Mondragon Unibertsitatea

Facultad de Empresariales

Grado en

Business Data Analytics

Informe final

Reto 04: Eroski

Curso: 1º 2024-2025

Equipo: Naranja

Autores:

- Uxue Duran
- Luken Larrea
- Daniel Alejandro Soponaru
- Jon Basarte
- Naia Parte
- Helene Urreta

Índice:

Índice de Figuras:

Índice de Tablas:

1 Introducción

En un entorno digital cada vez más competitivo, la **personalización** se ha convertido en un elemento clave para mejorar la experiencia de compra de los usuarios y fomentar su **fidelización**. Las empresas que logran adaptar sus servicios y recomendaciones a las necesidades reales de sus clientes consiguen no solo aumentar sus ventas, sino también **reforzar su relación** con ellos.

Este informe se enmarca en un proyecto propuesto por EROSKI, una de las principales compañías del sector de la distribución en España, que busca mejorar su canal de comercio electrónico a través del desarrollo de **sistemas de recomendación** personalizados. A partir del análisis de datos reales de transacciones, el objetivo es **construir modelos** capaces de **sugerir productos relevantes** a distintos perfiles de clientes en contextos concretos.

Para ello, se plantean cuatro situaciones prácticas en las que se pretende generar recomendaciones que respondan a necesidades diversas: **promocionar un producto concreto, sugerir artículos adicionales en el carrito, seleccionar el mejor producto de una lista en oferta y detectar posibles olvidos de compra**. La implementación de estas recomendaciones requiere un trabajo previo de exploración, tratamiento y modelado de datos, con el fin de obtener una matriz útil que sustente decisiones precisas y contextualizadas.

El presente informe detalla el enfoque adoptado para abordar este reto, describiendo tanto la lógica aplicada en cada situación como las técnicas utilizadas para extraer valor de los datos disponibles.

2 Identificar la problemática

2.1 Sobre el cliente

EROSKI es una de las principales empresas del sector de la distribución en España, perteneciente al Grupo Mondragón, y consolidada como una cooperativa de consumidores con una amplia implantación a nivel nacional. Fundada en 1969, EROSKI ha desarrollado una **sólida red comercial** que incluye supermercados, hipermercados, tiendas especializadas y una plataforma de comercio electrónico, a través de la cual ofrece una amplia gama de productos de alimentación, hogar, salud y cuidado personal.

Comprometida con la **innovación, la sostenibilidad y la cercanía con el cliente**, EROSKI orienta sus estrategias hacia la **mejora continua de la experiencia de compra**, tanto en tiendas físicas como en el entorno digital. En este contexto, la personalización de la oferta y el uso eficiente de los datos de cliente juegan un papel clave para adaptar sus servicios a las necesidades y preferencias de los consumidores.

Este reto cuadra con la apuesta de EROSKI por optimizar su canal de comercio electrónico mediante el desarrollo de sistemas de recomendación avanzados, que permitan ofrecer propuestas relevantes a cada cliente y reforzar su vínculo con la marca.

2.2 Sobre la problemática

Hoy en día, personalizar la experiencia de compra online se ha vuelto una prioridad para muchas empresas, especialmente en sectores tan competitivos como el de la distribución. Aunque los clientes tienen acceso a una gran variedad de productos, esto también puede generar cierta saturación o dificultad para encontrar lo que realmente necesitan. Aquí es donde los sistemas de recomendación juegan un papel fundamental, ya que ayudan a mostrar **opciones más ajustadas a los gustos e intereses de cada persona**.

En el caso de EROSKI, que cuenta con una gran cantidad de datos sobre el comportamiento de sus clientes, el reto está en cómo utilizar toda esa información de forma efectiva. **No se trata sólo de analizar lo que se ha comprado antes, sino de entender patrones, detectar necesidades no explícitas y hacer recomendaciones útiles en el momento adecuado.**

La problemática, por tanto, va más allá del simple análisis de datos: **implica transformar esa información en acciones concretas que mejoren la experiencia del usuario en la plataforma**. Esto requiere un trabajo cuidadoso de procesamiento, modelado y validación de resultados, siempre con el objetivo de aportar valor tanto para el cliente como para la empresa.

2.3 Objetivos

El objetivo principal del Reto 4 es **analizar los datos de transacciones** proporcionados por EROSKI para desarrollar sistemas de recomendación personalizados en su plataforma de comercio electrónico. A partir del comportamiento de compra de los clientes, se busca **generar recomendaciones relevantes** que mejoren su experiencia y fomenten la conversión.

Para ello, en primer lugar, se pretende identificar a los diez clientes para los que la promoción de un producto concreto sería más relevante, sin restricciones sobre compras

previas. Por otra parte, se busca recomendar a cada uno de diez clientes un producto adicional que aún no tengan en el carrito, basándose en patrones de consumo similares. A continuación, se trata de asignar a cada cliente el producto más adecuado dentro de una lista de veinte artículos en oferta. Finalmente, como último objetivo se pretende sugerir un producto posiblemente olvidado, eliminando de la cesta, pero relevante según el historial de compra.

La implementación de estos sistemas requerirá un análisis y procesamiento detallado de los datos, con el fin de construir una matriz que permita generar recomendaciones precisas y ajustadas a cada contexto planteado.

3 Recoger y almacenar los datos

3.1 Fuentes de datos utilizadas

Para llevar a cabo este análisis, se utilizaron datos extraídos de tres archivos en formato RDS, los cuales fueron proporcionados por Eroski, la empresa para la que se desarrolla este proyecto. Estos archivos constituyen la fuente principal de información y sirven como base para todo el estudio.

Dado que los datos provienen de una **fuente confiable**, se parte de que la información es **precisa y representativa** de la realidad de compra de los clientes analizados. No obstante, como parte del proceso de limpieza y preparación, se llevaron a cabo diversas verificaciones para asegurar la coherencia y calidad de los datos antes de su procesamiento.

Gracias a esta base de datos estructurada, ha sido posible realizar un análisis detallado, aplicando metodologías que permiten extraer información relevante y generar recomendaciones de compra personalizadas con un alto grado de precisión.

3.2 Procesamiento de los datos

En esta etapa se han preparado los datos de **interacción entre clientes y productos** con el objetivo de construir una matriz estructurada que sirva como entrada para modelos de recomendación. Para ello, se parte de la información de **tickets de compra**, **que se enriquece con atributos del maestro de productos**. Posteriormente, se genera una matriz donde se contabiliza cuántas veces cada cliente ha adquirido cada producto.

Esto significa que **se crea una matriz de interacción cliente-producto**, donde cada celda representa la cantidad de veces que un cliente compró un determinado artículo.

Para asegurar que la matriz final sea útil para los modelos posteriores, se aplica una **reducción progresiva de dimensionalidad**. Se eliminan productos con muy pocas compras y clientes con poca actividad, lo que ayuda a **reducir la dispersión de la matriz**. Este filtrado se realiza considerando tanto el número total de compras como la media de interacciones por producto y cliente. Sin embargo, **se garantiza la retención de un conjunto de productos y clientes considerados estratégicos porque forman parte de los objetivos del proyecto y deben permanecer en el análisis**.

Aunque en esta fase no se realiza un tratamiento específico de outliers, esta reducción mediante filtros extremos en la actividad actúa como un **control natural**, eliminando tanto casos muy inactivos como potencialmente excesivos. Por tanto, se considera que la matriz final queda depurada sin necesidad de un paso adicional de detección de valores atípicos.

4 Analizar y modelar los datos

4.1 Análisis descriptivos

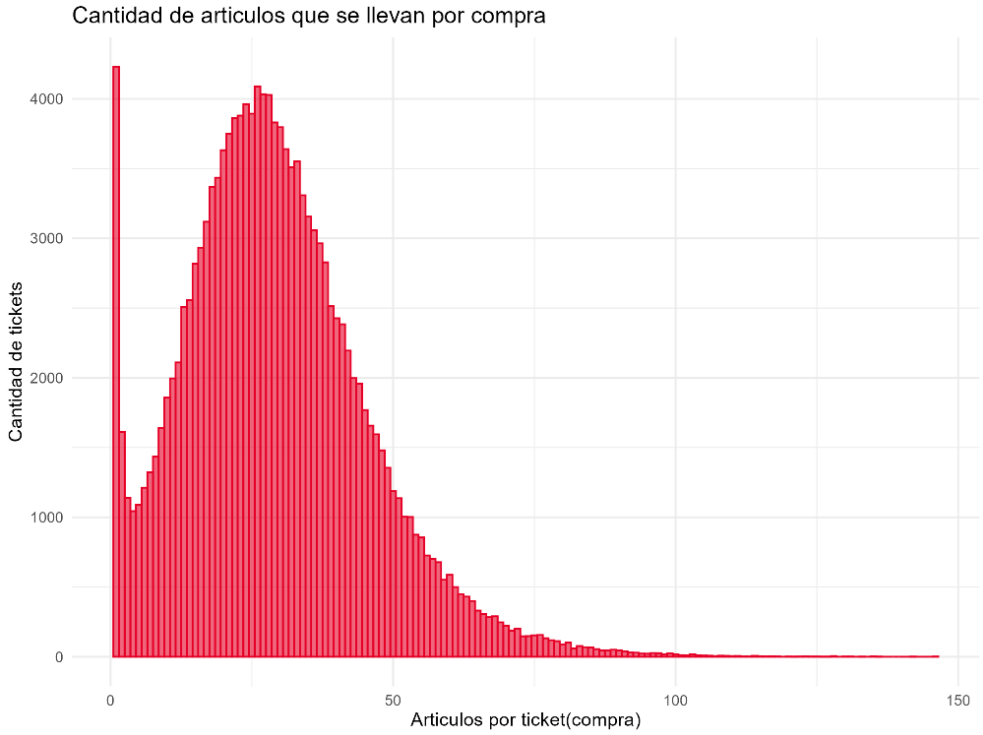

Figura 1: Cantidad de artículos que se llevan por compra. Fuente: Propia.

El siguiente histograma muestra la cantidad de artículos que los clientes suelen comprar por ticket (equivalente a una compra). Como se observa, la distribución es asimétrica hacia la derecha, con la mayoría de las compras concentradas entre los 20 y 50 artículos. Existe también un **pico significativo en compras con un solo artículo**, lo que puede asociarse a **compras rápidas o de urgencia**. A medida que aumenta la cantidad de artículos por ticket, la frecuencia disminuye progresivamente, aunque se registran casos de compras muy grandes, con más de 100 artículos. Estas compras, aunque son menos frecuentes, podrían corresponder a clientes profesionales, a compras familiares de gran volumen o a abastecimientos para eventos puntuales.

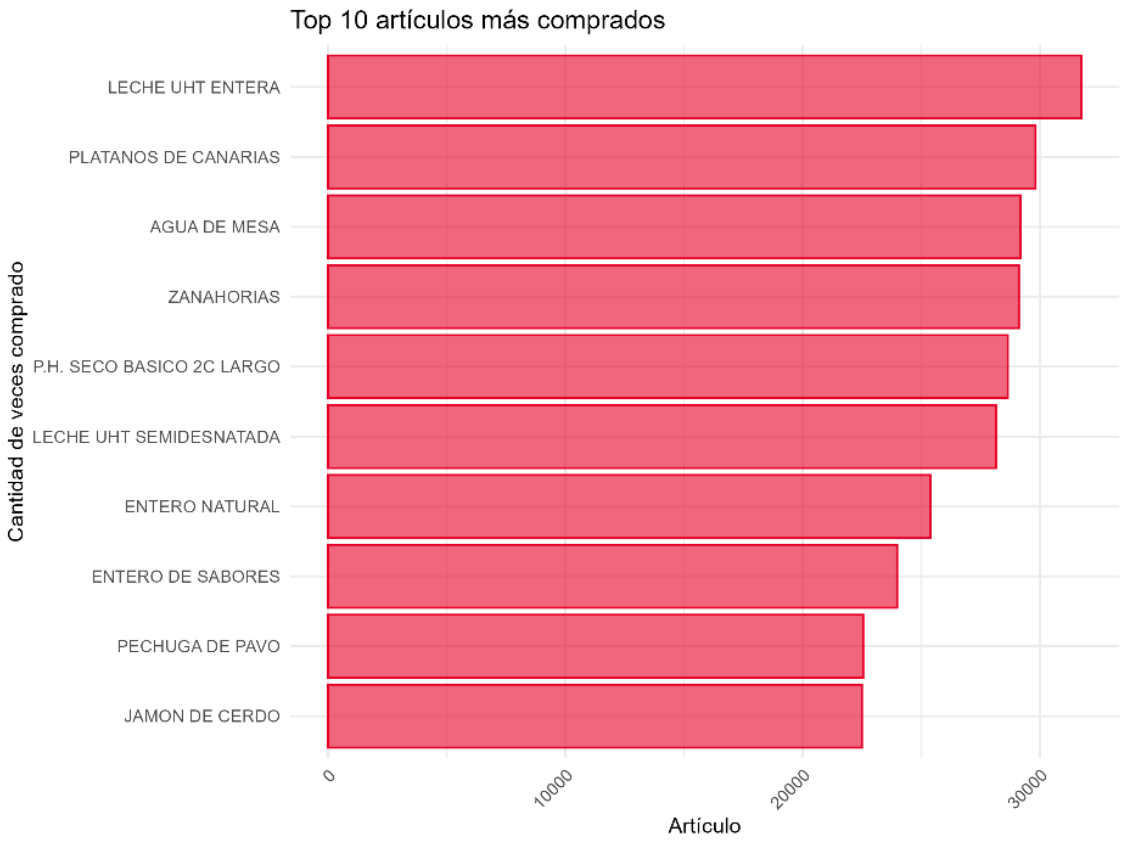

Figura 2: Artículos más comprados. Fuente: Propia.

El gráfico muestra el Top 10 de artículos más comprados según la cantidad de veces que cada uno fue adquirido. En el eje horizontal se indican los nombres o descripciones de los artículos, mientras que el eje vertical representa la cantidad de veces que estos artículos fueron comprados, a lo largo del período analizado.

Se observa que el artículo más comprado es "LECHE UHT ENTERA", con más de 31.000 registros de compra, seguido de productos como "PLÁTANOS DE CANARIAS", "AGUA DE MESA" y "ZANAHORIAS", todos con cifras cercanas a las 30.000 unidades compradas. Estos productos reflejan una preferencia por artículos de consumo cotidiano, tanto frescos como de larga duración.

El gráfico también incluye productos como "JAMÓN DE CERDO", "PECHUGA DE PAVO" y "ENTERO NATURAL", que, aunque tienen un volumen de compra menor, superan las 22.000 adquisiciones. En conjunto, este análisis permite identificar los **productos de mayor rotación** en los establecimientos de Eroski, información **clave para la planificación de inventarios, promociones y estrategias de fidelización**.

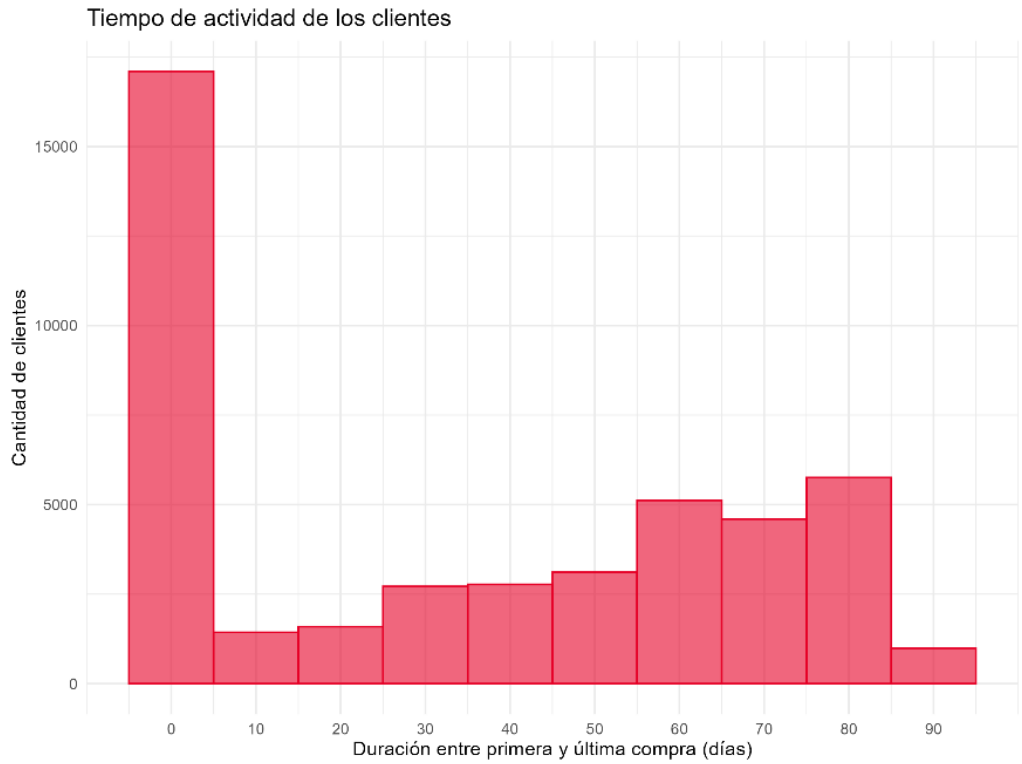

Figura 3: Tiempo de actividad de los clientes. Fuente: Propia.

El gráfico representa la duración de la actividad de los clientes, es decir, el tiempo en días entre su primera y última compra. La mayor parte de los clientes presenta una actividad muy corta, concentrada en los primeros diez días, lo que sugiere que muchos realizan solo una compra o abandonan rápidamente su relación con la empresa.

A partir de los diez días la cantidad de clientes disminuye, pero luego se estabiliza y muestra un ligero repunte entre los sesenta y noventa días. Esto indica la presencia de un **grupo de clientes más comprometido**, que mantiene una relación más prolongada con la marca y realiza compras de forma sostenida en el tiempo.

Este comportamiento sugiere la existencia de diferentes perfiles de clientes. Por un lado, clientes ocasionales o de una sola compra y por otro lado, **clientes recurrentes que podrían ser el objetivo de estrategias de fidelización o personalización para reforzar su vínculo con la marca**.

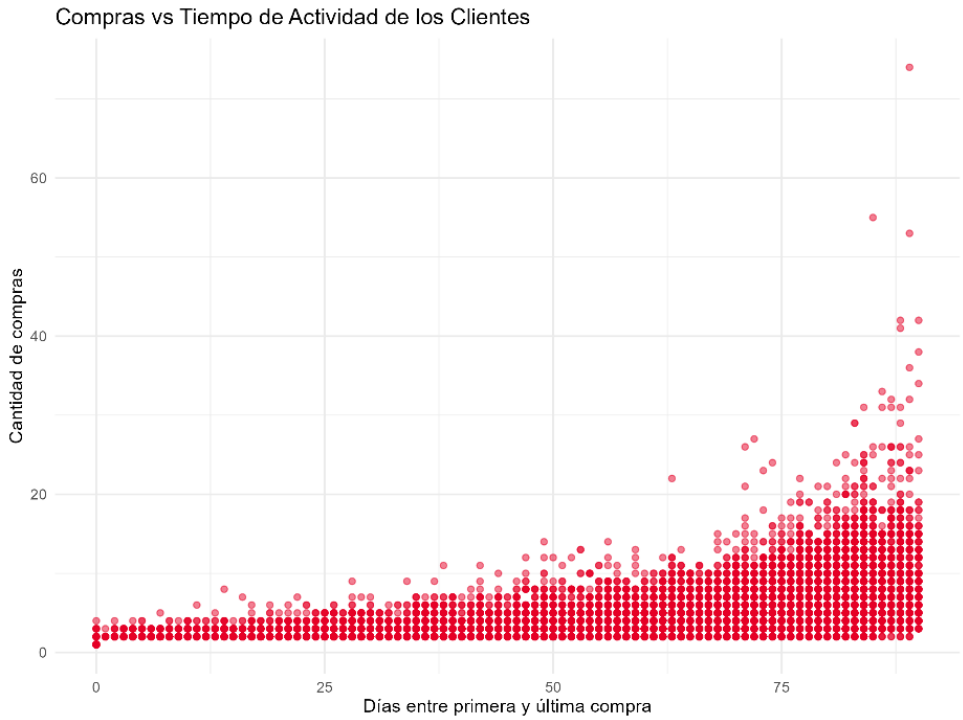

Figura 4: Fidelización de los clientes. Fuente: Propia.

El gráfico muestra la relación entre la duración de la actividad de los clientes, medida en días entre su primera y última compra, y la cantidad total de compras realizadas en ese período. Se observa que los clientes con un tiempo de actividad corto tienden a realizar pocas compras, en su mayoría entre una y cinco, mientras que a medida que aumenta la duración de actividad también crece la cantidad de compras.

A partir de los cincuenta días se empieza a notar una mayor dispersión, con algunos clientes que alcanzan volúmenes de compra considerablemente altos. Esto sugiere que **los clientes más activos en el tiempo tienden a ser también los más frecuentes en sus compras**, y por tanto pueden representar un grupo valioso para campañas de fidelización.

Este gráfico mide la fidelización de los clientes, ya que no solo refleja el tiempo durante el cual se mantienen activos, sino también la frecuencia con la que compran en ese periodo. **Cuanto mayor es la duración y la cantidad de compras, mayor es el compromiso del cliente con la marca**, lo que lo convierte en un **buen indicador** para identificar patrones de **lealtad** y diseñar estrategias orientadas a mantenerlos a largo plazo.

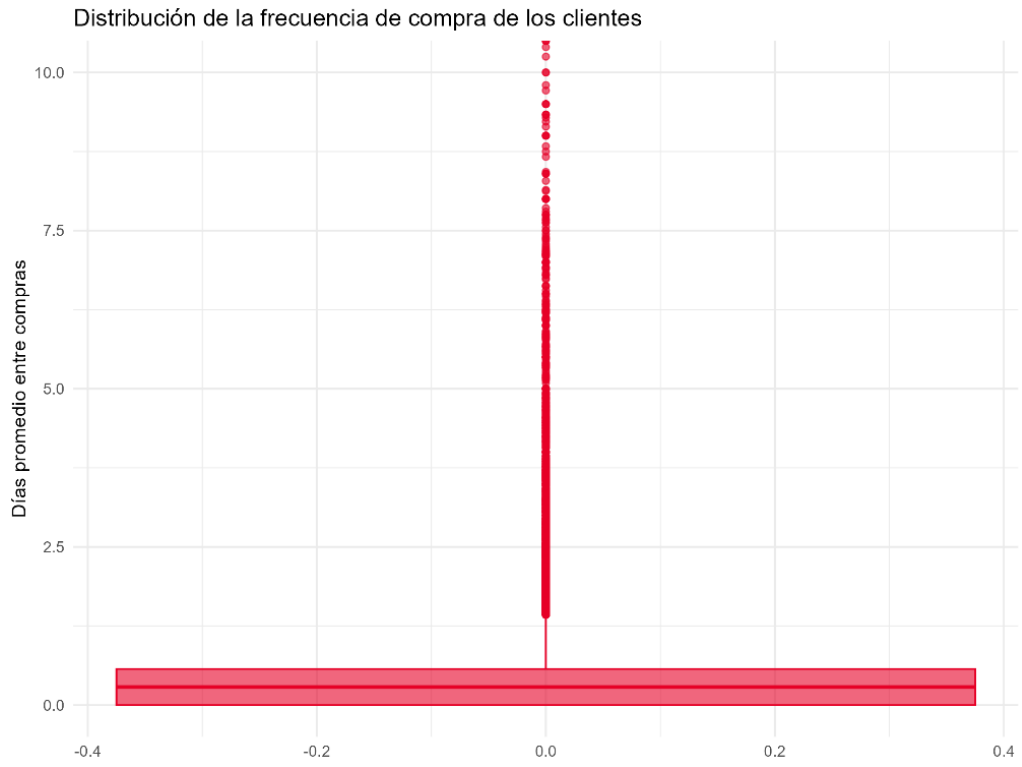

Figura 5: Distribución de la frecuencia de compra de clientes. Fuente: Propia.

Este gráfico de tipo boxplot muestra la distribución de la frecuencia de compra de los clientes, representada como el número promedio de días que pasan entre una compra y otra. La gran mayoría de los valores se concentra cerca del cero, lo que indica que muchos clientes realizan sus compras con muy poca separación en el tiempo.

El gráfico también presenta una gran cantidad de valores atípicos hacia arriba, lo que sugiere que hay clientes que tardan más días entre compras, aunque estos casos son menos frecuentes. Esta distribución altamente sesgada refuerza la idea de que **existen comportamientos de compra muy distintos entre los clientes**.

Este gráfico permite analizar la regularidad con la que los clientes interactúan con la empresa. Dado que muestra la frecuencia promedio entre compras, puede considerarse un **indicador útil para medir el grado de fidelización**. Los clientes que compran de forma más seguida, con menos días entre cada compra, tienden a estar más comprometidos con la marca, lo que los convierte en buenos candidatos para acciones de retención y personalización.

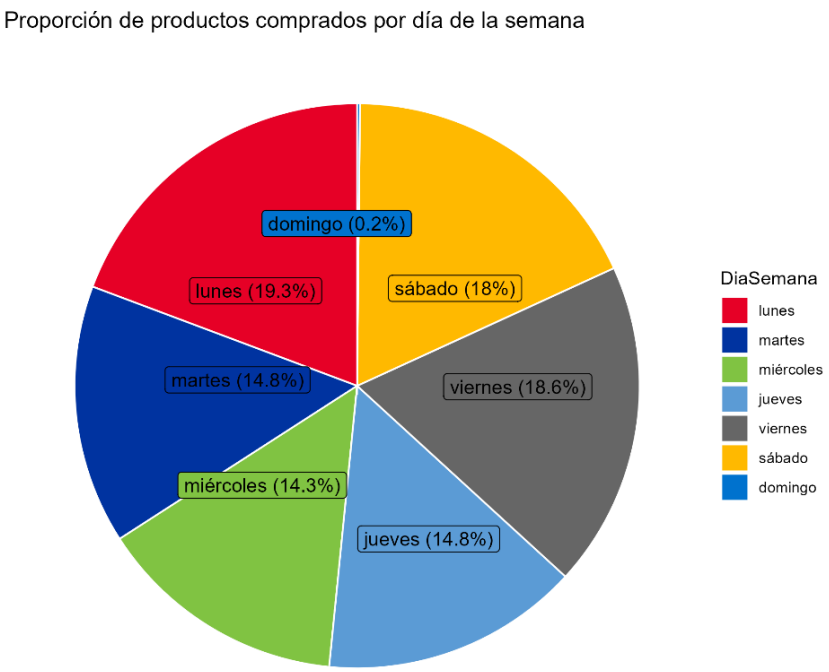

Figura 6: Proporción de productos por día de la semana. Fuente: Propia.

El gráfico de pastel muestra la proporción de productos comprados por día de la semana. Destacan **lunes (19.3%)**, **viernes (18.6%)** y **sábado (18%)** como los días con mayor volumen de compras, lo que puede estar relacionado con distintos hábitos de consumo.

Los lunes destacan como el día con mayor proporción de compras. Esto puede deberse a que muchas personas realizan su compra semanal al inicio de la semana, planificando sus comidas y necesidades del hogar tras el fin de semana.

Los viernes y sábados, concentran una alta actividad de compra asociada a la preparación para el fin de semana. Durante estos días, los clientes suelen adquirir productos para reuniones sociales, comidas familiares o descanso en el hogar. Además, muchas personas reciben su salario a final de semana, lo que incrementa su capacidad y disposición de gasto.

En cambio, **el domingo (0.2%)** presenta mínima actividad, posiblemente por el cierre de tiendas, lo que adelanta las compras al sábado.

Los días intermedios (**martes, miércoles y jueves**) muestran proporciones más equilibradas (entre 14.3% y 14.8%), reflejando compras de reposición o de menor volumen, sin un objetivo específico de planificación semanal.

Este análisis permite a Eroski **identificar patrones de comportamiento del consumidor** y puede ser clave para optimizar la logística, ajustar horarios de apertura, lanzar promociones estratégicas o reforzar el personal en días de mayor afluencia.

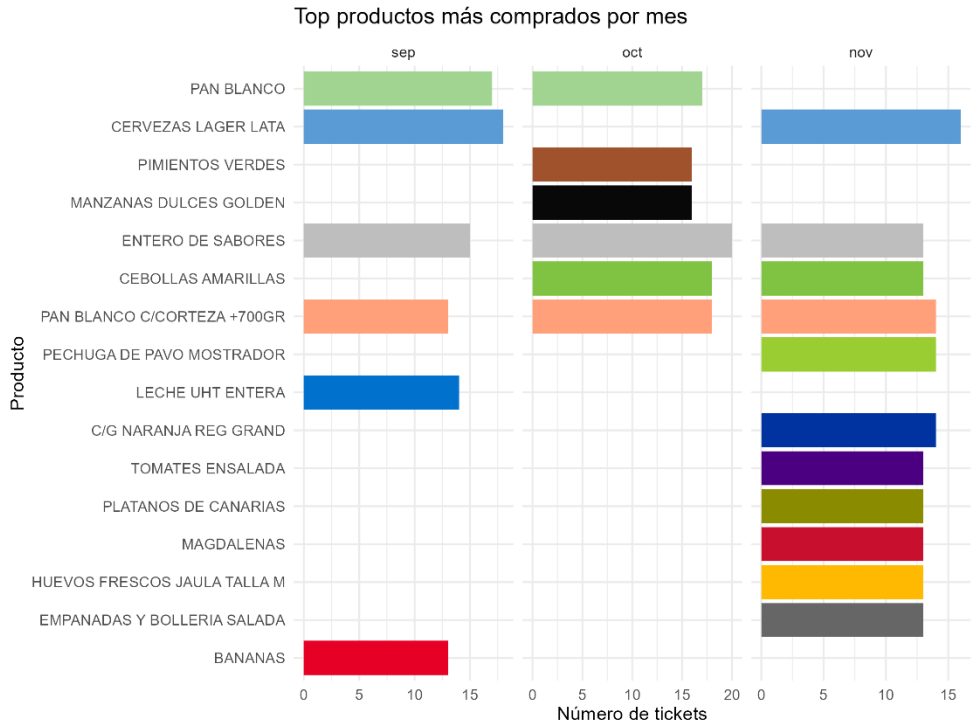

Figura 7: Productos más comprados por mes. Fuente: Propia.

Este gráfico de barras muestra los **productos más comprados por mes**, agrupando los resultados de septiembre, octubre y noviembre según el número de tickets en los que aparece cada producto. Se observan patrones de compra repetidos a lo largo del trimestre analizado, destacando artículos de **consumo básico y frecuente**.

Productos como el **pan blanco**, **cervezas en lata**, **cebollas amarillas** y **plátanos de Canarias** se mantienen entre los más vendidos en varios meses, lo que refleja su alta rotación y preferencia constante entre los consumidores. También destacan productos como **leche entera**, **huevos frescos**, **empanadas y bollería salada**, típicos de un consumo habitual y diario.

Noviembre muestra mayor diversidad de productos con niveles similares de compra, lo que podría indicar una **mayor dispersión del consumo** o el inicio de compras más variadas con vistas a la temporada prenavideña.

Este análisis permite a Eroski **identificar los productos clave para mantener en stock de forma prioritaria**, anticiparse a la demanda estacional y orientar promociones o campañas sobre artículos de alta rotación y preferencia estable.

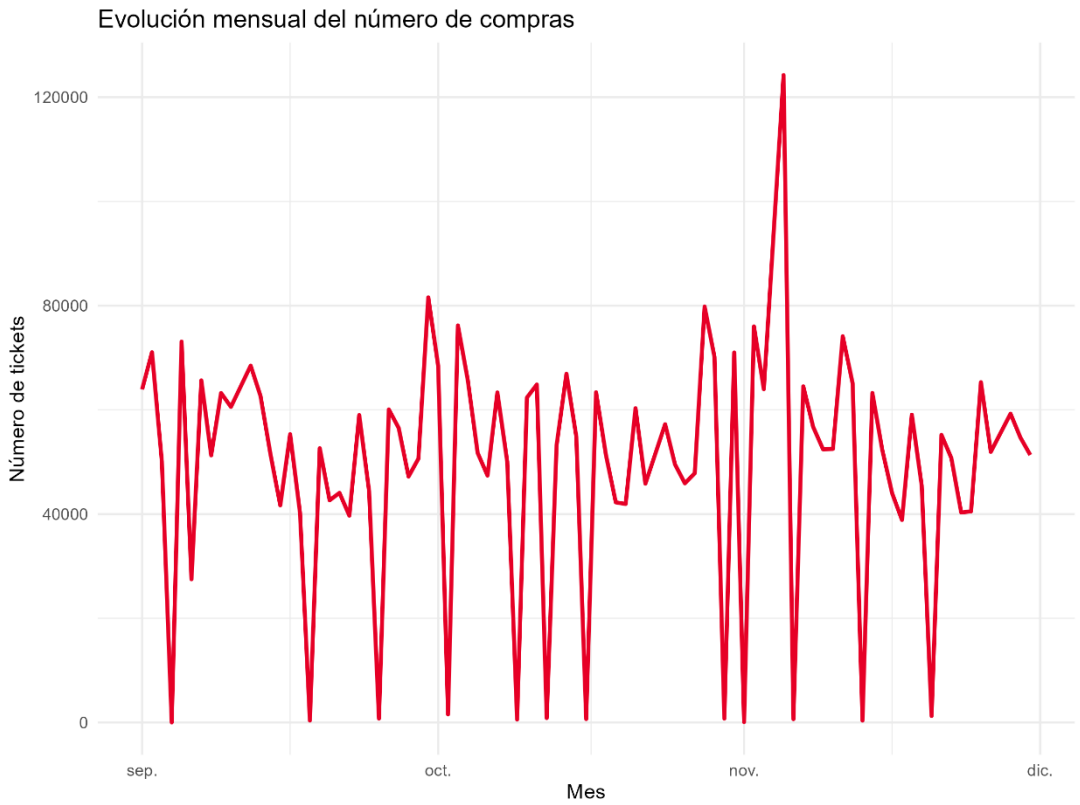

Figura 8: Evolución mensual del número de compras. Fuente: Propia.

Este gráfico de líneas muestra la **evolución mensual del número de compras** (tickets generados) entre los meses de **septiembre y noviembre**. Se observa una tendencia irregular, con picos y caídas marcadas que parecen corresponder a días específicos de mayor o menor actividad.

En general, **septiembre y octubre** presentan una evolución relativamente estable, con fluctuaciones frecuentes, posiblemente asociadas al comportamiento semanal del consumidor (por ejemplo, más compras los lunes o fines de semana y menos en domingos o festivos).

En **noviembre** se destaca un **pico muy pronunciado**, con el número de tickets superando los 120.000. Este aumento podría estar relacionado con promociones especiales, eventos como el **Black Friday**, o anticipación de compras de cara al periodo navideño. Tras este pico, el nivel de compras vuelve a estabilizarse.

Este análisis permite a Eroski **identificar momentos de alta demanda** para optimizar la operativa comercial, preparar inventario, reforzar personal y lanzar campañas en los días con mayor impacto en ventas.

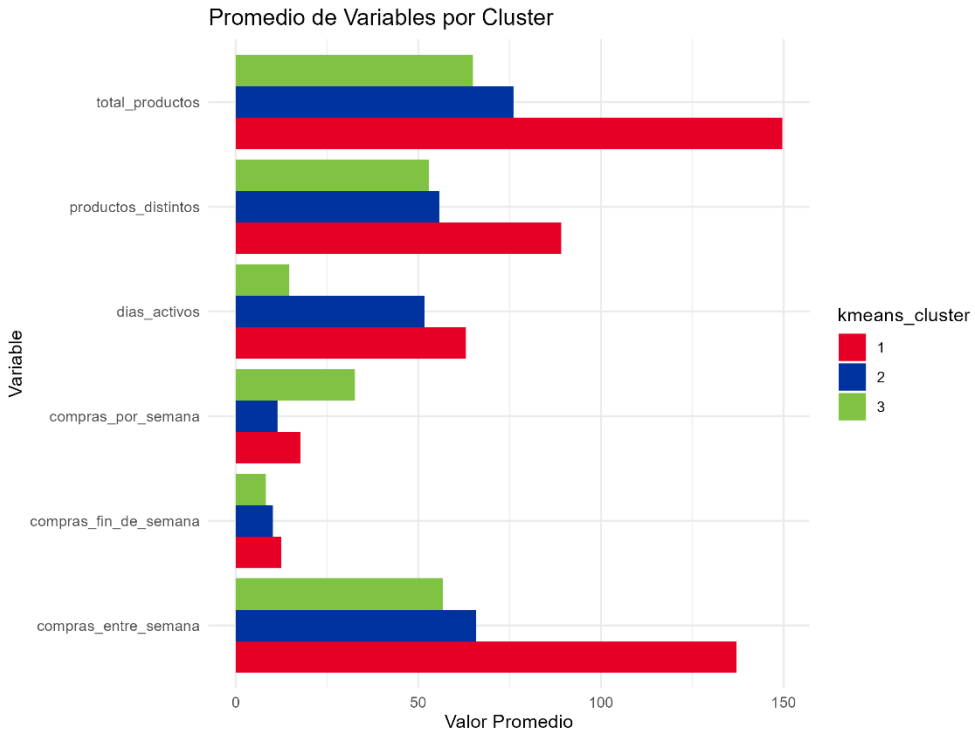

Figura 9: Promedio de Variables por Cluster. Fuente: Propia.

Posteriormente, se muestra un gráfico con el promedio de varias variables para cada clúster. Este tipo de visualización permite **identificar las principales diferencias entre los grupos de clientes** a nivel general.

El **Clúster 1** destaca claramente en todas las variables, especialmente en total_productos, productos_distintos y compras_entre_semana, lo que indica un **perfil de cliente muy activo y con un alto nivel de consumo**. El **Clúster 2** ocupa una **posición intermedia en la mayoría de las variables**, mientras que el **Clúster 3** muestra los valores promedio más bajos, especialmente en días_activos y compras_por_semana, lo que sugiere **un grupo con menor frecuencia y volumen de compra**. Esta comparación permite entender el nivel de actividad y diversidad de consumo de cada clúster, información clave para diseñar estrategias diferenciadas según el perfil del cliente.

4.2 Modelar los datos y Visualizar los resultados obtenidos

COMPARACION DE MODELOS:

Como **etapa fundamental del modelado** se realizó una **evaluación de diversos algoritmos de recomendación** con el objetivo de determinar cuál presentaba el mejor desempeño sobre la base de comportamiento de los clientes. Esta revisión comparativa fue esencial para seleccionar un modelo robusto y eficiente que permitiera identificar con precisión los productos más adecuados para recomendar a cada cliente.

Para ello se utilizó la librería “recommenderlab” en R, que permitió transformar la matriz de datos original en un formato adecuado para la evaluación de sistemas de recomendación. Se aplicó un **esquema de validación mediante particionado simple**, donde el 80 % de los datos se destinó al entrenamiento y el 20 % restante se utilizó para

la prueba, considerando un umbral que definía cuándo una interacción se consideraba positiva. Se evaluaron cinco algoritmos representativos: popularidad, aleatorio, filtrado colaborativo basado en usuario (UBCF) e ítem (IBCF), y un modelo basado en factorización mediante “Singular Value Decomposition” (SVDF). Para cada uno se realizaron predicciones tanto en términos de estimación de ratings como en la generación de listas top-N, utilizando métricas específicas para cada tipo de evaluación.

Tabla 1: Métricas de error para Ratings. Fuente: Propia.

Modelo	RMSE	MAE	MSE
POPULAR	1.516838	1.006700	2.300797
RANDOM	26.346051	22.595204	694.114390
UBCF	1.763597	1.217882	3.110275
IBCF	2.286617	1.374139	5.228618
SVDF	1.572607	1.010396	2.473093

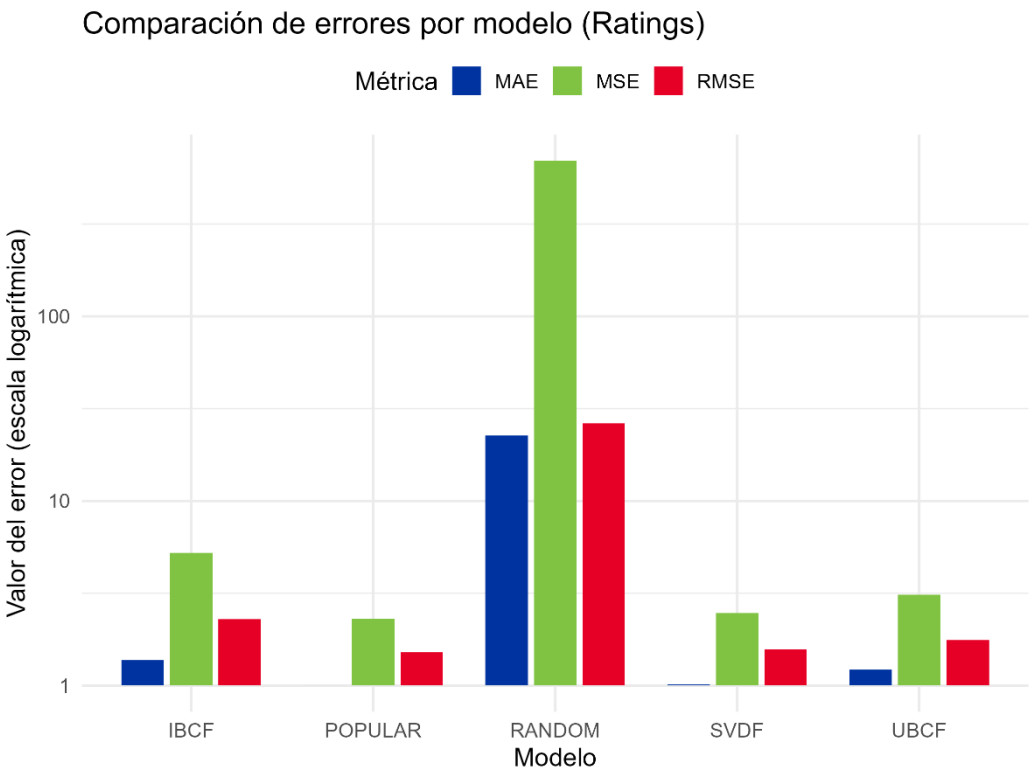

Figura 10: Comparación de errores por modelo en Ratings. Fuente: Propia.

Los resultados mostraron que, en cuanto a predicción de ratings, **el modelo basado en popularidad presentó el menor error según métricas como RMSE, MAE y MSE**, seguido de cerca por el modelo SVDF, lo que indica una **mejor capacidad para estimar**

con precisión las valoraciones implícitas de los usuarios. Por otro lado, los modelos UBCF e IBCF mostraron errores más elevados, reflejando una menor precisión en el contexto de estos datos dispersos y mayoritariamente implícitos, mientras que el modelo aleatorio sirvió como referencia inferior con errores muy altos.

Tabla 2: Metricas para TopNList. Fuente: Propia.

Modelo	Precision	Recall	TPR	FPR
POPULAR	0.18713252	0.080536548	0.080536548	0.005647215
RANDOM	0.01856206	0.006688243	0.006688243	0.006828117
UBCF	0.02367707	0.008622072	0.008622072	0.006790519
IBCF	0.13208926	0.037178347	0.037178347	0.004786598
SVDF	0.06508367	0.029148492	0.029148492	0.006501559

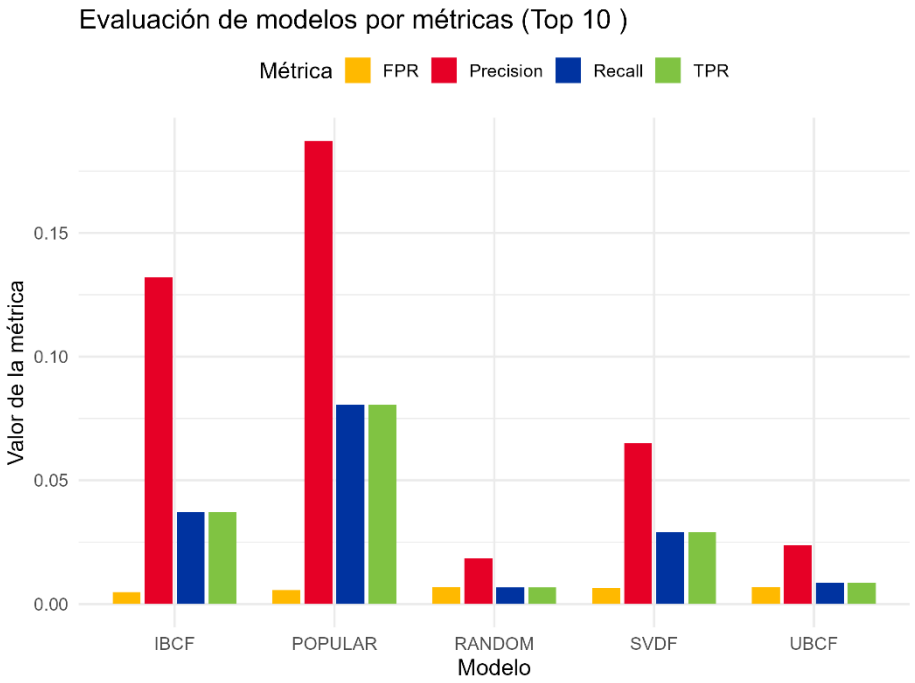

Figura 11: Comparación gráfica de métricas top-N-list por modelo. Fuente: Propia.

En la **evaluación top-N**, que mide la capacidad de los modelos para recomendar ítems relevantes dentro de un conjunto limitado, el algoritmo de popularidad nuevamente destacó, obteniendo los **valores más altos en precisión y recall**, seguido del IBCF, aunque con una menor capacidad de recuperación. El modelo SVDF, a pesar de su buen desempeño en ratings, presentó métricas top-N más moderadas, lo que sugiere que su precisión en la predicción no se traduce siempre en las posiciones más

relevantes dentro del ranking de recomendaciones. Los modelos UBCF y aleatorio mostraron los resultados menos favorables en este aspecto.

En conjunto, estas métricas permitieron identificar que, aunque el modelo popular mostró un desempeño consistente en ambas evaluaciones y una buena capacidad práctica para recomendaciones inmediatas, **los modelos basados en factorización como SVDF ofrecen un mejor balance entre precisión y capacidad para capturar patrones latentes en datos implícitos**. Esta conclusión guió la elección final hacia un modelo de factorización regularizada con pesos (WRMF), que, aunque no se incluyó en esta fase comparativa inicial, es conocido por su robustez en contextos similares. Esta selección responde a la necesidad de un sistema que no solo estime valoraciones con precisión, sino que también optimice la relevancia y personalización en la recomendación de productos para los clientes de Eroski.

OBJETIVO 1:

Para alcanzar el primer objetivo del proyecto se implementó un sistema de recomendación orientado a **identificar clientes con alta afinidad hacia un producto específico que aún no han adquirido**. Se utilizó un enfoque basado en **filtrado colaborativo**, aplicando un **modelo de retroalimentación** implícita conocido como **WRMF (Weighted Regularized Matrix Factorization)**, el cual permite estimar el nivel de interés de los clientes en función de patrones previos de comportamiento de compra.

A partir de una matriz de interacciones entre clientes y productos, el modelo fue entrenado para detectar afinidades latentes. Luego de realizar las predicciones, se filtraron aquellos clientes cuyo nivel de afinidad superaba un umbral establecido, y se eliminaron del conjunto final aquellos que ya habían comprado el producto.

Como resultado de este proceso, se obtuvo un conjunto de identificadores de clientes priorizados según su probabilidad de compra.

(Para consultar los resultados del objetivo ir a ANEXO 2).

OBJETIVO 2:

En el este objetivo se planteó como meta recomendar un único producto adicional a un conjunto de 10 clientes específicos, con el propósito de maximizar la probabilidad de que lo añadan a su próximo carrito de compra. Para lograr esto, se ha implementado un sistema de recomendación basado en un **modelo de filtrado colaborativo implícito**.

El proceso comienza con la preparación de los datos de compra, donde se construye una matriz cliente-producto que refleja las interacciones entre usuarios y artículos. En este caso, se ha utilizado una **matriz binarizada**, donde cada entrada indica si un cliente compró (1) o no compró (0) un producto. Esta binarización es clave porque el modelo WRMF (Weighted Regularized Matrix Factorization) está diseñado para trabajar con datos implícitos, es decir, información basada en señales indirectas de preferencia, como la presencia o ausencia de compra, en lugar de valoraciones explícitas o puntuaciones.

Una vez preparada la matriz binarizada, se entrena el **modelo ALS** para aprender representaciones latentes tanto de clientes como de productos. Posteriormente, se generan predicciones para los clientes objetivo, **identificando el producto con mayor afinidad no adquirido aún**. Finalmente, se obtienen los scores que justifican estas recomendaciones, permitiendo explicar la relevancia de cada producto sugerido.

(Para consultar los resultados del objetivo ir a ANEXO 3).

OBJETIVO 3:

En el objetivo tres se buscó recomendar un único producto a cada cliente de un grupo determinado, eligiendo solo entre una lista cerrada de productos en oferta. La intención era que estas recomendaciones fueran útiles y directamente **aplicables dentro de una campaña comercial activa**.

Para lograrlo, se partió de una matriz que recoge la relación entre clientes y productos, donde cada número indica cuántas veces un cliente ha comprado un producto. Los valores vacíos, que representan productos que nunca fueron comprados por un cliente, se reemplazaron por ceros. También se eliminaron valores muy altos para evitar que influyeran demasiado en el modelo. Después, la matriz se transformó a un formato más eficiente para trabajar con muchos datos a la vez.

Con la matriz lista, se entrenó un modelo de recomendación usando el **algoritmo ALS**, que permite encontrar patrones de comportamiento entre clientes y productos, incluso si no han comprado los mismos artículos. Este modelo asigna una puntuación a cada producto para cada cliente, y con esa información **se seleccionó el producto con mayor puntuación como la mejor opción para recomendar**.

Como solo interesaban los productos en oferta, se aplicó un filtro para asegurarse de que las recomendaciones coincidieran con esa lista. Cuando no se pudo recomendar directamente un producto en promoción, se aplicó una **estrategia de respaldo basada en selección aleatoria ponderada por popularidad**. Este enfoque permitió diversificar las recomendaciones y evitar que todos los casos sin predicción válida apuntaran siempre al producto más popular.

Finalmente, se preparó una tabla con las recomendaciones para cada cliente, incluyendo el nombre del producto recomendado. También se creó un resumen que muestra cuántas veces fue recomendado cada producto, lo que ayuda a entender si el reparto de recomendaciones está equilibrado y si hay productos que destacan especialmente.

(Para consultar los resultados del objetivo ir a ANEXO 4).

OBJETIVO 4:

En este proyecto se desarrolla un sistema de recomendación personalizado dirigido específicamente a un conjunto de clientes predefinidos, implementando técnicas avanzadas de **filtrado colaborativo mediante factorización matricial (WRMF)**. El sistema se fundamenta en un enfoque de retroalimentación que aprende de los patrones históricos de comportamiento de compra, distinguiéndose por su capacidad única de

generar sugerencias inteligentes que evitan recomendar productos ya comprados en la última compra de los clientes.

El modelo procesa una matriz de interacciones cliente-producto donde cada elemento representa exclusivamente la presencia o ausencia de una transacción, eliminando información de frecuencia para enfocarse en patrones de preferencia puros. Esta matriz se utiliza para entrenar un modelo WRMF que identifica características latentes del comportamiento de compra, permitiendo al sistema comprender las afinidades subyacentes entre clientes y productos sin depender de valoraciones explícitas.

Una característica distintiva fundamental del sistema es la implementación de un mecanismo de exclusión inteligente que analiza automáticamente las compras más recientes de cada cliente objetivo. El sistema identifica el último ticket de compra de cada cliente y marca todos los productos adquiridos en esa transacción como no recomendables, previniendo así sugerencias redundantes y optimizando la experiencia del usuario. Esta funcionalidad es particularmente valiosa en contextos comerciales donde la repetición de compras recientes puede generar insatisfacción o pérdida de confianza en el sistema.

El proceso de recomendación se configura para generar una única sugerencia por cliente, priorizando la calidad y relevancia sobre la cantidad. Esta decisión permite **concentrar los esfuerzos comerciales en las oportunidades más prometedoras, maximizando la probabilidad de conversión**. Los resultados finales se enriquecen con información descriptiva completa de los productos, proporcionando un output comprensible y accionable que facilita la implementación de estrategias comerciales dirigidas.

Este enfoque no solo garantiza recomendaciones personalizadas basadas en comportamientos reales de compra, sino que también **optimiza la efectividad comercial al dirigir los recursos hacia clientes específicos con productos verdaderamente relevantes**, evitando al mismo tiempo la fatiga del cliente por recomendaciones inapropiadas.

(Para consultar los resultados del objetivo ir a ANEXO 5).

4.3 Conclusiones

Los **sistemas de recomendación** representan una **herramienta clave** para mejorar la precisión y efectividad de las **campañas de marketing**. Al permitir identificar de manera anticipada a los clientes con mayor probabilidad de interés por determinados productos, facilitan una segmentación mucho más precisa y orientada a la acción, en contraste con enfoques masivos o genéricos.

Gracias a estas recomendaciones, es posible dirigir los esfuerzos comerciales hacia aquellos clientes que tienen mayor potencial de conversión, lo que contribuye directamente a un mejor uso de los recursos, una mayor tasa de respuesta y una

experiencia más personalizada para el cliente. Esto se traduce en comunicaciones más relevantes, campañas más efectivas y un incremento potencial en las ventas.

En conjunto, los sistemas de recomendación no solo aportan **eficiencia**, sino que también ayudan a **construir relaciones comerciales más inteligentes**, **oportunas** y **alineadas con las necesidades reales de los consumidores**.

5 Transformar los negocios

En el desarrollo de este reto se ha optado por utilizar el **email marketing** como estrategia principal de comunicación digital con los clientes. Esta herramienta ha sido seleccionada por su capacidad de ofrecer una comunicación directa, personalizada y eficaz. A través de una campaña simulada, se han promocionado productos destacados, ofertas especiales y ventajas del programa de fidelización, en línea con las estrategias comerciales y de comunicación empleadas por EROSKI.

Uno de los elementos clave de esta estrategia ha sido la **personalización del contenido**, con recomendaciones realizadas en función del perfil y comportamiento del cliente. Esta adaptación del mensaje, basada en sus intereses y hábitos de compra, permite aumentar significativamente la relevancia del contenido y mejorar la experiencia del usuario.

El email marketing no solo es útil para informar, sino también para **fidelizar a los clientes**. A través de envíos periódicos con promociones exclusivas, sugerencias útiles y contenido de valor, se refuerza el vínculo entre el consumidor y la marca. Esta relación se fortalece aún más al ofrecer ventajas exclusivas como descuentos para suscriptores, recordatorios personalizados o mensajes en fechas señaladas.

Además, esta estrategia permite un **control detallado** de los resultados (aperturas, clics, tasas de conversión), lo que facilita su evaluación y mejora continua. En definitiva, el email marketing es una herramienta potente y económica para mantener al cliente activo, satisfecho y comprometido a largo plazo.

(Para consultar sobre la herramienta utilizada para la creación del diseño del email ir al Anexo 7.)

6 Implicaciones legales y éticas

Los apartados éticos son una parte fundamental del informe, siendo aún más relevantes cuando el proyecto se trata de recomendadores enfocados a los clientes. Estos recomendadores son herramientas muy útiles a la hora de analizar y recomendar productos a los clientes según el segmento en el que están. Sin embargo, estos sistemas tienen un impacto directo sobre la experiencia del usuario, influyen en decisiones de compra y, en muchos casos, manejan información sensible.

De esta manera, este apartado tratará los posibles problemas éticos que podemos encontrarnos a la hora de implementar los recomendadores, las maneras de evitarlos, las soluciones y diferentes conceptos éticos relacionados a la ética de Eroski.

El primer tema que nos gustaría mencionar son los **sesgos algorítmicos**. Estos sesgos discriminatorios pueden no ser introducidos de manera intencionada. Sin embargo, si entrenamos el algoritmo con datos históricos en los que, durante ciertas épocas, se promocionaban más unos productos que otros, es posible que ese patrón provoque que el recomendador ni siquiera sugiera otras marcas, aunque estas sean mejores o generen un mayor margen de beneficio para Eroski.

Esto no solo afectaría a las marcas que venden a través de Eroski, sino que también impactaría en los propios usuarios del supermercado, quienes recibirían una experiencia más limitada al obtener recomendaciones repetitivas o, en algunos casos, ninguna.

Para evitar esto, es fundamental revisar cuidadosamente el proceso de entrenamiento del modelo e incorporar herramientas que permitan medir y controlar los sesgos presentes en los datos.

En segundo lugar, consideramos que tan importante como la ética en el entrenamiento de los modelos es la **recogida y tratamiento de los datos**, un aspecto que muchas veces se pasa por alto. Este proceso es especialmente delicado, ya que los datos personales como el nombre, teléfono, correo electrónico o DNI son altamente sensibles.

Si no se anonimizan adecuadamente o no se gestionan de forma segura, existe el riesgo de filtraciones o usos indebidos, lo cual no solo va en contra de la legalidad (por ejemplo, el Reglamento General de Protección de Datos), sino que también supone una falta ética grave.

Para evitar este tipo de riesgos, es fundamental implementar medidas de anonimización, cifrado y control de acceso, así como asegurarse de que solo se recojan los datos estrictamente necesarios para el funcionamiento del sistema.

El tercer aspecto que debemos tratar es el de la **compra abusiva y el ahorro**. Nuestro objetivo final es crear un recomendador de productos que fortalezca la relación comercial entre Eroski y sus clientes, a través de sugerencias que combinen calidad, margen para Eroski y equilibrio en los precios.

Sin embargo, si el sistema empuja constantemente promociones, productos en grandes cantidades o compras impulsivas, puede acabar fomentando hábitos poco sostenibles o incluso perjudiciales para el cliente, tanto desde un punto de vista económico como de salud.

Una forma de evitar las compras abusivas es introducir **filtros o reglas** que limiten recomendaciones excesivas o poco responsables, como evitar sugerir grandes

volúmenes de productos si no se ajustan al historial de consumo del usuario, o reducir la frecuencia de productos poco saludables. Además, se puede fomentar la **diversidad en las recomendaciones**, mostrando una gama equilibrada de productos que incluya marcas locales, opciones sostenibles o alternativas con mejor relación calidad-precio, promoviendo así una experiencia de compra más variada, saludable y responsable.

El cuarto punto a tener en cuenta es la **transparencia en la manera de recomendación**. La empresa debería ser clara y abierta respecto a estos proyectos, ya que los clientes podrían llegar a pensar que Eroski únicamente recomienda productos con el objetivo de maximizar su propio beneficio. Para evitar esa percepción, es importante **comunicar activamente**, a través de canales como redes sociales, "newsletters" u otros medios, cómo funciona el sistema de recomendación.

Se pueden compartir detalles como **qué tipo de productos se recomiendan**, **cómo se ha entrenado el sistema**, **de dónde provienen los datos utilizados** y **cómo han sido tratados para garantizar la privacidad y el cumplimiento legal**. Esta comunicación no solo refuerza la confianza del consumidor, sino que también posiciona a Eroski como una empresa ética, transparente y comprometida con el uso responsable de la tecnología.

A continuación, existe la necesidad de 'feedback', siendo este un aspecto clave para mantener la ética y la eficacia en un sistema de recomendación. Permitir que los usuarios expresen si una recomendación fue útil, adecuada o irrelevante no solo mejora la personalización, sino que también refuerza la idea de que el sistema está diseñado pensando en sus necesidades reales.

Desde un punto de vista ético, este enfoque convierte al usuario en una parte activa del proceso, fomentando la **participación y la mejora continua**. Además, integrar este feedback en el entrenamiento o ajuste del modelo ayuda a reducir sesgos, detectar errores o patrones no deseados, y adaptar las recomendaciones a los cambios en los hábitos de consumo.

Ofrecer canales visibles y accesibles para que el cliente pueda opinar, sea mediante botones, encuestas rápidas o valoraciones, también contribuye a la **transparencia y la confianza**, permitiendo a Eroski demostrar que escucha y valora la voz de sus clientes a lo largo del tiempo.

Figura 12: Gráfico ODS. *Fuente: sustainabledevelopment.report*

El sexto y último aspecto relevante en el desarrollo de sistemas de recomendación es su alineación con los Objetivos de Desarrollo Sostenible (ODS), que ofrecen un marco útil para evaluar el impacto social del proyecto. Este tipo de sistemas pueden llegar a contribuir directamente en varios de estos objetivos si se diseñan de forma apropiada. Por ejemplo, al priorizar productos saludables en las recomendaciones, se apoya el ODS 3 (Salud y bienestar); al fomentar opciones sostenibles, como productos locales o con menor impacto ambiental, se impulsa el ODS 12 (Producción y consumo responsables) y el ODS 13 (Acción por el clima). Además, si se da visibilidad a productos de comercio justo o se promueve la economía local, se refuerzan metas del ODS 8 (Trabajo decente y crecimiento económico). Incorporar criterios relacionados con los ODS no solo amplía la dimensión ética del sistema, sino que también permite conectar el proyecto con valores de sostenibilidad que EROSKI ya promueve, contribuyendo así a generar un impacto positivo más allá del ámbito comercial.

En conclusión, el proyecto de recomendación de productos para los clientes de EROSKI representa una gran oportunidad no solo para incrementar las ventas, sino también para promover un consumo más sostenible, consciente y alineado con los valores de la empresa. Si se diseña e implementa con responsabilidad, este sistema puede ofrecer recomendaciones que no solo beneficien a la empresa, sino que también ayuden a los clientes a tomar decisiones más informadas y equilibradas. Además, incorporar principios éticos como la transparencia, la protección de datos, la reducción de sesgos y la participación del cliente a través del 'feedback' permitirá fortalecer la confianza del consumidor y consolidar a EROSKI como una compañía innovadora, comprometida socialmente y líder en el uso responsable de la tecnología.

7 Glosario de Términos

A

ALS
Método de optimización utilizado en WRMF que alterna entre la actualización de vectores de usuario y producto para minimizar el error. 6

API de recomendación
Interfaz que permite a sistemas externos solicitar recomendaciones personalizadas a partir de un modelo entrenado. 7

B

Binarización de matriz
Proceso de convertir los datos de interacciones en valores 0 o 1, indicando ausencia o presencia de acción. 6

C

Clusterización (K-means)
Algoritmo de segmentación que agrupa objetos similares en clústeres según sus características. 6

cooperativa
unión voluntaria y democrática entre miembros para administrar y gestionar diversos acuerdos entre las partes, a fin de sacar adelante un proyecto. 6

Cross-Selling
Estrategia de venta que consiste en ofrecer productos complementarios a los que ya ha comprado el cliente. 6

F

factorización matricial
Técnica que descompone una matriz grande (como la de interacciones) en matrices más pequeñas que capturan relaciones latentes. 6

feedback implícito
Datos indirectos sobre preferencias del usuario, como compras o clics, en lugar de valoraciones explícitas. 6

fidelización
proceso y las estrategias que utilizan las empresas para retener a sus clientes 5

Filtrado colaborativo basado en ítem (IBCF)
Recomienda productos similares a los que el usuario ya ha consumido. 6

filtrado colaborativo basado en usuario (UBCF)
Recomienda productos basándose en similitudes entre usuarios. 6

filtro colaborativo
Método de recomendación que se basa en el comportamiento de usuarios similares o productos similares. 6

L

latent factors (factores latentes)
Variables no observadas directamente que representan características ocultas de usuarios y productos. 6

M

MAE (Mean Absolute Error)
Métrica que mide el error promedio absoluto entre las predicciones y los datos reales. 6

matriz
arreglo rectangular de números dispuestos en filas y columnas. 5

matriz cliente-producto
Representación tabular donde filas son clientes, columnas productos y celdas indican interacciones (como compras o valoraciones). 6

matriz dispersa (dgCMatrix)
Estructura optimizada para representar matrices con muchos ceros, ahorrando memoria y mejorando eficiencia computacional. 6

MSE (Mean Squared Error)
Promedio de los errores cuadrados entre predicciones y valores reales, penaliza más los errores grandes. 6

P

Precisión (Precision)
Porcentaje de productos recomendados que realmente son relevantes para el usuario. 6

R

recall
Porcentaje de productos relevantes que fueron correctamente recomendados al usuario. 6

recomendaciones top-N
Lista de N productos sugeridos con mayor relevancia para un usuario según el modelo. 6

recommenderlab
Paquete de R para construir, evaluar y comparar sistemas de recomendación. 6

regularización
Penalización añadida a una función de error para evitar que el modelo se ajuste demasiado a los datos de entrenamiento. 6

S

U

W

8 Bibliografía

- Eroski. (s.f.). *Misión, visión y valores*. Recuperado el 10 de mayo de 2025, de https://corporativo.eroski.es/quienes-somos/mision-y-valores/Eroski Corporativo+1Eroski Corporativo+1
- Heraldo. (2019, 5 de agosto). *Eroski desarrolla un plan de acciones dirigido a sus clientes más veteranos*. Recuperado el 10 de mayo de 2025, de https://www.heraldo.es/branded/eroski-desarrolla-un-plan-de-acciones-dirigido-a-sus-clientes-mas-veteranos/
- Eroski. (2017). *Memoria 2017: Consumidores y consumidoras*. Recuperado el 10 de mayo de 2025, de https://corporativo.eroski.es/memoria-2017/consumidores/
- Eroski. (s.f.). *Comprometidos con la salud y la sostenibilidad*. Recuperado el 10 de mayo de 2025, de https://www.eroski.es/salud-y-sostenibilidad/EROSKI
- Eroski. (2020, 29 de marzo). *Para esperar MENOS: Horas de mayor y menor afluencia a los supermercados*. Norte Exprés. Recuperado el 10 de mayo de 2025, de https://nortexpres.com/las-horas-de-mayor-y-menor-afluencia-a-los-supermercados/Norte Exprés+1Norte Exprés+1
- Eroski. (2024, 26 de marzo). EROSKI avanza en su transformación digital con la incorporación de nuevas funcionalidades en su app. https://corporativo.eroski.es/notas-de-prensa/eroski-avanza-en-su-transformacion-digital-con-la-incorporacion-de-nuevas-funcionalidades-en-su-app/

9 ANEXO

9.1 Anexo 1: Teoría de los algoritmos de recomendación de ALS y WRMF.

Tras haber presentado anteriormente una descripción general de los algoritmos WRMF y ALS, a continuación, se expone una explicación teórica más desarrollada.

"Weighted Regularized Matrix Factorization" (WRMF) es una técnica de recomendación especialmente útil en entornos como Eroski, donde los datos de comportamiento de los clientes (por ejemplo, productos comprados) son más frecuentes que las valoraciones explícitas. En este contexto, los sistemas deben basarse en valoraciones indirectas, es decir, inferir cuánto le ha gustado un producto a un cliente según cuánto lo ha comprado o con qué frecuencia lo ha incluido en su cesta.

WRMF trabaja con una matriz binaria de interacciones P, donde cada entrada indica si un cliente ha comprado un producto al menos una vez. Además, introduce una matriz de confianza C, que ajusta el peso de cada observación según el número de compras, con la fórmula $cui = 1+\alpha rui$. Aquí, rui es la cantidad de veces que el cliente u ha comprado el producto i, y α es un parámetro que regula cuánto influye esta frecuencia en la recomendación. Este enfoque permite distinguir entre compras ocasionales y hábitos de compra consolidados.

El objetivo es descomponer la matriz de interacciones en dos matrices de factores latentes: una para los clientes y otra para los productos. El producto de estas matrices aproxima los patrones de compra, permitiendo estimar qué productos podría estar interesado un cliente. Esta descomposición se consigue minimizando una función de error cuadrático ponderado, lo que guarda similitudes con la regresión lineal, aunque adaptada a un entorno matricial y regularizado para evitar el sobreajuste.

La implementación en R se basa en el algoritmo Alternating Least Squares (ALS), que itera alternando entre la actualización de los vectores de clientes y productos. Esto permite identificar relaciones latentes entre productos y tipos de clientes, muy útiles para generar recomendaciones personalizadas en supermercados como Eroski, donde los patrones de compra pueden variar significativamente entre perfiles de cliente.

En cuanto al algoritmo Alternating Least Squares (ALS), es una herramienta matemática que permite resolver de forma eficiente problemas de factorización de matrices, como los que se presentan en sistemas de recomendación para empresas de distribución como Eroski. Su objetivo es descomponer una matriz de compras o interacciones R, que puede estar incompleta, en dos matrices de menor dimensión: una para los clientes y otra para los productos.

Matemáticamente, ALS busca aproximar $R \approx XY\text{^}T$, donde X representa a los clientes y Y a los productos, ambos en un espacio de características ocultas o factores latentes. Cada fila de estas matrices es un vector que resume el comportamiento de un cliente o las características de un producto. Al calcular el producto escalar entre estos vectores se obtiene una estimación del interés del cliente por el producto, lo que equivale a una forma de regresión lineal que predice una puntuación o probabilidad de interacción.

El algoritmo minimiza el error cuadrático medio entre las compras observadas y las predichas, mediante la siguiente función de coste:

$$\sum_{(u,i)\in \text{datos}} (r_{ui} - x_u^T y_i)^2 + \lambda(\|x_u\|^2 + \|y_i\|^2)$$

El segundo término actúa como regularizador, evitando que los vectores aprendidos se ajusten demasiado a los datos conocidos (sobreajuste), lo que mejora la capacidad de generalización del modelo.

ALS alterna entre fijar Y (productos) y resolver X (clientes), y viceversa, hasta alcanzar la convergencia. Cada paso implica resolver múltiples regresiones lineales independientes, una por cliente o producto. Esto hace que ALS sea escalable y adecuado para grandes volúmenes de datos, como los que maneja Eroski en su red de tiendas físicas y online.

Aplicado a este entorno, ALS permite predecir qué productos podrían interesar a un cliente basándose en sus compras pasadas y en las de clientes similares. Esto es crucial para desarrollar estrategias de marketing personalizado, optimizar promociones, y mejorar la experiencia del cliente mediante recomendaciones relevantes en la tienda o en la app de Eroski.

9.2 Anexo 2: Recomendación de Producto Específico con WRMF

El primer objetivo del proyecto se centró en la generación de recomendaciones para un producto específico utilizando un enfoque de filtrado colaborativo con retroalimentación implícita. Para ello, se trabajó con una matriz de comportamiento cliente-producto que fue transformada a formato matricial y completada con ceros en los casos de ausencia de interacción.

El producto objetivo fue identificado y ajustado en su formato para coincidir con los nombres de fila de la matriz. Como el modelo WRMF requiere que los ítems estén representados en las filas, se procedió a transponer la matriz original, colocando los productos como filas y los clientes como columnas.

Sobre esta matriz transpuesta se entrenó el modelo WRMF utilizando una dimensión latente de 10, un parámetro de regularización de 0.1, y un total de 1000 iteraciones con una tolerancia de convergencia estricta (1e-6), asegurando la estabilidad del modelo. Una vez entrenado, se extrajo la fila correspondiente al producto objetivo para obtener las predicciones. En caso de que el producto no se encontrara en la matriz, el proceso se detenía automáticamente para evitar resultados inconsistentes.

La fila fue transformada en una matriz dispersa y utilizada como entrada del modelo para generar un conjunto de 500 predicciones, correspondientes a los clientes con mayor afinidad hacia el producto objetivo. A cada cliente se le asignó un score, y se aplicó un umbral mínimo de 0.9 para conservar únicamente aquellos con una alta probabilidad de interés.

Posteriormente, se compararon estos resultados con los datos de comportamiento original para eliminar del conjunto final aquellos clientes que ya habían adquirido el producto, asegurando que las recomendaciones se orientaran exclusivamente a nuevos posibles compradores.

El resultado final consistió en un dataframe con los identificadores de clientes y sus scores de afinidad, ordenados de mayor a menor. Esta información constituye una base precisa y priorizada para el desarrollo de campañas comerciales dirigidas, enfocadas en los clientes con mayor probabilidad de conversión.

Tabla 3: Resultados Objetivo 1: Fuente: Propia.

Cliente	Score
0249d2531b3b8b566abf7ba16e05d5ee	1.094
36ebf28ad81f030087dd9c0375951863	1.067
a6c71ce95eccaee02883e1f2161a182a	1.044
8d3c22af43a692067bc1a3394a86f47b	0.996
d797ebb24fc26f23e683f4e694de9b34	0.985
112f1167a8828efd56e2c5800400e483	0.975
dfcef2b1f7bbd46daebcdb2bac2b6222	0.972
5f2c89d91969dbe9c6f8329356f5497f	0.966
680fa56313ba3841344cb4cbde7de5d7	0.965
d623a73759353910301167c377f8de75	0.954

Por último, mirando estos resultados queda claro que el algoritmo ha hecho un buen trabajo identificando a los 10 clientes que más probablemente van a responder bien a la promoción de este producto. Los scores van desde 1.09 hasta 0.90, lo que muestra que hay una diferencia clara entre el cliente más prometedor y el que cierra la lista de los 10. El hecho de que varios clientes tengan puntuaciones por encima de 1.0 es bastante positivo, sugiere que hay usuarios con una afinidad muy alta hacia este artículo.

9.3 Anexo 3: Procedimiento y evaluación de las recomendaciones generadas del objetivo 2

Respecto al objetivo dos, en primer lugar, se procedió a la carga y preparación de los datos fundamentales para el modelo de recomendación. Se contó con una matriz histórica de compras que relaciona clientes y productos, donde cada valor refleja si un

producto fue comprado o no. Debido a que el modelo WRMF está diseñado para trabajar con datos implícitos, se binarizó la matriz, reemplazando los valores ausentes o NA por ceros y cualquier valor positivo por 1. Esto significa que el sistema solo considera la presencia o ausencia de compra, sin tener en cuenta la cantidad o valoración explícita. Esta elección es justificada porque los datos implícitos son más abundantes y reflejan mejor las preferencias reales en contextos comerciales donde no hay calificaciones, sino únicamente acciones (como comprar o no comprar).

Seguidamente, se entrenó un modelo de factorización matricial con el algoritmo Alternating Least Squares (ALS), configurado para maximizar la precisión en la predicción de la afinidad entre clientes y productos. El modelo aprende representaciones latentes que capturan patrones complejos en las interacciones de compra, permitiendo predecir qué productos son más relevantes para cada cliente en función de sus hábitos previos y los patrones generales de comportamiento de otros usuarios similares.

Para la generación de recomendaciones, se enfocó únicamente en los 10 clientes indicados en el objetivo, extrayendo para cada uno el producto con mayor score o afinidad según el modelo, excluyendo aquellos que ya se encontraban en su carrito. Este paso asegura que las recomendaciones sean novedosas y relevantes, incentivando la incorporación de un ítem adicional. Para justificar la recomendación, se obtuvo el score asociado, que cuantifica la fuerza de la preferencia implícita. Un score más alto indica que el modelo estima una mayor probabilidad de interés del cliente hacia ese producto.

Tabla 4: Resultados Objetivo 2. Fuente: Propia.

ID Cliente	Código producto	Descripción	Score
53ffb83e85fd51cf1ec2fdef3c78b4fd	01012310	Plátanos de Canarias	370
26f424b3bba6aaf97952ac599ed39f75	01027205	Puerros	136
b51353fcf07cb61280eda45e33679871	01027805	Lechugas	444
25d259d32a2bc254343715f2e347c518	01201005	Cebollas amarillas	81
32cc820ac27ff143c3ea976f3fe69d34	04032060	Lonchas Porción Nacional	63
8b9aa623b654a8be21b316a5fdf41007	04032065	Lonchas Porción Internacional	185
02ff5edaa057b63ea0a0010c5402205c	11040303	Agua de Mesa	513
e27ceb0a1576648212c4325fdf7d8002	12650101	P.H. Seco Básico 2C Normal	63
af30d404e282749ccd5f5ad0c8e834c7	12650103	P.H. Seco Básico 2C Largo	100
a57938025d714b65612bf2cfde12136d	12670111	Rollo Cocina Largo	63

El análisis de los resultados, como se muestra en la tabla anterior, confirma la coherencia y relevancia de las recomendaciones generadas por el modelo. Por ejemplo, al cliente con ID 53ffb83e85fd51cf1ec2fdef3c78b4fd se le recomienda "Plátanos de Canarias" con un score de 370, un valor alto que indica una fuerte afinidad implícita según su historial de compra y patrones generales. De forma similar, otros clientes reciben productos como "Lechugas" (score 444) o "Agua de Mesa" (score 513), que son productos comunes en los carritos y reflejan preferencias realistas. Los scores varían entre 63 y 513, mostrando que, aunque todos los productos recomendados son relevantes, algunos tienen una mayor probabilidad de interés, lo que justifica la selección individualizada. Además, las recomendaciones incluyen principalmente productos frescos y básicos, lo que coincide con categorías habituales en la cesta de compra, reforzando la validez práctica de las sugerencias. Esta disparidad en los scores también evidencia que las recomendaciones no son genéricas ni uniformes, sino adaptadas al perfil y comportamiento de cada cliente.

9.4 Anexo 4: Análisis recomendaciones objetivo 3.

Tabla 5: Resultados Objetivo 3. Fuente: Propia.

Nº	Producto	Veces recomendado
1	ZANAHORIAS	1648
2	PLATANOS DE CANARIAS	1395
3	ENTERO DE SABORES	1346
4	AGUA DE MESA	1331
5	LECHE UHT ENTERA	1298
6	ENTERO NATURAL	1288
7	JAMON DE CERDO	1259
8	PATATAS TODO USO	1212
9	PECHUGA DE PAVO	1188
10	ATUN CLARO EN ACEITE OLIVA	1159
11	LECHE UHT SEMIDESNATADA	1155
12	P.H. SECO BASICO 2C LARGO	1106
13	CALABACINES	1054
14	CEBOLLAS AMARILLAS	1009

15	PUERROS	962
16	MANDARINAS POSTRE	855
17	ARROZ REDONDO	835
18	TOMATE FRITO LISO BRIK PACK	764
19	ROLLO COCINA LARGO	762
20	P.H. SECO BASICO 2C NORMAL	486

Los resultados muestran que la gente recomienda principalmente productos básicos del día a día. Las zanahorias son las claras ganadoras con 1648 recomendaciones, seguidas de los plátanos canarios y productos lácteos, lo que indica que los consumidores van a lo seguro cuando hacen recomendaciones: frutas, verduras frescas y alimentos que todo el mundo usa habitualmente.

Lo interesante es cómo se distribuyen los números. Los primeros 10 productos están todos por encima de 1100 recomendaciones, pero después hay una caída notable hasta llegar a las 486 del último puesto. Esto sugiere que existe un grupo de productos "imprescindibles" que casi todo el mundo recomendaría, mientras que otros son más específicos según gustos o necesidades particulares. La mezcla entre productos frescos, procesados y hasta papel higiénico muestra que las recomendaciones cubren tanto la alimentación como las necesidades básicas del hogar.

9.5 Anexo 5: Implementación Sistema de Recomendación en Objetivo 4.

La implementación técnica del sistema del cuarto objetivo comienza con la carga y procesamiento de cuatro fuentes de datos fundamentales: una matriz reducida de interacciones cliente-producto, una base con información estructurada de productos, registros detallados de tickets de compra y la definición específica de clientes objetivo para el cuarto objetivo del proyecto. La matriz de entrada se somete a un proceso de preparación que incluye la conversión de valores faltantes a ceros, transformación a formato de matriz dispersa tipo “dgCMatrix” para optimizar el manejo de memoria, y binarización de todos los valores mayores o iguales a uno, eliminando así información de frecuencia para enfocarse exclusivamente en patrones de preferencia.

El proceso de preparación para el objetivo cuatro involucra la extracción de identificadores de clientes objetivo desde el archivo de objetivos, seguida del filtrado tanto de la matriz de interacciones como de los registros de tickets para trabajar exclusivamente con este subconjunto. Se implementa un procesamiento temporal que convierte las fechas de los tickets a formato estándar utilizando la función “ymd” de “lubridate”, y posteriormente identifica el ticket más reciente de cada cliente mediante operaciones de agrupación y filtrado con “dplyr”, información crítica para el mecanismo de exclusión posterior.

El modelo WRMF se configura con parámetros optimizados: un espacio latente de dimensión diez para capturar las principales características de comportamiento sin sobrecomplicar el modelo, un parámetro de regularización lambda de 0.1 para prevenir sobreajuste, y un proceso de entrenamiento extensivo de mil iteraciones con una tolerancia de convergencia extremadamente estricta de 1e-6 que garantiza la estabilidad y precisión de los vectores latentes aprendidos. El entrenamiento se ejecuta sobre la matriz completa de interacciones, permitiendo al modelo aprender patrones de comportamiento que luego se aplicarán específicamente a los clientes objetivo.

Para cada cliente objetivo, se construye meticulosamente una matriz de exclusión que comienza como una matriz de ceros con las mismas dimensiones que la matriz original de interacciones. Mediante un bucle, el sistema procesa cada cliente objetivo, extrae todos los productos presentes en su último ticket de compra, valida que estos productos existan en la matriz de productos disponibles, y marca con unos en la matriz de exclusión todas las posiciones correspondientes a productos que deben ser excluidos de las recomendaciones. Esta matriz de exclusión se convierte posteriormente a formato dgCMatrix y se utiliza como parámetro de restricción en el proceso de predicción.

El proceso de generación de recomendaciones utiliza el modelo WRMF entrenado aplicando la matriz de exclusión como filtro de productos no recomendables. La función de predicción se configura con “k” igual a uno para generar una única recomendación por cliente, y utiliza el parámetro “not_recommend” para aplicar las restricciones definidas en la matriz de exclusión. Las predicciones resultantes se extraen desde los atributos del objeto de predicción, se reestructuran en formato de “dataframe” asignando los identificadores de cliente como nombres de filas, y se procesan mediante operaciones de selección y renombrado de columnas para obtener una estructura limpia de cliente-producto.

El procesamiento final de resultados incluye el enriquecimiento de las predicciones mediante una operación de “left_join” con la base maestra de productos, utilizando el código de producto como clave de unión para agregar información descriptiva completa. Las columnas resultantes se renombran a un formato estándar que incluye identificadores de clientes, códigos de productos recomendados y descripciones detalladas. Todo termina con la exportación de los resultados a un archivo CSV, almacenado en la carpeta "Resultados" con el nombre específico "resulrados_objetivo4.csv", utilizando la función write.csv con el parámetro “row.names” configurado como “FALSE” para evitar la inclusión de índices de fila innecesarios en el archivo final.

Tabla 6: Resultados Objetivo 4. Fuente: Propia.

Código Cliente	Código Producto	Producto
fe234baf66f020e01feb5253dfb398f0	01012310	PLATANOS DE CANARIAS
d85ceefcf666f2b27e3e1e1252e5a1ac	01026810	BROCOLIS
a8a16b0b76cb14783348e920a59588ed	05040181	ENTERO DE SABORES
1d98f84a5f074ed9c7a47515d4f5f329	09130302	PASTA LARGA ESPAGUETI
528435b91691a75f5a60c6ccf4c6294c	01027205	PUERROS
8e8315ed119c1382c4d351bbb188510e	05040181	ENTERO DE SABORES
fe52311b246f88407a1142d891ad77ae	04200505	JAMON DE CERDO
503a6539df48964124fe026b9deb5d13	05030102	LECHE UHT SEMIDESNATADA
a809525fe25b3de695bc87e00bea215f	05040180	ENTERO NATURAL
ec926181c315b758d775ee64a6a8e033	12650103	P.H. SECO BASICO 2C LARGO

Por último, en la vista de resultados, los resultados muestran una buena variedad en las recomendaciones generadas para los 10 clientes. El 60% de los productos sugeridos son frescos (plátanos, brócolis, puerros, mandarinas), mientras que el resto incluye lácteos, pasta y cerveza. Es interesante ver que algunos productos como "PUERROS" y "ENTERO DE SABORES" aparecen recomendados para diferentes clientes, lo que sugiere que tienen un atractivo amplio entre distintos perfiles de consumo.

9.6 Anexo 6: Segmentación de Clientes y Estrategias de “Cross-Selling”.

A partir del análisis realizado sobre el comportamiento de los clientes, se ha llevado a cabo una segmentación que permite identificar tres grupos diferenciados con características específicas. Esta segmentación se basó en variables como la cantidad total de productos adquiridos, la diversidad de productos diferentes comprados, los días activos de compra, la frecuencia semanal de las transacciones y la proporción de compras realizadas entre semana y en fines de semana. El método utilizado para realizar esta segmentación fue el algoritmo de K-means, con lo cual se obtuvo una agrupación de los clientes en tres clusters.

Esta segmentación tiene como finalidad facilitar el diseño de campañas de marketing personalizado, adaptadas a las necesidades y patrones de consumo de cada tipo de cliente. A partir de esta base, se propone la implementación de estrategias de **cross selling**, que consisten en ofrecer productos complementarios a los que el cliente ya ha comprado, con el fin de aumentar el valor medio de la compra y fortalecer la relación con el cliente.

Para cada uno de los tres clusters identificados, se han definido tácticas y acciones concretas tanto en canales online como offline.

En primer lugar, para el **Cluster 1**, que corresponde a los clientes más activos y con una mayor variedad de productos en su historial de compra, se recomienda aprovechar su alto nivel de interacción con la marca para ofrecerles productos adicionales que complementen sus hábitos de consumo. En el canal online, se pueden utilizar sistemas de recomendación dentro de la página web o la aplicación móvil, especialmente en el momento del pago, para sugerir productos relacionados. También es recomendable enviar correos electrónicos con promociones personalizadas y recordatorios de productos nuevos o similares a los adquiridos anteriormente. En el canal offline, se pueden imprimir cupones personalizados en los tickets de compra o entregarlos directamente en tienda, además de capacitar al personal de ventas para que sugiera productos complementarios de forma proactiva. Estas acciones pueden ejecutarse de forma semanal, durante eventos promocionales o en momentos clave como fines de semana o campañas mensuales.

En segundo lugar, el **Cluster 2** agrupa a los clientes que presentan una actividad media, con cierta regularidad en sus compras, pero con espacio para incrementar su nivel de consumo. Para este grupo, las acciones deben enfocarse en incentivar compras adicionales mediante beneficios visibles. En el canal online, se pueden enviar campañas por correo electrónico que incluyan ofertas tipo “2x1”, descuentos por volumen o promociones de productos complementarios. También se pueden implementar banners personalizados dentro de la web que sugieran productos afines a los ya comprados. En el canal offline, se puede recurrir a promociones en tienda como descuentos por compras combinadas o a la entrega de folletos promocionales. El momento ideal para aplicar estas acciones es poco después de una compra reciente o en los días de mayor actividad comercial, como fines de semana o fechas especiales.

En tercer lugar, el **Cluster 3** está compuesto por clientes nuevos o con baja actividad, cuya relación con la marca aún es débil o incipiente. En este caso, el objetivo debe ser aumentar la interacción y familiarizar al cliente con la variedad de productos disponibles. En el canal online, se pueden enviar campañas de bienvenida que incluyan kits promocionales o muestras digitales, así como utilizar las redes sociales para mostrar recomendaciones personalizadas y testimonios de otros usuarios. Además, se puede implementar publicidad segmentada que los incentive a realizar una segunda compra. En el canal offline, es recomendable el envío de muestras físicas o cupones por correo tradicional, así como llamadas breves desde el área de atención al cliente para ofrecer asesoría personalizada. Estas acciones deben llevarse a cabo de forma inmediata después de la primera compra o del registro del cliente, así como cuando se detecta un periodo de inactividad prolongado.

Tabla 7: Centroides Cluster. Fuente: Propia.

	Total productos	Productos distintos	Dias activos	Compras en semana	Compras entre semana
1	1.6417430	1.52492622	1.0870397	-0.0727607	1.60988189
2	-0.0587975	-0.04724698	0.7049399	-0.7909348	-0.0794880
3	-0.6789663	-0.63554069	-0.9428914	0.5563623	-0.6513111

Terminando la parte de clusters, el análisis de los centroides de los tres clústeres ha permitido identificar patrones diferenciados en el comportamiento de compra de los clientes. El primer clúster agrupa a usuarios muy activos, que compran con frecuencia, adquieren una amplia variedad de productos y lo hacen principalmente entre semana. Estos clientes representan un perfil de alto valor para la empresa, ideal para estrategias de fidelización y mantenimiento. En contraste, el segundo clúster está compuesto por clientes con un comportamiento medio o estándar: compran con regularidad, pero sin destacar ni por volumen ni por frecuencia. Son usuarios estables, que constituyen la base del negocio y que pueden mantenerse con acciones comunicativas regulares y consistentes.

El tercer clúster reúne a los clientes menos activos, con un número bajo de productos comprados, poca diversidad y escasos días de actividad, centrando sus compras principalmente en el fin de semana. Estos usuarios pueden representar clientes nuevos, poco comprometidos o en riesgo de abandono, por lo que podrían beneficiarse de campañas específicas de reactivación o incentivos personalizados. En conjunto, esta segmentación permite diseñar estrategias diferenciadas para cada grupo, optimizando los recursos de marketing y aumentando la eficacia de las acciones dirigidas a mejorar la retención, la frecuencia de compra y el valor de vida del cliente.

Cambiando el enfoque, para evaluar el éxito de las estrategias de cross selling, se proponen una serie de **indicadores clave de rendimiento (KPIs)**. Entre ellos se encuentran: la tasa de conversión en “cross-selling”, que indica qué porcentaje de clientes adquieren productos recomendados; el aumento del ticket promedio, que mide si el valor medio de la compra crece tras la aplicación de estas tácticas; la frecuencia de compra, que evalúa si los clientes compran con mayor regularidad; la tasa de clics en correos electrónicos (CTR), que refleja el nivel de interés generado por las campañas; la tasa de redención de cupones, tanto físicos como digitales; la tasa de retención de clientes, que permite saber cuántos clientes permanecen activos tras las campañas; y finalmente, el número de productos nuevos o complementarios adquiridos como resultado de las recomendaciones.

En resumen, la segmentación de clientes ha permitido identificar grupos con comportamientos distintos que requieren estrategias diferenciadas. A través de tácticas específicas para cada cluster y el uso combinado de canales online y offline, se pueden diseñar campañas de marketing personalizadas que fomenten el cross selling de

manera efectiva. El seguimiento constante de los KPIs definidos permitirá ajustar y optimizar estas estrategias con base en los resultados obtenidos.

9.7 Anexo 7: Campaña de Email Marketing con Mailchimp.

Tal y como se ha mencionado anteriormente, el email marketing representa una herramienta eficaz para mantener el contacto con los clientes, ofrecer contenido personalizado y fomentar su fidelización. Para la realización de esta campaña se ha utilizado la plataforma Mailchimp, reconocida por su facilidad de uso y sus múltiples funcionalidades orientadas al marketing por correo electrónico.

A través de Mailchimp ha sido posible diseñar el contenido del correo electrónico mediante un editor visual que permite incorporar fácilmente bloques de texto, imágenes, botones y enlaces. El diseño ha sido adaptado a la identidad visual de EROSKI, incluyendo productos destacados, recomendaciones según el perfil del cliente y accesos directos a la tienda online. Además, se ha incluido una sección destinada a resaltar los beneficios del programa EROSKI Club.

En conjunto, Mailchimp ha facilitado la creación y gestión de una campaña de email marketing alineada con los objetivos de fidelización y comunicación directa con el cliente.

9.8 Anexo 8: API

Lo siguiente que hicimos fue crear una API que permite recomendar productos de forma personalizada a cada cliente, utilizando como base el modelo del objetivo dos. Elegimos ese modelo porque nos parecía el óptimo, ya que había sido desarrollado para un propósito muy similar: identificar patrones de compra y predecir afinidades entre clientes y productos.

El funcionamiento de la API es sencillo desde fuera, pero internamente aplica una lógica bastante afinada. Al recibir el identificador de un cliente, el sistema calcula qué productos podrían interesarle más, basándose en lo que ha aprendido previamente. A partir de ahí, se eliminan automáticamente todos los productos que ese cliente ya ha comprado, lo que garantiza que la recomendación no sea redundante, sino verdaderamente útil.

Una vez filtradas las opciones conocidas, se selecciona el producto con la mayor puntuación entre los restantes. Esa sugerencia se devuelve junto con su descripción y una medida del interés estimado. Así, conseguimos que cada cliente reciba una recomendación ajustada a sus preferencias, basada en datos reales y enfocada siempre a descubrir algo nuevo que podría gustarle.