Trabajo Fin de Máster

Análisis predictivo para la optimización de estrategias comerciales en el sector retail de moda

Paulina Traba Piñeiro

Máster en Técnicas Estadísticas

Curso 2025-2026

Propuesta de Trabajo Fin de Máster

Título en galego: Análise preditivo para a optimización de estratexias comerciais no sector retail de moda
Título en español: Análisis predictivo para la optimización de estrategias comerciales en el sector retail de moda
English title: Predictive analytics for optimizing commercial strategies in the fashion retail sector
Modalidad: Modalidad B
Autor/a: Paulina Traba Piñeiro, Universidad de Santiago de Compostela
Director/a: Javier Roca Pardiñas, Universidad de Vigo
Tutor/a: Nicolás Sánchez Roel, Plexus Tech; Carlos Varela Sanjurjo, Plexus Tech
Breve resumen del trabajo: Este trabajo responde a la propuesta de la empresa Plexus Tech ante la necesidad de modelar la demanda a nivel de talla en el sector retail. Entre las diferentes líneas abordadas, se encuentran: el procesamiento y análisis de bases de datos mediante Python y SQL, la aplicación de técnicas de aprendizaje no supervisado para la clasificación de atributos de prendas y países, así como para entender patrones de consumo; y, finalmente, el desarrollo de un modelo de regresión para la predicción de la proporción de ventas por talla.
Recomendaciones:
Otras observaciones:

Don/doña Javier Roca Pardiñas, Catedrático de Universidad del Departamento de Estadística e Investigación Operativa de la Universidad de Vigo, don/doña Nicolás Sánchez Roel, Director Área Data de Plexus Techy don/doña Carlos Varela Sanjurjo, Gerente Área Data de Plexus Techinforman que el Trabajo Fin de Máster titulado

Análisis predictivo para la optimización de estrategias comerciales en el sector retail de moda

fue realizado bajo su dirección por don/doña Paulina Traba Piñeiro para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Don/doña Javier Roca Pardiñas y don/doña Paulina Traba Piñeiro

√ sí ☐ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En A Coruña, a 04 de Mayo de 2026.

El/la director/a:
Don/doña Javier Roca Pardiñas

ROCA PARDIÑAS JAVIER - 77592181W
Firmado digitalmente por ROCA PARDIÑAS JAVIER - 77592181W
Fecha: 2026.05.29 09:28:43 +02'00'

El/la tutor/a:
Don/doña Carlos Varela Sanjurjo

VARELA SANJURJO CARLOS MANUEL - 32683355X
Firmado digitalmente por VARELA SANJURJO CARLOS MANUEL - 32683355X
Fecha: 2026.05.29 22:32:26 +02'00'

El/la tutor/a:
Don/doña Nicolás Sánchez Roel

Firmado por SANCHEZ ROEL NICOLAS - ***5602** el día 30/05/2026 con un certificado emitido por AC FNMT Usuarios

El/la autor/a:
Don/doña Paulina Traba Piñeiro

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen

Resumen en español

El sector retail se enfrenta a una gran volatilidad e inestabilidad de sus productos, lo que dificulta la previsión de la demanda. Entre sus numerosos desafíos, destaca la necesidad de gestionar el inventario de forma muy precisa, incluso a nivel de tienda, producto y talla.

Es por ello que desde la empresa Plexus Tech se planteó la necesidad de estudiar el comportamiento del consumidor y modelar la demanda por tallas. De esta forma, se desarrolló un sistema de asignación de tallas que modela la probabilidad de elección del comprador. Para ello, se aplicaron técnicas de agrupamiento o *Clustering* para estudiar las similitudes en los patrones de demanda de tallas de distintos mercados, y también un modelo de Regresión Logística para predecir las distribuciones.

Finalmente, se validaron los modelos empleados y se realizó una simulación para comparar el rendimiento del modelo predictivo propuesto frente a una alternativa tradicional básica. Los resultados demostraron el alto impacto económico de una estimación acertada del surtido de tallas, optimizando la rentabilidad de las prendas más allá del volumen total vendido.

English abstract

The retail sector faces significant volatility and instability in its products, making it difficult to forecast demand. Among its many challenges, one that stands out is the need to manage inventory with a high degree of precision, including at the store, product, and size levels.

That is why Plexus Tech recognized the need to study consumer behavior and model demand by size. As a result, a size allocation system that models the buyer's choice probability was developed. To this end, Clustering techniques were applied to study similarities in size demand patterns across different markets, and a Logistic Regression model was also used to predict the distributions.

Finally, the applied models were validated, and a simulation was carried out to compare the performance of the proposed predictive model against a traditional baseline. The results demonstrated the significant financial impact of an accurate size assortment estimation, optimizing garment profitability beyond the total sales volume.

Capítulo 1

Introducción

1.1. Contexto y motivación

En el entorno del sector retail, la industria de la moda gestiona procesos muy exigentes y complejos, caracterizados por su gran velocidad y por la brevedad de sus productos. Tal y como describen Caro y Gallien (2010) [1], este modelo de negocio se caracteriza por ciclos de vida de productos extremadamente cortos y una rotación constante del surtido en tienda, lo que provoca que la planificación del inventario sea muy impredecible.

A diferencia de otros sectores con una demanda más estable, el mercado de la moda actual, a menudo denominado *fast fashion*, está sujeto a una alta volatilidad. Estas oscilaciones vienen impulsadas por cambios en las tendencias sociales, la incertidumbre en las preferencias del cliente y factores estacionales como el clima o las campañas comerciales, que pueden variar drásticamente en cuestión de semanas. Esto hace que la previsión de la demanda sea una tarea compleja y crítica (Liu et al., 2020).

El principal desafío logístico y financiero de este sector reside en encontrar un equilibrio entre la disponibilidad de los productos y la eficiencia del almacenamiento del inventario. Por un lado, subestimar la demanda provoca roturas de stock, es decir, situaciones en las que el producto no está disponible en el momento en que el cliente desea comprarlo. Por otro lado, estimar por encima de la demanda real genera un sobrante de prendas al final de la campaña, lo que en ocasiones obliga a vender el stock restante mediante rebajas o aplicando descuentos, penalizando el beneficio y provocando pérdidas debidas al coste del almacenamiento y el transporte.

Este problema se incrementa por la necesidad de gestionar el inventario al máximo nivel de detalle: tienda, producto y talla (SKU). No llega con predecir la demanda de una prenda en general, sino que la planificación debe ser más precisa. A este nivel de SKU, el problema estadístico se vuelve más complejo, pues la venta de una talla específica es un evento mucho más disperso que la venta de un producto en general. Como señalan Caro y Gallien (2010) [1], la falta de disponibilidad de una talla central no solo conduce a la pérdida de la venta, sino que puede tener un impacto negativo en la experiencia de compra del consumidor. Además, esto puede forzar incluso a la retirada del producto de la zona de exposición, provocando que se anule la demanda de todo el artículo de forma errónea.

A esto se le suma la constante llegada de nuevos artículos sin historial de ventas previo, lo que impide predecir su comportamiento basándose simplemente en datos pasados.

En este contexto, surge la necesidad de superar las limitaciones de los enfoques clásicos. El objetivo es desarrollar una herramienta de soporte que, apoyándose en técnicas de aprendizaje no supervisado y modelos de regresión, permita predecir la proporción de tallas incluso para productos sin historial de ventas. De este modo, se busca ajustar la forma del surtido de prendas para maximizar la probabilidad de venta independientemente del volumen total de compra, mejorando así la toma de decisiones.

1.2. Definición del problema y enfoque metodológico

Desde una perspectiva analítica, el problema descrito en la sección anterior se traduce en que existe una gran dispersión en los datos (*sparsity*), especialmente en las curvas de tallas debido a la existencia de tallas centrales y extremas. Esto provoca que, al intentar modelar la demanda de una talla aislada mediante métodos tradicionales, la forma de la tendencia real no se distinga debido al ruido aleatorio. Esto lleva a que los algoritmos de predicción tradicionales sean inestables, ya que intentan encontrar patrones temporales en distribuciones de probabilidad condicionadas por el contexto del cliente y la tienda.

Como señalan Fildes et al. (2022) [3], para reponer el stock en las tiendas es necesario bajar a ese nivel de detalle, pero al hacerlo nos enfrentamos a un problema de dimensionalidad: la demanda de una talla no depende solo de su historia pasada (que suele ser escasa o nula), sino de múltiples variables de contexto (precio, tipo de la prenda, categoría). A esto se suma, como ya adelantamos, la continua entrada de nuevas colecciones sin historial previo.

Para abordar esta dificultad, la estrategia tomada consistió en separar el problema en dos etapas independientes, combinando técnicas de estadística y aprendizaje automático. Distinguimos aquí dos aspectos claros: mientras que factores externos (como el precio o el clima) influyen en el volumen de ventas, las características propias del cliente y los atributos del producto son los que determinan la elección de la talla. Por eso, en lugar de intentar predecir todo a la vez, lo habitual es tratar cada componente por separado para ganar precisión.

1. Estimación de la distribución de tallas (Fase 1): Dada la aleatoriedad de la venta de una talla específica, se modeló la probabilidad de venta de cada talla independientemente del volumen total.

   Para ello, se adoptó una estrategia diferenciada según el sistema de tallaje y la disponibilidad de datos del producto. Por un lado, para el sistema numérico se validó el uso de una curva promedio dada su alta homogeneidad y escaso volumen. Por otro lado, para el sistema alfabético, se desarrolló un único modelo predictivo basado en Regresión Logística Multinomial, apoyado en técnicas de aprendizaje no supervisado (*Clustering*). Siguiendo la propuesta de Thomassey (2010) [4], se agruparon los productos y tiendas en *clusters* basados en atributos descriptivos y patrones de comportamiento similares. Esto permite obtener una estimación de la curva de tallas de un artículo nuevo utilizando el comportamiento de su *cluster*, superando así la falta de datos históricos individuales.

   Esta combinación resultó ser una solución unificada tanto para referencias con histórico de ventas (ajustando la predicción a su contexto exacto) como para productos nuevos, resolviendo la falta de información previa.

2. Predicción del volumen de ventas (Fase 2): Una vez aislada la variable `talla`, el problema pasa a ser estimar el volumen total de ventas a nivel tienda-producto.

Esta división en dos fases permite gestionar mejor la información disponible: agregando los datos para afrontar su escasez en la fase de predicción del volumen y utilizando la similitud y buscando patrones para resolver la ausencia de historial en la fase de tallas. Y también permite aplicar técnicas especializadas a cada problema.

En este trabajo, se ha decidido implementar exclusivamente la Fase 1 (Estimación de distribución de tallas) debido a la diferencia entre la información disponible para cada problema.

La predicción del volumen de ventas depende en gran medida de variables externas (clima, promociones, tendencias) que en ocasiones no están disponibles en los datos históricos o no se controlan con facilidad. Por el contrario, la elección de la talla es un fenómeno relacionado con información demográfica y con los atributos del producto, variables sobre las que sí que se disponen de datos fiables.

Al aislar el problema de la talla (asumiendo el volumen de ventas constante y conocido), se pueden aplicar técnicas específicas a este problema sin que interfieran errores causados por la resolución del problema del volumen de ventas.

Finalmente, es necesario comprobar si la complejidad introducida compensa frente a los métodos simples, por lo que se contrastaron los modelos frente a estrategias más básicas como los promedios. Para medir este impacto, no solo se analizaron medidas de error, sino que se calcularon métricas financieras para medir el valor real que aportaría la solución al negocio.

1.3. Objetivos

El objetivo principal de este proyecto es el desarrollo de un sistema de asignación de curvas de tallas óptimas en un entorno de tiendas de moda. No obstante, debido a la complejidad y las limitaciones asociadas al sector *fast fashion*, este trabajo no pretende únicamente predecir valores exactos, sino estudiar el comportamiento y la probabilidad de elección del consumidor y obtener conocimiento sobre la estructura y variabilidad de las curvas de tallas según el contexto de venta.

Los objetivos específicos son los siguientes:

- Desarrollar un modelo de estimación de curvas de tallas basado en el contexto del producto, permitiendo gestionar la incertidumbre de los nuevos lanzamientos.
- Detectar patrones de consumo presentes en los datos, como diferencias en la distribución de tallas según la geografía o el tipo de prenda, para así demostrar la ineficacia de los promedios globales y comprobar si las técnicas de agrupamiento aportan validez.
- Contrastar el rendimiento del modelo frente a métodos tradicionales, como curvas de tallas estáticas o promedios.
- Medir los beneficios de la solución propuesta mediante una simulación sobre los datos disponibles, comparando los resultados obtenidos frente a la estrategia clásica de reponer el surtido basándose en el promedio histórico.

Capítulo 2

Datos y herramientas

En este capítulo se detallan la procedencia, estructura y características del conjunto de datos utilizado en el trabajo, así como el entorno tecnológico empleado para desarrollar los modelos.

2.1. Descripción del conjunto de datos

Los datos utilizados en este estudio provienen de un caso real facilitado por la empresa tecnológica Plexus Tech. En concreto, se trabajó con el histórico de ventas de una cadena de moda global, centrándose exclusivamente en la familia de prendas "pantalones de señora" para construir y validar los modelos.

El conjunto de datos original se divide en tres ficheros CSV que funcionan de forma conjunta: un archivo central que recoge el historial de transacciones de ventas y dos maestros que añaden información complementaria sobre los productos y las tiendas. En relación a la cantidad de datos, se partió de un volumen inicial de aproximadamente 2.18 millones de transacciones, repartidas a lo largo de 9 meses (entre febrero y octubre de 2025).

Se procede a describir más en detalle cada una de las fuentes:

- **Archivo de ventas diarias:** Este fichero representa la fuente principal de información. Cada fila describe una transacción diaria de una referencia concreta (desglosada por producto, tienda, talla y color). En este archivo es donde se encuentran las variables principales que se busca modelar: la cantidad de unidades vendidas (`quantity`) y la talla de cada producto (`size`).

- **Archivos de productos y tiendas:** Para añadir contexto a los datos de transacciones y mejorar la capacidad predictiva de los modelos, se dispone de dos catálogos:

  - El maestro de productos, que contiene información adicional y detallada de los 162 modelos únicos de pantalones disponibles en el estudio. Se trata de una fuente fundamental para la posterior estrategia de *clustering*, ya que permite agrupar prendas según sus atributos o características físicas (formas, tejidos o patrones) más allá de sus identificadores numéricos.
  - El maestro de tiendas, que dispone de un total de 1.460 puntos de venta físicos, proporcionando información geográfica necesaria para el mercado. Este archivo incluye variables como el tipo de ubicación (si la tienda se encuentra en la calle o en un centro comercial, por ejemplo), que influyen directamente tanto en el tráfico de clientes como en la curva de tallas que se espera.

Descripción de las variables

A continuación, se detallan las variables disponibles en cada fuente tras el proceso de estandarización y limpieza inicial:

Variables transaccionales (ventas)

Variable	Tipo	Descripción	Ejemplo
`base_product_id`	Categórica	Identificador único del modelo de pantalón (sin talla/color).	2, 54
`store_id`	Categórica	Identificador único del punto de venta físico.	416, 827
`date`	Fecha	Fecha (día) de la transacción.	2025-07-17
`size`	Categórica	Talla de la prenda.	XS, 38, 40
`color_name`	Categórica	Color de la prenda.	CRUDO, NEGRO
`quantity`	Numérica	Cantidad de unidades vendidas (números enteros).	1, 3, -1
`returns`	Numérica	Cantidad de unidades devueltas (números enteros positivos).	0, 1
`price`	Numérica	Precio base unitario fijado para el artículo.	24.95, 59.95
`hierarchy_1`	Categórica	Sección comercial.	SEÑORA
`hierarchy_2`	Categórica	Subsección comercial.	PANTALONES
`description`	Texto	Descripción corta de la prenda.	PANTALON JOGGERS ESTAMPADO

Tabla 2.1: Variables presentes en el archivo de transacciones diarias.

Variables del maestro de productos

Variable	Tipo	Descripción	Ejemplo
`category`	Categórica	Familia o tipo de producto.	PINZAS, PANA, ANCHO
`range`	Categórica	Gama comercial del producto.	FASHION
`department`	Categórica	Departamento.	TRAJE CORTO, DISEÑO

Tabla 2.2: Variables descriptivas del catálogo de productos.

Variables del maestro de tiendas

Variable	Tipo	Descripción	Ejemplo
`country`	Categórica	País donde se localiza el punto de venta.	ESPAÑA, MEXICO, JAPON
`store_type`	Categórica	Ubicación de la tienda.	CALLE, CENTRO COMERCIAL

Tabla 2.3: Variables descriptivas del catálogo de tiendas.

2.1.1. Confidencialidad y Anonimización

Dado que los datos disponibles provenían de operaciones reales, se llevó a cabo un proceso de anonimización antes de proceder al análisis y estudio del trabajo, con el fin de proteger la información comercial de la empresa colaboradora.

El proceso que se realizó abarca los siguientes puntos:

1. Se sustituyeron los identificadores de las referencias reales (IDs de producto y tienda) por códigos generados aleatoriamente para el estudio.
2. Se modificaron los precios de los productos de forma escalonada y controlada para ocultar la facturación real de las tiendas pero manteniendo los patrones originales para no afectar a la estructura de los datos.
3. Se simplificaron las descripciones y atributos del producto para evitar la identificación del catálogo comercial original.

No obstante, no se alteró ni la estructura de la base de datos, ni los ciclos temporales de ventas, ni la distribución de la demanda por tallas asegurando que los modelos estadísticos desarrollados fuesen aplicables, sólidos y fieles a la realidad.

2.2. Lenguajes y entorno tecnológico

Para el desarrollo de este proyecto se combinaron dos enfoques de trabajo: el procesamiento de datos mediante código y las consultas directas sobre la base de datos para tomar decisiones en cada paso. Para ello, se utilizó:

- **Python**: Se utilizó como lenguaje principal para desarrollar todo el flujo de datos y el modelado predictivo. Esto se ejecutó mediante notebooks de *JupyterLab*, un entorno de trabajo que permitió escribir, ejecutar y documentar el código.
- **SQL**: Se utilizaron consultas SQL como método complementario para examinar y filtrar la información de forma precisa para poder analizarla posteriormente.

Librerías de Python

Las principales librerías utilizadas para la manipulación, gestión y visualización de datos fueron:

- Pandas y NumPy: Para la manipulación de estructuras de datos, operaciones de limpieza, uniones y agregaciones.

- Matplotlib y Seaborn: Para la generación de visualizaciones.
- PyArrow: Para gestionar de forma eficiente archivos en formato Parquet, empleados ya que optimizan los tiempos de lectura y escritura de grandes volúmenes de datos.

2.3. Preprocesamiento y limpieza

Una vez analizadas en detalle las fuentes de información y su estructura, el siguiente paso del proyecto consistió en implementar un flujo de limpieza y preprocesamiento de los datos. El objetivo principal de esta etapa fue transformar los datos crudos, tanto el archivo de ventas diarias como los maestros, en un dataset unificado y limpio, preparado para el entrenamiento de los modelos.

Durante este proceso, surgieron diversos inconvenientes habituales al trabajar con datos reales del sector retail, como inconsistencias en las devoluciones o la necesidad de gestionar ventas negativas.

Para llevar a cabo este flujo de trabajo, se construyó el primer *notebook* del proyecto. A continuación, se detallan las distintas etapas:

2.3.1. Unificación de fuentes de datos

El proceso se inició con la carga en memoria de los tres archivos de datos originales (tanto las transacciones como los catálogos) en el entorno de trabajo. A continuación, se procedió a la normalización de los nombres de las variables para seguir una nomenclatura estándar interna (por ejemplo, se transformó `id_articulo` a `base_product_id`, `ventas` a `quantity`, etc). Esto se hizo para mejorar la claridad y legibilidad del código y también para que fuese posible volver a utilizarlo con otros conjuntos de datos sin más que actualizar el mapeo de entrada.

Con las variables normalizadas, se procedió a enriquecer la fuente de ventas, la cual se considera como núcleo del conjunto de datos. Esto se llevó a cabo mediante operaciones de unión de tipo `Left Join` con los maestros de productos y tiendas, ya que se eligió conservar todos los registros de venta originales y simplemente añadirles información de contexto más específica. Es decir, se intentó que no se perdiesen datos en el caso de que alguna referencia no apareciese en los maestros.

Posteriormente, se validó el resultado de esta operación, obteniendo que el total de los 2.184.140 registros se cruzaron correctamente durante el enriquecimiento, lo que confirmó la calidad de los datos. De esta forma, el dataset resultante integraba la venta pura junto con sus atributos explicativos, tanto del producto como de la tienda correspondientes.

Una vez completo el conjunto de datos y antes de proceder a crear los modelos predictivos, se realizó un análisis exploratorio de algunas de las variables más relevantes con el fin de garantizar que la información fuese coherente y que no hubiese inconsistencias en los datos. De este modo, se aseguró que los datos llegasen limpios a la fase de predicción y sin ruido que pudiese confundir en etapas posteriores. A continuación, se examinó el comportamiento y la estructura de las variables clave.

2.3.2. Tratamiento de la variable temporal

Lo primero que se abordó fue la variable temporal `date`. Al cargar el archivo CSV con Pandas, esta columna se lee por defecto como texto simple (*object*), así que se convirtió a formato `datetime`. Gracias a esto, fue posible calcular diferencias de tiempo, ordenar la serie temporal y extraer información más específica (como días o meses).

Por otro lado, se comprobó que no había fechas nulas ni errores de formato y se confirmó que el marco temporal del dataset completo comprendía desde el 6 de febrero hasta el 9 de octubre de 2025.

2.3.3. Gestión de ventas negativas y devoluciones

Tras validar la variable temporal, se procedió a examinar las variables de ventas (`quantity`) y devoluciones (`returns`). El objetivo fue entender su estructura y, sobre todo, identificar cualquier irregularidad antes de proceder a la limpieza.

Durante la exploración inicial se detectó que el 13.55 % de las transacciones diarias (casi 300.000 filas) tenían ventas negativas. A primera vista, un volumen tan elevado de valores negativos apuntaba a posibles inconsistencias en la forma en la que el sistema de origen registró estas operaciones.

Para averiguar la lógica de estos registros antes de manipularlos en Python, se revisó con detalle la estructura de datos aplicando Ingeniería Inversa mediante la herramienta SQL. Con este este análisis se dedujo que el sistema no seguía un criterio único, sino que combinaba diferentes maneras de registrar las operaciones:

- **Comportamiento general:** En la gran mayoría de los casos se observó una estructura definida, siendo las devoluciones siempre valores positivos (> 0) e incluso nulos en un elevado porcentaje (98.6 %). Por su parte, las ventas (`quantity`) tendían a ser unitarias (el 72.4 % eran iguales a 1) o cero (1.4 %), lo que sugirió un modelo de negocio mayoritariamente basado en la venta de artículos de forma individual.
- **Inconsistencia:** A pesar de que existe una columna dedicada a las devoluciones, se detectó una cantidad significativa (un 13.55 %) de operaciones registradas como "ventas negativas" en `quantity`. Lo interesante es que, de este grupo, prácticamente el total (un 13.54 % del global de filas) mantenía la columna `returns` a cero.

  Esto indicó que, en estos casos, las devoluciones se gestionaron de otra manera (quizás a través del canal online), lo que provocó que la información se guardase con una estructura diferente al resto de transacciones.
- **Excepciones:** Por último, se identificó un comportamiento residual (0.011 % del total) donde se registraron en la misma transacción diaria valores no nulos tanto en ventas como en devoluciones. Lo más llamativo dentro de este grupo fue un subconjunto (0.005 % del total) que presentó ventas negativas y devoluciones positivas al mismo tiempo, lo que sugirió que se realizaron correcciones manuales o ajustes para cuadrar el stock o la caja.

Tras el diagnóstico, se determinó que modelar la demanda basándose en las ventas netas (restando devoluciones) sería erróneo para gestionar el inventario ya que, para que una venta se produzca (independientemente de si el artículo se devolverá posteriormente) el producto debe estar disponible en la tienda. Por tanto, incluir valores negativos en el estudio subestimaría la necesidad real de stock, aumentando el riesgo de rotura.

En consecuencia, se decidió separar ambas cantidades (ventas y devoluciones) para tratar la variable `quantity` como demanda bruta positiva. Adicionalmente, por si fuese necesario validaciones posteriores, se creó la variable `net_sales` (ventas - devoluciones).

De acuerdo con este razonamiento, se implementó la siguiente estrategia de corrección en el *notebook*:

1. Primero, se identificaron los registros con `quantity` < 0.
2. A continuación, el valor absoluto de estas ventas negativas se sumó a la columna de devoluciones (`returns`), asignando un valor de cero a la venta bruta en esos registros para mantener la coherencia.
3. Por último, se generó una nueva columna, `net_sales`, resultante de la resta entre la venta bruta corregida y las devoluciones totales.

2.3.4. Clasificación y filtrado de tallas

Como se disponía de una gran variedad de tallas distintas en las prendas disponibles, se procedió a clasificar cada producto del catálogo en tres categorías según su sistema de tallaje para analizarlas por separado: *Letter* (alfabético), *Numeric* (numérico) e *Intermediate* (intermedio). Esta división mostró una estructura desbalanceada de las prendas, dominada por el sistema alfabético con 151 productos.

Para visualizar cómo estaba representado cada sistema, se generó un gráfico de barras de la distribución de ventas brutas (`quantity`) por talla empleando la librería Matplotlib de Python. Este análisis, representado en la Figura 2.1, permitió extraer dos conclusiones: que la demanda se concentró en las prendas con sistema de tallaje *Letter* (XS, S, M, L, XL, XXL), con la talla M superando las 600.000 unidades vendidas; y que el sistema *Numeric* (tallas 32 - 46) presentaba un volumen inferior, aunque manteniendo una actividad de ventas suficiente para que fuese posible modelar su comportamiento de forma independiente.

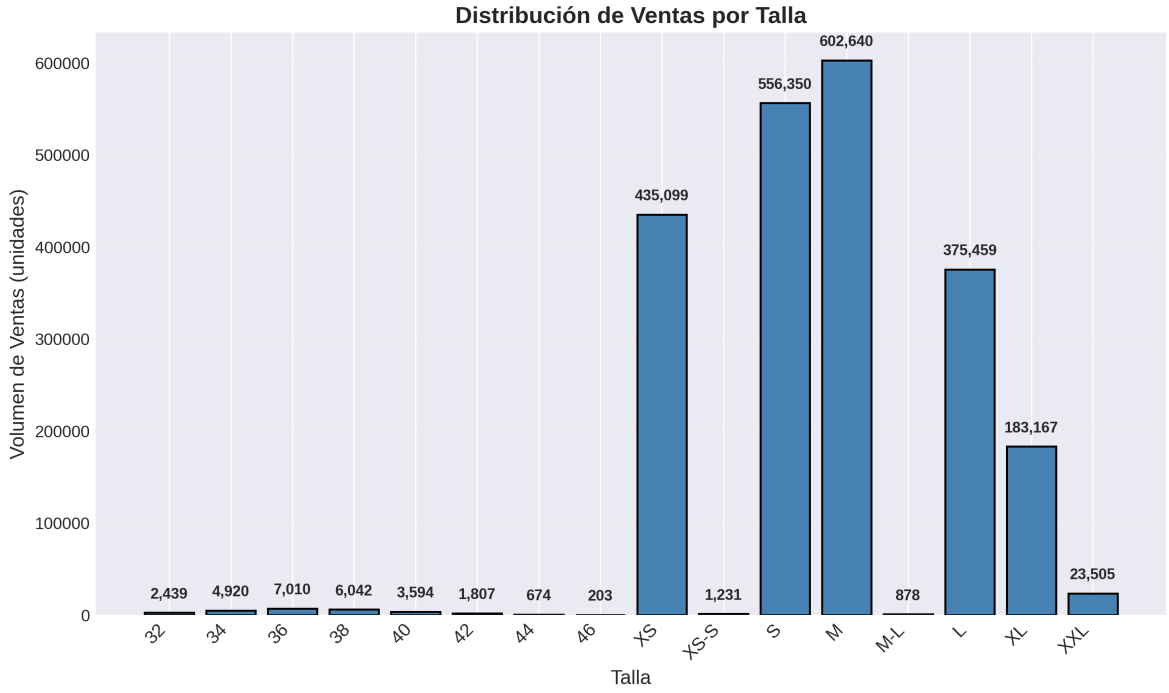

Figura 2.1: Distribución del volumen de ventas brutas por talla.

Debido a la escasa cantidad de ventas relativas al sistema de tallaje *Intermediate*, se decidió analizar las tasas de devolución por cada sistema, calculadas como el cociente entre devoluciones y ventas brutas:

$$\text{Rate} = \frac{\texttt{returns}}{\texttt{quantity}}.$$

Los cálculos revelaron que el sistema intermedio presentó tasas de devolución extredemadamente altas: un 59.5 % para la talla M-L y un 54.9 % para la XS-S. Esto significa que, para estas tallas, más de la mitad de las unidades vendidas fueron devueltas.

Como se detalla en la Tabla 2.4, este sistema estaba compuesto por un único producto y aportaba solo el 0.1 % de las transacciones (2.477 registros totales) pero introducía ruido en los datos debido a su alta tasa de devolución. Al no ser representativo del comportamiento general de la demanda, se decidió eliminarlo del estudio para evitar sesgos en el modelo.

Sistema de tallaje	Productos	Transacciones	% de transacciones
Letter	151	2.152.841	98.6 %
Numeric	10	28.822	1.3 %
Intermediate	1	2.477	0.1 %

Tabla 2.4: Análisis de rendimiento por sistema de tallaje.

Tras el filtrado, el resultado fue un conjunto de datos final compuesto por de 2.181.663 transacciones correspondientes a 161 productos únicos, simplificando el catálogo de pantalones a los dos sistemas de tallaje predominantes: *Letter* y Numeric.

2.3.5. Almacenamiento del dataset final

Finalmente, para cerrar esta etapa de preprocesamiento, el conjunto de datos final se almacenó utilizando el formato Parquet. A diferencia de los CSV tradicionales, este formato permite mantener los tipos de datos y presenta una alta tasa de compresión, reduciendo el tamaño de los archivos para facilitar su carga durante el entrenamiento de los modelos.

Capítulo 3

Análisis del comportamiento de la demanda por tallas

Una vez consolidado el conjunto de datos en el capítulo anterior, en este apartado y en el siguiente se llevó a cabo la construcción de los modelos. El proceso se estructuró siguiendo la descomposición del problema planteada en la introducción: primero se analizó el comportamiento de los datos y de la demanda, y, posteriormente se modeló la distribución de tallas mediante técnicas de regresión (resolviendo la heterogeneidad mediante *clustering*).

3.1. Estrategia de partición temporal

En primer lugar, se definió una estrategia de validación para simular un escenario de predicción realista. Como el objetivo del proyecto es estimar la proporción de tallas de la demanda futura, se descartaron los métodos de validación cruzada tradicionales para particionar el conjunto de datos ya que su aplicación en series temporales implicaría entrenar el modelo con información futura para predecir el pasado.

Por tanto, se decidió dividir el conjunto de datos siguiendo el orden cronológico y utilizando los datos más antiguos para entrenar, como sucedería en la realidad al hacer una predicción nueva. A continuación, se describen en profundidad ambos conjuntos de datos y en la Tabla 3.1 se presenta un resumen de los resultados obtenidos al ejecutar el proceso de partición del dataset original.

Conjunto de datos	**Rango de fechas**	**Nº de registros**	**% del total**
Entrenamiento (*Train*)	2025-02-06 – 2025-09-18	1.894.658	86.8 %
Validación (*Holdout*)	2025-09-19 – 2025-10-09	287.005	13.2 %

Tabla 3.1: División temporal del conjunto de datos preprocesado.

En definitiva, para simular un escenario realista de predicción, se asumió que el momento actual era sobre mediados de septiembre de 2025 (contando con un histórico de datos desde febrero) y el objetivo era predecir el comportamiento futuro de las siguientes tres semanas de datos (reservadas como conjunto de validación o *holdout*).

- Conjunto de Entrenamiento (*Train*): Datos históricos desde el inicio hasta el 18 de septiembre de 2025. Estos registros se utilizaron para todo el análisis exploratorio, cálculo de distribuciones y entrenamiento de algoritmos.

- Conjunto de Validación (*Holdout*): Datos correspondientes a las últimas 3 semanas del periodo. Este subconjunto permaneció reservado hasta la parte final del proyecto para evaluar el rendimiento real de los modelos sobre datos nuevos.

3.2. Análisis exploratorio de datos

A lo largo de esta sección se llevó a cabo un análisis exploratorio para comprender la naturaleza de la demanda y, específicamente, el comportamiento de las tallas. Esto se hizo utilizando el subconjunto de datos de entrenamiento definido en el apartado anterior. El flujo de trabajo se estructuró en cuatro etapas.

3.2.1. Agregación de datos transaccionales

Dado que el conjunto de datos original estaba agrupado por transacciones diarias por prenda, fue necesario agregar el volumen de ventas por producto y talla para poder analizar la distribución de las tallas. Se generó un nuevo conjunto de datos donde cada fila representaba una combinación única de las variables `base_product_id` y `size`, sumando el total de unidades vendidas.

El dataset resultante de la agrupación contenía 605 combinaciones únicas de producto-talla, reduciendo las dimensiones del problema y aumentando la precisión en el análisis de la demanda.

3.2.2. Análisis de volumen y filtrado

Antes de calcular las distribuciones de probabilidad, se analizó el volumen de ventas para cada artículo. Tanto el histograma como el diagrama de cajas representados en la Figura 3.1 muestran una fuerte asimetría, donde se observa que unos pocos productos concentraron volúmenes enormes de ventas (algunos más de 100.000 unidades vendidas en total), mientras que la gran mayoría presentaron muy poca demanda (entre 0 y 20.000 ventas aproximadamente).

Figura 3.1: Distribución de ventas acumuladas por producto. A la izquierda, el histograma que relaciona las ventas totales con el número de productos y, a la derecha, el diagrama de cajas (*boxplot*) representando las ventas totales en escala logarítmica.

Debido a este desbalance, se filtraron todos los productos que no superaran un volumen mínimo de ventas, ya que calcular distribuciones de probabilidad sobre una muestra muy pequeña podría generar resultados inestables y que no reflejan la realidad.

Por ejemplo, si un producto recién lanzado registra únicamente 2 ventas y ambas coinciden en ser de la talla M, el cálculo asignaría un 100 % de probabilidad a dicha talla y un 0 % al resto. Esto no implica que no exista demanda para las tallas S o L, sino que el artículo no ha tenido suficiente tiempo de exposición en tienda para mostrar su curva de demanda real.

Para evitar que este ruido afecte al modelo general, se excluyeron del cálculo únicamente aquellos productos con un volumen histórico inferior a 100 unidades. Este filtro garantizó que las curvas de tallas se basaran en datos robustos, descartando sólo el 0.01 % del volumen total de ventas y reduciendo la muestra de 123 a 111 productos (107 para el sistema de tallas *Letter* y 4 para el *Numeric*).

3.2.3. Cálculo de distribuciones de tallas

Una vez agregado y filtrado el conjunto de datos, el siguiente paso fue estandarizar las ventas. El objetivo fue determinar la proporción de ventas de cada talla dentro de un producto. Esto permitió analizar la estructura de la demanda y comparar el comportamiento de las tallas independientemente de si se trataba de productos muy populares o poco vendidos.

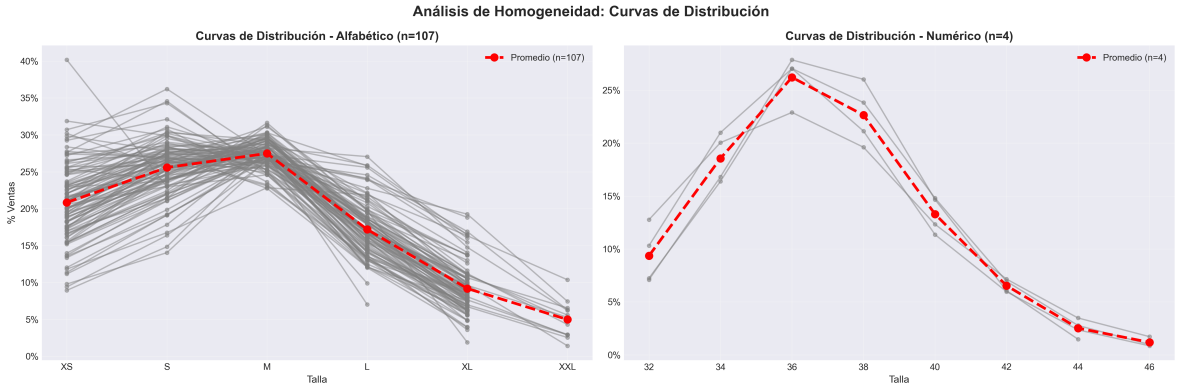

Figura 3.2: Análisis de homogeneidad de las curvas de tallas. Las líneas grises representan la distribución de cada producto individual y la línea roja discontinua indica el promedio del grupo.

Para ello, se calcularon las frecuencias relativas de cada producto, dividiendo las ventas de cada talla específica entre el total de ventas de ese producto:

$$f_{i,j} = \hat{\mathbb{P}}(Y = j \mid Z = i) = \frac{q_{i,j}}{\sum_{k \in \mathcal{A} \cup \mathcal{N}} q_{i,k}}, \quad \text{para todo } j \in \mathcal{A} \cup \mathcal{N},\ i \in \mathcal{I}, \tag{3.1}$$

donde las variables y conjuntos anteriores se definen como:

- $Y =$ `size` (talla)
- $Z =$ `base_product_id` (producto)
- $q =$ `quantity` (cantidad/unidades vendidas)
- $\mathcal{A} = \{XS, S, M, L, XL, XXL\}$ el conjunto de tallas alfabéticas disponibles
- $\mathcal{N} = \{32, 34, 36, 38, 40, 42, 44, 46\}$ el conjunto de tallas numéricas disponibles

- $\mathcal{I}$ el conjunto de todos los identificadores únicos de pantalones

Al realizar esta operación, se generaron los vectores con la proporción de tallas donde la suma de todos los pesos es siempre igual a 1. Para comprobar si existían diferencias a simple vista en la distribución de tallas de los productos, se representaron las curvas de todos los artículos filtrados, separándolos según su sistema de tallaje (alfabético o numérico).

En la Figura 3.2 se observa un patrón en ambos sistemas de tallaje:

- En el sistema alfabético (izquierda), la demanda se concentra en las tallas centrales (S y M), con una caída suave hacia los extremos.
- En el sistema numérico (derecha), que cuenta con menos referencias, la forma es todavía más marcada con un pico en las tallas 36 y 38.

Estas similitudes sugieren que los pantalones comparten una forma de demanda por tallas bastante genérica. Esto tiene sentido, pues existe una estructura equivalente entre cada rango de tallas de ambos sistemas (por ejemplo, una 36 equivaldría a una S, etc) lo que permitiría fusionar ambas realidades. Sin embargo, en este trabajo no se dispuso del mapeo específico entre los sistemas de tallaje y se prefirió llevar a cabo un estudio personalizado para cada uno.

Por otra parte, se calculó la curva promedio por cada grupo y representó en la Figura 3.2 en color rojo y con un estilo de línea discontinua. A simple vista, se puede apreciar una similitud aparente tanto entre pares de curvas de distribución como entre cada curva y el promedio del grupo. Es decir, a nivel global, la variación dentro de cada grupo no parece lo suficientemente evidente como para justificar el desarrollo de un modelo específico para cada referencia.

Desde el punto de vista del negocio y la planificación comercial, este hallazgo es clave. Cuando un nuevo pantalón se introduce en el catálogo y no tiene histórico de ventas, la estrategia más segura consiste en asignarle directamente la curva promedio de su sistema de tallaje (por ejemplo, para el sistema alfabético: 24.2 % para la talla S, 26.3 % para la M, etc.). Esto permite realizar un reparto inicial del stock que minimice el riesgo de rotura en las tallas centrales antes de disponer de datos reales del producto.

Limitaciones

Sin embargo, es necesario tener en cuenta algunos aspectos importantes acerca de la interpretación de la homogeneidad anterior. El análisis visual realizado evalúa el comportamiento de los productos agregando todas sus ventas a nivel mundial, lo que hace que se pierda información sobre el contexto y oculta los patrones de cada grupo.

Al agregar todas las transacciones sin distinguir por países, las diferencias entre mercados opuestos podrían compensarse entre sí (por ejemplo, la demanda de tallas pequeñas en mercados asiáticos frente a tallas grandes en el norte de Europa). Esto da lugar a una distribución genérica de cada producto que no se ajusta al comportamiento de ninguna región en concreto, aumentando el riesgo de rotura de stock en las tallas extremas.

Esto provoca que, aunque se haya deducido que el promedio global es una muy buena referencia general (especialmente ante la falta de histórico de ventas), no tiene la precisión necesaria para asignar de forma exacta el stock en cada tienda en específico. El efecto del país y la categoría de la prenda, que son variables que ahora no se están teniendo en cuenta por la agregación, son muy relevantes para la venta final. De hecho, en la Figura 3.2 se pueden detectar pequeñas variaciones en los picos de tallas de algunas prendas (sobre todo en el sistema alfabético), que actualmente están difuminadas por el efecto del promedio.

En consecuencia, en la siguiente fase del modelado se cambiará este enfoque global para aplicar técnicas de aprendizaje no supervisado (*Clustering*), con el objetivo de descubrir estos patrones y predecir la demanda de forma más específica.

Capítulo 4

Diseño e implementación de los modelos

Durante la fase exploratoria previa, se evaluó la viabilidad del uso de curvas de tallas promedio calculadas a nivel de producto (agregando las ventas de todos los países y tiendas). Aunque el análisis visual confirmó una gran homogeneidad a nivel global, también permitió identificar la presencia de un sesgo debido a esta agregación: al calcular promedios las diferencias locales se compensan entre sí. De esta forma, se obtienen curvas suavizadas que ocultan las características de cada mercado.

Como el objetivo de este estudio es la asignación precisa de stock por talla a nivel de tienda, un modelo basado en promedios globales podría servir como un primer modelo base, pero resultaría ineficiente en muchas ocasiones, provocando roturas de stock en las tallas más inusuales según la zona geográfica.

En esta fase se buscó estimar la probabilidad de venta de cada talla específica, pero teniendo en cuenta el contexto real en el que ocurre la venta. Por tanto, se buscó construir un modelo dinámico que consiguiese superar las limitaciones del enfoque estático.

Matemáticamente, la idea anterior se expresa mediante una probabilidad condicional:

$$P(Y = j \mid \mathbf{X}), \text{ donde } j \in \mathcal{A} \cup \mathcal{N}.$$

Es decir, denotamos por Y a la variable `size` y llamamos $\mathbf{X}$ al vector compuesto por el conjunto de características presentes en el estudio, en el que se agrupan las variables de contexto (`country`, `store_type`) y las características propias del producto (`category`, `prize` o `color_name`).

Antes de comenzar a detallar la estrategia empleada, es necesario justificar el alcance de su aplicación. Aunque el estudio se realizó sobre todos los artículos del catálogo de pantalones, se decidió implementar esta metodología más compleja solo a los productos del sistema de tallas alfabético. Esta decisión fue consecuencia de dos factores que surgieron previamente:

- El conjunto de datos está realmente desbalanceado, donde más del 96 % de las referencias (107 productos) corresponden al sistema alfabético, mientras que la presencia del sistema numérico es casi inexistente (4 productos). Aplicar un modelo complejo al grupo minoritario sería ineficiente, pues podría aumentar el riesgo de sobreajuste sin aportar una mejora real y no compensaría la escasa ganancia que obtendríamos frente a usar un simple promedio histórico.
- Como se detectó en el capítulo anterior, el sistema numérico resultó ser un grupo extremadamente estable y con una variabilidad muy escasa. Esto se evidenció al observar una distancia entre sus

curvas de tallas muy pequeña, siendo casi idénticas al promedio del grupo.

En consecuencia, las siguientes secciones se centrarán únicamente en la optimización de la predicción para el sistema alfabético, manteniendo el modelo base validado anteriormente como la solución final para el sistema numérico. Dicha curva promedio asigna el siguiente reparto de probabilidades por talla:

Talla	**32**	**34**	**36**	**38**	**40**	**42**	**44**	**46**
Probabilidad	9.0 %	17.7 %	25.8 %	23.0 %	13.9 %	6.7 %	2.6 %	1.3 %

Tabla 4.1: Distribución de probabilidad por talla numérica en el conjunto de datos.

Para llevar a cabo este nuevo enfoque para el sistema de tallas alfabético, se diseñó un flujo de trabajo en dos etapas implementadas en un tercer *notebook*, diseñadas tanto para cubrir las necesidades de los productos duraderos como aquellos sin histórico de ventas:

1. *Clustering (K-Means)*: Se utilizó como técnica de reducción de dimensionalidad para emplear la información disponible sobre el contexto geográfico y del tipo de prenda. Esto resulta crucial para tratar nuevos lanzamientos, ya que permite caracterizarlos en función de su similitud con grupos existentes.
2. Regresión Logística Multinomial: Se trata del modelo predictivo final. Integra las variables generadas por el *clustering* con el resto de atributos (precio, tipo de tienda) para calcular la probabilidad exacta de cada talla.

A continuación se detalla la formulación de estos modelos.

4.1. Caracterización de productos mediante Clustering

El desarrollo del modelo se inició con la carga en memoria del conjunto de entrenamiento dividido en el capítulo anterior. Es necesario recalcar que el conjunto formado por las últimas tres semanas de datos se mantuvo separado del resto, para garantizar que los algoritmos de esta sección aprendiesen con datos únicamente del pasado. A continuación, se filtró y estructuró la información para poder emplear el algoritmo de Regresión Logística Multinomial. Como ya se mencionó, en este punto se procedió a aislar las transacciones correspondientes al sistema de tallaje alfabético.

Como resultado del filtrado, el volumen de datos disponible para el modelado fue el siguiente:

- Registros Totales (Train): 1.894.658 transacciones.
- Registros Seleccionados (*Letter*): 1.875.582 transacciones.

Estos resultados confirman que la división anterior abarca el 99.0 % del volumen de datos de entrenamiento, validando la decisión de basar el modelo en este sistema.

Durante el análisis de las variables presentes en el dataset se identificó un gran obstáculo: la alta cardinalidad de las variables categóricas. Se contó con información de ventas relativas a 88 países distintos, 15 categorías de pantalones, 7 departamentos y 2 tipos de tienda. Al introducir esta diversidad directamente en el modelo de regresión logística (mediante *One-Hot Encoding*) se trabajaría con demasiadas variables binarias. Esto aumentaría la complejidad computacional del modelo y provocaría dificultades en su capacidad de predicción, provocando un riesgo de sobreajuste en aquellos mercados

o categorías con pocas transacciones.

Para tratar de solucionar estas dificultades, se utilizó el algoritmo *K-Means* como técnica de clasificación no supervisada, con el objetivo de comprimir esta información disponible agrupando las 88 localizaciones y las 15 familias. De este modo, se condensó la información sin perder capacidad explicativa.

Base teórica del algoritmo K-Means

El algoritmo *K-Means* es un algoritmo de aprendizaje no supervisado (en concreto, de *Clustering*) cuyo objetivo es particionar un conjunto de N observaciones en k grupos (*clusters*) disjuntos, buscando minimizar la varianza dentro de cada grupo.

En el contexto de este estudio, cada observación (país o categoría) se representa como un vector p-dimensional $\mathbf{x}_i \in \mathbb{R}^p$, donde p es el número de tallas disponibles en el sistema alfabético ($p = 6$ correspondientes a $\{XS, \ldots, XXL\}$).

El algoritmo empleado es la versión estándar (que es la empleada por defecto en la librería `Scikit-Learn` de Python) y busca encontrar un conjunto de k centroides, $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$, que representen el comportamiento promedio de cada grupo. Matemáticamente, el problema global del algoritmo consiste en minimizar la Suma de Errores al Cuadrado (RSE), definida como la suma de todas las distancias al cuadrado entre cada punto y el centro de su cluster asignado [5]:

$$J = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_j} ||\mathbf{x}_i - \boldsymbol{\mu}_j||^2. \tag{4.1}$$

Donde:

- $\mathcal{C}_j$ es el conjunto de observaciones asignadas al cluster j.
- $||\cdot||$ denota la norma Euclidiana (L_2). Esta métrica compara la forma de las curvas de demanda: dos países quedarán en el mismo clúster si, al superponer sus gráficos, sus picos de demanda coinciden en las mismas tallas y con intensidades similares.

En este estudio, el procedimiento empleado se basó en una regla de arranque y una de parada, y se estructuró en tres fases:

1. Inicialización (K-Means++): La elección de los centroides iniciales $\{\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_k^{(0)}\}$ es crítica para evitar óptimos locales deficientes. En lugar de una selección completamente aleatoria, se empleó el método de inicialización *K-Means++* [5]. Este algoritmo selecciona el primer centroide al azar y los siguientes con una probabilidad proporcional a su distancia al centroide más cercano ya elegido, garantizando que haya una mayor dispersión inicial.

2. Asignación: En la iteración t, cada observación $\mathbf{x}_i$ (curva de probabilidades de cada país o categoría) se asigna al cluster cuyo centroide $\boldsymbol{\mu}_j^{(t)}$ es el más cercano según la distancia euclidiana:

$$C_j^{(t)} = \{\mathbf{x}_i : ||\mathbf{x}_i - \boldsymbol{\mu}_j^{(t)}|| \leq ||\mathbf{x}_i - \boldsymbol{\mu}_l^{(t)}||, \quad \forall\, l \neq j\} \tag{4.2}$$

3. Actualización: Se recalculan los nuevos centroides $\boldsymbol{\mu}_j^{(t+1)}$ como el vector promedio de todas las observaciones asignadas a dicho cluster en el paso anterior.

El proceso iterativo (pasos 2 y 3) se repite hasta alcanzar el criterio de convergencia. Esto ocurre cuando al reasignar observaciones ya no se modifica la posición de los centroides ($\boldsymbol{\mu}_j^{(t+1)} \approx \boldsymbol{\mu}_j^{(t)}$) o

cuando la reducción en la función de coste J es inferior a un umbral de tolerancia predefinido.

Este enfoque permite simplificar la estructura del problema original, pasando de tener 88 países y 15 categorías a un número manejable de grupos.

Implementación en la práctica y resultados

Para implementar el algoritmo se utilizó la función `KMeans` de la librería `Scikit-Learn` de Python. La configuración del modelo se estableció con los siguientes hiperparámetros para garantizar la reproducibilidad de los resultados: se estableció el número de clústers como $k = 4$, se realizaron `n_init = 10` ejecuciones independientes con distintos centroides iniciales para mejorar la búsqueda y se fijó una semilla `random_state = 42` para asegurar que los resultados fuesen reproducibles e iguales en cada ejecución del código.

4.1.1. Agrupamiento geográfico (cluster de países)

La primera aplicación de la estrategia de reducción de la dimensión se centró en la variable `country`, que contaba con 88 mercados distintos. Para abordar esto, no se agruparon los países por cercanía o tamaño, sino por similitud en la curva de distribución de ventas por talla. La metodología para la construcción de las curvas de frecuencias fue la que se expone a continuación.

Para que el algoritmo de *Clustering* pudiera entender y comparar los países, fue necesaria una primera fase de preparación de los datos crudos de transacciones en la que se transformaron a una estructura vectorial estandarizada. Este proceso de transformación se llevó a cabo mediante tres pasos:

- Primero, se transformaron los registros de ventas en una matriz de frecuencias donde cada fila representaba un país y cada columna una talla (de la XS a la XXL).
- Como el volumen de ventas era muy distinta entre mercados (un país grande vendería mucho más que uno pequeño), fue necesario normalizar los datos para eliminar el efecto de la escala. Para ello, se transformó el vector de ventas brutas ($\mathbf{v}_i$) en un vector de proporciones ($\mathbf{p}_i$) dividiéndolo por su volumen total:

$$\mathbf{p}_i = \frac{\mathbf{v}_i}{\sum_j v_{i,j}}$$

  Donde el denominador representa la suma total de prendas vendidas en ese país.
- Finalmente, se aplicó el algoritmo *K-Means* configurado para detectar $k = 4$ grupos de comportamiento.

El algoritmo segmentó el territorio global en cuatro perfiles de comportamiento claros y definidos:

- **Cluster 0 (Mercados de tallas estándar con dominancia de talla M):** Formado por países como Alemania, Bélgica y Croacia. Se trata del grupo mayoritario con 39 países y funciona como el grupo estándar, mostrando una distribución equilibrada con un pico en la talla M.
- **Cluster 1 (Mercados de tallas grandes):** Este conjunto agrupa mercados como Arabia Saudita, India, Emiratos Árabes y Egipto, resumiendo mayoritariamente a los países árabes. Aunque comparten la M como talla central con otros clusters, se diferencian por presentar una mayor demanda de tallas más grandes, como la L y la XL.
- **Cluster 2 (Mercados de tallas pequeñas):** Es el grupo más diferenciado y crítico. Se corresponde a países asiáticos como Japón, China y Corea y se caracteriza por una alta concentración en la demanda de las tallas más pequeñas, dominando la XS y S (a la izquierda en el gráfico).

Este grupo evidencia la necesidad de segmentar en *clusters* en lugar de usar el promedio global, pues en este caso generaría un exceso de stock sobrante en las tallas más grandes. Además, se trata del grupo más reducido formado por solo 10 países.

- **Cluster 3 (Mercados de tallas estándar con dominancia de S):** Contiene a países como Estados Unidos, Italia, Austria y Suiza. Similar al cluster 0 pero con mayor volumen de ventas en la talla S, reflejando una morfología de la población ligeramente más pequeña.

Esta nueva variable adicional (`cluster_pais`) permite que, si la empresa abre mercado en un nuevo país, baste con asignarlo a uno de estos cuatro perfiles para tener una estimación inicial robusta.

Cluster	Talla dominante	Países
0	M	Albania, Alemania, Andorra, Armenia, Aruba, Belgica, Canada, Croacia, Dinamarca, El Salvador, Emiratos Arabes Unidos, España, Filipinas, Finlandia, Francia, Georgia, Grecia, Guatemala, Honduras, Indonesia, Irlanda, Jordania, Kosovo, Letonia, Lituania, Luxemburgo, Mexico, Montenegro, Nicaragua, Noruega, Panama, Países Bajos, Polonia, Portugal, Reino Unido, Republica Dominicana, República De Macedonia Del Norte, Suecia, Uzbekistan
1	M	Arabia Saudita, Argelia, Bahrein, Bosnia-Herzegovina, Egipto, India, Islandia, Kuwait, Marruecos, Oman, Qatar, Tunez
2	XS	Bulgaria, Camboya, Corea Del Sur, Hong Kong Sar, Japon, Macao Sar, Mainland China, Rumania, Taiwan, China, Vietnam
3	M	Austria, Azerbaiyan, Bielorrusia, Chipre, Colombia, Costa Rica, Ecuador, Eslovaquia, Eslovenia, Estados Unidos, Estonia, Hungria, Israel, Italia, Kazajstan, Libano, Malasia, Malta, Monaco, Republica Checa, Serbia, Singapur, Suiza, Tailandia, Turquia, Ucrania, Venezuela

Tabla 4.2: Asignación de países a los cuatro grupos identificados.

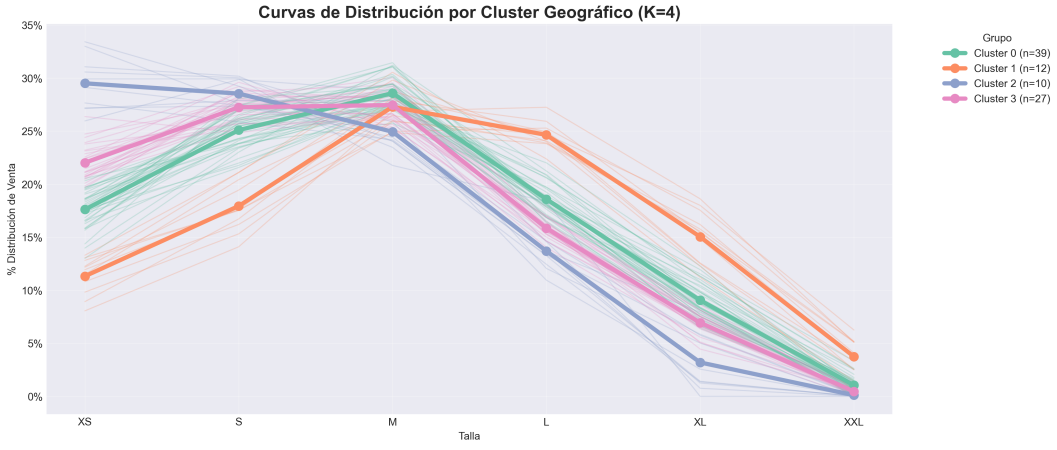

Figura 4.1: Perfiles de distribución de tallas obtenidos mediante K-Means ($k = 4$) sobre los países. Se observan diferencias estructurales en la demanda según la morfología del mercado.

El análisis de los resultados y la visualización de los gráficos generados tras ejecutar el algoritmo

(en la Figura 4.1 y la Tabla 4.2) confirman que la demanda de tallas se agrupa siguiendo patrones estructurales según la región.

4.1.2. Agrupamiento de productos (cluster de categorías)

Siguiendo la misma lógica, se agruparon las familias de productos. Esto es crítico ya que las diferencias físicas en la prenda (como el material que la compone o la holgura del patrón) generan diferencias en las tallas elegidas por los compradores. Con el objetivo de identificar estas familias de comportamiento, se replicó la estrategia del apartado anterior, esta vez aplicada sobre la variable `category`.

Al aplicar el algoritmo *K-Means* (con $k = 4$) sobre las 15 categorías originales de pantalones, se logró capturar cuatro patrones de comportamiento claramente diferenciados, confirmando la elección del número de grupos. La estructura resultante, observada en la Figura 4.2 relativa a las curvas de densidad, revela estos patrones:

- **Cluster 0 (Pantalones confort/elásticos):** Agrupa categorías como Joggers y Pijamero. Estos pantalones se caracterizan por tener patrones holgados y tejidos cómodos. Su curva de tallas presenta un pico en la talla M, aunque tiene un patrón bastante plano en general, ya que una misma talla puede utilizarse en varios tipos de cuerpo distintos.
- **Cluster 1 (Pantalones tendencia):** Son prendas con patrones más especiales que presentan ligeras variaciones en la distribución respecto a los básicos, como mayor presencia de tallas grandes como la L o la XL.
- **Cluster 2: (Pantalones piel/tejidos rígidos):** Se trata de un grupo formado únicamente por la categoría Piel. Cabe destacar que presenta el comportamiento más diferente en cuanto a distribución de tallas, con máximos alrededor de las tallas más pequeñas. Esto tiene sentido pues se trata de un tejido poco elástico, lo que obliga a un ajuste muy preciso. Además, al tratarse de tejidos más caros, podría deberse también a que no se fabriquen este tipo de pantalones en tallas más grandes, lo que evidencia que la proporción de la talla XXL sea del 0 %.
- **Cluster 3 (Pantalones holgados):** Es el grupo más numeroso, incluyendo categorías como Ancho o Pinzas. Presenta una forma de curva muy similar al cluster 0, pero con una mayor proporción en la talla S, seguramente debido a su patrón holgado.

Al examinar la composición de cada cluster en la Tabla 4.3, se observa una fuerte agrupación por atributos físicos y de forma. El algoritmo, basándose únicamente en la distribución de ventas, logró capturar características internas de cada producto, separando las prendas según su corte, holgura o forma sin necesidad de utilizar otras variables explicativas.

Cluster	Talla dominante	Categorías
0	M	Joggers, Lino, Liso, Pijamero
1	M	Culotte, Especial, Estampado
2	XS	Piel
3	M	Ancho, Crepe, Fantasia, Flojo, Maxi, Pinzas, Recto

Tabla 4.3: Asignación de categorías de producto a los cuatro grupos identificados.

Finalmente, la creación de estos cuatro grupos resuelve el problema de la alta cardinalidad y aporta una solución para tratar los nuevos lanzamientos. Cuando la marca saque a la venta un nuevo

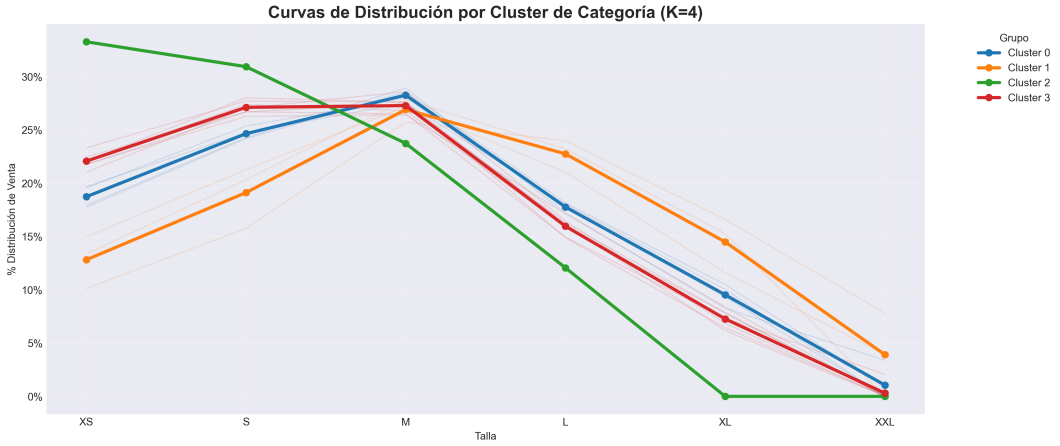

Figura 4.2: Perfiles de distribución de tallas por categoría obtenidos mediante K-Means ($k = 4$).

artículo sin histórico, el sistema lo asignará a su *cluster* correspondiente, heredando las características de ese grupo. Con las variables sintéticas ya definidas, se presenta el modelo de Regresión Logística Multinomial empleado en la fase de predicción final.

4.2. Predicción mediante Regresión Logística Multinomial

Una vez caracterizados el contexto geográfico y el tipo de producto mediante las variables generadas en la fase de Clustering (`cluster_pais` y `cluster_categoria`), se procedió a la construcción del modelo predictivo. El objetivo de esta etapa es entrenar un algoritmo capaz de calcular la probabilidad exacta de venta de cada talla condicionada a un vector de características. Para ello, se seleccionó el modelo de Regresión Logística Multinomial, preferido frente a otras alternativas por su interpretabilidad, pues permite conocer cómo afecta cada variable a la probabilidad de elegir una talla más grande o más pequeña estudiando sus coeficientes.

Base teórica Regresión Logística Multinomial

Como se describe en [6], la Regresión Logística Multinomial es la generalización del modelo logístico binario para tratar situaciones donde la variable dependiente Y es categórica y presenta más de dos categorías posibles (mutuamente excluyentes y no ordenadas). En el caso de este estudio, Y representa la talla del producto y toma valores en el conjunto $\mathcal{A} = \{XS, S, M, L, XL, XXL\}$ (donde el número de categorías es $K = 6$).

Explicación del modelo: En lugar de ajustar K modelos binarios independientes, la regresión logística multinomial permite estimar simultáneamente la probabilidad de pertenecer a cada categoría condicionada a un vector de características $\mathbf{X}$. Para lograrlo, la formulación matemática tradicional selecciona una categoría de referencia (denotada por J) y las probabilidades para las $J-1$ categorías restantes ($j = 1, 2, ..., J-1$) se modelan como:

$$P(Y = j|X) = \frac{\exp(\beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jk}X_k)}{1 + \sum_{h=1}^{J-1} \exp(\beta_{h0} + \beta_{h1}X_1 + \cdots + \beta_{hk}X_k)}. \tag{4.3}$$

Por su parte, la probabilidad para la categoría de referencia J queda definida por la diferencia necesaria para que la suma de todas las probabilidades sea exactamente igual a la unidad:

$$P(Y = J|X) = \frac{1}{1 + \sum_{h=1}^{J-1} \exp(\beta_{h0} + \beta_{h1}X_1 + \cdots + \beta_{hk}X_k)}. \tag{4.4}$$

Para predecir probabilidades, estas deben estar acotadas entre 0 y 1, lo cual dificulta el ajuste de una ecuación lineal. En ocasiones, esto supone un reto y, para solucionarlo, el modelo trabaja con la *log-odds* (el logaritmo neperiano de la razón de probabilidades respecto a la categoría de referencia). Esta transformación permite proyectar el espacio acotado de las probabilidades a una escala continua desde $-\infty$ hasta $+\infty$.

Para cualquier talla $j \neq J$, el modelo establece la siguiente relación lineal:

$$g_j(X) = \ln\left(\frac{P(Y = j|X)}{P(Y = J|X)}\right) = \beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jk}X_k. \tag{4.5}$$

Esta formulación permite una interpretación directa de los coeficientes β: un valor β_{jk} positivo indica que un aumento en la variable predictora X_k incrementa la probabilidad de que el artículo vendido sea de la talla j en comparación con la talla de referencia J.

Estimación de los parámetros: Dado que disponemos de n observaciones independientes, los parámetros del modelo se estiman mediante el método de Máxima Verosimilitud. La función de verosimilitud condicional que se maximiza es:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{J} P(Y = j|X_i)^{y_{ij}} \tag{4.6}$$

Donde y_{ij} es una variable indicadora que toma el valor 1 si la observación i es de la talla j, y 0 en caso contrario.

4.2.1. Transformación y preparación del dataset

Antes de entrenar el modelo, fue necesario preparar los datos mediante dos transformaciones para asegurar que el aprendizaje del modelo fuese adecuado:

Desagregación de transacciones

El dataset original agrupaba todas las ventas en una sola línea, sin embargo, para poder emplear los algoritmos de clasificación de la librería *Scikit-Learn* fue necesario que cada fila fuese una única observación. En consecuencia, se transformó cada registro con múltiples ventas en tantas filas como unidades vendidas indicase su variable `quantity`.

Se pasó a tener un dataset desagregado con 1.911.973 filas, donde cada línea representa la venta unitaria de un pantalón. Esta fase aseguró que el modelo asignase el peso correcto a aquellas combinaciones producto-tienda con mayor volumen de ventas.

Tratamiento de variables

Para construir la matriz de entrada $\mathbf{X}$, se seleccionaron y se procesaron las variables predictoras utilizando la herramienta `ColumnTransformer`. Su función es aplicar distintas reglas o transformaciones a cada tipo de variable, permitiendo tratar por separado las columnas numéricas de las categóricas.

1. **Variables categóricas:**
    - `cluster_pais` (4 grupos).

- `cluster_categoria` (4 grupos)
- `store_type` (2 clases: Calle vs Centro Comercial).
- `department` (7 clases: Fiesta, Básico, etc.).

Se aplicó `One-Hot Encoding` para tratar tanto los *clusters* como el resto de variables categóricas, convirtiendo cada categoría en una columna binaria independiente (0 o 1). Por ejemplo, `cluster_pais` se convirtió en cuatro columnas distintas (`cluster_pais_0`, `cluster_pais_1`, ...) y la variable `department` se dividió a su vez en tantas columnas como categorías distintas presentaba (`department_DISEÑO`, `department_FIESTA`, ...).

2. **Variable numérica:**

   - `price`: Precio unitario de la prenda.

   Se utilizó `StandardScaler` para reescalar el precio restando la media y dividiendo por la desviación típica. De esta forma, los valores de esta variable presentan un rango común, lo que facilita al algoritmo a converger más rápido.

Tras este proceso, la matriz de entrada resultante contó con 18 variables predictoras para explicar la variable objetivo Y (la talla).

4.2.2. Configuración e implementación del algoritmo

El modelo se implementó utilizando la clase `LogisticRegression` de la librería `Scikit-Learn` de Python [7]. Para garantizar el buen funcionamiento del modelo de clasificación multiclase se establecieron los siguientes hiperparámetros:

- `multi_class = 'multinomial'`: Se configuró el algoritmo seleccionando la función *Softmax*. Esta función permite que el modelo pueda tratar las tallas como opciones que se excluyen entre sí, asegurando que las probabilidades de las seis tallas sumen siempre el 100 %.
- `solver = 'lbfgs'`: Se eligió este algoritmo por ser la opción más rápida para conjuntos de datos de tamaño medio, pues minimiza el uso de memoria.
- `max_iter = 500`: Se aumentó el máximo de iteraciones ya que, al haber creado variables nuevas empleando el `One-Hot Encoding`, el algoritmo podría necesitar más pasos para alcanzar la solución óptima.
- `penalty = 'l2'`: El modelo aplica por defecto una regularización L_2 o Ridge. Con esto se añade un término de penalización a la función de coste durante el entrenamiento para evitar sobreajustes.
- `C = 1.0`: Se mantuvo el valor por defecto del parámetro C, el cual representa el inverso de la fuerza de regularización (C = 1 / λ). Un valor elevado de este parámetro reduce el efecto de la penalización del modelo, y viceversa.

4.2.3. Entrenamiento del modelo

Para validar la precisión del clasificador que se ha construido empleando los parámetros anteriores, se dividió el conjunto de datos en dos muestras: el 80 % de las transacciones se emplearon para entrenar y ajustar los coeficientes del modelo y el 20 % restante se reservó para evaluar métricas de error sobre datos no vistos por el algoritmo.

Para realizar esta partición, a diferencia de la separación cronológica realizada al inicio del proyecto, en este caso se aplicó un muestreo aleatorio estratificado (`stratify = y`). Esta técnica fuerza al

conjunto de test a mantener la misma proporción de tallas que el conjunto original. Así, se asegura que se esté trabajando con una muestra representativa de la demanda y se garantiza que sea posible evaluar si el modelo detecta correctamente las tallas minoritarias.

Capítulo 5

Evaluación del modelo e impacto de las variables

Una vez entrenado el algoritmo de Regresión Logística Multinomial, fue necesario evaluar su rendimiento utilizando el conjunto de datos de *test* reservado previamente. Esto se llevó a cabo desde tres perspectivas: identificando qué variables influyeron realmente en la elección de la talla y de qué forma, midiendo el error de las predicciones a nivel de grupos de comportamiento y comprobando la precisión del modelo al predecir.

5.1. Importancia e impacto de las variables predictoras

Como primera medida del rendimiento del algoritmo, fue fundamental interpretar los coeficientes extraídos (β) para comprender cómo las distintas variables influyeron en la venta de cada talla. Como se detalló en la base teórica, estos coeficientes no son probabilidades directas, sino que representan el cambio en la *log-odds*.

El signo del coeficiente indica la dirección del impacto. Un β positivo indica que la presencia de esa variable favorece la venta de dicha talla, mientras que un valor negativo la reduce.

Para cuantificar este impacto y poder interpretarlo, se aplicó además la función exponencial sobre el coeficiente (e^{β}), obteniendo así el *Odds Ratio*. Este valor funciona como un factor que multiplica la probabilidad inicial de venta. Por ejemplo, un coeficiente de $\beta = 0{,}893$ se traduciría en un *Odds Ratio* de $e^{0{,}893} \approx 2{,}44$. Esto significa que la presencia de dicha variable multiplica por más de dos veces la probabilidad de que el cliente compre esa talla en concreto.

Para facilitar el análisis, se extrajo la matriz completa de coeficientes y se representó visualmente mediante un mapa de calor (*heatmap*), el cual se muestra en la Figura 5.1.

Estos pesos revelan patrones de compra relevantes para la toma de decisiones. En concreto, el modelo de regresión mostró la existencia de una segmentación clara por grupos de países y por tipo de pantalones, lo cual ya se intuía en fases previas.

- **Validación de los *clusters* por países:** El modelo asigna mayor peso a las variables relativas a los grupos de países. Esto confirma que las diferencias físicas entre clientes de distintas regiones son la clave principal del tallaje de pantalones.

    - Como se observa en el mapa de calor, el `cluster_pais_2` (mercados asiáticos) presenta los coeficientes positivos más fuertes del modelo para la talla XS ($\beta_{XS} = 0{,}893$) y S ($\beta_S =$

0,656), mientras que presenta un valor muy negativo para la XXL ($\beta_{XXL} = -1{,}681$). Es decir, la *Odds Ratio* para la talla XS es de $e^{0{,}893} \approx 2{,}44$, lo que significa que, en los mercados asiáticos, la tendencia a que un pantalón vendido sea de la talla XS es 2.44 veces superior a la de una situación estándar.

- Por el contrario, el `cluster_pais_1` (mercados árabes) muestra un comportamiento opuesto, penalizando fuertemente la talla XS ($\beta_{XS} = -0{,}661$) y promoviendo las ventas de la XXL ($\beta_{XXL} = 0{,}741$).

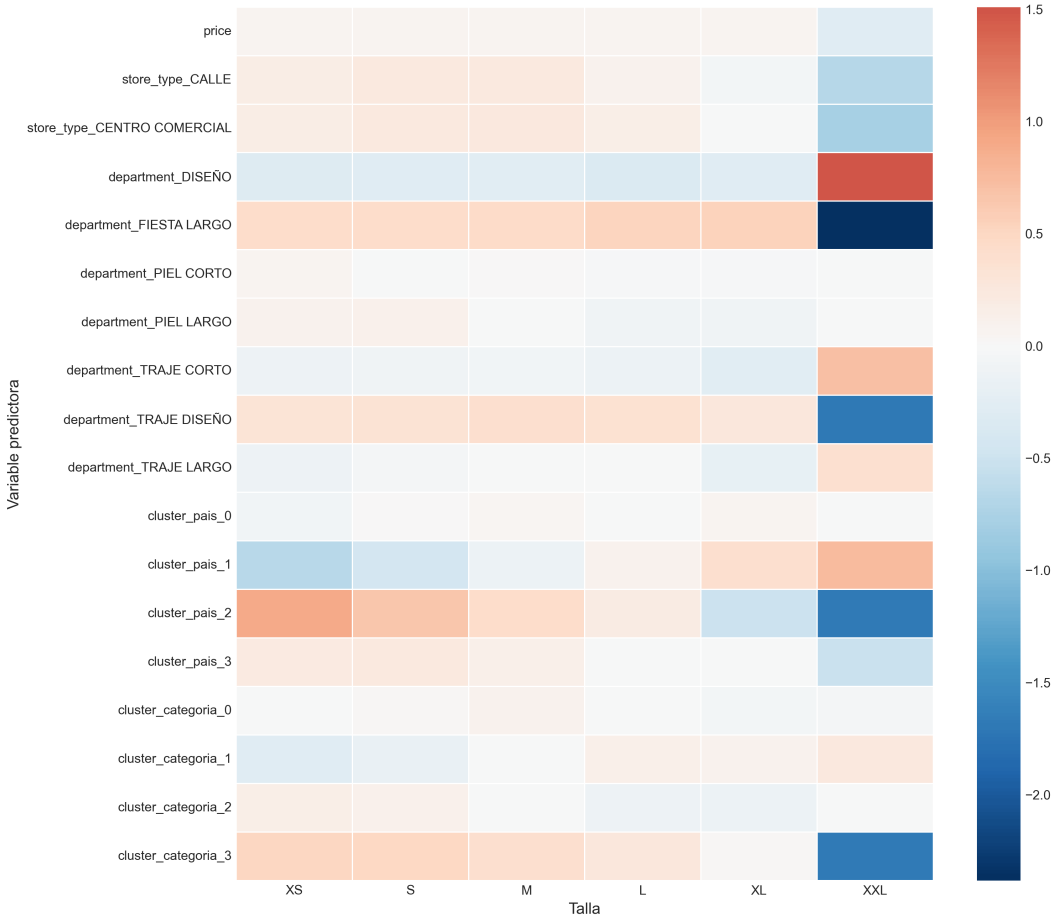

Figura 5.1: Mapa de calor de los coeficientes del modelo de Regresión Logística Multinomial. Los colores cálidos (rojos) indican un aumento en la probabilidad de venta de la talla, mientras que los colores fríos (azules) indican una disminución.

- **Validación de los *clusters* por categoría:** El modelo también detectó la forma del producto. En cuanto a las categorías de pantalón, destaca el hecho de que el `cluster_categoria_2` (pantalones de piel) desplaza la demanda hacia la izquierda, impulsando las tallas XS y S frente a las tallas más grandes.

  Por su parte, el `cluster_categoria_3` (pantalones holgados), presenta un fuerte impulso en las tallas pequeñas ($\beta_{XS} = 0{,}511$, multiplicando por 1,67 la probabilidad) y reduce la probabilidad de venta de la talla XXL ($\beta_{XXL} = -1{,}681$).

- **Variables poco influyentes:** Destaca el hecho de que hay variables que no aportan información al modelo para clasificar correctamente. La variable `price` presenta coeficientes casi nulos (con valores alrededor de 0,05 para la mayoría de tallas), lo que indica que las ventas casi no varían en función del precio de la prenda. Del mismo modo, las variables `store_type_CALLE` y `store_type_CENTRO COMERCIAL` muestran vectores de coeficientes casi idénticos, demostrando que el perfil de tallas del comprador no se ve alterado por la ubicación física del establecimiento.

  Aunque se podría proceder a eliminar estas variables no significativas para reentrenar un modelo más simple, al usar una Regresión Logística con penalización Ridge (L_2) no es estrictamente necesario este paso. Esta regularización se encarga de contraer hacia cero la magnitud de los coeficientes, y así evita que exista sobreajuste. Esto pretende mitigar el efecto de las variables con menos relevancia, sin llegar a eliminarlas del modelo completamente.

  En resumen, esto sugiere que el precio y el tipo de tienda no influyen en la elección de la talla.

- Efectos de los departamentos: A pesar de que los *clusters* agrupan el comportamiento general, la variable `department` captura comportamientos más específicos. Por ejemplo, el departamento de `DISEÑO` muestra una tendencia extrema hacia la talla XXL ($\beta_{XXL} = 1{,}508$), multiplicando su probabilidad de venta por 4.52, mientras que el departamento de `FIESTA LARGO` penaliza esta talla ($\beta_{XXL} = -2{,}383$).

En conclusión, el análisis de estos pesos confirma que el modelo ha sido capaz de aprender el comportamiento real de los clientes, permitiéndole predecir la asignación de tallas de forma personalizada adaptándose al contexto particular de cada compra. De esta forma, se garantiza una asignación de inventario mucho más inteligente y precisa.

5.2. Precisión en la estimación de la demanda agregada (MAE)

Una vez analizada la influencia de las variables, el siguiente paso fue verificar si el modelo era capaz de capturar el comportamiento real del mercado. En esta sección, se midió la capacidad del algoritmo para predecir la distribución agregada de la demanda, más allá de acertar una venta puntual. Así, se evaluó el valor estratégico de este modelo.

Se decidió evaluar el error a nivel de *cluster* para así replicar la realidad del sector retail, donde la planificación y flujo de inventario se organiza en ocasiones por regiones específicas o familias de productos con categorías comunes. En este contexto, no se utiliza el precio ya que no se considera un factor relevante a la hora de asignar el stock. Además, en el apartado anterior se descubrió que se trataba de una variable sin influencia en la toma de decisiones.

Base teórica del Error Absoluto Medio (MAE)

Para estudiar la precisión, se comparó la frecuencia relativa real observada de ventas en el conjunto de test frente a la estimada por el modelo para cada grupo de comportamiento. Para cada transacción, el modelo genera un vector de probabilidades. Se calculó el promedio de estos vectores para cada *cluster*, obteniendo así la curva de demanda estimada para ese grupo.

El Error Absoluto Medio (MAE) para un *cluster* concreto, pongamos c, se define como la desviación promedio en valor absoluto entre la proporción real observada y la predicción promedio para el conjunto de tallas $\mathcal{A} = \{XS, S, \ldots, XXL\}$:

$$MAE_c = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \left| f_{c,j} - \hat{\mathbb{P}}(Y = j \mid c) \right|, \tag{5.1}$$

donde

- $|\mathcal{A}|$ es el número de total de tallas que pertenecen al conjunto $\mathcal{A}$
- $f_{c,j}$ es la frecuencia relativa real de la talla j en el *cluster* c, calculada de forma análoga a como se detalla en la Ecuación (3.1), pero agregando las ventas a nivel de *cluster*.
- $\hat{\mathbb{P}}(Y = j \mid c)$ es la probabilidad media estimada por el modelo para la talla j.

Se calculó esta métrica para tratar de validar si el algoritmo es capaz de estimar correctamente la curva de demanda específica de cada mercado.

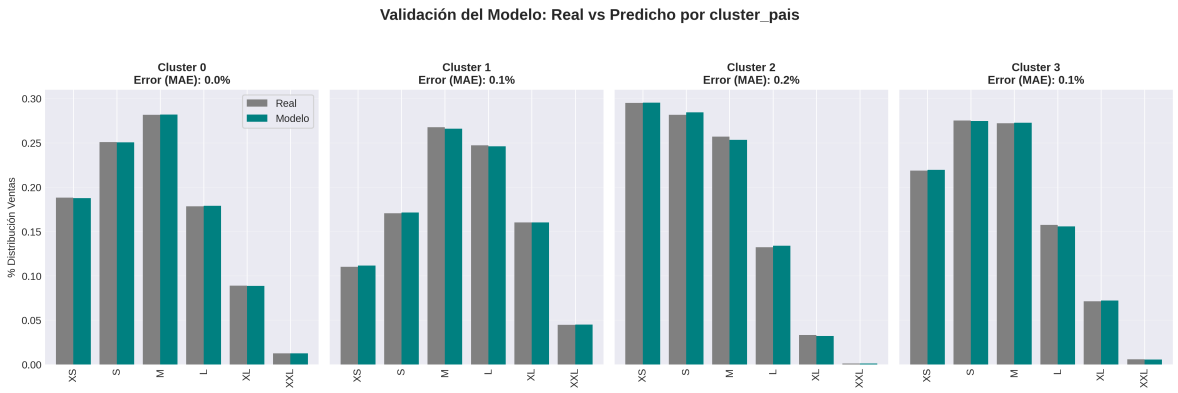

Figura 5.2: Comparación de la distribución real de tallas (en gris) frente a la predicha por el modelo (verde) por cluster de país.

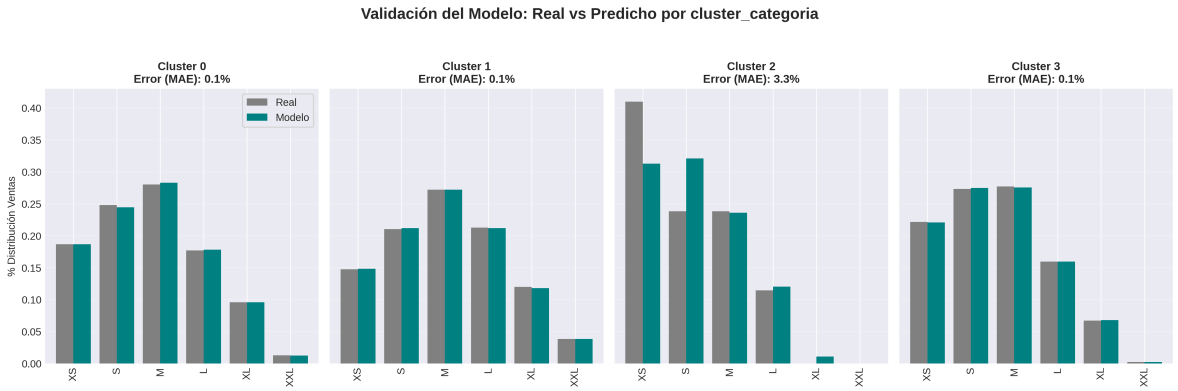

Figura 5.3: Comparación de la distribución real de tallas (en gris) frente a la predicha por el modelo (verde) por cluster de categorías.

Interpretación de los resultados

Los resultados obtenidos, representados visualmente en las Figuras 5.2 y 5.3, confirman la fiabilidad del modelo al capturar los matices de la demanda.

- En los cuatro *clusters* de países, el MAE presenta valores muy bajos, de entre 0,0 % y 0,2 %. Esto demuestra que el modelo no es genérico, sino que los coeficientes asignados a cada región (como la tendencia en el mercado asiático hacia la talla XS) lograron capturar las diferencias en la demanda por tallas de las distintas regiones.

- Con respecto a los grupos de comportamiento por familias de pantalones, la precisión se mantuvo constante en la mayoría de grupos. Sin embargo, en el `cluster_categoria_2` (correspondiente a los pantalones de piel), se observa un MAE del 3,3 %. Visualmente, se detecta que el modelo suaviza el comportamiento real de la demanda: subestima la concentración extrema de ventas en la talla XS (predice alrededor de un 30 % frente al 40 % real) y sobrestima ligeramente las tallas grandes. Esto sugiere que, al ser un grupo con un menor tamaño muestral, el algoritmo prefiere ser más conservador y tender hacia la media global para evitar el sobreajuste.

Incluso en su peor escenario (error del 3,3 %), la Regresión Logística Multinomial es una mejor alternativa frente al uso de un modelo de promedios globales. Mientras que el promedio global ignoraría los distintos comportamientos de los grupos, el modelo propuesto los identifica.

5.2.1. Análisis discriminante del modelo

Para complementar el análisis del error absoluto de las curvas agregadas, se evaluó el poder discriminante del modelo a nivel transaccional. Mientras que la sección anterior se centró en la precisión de la estimación de las proporciones de tallas a nivel agregado, el análisis discriminante busca cuantificar la capacidad del algoritmo para clasificar correctamente [8].

Base teórica del AUC-ROC multiclase

La curva ROC es una representación gráfica de la relación entre la sensibilidad (tasa de verdaderos positivos) y la especificidad (1 - tasa de falsos positivos) para todos los posibles umbrales de corte de un clasificador. El área bajo esta curva (AUC) proporciona una medida agregada del rendimiento del modelo en todos los umbrales posibles. Matemáticamente, el AUC representa la probabilidad de que el modelo asigne una puntuación (probabilidad) más alta a una instancia positiva elegida al azar que a una negativa elegida al azar [9].

Como la elección de la talla es un problema multiclase con seis categorías mutuamente excluyentes (XS a XXL), el cálculo del AUC necesita una estrategia de agregación. En este trabajo se optó por el uso de la medida AUC-ROC *One-vs-Rest (OvR)* ponderada. Esta metodología descompone el problema en seis comparaciones por pares (por ejemplo, la talla M frente al resto de tallas combinadas) y calcula el AUC para cada una, ponderando el resultado final según la presencia de cada clase en el conjunto de datos original [9].

Interpretación de los resultados

Para la implementación de esta medida se utilizó la función `roc_auc_score` de la librería `Scikit-Learn` de Python. En la configuración del algoritmo, se establecieron estos dos parámetros:

- `multi_class='ovr'`: Implementa la estrategia *One-vs-Rest*, descomponiendo el problema en seis comparaciones binarias.
- `average='weighted'`: Calcula la media ponderada de los AUC individuales según la proporción de cada clase. Esta configuración es clave ya que las tallas se encuentran desbalanceadas, y el rendimiento podría verse distorsionado por la presencia de tallas centrales mayoritarias.

El resultado global obtenido tras la evaluación fue un AUC de 0.5528. En la escala de interpretación de Hosmer y Lemeshow [6], un valor de 0.5 indica un poder discriminante nulo (equivalente al azar), mientras que valores cercanos a 0.7 sugieren una separabilidad aceptable.

Este resultado evidencia que el modelo no posee un gran poder de separación a nivel de transacción individual. Es decir, el algoritmo no es capaz de identificar qué talla exacta elegirá un cliente específico basándose únicamente en el contexto geográfico y de producto. No obstante, como se demostró en la

sección anterior con el análisis del MAE, el modelo es capaz de capturar la curva de demanda de cada grupo de comportamiento. Aunque no prediga con certeza cada venta individual, el sistema asegura con cierta precisión que la curva de tallas estimada coincidirá con el comportamiento agregado del mercado.

Adicionalmente, se generaron diagramas de cajas (*Boxplots*) para analizar la distribución de las probabilidades predichas en función de la talla real finalmente comprada. Dichas representaciones se observan en la Figura 5.4.

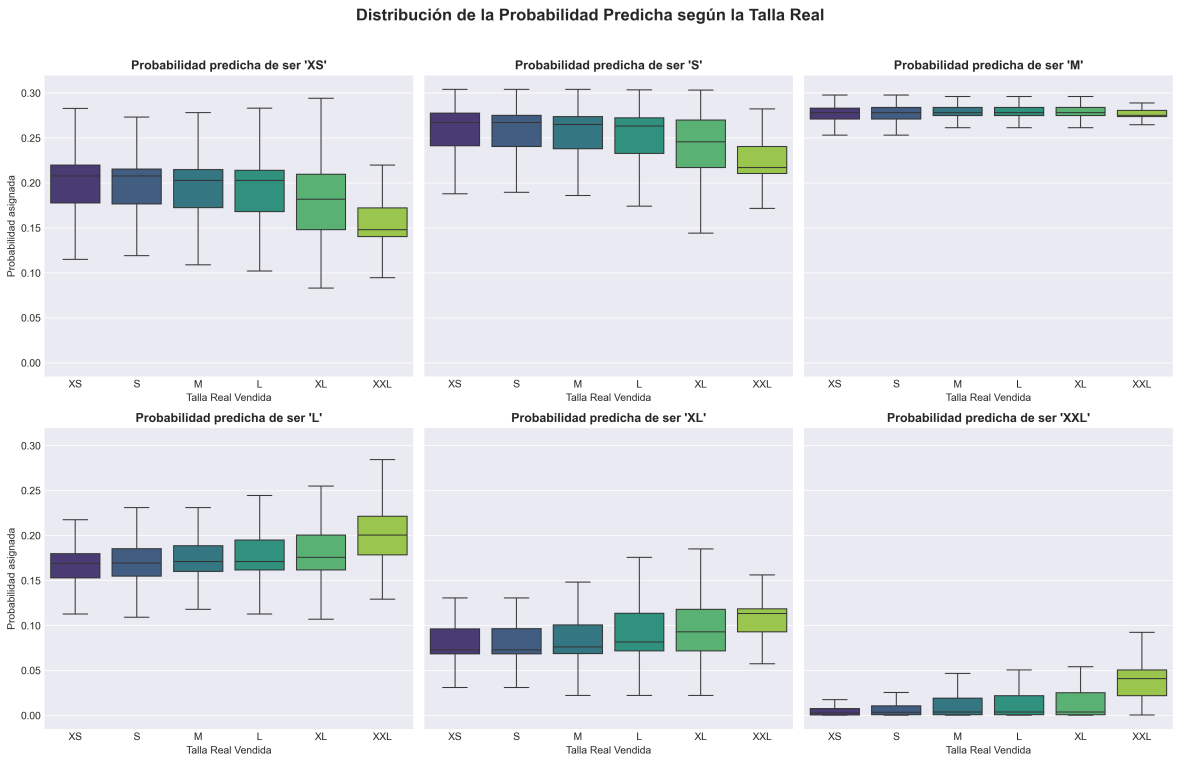

Figura 5.4: Distribución de la probabilidad predicha frente a la talla real adquirida.

En primer lugar, se observa que las cajas están planas y superpuestas en la mayoría de los gráficos anteriores. Esto, sumado a que el AUC ha devuelto un valor cercano a 0.5 podría interpretarse como una falta de capacidad predictiva del modelo. Sin embargo, al analizar los diagramas de cajas en detalle se obtienen matices valiosos que explican la naturaleza de los datos con los que se está trabajando.

- **Sensibilidad en las tallas extremas:** A pesar del solapamiento general de las cajas, el modelo sí logra asignar una probabilidad ligeramente superior a la talla correcta en los extremos de la curva (XS, XL y XXL). Como se observa en sus respectivas gráficas, la mediana de la probabilidad predicha para la talla real se sitúa por encima de las demás de forma sutil. Por el contrario, en las tallas centrales (S, M y L), esta distinción se difumina casi por completo.

  Este comportamiento refleja una realidad de los datos: ante la ausencia de información de contexto relativa al cliente, para el modelo es muy difícil distinguir entre tallas centrales y similares (como la S y la M), pero le resulta más fácil reconocer las situaciones en las que se venderán tallas extremas, ya que estas suelen estar vinculadas a factores geográficos o de categoría mucho más definidos.

- **Predicción a nivel individual:** Como se adelantó, el modelo es incapaz de predecir con exactitud qué talla elegirá un cliente individual en una transacción aislada. Esto es lógico desde el punto de vista del negocio, pues las variables predictoras disponibles (país, tipo de tienda, categoría del producto) caracterizan el entorno, pero se carece de información a nivel de usuario (historial de compras, medidas corporales, preferencias de ajuste holgado/ceñido). Ante la incertidumbre individual, el modelo asume un comportamiento conservador, asignando probabilidades cercanas a la distribución promedio para minimizar el error.
- **Predicción a nivel de grupo:** Como se demostró previamente, el modelo sí posee una buena capacidad para predecir la distribución global de la demanda (MAE $< 0.2\,\%$). Aunque no sabe qué talla comprará un cliente en particular, sí sabe predecir con alta fiabilidad qué proporción de tallas se venderán tras una cantidad mayor de clientes.

En conclusión, estas métricas confirman que el algoritmo desarrollado no está diseñado para actuar como un Sistema de Recomendación personalizado. Por el contrario, sí que funciona como una herramienta para la asignación de stock a nivel de grupos de comportamiento.

Capítulo 6

Resultados y simulación de negocio

El paso final del proyecto consistió en evaluar su impacto y su beneficio económico para la empresa de retail. Para ello, se diseñó una simulación sobre el conjunto de datos de validación reservado al inicio, y que agrupaba las transacciones reales entre el 19 de septiembre y el 9 de octubre de 2025. Estas semanas representan un escenario futuro desconocido para el algoritmo.

6.1. Diseño del experimento

El objetivo de esta simulación fue medir el beneficio que aporta la asignación correcta de tallas frente a una estrategia simplificada. Para que el análisis fuese riguroso y los resultados comparables, se definieron dos escenarios de gestión que se aplicaron sobre las mismas transacciones reales:

- **Modelo base:** Como modelo base de referencia se utilizó un modelo estadístico simple basado en promedios. Este modelo reparte el volumen total de prendas utilizando la media histórica global de ventas del producto, aplicando un porcentaje fijo para cada talla, el cual se puede observar en la Tabla 6.1. En este escenario, la distribución de tallas fue idéntica para todas las tiendas y países, ignorando por completo el contexto.
- **Modelo predictivo:** Asigna las tallas utilizando las probabilidades estimadas por el modelo de Regresión Logística Multinomial desarrollado en el proyecto, el cual adapta la curva al contexto específico (país, tienda y atributos del producto).

Talla	**XS**	**S**	**M**	**L**	**XL**	**XXL**
Probabilidad	19.8 %	24.2 %	26.3 %	16.5 %	8.7 %	4.5 %

Tabla 6.1: Distribución de probabilidad por talla alfabética (curva promedio global utilizada por el modelo base).

Tal y como se razonó en capítulos previos, los pantalones de tallaje numérico presentaban un patrón de ventas más homogéneo, por lo que se les asignó la curva promedio global y el experimento se centró en evaluar el rendimiento sobre el sistema de tallaje alfabético.

Por otra parte, se asumió que ambos modelos conocían de antemano el volumen exacto de pantalones a vender. Así, al comparar el modelo predictivo con el modelo base, la única variable que difiere es la distribución de tallas que cada sistema asigna a un mismo volumen de artículos.

6.1.1. Nivel de agregación

Es importante destacar una diferencia metodológica que se llevó a cabo entre la fase de entrenamiento y la fase de simulación. Mientras que los modelos predictivos del Capítulo 4 se entrenaron a nivel transaccional diario para capturar patrones con más precisión, la simulación de este capítulo agrupa la demanda real bajo el nivel de agregación Tienda-Producto-Semana. Esta decisión de diseño responde a tres motivos fundamentales:

1. En la industria del sector retail, la reposición del inventario no se gestiona diariamente para cada referencia individual, sino mediante envíos semanales o bisemanales desde los centros de distribución. Agrupar la demanda de forma semanal permite replicar con exactitud los tiempos de suministro, adaptando el experimento a la realidad.

2. A nivel diario, es muy común que una tienda venda una única unidad de un modelo concreto de pantalón. Si el algoritmo de asignación se aplicara sobre unidades aisladas, se vería forzado a enviar siempre la talla con la probabilidad máxima (generalmente la talla M), eliminando por completo la diversidad de la curva de tallas a lo largo del tiempo. Al agrupar por semanas, algoritmo presenta un volumen total suficiente para poder asignar una distribución de probabilidades. Además, de esta forma no se pierde información de las tallas extremas.

3. Las ventas diarias presentan una alta volatilidad, incluyendo días de demanda cero o picos anómalos de compras individuales. La semana actúa como la unidad temporal mínima y óptima para suavizar este ruido.

6.1.2. Asignación de unidades de producto

Una vez obtenidas las densidades de probabilidad por talla para cada sistema, fue necesario convertir esas proporciones en unidades enteras de stock para garantizar que el total de prendas a enviar coincidiese con el volumen real previsto.

Para resolver este problema de prorrateo y evitar errores, se aplicó el Método del Resto Mayor (también conocido como método de Hamilton [10]), un algoritmo de reparto inteligente. Se eligió este procedimiento ya que permite respetar de manera más fiel la proporción real de tallas minoritarias. El sistema asigna inicialmente las unidades correspondientes a la parte entera de cada talla y, a continuación, reparte las prendas sobrantes priorizando aquellas con la mayor parte decimal. Esto asegura que no haya diferencias entre el inventario asignado y la demanda (que en este caso se conoce de antemano).

Ejemplo ilustrativo:

Para ilustrar el funcionamiento de este algoritmo, se presenta un ejemplo práctico. Se supone un escenario en el que se conoce que se venderán exactamente 10 unidades de un pantalón concreto en una tienda específica, y se asignan las probabilidades del modelo base (promedio global). El proceso de asignación consta de tres pasos, detallados en la Tabla 6.2:

1. En primer lugar, se multiplica el volumen total (10) por la probabilidad de cada talla. La suma de las partes enteras de este cálculo determina 6 prendas iniciales, sobrando 4 prendas restantes.

2. Tras restar las partes enteras, se ordenan las tallas según su parte decimal de mayor a menor (XS: 0.98; XL: 0.87; L: 0.65; M: 0.63; XXL: 0.45; S: 0.42).

3. Finalmente, las 4 prendas sobrantes se asignan añadiendo una unidad a las 4 tallas con los mayores valores en la lista anterior.

Talla	Probabilidad	Stock inicial	Parte entera	Parte decimal	Asignación resto	Stock final
XS	19.8 %	1.98 uds.	1	0.98 (1º)	+1	**2 uds.**
S	24.2 %	2.42 uds.	2	0.42 (6º)	0	**2 uds.**
M	26.3 %	2.63 uds.	2	0.63 (4º)	+1	**3 uds.**
L	16.5 %	1.65 uds.	1	0.65 (3º)	+1	**2 uds.**
XL	8.7 %	0.87 uds.	0	0.87 (2º)	+1	**1 ud.**
XXL	4.5 %	0.45 uds.	0	0.45 (5º)	0	**0 uds.**
Total	**100 %**	**10 uds.**	**6 uds.**	**-**	**4 uds.**	**10 uds.**

Tabla 6.2: Ejemplo de asignación de 10 unidades de producto aplicando el Método del Resto Mayor sobre la curva promedio del modelo base de promedios.

Como se observa en el ejemplo, este método es clave para representar correctamente la presencia de las tallas minoritarias. Si se hubiese empleado una técnica de redondeo estándar se habría dejado a la talla XL con 0 unidades (al ser $0.87 < 1$), perdiendo la representación de la curva original.

6.1.3. Cálculo del impacto

A continuación, se procedió a comparar la asignación de unidades de stock propuesta por ambos modelos con la demanda real observada en el conjunto de *holdout*.

Una venta de un producto no queda determinada por tener demanda del mismo, sino por la disponibilidad física del artículo en la tienda y, en concreto, en la talla exacta. Por ello, para cada combinación de tienda, producto y talla (SKU), se define la Venta Real (V) como el mínimo entre la demanda (D) y el stock disponible (S): $V = \text{mín}(D, S)$. De esta forma, la simulación es realista:

- Si el modelo asigna más stock del que se pide ($S > D$), la venta se limita a D y el exceso de inventario se considera *sobrestock*.
- Si la demanda supera el stock disponible ($D > S$), la venta es S y se genera una rotura de stock o venta que no se llega a realizar.

Finalmente, se tienen en cuenta los precios unitarios correspondientes a cada referencia (P) para calcular las métricas de negocio de interés. De esta forma, se obtienen los dos KPIs que se midieron durante esta simulación:

1. **Ingresos totales (I):** Representan el ingreso económico de la demanda satisfecha.

$$I = \sum (V \times P)$$

2. **Ventas no realizadas (L):** Representa el ingreso que no se llega a generar al no tener la talla correcta en el momento adecuado.

$$L = \sum ((D - V) \times P)$$

6.2. Análisis de resultados económicos

Los resultados de la simulación, aplicados sobre un volumen total de 89.702 combinaciones únicas de Tienda-Producto-Semana, revelaron una mejora en la eficiencia del inventario.

Como se adelantó al inicio del capítulo, para evitar que los resultados se alterasen por posibles errores en la predicción del volumen total de ventas, se asumió que ambos sistemas (el modelo desarrollado y el de referencia) conocían de antemano las unidades exactas de artículos que se iban a vender en cada tienda. De esta forma, se garantizó que la variable objeto de estudio (`size`) quedase completamente aislada. Es decir, ambos modelos conocían con exactitud cuántas unidades de una referencia concreta necesitaría una tienda y semana determinadas. La diferencia en el rendimiento de cada sistema residiría, por tanto, en su capacidad para predecir la distribución de tallas.

Modelo / Escenario	**Eficiencia**	**Unidades vendidas**	**Ingresos totales**	**Ventas no realizadas**
Demanda real total	*100 %*	*264.247 uds.*	*13.489.303 €*	*0 €*
Modelo predictivo	**66.7 %**	**176.260 uds.**	**8.978.868 €**	**4.510.435 €**
Modelo base (promedios)	43.1 %	113.880 uds.	6.180.767 €	7.308.536 €
Diferencia (Impacto)	**+23.6 %**	**+62.380 uds.**	**+2.798.101 €**	**-2.798.101 €**

Tabla 6.3: Comparación del rendimiento económico en el periodo de validación (3 semanas). La eficiencia representa el porcentaje de la demanda real capturada por cada sistema.

Como se detalla en la Tabla 6.3, en esta simulación se tomó como referencia un escenario de demanda ideal total de aproximadamente 13.48 millones de euros (equivalente a 264.247 unidades vendidas), lo que define el límite máximo de eficiencia del sistema. Al comparar ambos modelos entre si, se pudieron extraer las siguientes conclusiones:

1. El modelo propuesto logró capturar 62.380 unidades adicionales respecto al método base, lo que supone un beneficio de 2.79 millones de euros (+45.3 %). Como el volumen total de prendas era el mismo para ambos sistemas, esto demuestra que la mejora se produjo debido a la optimización de la asignación de tallas.

2. El modelo base sugiere una distribución de tallas que no se adecua a ningún mercado concreto de forma fiel. Esto provoca roturas de stock al enviar tallas centrales que no se demandan, y se pierde oportunidad de ventas en tallas extremas. El modelo propuesto optimiza el flujo de mercancía eliminando estos desajustes.

En conclusión, los resultados demuestran que el éxito en el sector retail no depende únicamente de acertar con el volumen total de prendas enviadas a una tienda, sino de la precisión en la composición interna del surtido de prendas.

Capítulo 7

Conclusiones y líneas futuras

El presente Trabajo de Fin de Máster ha abordado uno de los desafíos logísticos más complejos y costosos de la industria del retail y del *fast fashion*: la optimización en la asignación de tallas. Tras el desarrollo, entrenamiento y validación de los modelos analíticos, se obtuvieron una serie de conclusiones técnicas y de negocio, así como la identificación de las limitaciones del estudio y las posibles líneas futuras de trabajo.

7.1. Conclusiones

Se han obtenido las siguientes conclusiones generales sobre la totalidad del proyecto:

1. Los resultados confirman que el uso de promedios generales para decidir el reparto de tallas es ineficiente. Se ha demostrado en este estudio que para maximizar la eficiencia no se deben usar modelos de distribución rígidos, sino usar sistemas más flexibles capaces de ajustar la oferta a las características y los hábitos de consumo de cada mercado y tienda de forma individual.

   En la industria de la moda, la ausencia de la talla demandada por el cliente se traduce en una venta perdida. Al no existir productos que lo sustituyan inmediatamente, cualquier error en el surtido de tallas provoca ventas perdidas y un impacto negativo en el consumidor.

   En este caso, el modelo propuesto no solo optimiza el rendimiento financiero, sino que mejora la experiencia del cliente al garantizar una mayor disponibilidad de producto en el punto de venta.

2. Uno de los mayores logros del sistema propuesto es capacidad para estimar la demanda de artículos que nunca se han vendido. Al añadir la información de contexto de mercados y productos al modelo estadístico, el sistema ha aprendido el comportamiento de productos similares. Esto soluciona el problema de la falta de datos históricos, permitiendo que una prenda nueva herede el comportamiento de sus grupos afines y tenga una curva de tallas precisa desde su primer día en tienda.

3. Al usar un escenario de prueba real se ha demostrado que el modelo predictivo desarrollado ha superado al sistema de promedios básico.

7.2. Limitaciones del estudio

A pesar de las mejoras obtenidas, las métricas de evaluación analizadas en capítulos anteriores revelan que el modelo actual presenta limitaciones. Esto podría deberse a la falta de disponibilidad y cantidad de datos de contexto. En particular, su bajo poder discriminante individual (AUC de 0.55) recalca este hecho.

Cabe destacar que el sistema ha logrado obtener una mejora del 45 % utilizando únicamente datos básicos de transacciones (identificador de tienda, país, categoría de producto y precio), pero no tiene conocimiento de variables y factores externos adicionales, entre los que destacan:

- **Datos del cliente y población:** Sería interesante tener datos sobre los perfiles y medidas corporales de la población local para refinar la curva de tallas de cada mercado. Incluso contar con el historial de compras individual de los consumidores.
- **Contexto:** Resulta esencial poseer información sobre factores externos como el clima local, el calendario de festivos o los periodos promocionales y de rebajas. Sin estos indicadores, el modelo ignora los picos de demanda provocados por eventos externos.
- **Restricciones físicas:** El sistema no tiene en cuenta restricciones del punto de venta, como la capacidad de las tiendas y del espacio de almacenamiento disponible, las cuales condicionan la disponibilidad del producto.

7.3. Líneas futuras de trabajo

El modelo diseñado sirve de base para construir una herramienta logística completa. Se proponen las siguientes líneas para continuar con el desarrollo del proyecto:

1. El siguiente paso sería desarrollar un modelo predictivo que estime la cantidad total de prendas a enviar. De esta forma, la empresa contaría con un sistema capaz de decidir tanto la cantidad de prendas a enviar como su composición por tallas.
2. También resultaría crucial añadir información externa al sistema. El uso de datos sobre el clima o las características de la población local permitiría al algoritmo ajustar mucho mejor sus predicciones en cada zona.
3. Explorar otras alternativas de modelado, esto es, probar otros algoritmos de *Machine Learning* y explorar nuevas configuraciones de hiperparámetros, como cambios en los métodos de regularización o aumento o disminución del número de iteraciones.
4. Estudiar la relación entre las devoluciones y el tallaje para entender qué perfiles de cliente no encuentran su talla ideal.
5. También se podría utilizar el modelo para agregar descuentos en prendas o tallas concretas con menor probabilidad de venderse.

En conclusión, en el proyecto desarrollado se ha creado una herramienta capaz de optimizar el reparto de tallas de una compañía de retail y se ha demostrado que es posible reducir el desperdicio de stock y mejorar el servicio del cliente frente a una alternativa básica.

Bibliografía

[1] Caro F, Gallien J (2010) Inventory Management of a Fast-Fashion Retail Network. Operations Research 58:257-273.

[2] Liu B, Ren T, Choi TM, Wee HM (2020) Forecasting in fashion operations: A literature review. International Journal of Production Economics 219:109-124.

[3] Fildes R, Ma S, Kolassa S (2022) Retail forecasting: Research and practice. International Journal of Forecasting 38:1283-1318.

[4] Thomassey S (2010) Sales forecasts in clothing industry: The key success factor of the supply chain management. International Journal of Production Economics 128:470-483.

[5] Murphy KP (2012) Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge.

[6] Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression (3rd ed.). John Wiley & Sons.

[7] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825-2830.

[8] Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27:861-874.

[9] Hand DJ, Till RJ (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45:171-186.

[10] Gallagher M (1991) Proportionality, Disproportionality and Electoral Systems. Electoral Studies 10:33-51.