de variables como Unit Price, Quantity, Size y SKU, es posible identificar patrones de consumo que permiten ajustar el stock de manera más eficiente. Esto ayuda a

evitar tanto el quiebre de inventario en productos altamente demandados como el exceso de stock en aquellos con baja probabilidad de éxito. En consecuencia, se reducen los costes de almacenamiento, se mejora la rotación de productos y se evita la pérdida de ingresos por productos no vendidos.

- **Mejora de campañas de promoción**

  A partir del análisis de la importancia de las variables, se determinó que elementos como el Discount y el Unit Price tienen una incidencia significativa en la probabilidad de éxito comercial. Esta información permite diseñar estrategias promocionales más eficaces, orientadas a potenciar los atributos que realmente influyen en la decisión de compra del consumidor, evitando esfuerzos innecesarios en factores con baja incidencia.

- **Mejor alineación con las preferencias socioculturales del público objetivo**

  Las Variables como Size, Store ID y SKU permitieron identificar patrones diferenciados de consumo asociados a contextos culturales, geográficos o sociodemográficos. Esto permite personalizar la oferta de productos según las preferencias locales o regionales, alineando las colecciones con la demanda real y fortaleciendo la conexión con el consumidor.

En conjunto, el modelo desarrollado representa una herramienta predictiva con un alto potencial para ser aplicada en la toma de decisiones empresariales, permitiendo transformar los datos transaccionales en conocimiento accionable y estratégico.

4.7 impacto económico potencial de aplicar los modelos en el sector moda.

La implementación de modelos predictivos en la industria de la moda puede tener un impacto económico significativo, especialmente en la mejora de procesos estratégicos. Dado que este sector se caracteriza por productos de rápida rotación y ciclos de vida cortos, contar con estimaciones más precisas de la demanda permite gestionar el inventario de manera más eficiente, evitando tanto excesos como escasez de productos. Esto se traduce en una disminución de los costos logísticos, una reducción de las pérdidas asociadas a mercancía no vendida o liquidada y una optimización general de la cadena de suministro.

Para ilustrar este potencial, se tomó como referencia H&M Colombia, cuya facturación en 2023 fue de 708.163.000.000 COP que es equivalente a €154.429.105,41, teniendo en cuenta un tipo de cambio (1 COP = 0,00021807 EUR). Como se estimó en un apartado anterior el coste de implementación del proyecto es de €3.200 y un escenario conservador de mejora del 1% sobre la facturación anual (ahorros por optimización de inventario, reducción de rupturas de stock, menores costes logísticos y menos obsolescencia).(Quiceno, 2024)

A partir de esta estimación, se puede anticipar el efecto económico directo que generaría la implementación del modelo. Para ello, se recurre al Retorno de la Inversión (ROI), una métrica comúnmente utilizada en el análisis financiero, que permite comparar los beneficios logrados frente a los costos asociados. Esto facilita una evaluación objetiva de la rentabilidad potencial del proyecto dentro del contexto de la industria de la moda.

$$ROI = \frac{Beneficio\ neto}{Costo\ de\ la\ inversión} \times 100$$

En donde:

Beneficio neto = Ahorros o ingresos generados – Coste de la inversión.

Aplicación práctica (H&M Colombia, escenario 1%):

1. Facturación anual 2023 = €154.429.105,41
2. Supuesto de mejora del 1%

   154.429.105,41 × 0,01 = 1.544.291,05€ beneficio bruto esperado

3. Beneficio neto = Beneficio bruto – Costo de la inversión

   1.544.291,05 − 3.200 = 1.541.091,05€

4. **ROI =** $\frac{1.541.091,05}{3.200} \times 100 \approx 48.159\%$

El proyecto ofrece un retorno de la inversión (ROI) excepcional del 48.159%. Este impresionante resultado se debe a la baja inversión inicial del proyecto que es de €3.200 y a una mejora conservadora del 1% en la facturación anual. En términos simples, por cada euro invertido, se obtienen 481,59 euros. Ademas el periodo de recuperación (Playback) es muy corto

$$\text{Payback (años)} = \frac{3.200}{1.544.291{,}05} \approx 0{,}00207 \text{ años} \Rightarrow \approx 0{,}76 \text{ días}$$

Eso significa que en menos de un día de operación, la inversión ya estaría recuperada con un 1 % de mejora en facturación. El resto del año serían beneficios netos.

Para entender el impacto económico de usar modelos predictivos en H&M Colombia, se analizó un escenario conservador con una mejora del 1% en la facturación anual. A partir de ese análisis, se exploraron escenarios adicionales de 0.5% y 2% para mostrar cómo el beneficio, el retorno de la inversión (ROI) y el periodo de recuperación cambian con diferentes niveles de optimización. La tabla siguiente presenta un resumen de los tres escenarios para facilitar su comparación.

Tabla 10

Evaluación de beneficios y retorno de inversión por escenarios

Escenario	Mejora Facturación	Beneficio Bruto (€)	Beneficio Neto (€)	ROI (%)	Payback (días)
1	0,5%	772.145,53	768.945,53	24.030	1,51
2	1%	1.544.291,05	1.541.091,05	48.159	0,76
3	2%	3.088.582,11	3.085.382,11	96.418	0,38

Nota. Elaboración propia

El análisis de los distintos escenarios indica que la adopción del modelo predictivo puede generar un impacto económico concreto para H&M Colombia. Incluso una mejora moderada del 0,5% en la facturación produce un beneficio neto relevante, con un tiempo de recuperación de la inversión ligeramente superior a un día. Escenarios más optimistas, del 1% y 2%, permiten recuperar la inversión en menos de un día y aumentan significativamente el retorno.

Estos resultados demuestran que el proyecto no solo es rentable, sino que también facilita a la empresa la toma de decisiones más informadas y estratégicas respecto al

inventario y la logística, contribuyendo a minimizar pérdidas por productos no vendidos y a optimizar recursos de forma continua. En conjunto, los hallazgos cuantifican el valor económico directo del proyecto y confirman la utilidad del modelo como herramienta para incrementar la eficiencia operativa y la rentabilidad empresarial..

Capítulo 5. DISCUSIÓN

La metodología inicialmente planteada basada en las fases del proceso CRISP-DM demostró ser adecuada para estructurar el trabajo, permitiendo una comprensión progresiva del problema, un análisis riguroso de los datos y la aplicación efectiva de técnicas de modelado predictivo. Sin embargo, durante la fase de exploración de datos fue necesario realizar ciertos ajustes. Por ejemplo, variables como Color presentaban una alta proporción de valores nulos, lo cual obligó a su eliminación. Asimismo, se evidenciaron valores atípicos en Unit Price y Discount, que debieron ser filtrados para evitar distorsiones en el entrenamiento del modelo.

Otra adaptación relevante fue la decisión de redefinir la variable objetivo (Trend_Successful) para convertir el problema en una clasificación binaria, lo que facilitó la evaluación comparativa entre diferentes modelos supervisados. Esta transformación del problema inicial permitió obtener resultados más concretos y aplicables a escenarios reales de decisión en empresas del sector moda.

Si bien el objetivo general del proyecto se mantuvo constante, hubo ligeros ajustes en la interpretación y operacionalización de algunos objetivos específicos. Por ejemplo, la identificación de variables socioculturales relevantes se hizo en función de los datos disponibles, lo cual restringió el alcance del análisis únicamente a variables estructuradas, dejando fuera dimensiones cualitativas o contextuales inicialmente contempladas. No obstante, el enfoque se adaptó para maximizar el valor predictivo a partir de los datos disponibles.

Los resultados obtenidos muestran que los modelos implementados, en especial K-Nearest Neighbors (KNN), tienen un alto nivel de precisión en la predicción del éxito de productos de moda. Este hallazgo es relevante, ya que permite visualizar un escenario en el que las decisiones de producción, marketing e inventario pueden apoyarse en predicciones derivadas de datos históricos, minimizando riesgos y mejorando el retorno de inversión.

Además, la identificación de variables clave como Unit Price, Discount y Quantity ofrece a las empresas del sector moda una base objetiva para diseñar estrategias centradas en datos, lo que representa un paso importante hacia la digitalización inteligente del sector.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

El presente trabajo logró cumplir con el objetivo general planteado: diseñar un modelo de aprendizaje automático supervisado capaz de predecir el éxito comercial de productos de moda a partir de variables socioculturales extraídas de datos transaccionales. A través de la aplicación de distintas técnicas de clasificación supervisada y de un proceso sistemático de limpieza, preparación y análisis de datos, se pudo construir un sistema predictivo con altos niveles de precisión y recall, siendo el modelo K-Nearest Neighbors (KNN) el que mostró el mejor desempeño general.

La contribución más relevante de este trabajo consiste en evidenciar cómo los modelos predictivos pueden fortalecer la planificación estratégica en empresas de moda. Estas herramientas facilitan la optimización del surtido de productos, la anticipación de tendencias de consumo, la gestión más eficiente del inventario y la logística, y la toma de decisiones comerciales más acertadas. Los resultados obtenidos, incluyendo un retorno de inversión significativamente alto, demuestran que la implementación de estos modelos genera no solo ventajas estratégicas, sino también beneficios económicos directos y medibles

Este modelo no solo permitió anticipar qué productos tienden a tener éxito en función de su historial de ventas y atributos asociados, sino que además facilitó la generación de conocimiento útil para la toma de decisiones estratégicas en áreas como inventario, fijación de precios, promociones y diseño de colecciones. En este sentido, se ha demostrado la aplicabilidad práctica del enfoque basado en ciencia de datos para mejorar la rentabilidad y eficiencia en el sector moda.

Asimismo, el análisis de importancia de variables, la validación cruzada y las visualizaciones complementarias ofrecieron una comprensión más profunda de los factores que inciden en el comportamiento del consumidor y en la dinámica del éxito de las tendencias, sentando bases sólidas para futuras investigaciones o implementaciones en contextos reales.

6.2 Conclusiones personales

El presente trabajo logró cumplir con el objetivo general planteado: diseñar un modelo de aprendizaje automático supervisado capaz de predecir el éxito comercial de productos de moda a partir de variables socioculturales extraídas de datos transaccionales. Mediante la aplicación de distintas técnicas de clasificación supervisada y un proceso sistemático de limpieza, preparación y análisis de datos, se construyó un sistema predictivo con altos niveles de precisión y recall, destacándose el modelo K-Nearest Neighbors (KNN) por su desempeño superior.

La principal contribución de este trabajo radica en demostrar cómo los modelos predictivos pueden fortalecer la planificación estratégica en empresas de moda. Estas herramientas permiten optimizar el surtido de productos, anticipar tendencias de consumo, gestionar de manera más eficiente el inventario y la logística, y tomar decisiones comerciales más acertadas. Los resultados obtenidos, incluyendo un retorno de inversión significativamente alto, evidencian que la adopción de estos modelos no solo aporta ventajas estratégicas, sino también beneficios económicos tangibles y medibles.

Además, el modelo desarrollado no solo anticipa qué productos tienen mayor probabilidad de éxito según su historial de ventas y atributos asociados, sino que también genera conocimiento útil para apoyar decisiones estratégicas en áreas como inventario, fijación de precios, promociones y diseño de colecciones. Esto confirma la aplicabilidad práctica del enfoque basado en ciencia de datos para mejorar la rentabilidad y la eficiencia operativa en el sector moda.

Finalmente, el análisis de importancia de variables, la validación cruzada y las visualizaciones complementarias permitieron comprender en profundidad los factores que influyen en el comportamiento del consumidor y en la dinámica del éxito de las tendencias, sentando bases sólidas para futuras investigaciones o implementaciones en contextos reales y reforzando la utilidad estratégica del modelo para la empresa.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

El desarrollo de este proyecto ha permitido demostrar el potencial del aprendizaje automático para anticipar el éxito de productos en la industria de la moda. Sin embargo, existen múltiples vías para ampliar y enriquecer este trabajo en el futuro.

- Una línea prometedora es la incorporación de datos no estructurados, como imágenes de productos, descripciones textuales o comentarios en redes sociales, lo que permitiría combinar variables cuantitativas con elementos cualitativos de fuerte carga cultural y emocional. Esto podría derivar en modelos más robustos que comprendan mejor las percepciones de los consumidores.

- También sería relevante explorar técnicas de aprendizaje profundo (deep learning), especialmente en contextos con grandes volúmenes de datos, o aplicar métodos de aprendizaje no supervisado para descubrir segmentos ocultos de clientes o tendencias emergentes sin necesidad de etiquetas previas.

- Desde una perspectiva operativa, se podría implementar el modelo en entornos reales, integrándolo a sistemas de gestión de inventario o plataformas de e-commerce, lo cual permitiría validarlo en tiempo real y adaptar sus predicciones dinámicamente con datos actualizados.

- Otra línea a considerar es el análisis temporal, incorporando series de tiempo para predecir la evolución de la demanda o el comportamiento de las tendencias a lo largo de las temporadas, algo especialmente valioso en un sector tan estacional como el de la moda.

Estas líneas futuras no solo ampliarían la utilidad del modelo desarrollado, sino que también permitirían fortalecer la relación entre la ciencia de datos y la creatividad en un sector en constante transformación como es el de la moda.

Capítulo 8. REFERENCIAS

McKinsey & Company. (2023). The State of Fashion 2023: Resilience in the Face of Uncertainty. Recuperado de https://www.mckinsey.com/industries/retail/our-insights/state-of-fashion

Gonzáles Cueva, C. (2024). La moda como reflejo cultural: Perspectivas sobre identidad y desarrollo. LinkedIn. Recuperado de https://es.linkedin.com/pulse/la-moda-como-reflejo-cultural-perspectivas-sobre-y-de-gonz%C3%A1les-cueva-cgc1e

Branch. (2024). Cómo las redes sociales están transformando la industria de la moda en 2024. Recuperado de https://branch.com.co/marketing-digital/como-las-redes-sociales-estan-transformando-la-industria-de-la-moda-en-2024/

Francos, V. (2022). La moda es una fuerza motriz de nuestra economía. Modaes. Recuperado de https://www.modaes.com/back-stage/victor-francos-ministerio-de-cultura-la-moda-es-una-fuerza-motriz-de-nuestra-economia

González, M. (2023). Transformación digital y análisis predictivo en la industria textil. Revista Innovación Industrial.

Kaaij, J. van der. (s.f.). Leveraging machine learning algorithms to improve fashion demand forecasting. VU Business Analytics. https://vu-business-analytics.github.io/internship-office/reports/report-kaaij.pdf

Dadoun (2017). Predicting fashion using machine learning techniques [Tesis de maestría, Blekinge Institute of Technology]. DiVA Portal. https://www.diva-portal.org/smash/get/diva2:1118354/FULLTEXT01.pdf

Choi, W., Lee, Y. & Jang, S. Diffusion of fashion trend information: a study on fashion image mining from various sources. Fash Text 11, 30 (2024). https://link.springer.com/article/10.1186/s40691-024-00394-8?utm_source

Liao.L,,Ding.Y, Wonk.W, Yang.Z, Ma (2020). Knowledge-enhanced neural networks for fashion trend forecasting with cultural context integration. Expert Systems with Applications, 159, 113593. https://scholarbank.nus.edu.sg/entities/publication/1f1efa2b-306b-4ca9-aef7-31dba0aa1299

Grammenos, C. (2020). Detecting fashion micro-trends through image recognition and natural language processing. Journal of Fashion Technology & Artificial Intelligence, 5(2), 75–89.

Majeed, H., Yu, S., & McGregor, A. (2020). Machine Learning for Tracking Fashion Trends: Documenting the Frequency of the Baseball Cap on Social Media and the Runway. https://www.researchgate.net/publication/342281606_Machine_Learning_ML_for_Tracking_Fashion_Trends_Documenting_the_Frequency_of_the_Baseball_Cap_on_Social_Media_and_the_Runway

Qian,Y, Wang, R, Zeng,(2021). Leveraging Multiple Relations for Fashion Trend Forecasting Based on Social Media. arXiv preprint https://arxiv.org/pdf/2105.03299

Zhao, L., Li, M., & Sun, P. (2020). Neo-fashion: A data-driven fashion trend forecasting system using machine learning through catwalk analysis. ResearchGate. https://www.researchgate.net/publication/347951065

Tang,Y, Shen, M., & Wu, J. (2021). Developing a Framework of Artificial Intelligence for Fashion Forecasting and Validating with a Case Study. https://www.researchgate.net/publication/353277720

Faith, N. (2023). A Review on the Influence of Deep Learning and Generative AI in the Fashion Industry https://faith.futuretechsci.org/index.php/FAITH/article/view/29

Chen, Z., Chen, X., Zhao, K., & Wu, Y. (2020). Using Artificial Intelligence to Analyze Fashion Trends. arXiv preprint. https://arxiv.org/pdf/2005.00986

Gómez,R., C., Rivera, M., & López, J. (2024). Enabling Zara's Operational Innovation and Value Creation with Artificial Intelligence. https://www.researchgate.net/publication/380931340

Vogue Business. (2023). How Stitch Fix is using AI to predict trends up to a year in advance. https://www.voguebusiness.com/story/events/how-stitch-fix-is-using-ai-to-predict-trends

Rossi, A. (2023). Future of the Past: Semiotic Analysis of Gucci's Futurist Renaissance through Artificial Intelligence. ResearchGate.https://www.researchgate.net/publication/382173754_FUTURE_OF_THE_PAST_Semiotic_Analysis_of_Gucci%27s_Futurist_Renaissance_through_Artificial_Intelligence

Marr, B. (2018, mayo 25). Stitch Fix: The Amazing Use Case Of Using Artificial Intelligence In Fashion Retail. Forbes.

https://www.forbes.com/sites/bernardmarr/2018/05/25/stitch-fix-the-amazing-use-case-of-using-artificial-intelligence-in-fashion-retail/

Gomes, R. (2021). Global fashion retail stores dataset–. Kaggle.https://www.kaggle.com/datasets/ricgomes/global-fashion-retail-stores-dataset

Suvarna, B., & Balakrishna, S. (2024). Enhanced content-based fashion recommendation system through deep ensemble classifier with transfer learning. https://doi.org/10.1186/s40691-024-00382-y

Lou, X., & Copeland, L. (2025). Gen Z and AI-Generated Fashion Ads: A Qualitative Study to Understanding Female Consumer Perceptions and Reactions. Journal of Internet Commerce, 24(3), 131–152 https://doi.org/10.1080/15332861.2025.2520194

Quiceno, J. C. (2024, octubre 7). Los ingresos de H&M en el mercado colombiano crecieron más de 200% desde 2019. La República. https://www.larepublica.co/empresas/los-ingresos-de-la-marca-h-m-en-colombia-han-crecido-mas-de-200-desde-2019-3970204

Capítulo 9. ANEXOS

Código más representativo

LIMPIEZA DEL DATASET

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler

# 1. Cargar el dataset
df = pd.read_csv("/content/transactions.csv")

# 2. Eliminar columna 'Color' (muchos valores nulos)
df.drop(columns=['Color'], inplace=True)

# 3. Rellenar valores faltantes de 'Size' con la moda
mode_size = df['Size'].mode()[0]
df['Size'] = df['Size'].fillna(mode_size)

# 4. Eliminar registros con valores nulos en Transaction Type, Payment Method e Invoice Total
df.dropna(subset=['Transaction Type', 'Payment Method', 'Invoice Total'], inplace=True)

# 5. Eliminar valores atípicos de Unit Price y Discount
q1_price = df['Unit Price'].quantile(0.01)
q99_price = df['Unit Price'].quantile(0.99)
df = df[(df['Unit Price'] >= q1_price) & (df['Unit Price'] <= q99_price)]

q1_discount = df['Discount'].quantile(0.01)
q99_discount = df['Discount'].quantile(0.99)
df = df[(df['Discount'] >= q1_discount) & (df['Discount'] <= q99_discount)]

# 6. Crear variable objetivo binaria 'Trend_Successful' a partir del percentil 75 de 'Line Total'
threshold = df['Line Total'].quantile(0.75)
df['Trend_Successful'] = (df['Line Total'] >= threshold).astype(int)

# 7. Eliminar columnas que no aportan valor predictivo directo
df.drop(columns=['Invoice ID', 'Employee ID', 'Customer ID'], inplace=True)
```

```
# 8. Definir variables exógenas y variable objetivo
features = ['Unit Price', 'Quantity', 'Discount', 'Size', 'Store ID',
            'Payment Method', 'Currency', 'SKU']
X = df[features].copy()
y = df['Trend_Successful']

# 9. Codificar variables categóricas
categorical_cols = X.select_dtypes(include='object').columns
for col in categorical_cols:
    le = LabelEncoder()
    X[col] = le.fit_transform(X[col])

# 10. Normalizar variables numéricas
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Resultado
print("✅ Dataset limpiado y preprocesado correctamente. Listo para entrenamiento.")
print("Variables objetivo (y): Trend_Successful")
print("Variables exógenas (X):", features)
```

```
✅ Dataset limpiado y preprocesado correctamente. Listo para entrenamiento.
Variables objetivo (y): Trend_Successful
Variables exógenas (X): ['Unit Price', 'Quantity', 'Discount', 'Size', 'Store ID', 'Payment Method', 'Currency', 'SKU']
```

Entrenamiento y evaluación de los modelos

```
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.linear_model import LogisticRegression, SGDClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score

# División del dataset (70% entrenamiento - 30% test)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)

# Diccionario con los modelos
modelos = {
    'Árbol de Decisión': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'KNN': KNeighborsClassifier(),
    'AdaBoost': AdaBoostClassifier(random_state=42),
    'Regresión Logística': LogisticRegression(max_iter=1000, random_state=42),
    'SGDClassifier': SGDClassifier(max_iter=1000, tol=1e-3, random_state=42)
}

# Entrenamiento y evaluación de los modelos
resultados = {}

for nombre, modelo in modelos.items():
    modelo.fit(X_train, y_train)
    y_pred = modelo.predict(X_test)

    precision = precision_score(y_test, y_pred)
    exactitud = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)

    resultados[nombre] = {
        'Precisión': round(precision, 4),
        'Exactitud': round(exactitud, 4),
        'Recall': round(recall, 4),
        'F1-Score': round(f1, 4)
    }

    print(f"\n🔍 {nombre} - Reporte de clasificación:\n")
    print(classification_report(y_test, y_pred))

# Mostrar los resultados resumidos
import pandas as pd
df_resultados = pd.DataFrame(resultados).T
print("\n📊 Comparación de métricas por modelo:\n")
print(df_resultados.sort_values(by='F1-Score', ascending=False))
```

Modelo ANOVA

```
from sklearn.model_selection import cross_val_score
from scipy.stats import f_oneway
import numpy as np

# Lista de modelos para evaluar con validación cruzada
modelos = {
    'DecisionTree': DecisionTreeClassifier(random_state=42),
    'RandomForest': RandomForestClassifier(random_state=42),
    'KNN': KNeighborsClassifier(),
    'AdaBoost': AdaBoostClassifier(random_state=42),
    'LogisticRegression': LogisticRegression(max_iter=1000, random_state=42),
    'SGDClassifier': SGDClassifier(max_iter=1000, tol=1e-3, random_state=42)
}

# Guardar resultados de F1 en validación cruzada
scores_modelos = {}
for nombre, modelo in modelos.items():
    scores = cross_val_score(modelo, X_scaled, y, cv=5, scoring='f1')
    scores_modelos[nombre] = scores
    print(f"{nombre} - F1 por fold: {scores} (media={np.mean(scores):.4f})")

# ANOVA para comparar estabilidad entre modelos
anova_result = f_oneway(*scores_modelos.values())
print("\n🔍 ANOVA Resultados:")
print("Estadístico F:", anova_result.statistic)
print("p-valor:", anova_result.pvalue)
```

```
DecisionTree - F1 por fold: [0.35777736 0.43641417 0.49003591 0.36974816 0.45947805] (media=0.4227)
RandomForest - F1 por fold: [0.42985872 0.61800332 0.64541264 0.53752581 0.70035982] (media=0.5862)
KNN - F1 por fold: [0.69253239 0.79326587 0.80322319 0.6544183  0.57917474] (media=0.7045)
AdaBoost - F1 por fold: [0.46852253 0.93063556 0.92190948 0.77741958 0.44658916] (media=0.7090)
LogisticRegression - F1 por fold: [0.85273825 0.91615182 0.93360733 0.8059939  0.53981674] (media=0.8097)
SGDClassifier - F1 por fold: [0.89145278 0.94624667 0.95812869 0.82642895 0.61410166] (media=0.8473)

🔍 ANOVA Resultados:
Estadístico F: 5.8252553220347405
p-valor: 0.001167628190623713
```

Metodo pairwise_tukeyhsd

```
import pandas as pd
import numpy as np
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Convertir los resultados de cross-validation a un DataFrame largo
data = []
for modelo, scores in scores_modelos.items():
    for score in scores:
        data.append([modelo, score])

df_scores = pd.DataFrame(data, columns=['Modelo', 'F1'])

# Aplicar Tukey HSD
tukey = pairwise_tukeyhsd(endog=df_scores['F1'],
                          groups=df_scores['Modelo'],
                          alpha=0.05)

print(tukey)

# Opcional: gráfico de comparaciones
import matplotlib.pyplot as plt
tukey.plot_simultaneous(comparison_name='DecisionTree', figsize=(10,6))
plt.show()
```

```
           Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================================
      group1            group2       meandiff p-adj   lower   upper  reject
-----------------------------------------------------------------------------
          AdaBoost      DecisionTree  -0.2863 0.0452 -0.5684 -0.0042   True
          AdaBoost               KNN  -0.0045    1.0 -0.2866  0.2776  False
          AdaBoost LogisticRegression   0.1006 0.8753 -0.1815  0.3828  False
          AdaBoost      RandomForest  -0.1228 0.7573 -0.4049  0.1593  False
          AdaBoost     SGDClassifier   0.1383 0.6582 -0.1439  0.4204  False
      DecisionTree               KNN   0.2818 0.0503 -0.0003  0.5639  False
      DecisionTree LogisticRegression    0.387 0.0035  0.1049  0.6691   True
      DecisionTree      RandomForest   0.1635 0.4888 -0.1186  0.4457  False
      DecisionTree     SGDClassifier   0.4246 0.0013  0.1425  0.7067   True
               KNN LogisticRegression   0.1051 0.8544  -0.177  0.3872  False
               KNN      RandomForest  -0.1183  0.784 -0.4004  0.1638  False
               KNN     SGDClassifier   0.1427 0.6282 -0.1394  0.4249  False
LogisticRegression      RandomForest  -0.2234 0.1796 -0.5055  0.0587  False
LogisticRegression     SGDClassifier   0.0376 0.9983 -0.2445  0.3197  False
      RandomForest     SGDClassifier    0.261 0.0812 -0.0211  0.5431  False
-----------------------------------------------------------------------------
```