Tipo de documento: Tesis de maestría

Escuela de Negocios. Master in Management + Analytics

Consolidación de promotores de dermocosméticos en plataformas de e-commerce

Autoría: Morello, Victoria
Año: 2024

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

Consolidación de promotores de dermocosméticos en plataformas de e-commerce

TESIS

Victoria Morello

Mayo 2024

Tutor: Santiago Cisco

Resumen

La industria del cuidado de la piel facial ha experimentado un crecimiento significativo en contextos virtuales, especialmente a partir de la pandemia, donde la demanda de productos online ha aumentado notablemente. Este estudio se enfoca en analizar reseñas y calificaciones de productos de cuidado facial en una plataforma de comercio electrónico líder en Latinoamérica para entender los factores que influyen en la promoción de productos y en la percepción de los consumidores. Para este fin, se entrenan y evalúan modelos de aprendizaje automático, incluyendo árboles de decisión, Extreme Gradient Boosting, Random Forest e Histogram Gradient Boosting. Los resultados revelan la importancia de ciertas características de un producto en la predicción de su promoción, así como la efectividad de diferentes modelos en este contexto. Este estudio ofrece perspectivas valiosas para empresas que buscan mejorar sus estrategias de marketing y desarrollo de productos en la industria del cuidado de la piel facial.

Abstract

The facial skincare industry has experienced significant growth in virtual contexts, especially after the pandemic, where the online demand for products has increased notably. This study focuses on analyzing reviews and ratings of facial care products in a leading e-commerce platform in Latin America to understand the factors that influence product promotion and consumer perception. In order to achieve this, machine learning models including decision trees, Extreme Gradient Boosting, Random Forest, and Histogram Gradient Boosting, have been trained and evaluated. The results reveal the importance of certain product characteristics in predicting their promotion, as well as the effectiveness of different models in this context. This study provides valuable insights for companies looking to improve their marketing strategies and product development in the facial skincare industry.

Índice de secciones

1. Introducción 6
   1.1. Contexto 6
   1.2. Problema 7
   1.3. Objetivo 8
2. Datos 9
   2.1. Resumen de las variables explicativas 9
   2.2. Detalle de las variables explicativas 12
   2.3. Consideraciones de la variable precio 19
   2.4. Variable dependiente 22
   2.5. Interacciones entre la variable dependiente y las variables explicativas 28
3. Metodología 32
   3.1. Armado del conjunto de datos 32
   3.2. Baseline 33
   3.3. Modelos candidatos 34
   3.4. Transformaciones y selección de variables 36
   3.5. Optimización de hiperparámetros 41
4. Resultados 41
   4.1. Selección del modelo 41
   4.2. Interpretación del modelo ganador 47
5. Reflexiones finales 53

Referencias 55

Anexo 1 57

Anexo 2 57

Índice de ecuaciones

Ecuación 1. Problema del consumidor. 7

Índice de tablas

Tabla 1. Ejemplos de títulos e interacción con otras variables. 37
Tabla 2. Resumen del rendimiento de modelos candidatos. 46
Tabla 3. Resultados de la optimización de hiperparámetros en Decision Tree. 57
Tabla 4. Resultados de la optimización de hiperparámetros en XGBoost. 60
Tabla 5. Resultados de la técnica de feature importance para XGBoost. 62
Tabla 6. Resultados de la segunda optimización de hiperparámetros en XGBoost. 69
Tabla 7. Resultados de la optimización de hiperparámetros para Random Forest imputando la media. 71
Tabla 8. Resultados de la optimización de hiperparámetros para Random Forest imputando la mediana. 74
Tabla 9. Resultados de la optimización de hiperparámetros para Random Forest imputando la moda. 76
Tabla 10. Resultados de la técnica de feature importance para Random Forest. 79
Tabla 11. Resultados de la segunda optimización de hiperparámetros para Random Forest imputando la mediana. 85
Tabla 12. Resultados de la optimización de hiperparámetros para HistGradientBoosting. 88
Tabla 13. Resultados de la técnica de permutación para HistGradientBoosting. 90
Tabla 14. Resultados de la segunda optimización de hiperparámetros para HistGradientBoosting. 97

Índice de figuras

Figura 1. Porcentaje de datos faltantes para variables explicativas. 13
Figura 2. Distribuciones de variables explicativas numéricas. 14
Figura 3. Distribución de factor_protección. 14
Figura 4. Distribución de variables con tres o menos categorías. 16
Figura 5. Distribución de variables con más de tres categorías. 17
Figura 6. Matriz de correlación entre variables explicativas. 18
Figura 7. Distribución de precio por ml de La Roche Posay. 20
Figura 8. Distribución de precio por ml de Garnier. 20
Figura 9. Distribución de precios de La Roche Posay. 21
Figura 10. Distribución de precios de Garnier. 21
Figura 11. Porcentaje de casos por cantidad de reseñas. 22
Figura 12. Distribución de la variable dependiente. 23
Figura 13. Distribución de puntaje de sentimiento utilizando VADER para no promotores y promotores. 24
Figura 14. Distribución de puntaje de sentimiento utilizando VADER para detractores, neutrales y promotores. 25
Figura 15. Distribución de puntaje de sentimiento utilizando RoBERTa para no promotores y promotores. 25
Figura 16. Distribución de puntaje de sentimiento utilizando RoBERTa para detractores, neutrales y promotores. 26
Figura 17. Tópicos para reseñas de promotores. 27
Figura 18. Tópicos para reseñas de detractores. 27
Figura 19. Diferencias entre promotores y no promotores para el logaritmo de variables numéricas. 28
Figura 20. Diferencias entre promotores y no promotores para el logaritmo de factor_proteccion. 29
Figura 21. Diferencias entre promotores y no promotores para variables con menos de tres categorías. 30
Figura 22. Diferencias entre promotores y no promotores para variables con más de tres categorías. 31
Figura 23. Distribuciones de variables independientes numéricas. 32
Figura 24. Distribuciones de variables independientes categóricas. 33
Figura 25. Distribuciones de variable dependiente. 33
Figura 26. Matriz de correlación entre línea y marca. 38
Figura 27. Matriz de correlación entre línea, tipos de piel y activos. 39
Figura 28. Matriz de correlación entre línea y funciones. 39
Figura 29. Matriz de correlación entre vegano, libre de crueldad y sustentable. 40
Figura 30. Área bajo la curva de ROC para árbol de decisión. 42
Figura 31. Feature importance para árbol de decisión. 43
Figura 32. Top 40 variables con mayor contribución al rendimiento del modelo. 47
Figura 33. Dependencia parcial entre características relevantes y variable objetivo. 49
Figura 34. Correlaciones entre características relevantes y variable objetivo. 52
Figura 35. Aumento en la tasa de promotores debido a cada variable. 53

1. Introducción

1.1. Contexto

La industria del cuidado de piel facial domina el segmento de cosmética con un 42% de market share (Statista, 2023). Se espera que para 2025, la industria de la dermocosmética genere hasta 177 mil millones de dólares (Statista, 2022). Desde el inicio de la pandemia y con un mayor tiempo dedicado en nuestros hogares, se ha observado una disminución en la demanda de maquillaje, mientras que la búsqueda de productos para el cuidado facial, como hidratantes, ha experimentado un incremento (Bakhati et al, 2022: 1835). Si bien ha habido una resistencia a la migración de ventas físicas a online (muchos consumidores aún prefieren poder probar los productos en persona), cada vez son más las marcas que ofrecen sus servicios de manera virtual. De acuerdo a Ascential, se espera que las ventas online dentro de este segmento de productos crezcan un 77% entre 2021 y 2026 (Howarth, 2023). A medida que crecen las ventas online, también surgen las recomendaciones y críticas de productos en la web. Un 74% de consumidores son más propensos a realizar una compra de una compañía que cuenta con reseñas en su página web (Nosto, 2021: 8). A través de las evaluaciones de clientes, las empresas reciben comentarios sinceros e imparciales sobre sus productos y los productos de su competencia. De esta manera, se pueden identificar oportunidades de mejora y las preferencias de los clientes. Al interpretar una gran cantidad de datos, las empresas pueden mejorar la toma de decisiones comerciales. Del lado del usuario, las reseñas en línea proveen información útil para facilitar el proceso de compra. De esta manera, las voces de los consumidores pueden influir en la percepción de la marca y moldear opiniones de otros consumidores. Aprovechando los conocimientos de sus clientes, las empresas podrían mejorar la calidad de sus productos, brindar un mejor servicio o incluso identificar nuevas oportunidades comerciales.

El intercambio de información entre consumidores sobre un producto es muchas veces englobado bajo el término 'word of mouth'. Una de las primeras referencias a este término fue la propuesta por Arndt (1967: 291), quien comprobó los efectos de conversaciones relacionadas a un producto sobre las ventas del mismo: la exposición a comentarios favorables colabora en la aceptación de un nuevo producto mientras que los comentarios desfavorables, dificultan su aceptación. Por otra parte, Katz y Lazarsfeld (2006: 27) aseguraron que las opiniones, actitudes y valores – las normas – de los amigos particulares y miembros de la familia que constituyen las relaciones interpersonales son tan importantes como los medios de comunicación masivos para comprender el comportamiento comunicativo (y el comportamiento en general) de los agentes. El componente distintivo del 'word of mouth' es la adjudicación de independencia de influencia comercial entre las fuentes. Teniendo todo esto en cuenta, el 'word of mouth' es ampliamente considerado como uno de los factores más influyentes que afectan el comportamiento del consumidor (Litvin et al., 2008: 458).

Con la adopción masiva de instrumentos digitales, surge un nuevo término que se contrapone al 'word of mouth' tradicional u offline: 'electronic word-of-mouth'. Mientras que la influencia del primero disminuye rápidamente con el tiempo y la distancia, el impacto del segundo, que es rápido, conveniente y está disponible por un período de tiempo indefinido, puede llegar a tener un mayor alcance y un efecto mucho más duradero. Esto hace que las empresas dependan más que nunca de cultivar el 'electronic word-of-mouth' positivo y eliminar el negativo. Considerando su relevancia en la

percepción de una marca frente a los métodos tradicionales de publicidad, poseer embajadores es fundamental ya que pueden hacer un gran aporte a su imagen. Según la Encuesta Mundial de Consumidores en Línea de Nielsen (2009) realizada a más de 25.000 personas en Internet de 50 países, el 90% de los encuestados señalaron que confían en las recomendaciones de personas que conocen, y el 70% confía en las opiniones publicadas en Internet.

Según el sentimiento subyacente a un comentario online sobre un producto, un individuo puede clasificarse como promotor (sentimiento positivo), neutral (indiferencia) o detractor (sentimiento negativo). Dicha clasificación busca determinar la probabilidad de que cierto individuo recomiende un producto a otra persona. Es muy probable que los promotores de un producto lo recomienden a otros y, por el contrario, es extremadamente improbable que los detractores lo hagan (Raassens et al, 2017: 323). Cuando el 'electronic word-of-mouth' es positivo, impacta favorablemente en la rentabilidad de la organización. En cambio, cuando el mensaje sobre una marca o empresa es negativo, su impacto puede provocar un gran daño a la salud financiera de la empresa (Villanueva et al, 2007: 8).

1.2. Problema

En economía, la teoría del consumidor es un enfoque que busca entender cómo las personas toman decisiones de compra en base a sus preferencias e ingresos y los precios de los bienes y servicios. Un consumidor busca maximizar su utilidad intertemporal sujeta a su restricción presupuestaria. En términos matemáticos, dicho problema puede ser expresado mediante la Ecuación 1:

Ecuación 1. Problema del consumidor

$$max\, V(x_{i,m}) = \sum_{m=1}^{n} \sum_{i=0}^{t} \frac{1}{(1+\beta)^{i}} \times U\left(x_{i,m}\right)$$

$$sa \sum_{m=1}^{n} p_{i,m} \times x_{i,m} \leq M_{i}\ para\ todo\ i$$

Es decir, se busca maximizar $V(x_{i,m})$ que, a su vez, está conformada por la función de utilidad $U\left(x_{i,m}\right)$. Dicha función $U\left(x_{i,m}\right)$ será personal y dependerá directamente de la cantidad x elegida para el producto m (en total hay n productos en la economía planteada) en el período i (que va desde el presente, $i = 0$, hasta algún valor genérico t). Dicha utilidad será descontada por un factor de descuento β ya que, en línea con lo que establece la teoría económica, consumir hoy da mayor satisfacción que postergarlo (o, por otra parte, un producto específico puede ser valorado de diferente manera en distintos períodos de tiempo; por ejemplo, una crema anti-age podría ser más apreciada a medida que pasa el tiempo). Un producto m tendrá una serie de atributos (precio, marca, etc.) que generarán cierta utilidad y la elección de cantidades x de cada producto por parte de un individuo en cada período deberá respetar la restricción presupuestaria del mismo. Es decir, la multiplicación entre el precio p del producto m bajo consideración y la cantidad elegida del mismo producto (ambos en el período i) debe ser menor o igual al ingreso M del individuo en el período i (bajo el supuesto de que no podrá tomar crédito/endeudarse para cubrir gastos mayores a sus ingresos en un período determinado).

El problema del consumidor tiene su impacto en el problema del productor (cómo las empresas deciden qué producir y a qué precio venderlo). Comprender las preferencias

generales de un período, es una parte fundamental en la estrategia comercial de un negocio. La optimización y selección de atributos para un nuevo producto podría hacer la diferencia entre un lanzamiento exitoso o uno fallido. Dado que los recursos también son limitados para las empresas, las mismas deberán priorizar la incorporación de las variantes más apreciadas por su público a un producto nuevo.

A modo de ejemplo, podemos considerar 5 tendencias que representan atributos valorados. En primer lugar, la sensibilización sobre los peligros relacionados con la exposición al sol ha derivado en un incremento en las ventas de productos de protección de la piel, así como productos 'anti-age' (CEyS, 2013: 8). En segundo lugar, la creciente apreciación por la multifuncionalidad de productos (productos con más de una función como las BB Creams) surge como respuesta al ahorro de tiempo y costo que suponen (CEyS, 2013: 11). En tercer lugar, comienzan a valorarse en mayor medida los productos masivos frente a los premium. Si bien las innovaciones a menudo son introducidas por las marcas en el segmento premium, la industria también está desarrollando la estrategia de incorporación de líneas de productos masivos que son percibidos como prestigiosos (masstige) en mercados maduros. Ello es el resultado de que los consumidores son cada vez más conscientes de que ciertos productos del mercado masivo pueden proporcionar la misma calidad que sus homólogos de lujo (CEyS, 2013: 12). En cuarto lugar, la digitalización de las transacciones coloca al precio en un lugar decisivo en la compra de cosméticos en línea (CEyS, 2013: 12). Por último, el cuidado del medio ambiente ha adquirido gran relevancia. Los cosméticos naturales, fabricados de acuerdo con la filosofía del comercio justo, surgieron de un nicho que antes ocupaba un pequeño número de empresas y consiguieron incorporarse en el mercado principal. Estos son distribuidos a través de los canales estándar, tales como supermercados y grandes almacenes. Su facturación anual a escala global es estimada en u$s 26.300 millones (2011) y la tasa de crecimiento promedio anual del 13,8 % desde 2006. De aquí surge el concepto de 'orgánico', que no ha sido todavía claramente definido y en muchos casos responde no tanto a la composición química del producto sino al proceso de elaboración amigable con la naturaleza. En la introducción de productos, los fabricantes de cosméticos hacen cada vez más hincapié en que se han producido de una manera que no es perjudicial para el medio ambiente, mientras que su eficacia es similar a las versiones anteriores. En algunos países, se han puesto en vigor cambios en la legislación para hacer que los consumidores estén más informados y para ayudar a distinguir los cosméticos orgánicos y no orgánicos, dejando que la gente aprecie más su valor (CEyS, 2013: 12). Está claro que las características mencionadas previamente (protección solar, multifuncionalidad, mayor calidad, menor precio y sustentabilidad) son valoradas positivamente en líneas generales. Sin embargo, no todas las dimensiones serán compatibles debido a un costo excesivo o por ser excluyentes (por ejemplo, que un producto brinde una humectación intensa, pero a la vez tenga una textura liviana). Las empresas deberán, por lo tanto, decidir qué atributos priorizar. Entonces, una comprensión sobre qué factores inciden en una valoración positiva resulta útil en caso de estar por lanzar un nuevo producto al mercado.

1.3. Objetivo

En este estudio se utiliza una base de datos de una plataforma dedicada al comercio electrónico en Latinoamérica que contiene reseñas para productos de variadas industrias, entre ellas la industria del cuidado de la piel facial. Si bien dicha plataforma tiene presencia en muchos mercados de América Latina, los datos utilizados en este estudio corresponden al mercado argentino en específico. Cada producto publicado en la plataforma está asociado a un rating promedio que surge de una serie de calificaciones de

usuarios individuales. En este estudio, se categorizarán los productos a partir de su rating promedio: casos con mayoría promotores (rating promedio 4 o 5) o mayoría no promotores (rating promedio 3 o menor). Es decir, si un producto en promedio alcanza un rating alto, entonces cuenta con promotores por sobre neutrales/detractores. Se busca determinar, por un lado, si es posible predecir si un producto será promocionado o no a partir de sus atributos, y, por el otro, analizar cuáles son las características del mismo que más inciden en esa predicción (por ejemplo: su marca, su función, si posee protección solar, su formato de venta, etc.). De esta manera, una empresa podría poner foco en aquellos atributos de un producto que lo vuelven más propenso a ser percibido positivamente entre sus potenciales consumidores a la hora de lanzar un nuevo producto.

2. Datos

2.1. Resumen de las variables explicativas

La base de datos bajo consideración está conformada por productos bajo la categoría de 'cuidado facial' en una plataforma de comercio electrónico líder en Latinoamérica[1]. En total, se cuenta con información sobre 10276 productos, identificados por un código único. Se poseen 36 atributos asociados a dichos productos. Estos son:

1. Title: breve descripción del producto bajo consideración. Es el título de la publicación del producto en la plataforma de e-commerce.
2. Len_title: largo del título que acompaña a un producto en la plataforma de e-commerce.
3. Unidades_pack: cantidad de unidades que se venden en conjunto en una publicación específica.
4. Volumen_ml: mililitros contenidos en una oferta puntual (si se trata de un pack, se considera el volumen total del set).
5. Precio_ars: precio del producto en pesos argentinos. Los datos en consideración corresponden a septiembre del 2023.
6. Precio_usd: precio del producto en dólares estadounidenses, utilizando el tipo de cambio oficial a la fecha de septiembre del 2023 (367 ARS/USD).
7. Precio_ml: indica el valor de mercado de cada mililitro de un producto. Esto permite normalizar por tamaño de producto. Un producto más grande será más caro que uno pequeño, pero quizás incorporando sus respectivas cantidades, termina siendo más conveniente el producto más grande. Esta variable se calcula dividiendo a la variable precio_usd por volumen_ml.
8. Zona_aplicación: un producto de dermocosmética puede aplicarse al rostro en su totalidad o a sectores más específicos del mismo como contorno de ojos, labios, cuello, mentón, etc. Si bien para la gran mayoría de los registros esta variable toma el valor 'rostro', vale la pena distinguir aquellos productos destinados a áreas más específicas ya que se valoran atributos diferentes. Para contorno de ojos, por ejemplo, es sobre todo importante si es hipoalergénico ya que se trata de una zona mucho más sensible.
9. Marca: la marca constituye un atributo que actúa como factor de diferenciación frente a la competencia. El hecho de que una misma empresa comercialice sus productos con diferentes marcas permite a la misma cubrir una demanda más amplia y heterogénea, segmentando el mercado por niveles de ingreso, franjas etarias, gustos, etc. Por

[1] Detalle sobre la obtención de la base de datos en el Anexo 1.

ejemplo, Procter & Gamble ofrece productos naturales y sustentables con foco en minimizar la huella de carbón generada durante su proceso de producción bajo la marca Snowberry, ofrece productos veganos y cruelty free bajo la marca Native, ofrece productos económicos y para hombres bajo la marca Old Spice, ofrece productos para el cuidado facial en torno al afeitado bajo la marca Gillette, y ofrece productos derivados a partir de la ciencia cosmética con presencia de activos como vitamina c y retinol en su formulación bajo la marca Olay. Según Brand Finance, las marcas más valuadas en el mercado global son Olay - Procter & Gamble, L'Oreal Paris - L'Oreal, Neutrogena – Johnson & Johnson, Nivea - Beiersdorf, Lancôme - L'Oreal y Avon - Natura, cada una de ellas con valores superiores a u$s 5.000 millones y en el caso particular de Olay, alcanza U$s 11.709 millones (CEyS, 2013: 9). El dataset cuenta con 144 marcas en total. Las 15 con mayor frecuencia de aparición son: ACF, Lidherma, Eucerin, La Roche Posay, Dermaglós, Eximia, Isdin, Vichy, Cetaphil, Cicatricure, Garnier, Clinique, Nivea, L'Oreal Paris y Estee Lauder. Bajo la marca 'Varios' se incluyen packs con productos de diferentes marcas (por ejemplo: Kit Crema Caviahue Hombre + Jabón Neutrogena + Botella Arg).

10. Gama_marca: variable categórica ordinal que determinará si una marca es de lujo, gama media o masiva. Para generar esta variable, se divide a la variable precio_ml (para normalizar diferentes tamaños de productos) en terciles y, si el precio por ml promedio de una marca cae en el tercil superior, entonces se la etiqueta como una marca de lujo. Si cae en el tercil intermedio, entonces es de gama media y, si cae en el último tercil, será masiva.
11. Gama_marca_nominal: análoga a gama_marca, pero generada a partir de precio_ars en lugar de precio_ml de manera tal que se logre capturar mejor los cálculos diarios que un consumidor argentino podría realizar de manera más directa a la hora de seleccionar un producto.
12. Gama_producto: variable categórica ordinal que refleja la naturaleza de un precio alto, medio o bajo para un producto dentro de una misma marca. Es decir, esta variable captura el hecho de que un mismo precio puede ser considerado económico para una marca de lujo, pero no serlo para una marca masiva. Para calcular esta variable, se separan los valores de precio_ml para cada marca en terciles. Si el precio de un producto de una marca puntual se encuentra en el primer tercil, será un producto de gama alta. Si está en el tercil intermedio, entonces será de gama intermedia. Si se encuentra en el último tercil, será de gama baja. En este caso, si una marca tiene menos de 5 productos en el dataset, esta variable no tomará ningún valor.
13. Gama_producto_nominal: análoga a gama_producto, pero generada a partir de precio_ars en lugar de precio_ml de manera tal que se logre capturar mejor los cálculos diarios que un consumidor argentino podría realizar de manera más directa a la hora de seleccionar un producto.
14. Funciones: cada producto está asociado a una finalidad u objetivo puntual. Por ejemplo, algunos productos son protectores solares, otros son destinados a combatir el acné, otros son productos anti-aging, etc. Nuevamente, 'Varios' abarca registros con packs que incluyen productos con diferente función en simultáneo. Por ejemplo, un hidratante combinado con otro producto que busca fomentar un aspecto uniforme en la piel (los llamados 'anti-manchas', que poseen niacinamida en su formulación).
15. Testeado_dermatologicamente: es un criterio compartido por diversas autoridades sanitarias e implica que el producto ha sido evaluado mediante ensayos clínicos efectuados en seres humanos bajo la supervisión de un médico dermatólogo. Es una variable categórica (si/no).
16. Libre_de_fragancia: variable categórica que establece si un producto tiene o no aroma.

17. Es_comedogénico: variable categórica que establece si un producto produce imperfecciones (puntos negros, barritos, espinillas) debido a que obstruye los poros de la piel.
18. Es_hipoalergénico: variable categórica que indica aquellos productos cuyos componentes están pensados para minimizar potenciales reacciones alérgicas en la piel. Hay ciertos ingredientes en los productos cosméticos que podrían causar reactividad en nuestra piel.
19. Es_libre_de_parabenos: variable categórica. Los parabenos son una amplia familia de compuestos derivados del ácido para-hidroxibenzoico (PHBA), muy utilizados como conservantes en productos cosméticos, fármacos y en alimentos debido a sus propiedades bactericidas y fungicidas. Sin embargo, ahora crece la controversia sobre sus beneficios por sus posibles efectos cancerígenos y las marcas han empezado a clarificar si sus productos son libres de este componente.
20. Condición: variable categórica (nuevo/usado).
21. Línea: dentro de una misma marca, hay diferentes líneas de productos. Por ejemplo, para La Roche Posay la línea anti-acné se llama 'Effaclar' y la destinada a pieles más sensibles y menos propensa a generar alergias se llama 'Toleriane'.
22. Formato_producto: textura del producto. Por ejemplo: serum, bruma, crema, gel, loción, emulsión, etc.
23. Formato_venta: packaging del envase. Por ejemplo: aerosol, gotero, pomo, tubo, pack, etc.
24. Tipo_piel: audiencia target del producto de acuerdo a las características y necesidades de la piel. Por ejemplo: piel grasa, piel mixta, piel seca, piel sensible, etc.
25. Resistente_al_agua: variable categórica (si/no). Esta variable es sobre todo relevante en caso de tratarse de una cobertura o un protector solar.
26. Factor_protección: aplica a productos con cierta protección solar (factor 10, factor 30, factor 50, etc.). De lo contrario, vale 0.
27. Libre_crueldad: variable categórica (si/no). Establece si un producto fue testeado en animales. Teniendo en cuenta que difícilmente una publicación especifique que un producto no es libre de crueldad, considero a la ausencia de dicho rótulo como un 'no'. Entonces, esta variable no indica tanto si un producto es o no 'cruelty free', sino si hace referencia explícitamente a este atributo y si eso fomenta su compra. El hecho de que existan marcas[2] que sean reconocidas por el público como libre de crueldad (The Ordinary, Veganis) o no libre de crueldad (Estee Lauder, L'Oreal, Biotherm, Clinique, Kiehls, Lancôme, La Roche Posay, MAC, Mary Kay, Neutrogena, Vichy) será controlado por la variable 'marca'.
28. Es_vegano: variable categórica (si/no). Establece si un producto tiene ingredientes o derivados animales. Un producto puede ser libre de crueldad (no testeado en animales), pero no vegano y viceversa por lo que estas variables guardan una relación, pero no tienen por qué ser siempre idénticas.
29. Sustentable: variable categórica (si/no). Vale 'si' en caso de que un producto sea libre de crueldad o vegano. Indica si un producto tiene algún tipo de etiqueta que podría considerarse como sustentable.
30. Con_protector: variable categórica (si/no). Si bien dentro de la variable 'función' se indica si un producto es protector solar, existen productos principalmente destinados a otra función (como humectación) que poseen cierto grado de protección solar.

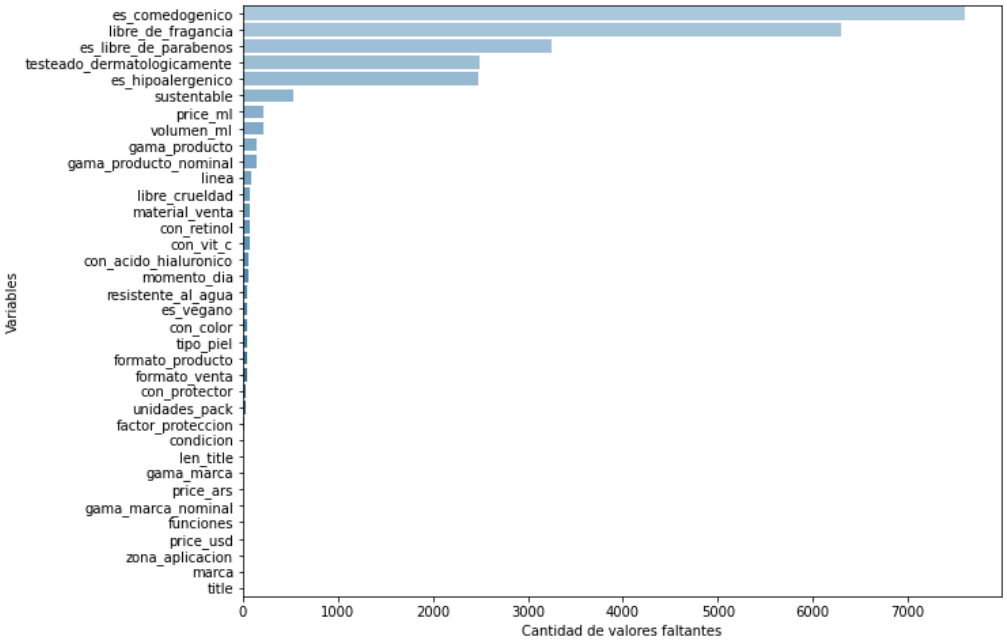
[2] El listado de marcas libres y no libres de crueldad fue realizado a partir del listado PETA (Personas por el Trato Ético de los Animales) 2023, una organización por los derechos de los animales.

31. Con_color: variable categórica (si/no). Hace referencia a si un producto tiene cierta tonalidad (existen protectores solares con color, por ejemplo, que apuntan a funcionar como una suerte de cobertura de maquillaje).
32. Con_ácido_hialurónico: variable categórica (si/no). Este activo está presente de forma natural en nuestras articulaciones, cartílagos y piel. En la piel actúa como agente hidratante y de soporte para que se mantenga un aspecto joven. Por sus propiedades hidratantes y anti-aging, empezó a aplicarse como ingrediente en tratamientos estéticos.
33. Con_vit_c: variable categórica (si/no). La vitamina C es un antioxidante potente que ayuda a la piel a combatir los radicales libres perjudiciales o dañinos. La vitamina C favorece la producción de colágeno, una de las sustancias naturales que rellenan la piel. La vitamina C es cada vez más un ingrediente clave en los productos para el cuidado de la piel debido a su efecto 'iluminador' en el aspecto facial.
34. Con_retinol: variable categórica (si/no). El retinol se trata de un derivado de la vitamina A que, al aplicarlo sobre la piel, es capaz de revertir los daños oxidativos y mostrar una piel más tersa y unificada. Los cosméticos que lo incluyen sirven para tratar pieles con acné, manchas o líneas de expresión. Sin embargo, ha estado involucrado en una polémica luego de que la Comisión Europea decidiera introducir restricciones en su comercialización, para reducir los porcentajes de retinol utilizados en productos comerciales para evitar posibles efectos adversos sobre la piel (en el caso de los productos faciales solo se permitirían concentraciones de hasta el 0,3%).
35. Momento_día: si se utiliza sobre todo de día, de noche, o ambos.
36. Material_venta: material del envase (plástico, metal, vidrio, cartón, etc.). Algunos consumidores prefieren no utilizar envases plásticos con un fundamento de sustentabilidad. Por otra parte, algunos materiales puntuales no funcionan bien con determinados formatos (por ejemplo el caso del retinol de La Roche Posay, que consistía en un pomo de aluminio muy quebradizo y luego fue modificado a una mezcla de plástico y aluminio).

2.2. Detalle de las variables explicativas

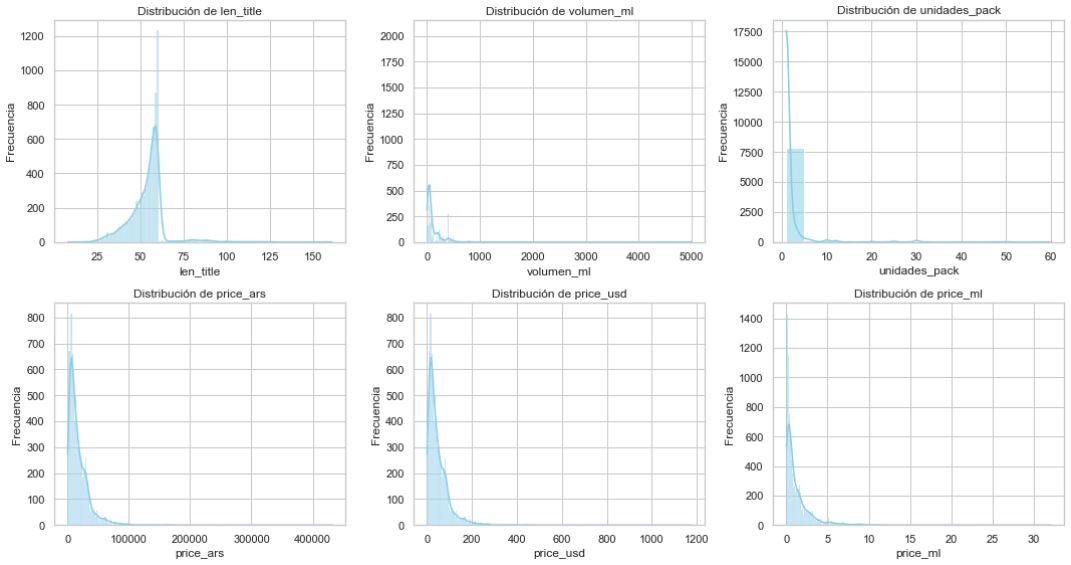
Al analizar porcentajes de datos faltantes para cada una de las variables listadas previamente, se logra visualizar a partir de la Figura 1 que son dos las variables con un porcentaje notable de valores missing. Dichas variables son libre_de_fragancia y es_comedogenico. En consecuencia y teniendo en cuenta que no serán indispensables para mi análisis, ambas variables serán excluidas.

Figura 1. Porcentaje de datos faltantes para variables explicativas



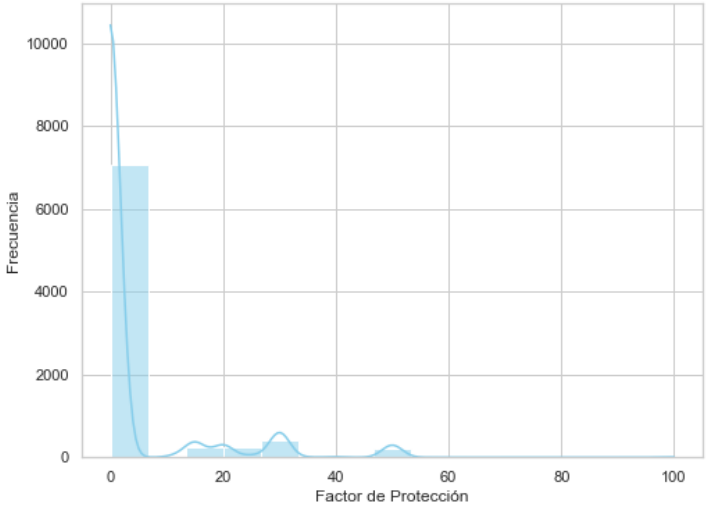
Entre las variables numéricas, hay una gran concentración en torno a valores específicos y, en menor medida, en torno a valores más extremos. La Figura 2 refleja estos puntos. En el caso de len_title, por ejemplo, la gran mayoría de las publicaciones tienen una longitud de alrededor de 60 caracteres. Sin embargo, también hay algunas publicaciones con largos tan bajos como 8 ('Vichy 98') y otros tan largos como 161 ('Caviahue Fango Termal Volcanico Mascara Facial Limpieza 50gr Tipo De Piel Todo Tipo De Piel Caviahue Fango Termal - Todo Tipo De Piel - 50 Ml - 1 - Unidad - 50 G'). Cuánto más largo el título, más detalle sobre el producto. Sin embargo, un título demasiado extenso también podría derivar en una pérdida de foco y atención por parte de la audiencia. En cuanto al volumen de los productos, la gran mayoría se centra en torno a los 50 ml. Sin embargo, también hay casos de productos con 2 ml (máscaras individuales o muestras de productos) y otros de 5000 (bidones de ácidos). Por otra parte, en general las publicaciones no superan las 10 unidades. En general, aquellas que sí superan las 10 unidades corresponden a paquetes de toallitas o pads. Los precios suelen ubicarse por debajo de los 100,000 pesos (y debajo de los 100 dólares y debajo de los 5 dólares por ml) aunque hay algunas marcas de lujo como Estee Lauder que pueden superar los 400,000 pesos (por ejemplo, la Máscara Estee Lauder Re Nutriv Ultimate Diamond 50ml).

Figura 2. Distribuciones de variables explicativas numéricas



Por otra parte, queda constatado en la Figura 3 que, para los productos que cuentan con protección UV, la gran mayoría se ubica en torno al factor 30 o 50. De todas maneras, hay algunos productos con protección muy elevada como 'Eucerin Sun Protector Solar Fps 100 Actinic Control' o 'Isdin Fotoprotector Fotoultra Fluido Antimanchas Spf99'.

Figura 3. Distribución de factor_protección



En el caso de las variables categóricas, es posible distinguir aquellas que cuentan dos o tres categorías[3] en contraposición a aquellas que presentan más de tres categorías.

Dentro del primer grupo, se observa en la Figura 4 que hay algunas variables con una distribución muy sesgada en favor a una categoría puntual. Este es el caso de las variables testeado_dermatologicamente, es_hipoalergenico, es_libre_de_parabenos, condicion, resistente_al_agua, con_color, con_vit_c, con_retinol. Resistente_al_agua y con_color son atributos aplicables al caso de protectores solares y por eso la gran mayoría

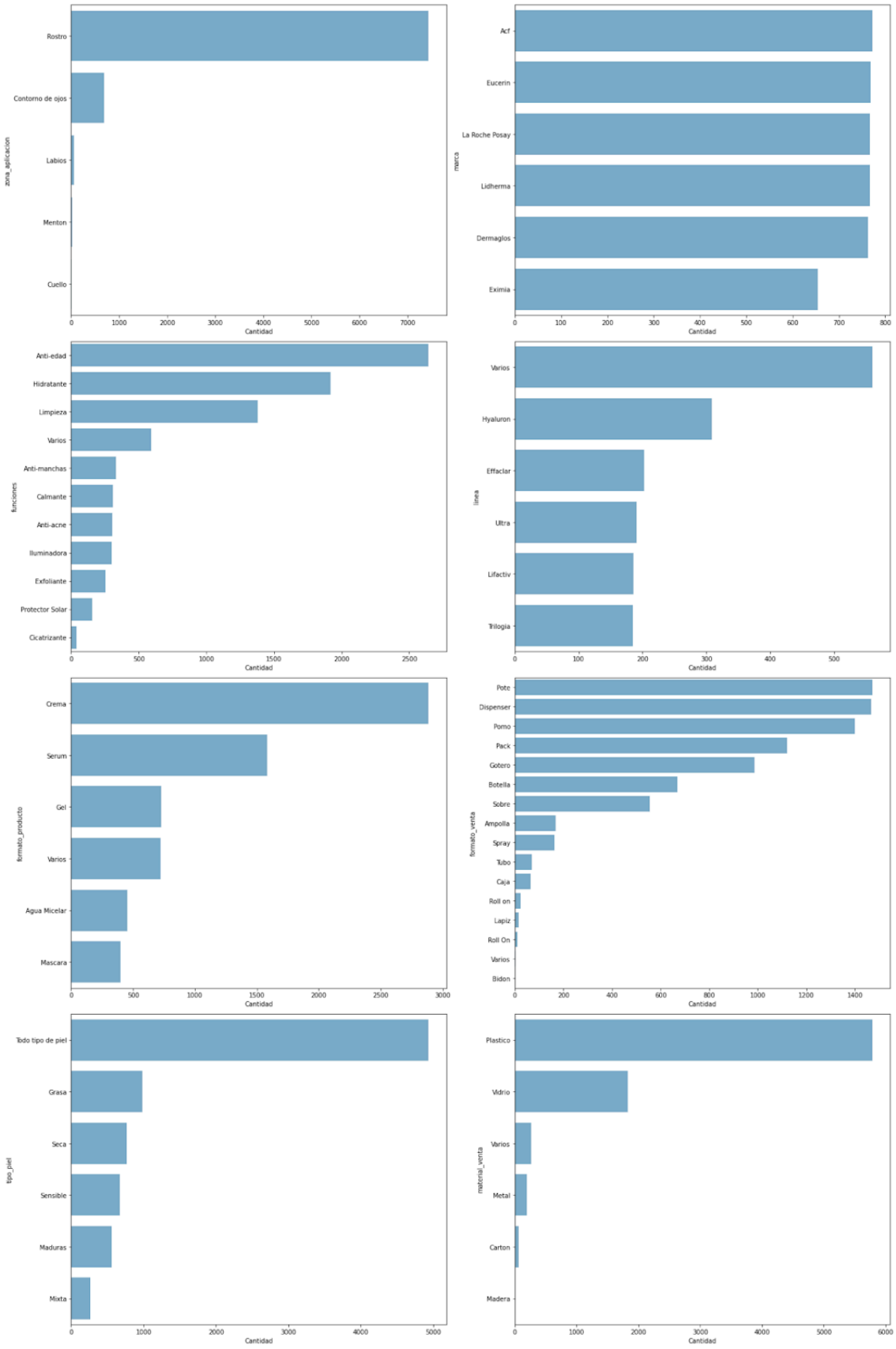
[3] Se excluyen del análisis en esta instancia las variables con un porcentaje elevado de faltantes (libre_de_fragancia y es_comedogenico).

de sus valores son 'no'. No obstante, ambas son características relevantes a la hora de comprar una crema solar por lo que van a ser incorporadas al análisis de todas maneras. Es_hipoalergenico y es_libre_de_parabenos se han vuelto características no negociables para aquellos consumidores con piel sensible o tendencia alérgica; por lo tanto también serán tenidas en cuenta a pesar de su tendencia al valor 'si'. A su vez, es posible notar que el ácido hialurónico se ha vuelto un activo frecuente en cremas faciales, mientras que la vitamina c y el retinol aún se encuentran presentes en productos con funciones muy específicas (iluminadores o anti-manchas en el caso de la vitamina c y anti-age en el caso del retinol). A su vez, el retinol es un activo costoso que refuerza el hecho de que no se haga presente de manera tan frecuente en productos cosméticos. Tanto la vitamina c como el retinol se han vuelto cada vez más populares por sus beneficios en la piel y por eso serán incorporadas al análisis. Contrariamente, se excluirán las variables testeado_dermatologicamente y condicion. El consumidor suele asumir que los productos en góndola han sido sometidos a las pruebas de salud correspondientes previo a su venta. Por esto, si bien un producto puede no especificar explícitamente si ha sido efectivamente testeado, esto no debería influir demasiado en el comportamiento consumidor. De manera similar, la condición 'usado' solamente aplica a elementos como máquinas para limpieza de piel, una porción mínima del dataset.

Dentro del segundo grupo, se visualiza en la Figura 5 que las variables marca, línea y formato_producto son las que más cantidad de categorías presentan (144, 555 y 45 respectivamente). Dada la imposibilidad de visualizar la distribución de registros entre la totalidad de categorías para estas tres variables, se analizarán las seis categorías más frecuentes para cada una de ellas con este fin. De esta manera, se corrobora que efectivamente ninguna categoría específica concentra la gran mayoría de registros. Para la variable 'marca' en específico existe una distribución variada entre marcas locales como Dermaglós, ACF, Eximia y Lidherma, y marcas importadas como La Roche Posay y Eucerin. La línea 'Varios' se refiere a publicaciones con múltiples líneas en simultáneo. Por ejemplo, si una publicación de La Roche Posay ofrece un pack de productos y uno es de la línea Effaclar y otro Toleriane, entonces la línea en este caso es 'Varios'. El formato de producto más popular en el dataset es 'Crema' seguido por 'Serum' y el formato de venta 'Pote' (el cual está más asociado a las cremas) y 'Dispenser' (más asociado a serums). La variable zona_aplicacion muestra que la gran mayoría de productos del dataset están destinados al rostro. De todas maneras, dicha variable permite distinguir aquellos productos dirigidos al contorno de ojos, una zona mucho más sensible que requiere un cuidado especial. Por este motivo, los consumidores suelen ser más exigentes y cuidadosos al comprar productos destinados a esta área. Entre las funciones más frecuentes de productos, se destacan limpieza, hidratación y anti-edad. El tipo de piel al cual más productos están dirigidos es 'todo tipo de piel'. Si bien esto resulta en un producto con una audiencia más genérica y amplia, también podría derivar en efectos menos notables debido a la falta de especificidad en el tratamiento. Por último, el material de los packagings más frecuente es el plástico, seguido por el vidrio, material más amigable con el medio ambiente debido a su naturaleza más duradera y, en consecuencia, la posibilidad de reciclaje para otros usos.

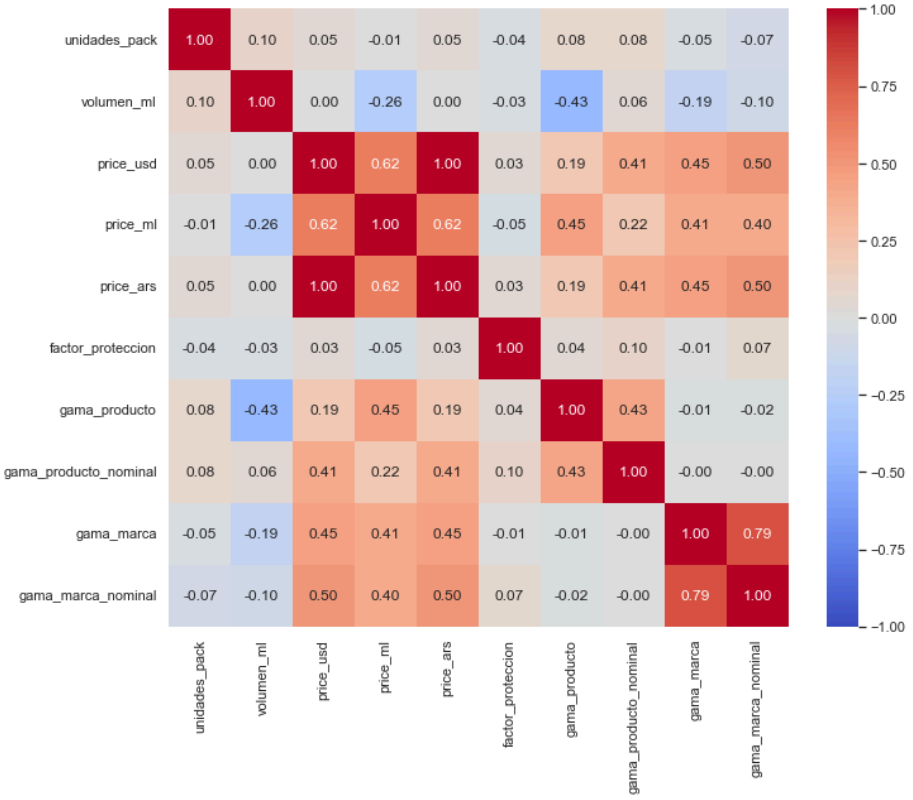
Figura 4. Distribución de variables con tres o menos categorías

Figura 5. Distribución de variables con más de tres categorías



Para las variables numéricas relacionadas al precio, packaging, factor de protección solar y, a su vez, para las variables categóricas ordinales gama_producto, gama_producto_nominal, gama_marca y gama_marca_nominal, se realiza un análisis de correlación entre las mismas, el cual es resumido a partir de la Figura 6. De esta manera, se logra visualizar una correlación positiva entre la gama de un producto o marca y su precio (las gamas más lujosas correlacionan positivamente con precios más elevados). Por otro lado, las marcas más lujosas suelen ofrecer productos más pequeños (menor volumen y menor cantidad de unidades por pack). El atractivo por los mini-packs representa una tendencia para nada desconocida en el mercado de productos cosméticos. En su reporte de tendencias en 2023, Cosmetics Business revela como cuarta tendencia 'Mini size, maxi appeal'. El encanto de los minis de belleza es cada vez mayor a medida que más consumidores buscan formas de experimentar la belleza de lujo por menos. Los minis de belleza y fragancias están volando de los estantes por ofrecer un punto de entrada asequible a marcas de alta gama (Allen, 2023).

Figura 6. Matriz de correlación entre variables explicativas



Teniendo en cuenta que la variable que contiene los precios de productos en pesos argentinos y la variable que contiene los precios de productos en dólares son altamente correlacionados, se descartará la primera ya que sus valores quedarán rápidamente desactualizados debido al proceso inflacionario que transcurre actualmente en el país. Por otra parte, gama_marca y gama_marca_nominal también muestran una alta correlación y por esto solamente la segunda formará parte del modelo. Gama_marca_nominal captura mejor la lógica de compra de un consumidor al considerar si una marca es o no lujosa. Gama_marca está generada a partir de la variable precio por ml mientras que gama_marca_nominal se genera a partir del precio de un producto en pesos argentinos. Usualmente, el consumidor piensa en términos nominales. De hecho, esta realidad se

conoce bajo el nombre de 'ilusión monetaria'. Este fenómeno se define como la tendencia de ser influenciado por valores nominales en lugar de reales cuando se hacen transacciones económicas, tanto en empresas como en la vida diaria (Galeas et al, 2005: 56). Teniendo esto en cuenta y, dado que gama_marca no está agregando información nueva por sobre gama_marca_nominal, se optará por excluir la primera.

2.3. Consideraciones de la variable precio

Debido al contexto inflacionario del país, es esperable que los valores absolutos en pesos rápidamente queden desactualizados. En caso de aplicar modelos de árboles, el valor absoluto no importa tanto como el orden relativo. Sin embargo, al contar con marcas tanto extranjeras como nacionales, la tasa de crecimiento en los precios podría no mantenerse constante entre dichos productos (dada la fluctuación del dólar en el país y las restricciones a las importaciones). Para comprobar si existe este diferencial, se busca determinar si los productos de las marcas locales verdaderamente son de fabricación únicamente nacional ya que, en muchos casos, las mismas importan gran parte de los insumos utilizados durante el proceso de producción. De ser así, todos los productos, de alguna manera, tendrían naturaleza de 'importados'. De acuerdo a un reporte del Centro de Estudios para la Producción (2004: 51), las principales materias primas utilizadas por el sector provienen de origen importado ya que Argentina cuenta con una escasa oferta local. Estados Unidos, Alemania y Suiza son los mayores proveedores de las mismas. La insuficiencia o falta de provisión local de determinados insumos (esencias, vitaminas, colorantes, etc.) ha impedido que los componentes nacionales sustituyeran en mayor medida a los importados, pese a que estos se han encarecido considerablemente. Para los materiales de empaque (envases de plástico, vidrio, hojalata), sin embargo, sí existe una numerosa red de productores nacionales. El Consejo Económico y Social de la Ciudad de Buenos Aires afirmó que 'el sector es altamente dependiente de la importación de insumos' (2013: 39). En la misma línea, un reporte del Ministerio de Producción y Trabajo de la Presidencia Nacional (2019: 6) sobre el sector cosmético, aseguró que los materiales de empaque son principalmente de fabricación nacional, pero las materias primas son sobre todo de origen extranjero mismo en las marcas nacionales. A su vez, la Cámara Argentina de Productos Químicos expresó a principios de 2023[4] su preocupación en torno a la traba de importaciones ya que derivarían en faltantes de insumos básicos para diferentes industrias -desde farmacéuticas, consumo masivo y minería, entre otras- ya que 'la mayor parte de la materia prima requerida no se produce en la Argentina'.

Teniendo todo esto en cuenta, se asumirá que el comportamiento de precios debido a la inflación será similar tanto para marcas extranjeras como locales (ya que en ambos casos se importa gran parte de las materias primas) y las relaciones relativas se mantendrán aproximadamente estables. De todas maneras y a modo de atenuar el impacto de la inflación que desactualiza los valores absolutos, se utiliza la variable Precio_usd que expresa los precios en dólares estadounidenses tomando la cotización del momento (septiembre 2023): 367 ARS/USD. Además, a modo de normalizar productos con diferentes tamaños, se incluye la variable Precio_ml que determina el precio promedio en dólares por cada ml del producto en cuestión.

Como control adicional, se realiza un análisis de precio por marca. Se incluye una variable categórica ordinal llamada gama_marca que determina si una marca es de lujo, de gama media o masiva. Si el precio por ml promedio en dólares de una marca se ubica en el

[4] Empresas advirtieron que las trabas a la importación de insumos químicos pueden frenar la producción

primer tercil de la variable Precio_ml, entonces será de lujo. De caer en el segundo tercil, será de gama media y, si cae en el último tercil, será masiva.

Se suma otra variable, además, que reflejará la naturaleza de un producto de precio alto, medio o bajo dentro de una misma marca. Para calcular esta variable, Gama_producto, se separan los valores de precio por ml para cada marca en terciles. Por ejemplo, para la Roche Posay la Figura 7 indica que un producto menor a 1.16 dólares por ml es etiquetado como precio bajo (ya que para esa marca es un precio bajo), pero en el caso de Garnier (marca masiva en lugar de premium) la Figura 8 muestra cómo dicho criterio cambia a 0.03. De no hacer este análisis separado por marca, todos los productos de La Roche Posay serían considerados de precio alto, y estaría capturando más la distinción de marca que la noción de precio. Por lo tanto, si una persona es particularmente leal a una marca, por ejemplo, La Roche Posay, es posible que considere un producto de 1.14 dólares por ml razonable mientras que una persona leal a otra marca, por ejemplo, Garnier, considere este precio excesivo.

Figura 7. Distribución de precio por ml de La Roche Posay

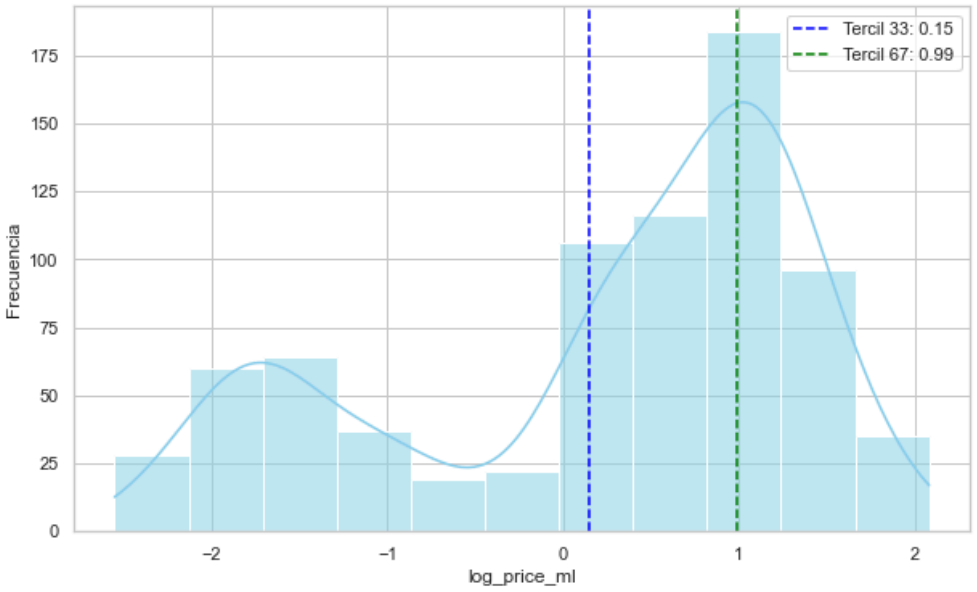
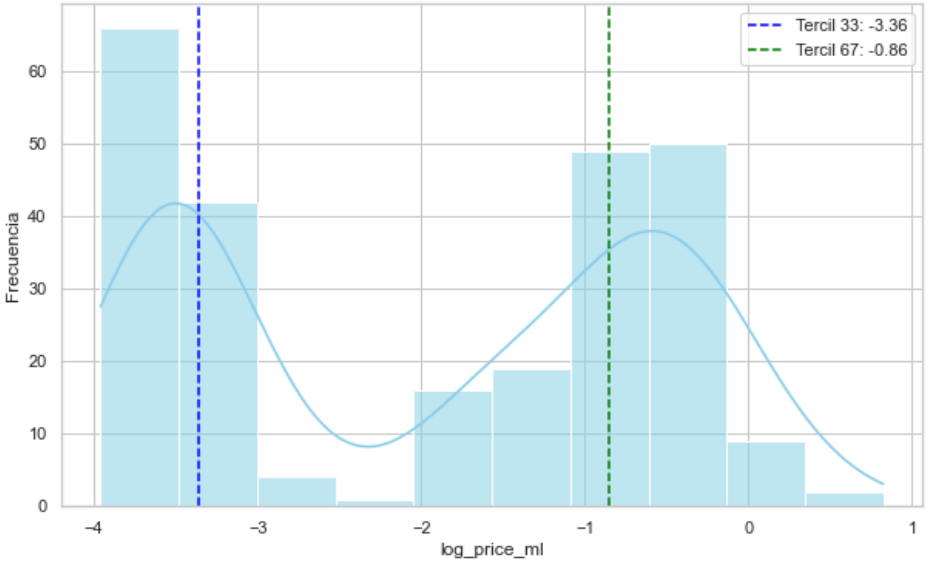


Figura 8. Distribución de precio por ml de Garnier



El motivo por el cual se considera a la variable precio_ml para generar la variable gama_producto es para controlar por el hecho de que una marca puede tener precios en términos nominales más elevados, pero solamente por ofrecer productos de mayor volumen o kits con varios productos incluidos. Sin embargo, teniendo en cuenta que un consumidor suele razonar en términos nominales (es decir, no estar dispuesto a gastar más de determinado monto en pesos para cierto producto de una marca particular independientemente de la variable volumen), también se incluye una variable categórica ordinal similar a gama_producto, pero generada a partir de precio_ars en lugar de precio_ml; la misma se denomina gama_producto_nominal. En este caso, para la Roche Posay un producto menor a 20,840 pesos es etiquetado como precio bajo (Figura 9), pero en el caso de Garnier, el criterio cambia a 2,818 (Figura 10).

Figura 9. Distribución de precios de La Roche Posay

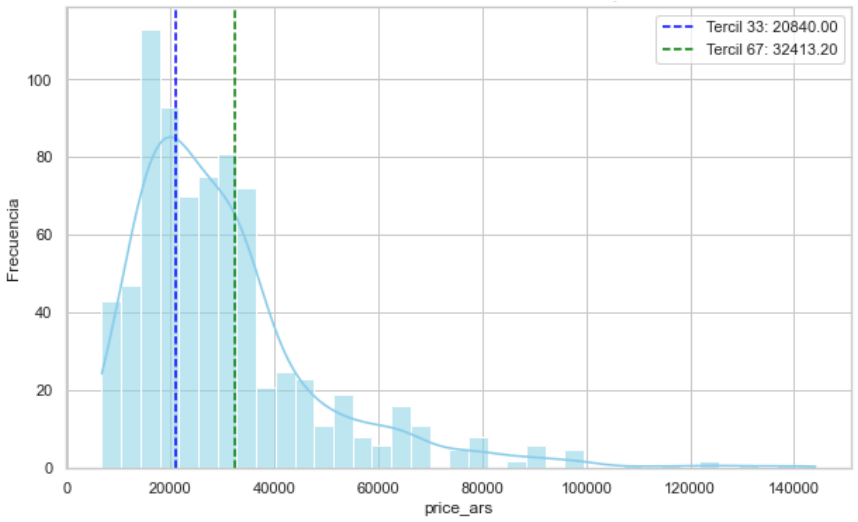
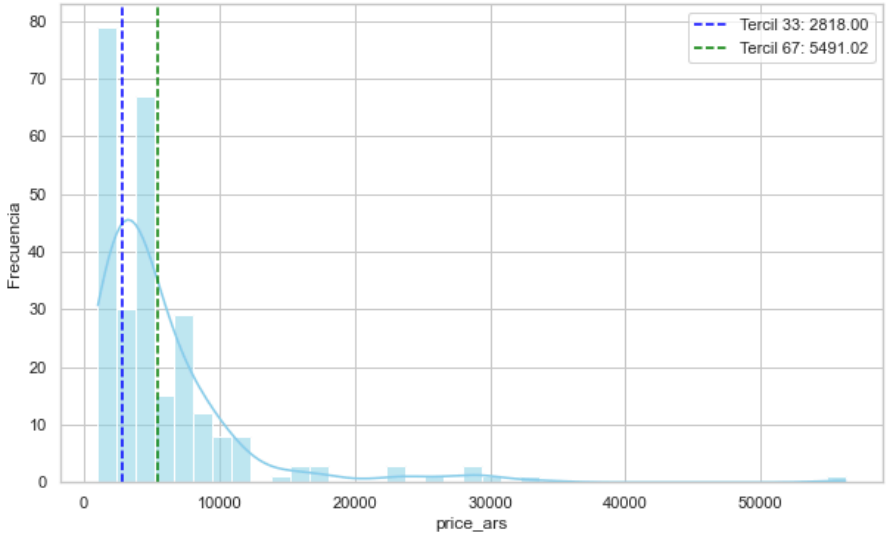


Figura 10. Distribución de precios de Garnier

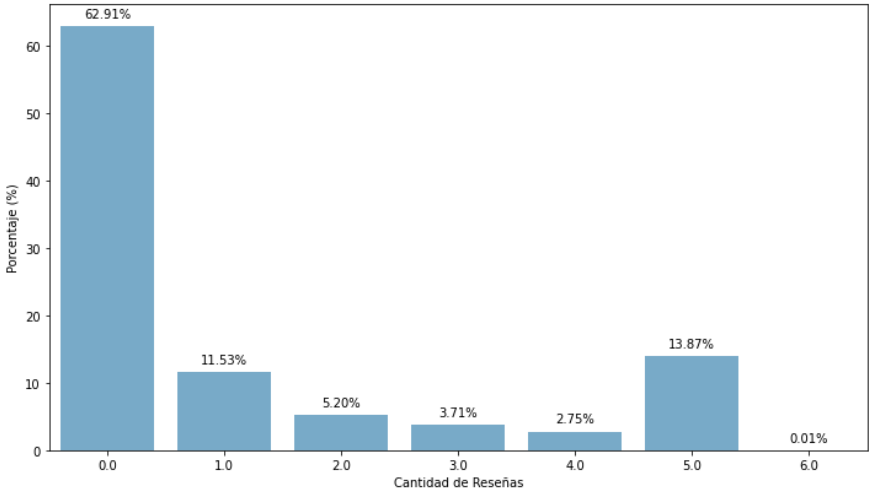


2.4. Variable dependiente

Del total de 10,276 productos bajo consideración, 3,797 presentan al menos una reseña. En total, hay 11,071 reseñas sobre estos 3,797 productos. Para cada uno, se posee información sobre: fecha y hora en la cual el producto reseñado fue comprado por el cliente, fecha y hora en la cual el producto fue reseñado, el título que lleva la reseña (un breve resumen del contenido de la reseña), el contenido de la reseña, cuántos 'me gusta' tuvo una reseña, cuántos 'no me gusta' tuvo una reseña, cantidad de palabras en la reseña y el rating individual asociado a la reseña.

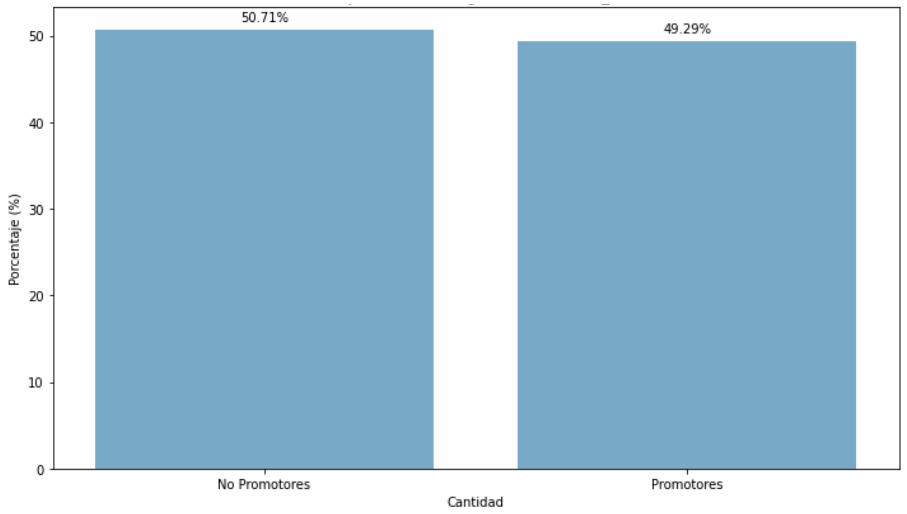
La variable rate_av_reviews indica, si un producto tuvo reseñas, el promedio de los ratings de dichas reseñas (si no tuvo reseñas, se mantiene en cero). Por otro lado, rating_average establece la calificación promedio de un producto independientemente si dichas calificaciones fueron acompañadas o no por un comentario. Si bien aquellos consumidores que califican un producto y además dejan una reseña, están más involucrados con la marca y probablemente sean promotores más efusivos en caso de tener una buena experiencia con un producto que aquellos que califican positivamente el producto, pero no dejan ninguna reseña, también es cierto que muchas personas optan por no redactar un comentario por tiempo, conveniencia o privacidad. En caso de considerar únicamente las calificaciones asociadas a una reseña, estaría dejando de lado a más de la mitad de los productos bajo análisis (Figura 11). Además, para aquellos productos que sí tienen asociadas reseñas, la cantidad de dichas reseñas no supera el número 6. Por lo tanto, estaría poniendo un peso excesivo en un grupo muy pequeño de calificaciones en lugar de considerar todo el conjunto.

Figura 11. Porcentaje de casos por cantidad de reseñas



Teniendo esto en cuenta, la variable dependiente será generada a partir de la variable rating_average y no rate_av_reviews. La variable dependiente será categórica y establecerá la presencia o no de promotores para un producto puntual. Si un producto calificó en promedio entre 4 y 5 (es decir, el valor de rating_average es mayor o igual a 4), entonces se considera que tiene, en gran medida, promotores. En caso contrario, no los tiene. Dentro del grupo 'sin promotores' tengo en cuenta aquellos productos tanto con detractores (calificación entre 1 y 2.99) como neutrales (calificación entre 3 y 3.99). Tal como se logra ver en la Figura 12, la distribución entre ambas categorías es realmente similar.

Figura 12. Distribución de la variable dependiente



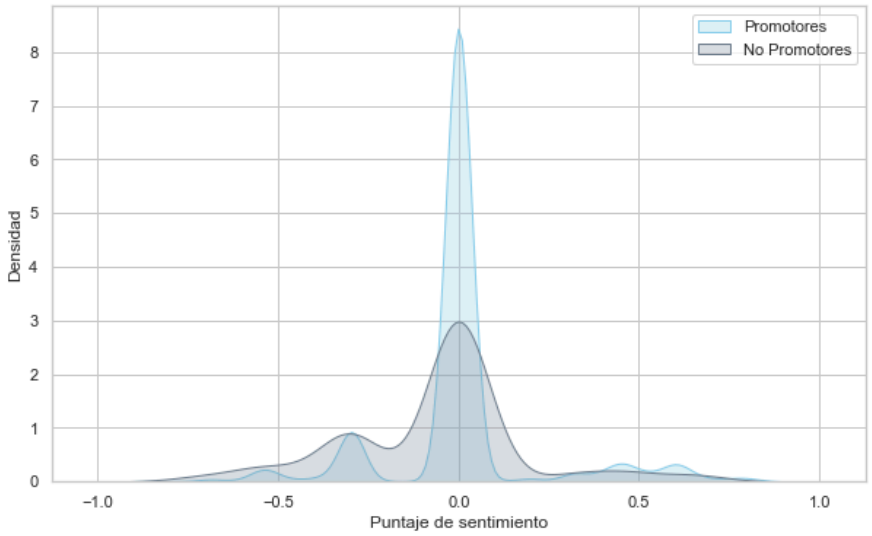
Para confirmar la robustez de la variable dependiente y en línea con el trabajo de Raassens et al (2017: 2), se busca encontrar la relación entre la categorización promotor-no promotor basada en el score promedio de un producto y el 'electronic word of mouth'. Es decir, ¿efectivamente un producto definido a partir de su calificación promedio dentro de la categoría 'con promotores' cuenta con reseñas, en gran medida, positivas? Y para aquellos definidos como parte de la categoría 'sin promotores', ¿se cuenta con reseñas, en gran medida, negativas? Para comprobarlo, se utiliza un enfoque de procesamiento de lenguaje natural para analizar sentimiento y también el algoritmo de asignación latente de Dirichlet para identificar los tópicos más comunes entre los promotores y los no promotores (Darad et al, 2023: 110).

Para analizar sentimientos se considerarán dos enfoques. En primer lugar, VADER basado en léxicos, una librería predesarrollada en código abierto. En segundo lugar, un modelo preentrenado: RoBERTa, que es una extensión de BERT. León-Sandoval et al. (2022: 14) compara VADER con BERT y concluye que muestran tendencias similares y reaccionan de manera similar a los eventos del mundo real, lo que hace que sean buenas opciones para el análisis de sentimiento a gran escala.

VADER, perteneciente a un análisis de sentimiento basado en léxicos, está muy asociado a entornos de redes sociales ya que logra interpretar abreviaciones, señales de énfasis como las mayúsculas o puntuación, vocabulario informal y emoticones. Determina no sólo la polaridad binaria (positiva versus negativa), sino también la fuerza del sentimiento expresado en el texto. Uno de sus resultados es la puntuación compuesta, que es calculado sumando las puntuaciones de valencia (es decir, intensidad) de cada palabra en el léxico, ajustando por algunas reglas y luego normalizado para estar entre -1, el más negativo, y 1, el más positivo (Darad et al, 2023: 114). De acuerdo a Hutto et al (2014: 216) VADER funciona, en la mayoría de los casos, mejor que otras herramientas de análisis de sentimientos de gran prestigio, supera a los evaluadores humanos individuales y funciona excepcionalmente bien en el ámbito de las redes sociales y micro-blogs. En comparación con técnicas de aprendizaje automático, VADER es más rápido y computacionalmente económico, sin sacrificar la precisión; y el léxico y las reglas que utiliza son directamente accesibles (Hutto et al, 2014: 224).

Como alternativa, se consideró el modelo RoBERTa [5] que genera valores entre 0 y 1 según la probabilidad de sentimiento positivo o negativo. El modelo RoBERTa es una extensión de BERT. De acuerdo a Darad et al (2023: 115), BERT se desempeña extremadamente bien en comparación con otros modelos de machine learning. Es un tipo de Transformer denominado 'encoder-only'. La arquitectura de Transformers fue primeramente introducida en 2017, con un foco principal en tareas de traducción. Dichos modelos son inicialmente entrenados a partir de un régimen auto-supervisado, por medio del cual ajustan una distribución estadística del lenguaje mediante el análisis de grandes volúmenes de texto sin la necesidad de etiquetado humano. Posteriormente, se lleva a cabo un proceso en el cual se adaptan los modelos pre-entrenados a tareas específicas mediante el empleo de etiquetas humanas. Este enfoque permite abordar una amplia gama de aplicaciones en el ámbito del procesamiento del lenguaje natural. El modelo RoBERTa refina el modelo BERT entrenándolo con más datos, secuencias más largas y más tiempo de entrenamiento (Tan et al, 2022: 21520).

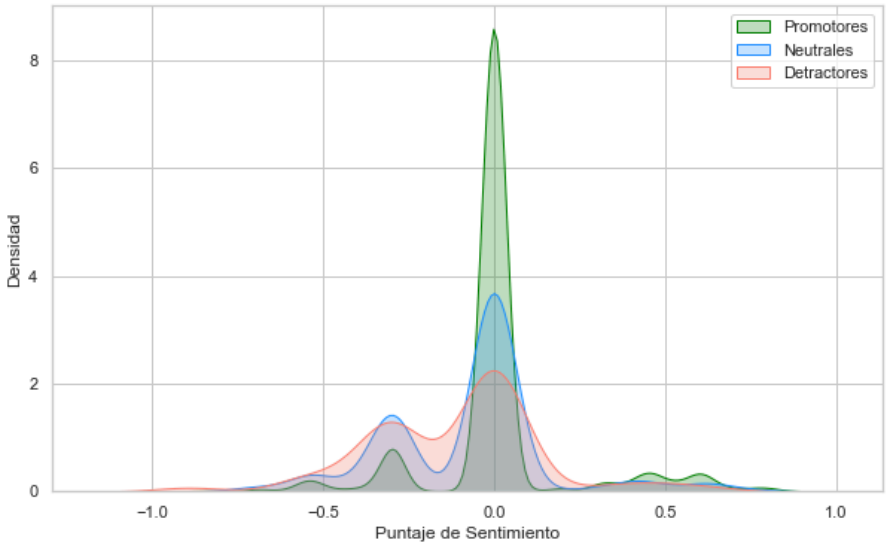
Figura 13. Distribución de puntaje de sentimiento utilizando VADER para no promotores y promotores



A partir de VADER, se logra visualizar (Figura 13) que las reseñas asociadas al grupo de promotores están ligeramente más asociadas a un sentimiento positivo mientras que las asociadas al grupo de no promotores, están más vinculadas a un sentimiento negativo. Desglosando los 'no promotores' en 'neutrales' y 'detractores' se observa en la Figura 14 que incluso los 'neutrales' presentan reseñas con cierto sentimiento negativo (aunque menos que los detractores) por lo que hace sentido agruparlos bajo la misma categoría.

[5] https://huggingface.co/edumunozsala/roberta_bne_sentiment_analysis_es

Figura 14. Distribución de puntaje de sentimiento utilizando VADER
para detractores, neutrales y promotores



Sin embargo, la tendencia no es del todo clara, dado que muchas reseñas reciben una puntuación de cero al aplicar VADER, por lo que se recurre a un método alternativo de clasificación de texto. De acuerdo a Saha et al (2022: 384), BERT supera a VADER en todos los algoritmos supervisados porque es un modelo profundo de izquierda a derecha y de derecha a izquierda basado en transformadores. RoBERTa, extensión de BERT, contempla la probabilidad de que una reseña sea positiva o negativa. Entonces, se establece que, si la probabilidad de que una reseña sea positiva supera a la de que sea negativa, esta reseña tiene sentimiento positivo (toma el valor 1). Por el contrario, si la probabilidad de que una reseña sea negativa es mayor, entonces es negativa (toma el valor -1). De esta manera, y a partir de su aplicación en las reseñas de productos, se logra establecer en la Figura 15 de forma más definitiva que las reseñas de los 'promotores' son, en gran medida, positivas y las de 'no promotores', negativas (acentuado en caso de 'detractores' en comparación a 'neutrales' en la Figura 16). Esto supone evidencia a favor de la definición promotor - no promotor establecida para la variable dependiente.

Figura 15. Distribución de puntaje de sentimiento
utilizando RoBERTa para no promotores y promotores

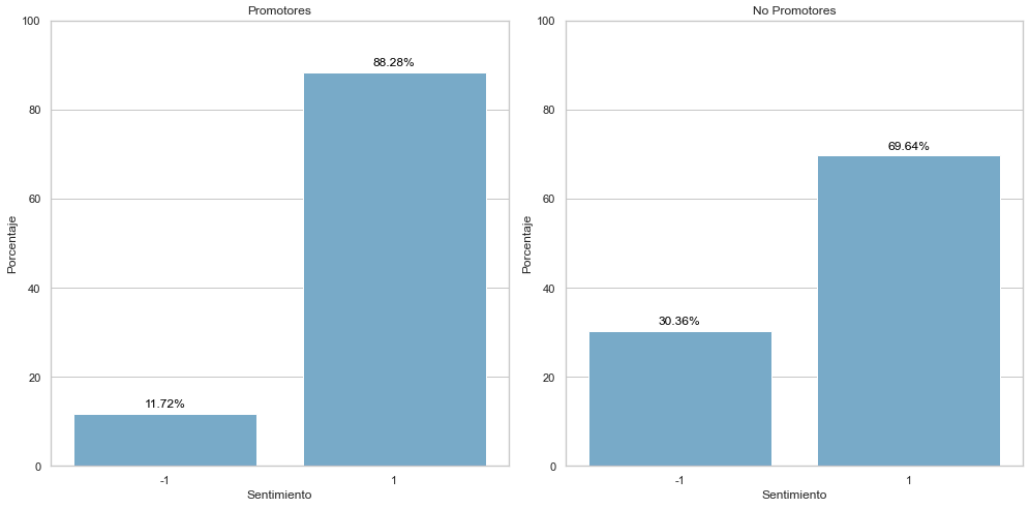
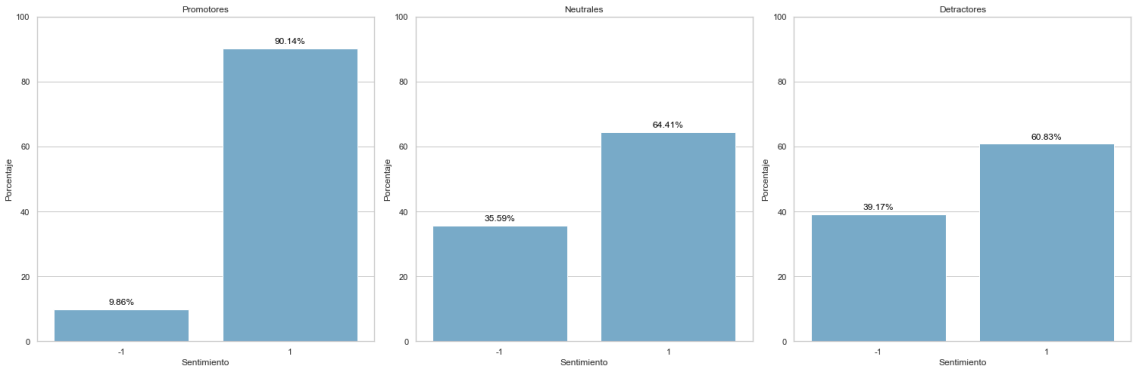


Figura 16. Distribución de puntaje de sentimiento utilizando RoBERTa
para detractores, neutrales y promotores



Para comprender mejor a los individuos dentro de cada grupo, se adoptará a modo complementario un enfoque de 'topic modeling' no supervisado (LDA) aplicado a las reseñas de productos, diferenciando por rating. Es decir, para ratings bajos (1 y 2), se interpretarán las temáticas principales referenciadas en la Figura 18. Lo mismo se hará para los ratings altos (4 y 5) en la Figura 17. De esta manera, quedan definidos los atributos mencionados en cada grupo de polaridad. LDA utiliza una estrategia probabilística para asignar palabras a tópicos y aprender la distribución de tópicos en los documentos. Según Kherwa y Bansal (2019: 7-8), cuando comparamos a fondo contra otras medidas estadísticas, los términos principales de todos los temas con LDA muestran claramente temas muy nítidos, claramente separados y también coherentes para contar la naturaleza del tema individual.

Dentro de los promotores, el primer tópico referenciado hace énfasis en la textura (suave) del producto. El segundo tópico se enfoca en la ausencia de irritación en la piel (producto apto para caras sensibles). El tercer tópico hace referencia a la calidad del producto (deriva en resultados visibles y diferenciales) y la apreciación por la marca. Por último, el cuarto tópico menciona el grado de absorción de las cremas (efecto seco en lugar de grasoso).

Dentro de los detractores, el principal tópico coloca el foco en fallas en el packaging del producto. Es decir, ocasiones en las que un producto era recibido con la faja de seguridad fallada o incluso con su respectiva caja abierta. El segundo tópico entre detractores consiste en desajustes entre precio y calidad (la calidad no se condice con el precio). Otros tópicos abarcan texturas grasosas, fragancias invasivas y rapidez en la oxidación de productos de vitamina C.

Figura 17. Tópicos para reseñas de promotores

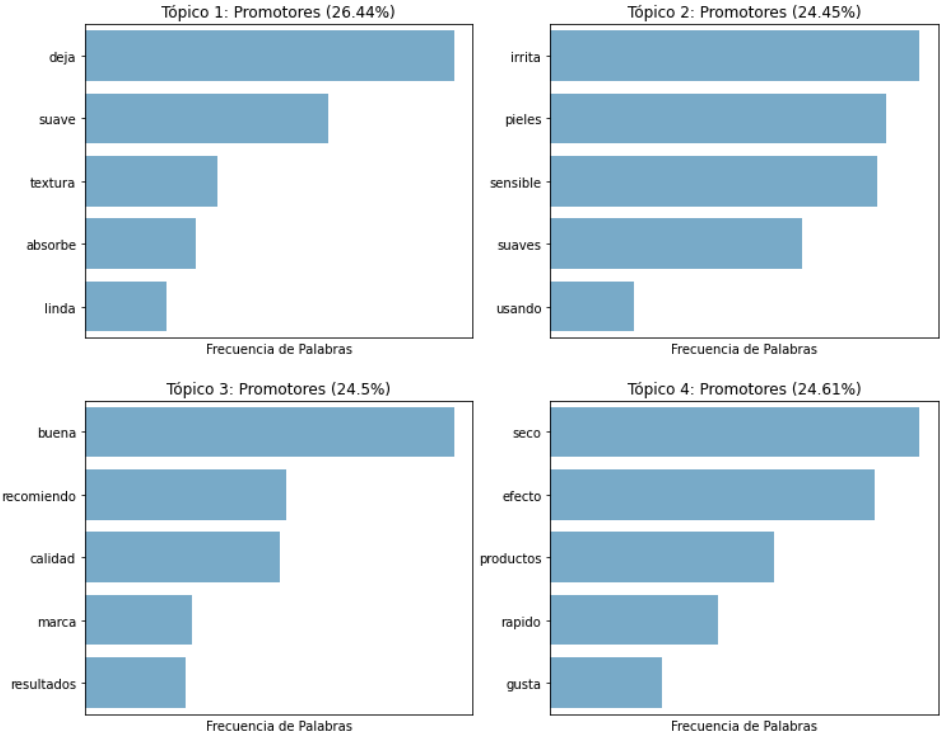
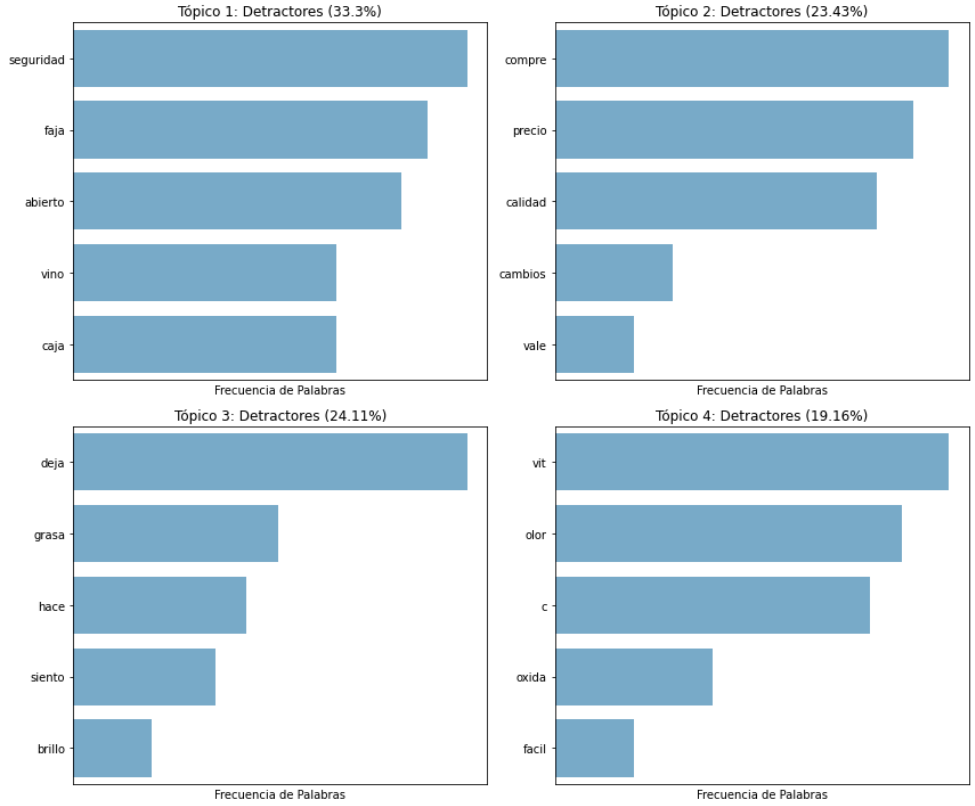


Figura 18. Tópicos para reseñas de detractores



2.5. Interacciones entre la variable dependiente y las variables explicativas

A primera vista y a partir de un análisis de medias, los productos con promotores parecieran estar asociados a títulos de publicaciones más extensos. Sin embargo, al analizar distribuciones a partir de curvas de densidad (Figura 19), se visualiza un comportamiento similar, a grandes rasgos, de la variable len_title entre categorías de la variable var_dep_3. Ambas categorías presentan una densidad máxima aproximadamente en 55 caracteres y el rango de caracteres oscila entre 20 y 125. La diferencia de la media resulta de un grupo de publicaciones con no promotores asociados con longitud aproximadamente de 30 caracteres (que supera la cantidad de publicaciones con promotores para esa longitud) y de un grupo de publicaciones con promotores asociados con longitud aproximadamente de 80 caracteres (que supera la cantidad de publicaciones con no promotores para esa longitud). En otras palabras, las publicaciones más extensas (entre 75 y 100 caracteres) están en gran medida asociadas a promotores, y en esta diferencia radica la mayor media de longitud del título asociada a dicho grupo. A su vez, los precios en términos nominales presentan un comportamiento diferente entre clases. Los precios extremadamente bajos (quizás asociados a una calidad precaria) o demasiado elevados, presentan sobre todo no promotores, mientras que los precios intermedios cuentan con más promotores. En términos reales, los precios más bajos (normalizados por volumen) cuentan con más promotores mientras que los más altos más no promotores (Figura 19). Frente a variables de packaging como unidades_pack y volumen_ml, no hay grandes diferencias entre grupos de var_dep_3 aunque se logra distinguir una ligera concentración mayor de promotores para productos más voluminosos y de promotores para packs con menos unidades (Figura 19). En el caso de factor_proteccion, hay una ligera preferencia por productos con mayor protección frente a los rayos UV (Figura 20).

Figura 19. Diferencias entre promotores y no promotores
para el logaritmo de variables numéricas

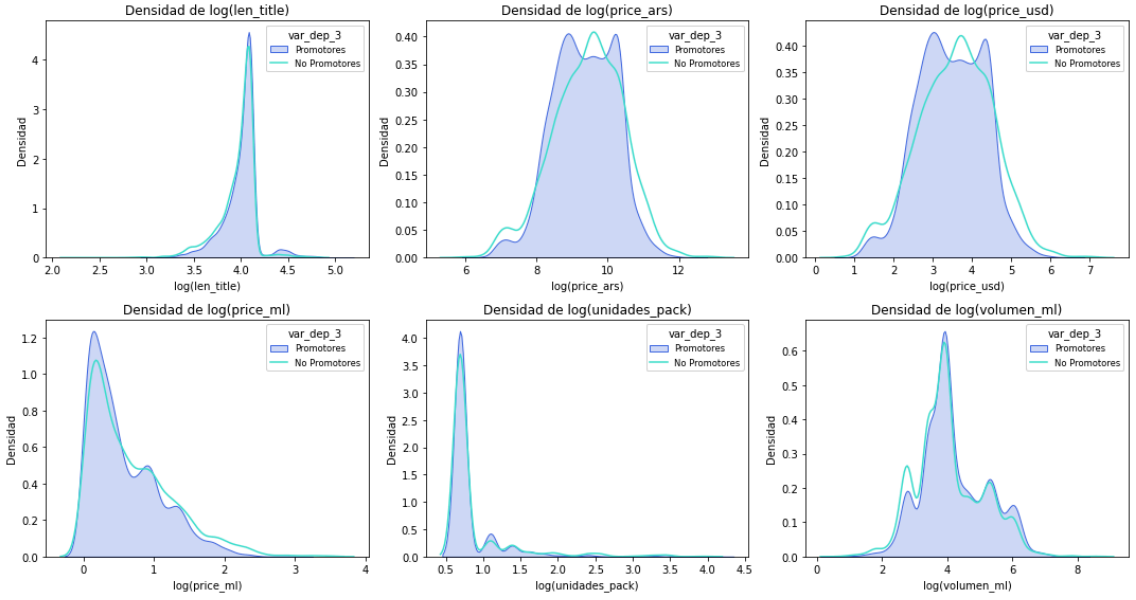
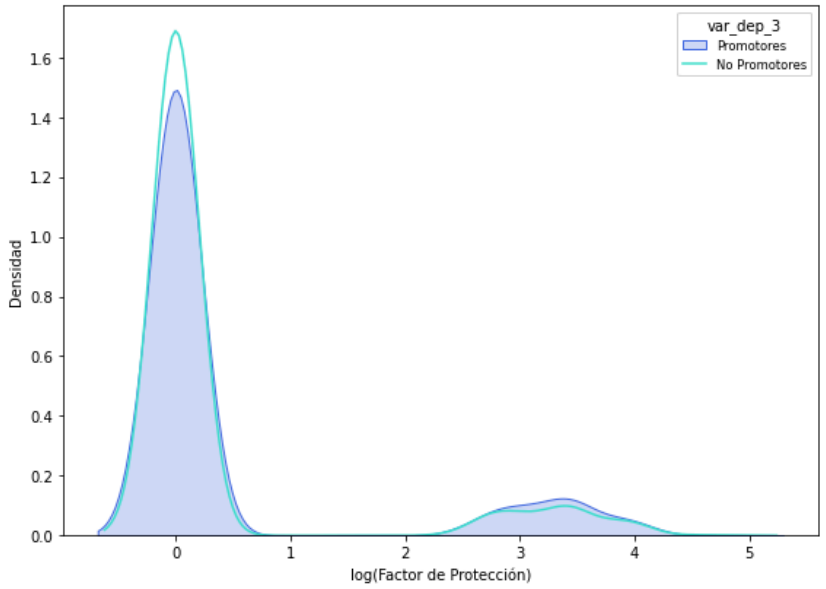


Figura 20. Diferencias entre promotores y no promotores
para el logaritmo de factor_proteccion



Entre las variables categóricas con más diferencia entre promotores y no promotores se encuentran (Figura 21 y 22): sustentable, es_vegano, libre_crueldad, gama_marca, gama_marca_nominal, gama_producto, gama_producto_nominal, marca, funciones, linea, formato_producto, formato_venta, tipo_piel y material_venta. Sorprendentemente, los productos caracterizados por ser libres de crueldad o veganos (la variable sustentable se forma por la unión entre esas dos variables independientes: si un producto cumple con una de los dos, es sustentable) poseen más no promotores. Entre las marcas con más adeptos se encuentra La Roche Posay, Eucerin, Dermaglós y Lidherma, todas con mucha trayectoria y renombre. Las más nuevas como ACF o Eximia aún cuentan con muchos no promotores. Las marcas más lujosas cuentan con más promotores que las masivas (gama_marca_nominal). Si bien se valoran productos con precios accesibles (gama_producto_nominal), la lealtad a empresas de mayor lujo también es un factor relevante en la decisión del consumidor. Así como cada marca presenta su grupo de consumidores fieles, cada línea de marcas específicas se comporta de la misma manera (Effaclar de La Roche Posay cuenta con muchos no promotores, pero Trilogía de ACF con muchos no promotores). En cuanto a la función de un producto, los hidratantes cuentan con más promotores, pero los productos de limpieza o anti-edad están más vinculados a no promotores. Esto podría deberse a que la audiencia de productos anti-edad es más sofisticada y exigente y espera efectos más notables. Las cremas hidratantes no suelen asociarse a cambios demasiado drásticos. Los productos de limpieza, por su parte, fácilmente pueden generar acné o brotes en caso de no ser el adecuado para cierta piel y esto podría dificultar la consolidación de seguidores. En la misma línea, el formato 'crema' para un producto (el cual usualmente está vinculado con hidratantes) tienen muchos promotores mientras que los 'serums' (que, en general, cumplen con funciones anti-edad), varios no promotores. A su vez, los envases estilo pote o pomos que suelen acompañar a las cremas presentan sobre todo promotores y los dispensers o goteros muy vinculados a serums, no promotores. La definición de audiencia de un producto a partir de un tipo de piel genérico como 'todo tipo de piel', suele tener más no promotores que los tipos de piel más específicos como 'grasa' o 'madura'. Una definición más precisa de audiencia de un producto puede derivar en un mejor match entre consumidor y oferta, permitiendo así ganar más simpatizantes. Con respecto al material del envase de productos, el plástico

presenta más no promotores, mientras que packagings reciclables como el vidrio, más promotores.

Figura 21. Diferencias entre promotores y no promotores
para variables con menos de tres categorías

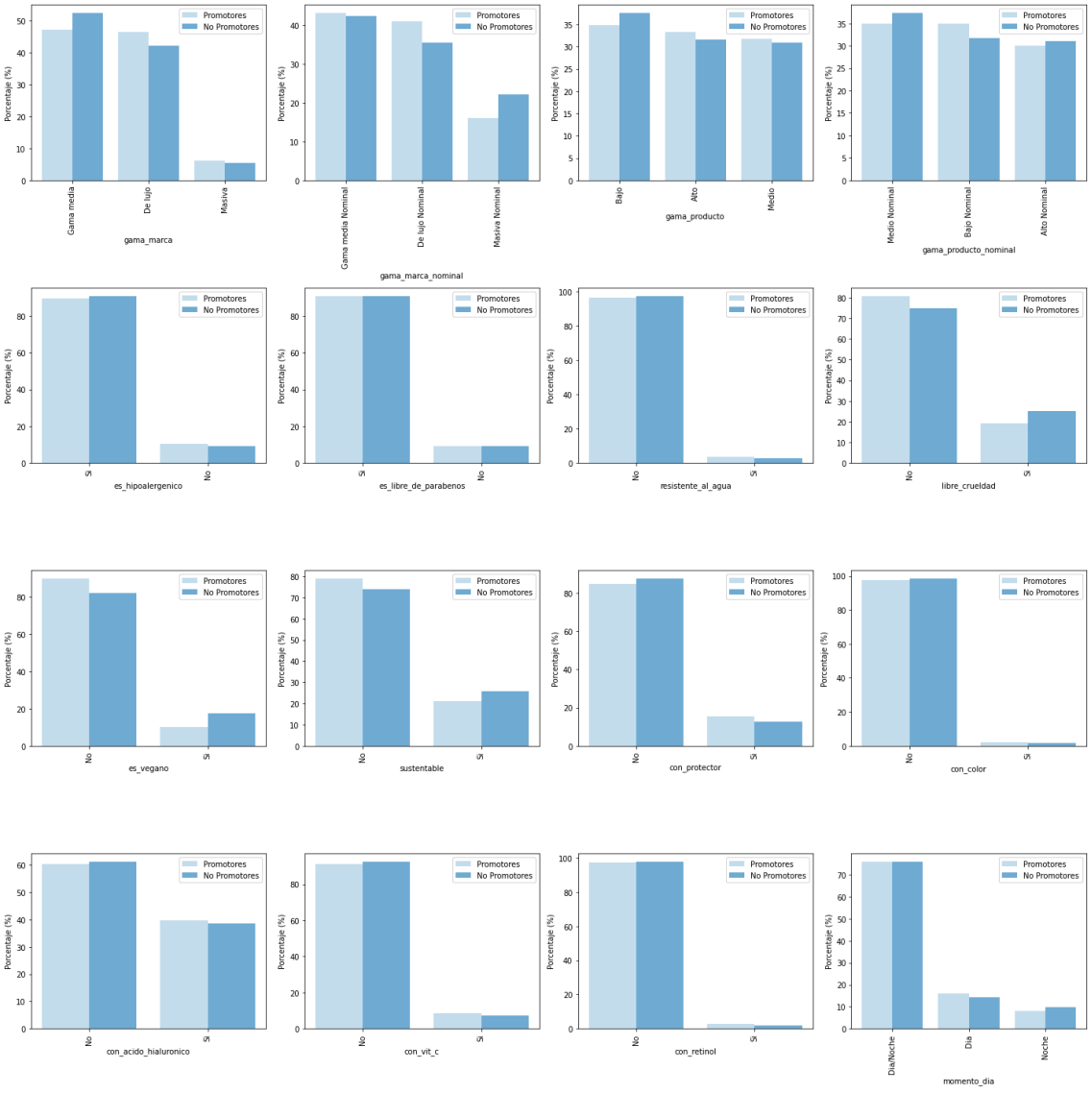
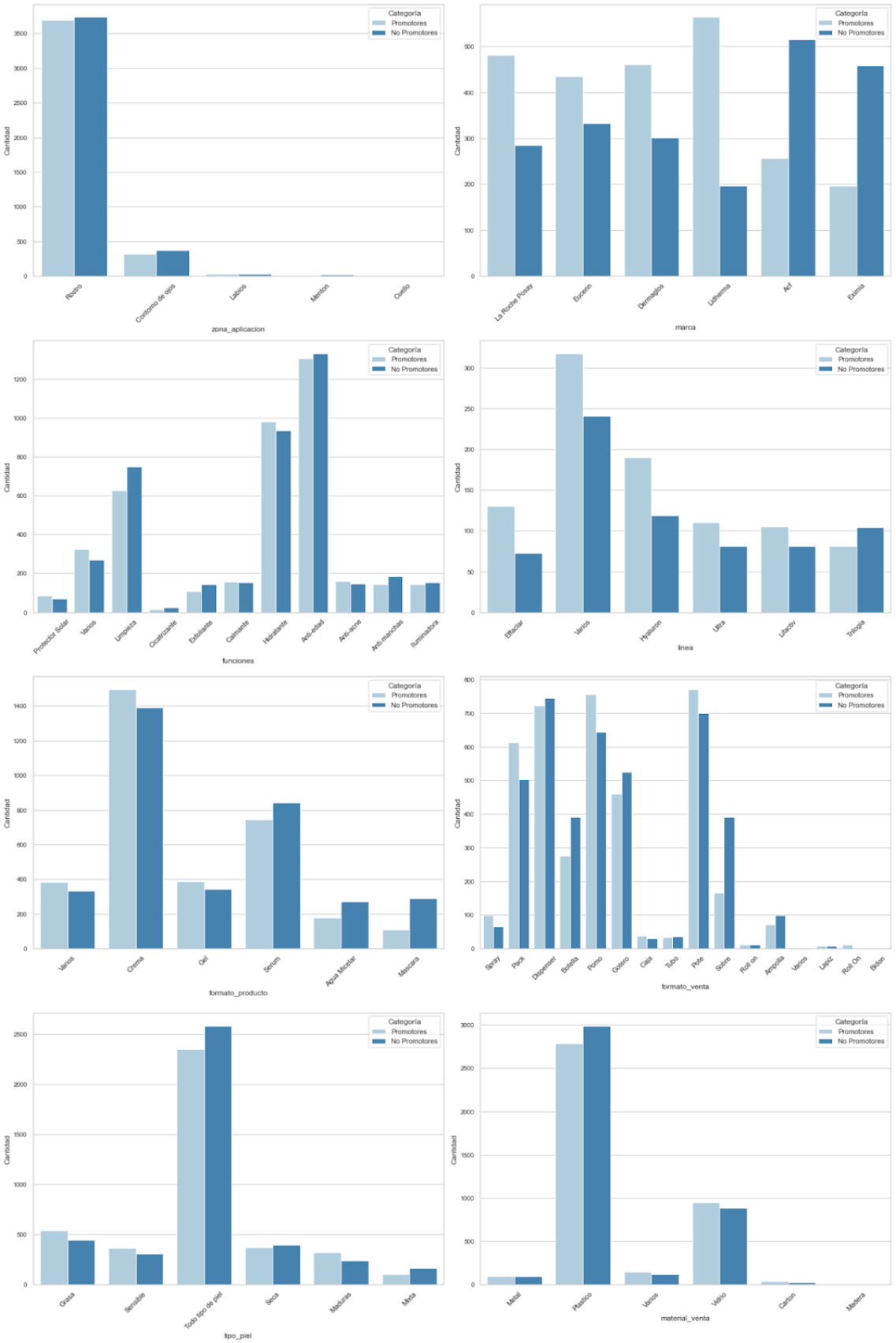


Figura 22. Diferencias entre promotores y no promotores
para variables con más de tres categorías



3. Metodología

3.1. Armado del conjunto de datos

Se definió el conjunto de entrenamiento y validación a partir de una división aleatoria de los datos, dejando así un 80% del total para el entrenamiento y el 20% restante para validación. Dado que la naturaleza de los datos en cuestión no depende de una secuencia temporal, no es necesario considerar un orden cronológico al conformar el conjunto para validación. Para corroborar que efectivamente el conjunto de validación presenta un comportamiento similar al de entrenamiento, se comparan las distribuciones de algunas variables en cada set en la Figura 23 y 24. De esta manera y al establecerse una semejanza notable entre ambos conjuntos de datos, se asume que lo mismo sucede para el resto de las variables.

Figura 23. Distribuciones de variables independientes numéricas

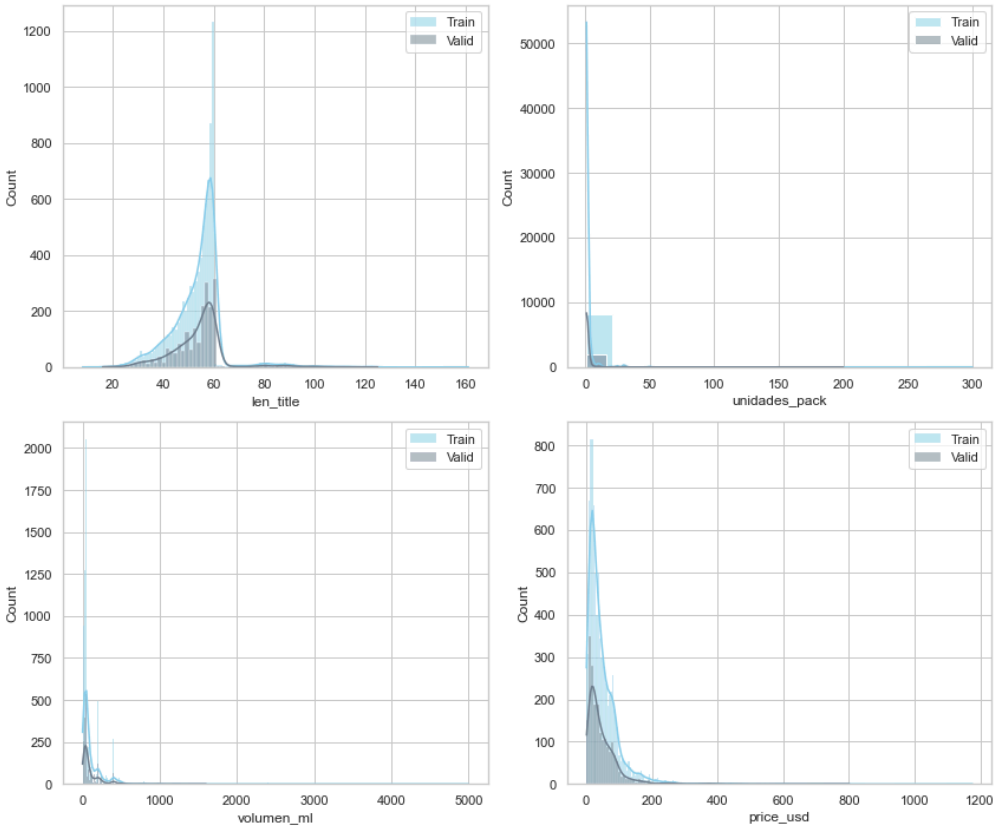


Figura 24. Distribuciones de variables independientes categóricas

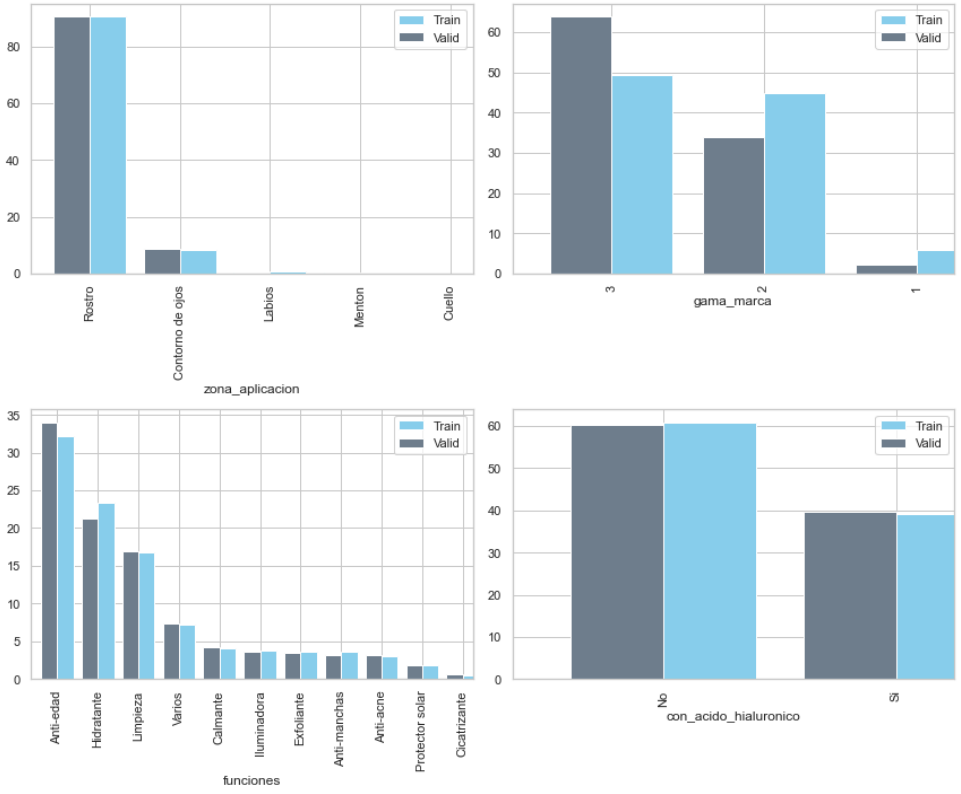
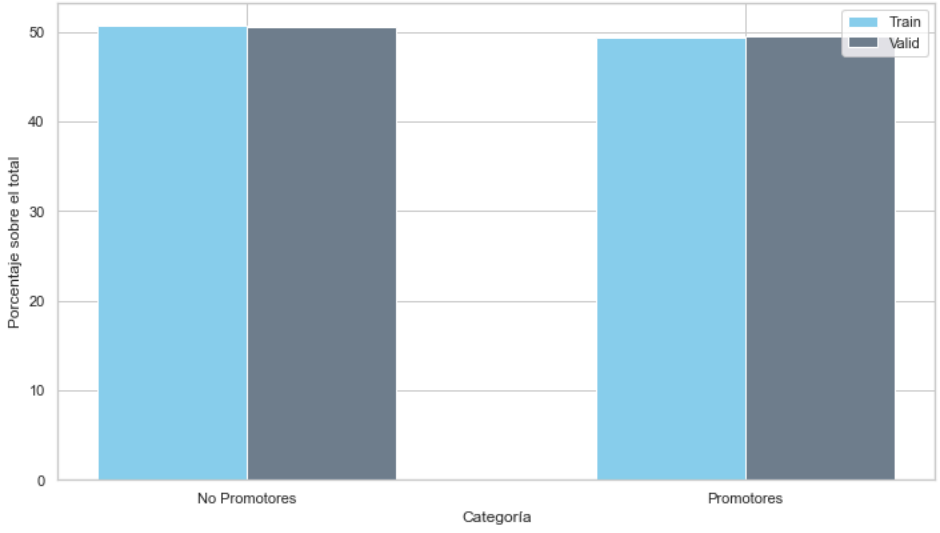


Figura 25. Distribuciones de variable dependiente



3.2. Baseline

Un modelo baseline, también conocido como modelo base, es un punto de referencia inicial utilizado en el desarrollo y evaluación de modelos más avanzados en aprendizaje automático y estadísticas. El mismo suele ser más básico o ingenuo en comparación con otros modelos más complejos que se están desarrollando. La idea detrás de un modelo baseline es proporcionar un punto de comparación para evaluar si los modelos más complejos tienen un desempeño significativamente mejor que una estrategia simple. Algunos ejemplos incluyen: modelos aleatorios (se elige la salida de forma

aleatoria y se utiliza para comparar si un modelo tiene un rendimiento significativamente mejor que simplemente elegir al azar), predicción constante (la predicción siempre es la misma; por ejemplo siempre predecir la clase mayoritaria en un problema de clasificación), modelo simple (puede ser un modelo lineal básico que no tiene en cuenta la complejidad del problema, pero sirve como punto de partida).

En el contexto de este estudio, la proporción de cada categoría que compone a la variable dependiente en el conjunto de entrenamiento está alrededor del 50% lo cual es análogo a una clasificación binaria basada en el azar. Este será el punto de partida para proporcionar un marco de referencia para evaluar si los modelos más complejos son realmente efectivos y están capturando patrones significativos en los datos. Si un modelo avanzado no supera el rendimiento del modelo baseline de manera significativa, puede indicar que algo está mal con la complejidad del modelo o que los datos no tienen suficiente información para mejorar la predicción.

3.3. Modelos candidatos

En una primera instancia, se utilizará un modelo básico de árboles de decisión, optimizando hiperparámetros. El método DecisionTreeClassifier en scikit-learn es un clasificador que crea un árbol de decisión recursivamente dividiendo el conjunto de datos en subconjuntos más puros. En cada paso de la construcción del árbol, se selecciona la característica que mejor separa las clases en los subconjuntos resultantes. Esto se hace maximizando alguna medida de pureza, como la ganancia de información o el índice Gini. La construcción del árbol continúa hasta que se alcanza algún criterio de parada, como la profundidad máxima del árbol, el número mínimo de muestras requeridas para dividir un nodo o el número mínimo de muestras requeridas en un nodo hoja. Una vez que se ha construido el árbol, se utiliza para predecir la clase de nuevas muestras. Esto se hace siguiendo las divisiones del árbol: cada muestra pasa por el árbol, se desplaza de nodo en nodo según los valores de sus características, hasta que alcanza un nodo hoja, donde se asigna la clase mayoritaria de las muestras de entrenamiento que llegaron a ese nodo. Se optimizarán hiperparámetros a partir de Random Search en conjunto con una validación cruzada estratificada.

Por otra parte, se probarán algoritmos de ensamble de árboles, optimizando hiperparámetros. Un método de ensamble es un enfoque que combina muchos modelos simples para obtener un modelo único y potencialmente muy poderoso. Estos modelos de bloques de construcción simples a veces se conocen como aprendices débiles, ya que pueden conducir a predicciones triviales por sí solas. Los árboles simples pueden ser poco robustos ya que un pequeño cambio en los datos puede causar un gran cambio en el valor estimado. Sin embargo, al agregar muchos árboles de decisión, el aumento del rendimiento predictivo de los árboles puede ser sustancialmente mejorado (James et al, 2023: 342). Se tendrá en cuenta, por un lado, algoritmos de boosting y, por el otro, de bagging. En ambos casos, se optimizarán hiperparámetros a partir de Random Search en conjunto con una validación cruzada estratificada de manera tal de asegurar de que el proceso de búsqueda se realice de manera imparcial y equitativa en todos los pliegues de validación.

Boosting es un método de ensamblado que construye una secuencia de modelos de árboles de decisión poco profundos. Cada nuevo árbol se enfoca en los errores cometidos por los árboles anteriores, lo que conduce a una reducción del error por sesgo y una mejora incremental del rendimiento del modelo. Al construir la secuencia utilizando árboles poco profundos, se evita caer en overfitting.

Bagging es un método de ensamblado que se basa en la combinación de múltiples modelos de árboles de decisión profundos entrenados en subconjuntos aleatorios del conjunto de datos de entrenamiento. Estos subconjuntos se obtienen mediante muestreo con reemplazo, lo que permite que algunas muestras aparezcan múltiples veces y otras no aparezcan en absoluto.

XGBClassifier se basa en el algoritmo de boosting de gradiente estándar. Funciona mediante la construcción de una serie de árboles de decisión débiles y combinándolos para formar un modelo predictivo más fuerte. Cada nuevo modelo trabaja sobre el output del anterior utilizando el gradiente de la loss function como función de las predicciones. En el contexto de los modelos de boosting, identificar la relevancia de características implica comprender qué variables contribuyen más significativamente a la reducción del error del modelo durante el proceso de entrenamiento (Chen et al, 2016: 786). XGBClassifier puede manejar valores NaN de manera nativa. Durante el proceso de entrenamiento, XGBoost utiliza una técnica llamada 'separación de árboles' para dividir los nodos de los árboles en función de si los valores faltantes son menores o mayores que un umbral específico (Chen et al, 2016: 787). De esta manera, los valores faltantes se distribuyen de manera equitativa a lo largo de los nodos del árbol durante la construcción del modelo. Para la predicción, XGBoost también puede manejar valores faltantes en los datos de entrada, asignando automáticamente los valores faltantes a las hojas correspondientes del árbol basándose en las divisiones realizadas durante el entrenamiento.

Por otra parte, HistGradientBoostingClassifier utiliza un algoritmo de boosting de gradiente basado en histogramas. Este algoritmo aprovecha el histograma de las características para realizar divisiones eficientes durante la construcción del árbol, lo que lo hace más rápido y eficiente en términos de memoria en comparación con los métodos tradicionales. El HistGradientBoostingClassifier es capaz de manejar datos faltantes de manera nativa sin necesidad de imputación previa. Durante el entrenamiento, el algoritmo trata los valores faltantes como un valor separado y los maneja de manera específica durante el proceso de división de nodos del árbol de decisión. El algoritmo clasifica los valores faltantes en función de su dirección en cada nodo del árbol de decisión durante el proceso de entrenamiento. Durante la predicción, si un nuevo ejemplo tiene un valor faltante en una característica utilizada para tomar una decisión en un nodo, el algoritmo seguirá la ruta correspondiente al hijo izquierdo o derecho del nodo según la clasificación de valores faltantes que aprendió durante el entrenamiento.

Finalmente, RandomForestClassifier crea un conjunto de árboles de decisión durante el entrenamiento. Estos árboles se construyen utilizando un subconjunto aleatorio de las muestras de entrenamiento (con reemplazo) y un subconjunto aleatorio de las variables independientes disponibles en cada división del árbol. Esta aleatorización ayuda a decorrelacionar los árboles y aumentar su diversidad, lo que mejora la capacidad de generalización del modelo. Una vez que se construyen todos los árboles en el conjunto, el RandomForestClassifier realiza predicciones combinando las predicciones individuales de cada árbol. Para la clasificación, se utiliza un esquema de votación, donde cada árbol vota por la clase más frecuente en su hoja de decisión. Por otra parte, este modelo proporciona una medida de la importancia de las características que se calcula observando cuánto se reduce la impureza (Gini o Entropía) en los árboles de decisión debido a las divisiones de cada característica. Las características que más reducen la impureza (ponderadas por el número de muestras) se consideran más importantes para la clasificación. Antes de aplicar RandomForestClassifier, se prueban distintos métodos de imputación para el tratamiento

de datos faltantes ya que, al igual que algunos otros algoritmos, no maneja nativamente datos faltantes (NaNs) en las características.

3.4. Transformaciones y selección de variables

Para aplicar modelos de árboles, es necesario transformar variables a un formato numérico. Las variables existentes pueden agruparse en dos casuísticas: categóricas y numéricas.

Variables categóricas

Entre los datos categóricos, se distinguen características ordinales y nominales. Los rasgos ordinales pueden entenderse como valores que pueden ordenarse. Por ejemplo, el talle de una remera sería un rasgo ordinal, porque podemos definir un orden: XL > L > M. Por el contrario, los rasgos nominales no implican ningún orden y, siguiendo con el ejemplo anterior, podríamos pensar que el color de la camiseta es un rasgo nominal (Raschka et al, 2019:115). Para asegurarnos de que el algoritmo de aprendizaje interpreta correctamente las características ordinales, tenemos que convertir los valores categóricos de cadena en números enteros (Raschka et al, 2019:117). En el caso de características nominales, una solución habitual es utilizar una técnica denominada one-hot-encoding. El proceso de one-hot encoding consiste en crear una nueva columna binaria para cada categoría única presente en la variable categórica original. Cada columna binaria representa una categoría y toma el valor 1 si la observación pertenece a esa categoría, y 0 en caso contrario. Esto se realiza para todas las categorías únicas, de manera que cada observación esté representada por un vector de ceros y unos, donde hay un 1 en la posición correspondiente a la categoría a la que pertenece y ceros en todas las demás posiciones (Raschka et al, 2019:119).

Entre las variables categóricas ordinales se encuentran gama_marca, gama_marca_nominal, gama_producto, gama_producto_nominal. Para respetar el concepto de orden en estos casos y no perder esa relación entre categorías, estas variables serán convertidas a números subsiguientes. En el caso de gama_marca, se mapearon los productos ‘De lujo’ al número 3, ‘Gama media’ al 2, y ‘Masiva’ al 1. En el caso de gama_marca_nominal, los productos ‘De lujo nominal’ se mapearon al 3, ‘Gama media’ al 2, y ‘Masiva nominal’ al 1. Para gama_producto, se mapeó ‘Alto’ al 3, ‘Medio’ al 2 y ‘Masiva’ al 1. Para gama_producto_nominal, se mapeó ‘Alto nominal’ al 3, ‘Medio nominal’ al 2 y ‘Masiva nominal’ al 1.

Entre las variables categóricas nominales se encuentran zona_aplicacion, marca, funciones, testeado_dermatologicamente, libre_de_fragancia, es_comedogenico, es_hipoalergenico, es_libre_de_parabenos, condicion, linea, titulo, formato_producto, formato_venta, tipo_piel, resistente_al_agua, libre_crueldad, es_vegano, sustentable, con_protector, con_color, con_acido_hialuronico, con_vit_c, con_retinol, momento_dia, material_venta. Tal como fue comentado previamente en la sección 2, testeado_dermatologicamente y condicion serán excluidas debido a una distribución muy sesgada hacia una categoría puntual. Por otra parte, libre_de_fragancia y es_comedogenico también serán excluidas debido a un alto número de datos faltantes. Para el resto, se aplicará one-hot-encoding en todos los casos. Por ejemplo, para la variable 'con_protector' con las categorías 'Si' y 'No', el proceso de one-hot encoding crearía dos nuevas columnas llamadas 'con_protector_si' y 'con_protector_no'. Para una observación que tenga el valor 'Si' en la variable original, el vector resultante sería [1, 0], indicando que pertenece a la categoría 'Si' y no a la otra. La variables texto ‘Title’, puede ser contemplada como una

categórica nominal, donde cada fila representa una categoría única. Aplicar one-hot-encoding en este caso derivaría en un número muy elevado de nuevas columnas debido a la ausencia de valores duplicados. Esto generaría un esfuerzo computacional excesivo al momento de entrenar el modelo. Teniendo en cuenta que el título ofrece información sobre características contenidas en otros atributos (por ejemplo, el volumen, la marca o el tipo de piel al cual está destinado el producto), esta variable será descartada y solamente se guardará información sobre el largo de caracteres del título para cada publicación. De esta manera, la variable previamente en formato texto pasa a un formato numérico. El largo del título representa el detalle o el grado de descripción disponible para un producto. El resto de la información contenida en el título guarda una relación estrecha con otras columnas, por lo que concentra información redundante. En la Tabla 1 se ofrecen algunos ejemplos:

Tabla 1. Ejemplos de títulos e interacción con otras variables

Título	Otras variables que contienen información presente en el título
Protector Solar La Roche-posay Anthelios Fps 50 Bruma De Rostro En Bruma De 75 Ml	Funciones: Protector solar Marca: La Roche-posay Línea: Anthelios Con_protector: Si Formato_producto: Bruma Volumen_ml: 75 Factor_proteccion: 50 Zona_aplicacion: Rostro Momento_dia: Dia
Locion Desmaquillante Dual Ponds Bio-hydratante 200ml	Funciones: Limpieza Marca: Ponds Línea: Bio Hydratante Formato_producto: Loción Volumen_ml: 200
Combo Solar Vichy La Roche Posay Fps 50 Corporal + Facial Ts	Marca: Varios Funciones: Varios Con_protector: Si Factor_proteccion: 50

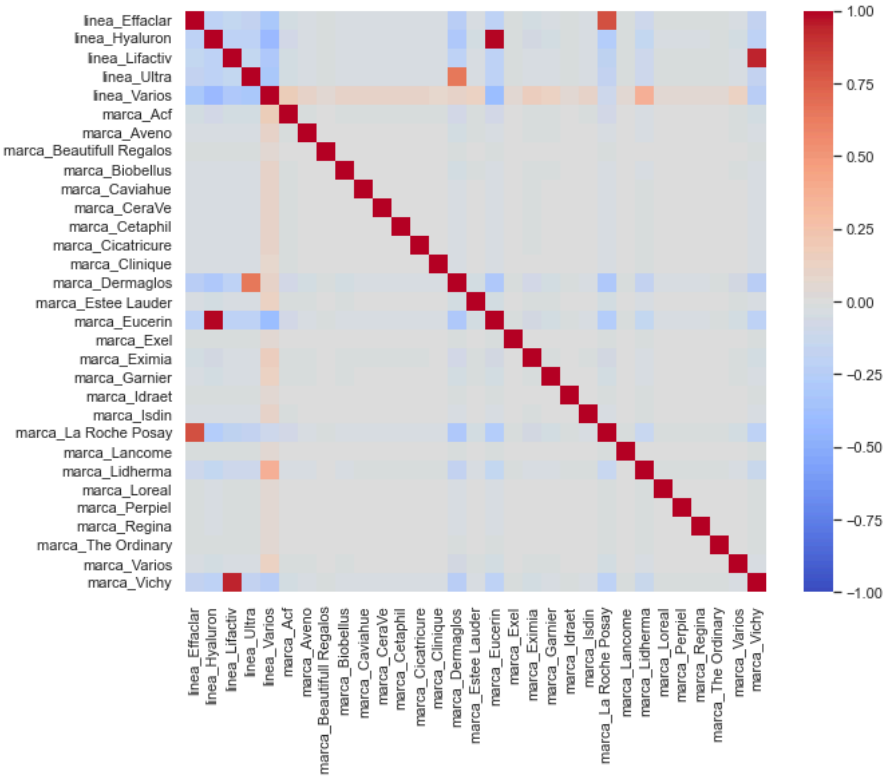
De manera similar, la variable línea cuenta con un número muy elevado de categorías y también será excluida del análisis. Dicha variable posee una alta correlación con otras variables como tipo_piel, marca, funciones, con_acido_hialuronico, con_vit_c y con_retinol. De acuerdo al European Business School[6], una línea representa una colección de productos ofrecidos al mercado bajo una misma marca. Los productos que componen una línea son similares entre ellos y se dirigen a un mismo sector de consumidores. Al desarrollar una línea de productos, una marca puede ampliar su oferta y llegar a nuevos segmentos de mercado. A su vez, si una marca ya es conocida en el mercado, puede aprovechar su reputación para introducir nuevos productos. Al tener una base sólida de

6 https://www.ceupe.com/blog/linea-de-productos.html#:~:text=Una%20l%C3%ADnea%20de%20productos%20es,un%20mismo%20sector%20de%20consumidores

clientes fieles, la marca puede tener una mejor oportunidad de éxito en el lanzamiento de nuevos productos.

Si se coloca el foco en las 5 líneas más frecuentes (Effaclar, Hyaluron, Liftactiv, Ultra y Varios) a modo de analizar las interacciones entre 'linea' y otras variables, se logra reflejar cómo dicha variable es capturada por otras. En primer lugar, la variable 'marca' podrá agrupar a varias líneas bajo su nombre, pero dichas líneas no se asociarán a otras marcas (Figura 26). Por ejemplo, Effaclar está vinculada a La Roche Posay. La Roche Posay podrá tener otras líneas bajo su nombre, pero no habrán productos Effaclar ligados a marcas diferentes a La Roche Posay. A su vez, Hyaluron está ligada a Eucerin y Liftactiv a Vichy. Es importante resaltar la casuística multi-línea que, al involucrar una bolsa genérica, sí podrá darse en varias marcas en simultáneo. Cualquier set de una marca que involucre productos de más de una línea, será catalogado como 'Varios' bajo la columna 'linea'.

Figura 26. Matriz de correlación entre línea y marca



En segundo lugar, una línea de un producto está vinculada a un segmento específico y esto puede determinarse por medio de la función del producto, el tipo de piel a quién está destinado, o los activos que incluye en su fabricación (Figura 27 y 28). Los productos Effaclar están destinados, sobre todo, a pieles grasas. Entre sus principales funciones están los tratamientos anti-acné, aunque también cuenta con algunos exfoliantes y productos de limpieza especiales para pieles grasas. Hyaluron está diseñado para pieles maduras debido a su función anti-edad, la cual es construida a partir de la incorporación de ácido hialurónico a su fórmula. La línea Liftactiv de Vichy[7] apunta a combatir el envejecimiento mediante tres pilares: recuperar la luminosidad, la elasticidad y la

7

https://www.vogue.mx/belleza/articulo/vichy-liftactiv-super-serums-la-nueva-linea-de-sueros-para-el-rostro

uniformidad. Para recuperar la luminosidad, ofrece productos con vitamina C. Para la uniformidad (anti-manchas), ofrece productos con niacinamida. Para fomentar la elasticidad, productos con hialurónico. Detrás de la línea Ultra, por su parte, se busca priorizar la hidratación de la piel a partir de hialurónico para así crear un aspecto jovial.

Figura 27. Matriz de correlación entre línea, tipos de piel y activos

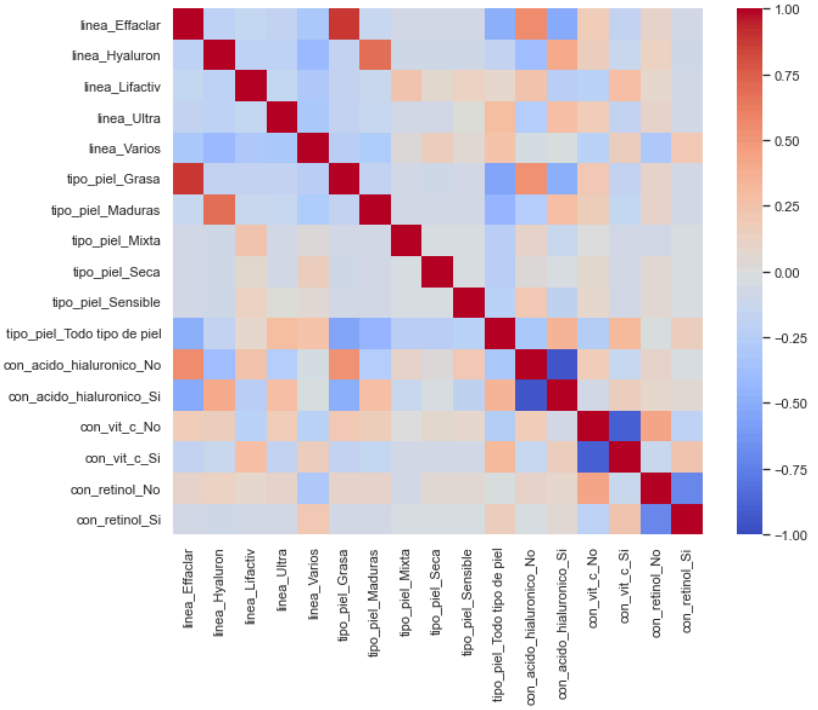
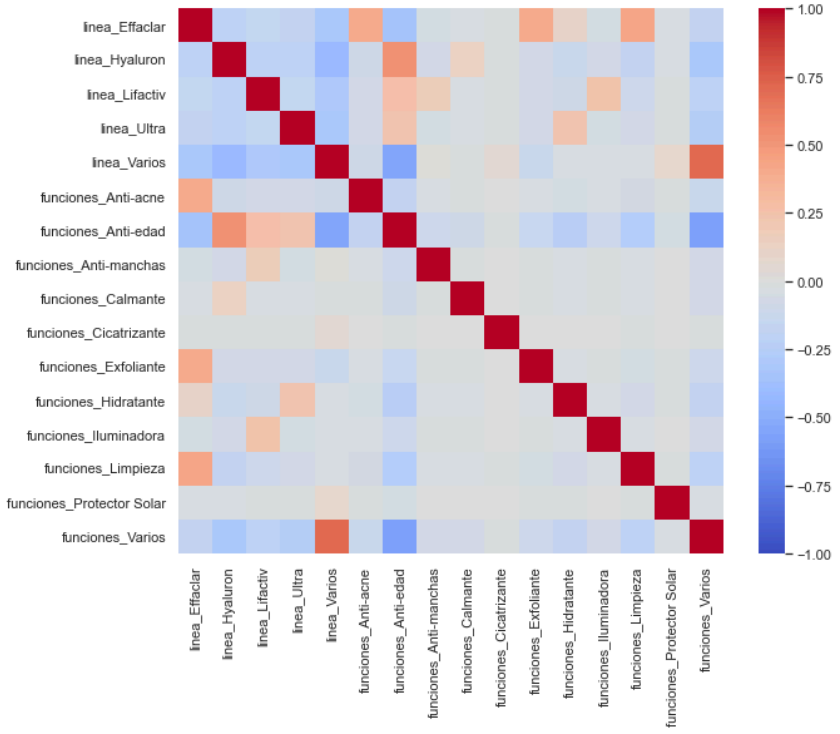
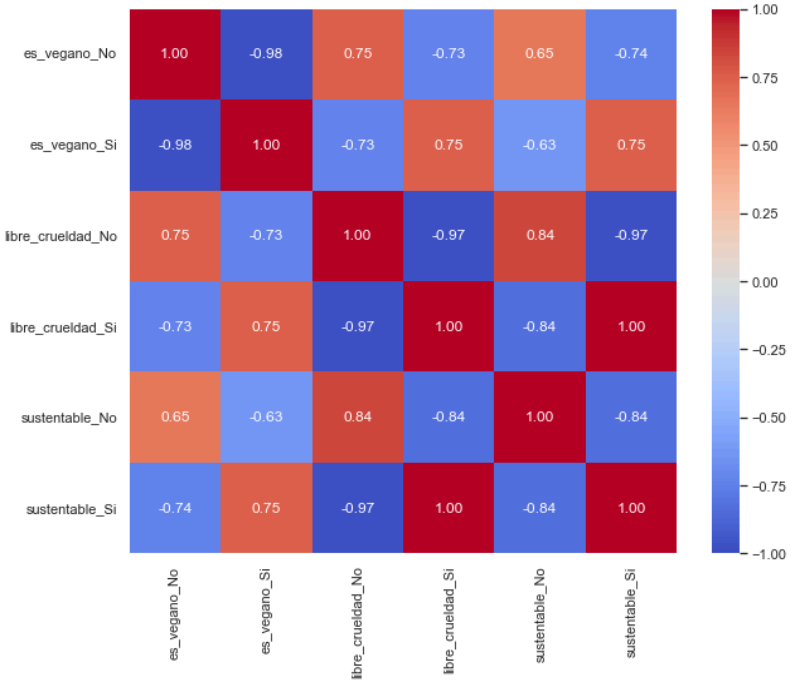


Figura 28. Matriz de correlación entre línea y funciones



Por otra parte, 'libre_crueldad' y 'es_vegano' serán descartadas debido a presentar una alta correlación con la variable 'sustentable', que es la unión entre las dos primeras (Figura 29). Un producto calificado como 'vegano' no contiene ingredientes animales ni derivados de estos (colágeno, miel, etc.). Esta etiqueta no garantiza que el producto en cuestión, a su vez, no haya sido testeado en animales. Un producto libre de crueldad asegura que, en ninguna de las fases del desarrollo de un producto, se haya testeado sobre animales. Esto no significa que no contenga ingredientes animales. En general, los términos 'vegano' y 'libre de crueldad' suelen solaparse mucho en el mercado y los consumidores suelen tomarlos como sinónimos. Por este motivo, se diferencian aquellos productos que hacen mención a alguno de estos dos atributos a partir de la variable 'sustentable', en lugar de hacer referencia estrictamente a uno u otro.

Figura 29. Matriz de correlación entre vegano, libre de crueldad y sustentable



Variables numéricas

Entre las variables numéricas, se encuentran len_title, precio_ars, precio_usd, precio_ml, unidades_pack, volumen_ml y factor_proteccion. Al trabajar con modelos de árboles, no es necesario aplicar un método para escalar variables ya que dichos algoritmos son invariantes al escalado de características (Raschka et al, 2019:124). Tal como fue indicado en la sección 2, hay algunos valores extremos para las variables numéricas. Sin embargo, ninguno supone una caracterización errónea de un producto (por ejemplo, un producto etiquetado con 5000 ml de volumen, efectivamente tiene ese volumen a pesar de ser un valor extremo para la distribución general de esa variable). Se busca entender también qué ocurre con estos casos extremos: productos con volumen excesivo o, por el contrario, muy por debajo de lo usual, o bien productos con costos sumamente altos o sorprendentemente bajos. Por este motivo, no se aplicarán transformaciones adicionales a variables numéricas. Se descartará la variable precio_ars que posee una relación lineal con precio_usd y se dejará precio_usd a modo de una interpretación más sencilla a través del tiempo dado que el peso argentino se encuentra en un momento de fuerte pérdida de

valor. A su vez, se mantendrá la variable precio_ml para tener una suerte de normalización entre productos con diferente volumen.

Sumados a los criterios de selección de variables hasta ahora mencionados, se considerarán dos métodos adicionales que derivan del proceso de entrenamiento de modelos: importancia de características y permutación. El primer método será utilizado en todos los modelos candidatos excepto para Histogram Boosting Gradient Classifier, dónde dicha técnica no está disponible y se utilizará la de permutación en su reemplazo. Si bien en corridas iniciales se considerarán todas las variables explicativas mencionadas, se refinará luego dicha selección a partir de ambas técnicas. Las características con menor relevancia para el rendimiento del modelo pueden generar ruido para la tarea de predicción. Esto puede confundir al modelo y llevar a un rendimiento deficiente cuando se incluyen en el proceso de entrenamiento.

Por un lado, la técnica de la importancia de características asigna un puntaje a cada variable según su contribución al rendimiento del modelo. Al observar qué características tienen una mayor importancia relativa, se puede priorizar la optimización de estas variables. Este enfoque permite una mejor comprensión de cómo las características individuales afectan la capacidad del modelo para generalizar y predecir con precisión.

Por otro lado, la técnica de permutación[8] funciona perturbando aleatoriamente el valor de una característica y observando cómo afecta al rendimiento del modelo. Implica mezclar aleatoriamente los valores de una sola característica y observar la degradación resultante de la puntuación del modelo. Al romper la relación entre la característica y el objetivo, se determina en qué medida el modelo depende de esa característica en particular.

3.5. Optimización de hiperparámetros

Se utiliza RandomizedSearchCV para la optimización de hiperparámetros. Este método realiza una búsqueda aleatoria en el espacio de hiperparámetros especificado para encontrar la combinación óptima que maximiza o minimiza una métrica de evaluación. En este caso, se busca maximizar el área bajo la curva ROC (AUC-ROC). Se definen los hiperparámetros a explorar y se establece un rango para cada uno. Por ejemplo, max_iter, learning_rate, max_depth, etc., tienen rangos específicos dentro de los cuales se realizará la búsqueda aleatoria. Se utiliza validación cruzada estratificada (StratifiedKFold) para dividir el conjunto de entrenamiento en varios subconjuntos de entrenamiento y validación. Esto ayuda a evaluar el rendimiento del modelo de forma más robusta, al asegurarse de que cada subconjunto tenga una distribución similar de clases.

4. Resultados

4.1. Selección del modelo

Para contrastar el desempeño entre modelos, se utilizará como métrica el área bajo la curva ROC. Esta métrica es utilizada comúnmente en problemas de clasificación binaria. La curva ROC es una representación gráfica de la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes umbrales de clasificación. Es decir, cuantifica la capacidad de discriminación de un modelo de clasificación. En otras palabras, mide la

[8] https://scikit-learn.org/stable/modules/permutation_importance.html

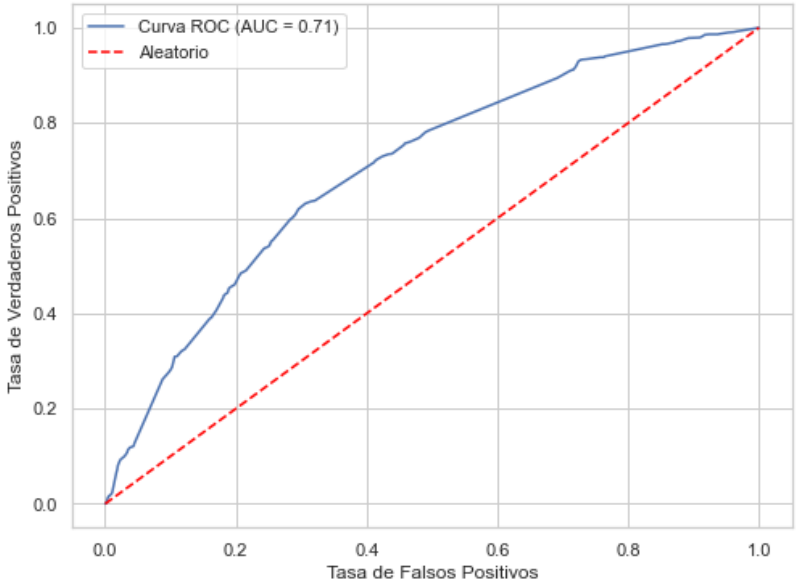
probabilidad de que el modelo clasifique correctamente una instancia positiva (clase positiva) por encima de una instancia negativa (clase negativa). Un AUC-ROC de 1 indica un modelo que puede distinguir perfectamente entre la clase positiva y negativa. Un AUC-ROC de 0.5 indica un modelo que no tiene capacidad de discriminación y realiza clasificaciones aleatorias.

En primer lugar, se corre un modelo simple de árbol de decisión, optimizando hiperparámetros[9], pero sin ninguna selección de variables. Los hiperparámetros para DecisionTreeClassifier se optimizan entre los siguientes intervalos:

1. criterion: 'gini' (criterio de impureza de Gini).
2. splitter: 'best' (elige la mejor división) por defecto.
3. max_depth: randint(3, 15).
4. min_samples_split: randint(2, 20).
5. min_samples_leaf: randint(1, 20).
6. min_weight_fraction_leaf: 0.0 (fracción mínima del peso total de las muestras requeridas en un nodo hoja).
7. max_features: ['sqrt', 'log2', None] (número de características a considerar para la mejor división).
8. max_leaf_nodes: None (sin límite en el número de nodos hoja).
9. random_state: None (semilla aleatoria).

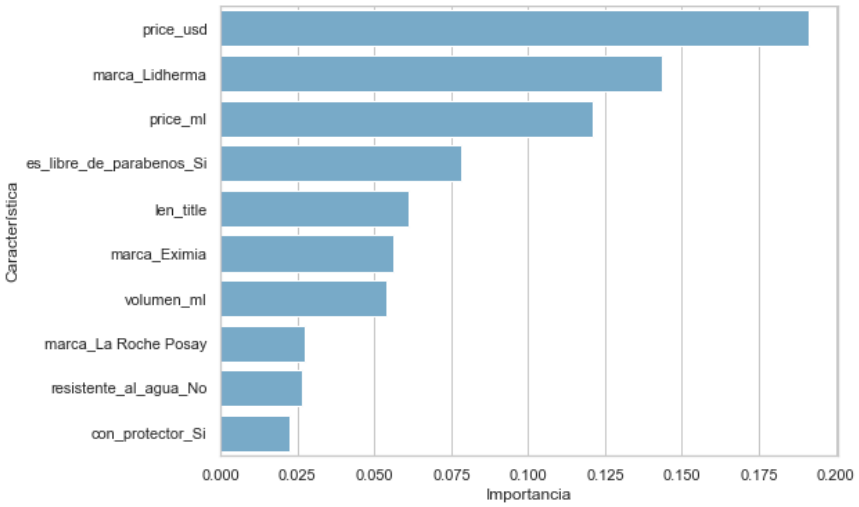
Como resultado, se obtiene un área bajo la curva de 0.7055, por encima del 0.50 que representa el criterio de aleatoriedad (Figura 30). Entre las características más relevantes para el modelo se destacan marcas y atributos de tamaño y precio (Figura 31).

Figura 30. Área bajo la curva de ROC para árbol de decisión



[9] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 3.

Figura 31. Feature importance para árbol de decisión



Pasando luego a un modelo más complejo de ensamble, se implementa XG Boost. Se optimizan hiperparámetros y también se hace un filtro de características según el feature importance calculado a partir del entrenamiento del modelo. La optimización de hiperparámetros[10] se realiza mediante la técnica de búsqueda aleatoria (RandomizedSearchCV) en el espacio de búsqueda definido por los siguientes parámetros:

1. n_estimator: Número de árboles en el modelo XGBoost. Se selecciona aleatoriamente un valor entre 100 y 300.
2. learning_rate: Tasa de aprendizaje del modelo. Se selecciona aleatoriamente un valor entre 0.01 y 0.3.
3. max_depth: Profundidad máxima de cada árbol. Se selecciona aleatoriamente un valor entre 3 y 15.
4. min_child_weight: Peso mínimo requerido para crear un nuevo nodo en el árbol. Se selecciona aleatoriamente un valor entre 1 y 20.
5. gamma: Parámetro de regularización para controlar la complejidad del modelo. Se selecciona aleatoriamente un valor entre 0.0 y 0.2.
6. subsample: Proporción de muestras utilizadas para entrenar cada árbol. Se selecciona aleatoriamente un valor entre 0.5 y 1.0.
7. colsample_bytree: Proporción de columnas (características) utilizadas para entrenar cada árbol. Se selecciona aleatoriamente un valor entre 0.5 y 1.0.
8. reg_lambda: Parámetro de regularización L2 (Ridge). Se selecciona aleatoriamente un valor entre 0.0 y 1.0.

La función `RandomizedSearchCV` realiza una búsqueda aleatoria en este espacio de hiperparámetros durante 100 iteraciones (definidas por n_iter = 100) utilizando validación cruzada estratificada con 5 divisiones (definidas por cv = StratifiedKFold(n_splits = 5, shuffle = True, random_state = 42)). Luego, se selecciona el mejor modelo encontrado según la métrica de evaluación definida, que en este caso es el área bajo la curva ROC (ROC AUC).

Sin realizar ningún tipo de selección de variables, el área bajo la curva ROC es de 0.7511. Sin embargo, al recurrir al cálculo de feature importance asociado al algoritmo de XG Boost, se logra observar que varias de las variables explicativas tienen una contribución

[10] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 4.

nula al modelo[11]. Para XGBoost, la importancia de una característica se basa en cómo contribuye a la ganancia de información en cada árbol de decisión construido por el algoritmo de boosting. Este cálculo se realiza internamente durante el entrenamiento del modelo. Una vez definida la importancia de cada variable, se excluyen aquellas cuya contribución al modelo fue cero. De esta manera, se logra simplificar el modelo de 247 variables a 87. Se vuelve a optimizar hiperparámetros[12] y entrenar el modelo esta vez con un conjunto de características filtradas por su relevancia superior al umbral mencionado. Esto permite potencialmente mejorar su rendimiento al eliminar características menos relevantes. Realizando estos pasos, la métrica de AUC crece a 0.758.

El tercer modelo candidato es Random Forest. En esta instancia no solamente se optimizarán hiperparámetros y se filtrarán variables explicativas, sino que también se imputarán datos faltantes. En una primera instancia, se imputará la media. En una segunda instancia, la mediana. Por último, se imputará la moda. En los tres casos, se optimizarán hiperparámetros mediante Random Search:

1. n_estimators: Número de árboles en el bosque. Se selecciona aleatoriamente un valor entre 100 y 300.
2. max_depth: Profundidad máxima de los árboles. Se selecciona aleatoriamente un valor entre 3 y 15.
3. min_samples_split: Número mínimo de muestras requeridas para dividir un nodo interno. Se selecciona aleatoriamente un valor entre 2 y 20.
4. min_samples_leaf: Número mínimo de muestras requeridas en un nodo hoja. Se selecciona aleatoriamente un valor entre 1 y 20.
5. max_features: Número máximo de características a considerar en cada split. Se selecciona aleatoriamente un valor entre 0.1 y 0.9.
6. bootstrap: Si se debe utilizar bootstrap samples para construir árboles. Se alterna entre True y False.

Estos hiperparámetros se definen con distribuciones específicas (randint para enteros y uniform para floats) que RandomizedSearchCV utilizará para explorar diferentes combinaciones de hiperparámetros durante la búsqueda. Se utiliza validación cruzada estratificada (`StratifiedKFold`) con 5 pliegues para realizar una búsqueda aleatoria en el espacio de hiperparámetros definido para encontrar la mejor combinación de que optimiza una métrica de rendimiento (en este caso, el área bajo la curva ROC).

En el caso del modelo que utilizó la media como método de imputación, el AUC fue de 0.7450[13]. Para el modelo que utilizó la mediana, 0.7526[14]. Para el que utilizó la moda, fue de 0.7470[15]. Una vez seleccionado el mejor método de imputación (la mediana), se pasa a dilucidar nuevamente las características más relevantes. De esta manera, se filtran aquellas con contribución nula[16] (de 247 variables, se pasan a tener 143) y se vuelve a optimizar hiperparámetros[17] y entrenar el modelo. Como resultado, el AUC incrementa a 0.7533.

Por último, se corre el modelo HistGradientBoostingClassifier. A partir de Random Search y validación cruzada, se optimizan los siguientes hiperparámetros[18]:

[11] Detalle de resultados en el Anexo 2 - Tabla 5.
[12] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 6.
[13] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 7.
[14] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 8.
[15] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 9.
[16] Detalle de resultados en el Anexo 2 - Tabla 10.
[17] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 11.
[18] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 12.

1. max_iter: Número máximo de iteraciones, se elige aleatoriamente entre 100 y 300.
2. learning_rate: Tasa de aprendizaje, se elige aleatoriamente entre 0.01 y 0.3.
3. max_depth: Profundidad máxima de los árboles, se elige aleatoriamente entre 3 y 15.
4. min_samples_leaf: Número mínimo de muestras por hoja, se elige aleatoriamente entre 1 y 20.
5. l2_regularization: Término de regularización L2, se elige aleatoriamente entre 0.0 y 0.2.
6. max_leaf_nodes: Número máximo de nodos hoja, se elige aleatoriamente entre 15 y 50.

En este caso, el área bajo la curva es de 0.7542. Utilizando la función permutation_importance de scikit-learn, se calcula la importancia de características mediante la técnica de permutación[19]. Para cada característica en los datos de entrenamiento, se permutan los valores de esa característica mientras se deja las demás intactas. Esto crea un conjunto de datos con una característica permutada. Luego, se utiliza el modelo base (que surgió de un conjunto de entrenamiento sin características permutadas) para hacer predicciones en este conjunto de datos permutado y se calcula nuevamente el rendimiento. La diferencia entre el rendimiento del modelo en los datos permutados y el rendimiento en los datos originales (modelo base) se utiliza para determinar la importancia de esa característica. Si permutar una característica afecta significativamente el rendimiento del modelo, se concluye que esa característica es importante para el modelo original. Entonces, la importancia de características por permutación evalúa cómo el rendimiento de un modelo se ve afectado al permutar las características en los datos de entrenamiento. Esto proporciona información sobre la relevancia de cada característica para el modelo en términos de su contribución al rendimiento predictivo. Filtrando nuevamente por la relevancia de características y excluyendo a aquellas que poseen una contribución nula o negativa al rendimiento del modelo, el modelo pasa de 247 variables a 107. De esta manera, el área bajo la curva pasa a ser de 0.75212[20]. La precisión, por otra parte, es de 0.6904.

A modo de resumen, se consolida el desempeño de cada modelo en la Tabla 2. Si bien el desempeño de un modelo será determinado por la métrica del área bajo la curva ROC para así lograr una visión más completa del rendimiento del modelo a través de diferentes umbrales de decisión, se adjunta también el detalle de una métrica alternativa, precisión, a modo de una comparación más directa con el baseline establecido. La precisión en modelos de machine learning se refiere a la proporción de predicciones correctas realizadas por el modelo en comparación con el número total de predicciones. En otras palabras, mide qué tan bien un modelo clasifica correctamente las etiquetas de los datos. Si un modelo tiene alta precisión, significa que está haciendo muchas predicciones correctas en relación con las incorrectas. Es una métrica útil para entender la efectividad general del modelo, especialmente cuando las clases están balanceadas.

[19] Detalle de resultados en el Anexo 2 - Tabla 13.
[20] Detalle de resultados de cada iteración en el Anexo 2 - Tabla 14.

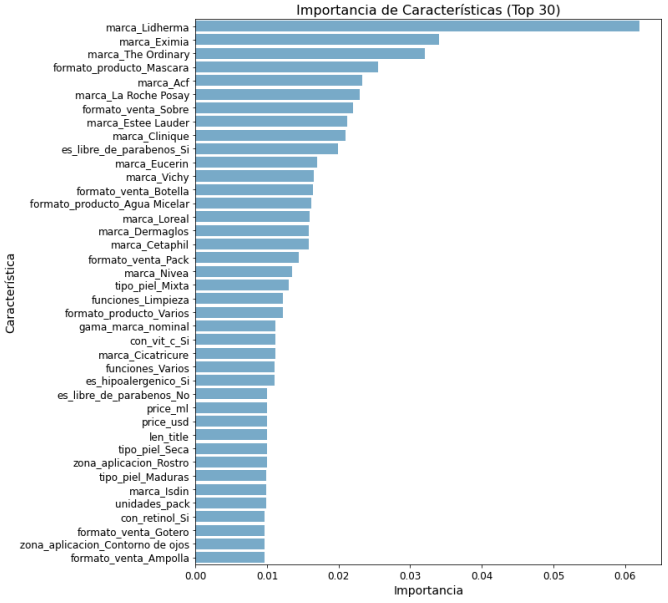
Tabla 2. Resumen del rendimiento de modelos candidatos

Variantes de modelos	Área bajo la curva ROC	Precisión
Algoritmo: DecisionTreeClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature Selection: se incluyen todas las variables	0.7055	0.6577
Algoritmo: RandomForestClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se incluyen todas las variables Método de imputación de valores faltantes: media	0.7450	0.6841
Algoritmo: RandomForestClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se incluyen todas las variables Método de imputación de valores faltantes: moda	0.7470	0.6870
Algoritmo: XGBClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se incluyen todas las variables	0.7511	0.6855
Algoritmo: HistGradientBoostingClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature Selection: se excluyen las variables con contribución nula o negativa al rendimiento del modelo (calculado a partir de permutación)	0.75212	0.6904
Algoritmo: RandomForestClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se incluyen todas las variables Método de imputación de valores faltantes: mediana	0.7526	0.68357
Algoritmo: RandomForestClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se excluyen variables con una contribución nula al rendimiento del modelo (calculado a partir de la técnica feature importance) Método de imputación de valores faltantes: mediana	0.7533	0.6865
Algoritmo: HistGradientBoostingClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature Selection: se incluyen todas las variables	0.7542	0.6938
Algoritmo: XGBClassifier Hiperparámetros: optimizados vía Random Search y validación cruzada estratificada Feature selection: se excluyen variables con una contribución nula al rendimiento del modelo (calculado a partir de la técnica feature importance)	0.758	0.688

4.2. Interpretación del modelo ganador

El modelo con mejor rendimiento deriva del algoritmo XG Boost, optimizando hiperparámetros, y filtrando por relevancia de características. A partir del cálculo de feature importance, se determina que, entre las variables más relevantes para predecir si un producto presenta promotores, se encuentran, sobre todo, marcas (Figura 32). Si bien la mayoría de las mismas son extranjeras (Clinique, The Ordinary, La Roche Posay, L'Oreal, Vichy, Eucerin, Cetaphil, Nivea, Estee Lauder, Isdin), también hay varias locales (Lidherma, Eximia, ACF, Cicatricure, Dermaglos). También se encuentran presentes algunas cualidades asociadas a la formulación del producto (si es libre de parabenos, si es hipoalergénico, si tiene vitamina C o retinol, cuál es su zona de aplicación), otras al formato del envase que contiene al producto (sobre, gotero, ampolla, botella, si es un pack y unidades por pack, pomo), otras al formato del producto en sí (mascara, agua micelar), a la función del producto (limpieza), al tipo de piel al cual está dirigido el producto (mixta, maduras, seca), o a variables relacionadas al precio del producto (precio en dólares, precio por mililitro, si es una marca de lujo). El largo del título de la publicación también es relevante.

Figura 32. Top 40 variables con mayor contribución
al rendimiento del modelo



La relevancia de características proporciona una visión general sobre qué variables son más influyentes en las predicciones del modelo. Esta métrica es valiosa para identificar las características más destacadas en términos de contribución al modelo. Para comprender las relaciones más sutiles entre variables específicas y las predicciones del modelo, se recurrirá a la técnica de dependencia parcial. Dicha herramienta examina cómo una variable específica afecta las predicciones del modelo mientras se mantienen constantes todas las demás variables. Esto permite revelar, por ejemplo, si una variable tiene un efecto lineal o no lineal en las predicciones, o si existen puntos de inflexión en las relaciones. La dependencia parcial es especialmente útil para interpretar modelos complejos como modelos de ensamble ya que ofrece información sobre cómo cada variable individualmente contribuye a las predicciones del modelo (Figura 33).

Aplicando dependencia parcial se logra visualizar una relación positiva entre la variable dependiente[21] y las siguientes variables explicativas:

- marca_Lidherma
- marca_La Roche Posay
- es_libre_de_parabenos_si
- marca_Loreal
- marca_Vichy
- marca_Eucerin
- formato_venta_pack
- gama_marca_nominal
- con_vit_c_si
- es_hipoalergenico_si
- es_libre_de_parabenos_no
- len_title
- tipo_piel_maduras
- marca_isdin
- con_retinol_si
- zona_aplicacion_contorno de ojos

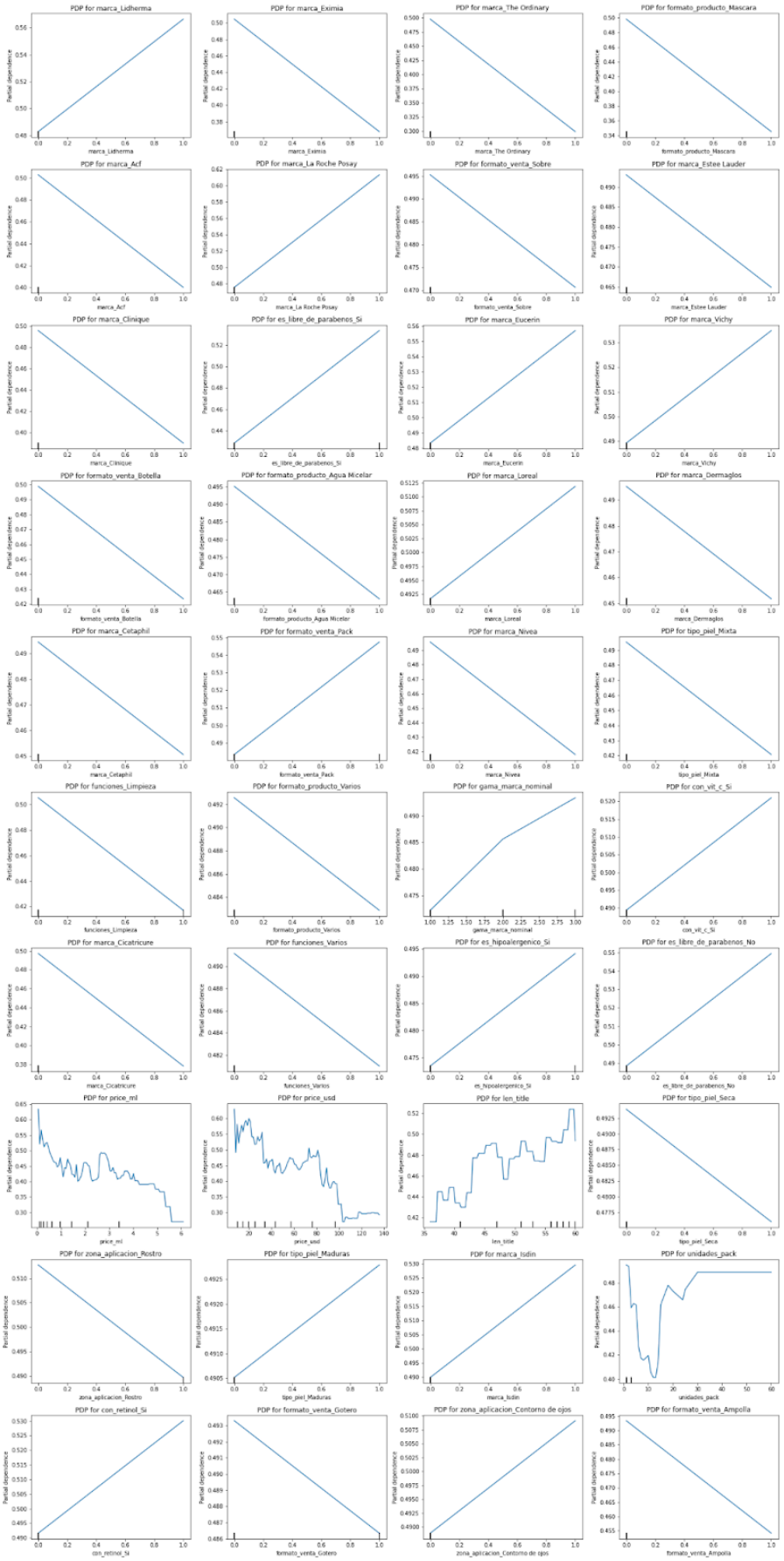
Por otra parte, se determina una relación negativa entre la variable dependiente y las siguientes variables explicativas:

- marca_Eximia
- marca_The Ordinary
- formato_producto_mascara
- marca_acf
- formato_venta_sobre
- marca_Estee Lauder
- marca_Clinique
- formato_venta_botella
- formato_producto_agua_micelar
- marca_dermaglos
- marca_cetaphil
- marca_nivea
- tipo_piel_mixta
- funciones_limpieza
- formato_producto_varios
- marca_cicatricure
- funciones_varios
- price_ml
- price_usd
- tipo_piel_seca
- zona_aplicacion_rostro
- formato_venta_ampolla
- formato_venta_gotero

Finalmente, unidades_pack presenta al principio una relación negativa y luego positiva.

[21] La variable dependiente toma valor 1 en caso de tratarse de un producto con promotores y 0 en caso contrario.

Figura 33. Dependencia parcial entre características
relevantes y variable objetivo



Para dar cierto contexto a las relaciones entre variables explicativas relevantes y la variable dependiente establecidas a partir de la técnica de dependencia parcial, es posible destacar 3 elementos fundamentales: el seteo de expectativas, el nivel de personalización de un producto y el posicionamiento de marca.

Las variables len_title y las variables asociadas al precio (price_ml y price_usd) están relacionadas al primer elemento: el seteo de expectativas. Cuánto más largo el título de una publicación y más detallada la descripción del producto que la misma ofrece, menor la probabilidad de que ocurra una falta de alineación entre lo que busca el cliente y lo que se ofrece en la plataforma de comercio electrónico. Explicaciones más vagas sobre un producto en venta, pueden derivar en realizar suposiciones imprecisas sobre la condición del mismo por parte del consumidor, derivando así en una mala experiencia. Además, el precio también juega un rol importante en el seteo de expectativas. Cuando un producto es costoso, se suele asumir que el resultado del mismo será mejor. Se paga más, y por ende también se exige más. Cuánto más estricto se vuelve el usuario, más chances existen de caer en decepciones en torno a compras.

En segundo lugar, un grado de personalización de la oferta pareciera estar vinculado a una mayor probabilidad de presentar promotores. Productos con activos con funciones específicas como el caso de la vitamina c para funciones de iluminación del rostro o el retinol para funciones anti-edad, permiten un mejor match entre atributos demandados y aquellos ofrecidos. Por otra parte, descripciones de tipo de piel más específicas como pieles maduras o hipoalergénicas, tienen más éxito que descripciones genéricas como el caso de piel mixta. En la misma línea, la aclaración de zonas más específicas al cual un producto está destinado (como contorno de ojos) es mejor recibido que zonas más vagas o generales (rostro).

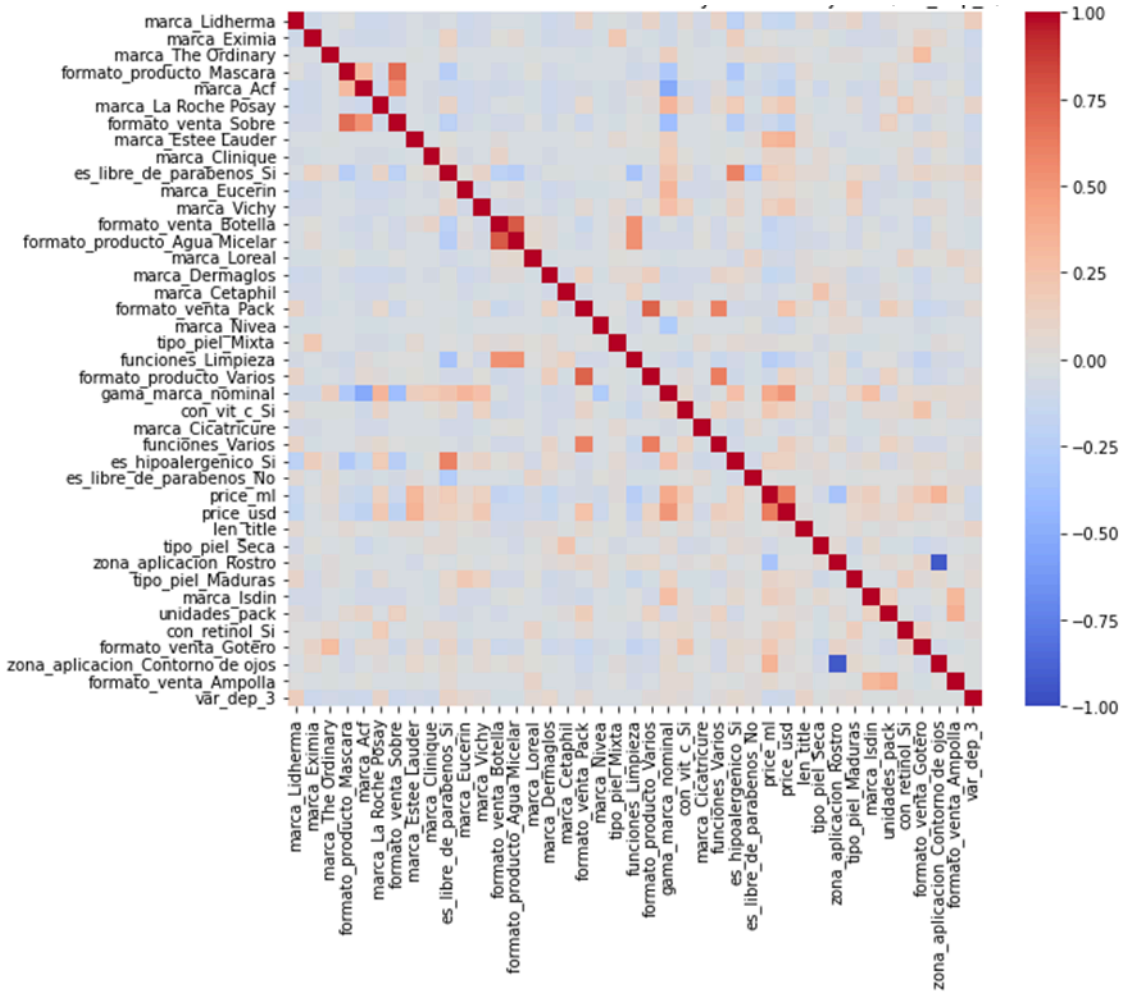
Las marcas más relacionadas a promotores se caracterizan por tener una trayectoria confiable de varios años y un buen posicionamiento en el mercado local. Entre las reseñas ligadas a dichas marcas, se destacan la buena relación entre marca y calidad (si bien varias son marcas de gama más bien alta, se puntualizan también sus notorios resultados), la buena presentación de productos, la capacidad de adaptación frente a pieles sensibles, las texturas livianas, los aromas sutiles, la buena absorción, la practicidad en la aplicación del producto (lo cual tiene que ver con su formato) y las experiencias previas (buenas experiencias previas con la marca, retroalimentan el buen posicionamiento de la misma). En cambio, las marcas que tuvieron sobre todo no promotores son, usualmente, marcas más nuevas en el mercado. Entre las reseñas de estas marcas, se destacan la mala presentación de productos (envases abiertos o fallados, derrames de productos, productos vencidos), los aplicadores de productos poco prácticos (formato del producto), los aromas demasiado invasivos, una mala relación entre precio y calidad (ausencia de resultados notorios y/o precios exorbitantes), falsificaciones en caso de marcas que solo se venden en el exterior (como The Ordinary), texturas grasosas, generación de reacciones alérgicas o cualidades irritantes del producto.

Fuera de estos 3 elementos mencionados, hay otros efectos indicados por medio de la dependencia parcial que también pueden ser interpretados (Figura 34). El formato de producto ‘máscara’ está muy correlacionado al formato de venta ‘sobre’ mientras que el ‘agua micelar’, está muy asociado al formato ‘botella’ y la función ‘limpieza’. Aparentemente, varios productos que caen en las categorías mencionadas, cuentan con muchos no promotores. Algunos de los comentarios en torno a las máscaras destacan la falta de resultados notorios o texturas diferentes a las esperadas (crema versus fango, por ejemplo). En el caso de las aguas micelares, se menciona si son irritantes o se expresan

dificultades en torno a remover efectivamente el maquillaje. A su vez, los goteros y ampollas cuentan con varios no promotores y, entre sus reseñas principales, se encuentra la dificultad de mantener el envase hermético / cerrado luego del primer uso (lo que muchas veces deriva en la oxidación del producto). Además, los goteros y ampollas suelen contener serums, productos en general más costosos. En cuánto a los packs (publicaciones con más de un producto), parecieran tener varios promotores. Sin embargo, la relación entre la cantidad de unidades dentro del pack y la presencia de promotores, no es tan lineal. Hasta 10 unidades por pack, una mayor cantidad de unidades repercute negativamente en la presencia de promotores. Sin embargo, a partir de dicho valor, la tendencia se revierte y más unidades pasan a repercutir positivamente en la variable dependiente.

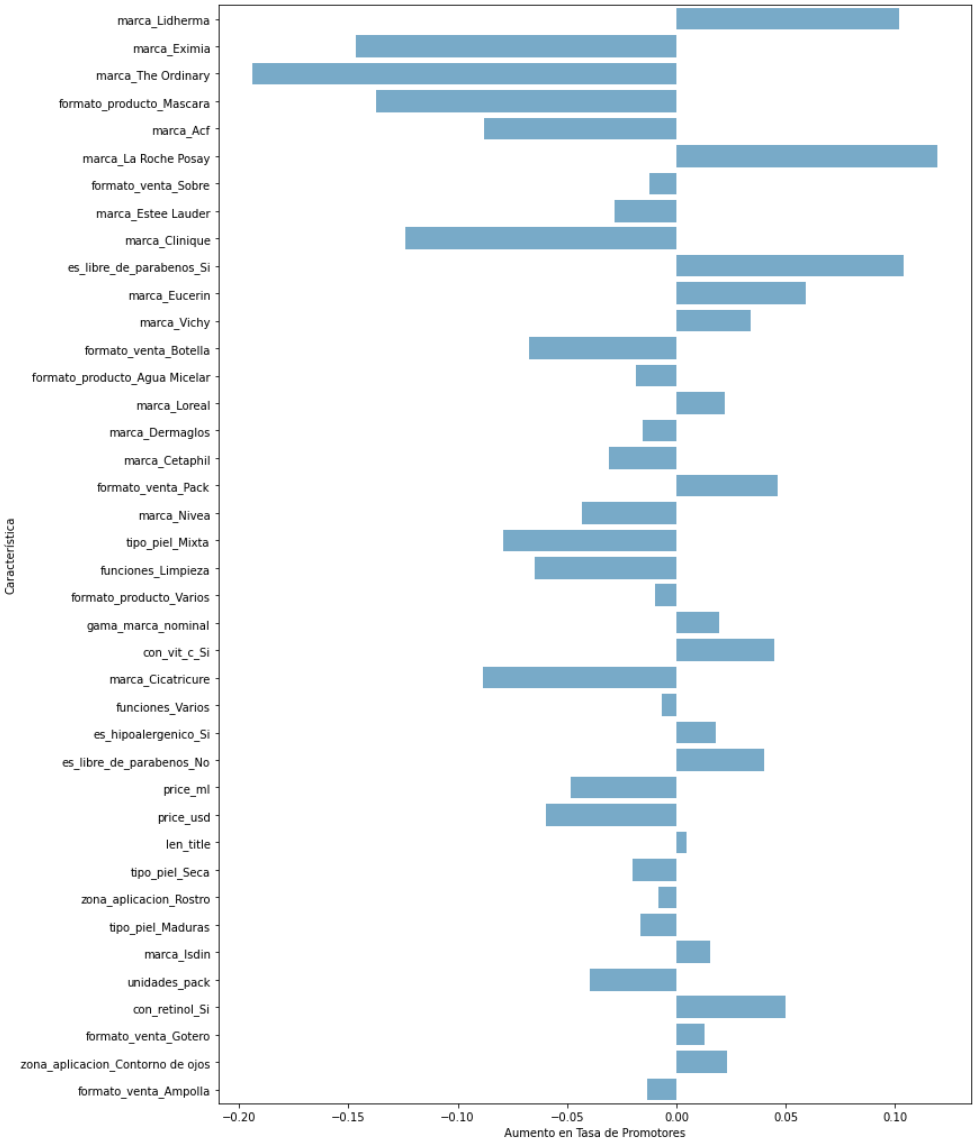
La variable es_libre_de_parabenos presenta cierta contradicción ya que tanto la categoría 'si' como 'no' guardan una relación positiva con la variable dependiente al visualizar su dependencia parcial. Esto podría deberse a interacciones con otras variables en el modelo. A primera vista, es_libre_de_parabenos_si se ve fuertemente correlacionada con es_hipoalergenico_si, cualidad asociada a la presencia de promotores. A su vez, es_libre_de_parabenos_no guarda una relación con la marca L'Oreal, también poseedora de promotores. Por otro lado, el hecho de que la variable tipo_piel_seca muestre una vinculación negativa con respecto a la variable dependiente pareciera ir en contra del análisis realizado previamente que defiende un mayor grado de personalización a la hora de consolidar promotores. Sin embargo, este tipo de piel establece una correlación positiva con una marca con muchos no promotores, Cetaphil, lo cual podría explicar este escenario. Por último, las variables formato_producto_varios y funciones_varios están muy correlacionadas a formato_venta_pack y, en menor medida a unidades_packs. Tal como fue discutido previamente, la relación entre promotores y packs no es tan lineal: arranca negativa para packs con pocas unidades y luego se vuelve positiva. Usualmente, los packs con pocas unidades están compuestos por productos de distinta índole y función mientras que los que presentan ya muchas unidades en general se tratan de un paquete puntual que contiene muchas unidades con la misma función (por ejemplo una caja de ampollas de serum o un sobre de toallitas desmaquillantes). El primer grupo tendrá varias funciones y varios formatos de productos y tendrá menos promotores mientras que el segundo se reducirá a una función y formato y tendrá más promotores.

Figura 34. Correlaciones entre características relevantes
y variable objetivo



Alternativamente, es posible calcular el aumento específico en la tasa de promotores debido a cada variable. El objetivo es medir cómo la tasa de promotores cambia cuando la variable cambia de su valor de referencia a un valor de interés. Para aquellas variables binarias, el valor de referencia será 0 (ausencia de la característica) y el valor de interés 1 (presencia de la característica). Para aquellas variables continuas, el valor de referencia será la mediana de la variable en el conjunto de entrenamiento y el valor de interés el máximo de la variable en el conjunto de entrenamiento. Finalmente, se compara el valor predicho cuando una característica toma su valor de referencia y su valor de interés. Los resultados de este procedimiento se visualizan en la Figura 35 y se alinean con los resultados generados a partir de los demás enfoques mencionados anteriormente.

Figura 35. Aumento en la tasa de promotores debido a cada variable



5. Reflexiones finales

En esta tesis, se exploraron las dinámicas de comportamiento del consumidor en el contexto de productos de cuidado facial, utilizando técnicas de modelado predictivo como árboles de decisión, Extreme Gradient Boosting, Random Forest e Histogram Gradient Boosting Classifier. El objetivo era predecir la promoción de productos e identificar las características más influyentes a la hora de realizar dicha predicción.

Los resultados obtenidos respaldaron la hipótesis inicial, demostrando que ciertas características específicas de los productos tienen un impacto significativo en su promoción. Por ejemplo, la presencia de ingredientes como el ácido hialurónico y la vitamina C se relaciona positivamente con una mayor probabilidad de promoción. Este tipo de hallazgos resulta crucial para las empresas de cosméticos, ya que les proporciona información valiosa sobre cómo diseñar y promover sus productos de manera efectiva.

En términos comerciales, comprender las preferencias del consumidor es esencial para la estrategia de producto y precios de una empresa. La optimización y selección de atributos para un nuevo producto pueden marcar la diferencia entre un lanzamiento exitoso y uno fallido, especialmente en un entorno digital donde el 'electronic word-of-mouth' juega un papel crucial en la percepción de la marca y la aceptación de productos.

Sin embargo, es importante reconocer las limitaciones de este estudio. Aunque se lograron resultados prometedores, el análisis se basó en datos específicos de una plataforma puntual y únicamente para el mercado argentino, lo que limita la generalización de las conclusiones a otros mercados. Las tendencias en el mundo cosmético son cambiantes, y los resultados obtenidos podrían no mantenerse constantes a medida que pasa el tiempo. Además, la disponibilidad de datos de alta calidad y representativos sigue siendo un desafío en el análisis del comportamiento del consumidor. Las interacciones entre variables explicativas y la variable dependiente resultan un buen punto de partida para entender las lógicas subyacentes de un consumidor a la hora de optar por comprar un producto, pero aún pueden ser sometidas a técnicas más exhaustivas para lograr mayor robustez.

Para futuros estudios, se sugiere ampliar el alcance del análisis a otros mercados a partir del uso de datos de otras plataformas, otras localidades y otros puntos de venta para mayor capacidad de generalización de los resultados. Además, se recomienda explorar técnicas de modelado más avanzadas a ensambles de árboles como modelos de aprendizaje profundo de contar con conjuntos de datos grandes. En caso de contar con datos demográficos de consumidores, es posible realizar análisis de clustering para identificar segmentos de clientes e incorporar al análisis otras variables socioeconómicas relevantes para obtener una comprensión más completa de las preferencias y comportamientos de los consumidores. A su vez, datos temporales permiten explorar cómo las tendencias y patrones a lo largo del tiempo afectan las métricas de promoción de productos. Sumado a esto, podrían explorarse otros métodos de imputación de datos faltantes (como KNN imputation) para manejar estos valores faltantes de manera más efectiva y podrían realizarse análisis de sensibilidad en la proporción de la partición de datos train y valid para evaluar cómo diferentes configuraciones afectan el rendimiento del modelo y la estabilidad de las predicciones.

En conclusión, esta investigación ha proporcionado una visión profunda de cómo las opiniones y características de los productos influyen en su éxito en el mercado. Estos resultados son fundamentales para las estrategias de marketing y desarrollo de productos en la industria del cuidado facial y pueden sentar las bases para investigaciones futuras en el campo del comportamiento del consumidor y el análisis predictivo.

Referencias

1. Allen, Jo. (2023). Cosmetics Business reveals the top 5 budget beauty trends of 2023 in new report. https://cosmeticsbusiness.com/news/article_page/Cosmetics_Business_reveals_the_top_5_budget_beauty_trends_of_2023_in_new_report/205855
2. Arndt, J. (1967). Role of Product-Related Conversations in the Diffusion of a New Product. Journal of Marketing Research, 4(3), 291–295. https://doi.org/10.2307/3149462
3. Bakhati, D., & Agrawal, S. (2022). COVID-19 pandemic lockdown—Is it affecting our skin hygiene and cosmetic practice?. Journal of Cosmetic Dermatology, 21(5), 1830-1836. https://doi.org/10.1111/jocd.14894
4. Centro de Estudios para la Producción. (2004). El Sector de Artículos de Tocador, Cosmética y Perfumería en Argentina. Notas de la Economía Real, 49-79. https://www.funcex.org.br/material/redemercosul_bibliografia/biblioteca/ESTUDOS_ARGENTINA/ARG_27.pdf
5. Ceupe | European Business School. https://www.ceupe.com/blog/linea-de-productos.html#:~:text=Una%20l%C3%ADnea%20de%20productos%20es,un%20mismo%20sector%20de%20consumidores.
6. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://dl.acm.org/doi/10.1145/2939672.2939785
7. Consejo Económico y Social de la Ciudad de Buenos Aires. (2013). Informe 'Sector Perfumería y Cosméticos en la Ciudad de Buenos Aires'. https://cdi.mecon.gob.ar/bases/docelec/ceys/industria/5.pdf
8. Darad, S., & Krishnan, S. (2023). Análisis de sentimiento de los datos de twitter de COVID-19 utilizando modelos de aprendizaje profundo y aprendizaje máquina. Ingenius. Revista de Ciencia y Tecnología, (29), 108-117. https://doi.org/10.17163/ings.n29.2023.10
9. Galeas Varas, M. E., & Rovner Berant, I. (2005). Aplicaciones de conducta del consumidor. https://repositorio.uchile.cl/handle/2250/115023
10. Howarth, J. (2023). The Ultimate List of Beauty Industry Stats (2023). https://explodingtopics.com/blog/beauty-industry-stats
11. Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225). https://doi.org/10.1609/icwsm.v8i1.14550
12. Infobae (2023) Empresas advirtieron que las trabas a la importación de insumos químicos pueden frenar la producción https://www.infobae.com/economia/2023/01/03/empresas-advirtieron-que-las-trabas-para-importar-insumos-quimicos-pueden-frenar-la-produccion/
13. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature. An Introduction to Statistical Learning (statlearning.com)
14. Kafka, F. (1997). Teoría económica. Universidad del Pacífico. Centro de Investigación.

15. Katz, E., Lazarsfeld, P.F., & Roper, E. (2006). Personal Influence: The Part Played by People in the Flow of Mass Communications (1st ed.). Routledge. https://doi.org/10.4324/9781315126234
16. Kherwa, P., & Bansal, P. (2019). Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems, 7(24). https://eudl.eu/pdf/10.4108/eai.13-7-2018.159623
17. León-Sandoval, E., Zareei, M., Barbosa-Santillán, L. I., & Falcón Morales, L. E. (2022). Measuring the Impact of Language Models in Sentiment Analysis for Mexico's COVID-19 Pandemic. Electronics, 11(16), 2483. https://doi.org/10.3390/electronics11162483
18. Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. Tourism management, 29(3), 458-468. https://doi.org/10.1016/j.tourman.2007.05.011
19. Ministerio de Producción y Trabajo Presidencia de la Nación (2019). Argentina Exporta Cosméticos. https://www.argentina.gob.ar/sites/default/files/analisis_sector_cosmeticos.pdf
20. Nielsen. (2009). Los consumidores mundiales de publicidad confían más en los amigos reales y en los desconocidos virtuales. https://www.nielsen.com/es/insights/2009/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most/
21. Nosto. (2021). The Future of Beauty and Skincare Ecommerce: EMERGING TRENDS TO WATCH IN 2021. https://www.nosto.com/wp-content/uploads/beauty-skincare-consumer-report-2021.pdf
22. Raassens, N., & Haans, H. (2017). NPS and Online WOM: Investigating the Relationship Between Customers' Promoter Scores and eWOM Behavior. Journal of Service Research, 20(3), 322-334. https://doi.org/10.1177/1094670517696965
23. Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd. https://books.google.com.ar/books?id=sKXIDwAAQBAJ
24. Saha, S., Showrov, M. I. H., Rahman, M. M., & Majumder, M. Z. H. (2022, September). VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Coronavirus Outbreak. In International Conference on Machine Intelligence and Emerging Technologies (pp. 371-385). Cham: Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-34619-4_30
25. Statista. (2022). Global skin care market size 2012-2025. https://www.statista.com/topics/3137/cosmetics-industry/#dossierKeyfigures
26. Statista. (2023). Cosmetics industry - statistics & facts. https://www.statista.com/topics/3137/cosmetics-industry/#topicOverview
27. Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access, 10, 21517-21525. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9716923
28. Villanueva, J., & Armelini, G. (2007). El boca oreja electrónico: ¿Qué sabemos de esta poderosa herramienta de marketing. Cuadernos del ebcenter, e-business Center PricewaterhouseCooper and IESE. https://media.iese.edu/research/pdfs/ESTUDIO-55.pdf

Anexo 1

Para la recolección de datos de productos y reseñas, se utilizó la API de Mercado Libre[22], enfocándose en productos de cuidado facial de diversas marcas. El proceso comenzó con la definición de una lista de términos de búsqueda específicos que incluían nombres de marcas y categorías de productos. Utilizando estos términos, se realizaron consultas a la API para obtener información detallada sobre los productos.

Se implementó un bucle para manejar la paginación y así recolectar todos los resultados disponibles para cada término de búsqueda. Para cada producto se extrajeron atributos clave (como el ID del producto, el título, el precio y la cantidad vendida), reseñas asociadas y calificaciones. Si bien muchos de los atributos se mantuvieron tal como fueron obtenidos a partir de la API (como la marca, el precio, el título, si es hipoalergénico) muchos otros necesitaron de un refinamiento adicional para unificar categorías y realizar chequeos de etiquetado adicionales (como funciones, zona de aplicación, formato del producto).

Anexo 2

Tabla 3. Resultados de la optimización de hiperparámetros en Decision Tree

mean_test_score	std_test_score	param_max_depth	param_min_samples_split	param_min_samples_leaf	param_max_features
594.641	14.155	9	12	15	sqrt
608.526	13.406	10	12	7	sqrt
618.879	2.761	13	9	4	sqrt
628.258	17.858	10	13	2	
572.838	15.236	8	13	1	log2
587.576	15.557	12	18	12	log2
630.207	1.772	13	17	10	
616.687	10.253	5	6	3	
616.443	10.374	5	8	9	
563.825	24.589	4	19	14	sqrt
609.988	16.019	12	16	2	sqrt
623.508	8.037	14	9	12	
559.805	5.439	5	5	17	log2
56.419	4.664	4	11	6	log2
571.985	12.048	6	13	18	log2
56.419	4.664	4	15	4	log2
589.281	21.551	10	17	8	log2
611.815	13.654	4	10	13	

[22] Mercado Libre es una empresa argentina de comercio electrónico fundada en 1999. Es la plataforma de comercio electrónico más grande de América Latina y ofrece una variedad de servicios y productos, incluyendo la compra y venta de bienes nuevos y usados.

549.574	27.772	3	10	7	sqrt
605.725	1.639	10	9	12	sqrt
632.156	21.093	10	18	11	
623.264	16.823	10	2	3	
605.116	8.453	13	8	10	sqrt
61.084	20.771	12	10	7	sqrt
595.737	22.464	10	17	1	log2
619.488	11.793	9	4	5	
587.454	1.161	14	2	3	log2
576.005	15.794	5	15	15	sqrt
572.838	2.105	5	15	5	sqrt
624.239	13.946	9	16	9	
589.647	10.992	12	8	13	log2
549.695	27.619	3	6	4	sqrt
625.335	11.383	9	16	13	
611.937	12.666	13	14	4	sqrt
62.229	9.571	9	3	19	
617.174	1.639	12	13	6	sqrt
586.358	17.715	14	8	11	log2
54.933	2.817	3	10	13	sqrt
616.565	10.202	5	9	6	
609.988	14.145	13	2	5	sqrt
5.581	4.896	5	16	12	log2
596.468	1.426	13	18	9	log2
55.335	3.501	3	8	12	log2
611.815	13.654	4	6	17	
54.933	2.817	3	3	17	sqrt
56.419	4.664	4	2	5	log2
604.019	5.972	3	13	2	
594.762	14.312	12	12	4	log2
625.579	14.435	10	7	17	
59.866	16.437	10	7	2	sqrt
624.239	18.058	8	17	16	
549.574	27.772	3	17	6	sqrt
616.687	10.253	5	4	4	
616.443	10.374	5	2	9	
629.963	15.352	10	9	18	
603.654	6.156	3	11	18	
617.418	10.767	5	17	16	
594.519	15.278	11	2	2	sqrt

614.495	12.314	14	10	5	sqrt
630.938	19.655	11	17	19	
560.414	513	5	12	1	log2
55.335	3.501	3	4	8	log2
634.348	14.589	11	15	18	
564.068	4.629	4	17	3	log2
604.019	21.732	11	2	4	sqrt
617.905	11.869	7	18	3	
603.654	6.156	3	15	19	
578.197	15.682	8	9	13	log2
626.553	17.236	13	2	2	
632.643	13.895	14	6	1	
613.398	18.993	14	5	16	sqrt
576.127	15.679	5	13	17	sqrt
580.512	12.011	14	7	14	log2
570.402	18.547	5	18	5	sqrt
628.015	18.554	7	2	3	
617.905	17.007	6	4	1	
564.677	25.865	4	4	10	sqrt
592.448	16.115	10	16	18	log2
626.553	1.316	8	11	2	
563.459	5.834	4	9	17	log2
549.208	28.237	3	17	11	sqrt
59.123	24.577	9	4	10	sqrt
563.825	24.589	4	8	13	sqrt
609.622	15.091	11	9	1	sqrt
629.354	13.512	14	3	15	
612.058	13.382	4	18	19	
568.453	8.695	6	16	6	log2
627.527	12.673	8	2	5	
605.725	1.639	10	13	12	sqrt
62.704	20.201	7	7	4	
57.162	11.775	6	4	15	log2
598.417	14.244	9	11	18	sqrt
576.492	16.157	5	6	19	sqrt
609.622	16.558	12	13	9	sqrt
58.782	11.943	7	16	1	sqrt
563.459	5.834	4	9	16	log2
585.384	17.309	7	8	16	sqrt
568.331	17.728	7	13	3	log2

624.604	14.543	12	15	5	
614.007	23.051	14	18	15	sqrt

Tabla 4. Resultados de la optimización de hiperparámetros en XGBoost

mean_test_score	std_test_score	param_n_estimators	param_learning_rate	param_max_depth	param_min_child_weight	param_gamma	param_subsample	param_colsample_bytree	param_reg_lambda
0.652984166	0.010467	221	0.229598	7	7	0.190143	0.529042	0.68727	0.155995
0.648355664	0.008085	187	0.222422	8	2	0.120223	0.60617	0.933088	0.832443
0.677222899	0.015417	188	0.101273	8	12	0.036681	0.805926	0.590912	0.291229
0.676735688	0.01841	154	0.119909	5	12	0.058429	0.733381	0.569747	0.983231
0.669305725	0.014262	188	0.14515	4	4	0.136062	0.904199	0.92997	0.965632
0.644336175	0.009041	271	0.21527	14	7	0.019534	0.517194	0.652307	0.495177
0.664799026	0.014176	153	0.208757	4	6	0.051756	0.592427	0.95466	0.54671
0.657125457	0.009948	194	0.29185	6	14	0.155027	0.544246	0.984792	0.921874
0.672107186	0.017059	152	0.107599	4	15	0.009045	0.982628	0.597991	0.586751
0.665286236	0.015153	108	0.098882	3	7	0.0552	0.599358	0.803517	0.772245
0.667600487	0.011843	132	0.222057	5	3	0.163092	0.96315	0.502761	0.60596
0.639951279	0.009778	132	0.265012	14	2	0.182992	0.655491	0.825539	0.063558
0.655663825	0.004087	292	0.201267	13	3	0.145921	0.985856	0.662592	0.382927
0.66954933	0.014821	114	0.080795	9	9	0.144346	0.719668	0.924457	0.110891
0.661144945	0.01405	250	0.152611	6	5	0.179153	0.769921	0.60086	0.604417
0.665408039	0.010905	269	0.18966	5	2	0.188571	0.816702	0.601531	0.80812
0.660170524	0.007102	289	0.065971	13	7	0.160734	0.613968	0.93573	0.110052
0.643118149	0.008639	299	0.268219	9	6	0.163603	0.742415	0.713554	0.534089
0.664433618	0.012076	279	0.083238	13	9	0.053882	0.985891	0.846218	0.36363
0.670645554	0.019558	212	0.159175	3	5	0.050356	0.633391	0.981224	0.99774
0.660779537	0.010227	246	0.019915	3	1	0.082207	0.765467	0.988307	0.680705
0.655663825	0.012497	283	0.187809	9	11	0.110579	0.864108	0.723892	0.237638
0.655785627	0.008092	243	0.200059	13	16	0.126461	0.66039	0.683892	0.835302
0.651400731	0.006897	246	0.187268	13	4	0.008155	0.904751	0.593259	0.71115
0.651157125	0.004398	181	0.292157	10	7	0.019235	0.837845	0.674333	0.83771
0.656638246	0.008111	106	0.172434	12	3	0.041814	0.908611	0.867608	0.659984
0.662606577	0.009292	136	0.082556	14	5	0.10593	0.594354	0.7776	0.882636
0.658587089	0.006897	251	0.263998	6	1	0.140072	0.967817	0.639436	0.850928
0.649451888	0.004835	252	0.184206	11	3	0.133798	0.580404	0.89267	0.005062
0.645188794	0.003901	260	0.205588	11	4	0.138379	0.746947	0.774367	0.018075
0.66090134	0.011066	227	0.233251	5	17	0.073294	0.546837	0.589411	0.568309
0.668696711	0.009332	208	0.083197	8	6	0.05304	0.953049	0.683858	0.486742
0.671132765	0.016908	160	0.203531	3	5	0.070016	0.786002	0.717197	0.499193
0.655785627	0.012608	198	0.308365	7	6	0.008721	0.976536	0.884277	0.747719

0.6454324	0.007343	215	0.181688	14	1	0.110553	0.692549	0.665375	0.294449
0.673081608	0.013516	153	0.060848	11	10	0.063384	0.785031	0.925568	0.69603
0.666626066	0.015328	229	0.307016	4	10	0.123001	0.924335	0.548588	0.322551
0.656151035	0.005373	149	0.175846	12	3	0.141782	0.95662	0.568311	0.867072
0.670401949	0.01166	243	0.249489	3	8	0.100303	0.943309	0.755671	0.050769
0.67088916	0.015753	252	0.141542	8	19	0.115773	0.99092	0.513808	0.155042
0.666991474	0.00846	114	0.085075	12	6	0.172081	0.56353	0.919467	0.360191
0.668818514	0.012584	295	0.074746	7	7	0.153999	0.59719	0.761122	0.842023
0.671619976	0.016239	257	0.051506	4	10	0.139902	0.69941	0.705677	0.041068
0.667356882	0.009077	239	0.085258	12	9	0.148809	0.586647	0.71676	0.408953
0.657247259	0.00871	216	0.174768	11	8	0.050049	0.758348	0.578219	0.638271
0.66406821	0.016677	253	0.229012	5	12	0.087135	0.878923	0.828556	0.355973
0.658465286	0.010488	289	0.023801	3	14	0.023215	0.737087	0.507197	0.703658
0.666869671	0.015575	122	0.152042	3	5	0.098323	0.974573	0.548917	0.048059
0.661144945	0.014332	168	0.014591	5	1	0.052179	0.807926	0.94334	0.683964
0.643727162	0.015169	277	0.27016	13	12	0.18885	0.56425	0.971946	0.573367
0.672107186	0.0158	267	0.197782	3	1	0.164128	0.597896	0.905602	0.905351
0.653593179	0.013549	134	0.015467	5	18	0.020156	0.827861	0.534681	0.336554
0.671985384	0.015525	234	0.112187	5	9	0.136323	0.848369	0.692698	0.118165
0.656272838	0.008011	237	0.230521	7	4	0.175494	0.78426	0.814471	0.423471
0.666991474	0.012201	195	0.048307	3	7	0.14633	0.965379	0.787962	0.340804
0.646772229	0.006053	204	0.235261	11	3	0.085799	0.767164	0.929206	0.377729
0.676857491	0.01607	199	0.099291	5	17	0.077924	0.952691	0.748281	0.010838
0.662484775	0.016564	210	0.295019	6	19	0.063863	0.545852	0.545643	0.083284
0.67320341	0.014811	156	0.073818	7	16	0.110741	0.965465	0.80122	0.113465
0.667356882	0.014935	174	0.026761	5	11	0.199186	0.571496	0.987124	0.117067
0.674421437	0.01762	240	0.040337	9	14	0.123644	0.698326	0.880755	0.222576
0.663459196	0.014082	235	0.163998	8	19	0.029321	0.685321	0.945948	0.374271
0.65408039	0.009645	169	0.3058	6	4	0.18945	0.740833	0.9064	0.708181
0.673325213	0.017883	225	0.084617	5	7	0.141017	0.505677	0.688994	0.492625
0.673812424	0.018265	100	0.045645	11	7	0.011261	0.618753	0.73433	0.681039
0.66455542	0.013626	216	0.034867	3	6	0.095543	0.506077	0.700111	0.963223
0.651644336	0.009544	263	0.277343	8	11	0.008632	0.984651	0.984939	0.553854
0.66772229	0.012673	243	0.218725	3	3	0.12588	0.957627	0.761549	0.737101
0.662971985	0.00715	221	0.128357	10	8	0.011573	0.731849	0.979351	0.188121
0.673812424	0.009502	291	0.03332	11	4	0.116731	0.516158	0.676676	0.453241
0.668453106	0.014482	181	0.190835	5	6	0.082241	0.790581	0.639882	0.416639
0.650426309	0.012211	104	0.272998	13	13	0.01655	0.688232	0.959588	0.026367
0.673081608	0.018337	260	0.055125	7	17	0.197455	0.70741	0.905277	0.468693
0.666260658	0.012416	194	0.269417	3	5	0.011275	0.956583	0.636704	0.749911

0.671010962	0.013702	184	0.237124	5	17	0.145224	0.56458	0.792575	0.450544
0.66589525	0.012591	131	0.078593	3	8	0.121235	0.634336	0.977026	0.046896
0.669305725	0.014838	111	0.152863	5	18	0.099633	0.983987	0.511092	0.816386
0.664677223	0.013435	184	0.186987	9	15	0.158364	0.902522	0.544204	0.639361
0.662119367	0.012237	181	0.304139	4	2	0.123453	0.718737	0.951576	0.392244
0.672228989	0.01593	129	0.164197	10	15	0.069651	0.931182	0.952079	0.622087
0.653471376	0.013839	183	0.287976	12	17	0.029415	0.990016	0.97476	0.459136
0.66638246	0.012346	210	0.20002	13	19	0.06575	0.564023	0.746309	0.12888
0.650548112	0.006775	283	0.202262	10	4	0.027765	0.937159	0.575951	0.411661
0.670401949	0.015785	120	0.190581	5	19	0.194622	0.639295	0.757618	0.168935
0.677344702	0.01446	189	0.046191	11	12	0.017741	0.564374	0.588505	0.894217
0.673690621	0.017479	155	0.037687	11	15	0.064317	0.960436	0.66505	0.873579
0.66954933	0.012129	279	0.25186	3	3	0.055376	0.685236	0.530539	0.209349
0.651766139	0.007001	208	0.120674	10	2	0.123651	0.57387	0.742261	0.062341
0.658952497	0.011361	117	0.263332	13	15	0.137433	0.766309	0.566558	0.447412
0.658343484	0.011587	284	0.123185	12	7	0.053849	0.544602	0.621235	0.84071
0.6545676	0.009279	237	0.112878	11	4	0.046643	0.749221	0.767668	0.789171
0.658099878	0.012997	212	0.186052	10	8	0.107421	0.689386	0.54346	0.564459
0.650548112	0.005245	297	0.192267	13	13	0.179929	0.618613	0.668723	0.605775
0.67088916	0.015381	289	0.083787	6	11	0.030572	0.586687	0.550891	0.285095
0.661266748	0.007443	135	0.167353	10	5	0.016047	0.948305	0.948383	0.997256
0.664920828	0.016402	193	0.01159	8	17	0.183479	0.834322	0.787999	0.170888
0.671498173	0.015682	143	0.181484	5	13	0.111353	0.983997	0.964688	0.205078
0.660170524	0.010513	100	0.230874	14	15	0.039901	0.644315	0.855476	0.61062
0.648721072	0.008211	260	0.154342	13	3	0.030873	0.727134	0.790619	0.36923
0.663215591	0.008052	183	0.070519	10	2	0.109784	0.840751	0.774302	0.254641

Tabla 5. Resultados de la técnica de feature importance para XGBoost

Característica	Importancia
marca_Lidherma	0.054231
marca_Eximia	0.032696
marca_Clinique	0.032568
marca_The Ordinary	0.028484
marca_Acf	0.026088
marca_Estee Lauder	0.020486
es_libre_de_parabenos_Si	0.019042
marca_La Roche Posay	0.018431
formato_producto_Mascara	0.017932
marca_Cetaphil	0.017788
marca_Loreal	0.017724

marca_Vichy	0.015415
marca_Cicatricure	0.015376
tipo_piel_Mixta	0.01463
marca_Nivea	0.013904
marca_Eucerin	0.013641
formato_venta_Sobre	0.013582
funciones_Varios	0.013203
formato_producto_Agua Micelar	0.012938
sustentable_Si	0.012819
marca_Dermaglos	0.01262
formato_venta_Pack	0.012295
formato_venta_Botella	0.01223
es_libre_de_parabenos_No	0.012101
funciones_Limpieza	0.011316
con_vit_c_Si	0.011085
unidades_pack	0.010724
gama_marca_nominal	0.010573
factor_proteccion	0.010563
con_retinol_No	0.010361
formato_producto_Balsamo	0.010286
tipo_piel_Seca	0.010279
marca_Garnier	0.010241
funciones_Iluminadora	0.010234
price_usd	0.010168
con_retinol_Si	0.010053
con_vit_c_No	0.009818
sustentable_No	0.00976
material_venta_Varios	0.009632
marca_Isdin	0.009508
formato_producto_Varios	0.009442
funciones_Anti-acne	0.009406
material_venta_Metal	0.00936
zona_aplicacion_Rostro	0.009237
gama_producto_nominal	0.009108
material_venta_Vidrio	0.009105
formato_venta_Pomo	0.009091
volumen_ml	0.009083
gama_producto	0.009074
len_title	0.009
funciones_Anti-edad	0.00894

tipo_piel_Maduras	0.008846
material_venta_Plastico	0.008775
es_hipoalergenico_Si	0.008718
funciones_Hidratante	0.008659
formato_venta_Ampolla	0.008599
con_protector_Si	0.008456
resistente_al_agua_Si	0.008433
price_ml	0.008412
formato_producto_Emulsion	0.008411
formato_venta_Spray	0.008286
formato_venta_Pote	0.00813
momento_dia_Dia	0.008075
es_hipoalergenico_No	0.007935
con_acido_hialuronico_Si	0.007779
con_color_No	0.00773
formato_venta_Gotero	0.007702
formato_producto_Serum	0.007638
resistente_al_agua_No	0.007549
formato_venta_Dispenser	0.00751
con_color_Si	0.007449
formato_producto_Crema	0.007413
zona_aplicacion_Contorno de ojos	0.007406
momento_dia_Dia/Noche	0.007406
tipo_piel_Sensible	0.0074
momento_dia_Noche	0.007167
funciones_Exfoliante	0.007165
tipo_piel_Grasa	0.007057
formato_producto_Locion	0.007048
con_acido_hialuronico_No	0.006941
formato_producto_Gel	0.006857
tipo_piel_Todo tipo de piel	0.006843
marca_Avene	0.006841
funciones_Calmante	0.006784
funciones_Protector Solar	0.006716
funciones_Anti-manchas	0.00612
con_protector_No	0.006073
zona_aplicacion_Cuello	0
zona_aplicacion_Labios	0
zona_aplicacion_Menton	0
marca_Adermicina	0

marca_Algabo	0
marca_Arex	0
marca_Artez Westerley	0
marca_Asepxia	0
marca_Aveno	0
marca_Avery Rose	0
marca_Avon	0
marca_Ayurdeva's	0
marca_BIU	0
marca_Bagovit	0
marca_Basicare	0
marca_Beautifull Regalos	0
marca_Biobellus	0
marca_Biocom	0
marca_Bioderma	0
marca_Biotherm	0
marca_Bling Pop	0
marca_Botik	0
marca_By Derm	0
marca_By She	0
marca_Calypso	0
marca_Carthage	0
marca_Caviahue	0
marca_Cepage	0
marca_CeraVe	0
marca_Collage	0
marca_Coony	0
marca_DD2	0
marca_Dekka	0
marca_Derm's	0
marca_Dermastore	0
marca_Dorothy Gray	0
marca_Dr Duval	0
marca_Exel	0
marca_Farmaclean	0
marca_Filorga	0
marca_Fiorel'a	0
marca_Formuly Piel	0
marca_Goicoechea	0
marca_Heburn	0

marca_Herbivore	0
marca_Icono	0
marca_Idraet	0
marca_Jactan's	0
marca_Jessamy	0
marca_Just	0
marca_Key Elements	0
marca_Kiehl's	0
marca_KoTaping	0
marca_Konjac	0
marca_Laboratorio Once	0
marca_Laca	0
marca_Lagos	0
marca_Laikou	0
marca_Lanbena	0
marca_Lancome	0
marca_Las Anittas	0
marca_Latour	0
marca_Le Lab de Beaute	0
marca_Libelle	0
marca_Libra	0
marca_Linfar	0
marca_M&Q Regalos	0
marca_MAC	0
marca_Maria T	0
marca_MegaCuper	0
marca_Microsule	0
marca_NAMECO	0
marca_Natura	0
marca_Naturaleza Organica	0
marca_Neostrata	0
marca_Neutrogena	0
marca_Niza	0
marca_Norma Bustos	0
marca_Nort	0
marca_OMS	0
marca_Orihens	0
marca_Orlane	0
marca_Ouroboros	0
marca_Paula's Choice	0

marca_Perpiel	0
marca_Pond's	0
marca_Prodermic	0
marca_Purederm	0
marca_RUH	0
marca_Regina	0
marca_Rtopr	0
marca_SHAGMIE	0
marca_Saiku	0
marca_Selecta	0
marca_Sentida Botanica	0
marca_Silfab	0
marca_Simple & Beauty	0
marca_Sir Fausto	0
marca_Sri Sri	0
marca_Sulderm	0
marca_Teatrical	0
marca_Top Choice	0
marca_Tortulan	0
marca_Ultracomb	0
marca_Valuge	0
marca_Varios	0
marca_Veganis	0
marca_Veritas	0
marca_Violetta	0
marca_Vitalis Navitas	0
marca_Weleda	0
marca_Xerotic	0
marca_Xulu Cosmeticos	0
marca_YDARIS	0
marca_Zine	0
marca_arv-lab	0
marca_bioclean	0
funciones_Cicatrizante	0
formato_producto_Aceite	0
formato_producto_Agua Termal	0
formato_producto_Agua micelar	0
formato_producto_Agua termal	0
formato_producto_Antifaz	0
formato_producto_Arcilla	0

formato_producto_Barra	0
formato_producto_Bruma	0
formato_producto_Capsulas	0
formato_producto_Cepillo	0
formato_producto_Cinta	0
formato_producto_Cobertura	0
formato_producto_Copas	0
formato_producto_Esponja	0
formato_producto_Espuma	0
formato_producto_Fango	0
formato_producto_Ice Globes	0
formato_producto_Jabon Liquido	0
formato_producto_Jabon Solido	0
formato_producto_Lapiz	0
formato_producto_Leche	0
formato_producto_Maquina	0
formato_producto_Mousse	0
formato_producto_Pad	0
formato_producto_Papel	0
formato_producto_Pastilla	0
formato_producto_Polvo	0
formato_producto_Pomada	0
formato_producto_Scrub	0
formato_producto_Solucion	0
formato_producto_Toalla	0
formato_producto_Toallitas	0
formato_producto_Tonico	0
formato_producto_Unguento	0
formato_venta_Bidon	0
formato_venta_Caja	0
formato_venta_Lapiz	0
formato_venta_Roll On	0
formato_venta_Tubo	0
formato_venta_Varios	0
material_venta_Carton	0
material_venta_Madera	0

Tabla 6. Resultados de la segunda optimización de hiperparámetros en XGBoost

mean_test_score	std_test_score	param_n_estimators	param_learning_rate	param_max_depth	param_min_child_weight	param_gamma	param_subsample	param_colsample_bytree	param_reg_lambda
0.654323995	0.00814	221	0.229598	7	7	0.190143	0.529042	0.68727	0.155995
0.639585871	0.010958	187	0.222422	8	2	0.120223	0.60617	0.933088	0.832443
0.670280146	0.012544	188	0.101273	8	12	0.036681	0.805926	0.590912	0.291229
0.673081608	0.018598	154	0.119909	5	12	0.058429	0.733381	0.569747	0.983231
0.673934227	0.014087	188	0.14515	4	4	0.136062	0.904199	0.92997	0.965632
0.637149817	0.003992	271	0.21527	14	7	0.019534	0.517194	0.652307	0.495177
0.669062119	0.014286	153	0.208757	4	6	0.051756	0.592427	0.95466	0.54671
0.657369062	0.011384	194	0.29185	6	14	0.155027	0.544246	0.984792	0.921874
0.670767357	0.015113	152	0.107599	4	15	0.009045	0.982628	0.597991	0.586751
0.664433618	0.013613	108	0.098882	3	7	0.0552	0.599358	0.803517	0.772245
0.667844093	0.015294	132	0.222057	5	3	0.163092	0.96315	0.502761	0.60596
0.642509135	0.005066	132	0.265012	14	2	0.182992	0.655491	0.825539	0.063558
0.654811206	0.009709	292	0.201267	13	3	0.145921	0.985856	0.662592	0.382927
0.674421437	0.014589	114	0.080795	9	9	0.144346	0.719668	0.924457	0.110891
0.666626066	0.011061	250	0.152611	6	5	0.179153	0.769921	0.60086	0.604417
0.659683313	0.012128	269	0.18966	5	2	0.188571	0.816702	0.601531	0.80812
0.665651644	0.01105	289	0.065971	13	7	0.160734	0.613968	0.93573	0.110052
0.649817296	0.006616	299	0.268219	9	6	0.163603	0.742415	0.713554	0.534089
0.660292326	0.009632	279	0.083238	13	9	0.053882	0.985891	0.846218	0.36363
0.673447016	0.013916	212	0.159175	3	5	0.050356	0.633391	0.981224	0.99774
0.6590743	0.012237	246	0.019915	3	1	0.082207	0.765467	0.988307	0.680705
0.655785627	0.011422	283	0.187809	9	11	0.110579	0.864108	0.723892	0.237638
0.656638246	0.010537	243	0.200059	13	16	0.126461	0.66039	0.683892	0.835302
0.644823386	0.006654	246	0.187268	13	4	0.008155	0.904751	0.593259	0.71115
0.647503045	0.007803	181	0.292157	10	7	0.019235	0.837845	0.674333	0.83771
0.65091352	0.008829	106	0.172434	12	3	0.041814	0.908611	0.867608	0.659984
0.66589525	0.014066	136	0.082556	14	5	0.10593	0.594354	0.7776	0.882636
0.656760049	0.007233	251	0.263998	6	1	0.140072	0.967817	0.639436	0.850928
0.645188794	0.008158	252	0.184206	11	3	0.133798	0.580404	0.89267	0.005062
0.647990256	0.00779	260	0.205588	11	4	0.138379	0.746947	0.774367	0.018075
0.666869671	0.008867	227	0.233251	5	17	0.073294	0.546837	0.589411	0.568309
0.670036541	0.014326	208	0.083197	8	6	0.05304	0.953049	0.683858	0.486742
0.670645554	0.015874	160	0.203531	3	5	0.070016	0.786002	0.717197	0.499193
0.659196102	0.012437	198	0.308365	7	6	0.008721	0.976536	0.884277	0.747719
0.638976857	0.003345	215	0.181688	14	1	0.110553	0.692549	0.665375	0.294449
0.673325213	0.016985	153	0.060848	11	10	0.063384	0.785031	0.925568	0.69603
0.666869671	0.015042	229	0.307016	4	10	0.123001	0.924335	0.548588	0.322551

0.658952497	0.006053	149	0.175846	12	3	0.141782	0.95662	0.568311	0.867072
0.66954933	0.01294	243	0.249489	3	8	0.100303	0.943309	0.755671	0.050769
0.672838002	0.013339	252	0.141542	8	19	0.115773	0.99092	0.513808	0.155042
0.671254568	0.013505	114	0.085075	12	6	0.172081	0.56353	0.919467	0.360191
0.67454324	0.013501	295	0.074746	7	7	0.153999	0.59719	0.761122	0.842023
0.673447016	0.016339	257	0.051506	4	10	0.139902	0.69941	0.705677	0.041068
0.664190012	0.01044	239	0.085258	12	9	0.148809	0.586647	0.71676	0.408953
0.657490865	0.011623	216	0.174768	11	8	0.050049	0.758348	0.578219	0.638271
0.671741778	0.015007	253	0.229012	5	12	0.087135	0.878923	0.828556	0.355973
0.659683313	0.010739	289	0.023801	3	14	0.023215	0.737087	0.507197	0.703658
0.671498173	0.016111	122	0.152042	3	5	0.098323	0.974573	0.548917	0.048059
0.661997564	0.013667	168	0.014591	5	1	0.052179	0.807926	0.94334	0.683964
0.649451888	0.007573	277	0.27016	13	12	0.18885	0.56425	0.971946	0.573367
0.669183922	0.013647	267	0.197782	3	1	0.164128	0.597896	0.905602	0.905351
0.650426309	0.012002	134	0.015467	5	18	0.020156	0.827861	0.534681	0.336554
0.673690621	0.009279	234	0.112187	5	9	0.136323	0.848369	0.692698	0.118165
0.65590743	0.010723	237	0.230521	7	4	0.175494	0.78426	0.814471	0.423471
0.665773447	0.013658	195	0.048307	3	7	0.14633	0.965379	0.787962	0.340804
0.648964677	0.005219	204	0.235261	11	3	0.085799	0.767164	0.929206	0.377729
0.674786845	0.018136	199	0.099291	5	17	0.077924	0.952691	0.748281	0.010838
0.663702801	0.010788	210	0.295019	6	19	0.063863	0.545852	0.545643	0.083284
0.67771011	0.015775	156	0.073818	7	16	0.110741	0.965465	0.80122	0.113465
0.670280146	0.015323	174	0.026761	5	11	0.199186	0.571496	0.987124	0.117067
0.677466504	0.018208	240	0.040337	9	14	0.123644	0.698326	0.880755	0.222576
0.66772229	0.009313	235	0.163998	8	19	0.029321	0.685321	0.945948	0.374271
0.659926918	0.010325	169	0.3058	6	4	0.18945	0.740833	0.9064	0.708181
0.671619976	0.01234	225	0.084617	5	7	0.141017	0.505677	0.688994	0.492625
0.676613886	0.014094	100	0.045645	11	7	0.011261	0.618753	0.73433	0.681039
0.665773447	0.01231	216	0.034867	3	6	0.095543	0.506077	0.700111	0.963223
0.653227771	0.008302	263	0.277343	8	11	0.008632	0.984651	0.984939	0.553854
0.672472594	0.012048	243	0.218725	3	3	0.12588	0.957627	0.761549	0.737101
0.662362972	0.008382	221	0.128357	10	8	0.011573	0.731849	0.979351	0.188121
0.668940317	0.009361	291	0.03332	11	4	0.116731	0.516158	0.676676	0.453241
0.670645554	0.013712	181	0.190835	5	6	0.082241	0.790581	0.639882	0.416639
0.653349574	0.013115	104	0.272998	13	13	0.01655	0.688232	0.959588	0.026367
0.677344702	0.016207	260	0.055125	7	17	0.197455	0.70741	0.905277	0.468693
0.666626066	0.020186	194	0.269417	3	5	0.011275	0.956583	0.636704	0.749911
0.668331303	0.017142	184	0.237124	5	17	0.145224	0.56458	0.792575	0.450544
0.670401949	0.012754	131	0.078593	3	8	0.121235	0.634336	0.977026	0.046896
0.672838002	0.013723	111	0.152863	5	18	0.099633	0.983987	0.511092	0.816386

0.663093788	0.011311	184	0.186987	9	15	0.158364	0.902522	0.544204	0.639361
0.660170524	0.017316	181	0.304139	4	2	0.123453	0.718737	0.951576	0.392244
0.668209501	0.015883	129	0.164197	10	15	0.069651	0.931182	0.952079	0.622087
0.65773447	0.009172	183	0.287976	12	17	0.029415	0.990016	0.97476	0.459136
0.658099878	0.01272	210	0.20002	13	19	0.06575	0.564023	0.746309	0.12888
0.647137637	0.007088	283	0.202262	10	4	0.027765	0.937159	0.575951	0.411661
0.670036541	0.015428	120	0.190581	5	19	0.194622	0.639295	0.757618	0.168935
0.675883069	0.011698	189	0.046191	11	12	0.017741	0.564374	0.588505	0.894217
0.674056029	0.016578	155	0.037687	11	15	0.064317	0.960436	0.66505	0.873579
0.669305725	0.013404	279	0.25186	3	3	0.055376	0.685236	0.530539	0.209349
0.652375152	0.010244	208	0.120674	10	2	0.123651	0.57387	0.742261	0.062341
0.663215591	0.011426	117	0.263332	13	15	0.137433	0.766309	0.566558	0.447412
0.655420219	0.012017	284	0.123185	12	7	0.053849	0.544602	0.621235	0.84071
0.662119367	0.010128	237	0.112878	11	4	0.046643	0.749221	0.767668	0.789171
0.655663825	0.010063	212	0.186052	10	8	0.107421	0.689386	0.54346	0.564459
0.651766139	0.010113	297	0.192267	13	13	0.179929	0.618613	0.668723	0.605775
0.67137637	0.015405	289	0.083787	6	11	0.030572	0.586687	0.550891	0.285095
0.6590743	0.008804	135	0.167353	10	5	0.016047	0.948305	0.948383	0.997256
0.66406821	0.014614	193	0.01159	8	17	0.183479	0.834322	0.787999	0.170888
0.671254568	0.016234	143	0.181484	5	13	0.111353	0.983997	0.964688	0.205078
0.658708892	0.00916	100	0.230874	14	15	0.039901	0.644315	0.855476	0.61062
0.649451888	0.005747	260	0.154342	13	3	0.030873	0.727134	0.790619	0.36923
0.663702801	0.005834	183	0.070519	10	2	0.109784	0.840751	0.774302	0.254641

Tabla 7. Resultados de la optimización de
hiperparámetros para Random Forest imputando la media

mean_test_score	std_test_score	param_n_ estimators	param_max_ depth	param_min_ samples_split	param_min_ samples_leaf	param_max_ features	param_bootstrap
0.636541	0.012962	288	6	9	11	0.955643	TRUE
0.658831	0.013007	187	9	12	11	0.501249	TRUE
0.647016	0.013528	187	6	3	3	0.22858	TRUE
0.666139	0.0128	157	8	13	1	0.100701	FALSE
0.67296	0.014178	114	14	17	10	0.120756	FALSE
0.629111	0.011387	150	5	6	3	0.444216	FALSE
0.645432	0.009485	188	7	5	18	0.505449	TRUE
0.624604	0.012257	287	4	16	2	0.446875	FALSE
0.671255	0.014017	180	14	4	15	0.545659	TRUE
0.616322	0.009419	153	4	7	2	0.69627	FALSE
0.643727	0.013079	261	6	13	18	0.266369	FALSE
0.633496	0.016425	299	6	16	16	0.63811	FALSE
0.648721	0.016622	181	9	17	8	0.568751	FALSE
0.64592	0.01216	256	7	10	13	0.628076	TRUE

0.639099	0.011339	235	3	2	9	0.114073	TRUE
0.669915	0.017048	262	13	4	8	0.278958	FALSE
0.670037	0.015073	127	13	10	7	0.16664	TRUE
0.663337	0.014925	147	11	2	2	0.660968	TRUE
0.659196	0.016641	271	9	4	5	0.701957	TRUE
0.632643	0.014979	200	8	4	1	0.898491	FALSE
0.629354	0.01216	126	5	8	14	0.784707	TRUE
0.666382	0.016064	151	12	8	13	0.122877	TRUE
0.63715	0.012812	242	6	14	7	0.91681	TRUE
0.635079	0.012134	269	6	3	7	0.948568	TRUE
0.667844	0.013501	183	11	13	12	0.194945	TRUE
0.663094	0.012996	228	13	2	7	0.391911	FALSE
0.669062	0.015767	102	13	10	13	0.484397	TRUE
0.652619	0.014394	150	8	6	9	0.559673	TRUE
0.668575	0.014501	161	14	18	9	0.319713	FALSE
0.666382	0.016022	230	12	3	7	0.966203	TRUE
0.649695	0.013461	101	7	18	17	0.356356	TRUE
0.67162	0.010388	228	14	2	5	0.146331	FALSE
0.625213	0.015148	282	4	5	6	0.577841	TRUE
0.664312	0.015803	215	10	6	6	0.313874	TRUE
0.651157	0.016138	196	8	17	16	0.458942	FALSE
0.670524	0.016164	247	14	4	16	0.557379	TRUE
0.666382	0.012783	151	13	8	3	0.298217	FALSE
0.620706	0.012225	203	3	19	7	0.457815	TRUE
0.667844	0.016344	198	13	11	18	0.40696	TRUE
0.672107	0.015611	101	14	18	16	0.257459	TRUE
0.670524	0.014823	198	14	10	9	0.602464	TRUE
0.626553	0.01139	212	5	12	1	0.870692	TRUE
0.624604	0.008618	198	6	17	3	0.908699	FALSE
0.635566	0.013751	153	4	3	18	0.244727	TRUE
0.665652	0.015755	241	11	2	4	0.258333	TRUE
0.645189	0.013324	294	6	8	8	0.429822	TRUE
0.616809	0.011799	257	3	15	19	0.588286	TRUE
0.650426	0.017048	120	8	3	8	0.453788	FALSE
0.607795	0.008934	160	3	6	1	0.619213	FALSE
0.62972	0.012259	271	5	18	17	0.614804	FALSE
0.630451	0.014608	123	7	6	9	0.895145	FALSE
0.616809	0.010417	215	3	2	3	0.597488	TRUE
0.652375	0.013804	227	7	4	1	0.27182	TRUE
0.661145	0.017978	273	11	9	3	0.455222	FALSE

0.615956	0.009915	229	4	11	2	0.653507	FALSE
0.59659	0.010195	275	3	12	9	0.863803	FALSE
0.667113	0.013912	252	12	19	3	0.828425	TRUE
0.66078	0.010492	119	14	14	4	0.560208	FALSE
0.659805	0.018134	214	10	15	16	0.73177	TRUE
0.667844	0.01634	279	14	18	19	0.620978	TRUE
0.652253	0.017431	107	8	6	11	0.372939	FALSE
0.672107	0.010496	295	14	8	5	0.294239	TRUE
0.639951	0.012904	151	6	9	3	0.470219	FALSE
0.619976	0.010335	219	4	18	19	0.56467	FALSE
0.633252	0.012189	239	5	10	5	0.495074	TRUE
0.635688	0.013702	133	3	16	1	0.255965	TRUE
0.642753	0.017561	216	8	14	8	0.743136	FALSE
0.65944	0.015385	175	9	4	5	0.6914	TRUE
0.625579	0.009648	146	5	6	14	0.782061	FALSE
0.64458	0.013402	184	6	17	16	0.257797	TRUE
0.647259	0.013284	168	6	11	5	0.255882	TRUE
0.629598	0.013149	168	4	2	10	0.334804	FALSE
0.640073	0.012523	227	6	3	16	0.654266	TRUE
0.649817	0.012033	262	13	19	12	0.820854	FALSE
0.612911	0.008293	200	3	12	19	0.512428	FALSE
0.622412	0.008942	142	5	11	8	0.838384	FALSE
0.674056	0.014491	134	14	19	2	0.1907	FALSE
0.61486	0.010181	214	3	3	10	0.69015	TRUE
0.659074	0.016101	174	9	14	18	0.206348	TRUE
0.661876	0.01398	200	10	11	4	0.761564	TRUE
0.659927	0.017455	135	10	4	17	0.826151	TRUE
0.600122	0.012117	212	3	13	5	0.798772	FALSE
0.657734	0.01682	212	9	18	9	0.779089	TRUE
0.625457	0.011542	199	6	18	13	0.843812	FALSE
0.646163	0.014818	127	11	8	9	0.862429	FALSE
0.63447	0.014465	216	6	17	15	0.616094	FALSE
0.631425	0.013267	156	5	17	4	0.642197	TRUE
0.647625	0.014301	221	7	15	7	0.544978	TRUE
0.626797	0.012812	245	5	12	11	0.898934	TRUE
0.671864	0.015576	106	14	18	12	0.719227	TRUE
0.666017	0.019388	104	10	3	18	0.165487	FALSE
0.649574	0.016317	208	7	17	8	0.309905	TRUE
0.652862	0.017939	231	8	14	7	0.83152	TRUE
0.639586	0.018534	281	8	13	19	0.737363	FALSE

0.666626	0.01594	225	13	8	19	0.323852	TRUE
0.626188	0.014274	266	4	10	8	0.521795	TRUE
0.62838	0.013906	265	5	2	17	0.625032	TRUE
0.649574	0.013218	160	8	14	13	0.821898	TRUE
0.655542	0.014082	263	7	6	11	0.139061	TRUE
0.627649	0.011456	148	5	16	18	0.972372	TRUE

Tabla 8. Resultados de la optimización de hiperparámetros para Random Forest imputando la mediana

mean_test_score	std_test_score	param_n_ estimators	param_max_ depth	param_min_ samples_split	param_min_ samples_leaf	param_max_ features	param_bootstrap
0.635445	0.013735	288	6	9	11	0.955643	TRUE
0.657004	0.015719	187	9	12	11	0.501249	TRUE
0.645798	0.015525	187	6	3	3	0.22858	TRUE
0.663946	0.012454	157	8	13	1	0.100701	FALSE
0.672107	0.011888	114	14	17	10	0.120756	FALSE
0.63106	0.013395	150	5	6	3	0.444216	FALSE
0.646285	0.010309	188	7	5	18	0.505449	TRUE
0.623386	0.012245	287	4	16	2	0.446875	FALSE
0.66894	0.016498	180	14	4	15	0.545659	TRUE
0.616078	0.008989	153	4	7	2	0.69627	FALSE
0.643484	0.01329	261	6	13	18	0.266369	FALSE
0.633009	0.017335	299	6	16	16	0.63811	FALSE
0.648721	0.019912	181	9	17	8	0.568751	FALSE
0.646285	0.01237	256	7	10	13	0.628076	TRUE
0.637759	0.012466	235	3	2	9	0.114073	TRUE
0.666382	0.015053	262	13	4	8	0.278958	FALSE
0.668331	0.01404	127	13	10	7	0.16664	TRUE
0.664434	0.016983	147	11	2	2	0.660968	TRUE
0.656638	0.016275	271	9	4	5	0.701957	TRUE
0.631669	0.014264	200	8	4	1	0.898491	FALSE
0.629598	0.011868	126	5	8	14	0.784707	TRUE
0.66687	0.017503	151	12	8	13	0.122877	TRUE
0.636419	0.013807	242	6	14	7	0.91681	TRUE
0.634714	0.014217	269	6	3	7	0.948568	TRUE
0.666139	0.013729	183	11	13	12	0.194945	TRUE
0.664921	0.013284	228	13	2	7	0.391911	FALSE
0.668697	0.015755	102	13	10	13	0.484397	TRUE
0.65335	0.016321	150	8	6	9	0.559673	TRUE
0.669306	0.014682	161	14	18	9	0.319713	FALSE
0.666504	0.016605	230	12	3	7	0.966203	TRUE
0.649939	0.012643	101	7	18	17	0.356356	TRUE

0.675274	0.013245	228	14	2	5	0.146331	FALSE
0.62497	0.015166	282	4	5	6	0.577841	TRUE
0.665408	0.016258	215	10	6	6	0.313874	TRUE
0.648843	0.019137	196	8	17	16	0.458942	FALSE
0.669184	0.015254	247	14	4	16	0.557379	TRUE
0.66821	0.014112	151	13	8	3	0.298217	FALSE
0.621072	0.01299	203	3	19	7	0.457815	TRUE
0.667235	0.013681	198	13	11	18	0.40696	TRUE
0.669549	0.015663	101	14	18	16	0.257459	TRUE
0.66821	0.015346	198	14	10	9	0.602464	TRUE
0.625822	0.012275	212	5	12	1	0.870692	TRUE
0.624604	0.009603	198	6	17	3	0.908699	FALSE
0.634348	0.012685	153	4	3	18	0.244727	TRUE
0.667479	0.015272	241	11	2	4	0.258333	TRUE
0.645189	0.01472	294	6	8	8	0.429822	TRUE
0.616809	0.011717	257	3	15	19	0.588286	TRUE
0.649208	0.01854	120	8	3	8	0.453788	FALSE
0.607552	0.009185	160	3	6	1	0.619213	FALSE
0.630572	0.012341	271	5	18	17	0.614804	FALSE
0.629963	0.015444	123	7	6	9	0.895145	FALSE
0.616443	0.010417	215	3	2	3	0.597488	TRUE
0.652375	0.014547	227	7	4	1	0.27182	TRUE
0.659196	0.017569	273	11	9	3	0.455222	FALSE
0.615469	0.008601	229	4	11	2	0.653507	FALSE
0.597686	0.01044	275	3	12	9	0.863803	FALSE
0.665408	0.015294	252	12	19	3	0.828425	TRUE
0.663459	0.009937	119	14	14	4	0.560208	FALSE
0.660658	0.015624	214	10	15	16	0.73177	TRUE
0.669184	0.014065	279	14	18	19	0.620978	TRUE
0.650061	0.017958	107	8	6	11	0.372939	FALSE
0.672229	0.010244	295	14	8	5	0.294239	TRUE
0.638977	0.011679	151	6	9	3	0.470219	FALSE
0.619732	0.010871	219	4	18	19	0.56467	FALSE
0.633861	0.011843	239	5	10	5	0.495074	TRUE
0.633739	0.012037	133	3	16	1	0.255965	TRUE
0.640804	0.01489	216	8	14	8	0.743136	FALSE
0.658952	0.017835	175	9	4	5	0.6914	TRUE
0.625579	0.009159	146	5	6	14	0.782061	FALSE
0.646772	0.013845	184	6	17	16	0.257797	TRUE
0.646041	0.015073	168	6	11	5	0.255882	TRUE

0.62972	0.013068	168	4	2	10	0.334804	FALSE
0.639099	0.012117	227	6	3	16	0.654266	TRUE
0.651279	0.010885	262	13	19	12	0.820854	FALSE
0.612546	0.008345	200	3	12	19	0.512428	FALSE
0.622533	0.009319	142	5	11	8	0.838384	FALSE
0.672594	0.014312	134	14	19	2	0.1907	FALSE
0.614982	0.010062	214	3	3	10	0.69015	TRUE
0.658831	0.015063	174	9	14	18	0.206348	TRUE
0.662728	0.017936	200	10	11	4	0.761564	TRUE
0.659805	0.016774	135	10	4	17	0.826151	TRUE
0.6	0.011961	212	3	13	5	0.798772	FALSE
0.658709	0.017561	212	9	18	9	0.779089	TRUE
0.625579	0.011106	199	6	18	13	0.843812	FALSE
0.645554	0.012107	127	11	8	9	0.862429	FALSE
0.632887	0.016316	216	6	17	15	0.616094	FALSE
0.631425	0.011999	156	5	17	4	0.642197	TRUE
0.647503	0.014481	221	7	15	7	0.544978	TRUE
0.626675	0.013415	245	5	12	11	0.898934	TRUE
0.670402	0.016027	106	14	18	12	0.719227	TRUE
0.663581	0.017616	104	10	3	18	0.165487	FALSE
0.648721	0.014845	208	7	17	8	0.309905	TRUE
0.653959	0.017581	231	8	14	7	0.83152	TRUE
0.638855	0.017831	281	8	13	19	0.737363	FALSE
0.667966	0.013511	225	13	8	19	0.323852	TRUE
0.625944	0.014393	266	4	10	8	0.521795	TRUE
0.628989	0.014264	265	5	2	17	0.625032	TRUE
0.649939	0.015196	160	8	14	13	0.821898	TRUE
0.656273	0.016203	263	7	6	11	0.139061	TRUE
0.627649	0.012135	148	5	16	18	0.972372	TRUE

Tabla 9. Resultados de la optimización de
hiperparámetros para Random Forest imputando la moda

mean_test_score	std_test_score	param_n_ estimators	param_max_ depth	param_min_ samples_split	param_min_ samples_leaf	param_max_ features	param_bootstrap
0.63581	0.012896	288	6	9	11	0.955643	TRUE
0.659318	0.015573	187	9	12	11	0.501249	TRUE
0.646894	0.01418	187	6	3	3	0.22858	TRUE
0.665286	0.013357	157	8	13	1	0.100701	FALSE
0.674056	0.014065	114	14	17	10	0.120756	FALSE
0.62972	0.011689	150	5	6	3	0.444216	FALSE
0.646407	0.010858	188	7	5	18	0.505449	TRUE

0.624604	0.012732	287	4	16	2	0.446875	FALSE
0.669793	0.014979	180	14	4	15	0.545659	TRUE
0.617052	0.009883	153	4	7	2	0.69627	FALSE
0.642631	0.011666	261	6	13	18	0.266369	FALSE
0.63313	0.016924	299	6	16	16	0.63811	FALSE
0.650792	0.017804	181	9	17	8	0.568751	FALSE
0.646041	0.012765	256	7	10	13	0.628076	TRUE
0.637759	0.013115	235	3	2	9	0.114073	TRUE
0.669549	0.014881	262	13	4	8	0.278958	FALSE
0.671133	0.012988	127	13	10	7	0.16664	TRUE
0.665043	0.015726	147	11	2	2	0.660968	TRUE
0.657491	0.017159	271	9	4	5	0.701957	TRUE
0.631912	0.013156	200	8	4	1	0.898491	FALSE
0.628745	0.012098	126	5	8	14	0.784707	TRUE
0.667357	0.014362	151	12	8	13	0.122877	TRUE
0.637028	0.013797	242	6	14	7	0.91681	TRUE
0.634957	0.013043	269	6	3	7	0.948568	TRUE
0.666748	0.015832	183	11	13	12	0.194945	TRUE
0.664068	0.011107	228	13	2	7	0.391911	FALSE
0.669549	0.015468	102	13	10	13	0.484397	TRUE
0.652132	0.015925	150	8	6	9	0.559673	TRUE
0.668331	0.011565	161	14	18	9	0.319713	FALSE
0.666139	0.016734	230	12	3	7	0.966203	TRUE
0.64933	0.013696	101	7	18	17	0.356356	TRUE
0.672838	0.011186	228	14	2	5	0.146331	FALSE
0.626188	0.014684	282	4	5	6	0.577841	TRUE
0.663946	0.012847	215	10	6	6	0.313874	TRUE
0.649695	0.01687	196	8	17	16	0.458942	FALSE
0.669184	0.015681	247	14	4	16	0.557379	TRUE
0.664434	0.013667	151	13	8	3	0.298217	FALSE
0.620585	0.011778	203	3	19	7	0.457815	TRUE
0.665408	0.014796	198	13	11	18	0.40696	TRUE
0.667722	0.017417	101	14	18	16	0.257459	TRUE
0.66687	0.013866	198	14	10	9	0.602464	TRUE
0.625579	0.012023	212	5	12	1	0.870692	TRUE
0.624361	0.008363	198	6	17	3	0.908699	FALSE
0.635201	0.014443	153	4	3	18	0.244727	TRUE
0.664555	0.015009	241	11	2	4	0.258333	TRUE
0.643605	0.011999	294	6	8	8	0.429822	TRUE
0.617296	0.011623	257	3	15	19	0.588286	TRUE

0.648965	0.017142	120	8	3	8	0.453788	FALSE
0.608039	0.009337	160	3	6	1	0.619213	FALSE
0.629354	0.012311	271	5	18	17	0.614804	FALSE
0.629963	0.013451	123	7	6	9	0.895145	FALSE
0.617052	0.011252	215	3	2	3	0.597488	TRUE
0.649086	0.015438	227	7	4	1	0.27182	TRUE
0.660171	0.016228	273	11	9	3	0.455222	FALSE
0.616078	0.008788	229	4	11	2	0.653507	FALSE
0.597199	0.010637	275	3	12	9	0.863803	FALSE
0.667113	0.014471	252	12	19	3	0.828425	TRUE
0.661389	0.010428	119	14	14	4	0.560208	FALSE
0.661267	0.018056	214	10	15	16	0.73177	TRUE
0.671011	0.01715	279	14	18	19	0.620978	TRUE
0.652253	0.016576	107	8	6	11	0.372939	FALSE
0.673569	0.011928	295	14	8	5	0.294239	TRUE
0.639708	0.011482	151	6	9	3	0.470219	FALSE
0.619488	0.010017	219	4	18	19	0.56467	FALSE
0.634836	0.012807	239	5	10	5	0.495074	TRUE
0.637028	0.012722	133	3	16	1	0.255965	TRUE
0.641657	0.01629	216	8	14	8	0.743136	FALSE
0.659683	0.015624	175	9	4	5	0.6914	TRUE
0.625944	0.009477	146	5	6	14	0.782061	FALSE
0.645798	0.014045	184	6	17	16	0.257797	TRUE
0.647016	0.01443	168	6	11	5	0.255882	TRUE
0.630938	0.012588	168	4	2	10	0.334804	FALSE
0.638124	0.013194	227	6	3	16	0.654266	TRUE
0.651523	0.011955	262	13	19	12	0.820854	FALSE
0.613033	0.008418	200	3	12	19	0.512428	FALSE
0.622655	0.009974	142	5	11	8	0.838384	FALSE
0.676005	0.013181	134	14	19	2	0.1907	FALSE
0.614982	0.009973	214	3	3	10	0.69015	TRUE
0.661023	0.015582	174	9	14	18	0.206348	TRUE
0.661023	0.013368	200	10	11	4	0.761564	TRUE
0.660901	0.017496	135	10	4	17	0.826151	TRUE
0.6	0.011961	212	3	13	5	0.798772	FALSE
0.65944	0.017279	212	9	18	9	0.779089	TRUE
0.626431	0.010658	199	6	18	13	0.843812	FALSE
0.646163	0.014251	127	11	8	9	0.862429	FALSE
0.634592	0.013016	216	6	17	15	0.616094	FALSE
0.63106	0.011711	156	5	17	4	0.642197	TRUE

0.647259	0.014097	221	7	15	7	0.544978	TRUE
0.627406	0.013636	245	5	12	11	0.898934	TRUE
0.667235	0.014766	106	14	18	12	0.719227	TRUE
0.662972	0.016394	104	10	3	18	0.165487	FALSE
0.648356	0.013379	208	7	17	8	0.309905	TRUE
0.651644	0.016772	231	8	14	7	0.83152	TRUE
0.6419	0.016264	281	8	13	19	0.737363	FALSE
0.668819	0.015346	225	13	8	19	0.323852	TRUE
0.626309	0.014061	266	4	10	8	0.521795	TRUE
0.630085	0.013599	265	5	2	17	0.625032	TRUE
0.65067	0.015197	160	8	14	13	0.821898	TRUE
0.653228	0.014841	263	7	6	11	0.139061	TRUE
0.627649	0.011508	148	5	16	18	0.972372	TRUE

Tabla 10. Resultados de la técnica de feature importance para Random Forest

Característica	Importancia
price_usd	0.138306
price_ml	0.113863
len_title	0.070456
marca_Lidherma	0.066728
es_libre_de_parabenos_Si	0.042574
marca_Eximia	0.040569
volumen_ml	0.03688
gama_producto_nominal	0.025575
marca_La Roche Posay	0.022445
marca_The Ordinary	0.020234
marca_Clinique	0.017924
es_hipoalergenico_Si	0.016039
marca_Acf	0.015902
formato_venta_Sobre	0.015624
gama_marca_nominal	0.015328
unidades_pack	0.015042
gama_producto	0.014349
formato_producto_Mascara	0.013522
marca_Eucerin	0.010194
con_acido_hialuronico_Si	0.008476
factor_proteccion	0.008182
material_venta_Plastico	0.008118
marca_Dermaglos	0.007858
con_acido_hialuronico_No	0.0073

funciones_Anti-edad	0.007134
sustentable_No	0.007099
funciones_Hidratante	0.007035
material_venta_Vidrio	0.006994
tipo_piel_Todo tipo de piel	0.006906
funciones_Limpieza	0.006508
sustentable_Si	0.006382
formato_venta_Pack	0.006355
formato_venta_Botella	0.005795
marca_Estee Lauder	0.005758
tipo_piel_Seca	0.005648
formato_producto_Varios	0.005642
formato_producto_Agua Micelar	0.005435
formato_venta_Dispenser	0.005314
formato_producto_Gel	0.005284
formato_venta_Pomo	0.005243
marca_Vichy	0.005061
formato_producto_Crema	0.005051
tipo_piel_Grasa	0.00484
funciones_Varios	0.004782
marca_Cetaphil	0.004531
con_vit_c_Si	0.004528
marca_Cicatricure	0.004423
zona_aplicacion_Contorno de ojos	0.004288
con_vit_c_No	0.004253
momento_dia_Dia/Noche	0.004238
marca_Loreal	0.004183
formato_venta_Gotero	0.004175
formato_producto_Serum	0.004136
momento_dia_Dia	0.003974
formato_venta_Pote	0.003963
marca_Nivea	0.003893
es_libre_de_parabenos_No	0.003757
con_protector_No	0.003594
zona_aplicacion_Rostro	0.003486
tipo_piel_Maduras	0.00346
con_protector_Si	0.003418
tipo_piel_Sensible	0.003307
es_hipoalergenico_No	0.003043
tipo_piel_Mixta	0.002998

marca_Isdin	0.002756
con_retinol_Si	0.002496
resistente_al_agua_No	0.002298
marca_Garnier	0.00229
momento_dia_Noche	0.002263
con_retinol_No	0.002203
funciones_Anti-acne	0.001924
funciones_Anti-manchas	0.00187
funciones_Iluminadora	0.001779
formato_producto_Emulsion	0.001694
formato_venta_Spray	0.001676
resistente_al_agua_Si	0.001647
material_venta_Metal	0.001634
material_venta_Varios	0.00156
funciones_Exfoliante	0.001524
formato_producto_Polvo	0.001477
marca_Regina	0.001422
marca_Kiehl's	0.001397
funciones_Calmante	0.001387
con_color_No	0.001384
formato_venta_Ampolla	0.001359
formato_producto_Locion	0.001255
formato_producto_Balsamo	0.001106
marca_Avene	0.000939
formato_producto_Solucion	0.000867
con_color_Si	0.000832
funciones_Cicatrizante	0.000818
formato_producto_Aceite	0.000814
formato_venta_Tubo	0.000648
formato_producto_Tonico	0.00058
zona_aplicacion_Labios	0.000472
formato_producto_Pad	0.000463
marca_Goicoechea	0.000455
marca_Varios	0.000419
formato_producto_Antifaz	0.000394
funciones_Protector Solar	0.000349
formato_venta_Roll On	0.000272
marca_Konjac	0.000253
formato_producto_Leche	0.000232
marca_Laca	0.000231

formato_producto_Esponja	0.000214
formato_producto_Espuma	0.000192
marca_CeraVe	0.000191
marca_Idraet	0.000191
marca_Biobellus	0.000188
formato_producto_Bruma	0.000171
marca_Exel	0.000171
material_venta_Carton	0.00016
marca_Derm's	0.000149
marca_Biotherm	0.000121
marca_Ayurdeva's	0.000119
marca_Pond's	0.000118
marca_Caviahue	0.000107
marca_Lancome	0.000105
formato_producto_Papel	9.86E-05
formato_venta_Caja	9.55E-05
zona_aplicacion_Menton	8.83E-05
marca_Neutrogena	7.92E-05
formato_producto_Agua termal	7.38E-05
formato_producto_Toallitas	6.88E-05
marca_Aveno	6.76E-05
formato_producto_Pomada	5.80E-05
marca_Coony	4.16E-05
formato_producto_Arcilla	4.01E-05
marca_Paula's Choice	3.59E-05
formato_producto_Jabon Solido	2.88E-05
formato_venta_Lapiz	2.81E-05
formato_producto_Jabon Liquido	2.50E-05
formato_producto_Cobertura	2.01E-05
marca_Prodermic	2.00E-05
formato_producto_Toalla	1.94E-05
marca_Perpiel	1.58E-05
marca_Carthage	1.51E-05
marca_Tortulan	9.76E-06
zona_aplicacion_Cuello	9.18E-06
marca_Libra	3.73E-06
marca_Adermicina	0
marca_Algabo	0
marca_Arex	0
marca_Artez Westerley	0

marca_Asepxia	0
marca_Avery Rose	0
marca_Avon	0
marca_BIU	0
marca_Bagovit	0
marca_Basicare	0
marca_Beautifull Regalos	0
marca_Biocom	0
marca_Bioderma	0
marca_Bling Pop	0
marca_Botik	0
marca_By Derm	0
marca_By She	0
marca_Calypso	0
marca_Cepage	0
marca_Collage	0
marca_DD2	0
marca_Dekka	0
marca_Dermastore	0
marca_Dorothy Gray	0
marca_Dr Duval	0
marca_Farmaclean	0
marca_Filorga	0
marca_Fiorel'a	0
marca_Formuly Piel	0
marca_Heburn	0
marca_Herbivore	0
marca_Icono	0
marca_Jactan's	0
marca_Jessamy	0
marca_Just	0
marca_Key Elements	0
marca_KoTaping	0
marca_Laboratorio Once	0
marca_Lagos	0
marca_Laikou	0
marca_Lanbena	0
marca_Las Anittas	0
marca_Latour	0
marca_Le Lab de Beaute	0

marca_Libelle	0
marca_Linfar	0
marca_M&Q Regalos	0
marca_MAC	0
marca_Maria T	0
marca_MegaCuper	0
marca_Microsule	0
marca_NAMECO	0
marca_Natura	0
marca_Naturaleza Organica	0
marca_Neostrata	0
marca_Niza	0
marca_Norma Bustos	0
marca_Nort	0
marca_OMS	0
marca_Orihens	0
marca_Orlane	0
marca_Ouroboros	0
marca_Purederm	0
marca_RUH	0
marca_Rtopr	0
marca_SHAGMIE	0
marca_Saiku	0
marca_Selecta	0
marca_Sentida Botanica	0
marca_Silfab	0
marca_Simple & Beauty	0
marca_Sir Fausto	0
marca_Sri Sri	0
marca_Sulderm	0
marca_Teatrical	0
marca_Top Choice	0
marca_Ultracomb	0
marca_Valuge	0
marca_Veganis	0
marca_Veritas	0
marca_Violetta	0
marca_Vitalis Navitas	0
marca_Weleda	0
marca_Xerotic	0

marca_Xulu Cosmeticos	0
marca_YDARIS	0
marca_Zine	0
marca_arv-lab	0
marca_bioclean	0
formato_producto_Agua Termal	0
formato_producto_Agua micelar	0
formato_producto_Barra	0
formato_producto_Capsulas	0
formato_producto_Cepillo	0
formato_producto_Cinta	0
formato_producto_Copas	0
formato_producto_Fango	0
formato_producto_Ice Globes	0
formato_producto_Lapiz	0
formato_producto_Maquina	0
formato_producto_Mousse	0
formato_producto_Pastilla	0
formato_producto_Scrub	0
formato_producto_Unguento	0
formato_venta_Bidon	0
formato_venta_Varios	0
material_venta_Madera	0

Tabla 11. Resultados de la segunda optimización de hiperparámetros para Random Forest imputando la mediana

mean_test_score	std_test_score	param_n_estimators	param_max_depth	param_min_samples_split	param_min_samples_leaf	param_max_features	param_bootstrap
0.636419	0.012752	288	6	9	11	0.955643	TRUE
0.6581	0.015893	187	9	12	11	0.501249	TRUE
0.649695	0.011849	187	6	3	3	0.22858	TRUE
0.663825	0.012994	157	8	13	1	0.100701	FALSE
0.670767	0.011769	114	14	17	10	0.120756	FALSE
0.62972	0.012705	150	5	6	3	0.444216	FALSE
0.645311	0.013686	188	7	5	18	0.505449	TRUE
0.623143	0.010424	287	4	16	2	0.446875	FALSE
0.670158	0.014255	180	14	4	15	0.545659	TRUE
0.617661	0.009061	153	4	7	2	0.69627	FALSE
0.644214	0.013295	261	6	13	18	0.266369	FALSE
0.631912	0.015828	299	6	16	16	0.63811	FALSE
0.649208	0.017096	181	9	17	8	0.568751	FALSE
0.645189	0.015197	256	7	10	13	0.628076	TRUE

0.638002	0.012923	235	3	2	9	0.114073	TRUE
0.668819	0.016591	262	13	4	8	0.278958	FALSE
0.668088	0.013124	127	13	10	7	0.16664	TRUE
0.661632	0.017569	147	11	2	2	0.660968	TRUE
0.658709	0.016757	271	9	4	5	0.701957	TRUE
0.632278	0.01523	200	8	4	1	0.898491	FALSE
0.628624	0.012189	126	5	8	14	0.784707	TRUE
0.66687	0.017638	151	12	8	13	0.122877	TRUE
0.635688	0.012146	242	6	14	7	0.91681	TRUE
0.636541	0.013239	269	6	3	7	0.948568	TRUE
0.664312	0.015048	183	11	13	12	0.194945	TRUE
0.662728	0.012631	228	13	2	7	0.391911	FALSE
0.66687	0.016847	102	13	10	13	0.484397	TRUE
0.65335	0.013658	150	8	6	9	0.559673	TRUE
0.66894	0.015059	161	14	18	9	0.319713	FALSE
0.66553	0.015161	230	12	3	7	0.966203	TRUE
0.64799	0.011753	101	7	18	17	0.356356	TRUE
0.676248	0.011074	228	14	2	5	0.146331	FALSE
0.62704	0.015587	282	4	5	6	0.577841	TRUE
0.663703	0.016013	215	10	6	6	0.313874	TRUE
0.648965	0.018595	196	8	17	16	0.458942	FALSE
0.672838	0.017915	247	14	4	16	0.557379	TRUE
0.663946	0.012286	151	13	8	3	0.298217	FALSE
0.619245	0.011239	203	3	19	7	0.457815	TRUE
0.66553	0.013478	198	13	11	18	0.40696	TRUE
0.673934	0.016198	101	14	18	16	0.257459	TRUE
0.670889	0.015487	198	14	10	9	0.602464	TRUE
0.627893	0.012769	212	5	12	1	0.870692	TRUE
0.622777	0.008889	198	6	17	3	0.908699	FALSE
0.637272	0.01524	153	4	3	18	0.244727	TRUE
0.666261	0.016058	241	11	2	4	0.258333	TRUE
0.643484	0.013561	294	6	8	8	0.429822	TRUE
0.613642	0.010609	257	3	15	19	0.588286	TRUE
0.650548	0.017682	120	8	3	8	0.453788	FALSE
0.60877	0.007661	160	3	6	1	0.619213	FALSE
0.627162	0.010502	271	5	18	17	0.614804	FALSE
0.631181	0.013784	123	7	6	9	0.895145	FALSE
0.614251	0.010489	215	3	2	3	0.597488	TRUE
0.65201	0.014685	227	7	4	1	0.27182	TRUE
0.658222	0.015953	273	11	9	3	0.455222	FALSE

0.616565	0.00987	229	4	11	2	0.653507	FALSE
0.598051	0.010414	275	3	12	9	0.863803	FALSE
0.666261	0.016665	252	12	19	3	0.828425	TRUE
0.662363	0.012049	119	14	14	4	0.560208	FALSE
0.659927	0.016906	214	10	15	16	0.73177	TRUE
0.667479	0.015267	279	14	18	19	0.620978	TRUE
0.651157	0.016888	107	8	6	11	0.372939	FALSE
0.671255	0.012103	295	14	8	5	0.294239	TRUE
0.635688	0.014248	151	6	9	3	0.470219	FALSE
0.617783	0.009471	219	4	18	19	0.56467	FALSE
0.632643	0.012394	239	5	10	5	0.495074	TRUE
0.632887	0.011818	133	3	16	1	0.255965	TRUE
0.641535	0.016686	216	8	14	8	0.743136	FALSE
0.658465	0.016362	175	9	4	5	0.6914	TRUE
0.6257	0.008479	146	5	6	14	0.782061	FALSE
0.646163	0.011727	184	6	17	16	0.257797	TRUE
0.651888	0.012597	168	6	11	5	0.255882	TRUE
0.63106	0.013576	168	4	2	10	0.334804	FALSE
0.636175	0.012638	227	6	3	16	0.654266	TRUE
0.654446	0.010056	262	13	19	12	0.820854	FALSE
0.612546	0.008685	200	3	12	19	0.512428	FALSE
0.624361	0.006709	142	5	11	8	0.838384	FALSE
0.672716	0.012341	134	14	19	2	0.1907	FALSE
0.615956	0.011559	214	3	3	10	0.69015	TRUE
0.660658	0.012442	174	9	14	18	0.206348	TRUE
0.66078	0.017872	200	10	11	4	0.761564	TRUE
0.660658	0.017488	135	10	4	17	0.826151	TRUE
0.600731	0.011121	212	3	13	5	0.798772	FALSE
0.659074	0.019191	212	9	18	9	0.779089	TRUE
0.627527	0.010892	199	6	18	13	0.843812	FALSE
0.643971	0.012037	127	11	8	9	0.862429	FALSE
0.631912	0.015192	216	6	17	15	0.616094	FALSE
0.629963	0.011763	156	5	17	4	0.642197	TRUE
0.644945	0.012908	221	7	15	7	0.544978	TRUE
0.627649	0.012749	245	5	12	11	0.898934	TRUE
0.669306	0.015598	106	14	18	12	0.719227	TRUE
0.663216	0.011953	104	10	3	18	0.165487	FALSE
0.650548	0.013284	208	7	17	8	0.309905	TRUE
0.652741	0.015553	231	8	14	7	0.83152	TRUE
0.641413	0.018965	281	8	13	19	0.737363	FALSE

0.664921	0.013712	225	13	8	19	0.323852	TRUE
0.628502	0.013343	266	4	10	8	0.521795	TRUE
0.628867	0.014229	265	5	2	17	0.625032	TRUE
0.651401	0.014965	160	8	14	13	0.821898	TRUE
0.658465	0.014217	263	7	6	11	0.139061	TRUE
0.628624	0.011539	148	5	16	18	0.972372	TRUE

Tabla 12. Resultados de la optimización de hiperparámetros para HistGradientBoosting

mean_test_score	std_test_score	param_max_iter	param_learning_rate	param_max_depth	param_min_samples_leaf	param_l2_regularization	param_max_leaf_nodes
0.651279	0.005785	171	0.295214	13	7	0.074908	35
0.674056	0.010302	187	0.039992	13	3	0.089167	38
0.655177	0.004612	257	0.300973	14	1	0.004117	16
0.656151	0.00789	188	0.167427	14	10	0.060848	41
0.664434	0.003336	207	0.302127	5	5	0.009333	17
0.670889	0.015245	117	0.061157	9	14	0.121509	18
0.662241	0.004583	229	0.101384	7	15	0.161679	34
0.651035	0.003694	274	0.192999	10	14	0.136653	49
0.670767	0.014492	101	0.064671	6	10	0.078212	20
0.651644	0.013699	143	0.019394	4	10	0.11354	48
0.656273	0.002572	289	0.18937	10	16	0.178965	35
0.655177	0.002969	253	0.171908	10	13	0.149464	39
0.667235	0.013616	188	0.250659	3	9	0.028185	21
0.671864	0.012483	162	0.069615	10	19	0.154449	25
0.669062	0.010939	222	0.228702	3	10	0.141371	19
0.671011	0.019241	171	0.044761	9	2	0.071693	26
0.665286	0.006024	136	0.121245	10	12	0.019082	49
0.668697	0.01435	200	0.276164	3	3	0.127511	28
0.663459	0.00509	126	0.178383	9	15	0.152157	23
0.661754	0.013523	162	0.017626	5	7	0.085508	46
0.658831	0.009255	250	0.178983	6	11	0.095074	29
0.653715	0.006993	286	0.292856	9	2	0.040612	36
0.664799	0.010173	233	0.09869	7	12	0.124871	42
0.650426	0.004717	227	0.277768	13	7	0.037314	42
0.663459	0.006993	289	0.09164	13	9	0.181366	27
0.669428	0.006119	236	0.059397	13	19	0.060956	47
0.665895	0.004551	217	0.083238	13	9	0.053882	46
0.672229	0.017598	151	0.029468	12	7	0.080767	26
0.669671	0.014881	283	0.100263	3	17	0.09945	31
0.66821	0.002867	228	0.133311	9	2	0.195323	33
0.66821	0.009943	259	0.144335	8	11	0.106187	18

0.669793	0.015093	123	0.082648	8	2	0.073931	19
0.66687	0.011315	223	0.170732	3	16	0.126706	20
0.652253	0.004272	263	0.187268	13	3	0.008155	33
0.666139	0.005245	151	0.1146	9	8	0.1619	47
0.667966	0.011874	110	0.291019	3	18	0.077347	42
0.665043	0.00988	106	0.273202	5	16	0.184939	30
0.671133	0.010942	228	0.168895	4	5	0.11104	26
0.658343	0.007237	198	0.19993	11	16	0.180084	33
0.64933	0.006131	219	0.266897	8	11	0.169332	47
0.648599	0.005295	183	0.245602	10	16	0.187127	49
0.66553	0.010815	227	0.209051	4	2	0.020294	32
0.663825	0.009139	260	0.205588	11	1	0.138379	18
0.6743	0.015846	227	0.063647	10	7	0.098779	22
0.662607	0.00956	286	0.102418	14	14	0.144188	36
0.662363	0.007805	281	0.303668	5	2	0.117974	22
0.67028	0.010912	296	0.203531	3	4	0.070016	33
0.666261	0.01123	271	0.203642	3	14	0.004863	42
0.66419	0.011907	136	0.121048	11	17	0.182973	38
0.657125	0.010485	215	0.181688	14	1	0.110553	35
0.66894	0.016994	153	0.060848	11	8	0.063384	17
0.66821	0.013604	250	0.072676	8	10	0.014083	16
0.669184	0.014664	236	0.232231	3	10	0.175475	43
0.653106	0.005798	251	0.253034	11	7	0.161872	27
0.669428	0.012537	107	0.249489	3	14	0.100303	30
0.670767	0.0138	197	0.111399	14	19	0.178001	16
0.659683	0.011173	157	0.172793	8	17	0.09312	36
0.649208	0.006954	287	0.25678	8	8	0.00747	47
0.672716	0.013767	202	0.074746	7	6	0.153999	18
0.66894	0.013659	154	0.169406	5	13	0.010336	30
0.6676	0.006345	244	0.224379	5	5	0.193907	33
0.669793	0.013494	239	0.034262	11	1	0.036867	15
0.670158	0.01175	147	0.081936	8	8	0.08847	39
0.661023	0.003626	291	0.201481	9	3	0.186923	19
0.66687	0.00985	185	0.116792	7	14	0.049546	43
0.672716	0.013413	296	0.038247	6	16	0.01836	26
0.660536	0.005119	199	0.157485	9	5	0.019567	47
0.660414	0.007171	199	0.194755	7	10	0.079701	48
0.66687	0.016382	168	0.160941	3	16	0.125172	18
0.670646	0.017067	227	0.202726	3	12	0.014114	46
0.671255	0.012005	288	0.202986	3	3	0.077634	26

0.662241	0.006348	143	0.205445	13	13	0.164085	43
0.672229	0.012054	134	0.123178	5	17	0.042918	37
0.660536	0.009217	189	0.125619	12	19	0.131145	48
0.671864	0.014714	213	0.045449	7	4	0.056371	25
0.664068	0.005345	237	0.300274	6	3	0.083589	23
0.673203	0.011913	195	0.133785	4	5	0.198101	38
0.657369	0.006049	198	0.213953	9	17	0.047719	44
0.661389	0.010169	185	0.158968	11	19	0.106865	27
0.669062	0.016659	246	0.013251	11	9	0.07784	26
0.665652	0.010584	182	0.295019	6	5	0.063863	30
0.670158	0.010461	152	0.108599	13	16	0.058642	37
0.671376	0.012287	285	0.037362	12	5	0.157924	28
0.657125	0.007828	245	0.173775	13	9	0.147407	26
0.66894	0.007824	106	0.195465	8	8	0.152302	27
0.660171	0.003947	293	0.256558	12	5	0.014553	19
0.672351	0.012523	185	0.184392	3	7	0.046646	31
0.653471	0.00511	258	0.236013	8	12	0.1972	33
0.672229	0.01446	118	0.137267	12	13	0.111681	17
0.653959	0.002016	247	0.150598	12	8	0.002271	42
0.663216	0.011721	230	0.204763	3	17	0.023505	27
0.659683	0.007033	233	0.168511	7	13	0.016578	46
0.650914	0.003496	263	0.277343	8	19	0.008632	49
0.666748	0.006665	148	0.166929	11	15	0.193861	28
0.664921	0.009255	102	0.198267	8	11	0.090908	30
0.654811	0.005718	201	0.295123	10	8	0.056193	26
0.650305	0.004564	285	0.204055	10	5	0.033936	42
0.649086	0.00384	290	0.305863	8	9	0.194879	31
0.653593	0.008929	256	0.254139	14	14	0.061906	27
0.656638	0.005972	296	0.29216	5	2	0.01524	39

Tabla 13. Resultados de la técnica de permutación
para HistGradientBoosting

Característica	Importancia
price_usd	0.082022
price_ml	0.04782
len_title	0.041486
es_libre_de_parabenos_Si	0.034555
volumen_ml	0.01771
marca_La Roche Posay	0.013459
formato_producto_Mascara	0.010974

gama_producto_nominal	0.008916
marca_Eximia	0.008867
marca_Nivea	0.008392
formato_venta_Pack	0.007674
marca_Clinique	0.007406
factor_proteccion	0.007406
marca_The Ordinary	0.007345
marca_Cicatricure	0.006882
funciones_Limpieza	0.006626
gama_marca_nominal	0.006273
marca_Lidherma	0.006163
marca_Acf	0.006017
tipo_piel_Mixta	0.005603
con_vit_c_Si	0.005603
funciones_Anti-edad	0.005554
marca_Eucerin	0.005006
gama_producto	0.004848
material_venta_Vidrio	0.004336
tipo_piel_Seca	0.003837
tipo_piel_Todo tipo de piel	0.003569
formato_venta_Botella	0.003557
es_hipoalergenico_Si	0.00352
unidades_pack	0.003313
es_libre_de_parabenos_No	0.002728
material_venta_Plastico	0.002326
sustentable_Si	0.002217
formato_venta_Pomo	0.002192
formato_venta_Dispenser	0.001985
marca_Isdin	0.001827
marca_Cetaphil	0.001803
formato_producto_Balsamo	0.001644
tipo_piel_Grasa	0.001583
marca_Vichy	0.00151
con_retinol_Si	0.001474
zona_aplicacion_Contorno de ojos	0.001449
momento_dia_Dia/Noche	0.00134
sustentable_No	0.001303
funciones_Iluminadora	0.001303
formato_venta_Ampolla	0.001303
marca_Varios	0.001291

funciones_Varios	0.001242
momento_dia_Dia	0.001194
formato_producto_Polvo	0.001169
marca_Regina	0.001157
formato_venta_Gotero	0.001121
es_hipoalergenico_No	0.001072
marca_Avene	0.00106
marca_Kiehl's	0.001048
formato_producto_Solucion	0.001023
funciones_Hidratante	0.000999
marca_Loreal	0.000926
formato_venta_Tubo	0.000877
formato_producto_Gel	0.000877
formato_producto_Serum	0.000828
marca_Lancome	0.000792
con_vit_c_No	0.000792
formato_venta_Sobre	0.000792
con_acido_hialuronico_No	0.000792
tipo_piel_Sensible	0.000755
momento_dia_Noche	0.000743
material_venta_Varios	0.000743
formato_venta_Spray	0.000719
marca_Garnier	0.000682
funciones_Calmante	0.000646
marca_Estee Lauder	0.000646
con_color_Si	0.000633
formato_producto_Tonico	0.000585
con_acido_hialuronico_Si	0.000585
funciones_Cicatrizante	0.000572
formato_producto_Crema	0.000536
formato_producto_Emulsion	0.000536
con_retinol_No	0.000475
formato_venta_Pote	0.000438
funciones_Anti-acne	0.000414
marca_Neostrata	0.000317
formato_venta_Roll On	0.000305
marca_Goicoechea	0.000292
formato_producto_Agua Micelar	0.000268
marca_Exel	0.000268
marca_Coony	0.000268

marca_Bioderma	0.000256
marca_Derm's	0.000219
zona_aplicacion_Rostro	0.000195
zona_aplicacion_Labios	0.000171
marca_Neutrogena	0.000171
formato_producto_Antifaz	0.000171
marca_Laca	0.000158
marca_Idraet	0.000134
tipo_piel_Maduras	9.74E-05
marca_Dermaglos	9.74E-05
marca_Caviahue	8.53E-05
con_protector_Si	8.53E-05
formato_producto_Locion	6.09E-05
funciones_Anti-manchas	6.09E-05
con_color_No	4.87E-05
funciones_Protector Solar	3.65E-05
marca_Biotherm	2.44E-05
formato_producto_Varios	1.22E-05
zona_aplicacion_Menton	0
zona_aplicacion_Cuello	0
material_venta_Madera	0
material_venta_Carton	0
marca_bioclean	0
marca_arv-lab	0
marca_Zine	0
marca_YDARIS	0
marca_Xulu Cosmeticos	0
marca_Xerotic	0
marca_Weleda	0
marca_Vitalis Navitas	0
marca_Violetta	0
marca_Veritas	0
marca_Veganis	0
marca_Ultracomb	0
marca_Tortulan	0
marca_Top Choice	0
marca_Teatrical	0
marca_Sulderm	0
marca_Sri Sri	0
marca_Sir Fausto	0

marca_Simple & Beauty	0
marca_Silfab	0
marca_Sentida Botanica	0
marca_Selecta	0
marca_Saiku	0
marca_SHAGMIE	0
marca_Rtopr	0
marca_RUH	0
marca_Purederm	0
marca_Prodermic	0
marca_Perpiel	0
marca_Paula's Choice	0
marca_Ouroboros	0
marca_Orlane	0
marca_Orihens	0
marca_OMS	0
marca_Nort	0
marca_Norma Bustos	0
marca_Niza	0
marca_Naturaleza Organica	0
marca_Natura	0
marca_NAMECO	0
marca_Microsule	0
marca_MegaCuper	0
marca_Maria T	0
marca_MAC	0
marca_M&Q Regalos	0
marca_Linfar	0
marca_Libra	0
marca_Libelle	0
marca_Le Lab de Beaute	0
marca_Latour	0
marca_Las Anittas	0
marca_Lanbena	0
marca_Laikou	0
marca_Lagos	0
marca_Laboratorio Once	0
marca_Konjac	0
marca_KoTaping	0
marca_Key Elements	0

marca_Just	0
marca_Jessamy	0
marca_Jactan's	0
marca_Icono	0
marca_Herbivore	0
marca_Heburn	0
marca_Formuly Piel	0
marca_Fiorel'a	0
marca_Filorga	0
marca_Farmaclean	0
marca_Dr Duval	0
marca_Dorothy Gray	0
marca_Dermastore	0
marca_Dekka	0
marca_DD2	0
marca_Collage	0
marca_CeraVe	0
marca_Cepage	0
marca_Carthage	0
marca_Calypso	0
marca_By She	0
marca_By Derm	0
marca_Botik	0
marca_Bling Pop	0
marca_Biocom	0
marca_Biobellus	0
marca_Beautifull Regalos	0
marca_Basicare	0
marca_Bagovit	0
marca_BIU	0
marca_Ayurdeva's	0
marca_Avon	0
marca_Avery Rose	0
marca_Aveno	0
marca_Asepxia	0
marca_Artez Westerley	0
marca_Arex	0
marca_Algabo	0
marca_Adermicina	0
funciones_Exfoliante	0

formato_venta_Varios	0
formato_venta_Lapiz	0
formato_venta_Caja	0
formato_venta_Bidon	0
formato_producto_Unguento	0
formato_producto_Toallitas	0
formato_producto_Toalla	0
formato_producto_Scrub	0
formato_producto_Pomada	0
formato_producto_Pastilla	0
formato_producto_Papel	0
formato_producto_Pad	0
formato_producto_Mousse	0
formato_producto_Maquina	0
formato_producto_Leche	0
formato_producto_Lapiz	0
formato_producto_Jabon Solido	0
formato_producto_Jabon Liquido	0
formato_producto_Ice Globes	0
formato_producto_Fango	0
formato_producto_Copas	0
formato_producto_Cobertura	0
formato_producto_Cinta	0
formato_producto_Cepillo	0
formato_producto_Capsulas	0
formato_producto_Bruma	0
formato_producto_Barra	0
formato_producto_Arcilla	0
formato_producto_Agua termal	0
formato_producto_Agua micelar	0
formato_producto_Agua Termal	0
con_protector_No	0
marca_Pond's	-1.22E-05
resistente_al_agua_Si	-1.22E-05
resistente_al_agua_No	-1.22E-05
formato_producto_Espuma	-1.22E-05
marca_Valuge	-3.65E-05
formato_producto_Esponja	-6.09E-05
formato_producto_Aceite	-8.53E-05
material_venta_Metal	-0.00013

Tabla 14. Resultados de la segunda optimización de hiperparámetros para HistGradientBoosting

mean_test_score	std_test_score	param_max_iter	param_learning_rate	param_max_depth	param_min_samples_leaf	param_l2_regularization	param_max_leaf_nodes
0.648721	0.009793	171	0.295214	13	7	0.074908	35
0.672107	0.01315	187	0.039992	13	3	0.089167	38
0.654568	0.008787	257	0.300973	14	1	0.004117	16
0.661754	0.007534	188	0.167427	14	10	0.060848	41
0.668453	0.006853	207	0.302127	5	5	0.009333	17
0.670158	0.01441	117	0.061157	9	14	0.121509	18
0.665043	0.008884	229	0.101384	7	15	0.161679	34
0.652253	0.003871	274	0.192999	10	14	0.136653	49
0.672594	0.017771	101	0.064671	6	10	0.078212	20
0.651157	0.01269	143	0.019394	4	10	0.11354	48
0.648965	0.007786	289	0.18937	10	16	0.178965	35
0.656151	0.003706	253	0.171908	10	13	0.149464	39
0.66821	0.009493	188	0.250659	3	9	0.028185	21
0.671011	0.013002	162	0.069615	10	19	0.154449	25
0.673203	0.010926	222	0.228702	3	10	0.141371	19
0.671255	0.016645	171	0.044761	9	2	0.071693	26
0.659562	0.007818	136	0.121245	10	12	0.019082	49
0.674421	0.010425	200	0.276164	3	3	0.127511	28
0.663581	0.012018	126	0.178383	9	15	0.152157	23
0.660171	0.013999	162	0.017626	5	7	0.085508	46
0.659683	0.010958	250	0.178983	6	11	0.095074	29
0.647747	0.006689	286	0.292856	9	2	0.040612	36
0.664312	0.011383	233	0.09869	7	12	0.124871	42
0.643605	0.003345	227	0.277768	13	7	0.037314	42
0.666504	0.010871	289	0.09164	13	9	0.181366	27
0.671742	0.008578	236	0.059397	13	19	0.060956	47
0.663094	0.008398	217	0.083238	13	9	0.053882	46
0.67162	0.018019	151	0.029468	12	7	0.080767	26
0.668331	0.014129	283	0.100263	3	17	0.09945	31
0.664921	0.006884	228	0.133311	9	2	0.195323	33
0.666017	0.010063	259	0.144335	8	11	0.106187	18
0.670037	0.015127	123	0.082648	8	2	0.073931	19
0.670889	0.011944	223	0.170732	3	16	0.126706	20
0.655055	0.003863	263	0.187268	13	3	0.008155	33
0.668819	0.004548	151	0.1146	9	8	0.1619	47
0.671742	0.015489	110	0.291019	3	18	0.077347	42
0.665043	0.013102	106	0.273202	5	16	0.184939	30

0.668697	0.011387	228	0.168895	4	5	0.11104	26
0.657856	0.006083	198	0.19993	11	16	0.180084	33
0.656151	0.005523	219	0.266897	8	11	0.169332	47
0.655786	0.005329	183	0.245602	10	16	0.187127	49
0.663094	0.006083	227	0.209051	4	2	0.020294	32
0.664921	0.010808	260	0.205588	11	1	0.138379	18
0.673447	0.013527	227	0.063647	10	7	0.098779	22
0.662728	0.005808	286	0.102418	14	14	0.144188	36
0.658587	0.007435	281	0.303668	5	2	0.117974	22
0.671864	0.009951	296	0.203531	3	4	0.070016	33
0.669062	0.013632	271	0.203642	3	14	0.004863	42
0.668331	0.007202	136	0.121048	11	17	0.182973	38
0.661754	0.002998	215	0.181688	14	1	0.110553	35
0.669306	0.013925	153	0.060848	11	8	0.063384	17
0.670037	0.013658	250	0.072676	8	10	0.014083	16
0.66821	0.012183	236	0.232231	3	10	0.175475	43
0.655907	0.004887	251	0.253034	11	7	0.161872	27
0.664799	0.012579	107	0.249489	3	14	0.100303	30
0.673082	0.013007	197	0.111399	14	19	0.178001	16
0.660292	0.007954	157	0.172793	8	17	0.09312	36
0.646041	0.008625	287	0.25678	8	8	0.00747	47
0.672229	0.015634	202	0.074746	7	6	0.153999	18
0.669915	0.010577	154	0.169406	5	13	0.010336	30
0.661876	0.00537	244	0.224379	5	5	0.193907	33
0.669915	0.015115	239	0.034262	11	1	0.036867	15
0.672107	0.011207	147	0.081936	8	8	0.08847	39
0.65944	0.005872	291	0.201481	9	3	0.186923	19
0.666261	0.008587	185	0.116792	7	14	0.049546	43
0.675274	0.015888	296	0.038247	6	16	0.01836	26
0.661876	0.005808	199	0.157485	9	5	0.019567	47
0.659318	0.008048	199	0.194755	7	10	0.079701	48
0.666626	0.013869	168	0.160941	3	16	0.125172	18
0.672107	0.014811	227	0.202726	3	12	0.014114	46
0.671133	0.011638	288	0.202986	3	3	0.077634	26
0.656395	0.00564	143	0.205445	13	13	0.164085	43
0.672473	0.016954	134	0.123178	5	17	0.042918	37
0.66553	0.010445	189	0.125619	12	19	0.131145	48
0.676492	0.013647	213	0.045449	7	4	0.056371	25
0.658709	0.004159	237	0.300274	6	3	0.083589	23
0.67503	0.01279	195	0.133785	4	5	0.198101	38

0.657978	0.006521	198	0.213953	9	17	0.047719	44
0.663094	0.008504	185	0.158968	11	19	0.106865	27
0.670524	0.016983	246	0.013251	11	9	0.07784	26
0.655055	0.006678	182	0.295019	6	5	0.063863	30
0.665652	0.007081	152	0.108599	13	16	0.058642	37
0.670037	0.012454	285	0.037362	12	5	0.157924	28
0.657247	0.003762	245	0.173775	13	9	0.147407	26
0.670767	0.008906	106	0.195465	8	8	0.152302	27
0.655542	0.005401	293	0.256558	12	5	0.014553	19
0.671864	0.013166	185	0.184392	3	7	0.046646	31
0.654202	0.006107	258	0.236013	8	12	0.1972	33
0.674056	0.01178	118	0.137267	12	13	0.111681	17
0.657978	0.001569	247	0.150598	12	8	0.002271	42
0.671011	0.012818	230	0.204763	3	17	0.023505	27
0.661023	0.005016	233	0.168511	7	13	0.016578	46
0.650548	0.00684	263	0.277343	8	19	0.008632	49
0.666017	0.012796	148	0.166929	11	15	0.193861	28
0.66285	0.007514	102	0.198267	8	11	0.090908	30
0.651035	0.005351	201	0.295123	10	8	0.056193	26
0.65408	0.007577	285	0.204055	10	5	0.033936	42
0.651401	0.008131	290	0.305863	8	9	0.194879	31
0.649695	0.00537	256	0.254139	14	14	0.061906	27
0.664434	0.004948	296	0.29216	5	2	0.01524	39