COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales (ICADE)

ANÁLISIS PREDICTIVO EN EL MARKETING DIRIGIDO: UN ENFOQUE BASADO EN LA INTELIGENCIA ARTIFICIAL

Autor: Alejandra García Boullosa

Director: Raúl González Fabre

MADRID | Abril, 2024

Resumen

El trabajo explora la convergencia de la inteligencia artificial (IA) con el campo del marketing dirigido, enfatizando la capacidad predictiva del 'machine learning' en relación con la segmentación del mercado y la personalización de la experiencia del consumidor. Se realiza un análisis de la literatura, que establece un marco teórico en el que se sitúan la IA y sus implicaciones en el marketing contemporáneo, así como el análisis predictivo y las técnicas de aprendizaje automático mayormente utilizadas en esta herramienta predictiva. Asimismo, se complementa esta teoría con la ilustración de dos casos de éxito empresariales donde se implementan de forma efectiva herramientas analíticas avanzadas. El abordaje metodológico combina el estudio cualitativo de experiencias empresariales con el tratamiento cuantitativo de datos, estudiando empíricamente cómo las técnicas predictivas pueden optimizar la relación entre marca y consumidor. El análisis concluye resaltando la indispensable integración de la IA en el marketing como catalizador de innovación, subrayando que si bien su aplicación, no está exenta de desafíos técnicos y éticos, es crucial para la personalización y eficiencia en un mercado cada vez más impulsado por datos.

Palabras clave: inteligencia artificial; marketing dirigido; machine learning; análisis predictivo; segmentación del consumidor; personalización; estrategias de marketing.

Abstract

The paper explores the convergence of artificial intelligence (AI) with the field of targeted marketing, emphasizing the predictive capability of 'machine learning' in relation to market segmentation and personalization of the consumer experience. A literature review is conducted, which establishes a theoretical framework in which AI and its implications in contemporary marketing are situated, as well as the predictive analytics and 'machine learning' techniques most commonly used in this predictive tool. It also complements this theory with the illustration of two business success cases where advanced analytical tools are effectively implemented. The methodological approach combines the qualitative study of business experiences with quantitative data processing, empirically studying how predictive techniques can optimize the relationship between brand and consumer. The analysis concludes by highlighting the indispensable integration of AI in marketing as a catalyst for innovation, underlining that while its application is not without technical and ethical challenges, it is crucial for personalization and efficiency in an increasingly data-driven marketplace.

Keywords: artificial intelligence; targeted marketing; machine learning; predictive analytics; consumer segmentation; personalization; marketing strategies.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

LISTADO DE ABREVIATURAS

AI Artificial intelligence

AR Augmented Reality

IoT Internet of Things

KNN K-Nearest Neighbour

IA Inteligencia Artificial

ML Machine Learning

MSE Mean Squared Error

NLP Natural Language Processing

PCA Principal Component Analysis

ROI Return on Investment

SVM Support Vector Machines

VR Virtual Reality

CAPITULO I. INTRODUCCIÓN

1. CONTEXTO Y JUSTIFICACIÓN DEL TEMA

En el dilatado mundo de la tecnología, la inteligencia artificial emerge como una protagonista fascinante que desafía las fronteras de la imaginación y redefine la forma en que interactuamos con la información.

La inserción de la inteligencia artificial en la cotidianidad de nuestras vidas ha transcendido lo que una vez fue catalogado propio de un mundo de ciencia ficción, transformándose en una realidad innegable. Los límites que permitían discernir en un pasado la humanidad de la tecnología se han visto disipados dando paso a una nueva normalidad donde ambos conceptos conviven y se complementan.

Su irrupción ha marcado la revolución tecnológica que el mundo de los datos lleva experimentando la última década. Un crecimiento vertiginoso sumado a un extraordinario potencial para transformar la economía y la sociedad han generado que la inteligencia artificial haya sido establecida como un eje estratégico de la agenda España Digital 2026 (Observatorio Nacional de Tecnología y Sociedad, 2023).

El desarrollo de algoritmos más complejos y el incremento en la capacidad de procesamiento de datos han generado nuevas oportunidades de aplicación para la inteligencia artificial en diversos campos disciplinares.

En el contexto empresarial, la integración de la inteligencia artificial se ha posicionado como un factor disruptivo y transformador. Las empresas se enfrentan a una nueva realidad donde la capacidad para aprovechar los beneficios de la inteligencia artificial se convierte en un componente distintivo entre la permanencia y el liderazgo en el mercado.

Las herramientas facilitadas por la inteligencia artificial proporcionan un marco para abordar desafíos magnos desde la toma de decisiones eficientes y el desarrollo de propuestas competitivas hasta la optimización de procesos, una reestructuración de sus modelos de negocios y el impulso de iniciativas innovadoras.

Dentro de esta coyuntura empresarial el marketing, en consonancia con los instrumentos facilitados por la inteligencia artificial, debe proporcionar enfoques resolutivos y atípicos a las diversas dinámicas de un mercado caracterizado por la competitividad y la multiplicidad de datos.

Un estudio publicado por Randstad Research a comienzos de 2024 referente a las perspectivas de las empresas y los profesionales con respecto a la incorporación de la inteligencia artificial en el ecosistema empresarial, respalda la integración de la inteligencia artificial como una estrategia esencial en el ámbito laboral contemporáneo (Randstad, 2024).

La suficiencia de esta herramienta permite tanto la consecución de una mejora en las predicciones, como una optimización de las operaciones y la personalización de determinados servicios. Estos beneficios generan resultados tan evidentes que un porcentaje significativo de las 300 empresas españolas y más de 1.500 profesionales encuestados para el estudio, en concreto un 45.5%, ya han incorporado la inteligencia artificial en sus dinámicas de negocio (Randstad, 2024).

Entre los diversos sectores del tejido empresarial, la optimización de las acciones de Marketing (27.9%) y el análisis de datos y predicción (49.2%), se erigen como dos de las múltiples funcionalidades sobre las que las entidades corporativas hacen uso de la inteligencia artificial (Randstad, 2024). La toma de decisiones más inteligentes basadas en datos emerge como una necesidad crucial para mantenerse dentro del ya establecido mercado competitivo.

Este imperativo precisa que las compañías de marketing aprovechen la capacidad del análisis predictivo para analizar cantidades masivas de datos y transformar los mismos en estrategias empresariales capaces de prever tendencias, comprender los comportamientos actuales y futuros de los consumidores y por ende, mejorar notablemente la eficacia de las campañas publicitarias.

2. OBJETIVOS

El presente trabajo de investigación tiene como objetivo principal examinar la implementación de la inteligencia artificial en el campo del marketing a través de las técnicas de análisis predictivo.

Los objetivos secundarios de este trabajo se sintetizan en:

- Explorar y definir los conceptos de inteligencia artificial y marketing dirigido, proporcionando una base teórica para la comprensión de la investigación.

- Investigar, a través del análisis y resultados de diversos estudios, literatura e informes, la aplicación de la inteligencia artificial en el contexto empresarial, tanto a nivel global.

- Examinar el impacto del 'machine learning' como herramienta principal de la inteligencia artificial en la resolución de desafíos en el marketing dirigido explorando, en concreto, la aplicación del análisis predictivo.

- Ilustrar mediante casos de éxito empresariales la implementación real y efectiva de las herramientas del análisis predictivo en el sector del marketing.

- Implementar técnicas de aprendizaje automático en una base de datos concreta para estudiar las aplicaciones reales del análisis predictivo en la mejora de la eficacia de las estrategias de marketing.

3. METODOLOGÍA

Con el propósito de alcanzar los objetivos presentados en esta investigación, se ha optado por seguir un enfoque deductivo y una metodología de índole tanto cualitativa como cuantitativa.

La metodología de investigación cualitativa se estructurará en los siguientes pasos: Como punto de partida se realiza una revisión bibliográfica en el contexto de investigaciones previas para establecer un fundamento teórico firme a través de la exposición de nociones fundamentales de los conceptos objeto del tema de investigación.

De la misma forma, se recopilan informes relevantes y representativos de diversas fuentes, incluyendo estudios de casos existentes. Dado que se trata de examinar las aplicaciones de la inteligencia artificial en el marketing dirigido, estos datos son recolectados con el fin de obtener una perspectiva crítica sobre el tema tratado.

Para las etapas ulteriores del proyecto, se aplica la metodología de investigación cuantitativa. En virtud de que su fin es analizar el comportamiento y el perfil de los clientes, y así potenciar las estrategias de marketing, esta metodología se erige como la más pertinente. La misma está fundamentada en la recopilación y el análisis de datos numéricos con miras a obtener conclusiones objetivas y aplicables.

Para ello, se selecciona una base de datos significativa para la investigación, así como dos técnicas de aprendizaje supervisado. Estos mismo datos son analizados de forma predictiva para identificar posibles patrones que contribuyan a la mejora de la eficiencia de las estrategias de marketing.

Por último, se analizan las implicaciones tanto teóricas como prácticas de los resultados. Se sintetizan asimismo las conclusiones del estudio, poniendo de relieve su importancia para la comprensión de la aplicación de la inteligencia artificial en el marketing dirigido.

4. DESARROLLO

Este trabajo se estructura en cinco capítulos.

En el primer capítulo se presenta la introducción del presente trabajo de investigación.

En el capítulo segundo se proporciona un marco teórico relacionado con las diversas concepciones de la inteligencia artificial. Se realiza un análisis que abarca desde la génesis de este concepto hasta su concepción actual. Del mismo modo, se examina desde un enfoque generalizado la implementación de la inteligencia artificial en el contexto empresarial.

El tercer capítulo, profundiza en la adopción de la inteligencia artificial en el sector del marketing, abordando el papel del análisis predictivo dentro de este contexto. Esta sección inicia con unas consideraciones previas acerca de cómo la inteligencia artificial está transformando el campo del marketing. A continuación, se introduce la noción del marketing dirigido. Posteriormente, se fundamenta la concepción del análisis predictivo, así como se explora su aplicación en el marketing a través del estudio de diferentes técnicas de aprendizaje supervisado y no supervisado propias del análisis predictivo. El capítulo termina con el análisis de dos estudios de caso, los cuales demuestran de manera

concreta el éxito y el impacto positivo resultante de la integración de métodos de análisis predictivo en las estrategias empresariales.

No obstante, esta pesquisa busca lograr trascender de un mero análisis documental del provecho de esta técnica de inteligencia artificial en el ámbito del marketing. Por ello, en el cuarto capítulo, a través de la realización de un caso práctico se implementan y analizan dos técnicas de aprendizaje automático supervisado asociadas al análisis predictivo, una regresión lineal y un random forest, en una base de datos que contiene diversas variables en relación con el perfil de los consumidores. Asimismo, los resultados alcanzados derivados de la aplicación práctica son interpretados.

En la fase final del trabajo, en el quinto capítulo, se extraen conclusiones que destacan el valor agregado de la inteligencia artificial en el ámbito del marketing, así como se aprecian las limitaciones encontradas en el estudio y las futuras líneas de investigación del mismo.

CAPITULO II. FUNDAMENTOS TEÓRICOS

1. INTELIGENCIA ARTIFICIAL

1.1. Conceptos fundamentales

Aludir a la inteligencia artificial implica hacer referencia a un concepto inconcluso, desigual; por un lado, podemos encontrar definiciones como la proporcionada por la Real Academia Española que define a la inteligencia artificial como la "*disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico*" (Real Academia Española, s.f.).

El Parlamento Europeo en la enmienda 18 considerando 6 de su proyecto de resolución legislativa sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial), en cambio, determina que la inteligencia artificial es "*un sistema basado en máquinas diseñado para funcionar con diversos niveles de autonomía y capaz, para objetivos explícitos o implícitos, de generar información de salida, como*

predicciones, recomendaciones o decisiones, que influya en entornos reales o virtuales" (Parlamento Europeo, 2021).

John McCarthy, ilustre informático que realizó importantes contribuciones en el campo de la inteligencia artificial, establece que la inteligencia artificial "*es la ciencia y la ingeniería para crear máquinas inteligentes, especialmente programas informáticos inteligentes. Está relacionada con la tarea similar de utilizar ordenadores para comprender la inteligencia humana, pero la inteligencia artificial no tiene por qué limitarse a métodos que sean biológicamente observables*" (McCarthy, 2007)

Numerosas son las definiciones que encontramos, así como abundantes son los retos que la inteligencia artificial genera para establecer una única definición que esclarezca esta noción. El estado de cambio constante como consecuencia de los diversos avances tecnológicos, así como la complejidad de sus técnicas; la discrepancia entre los expertos en la materia en relación con la constitución exacta del término; o los dilemas éticos y sociales que suscitan, forman parte de los desafíos actuales. No obstante, a pesar de estas dificultades, es posible identificar ciertos aspectos en común en las distintas formulaciones del término.

Por una parte, encontramos que la inteligencia artificial se rige por la autonomía. En este contexto, al hablar de autonomía entendemos la capacidad que tienen las máquinas de tomar decisiones y desempeñar funciones de manera independiente sin el requerimiento de la presencia humana ya sea de forma manual o intelectual.

Ligado con esta primera propiedad, se distingue asimismo el aprendizaje automático. Previo a lograr realizar tareas indistintamente, las máquinas por medio de algoritmos aprenden y mejoran a ejecutar las mismas, sin precisar ser programadas expresamente. Esta capacidad de aprendizaje en el contexto de la transformación digital actual y del incremento de grandes volúmenes de datos, ha impulsado al aprendizaje automático o *'machine learning'* a erigirse como la principal herramienta de la inteligencia artificial para abordar desafíos y proporcionar soluciones en múltiples ámbitos, incluido el marketing.

El 'machine learning' se fundamenta en la idea de que un sistema informático pueda aprender y adaptarse a nuevos datos de manera autónoma, sin necesidad de intervención

humana (Frankenfield, 2022). Su funcionamiento está basado en algoritmos y modelos que analizan e identifican patrones en los datos con el propósito de realizar predicciones o adoptar decisiones, sin requerir una programación explícita para ello. Dentro de la rama del *'machine learning'*, distinguimos a su vez tres modelos:

- Aprendizaje supervisado. Se caracteriza por emplear conjuntos de datos conocidos (datos de entrenamiento) para entrenar a los algoritmos en la clasificación de datos o la predicción de rigurosos resultados.

- Aprendizaje no supervisado. Técnica en la que los modelos no aprenden a partir de los datos entrenamiento sino que, mediante conjuntos de datos sin etiquetar, son los propios modelos los que de forma autónoma identifican patrones encubiertos en los datos a analizar.

- *Reinforcement learning* (Aprendizaje por esfuerzo). Técnica en la que los modelos aprenden a realizar una tarea mediante interacciones repetitivas de ensayo y error dentro de un entorno cambiante. Esta metodología de aprendizaje facilita la toma de decisiones por parte del modelo que incrementen al máximo un indicador de recompensa por el trabajo realizado, de manera autónoma y sin estar previamente preparado para la ejecución de la tarea.

En suma, en virtud de la esencialidad de las interpretaciones, la inteligencia artificial se reduciría a la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones autónomas (Rouhiainen, 2018).

1.2. Evolución histórica

Si bien los hitos y avances más relevantes en lo concerniente a la inteligencia artificial han acontecido en la última década, los mismos no son más que los frutos de una larga serie de progresos y estudios que han culminado en lo que hoy en día conocemos como inteligencia artificial.

Para establecer el origen de la inteligencia artificial, hemos de remontarnos al año 1940 cuando Alan Turing, un matemático británico considerado como uno de los

científicos más extraordinarios del siglo XX, estableció la llamada "Teoría de la Computación y Máquinas de Turing". La misma se basaba en el principio de que una máquina podía imitar a cualquier otra máquina, es decir, podía realizar cualquier tarea computable. Turing planteó una prueba elemental para probar la habilidad de una máquina para exhibir un comportamiento inteligente, semejante al de un ser humano

Una década más tarde, en el verano de 1956, se llevó a cabo una reunión en el Dartmouth College de Hanover. En este evento participaron un grupo de diez matemáticos y lógicos, entre ellos pioneros como John McCarthy, Marvin Minsky y Claude Shannon. La premisa central de este encuentro era la idea de que los rasgos característicos de la inteligencia humana podían ser descritos de forma tan precisa que pudieran ser simulados por un computador (Wachsmuth, 2000).

Fue en dicha reunión, conocida desde entonces como "la Conferencia de Dartmouth", donde se habló por primera vez del término "inteligencia artificial".

En el transcurso de los años 1960 y 1970, se llevó a cabo el diseño y desarrollo de sistemas expertos mediante la aplicación de la tecnología, matemáticas y la lógica. En 1960 Marvin Minsky en el MIT y John McCarthy en Stanford, divergen en sus enfoques sobre la IA, el primero enfocado en hacer que los programas funcionen y el segundo en la representación y razonamiento lógico. En este mismo año, se destacan avances significativos en redes neuronales, incluyendo el trabajo de Widrow y Hoff con adalines y el de Frank Rosenblatt con perceptrones, demostrando el aprendizaje automático en estas estructuras (Russell & Norvig, 2004).

En los años 80, el éxito comercial de sistemas expertos como R1 de *Digital Equipment Corporation* marca el inicio de la IA como una industria, impulsada por proyectos significativos en Japón y Estados Unidos. Asimismo, se produce un resurgimiento de interés por las redes neuronales gracias a nuevos algoritmos de aprendizaje y aplicaciones prácticas, así como se introduce un enfoque más riguroso y científico en la investigación de la IA, integrando métodos de campos establecidos como la estadística y la teoría de la información, y avanzando en áreas como el reconocimiento del habla y la minería de datos (Russell & Norvig, 2004).

En torno a los años 1990 y 2000, la inteligencia artificial se empezó a implementar en numerosos campos disciplinares a razón de los avances producidos en la década pasada, así como al auge de la disponibilidad de cantidades masivas de datos (BigData), adquiriendo significativa relevancia en la sociedad. En concreto, 1997 se estima como el año en el que se evidenció un momento de transformación impulsado por la Inteligencia artificial cuando el ordenador Deep Blue de IBM fue capaz de ganar al célebre y campeón mundial de ajedrez, Garry Kasparov (Abeliuk y Gutiérrez, 2021).

En la actualidad nos encontramos en la era denominada como Industria 4.0, también llamada cuarta revolución industrial Esta era se caracteriza por la premura en la creación de nuevas herramientas de inteligencia artificial que puedan ser aplicadas en todos los entornos posibles, así como por la celeridad en los progresos tecnológicos marcados por la mejora en el procesamiento del Big Data, capaz de llevar a cabo análisis de complejos modelos de aprendizajes y millones de parámetros.

2. INTELIGENCIA ARTIFICIAL EN EL CONTEXTO EMPRESARIAL

La inteligencia artificial ha emergido como una poderosa herramienta que se ha incorporado en un amplia gama de funciones empresariales, transformando la forma en que las organizaciones modernizan sus operaciones y sistemas, gestionan la toma de decisiones estratégicas y personalizan la experiencia del cliente.

La disrupción de esta tecnología abre un nuevo panorama en el modelo empresarial y desarrollo de las actividades productivas de las organizaciones. Las organizaciones están explorando el potencial que ofrece la inteligencia artificial para desencadenar valor en el ámbito empresarial, incrementar tanto la eficacia como la productividad, y propiciar la creación de nuevos productos, servicios y modelos de negocio. En este sentido, los resultados del estudio ‘Trust in artificial intelligence. 2023 Global study on the shifting public perceptions of IA’, publicado por KPMG en colaboración con la Universidad de Queensland en octubre de 2023 que recoge las opiniones de 17.000 personas de 17 países de diferentes regiones del mundo, demuestran que el 85% de los encuestados consideran que la IA ofrece diversos beneficios como la mejora de la eficiencia, innovación, eficacia, utilización de recursos y costes reducidos (Gillespie et al., 2023).

Sin embargo, cabe mencionar que el 73% de los encuestados globales están preocupados por los riesgos potenciales de la IA. Estos riesgos incluyen violaciones de la ciberseguridad y la privacidad, manipulación y uso nocivo de los datos o pérdida de puestos de trabajo, entre otros (Gillespie et al., 2023).

En cuanto a la implementación de la inteligencia artificial en las tareas empresariales, el informe "IBM Global AI Adoption Index 2023", elaborado por la multinacional tecnológica IBM y publicado en noviembre de 2023, recoge las opiniones de 2.343 profesionales del sector de TI a nivel global acerca de la adopción de la inteligencia artificial en sus empresas. Dentro de los hallazgos del estudio, se revela que a medida que la IA asume diversas funciones dentro de las empresas que la implementan, entre las aplicaciones más comunes de esta tecnología encontramos la automatización de procesos de TI (33%); seguridad y detección de amenazas (26%); automatización de procesos de negocio (22%); marketing y ventas (22%) y la toma de decisiones predictivas (18%), entre otros (IBM, 2023).

Por otro lado, dentro del 42% de los encuestados que ya han implementado la IA como parte de sus operaciones comerciales en sus respectivas empresas, el 80% de los profesionales valoran el cumplimiento de ciertos aspectos y valores éticos cuando se lleva a cabo la adopción de esta tecnología. Entre estos valores se destaca mantener la integridad de su marca y la confianza de sus clientes (87%), así como cumplir con las obligaciones regulatorias y de cumplimiento externas (87%). Del mismo modo, la salvaguarda de la privacidad de los datos del cliente (44%), el desarrollo de políticas éticas de la IA (44%) y el seguimiento de la procedencia y los cambios de los datos (37%), entre otros, son las formas más comunes en que las empresas garantizan una IA confiable (IBM, 2023).

En relación con las perspectivas futuras de implantación, según los datos procedentes de la encuesta ‘Future of Jobs Report 2023’ realizada por el World Economic Forum y publicada en mayo de 2023 que recoge las opiniones y perspectivas de 803 compañías de todas las regiones del mundo, el 74.9% de las empresas encuestadas prevén adoptar la inteligencia artificial en los próximos cinco años (Word Economic Forum, 2023).

Respecto de los sectores líderes en la utilización de inteligencia artificial en España en 2022, observamos que la adopción de esta tecnología adquiere una elevada presencia

en el sector de información y comunicaciones, con una implantación del 41.9% de la IA en sus actividades. Le sigue, muy de cerca, el sector de las tecnologías de la información y las comunicaciones (TIC) con un porcentaje del 41.3%. Respecto del tercer sector con más fuerza, encontramos las actividades profesionales, científicas y técnicas con un porcentaje considerablemente alejado respecto del segundo sector, del 21.8%. Entre los sectores en los que la incorporación de la inteligencia artificial es menos notoria, distinguimos el sector de alimentación, bebidas, tabaco y textil, con un porcentaje de instauración del 7%, y el sector de la construcción, con tan solo el 6.3% (Observatorio Nacional de Tecnología y Sociedad, 2023).

CAPITULO III. INTELIGENCIA ARTIFICIAL EN EL MARKETING

1. CONSIDERACIONES PREVIAS

El marketing está experimentando un rápido desarrollo, con alteraciones constantes tanto en el diseño como en los métodos de análisis utilizados por los investigadores en este campo. Estos cambios están influenciados por las transformaciones en las capacidades de gestión, tecnologías de la información, y principalmente, en el comportamiento del consumidor. Para poder adaptarse a esta evolución tecnológica, es esencial dotarse de las herramientas adecuadas, siendo la inteligencia artificial una de las más notables y fundamentales en esta era de transformación.

La implementación de la inteligencia artificial en el campo del marketing tiene el potencial de revolucionar y redefinir el enfoque tradicional que las agencias de publicidad adoptan en sus campañas. La tecnología de la IA, al estar habilitada para analizar datos a gran escala y predecir tendencias de consumo, es capaz de proporcionar recomendaciones estratégicas que los profesionales encargados de llevar el sector del marketing en las empresas podrían utilizar para informar y guiar sus decisiones futuras (Gaikwad & Gautam, 2023).

Estas decisiones influirán sustancialmente en cómo las marcas se comunican y relacionan con su audiencia, potenciando una interacción más individualizada y efectiva con los consumidores mejorando, consecuentemente, la eficiencia operativa y proporcionando *insights* más profundos sobre el comportamiento de estos últimos.

En este sentido el 'machine learning' como disciplina de la inteligencia artificial, juega un papel fundamental en potenciar y optimizar las aplicaciones de la IA en el ámbito del marketing. Mediante algoritmos que aprenden de los datos, el ML permite que las estrategias de marketing se vuelvan más inteligentes y eficientes con el tiempo. En relación con estas aplicaciones, encontramos (Gaikwad & Gautam, 2023):

- Segmentación de clientes. El *'machine learning'* analiza patrones complejos en los comportamientos y preferencias de los usuarios para identificar segmentos de mercado de forma más precisa. Esta segmentación más efectiva facilita la agrupación de clientes en segmentos específicos, permitiendo a las empresas personalizar sus estrategias de marketing (Reddy et al., 2023). En este sentido, entendemos por segmentar "*diferenciar el mercado total de un producto o servicio en grupos diferentes de consumidores, homogéneos entre sí y diferentes a los demás, en cuanto a hábitos, necesidades y gustos, que podrían requerir productos o combinaciones de marketing diferentes*" (Monferrer, 2013).

- Personalización. El ML ajusta los sistemas de recomendación de productos, contenido y mensajes a medida, basados en el comportamiento individual del consumidor, para ofrecer a los clientes ofertas que se ajustan a sus intereses y comportamientos pasados.

- Análisis predictivo. Utiliza el procesamiento de conjuntos de datos históricos para prever tendencias y patrones futuros en el comportamiento del cliente. Ofrece a los especialistas en marketing *insights* profundos derivados del análisis del comportamiento pasado de sus audiencias, lo cual les permite optimizar estrategias y distribuir sus recursos con una mayor precisión y eficacia.

- *Sentiment Analysis.* Las herramientas impulsadas por la IA juegan un papel muy importante al procesar e interpretar las opiniones y sentimientos de los clientes a partir de textos en las redes sociales y diversas plataformas online para capturar el *feedback* del público hacia una marca o producto, permitiendo ajustes estratégicos en respuesta a este último.

- NLP. El *Natural Language Processing* o procesamiento del lenguaje natural es una tecnología de 'machine learning' que capacita a las computadoras

para entender, procesar e interpretar el lenguaje humano. De esta manera, a través de *chatbots,* programas informáticos que utilizan IA y NLP, facilitan interacciones más naturales y eficientes con los clientes. Asimismo, estos *chatbots* pueden aprender de las interacciones con los usuarios para ofrecer respuestas de forma automatizada más rápidas, precisas y personalizadas.

En el contexto del objeto estudio de la investigación, se ha de destacar el denominado *Data Driven maketing*, o marketing basado en datos, un enfoque fundamental que proporciona los datos que el *'machine learning'* y la IA necesitan para operar. Se establece como una estrategia que se enfoca en aprovechar la información de los clientes para desarrollar y ejecutar estrategias de marketing efectivas. Mediante la utilización de una amplia variedad de canales, tanto en línea (SEM[1], E-Mail) como fuera de línea (radio, prensa escrita), se recopilan datos específicos sobre los clientes para una mejor comprensión de los mismos, entre los que se incluyen sus preferencias y comportamientos de compra, así como sus hábitos de consumo (Nadler & McGuigan, 2018).

Una vez recopilados los datos, estos son analizados minuciosamente para obtener una comprensión más profunda de los clientes y sus necesidades. Esta comprensión permite al equipo de marketing identificar patrones de compra y tendencias, lo que a su vez les permite diseñar y aplicar estrategias de marketing altamente personalizadas y dirigidas específicamente a cada segmento de clientes.

Añadido a desempeñar un papel fundamental en el desarrollo de las estrategias de marketing, los datos del mismo modo evalúan la eficacia y rendimiento de dichas estrategias en el mercado. Las empresas pueden supervisar métricas como las conversiones de ventas o la participación en las redes sociales para medir el impacto de sus campañas de marketing. Además, estas métricas ayudan a identificar los canales y estrategias de marketing más productivos, facilitando la asignación de recursos para las campañas futuras (Rosário y Dias, 2023).

Aunque la incorporación de la inteligencia artificial en el ámbito del marketing conlleva una diversidad de ventajas, es posible reconocer distintos retos asociados a la integración de esta tecnología. Uno de estos desafíos lo encontramos en la obtención de

[1] *Search Engine Marketing*: conjunto de herramientas, técnicas y estrategias que ayudan a optimizar la visibilidad de sitios y páginas web a través de los motores de los buscadores.

datos de alta calidad y protección de la privacidad del consumidor. La inteligencia artificial requiere, para el máximo aprovechamiento de sus facultades, acceder a una gran cantidad de datos, entre los que se incluyen datos personales de los clientes. Consecuentemente, la obtención de estos datos genera una cierta intranquilidad en cuanto a la privacidad y seguridad de estos mismos. Es por eso por lo que las empresas han de realizar un compromiso con sus clientes en mostrarse transparentes en relación con la recopilación y uso de estos mismos datos para que no se quebrante la protección de los datos personales de los consumidores (Rivera-Montaño, 2023).

Asimismo, la implementación y mantenimiento de la inteligencia artificial en el marketing requiere que las empresas estén dotadas de recursos financieros suficientes para hacer frente al alto coste de las inversiones en tecnologías especializadas requeridas para la adopción y gestión de estas herramientas.

En suma, la integración de la inteligencia artificial en el marketing constituye un punto de inflexión estratégico, permitiendo a las empresas enfrentar con mayor eficacia los retos de un entorno digital en constante cambio. A través de la implementación de tecnologías de IA y aprendizaje automático, el marketing se transforma hacia una práctica más analítica, personalizada y eficiente, aunque esta evolución conlleva la necesidad de abordar de manera ética los diferentes desafíos que dicha implementación lleva aparejada.

2. APROXIMACIÓN AL MARKETING DIRIGIDO EN LA ERA DIGITAL

El marketing dirigido, también conocido como marketing personalizado o marketing uno a uno, se ha establecido como una estrategia fundamental en el panorama empresarial contemporáneo.

El planteamiento estratégico de este proceso de identificación de clientes y promoción de productos y servicios se sustenta en la singularidad de cada cliente, demandando un enfoque personalizado en su estrategia de marketing. Para alcanzar dicha personalización, la estrategia se enfoca en la segmentación del mercado en grupos de tamaño reducido que compartan características distintivas similares con el propósito de, posteriormente, ofrecer a esos segmentos mensajes y promociones que se alineen con las necesidades y expectativas particulares de cada grupo identificado.

Cuando nos referimos a personalización en el contexto del objeto estudio de la investigación, el marketing, es crucial precisar el concepto que subyace de esta noción. Se entiende como personalización la habilidad de poder ofrecer los productos y servicios adecuados en el momento y lugar apropiados a los clientes correctos (Sunikka & Bragge, 2012). Del mismo modo, desde la perspectiva del marketing uno a uno, hablamos de personalización cuando, tras un estudio singularizado del cliente diferenciado, proporcionamos al consumidor experiencias individualizadas.

No obstante, es esencial no equiparar la personalización con la posibilidad otorgada al cliente de agregar complementos adicionales a un producto a su gusto, ni tampoco con aquel proceso iniciado exclusivamente por parte de la empresa. La personalización implica una comprensión integral y holística de las necesidades, preferencias y comportamientos de los clientes, que va más allá de simples opciones de personalización superficiales.

A su vez, la transformación hacia la digitalización en las recientes décadas ha ejercido una influencia revolucionaria en la práctica del marketing, marcando la transición de estrategias de difusión masivas (*broadcasting*) hacia enfoques altamente dirigidos y personalizados (*narrowcasting*). Este cambio se debe en gran medida al surgimiento e integración de las nuevas plataformas digitales que permiten una comunicación más específica y relevante con segmentos de audiencia menores.

Mientras que medios tradicionales como la televisión y la prensa siguen siendo relevantes, su predominio disminuye frente a medios más especializados y dirigidos, utilizados por publicistas para conectar con nichos específicos mediante mensajes interactivos y personalizados. La diversificación de estos canales incluye desde videos en internet hasta blogs, email marketing y redes sociales, favoreciendo una estrategia de comunicación selectiva sobre la difusión masiva, adaptándose así a las demandas de una audiencia fragmentada.

3. ANÁLISIS PREDICTIVO

3.1. Conceptos clave

El análisis predictivo se posiciona como una disciplina clave dentro del ámbito de la analítica avanzada, cuyo propósito radica en anticipar y pronosticar resultados futuros a

partir de la integración y análisis de datos históricos. Esta práctica abarca la implementación de métodos de aprendizaje automático, análisis estadístico y técnicas de extracción de datos, con el fin de identificar patrones existentes y anticipar tendencias basadas en dichos patrones extraídos (Bender y Mazza, 2017).

Las organizaciones recurren al análisis predictivo como una herramienta fundamental para no solo comprender el pasado, sino también para proyectarse hacia el futuro, identificando tanto riesgos potenciales como oportunidades latentes que puedan influir en su desempeño y estrategias empresariales. Al examinar datos históricos y discernir patrones y tendencias, los sistemas de inteligencia artificial pueden predecir las ventas futuras y el comportamiento de los consumidores con extraordinaria precisión. Esta destreza predictiva permite a los especialistas en marketing prever las fluctuaciones del mercado, diseñar estrategias para la gestión de inventario y optimizar sus campañas publicitarias para armonizarlas con las preferencias proyectadas de los consumidores. (Gaikwad & Gautam, 2023).

En relación con los beneficios que esta herramienta proporciona, el análisis predictivo desempeña un papel crucial en la mejora de la comprensión de las situaciones y la toma de decisiones informadas al poder abordar problemas complejos mediante la revelación de patrones que se encontraban ocultos en los datos de una manera más rápida y precisa. Del mismo modo, la capacidad de respuesta al igual que la eficiencia de las operaciones, se ve significativamente mejorada debido a la suficiencia de los modelos entrenados para procesar datos en tiempo real y proporcionar respuestas de forma instantánea.

En este contexto, se erige como un pilar fundamental en el proceso de toma de decisiones, ofreciendo soluciones óptimas derivadas del análisis predictivo. Este análisis no solo proporciona el conocimiento de las tendencias de marketing y distingue mercados con potencial. Al facilitar el procesamiento de grandes volúmenes de datos que el campo del marketing requiere para poder operar, la IA aporta los algoritmos matemáticos y probabilísticos necesarios para poder realizar el análisis de los diferentes escenarios que se presentan, así como la creación de mercados potenciales optimizando de manera significativa tanto el tiempo como los recursos invertido (Murillo-Andrade y Vizuete-Muñoz, 2024).

Establecido como una capacidad revolucionaria de la inteligencia artificial, el análisis predictivo emplea datos históricos para anticipar comportamientos o eventos específicos. Anteriormente, solo era posible identificar tendencias generales basadas en conjuntos de datos pasados. Sin embargo, gracias al análisis predictivo, los profesionales del marketing pueden comprender las experiencias de los clientes y evaluar cómo perciben las estrategias de marketing, así como los resultados derivados de estas últimas.

Con el propósito de adquirir una ventaja competitiva, las empresas realizan una evaluación constante para discernir los segmentos de audiencia que tiene más probabilidades de convertirse en clientes. Los algoritmos de aprendizaje automático, al integrar herramientas precisas de inteligencia artificial, son capaces de analizar la información sin procesar para distinguir a los clientes potenciales con mayor probabilidad de compra.

Adicionalmente, la clasificación predictiva de clientes potenciales permite determinar a través de qué canales de comunicación las estrategias son más efectivas, lo que permite dirigir los esfuerzos hacia aquellos clientes potenciales de mayor calidad en lugar de distribuirlos uniformemente en todos los canales disponibles. La implementación de un sólido y eficiente sistema de captación de clientes basado en inteligencia artificial ayuda a identificar qué clientes deben ser priorizados y a focalizar la atención en los canales más prometedores. Cuando todas estas funcionalidades se integran y ejecutan adecuadamente, estas aplicaciones contribuyen significativamente dentro de los equipos de ventas y marketing a la hora de adaptación de las estrategias, así como en la aceleración de los ciclos de ventas dirigidos (Zulaikha et al., 2020).

En el campo del marketing, el “Marketing predictivo” surge como el resultado de la sinergia entre la metodología del análisis predictivo y el marketing. Este concepto se revela como el producto de la evolución del marketing relacional, una práctica que ha sido adoptada con fervor por numerosos especialistas en el ámbito del marketing directo a lo largo de las últimas décadas (Artun & Levin, 2015).

El impulso detrás del ascendente interés en el marketing predictivo radica en la creciente demanda por parte de los clientes de un trato más personalizado e integrador. Esta exigencia surge en un contexto donde la interacción entre los consumidores y las marcas se produce a través de una multiplicidad de canales. Además, el advenimiento de

nuevas tecnologías ha facilitado la captación y el análisis de datos de clientes, tanto nuevos como preexistentes, permitiendo la identificación de patrones y la utilización de estos datos con una facilidad sin precedentes en la confluencia de los ámbitos físico y digital (Artun & Levin, 2015).

En consecuencia, el marketing predictivo está provocando una transformación radical en las estrategias de marketing tanto para empresas como para consumidores, a lo largo de todo el proceso de interacción con el cliente. Este enfoque innovador está redefiniendo la manera en que se conciben los productos y se gestionan los canales de comunicación, colocando al cliente en el centro de la estrategia (Artun & Levin, 2015).

3.2. Aplicación en marketing: Técnicas de análisis predictivo

El marketing predictivo emplea una diversidad de técnicas avanzadas de análisis de datos para identificar con precisión aquellas estrategias y acciones de marketing que poseen la mayor probabilidad de éxito. Mediante la integración de datos comerciales, de marketing y de ventas, junto con algoritmos matemáticos sofisticados, las organizaciones son capaces de descubrir patrones subyacentes, lo que les facilita a las empresas la toma de decisiones más informadas y acertadas sobre qué acciones implementar en sus futuras estrategias de marketing. Al conjunto de estas técnicas y herramientas lo denominamos modelo predictivo. Entre las técnicas de análisis predictivo más utilizadas y destacadas en el ámbito del marketing distinguimos:

3.2.1. Técnicas de aprendizaje supervisado

- KNN (k-vecinos más cercanos). Es un método no paramétrico empleado tanto para la clasificación como para la regresión. Su funcionamiento implica definir una función de similitud que establezca una puntuación entre pares, junto con la variable de respuesta y el número de vecinos más cercanos (k). En términos de clasificación, el objeto se clasifica por el voto mayoritario de sus vecinos más cercanos, es decir, se asigna a la clase del vecino más cercano. Para la regresión, se calcula la media de los valores de los k vecinos más cercanos del objeto. Es una técnica popular para predecir valores numéricos basándose en la similitud medida por una función de distancia. Se utiliza en diferentes áreas como redes sociales,

evaluación de productos y personalidad de marcas, entre otros. (Duarte et al., 2022).

- Regresión. El modelo emplea las conexiones entre las variables para calcular un valor de predicción para un nuevo conjunto de datos de entrada, donde una variable principal (variable dependiente) está influenciada por una serie de otras variables (variables independientes) que son a su vez fundamentales para realizar una predicción precisa o un pronóstico adecuado (Mackay-Castro et al., 2023).

- Árboles de decisión. La salida está formada por un nodo de decisión con dos o más ramas y un nodo hoja. Esta técnica es adecuada para describir secuencias de decisiones interrelacionadas o predecir tendencias futuras de los datos, y puede clasificar entidades concretas en clases específicas basándose en sus características. Cabe destacar que es una de las técnicas supervisadas más utilizadas en los sistemas de recomendación (Duarte et al., 2022). En relación con esta técnica, encontramos los modelos random forest,en los cuales la predicción de resultados para una nueva observación se realiza a través de la síntesis de las contribuciones de una pluralidad de árboles de decisión únicos donde cada uno ha sido entrenado a partir de un conjunto de datos de entrenamiento con sutiles variaciones. Este ensamblaje de árboles unidos conforma el núcleo del modelo predictivo.

- Redes neuronales. Se emplean con el propósito de explicar la correlación existente entre los conjuntos de datos de entrada y salida, lo cual habilita la capacidad de efectuar predicciones y anticipaciones. En el campo del marketing, se emplea en la predicción del comportamiento del consumidor, la elaboración y comprensión de segmentos de compradores más refinados, la automatización del marketing, la generación de contenido, así como en la estimación de ventas y el análisis del comportamiento del cliente. En relación con el proceso de ventas, las compañías llevan a cabo labores de búsqueda y evaluación de clientes, clasificándolos en función de su predisposición a la compra (Prasad & Ghosal, 2022).

- SVM (*Support Vector Machines*). Es un algoritmo de clasificación que transforma los datos de entrada en un espacio de características de mayor dimensión y, a continuación, construye un modelo lineal que crea límites de clase no lineales en el espacio original. Los datos se clasifican mediante un tipo especial de modelo lineal, el hiperplano óptimo, que maximiza la distancia entre las observaciones que pertenecen a cada categoría (Duarte et al., 2022).

3.2.2. Técnicas de aprendizaje no supervisado

- K-means. Destaca como uno de los métodos de *clustering* más populares debido a su eficacia en la obtención de resultados. Para su funcionamiento, K-means necesita un conjunto inicial de centroides, que deben tener la misma dimensión que el vector de entrada y corresponder al número deseado de *clusters* (agrupaciones) a formar. Durante el proceso de aprendizaje, estos centroides se ajustan iterativamente para aproximarse mejor a la distribución real de los datos (Duarte et al., 2022). En el campo del marketing, esta técnica es ampliamente utilizada debido a su eficacia en la segmentación de clientes.

- PCA (*Principal Component Analysis*). Es un método que permite explicar la estructura de varianza-covarianza de un conjunto de p-variables a través de una serie de combinaciones lineales de estas variables -componentes- (Sylvester et al., 2017). En el campo del marketing, esta técnica se utiliza para transformar grandes conjuntos de datos en formatos más manejables sin perder información clave. Esta simplificación de los datos, habilita a los analistas para discernir las características más influyentes de los consumidores. Como ejemplo, puede ayudar a determinar qué atributos de los clientes ejercen el mayor influjo en su conducta de compra, facilitando así una segmentación de mercado efectiva y el desarrollo de campañas publicitarias más precisas y personalizadas.

En resumen, la integración de modelos de análisis predictivo y algoritmos en la toma de decisiones de marketing contribuye significativamente a la capacidad de realizar elecciones basadas en información detallada y precisa. Mediante la adopción de técnicas de modelado y algoritmos de vanguardia, las organizaciones pueden descubrir *insights*

críticos sobre el comportamiento del consumidor y las dinámicas del mercado, lo que permite optimizar estrategias y alcanzar resultados superiores.

3.3. Casos de éxito

3.3.1. Netflix

Netflix, mediante la utilización de análisis predictivos basados en IA, puede anticipar las preferencias de visualización de sus usuarios. Al entender los patrones de consumo, la plataforma tiene la capacidad de sugerir contenido con un alto grado de adaptación a cada usuario, lo cual no solo incrementa la retención de sus usuarios, sino que del mismo modo, ofrece información valiosa sobre las tendencias de consumo (García et al., 2024).

Según (Chesñevar y Estevez, 2018), más del 80% de los programas de TV que la gente mira en Netflix surgen del sistema de recomendación que posee la plataforma. Esto significa, implícitamente, que la mayoría de lo que un usuario promedio de Netflix elige mirar es el resultado de un algoritmo basado en IA.

Todd Yellin, exvicepresidente de innovación de Netflix, detalla que para cada perfil de usuario se analizan múltiples variables como programas que el usuario ha visto, aquellos vistos antes y después de cada uno, y la hora en que se visualizaron, entre otros. Esta información se enriquece con datos provenientes de colaboradores independientes que visualizan y etiquetan cada programa de Netflix, asignándoles etiquetas que pueden abarcar desde el género del contenido (romántico, comedia, etc.), hasta los actores del elenco de la película o la serie (Chesñevar y Estevez, 2018).

"*Tomamos todas estas anotaciones y los datos de comportamiento del usuario y usamos algoritmos de aprendizaje automatizado que infieren qué es lo más importante y cómo deberíamos sopesarlo*", explica Yellin. "*Lo que logramos crear así son 'comunidades de gustos' (taste communities) en todo el mundo. Se trata de descubrir quiénes son las personas que miran el mismo tipo de cosas que usted mira*" (Chesñevar y Estevez, 2018).

3.3.2. Amazon

Amazon emplea estrategias de marketing predictivo para brindar sugerencias altamente personalizadas a sus consumidores. Este enfoque se sustenta en el análisis meticuloso de diversos parámetros relacionados con el comportamiento del usuario en la plataforma. Entre estos parámetros se incluyen el historial de búsquedas realizadas, las adquisiciones previas, el tiempo invertido en páginas específicas, o los artículos que han sido agregados al carrito de compras pero no han sido finalmente comprados.

Mediante la interpretación inteligente de estos datos, Amazon logra anticipar las necesidades y preferencias de sus usuarios mejorando significativamente la experiencia de compra en línea, así como reforzando la relación entre la compañía y sus clientes, al proporcionar un servicio que se percibe como único y ajustado a las preferencias individuales de cada usuario.

En este sentido, encontramos "Amazon Personalize", un servicio de aprendizaje automático que emplea sus datos para elaborar recomendaciones de productos específicos para sus usuarios. Asimismo, tiene la capacidad de crear segmentos de usuarios basándose en la proximidad que estos demuestran hacia ciertos productos o en los metadatos relacionados con los artículos. Entre los distintas utilidades de este servicio, encontramos la creación de una campaña de marketing dirigida. Gracias a esta herramienta, se pueden identificar segmentos de usuarios que interactúen con determinados productos del catálogo. Posteriormente, es posible utilizar un servicio de AWS o un servicio externo para elaborar una campaña de marketing específica que promocione distintos productos a diversos segmentos de usuarios (Amazon Webservice, s.f.).

Por otro lado, en febrero de 2024, Amazon presentó en versión beta el *chatbot* "Rufus", cuya implementación ha sido gradualmente extendida a un número creciente de usuarios. "Rufus" se presenta como un asistente de compras especializado que está formado en el catálogo de productos de Amazon y en conocimientos extraídos de toda la web. Su función es ayudar a los clientes a resolver dudas sobre productos, necesidades de compra y comparaciones; ofrecer recomendaciones personalizadas; y mejorar la experiencia de descubrimiento de productos dentro de la misma plataforma de compra habitual de Amazon (Rajiv & Chilimbi, 2024).

CAPITULO IV. CASO PRÁCTICO

1. REGRESIÓN LINEAL

La regresión es un proceso estadístico predictivo en el que los modelos de aprendizaje automático se esfuerzan por establecer una conexión entre las variables dependientes e independientes. Estos modelos pueden emplearse para estimar la influencia que tienen los predictores sobre la variable objetivo. En relación con el objetivo del algoritmo de regresión lineal, este se concreta en devolver un número continuo como pueden ser los ingresos o las ventas, entre otros.

El modelo de regresión lineal que se llevará a cabo en las secciones posteriores intentará predecir la variable 'Spending Score (1-100)', que es un indicador numérico de cuánto gasta un cliente. El objetivo principal de esta regresión es entender cómo diversas características demográficas y profesionales de los clientes (como edad, ingreso anual, experiencia laboral, tamaño de la familia, género y profesión) pueden influir en su comportamiento de gasto.

En el contexto de este trabajo de investigación, esta información puede ser muy útil para la toma de decisiones en marketing, permitiendo a la empresa diseñar estrategias de mercado más efectivas y personalizadas, basadas en el perfil de los clientes y sus patrones de gasto. Además, el modelo puede ayudar a identificar segmentos específicos de clientes que podrían ser el objetivo de campañas de marketing o promociones especiales.

Tal y como se plantea el problema expuesto, la aplicación del modelo de regresión lineal se presenta como idóneo debido a que se trata de un proceso estadístico predictivo cuyo objetivo es estimar valores número continuos.

1.1. Análisis descriptivo de los datos

La base de datos utilizada "Customers" ha sido extraída de la plataforma 'Kaggle' y la misma contiene un análisis detallado de los clientes de una tienda imaginaria. Este conjunto de datos consta de 2000 registros y ocho columnas entre las que distinguimos:

Figura 1: Tabla descriptiva de las variables del conjunto de datos 'Customers'

NOMBRE VARIABLE	DESCRIPCIÓN VARIABLE	TIPO VARIABLE
CustomerID	Identificación del cliente	Variable numérica
Gender	Género del cliente	Variable categórica
Age	Edad del cliente	Variable numérica
Annual Income ($)	Ingresos anuales de un cliente	Variable numérica
Spending Score (1-100)	Puntuación asignada por la tienda, basada en el comportamiento del cliente y la naturaleza del gasto	Variable numérica
Profession	Profesión del cliente	Variable categórica
Work Experience	Experiencia laboral en años del cliente	Variable numérica
Family Size	Número de miembros de una familia del cliente	Variable numérica

1.2. Análisis prescriptivo de los datos

En primer lugar, se ha preparado el entorno y cargado los datos. Para ello, se ha importado la librería 'pandas', necesaria para la manipulación de datos. Del mismo modo, se he cargado el conjunto de datos con el que se trabajará desde un archivo CSV en un DataFrame 'csv_data'[2].

Posteriormente se ha procedido a la limpieza y análisis de datos. Se elimina la columna 'CustomerID' porque es un identificador único que no aporta información útil para el modelado[3]. A continuación, se explora la descripción general del DataFrame con el que vamos a trabajar[4]. Como resultado, Figura 2, nos encontramos que en la variable 'Profession' existen valores faltantes, obteniendo 1965 observaciones frente a las 2000 del resto de las variables. Para el tratamiento de estos datos, se decide rellenar los valores faltantes en variable 'Profession' con el valor más repetido en esa columna[5], el cual es 'Artist'. Eliminar registros puede suponer pérdida de información, pero a su vez al modelo

[2] Véase anexo 1
[3] Véase anexo 2
[4] Véase anexo 3
[5] Véase anexo 4

no pueden entrar valores nulos, por ello se decide rellenarlos. Dentro de los estadísticos de centralización, la moda es el más común para imputar valores nulos en variables categóricas, dado que la media o mediana no se pueden calcular.

Figura 2: Tabla descriptiva de los registros de las variables

#	Column	Non-Null Count	Dtype
0	CustomerID	2000 non-null	int64
1	Gender	2000 non-null	object
2	Age	2000 non-null	int64
3	Annual Income ($)	2000 non-null	int64
4	Spending Score (1-100)	2000 non-null	int64
5	Profession	1965 non-null	object
6	Work Experience	2000 non-null	int64
7	Family Size	2000 non-null	int64

Fuente: Elaboración propia realizada en Python

Seguidamente, tiene lugar la exploración de *outliers*. Para identificar *outliers*, se utiliza la técnica del rango intercuartílico (IQR)[6]. El IQR se calcula utilizando los percentiles 25 (Q1) y 75 (Q3) de los datos. Estos percentiles dividen el conjunto de datos en cuatro partes iguales, lo que significa que el 25% de los datos están por debajo de Q1, el 25% están por encima de Q3, y el 50% están entre Q1 y Q3. Dado que el IQR se basa en los percentiles 25 y 75 en lugar de la media y la desviación estándar, es menos sensible a valores extremos en los datos. Esto se debe a que se calculan a partir de los datos ordenados, no de su magnitud absoluta. Debido a su menor sensibilidad a los valores extremos, los *outliers* extremos tienen menos impacto en el cálculo del IQR en comparación con otras medidas de dispersión, lo que significa que el IQR proporciona una mejor estimación de la variabilidad de los datos cuando estos contienen valores atípicos.

Como resultado, Figura 3,esta ejecución nos devuelve 5 valores atípicos en la variable 'Work Experience', todos con el mismo valor de 17 años. Para visualizar mejor estos *outliers* y tomar una decisión sobre el tratamiento de los mismos, se grafican los valores atípicos a través de un boxplot, Figura 4[7]. Esta representación gráfica muestra que la

[6] Véase anexo 5
[7] Véase anexo 6

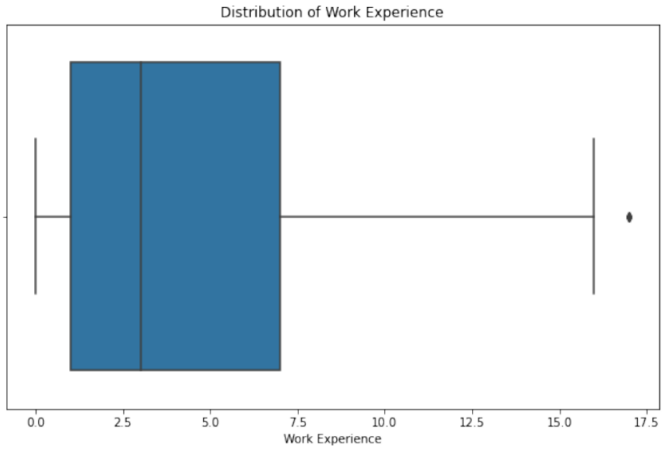
mayoría de los valores en la columna 'Work Experience' están bastante agrupados. Dado que estos valores son pocos y están muy separados del resto, podemos considerar la opción de "capar" estos valores, esto es, limitar los valores extremos a un valor máximo aceptable, al percentil más alto sin *outliers* para la construcción de nuestro modelo y así asegurar que los valores extremadamente altos no tengan un impacto desproporcionado en el modelo

Figura 3: Tabla descriptiva de los *outliers* de la variable 'Work Experience'

	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
392	Male	21	119116	30	Artist	17	4
405	Female	65	119889	11	Artist	17	6
473	Male	20	130813	92	Artist	17	5
566	Female	19	180331	14	Artist	17	5
603	Female	91	69720	78	Lawyer	17	6

Fuente: Elaboración propia realizada en Python

Figura 4: Boxplot de los *outliers* de la variable 'Work Experience'



Fuente: Elaboración propia realizada en Python

A continuación, se procede al procesamiento de los datos. En primer lugar se codifican las variables categóricas con 'OneHotEncoder'[8] para convertir los datos categóricos en un formato que el modelo de aprendizaje automático pueda usar. Esto genera lo que se

[8] Véase anexo 7

conoce como variables “dummy” o ficticias, que son variables que surgen para cada uno de los niveles del predictor categórico y que pueden tomar el valor de 0 o 1. Al convertirlos en dummies, el modelo no aporta mayor importancia a unas categorías sobre otras. Asimismo, se escalan las variables numéricas con ‘StandardScaler’[9]. La razón del escalamiento de las variables numéricas corresponde a que estas variables pueden tener diferentes rangos, por ejemplo la edad (‘Age’) puede variar de 0 a 100, mientras que el ingreso (‘Annual Income ($)’) puede variar de miles a cientos de miles. Si las variables no se escalan, las variables con mayor magnitud pueden dominar el proceso de entrenamiento del modelo, lo que puede llevar a un rendimiento subóptimo. Por otro lado, el escalado también puede ayudar a la interpretación de los coeficientes del modelo. Si las variables están en la misma escala, es más fácil comparar la magnitud de los coeficientes para ver cuál variable tiene más influencia en la variable objetivo.

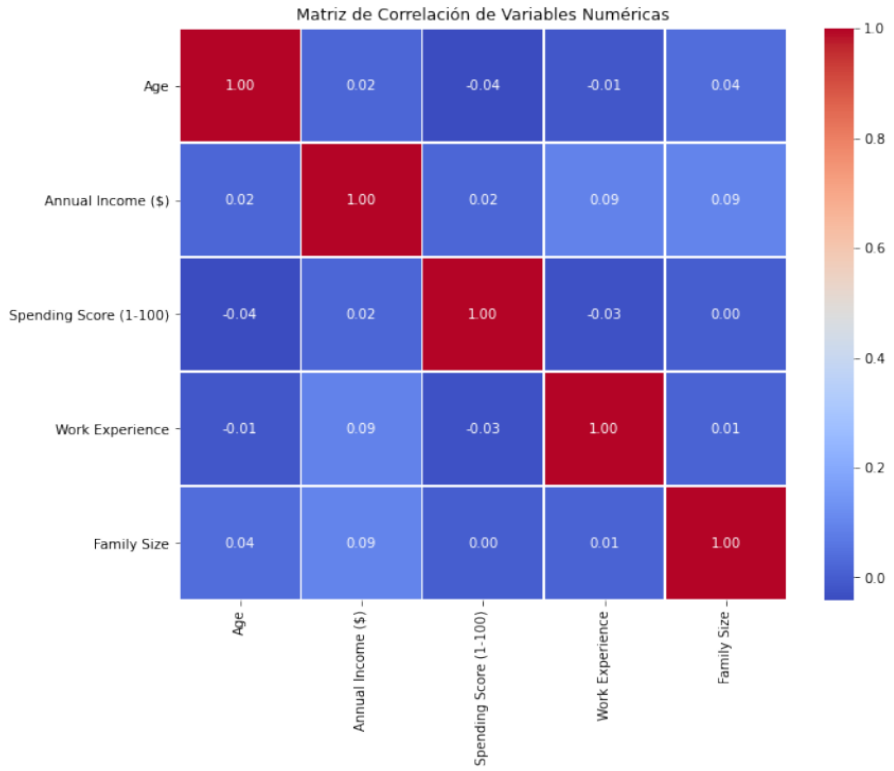
Posteriormente, se aplican diferentes transformaciones a las columnas de datos mediante la herramienta ‘ColumTransformer’[10] tanto a las variables numéricas, como a las variables categóricas ya que nos permite automatizar el proceso y tener los pasos definidos en un pipeline. Asimismo, se ajusta el ‘preprocessor’ previamente definido al conjunto de datos ‘csv_data’, modificándolo según las transformaciones especificadas. Tras el preprocesamiento de datos, el conjunto de datos ahora contiene 2000 filas y 14 columnas.

Una vez procesados los datos, se lleva a cabo una exploración de los mismos. Para visualizar y comprender las relaciones entre las diferentes variables y así visualizar cómo se distribuyen las mismas, se crea en primer lugar una matriz de correlación[11], Figura 5. Cabe mencionar que la matriz de correlación se ha realizado únicamente con las variables numéricas de nuestro DataFrame debido a que, aunque las variables categóricas fueron codificadas usando *one-hot encoding*, esta transformación convierte en columnas separadas las variable con valores de 0 a 1, haciendo que las correlaciones de estas variables puedan no ser muy informativas o incluso poco certeras debido a la naturaleza binaria de los datos. Del mismo modo, la codificación ‘one-hot’ introduce multicolinealidad (una alta correlación entre variables predictoras) debido a la columna

[9] Véase anexo 8
[10] Véase anexo 9
[11] Véase anexo 10

"dummy" que se crea, lo cual puede afectar al modelo de regresión. Por último, la interpretación de las correlaciones entre variables categóricas codificadas y otras variables numéricas puede ser complicada. Mientras que una correlación entre variables numéricas se interpreta como una relación lineal, una "correlación" entre una variable numérica y una categórica codificada no tiene una interpretación lineal clara.

Figura 5: Matriz de correlación de las variables numéricas de 'Customers'



Fuente: Elaboración propia realizada en Python

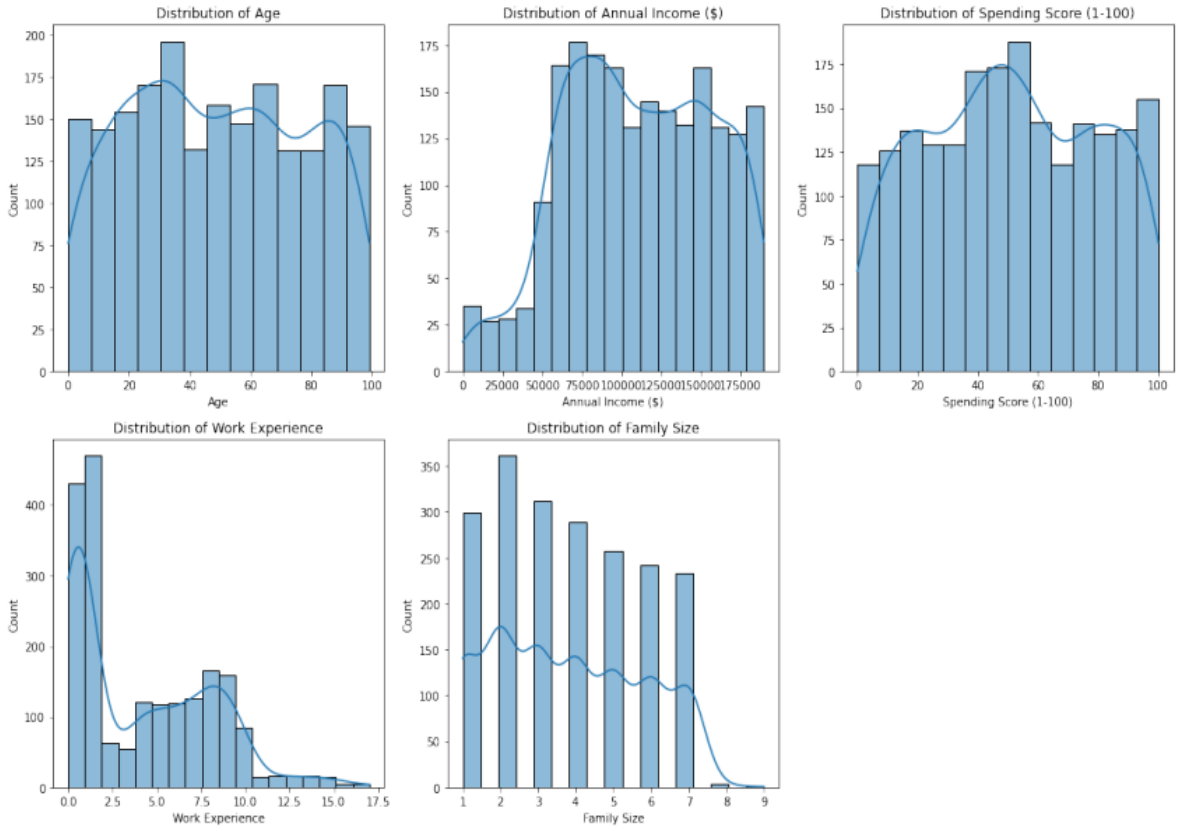
La correlación mide la relación lineal entre dos variables y varía de -1 a +1 donde +1 indica una correlación positiva perfecta (cuando una variable aumenta, la otra también lo hace); 0 indica que no hay correlación entre las variables; y -1 indica una correlación negativa perfecta (cuando una variable aumenta, la otra disminuye).

En la matriz de correlación efectuada sobre las variables numéricas de la investigación observamos que hay una falta de correlaciones sobre las mismas, lo cual sugiere que es poco probable que un modelo de regresión lineal simple, utilizando solo estas variables numéricas, sea eficaz para predecir la variable 'Spending Score (0-100)'. La correlación entre 'Age' y 'Spending Score (0-100)' es ligeramente negativa (-0.04) pero muy débil

como para considerarla significativa. Por otro lado, 'Annual Income ($)' tiene una correlación muy baja (0.02) con ‘Spending Score (0-100)’, lo que indica que no hay una relación lineal clara entre el ingreso anual y la puntuación de gastos. 'Work Experience' asimismo tiene una correlación casi nula con ‘Spending Score (0-100)’ (-0.03), lo que implica de igual forma que la experiencia laboral no tiene prácticamente ninguna relación lineal con la puntuación de gastos. Por último, no hay una relación aparente entre 'Family Size' y ‘Spending Score (0-100)’ en base a su correlación (0.00), lo que significa que el tamaño de la familia no influye en la puntuación de gastos de manera lineal.

Como a través de la matriz de correlación no hemos podido observar ninguna relación relevante entre las variables, vamos a explorar a través de diferentes gráficos la distribución y relación de cada variable numérica con el ‘Spending Score (0-100)’. En primer lugar, generamos una serie de histogramas para visualizar las distribuciones de cada una de las variables[12], Figura 6.

Figura 6: Histograma de distribución de las variables numéricas de ‘Customers’



Fuente: Elaboración propia realizada en Python.

[12] Véase anexo 11

En relación con las distribuciones visualizadas distinguimos:

- El histograma de la edad (‘Age’) muestra una distribución uniforme, esto es, el conjunto de datos abarca una amplia gama de grupos de edad, con una mayor concentración de individuos en el rango de 20-30 años.

- En relación con los ingresos anuales (‘Annual Income($)’), estos presentan una distribución con varias modas, observando una leve concentración en el rango de valores de 0-50.000$ anuales, experimentado posteriormente una crecida en el salario que se mantiene uniformemente en el resto de la gráfica.

- Por su parte, la puntuación de gasto (‘Spending Score (1-100)’) parece tener una distribución aproximadamente uniforme con ligeras concentraciones en los extremos bajos y altos.

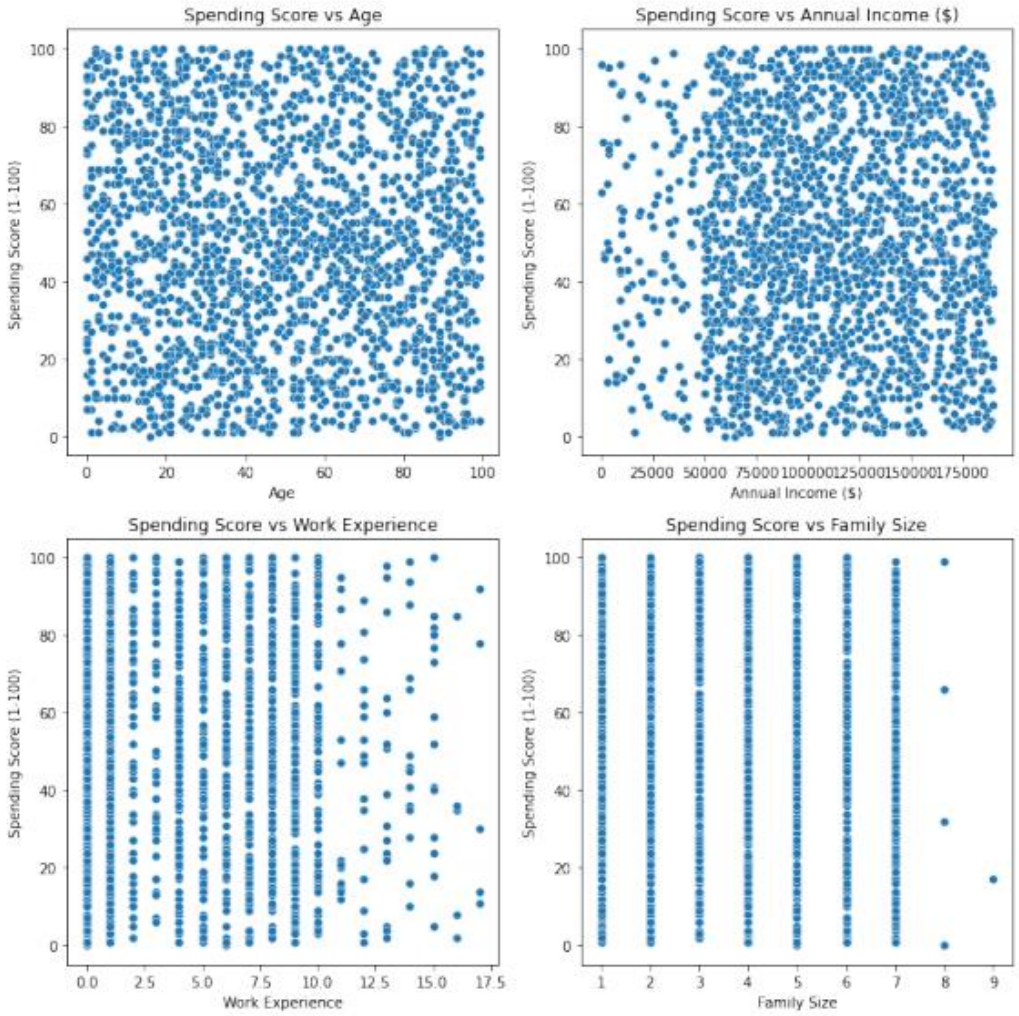
- La experiencia laboral (‘Work Experience’) muestra una distribución concentrada a la izquierda del histograma, con la mayoría de los individuos teniendo poca experiencia laboral y unos pocos con mucha experiencia. La concentración de datos cerca del valor cero sugiere que muchos individuos en la población pueden estar en las etapas iniciales de su carrera profesional.

- Por último, el tamaño de la familia (‘Family Size’) tiende a ser pequeño, con la mayoría de las familias compuestas por 1 a 4 miembros. Las familias más grandes son menos comunes.

Una vez hemos analizado las distribuciones de las variables, mediante la generación de gráficos de dispersión, vamos a estudiar cómo se relaciona la variable ‘Spending Score (0-100)’ con el resto de las variables[13], Figura 7.

[13] Véase anexo 12

Figura 7: Pairplot de correlación entre ‘Spending Score (1-100)’ y el resto de las variables numéricas



Fuente: Elaboración propia realizada en Python

En cuanto a la relación entre ‘Age’ y ‘Spending Score (0-100)’ no se aprecia una relación clara o lineal entre la edad y la puntuación de gasto, ya que los puntos están ampliamente dispersos y no muestran una tendencia específica, sugiriendo que la edad por sí sola puede no ser un buen predictor de la puntuación de gasto. Respecto de las variables ‘Annual Income ($)’ y ‘Spending Score (0-100)’, aunque también dispersos, hay una ligera concentración de puntos más altos de puntuación de gasto en torno a los ingresos medios. Sin embargo, no hay una tendencia lineal clara. En relación con la variable ‘Work Experience’ y ‘Spending Score (0-100)’, los datos muestran una serie de líneas horizontales, lo cual es característico de una variable numérica con valores

discretos y específicos. En este caso, podemos observar que hay pocas personas con una experiencia laboral de más de 10 años pero su puntuación de gasto está dispersa en todo el gráfico, no siendo suficientemente fuerte para indicar una relación lineal consistente entre estas dos variables. Por último, similar al caso de la experiencia laboral, entre las variables 'Family Size' y 'Spending Score (0-100)', vemos una dispersión de puntos a lo largo de valores específicos del tamaño de la familia pero no hay un patrón lineal obvio.

Una vez procesados y explorados los datos, se preparan los datos para realizar el modelo de regresión lineal. En primer lugar se importa la librería 'statsmodels.api', la cual ofrece una amplia gama de funcionalidades para realizar regresiones lineales, y se convierten los datos preprocesados y transformados de vuelta en un formato de DataFrame para facilitar el manejo y análisis posterior en el modelo de regresión[14]. Esta conversión se realiza debido a que los datos se ajustaron a un formato array en una de las etapas previas del preprocesamiento de datos, véase anexo 9. Este formato de datos no es adecuado en un modelo de regresión lineal ya que estos modelos esperan que los datos se presenten en forma de matriz bidimensional, donde cada fila representa una observación y cada columna representa una característica. Si los datos se presentan en un formato unidimensional (como una lista o un arreglo unidimensional), el modelo de regresión lineal no podrá interpretar correctamente las características. Del mismo modo, los modelos de regresión lineal generalmente requieren que los datos sean numéricos. Si los datos están en un formato array pero contienen tipos de datos no numéricos (como cadenas de texto o valores categóricos), es posible que el modelo de regresión lineal no pueda manejarlos directamente sin una adecuada codificación o transformación.

A continuación se preparan las variables para el modelo, separando las variables independientes que se utilizarán para entrenar al modelo, esto es todas las variables del conjunto de datos a excepción del 'Spending Score (0-100)' (X), de la variable objetivo, variable dependiente (y), 'Spending Score (0-100)'[15]. Posteriormente, se divide el conjunto de datos en subconjuntos de entrenamiento y prueba, para evaluar la capacidad del modelo para generalizar a nuevos datos[16].

[14] Véase anexo 13
[15] Véase anexo 14
[16] Véase anexo 15

Se lleva a cabo la creación de la instancia del modelo de regresión lineal, a la vez que se entrena el modelo con los datos de entrenamiento[17]. Del mismo modo, se generan las líneas de código que generan predicciones utilizando el modelo entrenado tanto del conjunto de datos de entrenamiento como de prueba[18].

Por otra parte, la regresión lineal es aplicada y se calcula el rendimiento del modelo mediante métricas de ajuste del modelo como el cálculo del Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2)[19]. El MSE proporciona una medida del promedio de los cuadrados de los errores, esencialmente ofreciendo una visión de la calidad del estimador. El R^2 ofrece una visión de cuánta variabilidad en la variable dependiente puede ser explicada por el modelo.

A continuación, se ejecuta la funcion 'sm.add_constant(X)'[20], que agrega una columna de unos al conjunto de datos de las variables independientes (X), proporcionando una nueva matriz de características 'X_const' la cual se puede utilizar como entrada en un modelo de regresión. De esta forma, se incluye el término de intercepción (término constante o bias) en el modelo de regresión lineal. Sin esta columna de unos, el modelo solo podría pasar por el origen (0,0) y no podría ajustar el término de intercepción. Asimismo, se ajusta el modelo de regresión lineal con la variable independiente (y) y la matriz de variables independientes 'X_const', que ahora incluye la columna constante[21]. El método .fit() realiza el ajuste del modelo, es decir, encuentra los coeficientes (betas) que minimizan la suma de cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo lineal, así como los p-valores, que prueban la hipótesis de que cada coeficiente es igual a cero.

Por último, se muestran los resultados para el MSE y el R^2 tanto para el entrenamiento como la prueba, así como los coeficientes (betas) y los p-valores del modelo[22], Figura 8.

[17] Véase anexo 16
[18] Véase anexo 17
[19] Véase anexo 18
[20] Véase anexo 19
[21] Véase anexo 20
[22] Véase anexo 21

Figura 8: Tabla descriptiva de los resultados del MSE, R^2 , betas y los p-valores

```
{'MSE Train': 0.9935920375954944,
 'MSE Test': 1.0038361377037253,
 'R² Train': 0.0073098227436461105,
 'R² Test': -0.015276217736021058,
 'Model Summary': <class 'statsmodels.iolib.summary.Summary'>
 """
                            OLS Regression Results
 ==============================================================================
 Dep. Variable:     Spending Score (1-100)   R-squared:                       0.007
 Model:                                OLS   Adj. R-squared:                  0.000
 Method:                     Least Squares   F-statistic:                     1.038
 Date:                    Sun, 21 Apr 2024   Prob (F-statistic):              0.411
 Time:                            17:59:07   Log-Likelihood:                -2831.1
 No. Observations:                    2000   AIC:                             5690.
 Df Residuals:                        1986   BIC:                             5769.
 Df Model:                              13
 Covariance Type:                nonrobust
 ===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
 -----------------------------------------------------------------------------------------------
 const                          0.0479      0.043      1.113      0.266      -0.036       0.132
 Age                           -0.0438      0.022     -1.951      0.051      -0.088       0.000
 Annual Income ($)              0.0272      0.023      1.206      0.228      -0.017       0.071
 Work Experience               -0.0299      0.023     -1.322      0.186      -0.074       0.014
 Family Size                    0.0021      0.023      0.094      0.925      -0.042       0.046
 Gender_Male                -7.109e-05      0.046     -0.002      0.999      -0.089       0.089
 Profession_Doctor             -0.0167      0.088     -0.189      0.850      -0.190       0.156
 Profession_Engineer           -0.1110      0.085     -1.312      0.190      -0.277       0.055
 Profession_Entertainment       0.0217      0.076      0.284      0.777      -0.128       0.172
 Profession_Executive          -0.0904      0.090     -1.004      0.316      -0.267       0.086
 Profession_Healthcare         -0.0675      0.067     -1.006      0.315      -0.199       0.064
 Profession_Homemaker          -0.2010      0.136     -1.483      0.138      -0.467       0.065
 Profession_Lawyer             -0.1294      0.093     -1.394      0.163      -0.311       0.053
 Profession_Marketing          -0.1298      0.115     -1.125      0.261      -0.356       0.097
 ==============================================================================
```

Fuente: Elaboración propia realizada en Python

Los p-valores son una medida que ayuda a determinar si los resultados de una hipótesis estadística son significativos. En la regresión, la hipótesis nula (H0) típicamente afirma que no hay relación entre la variable independiente y la variable dependiente, o que el coeficiente (beta) correspondiente a la variable independiente es igual a cero (no tiene efecto). Un p-valor bajo (comúnmente menos de 0.05) sugiere que hay suficiente evidencia en los datos para rechazar la hipótesis nula. En otras palabras, un p-valor bajo indica que es improbable obtener la estadística de prueba observada si la hipótesis nula fuera verdadera, y por lo tanto, se considera que la variable independiente tiene un efecto estadísticamente significativo sobre la variable dependiente.

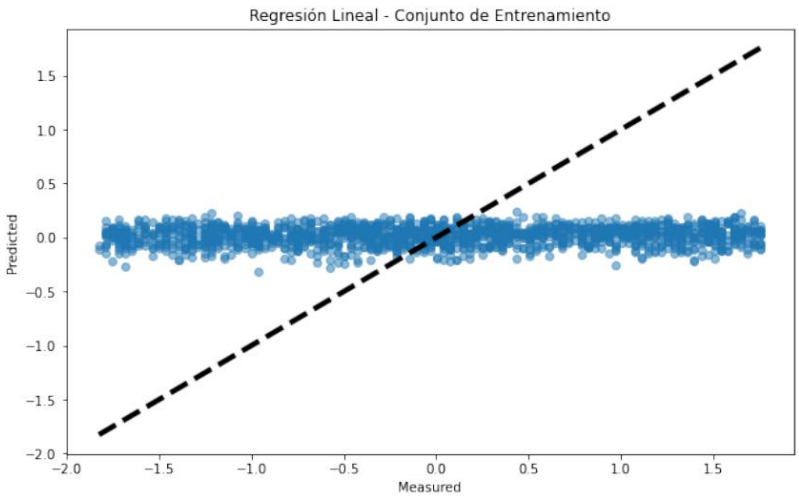
En relación con los resultados obtenidos en la Figura , observamos que la mayoría de las variables no resultan significativas (los p-valores (P>|t|) son más altos de 0.05). Asimismo, apreciamos que la variable edad es la que tiene más impacto sobre nuestra variable objetivo, con casi un 95% de confianza (p-valor de 0.051). Por esta razón,

podemos decir que, manteniendo el resto de las variables constante (esto es, con el mismo valor), si incrementamos en 1 año la edad de la persona, el 'Spending Score (0-100)' disminuye en 0.0438 puntos.

En cuanto al MSE, la similitud entre el MSE de entrenamiento (0.9935) y el de prueba (1.0038) sugiere que el modelo no sufre de sobreajuste, el cual ocurre cuando el modelo se ajusta demasiado bien a los datos con los que se entrena no solo aprendiendo de las relaciones subyacentes entre las variables en los datos de entrenamiento, sino también del ruido y las fluctuaciones aleatorias presentes en ese conjunto de datos específico, pero los valores altos pueden indicar que el modelo no predice muy bien. Por otro lado, el R^2 de entrenamiento (0.0073) significa que el modelo explica solo el 0.73% de la varianza en el conjunto de entrenamiento, lo cual es extremadamente bajo y sugiere que el modelo no tiene un buen ajuste. El R^2 de la prueba (-0.0152) es aún peor en términos de predicción, porque un R^2 negativo implica que el modelo es peor que simplemente predecir la media de la variable dependiente para todos los casos.

Por último, se define la función 'plot_regression_results', que grafica mediante gráficos de dispersión los valores predichos por el modelo contra los valores reales para los conjuntos de entrenamiento[23], Figura 9, y prueba[24], Figura 10.

Figura 9: Visualización modelo de regresión lineal con los datos de entrenamiento

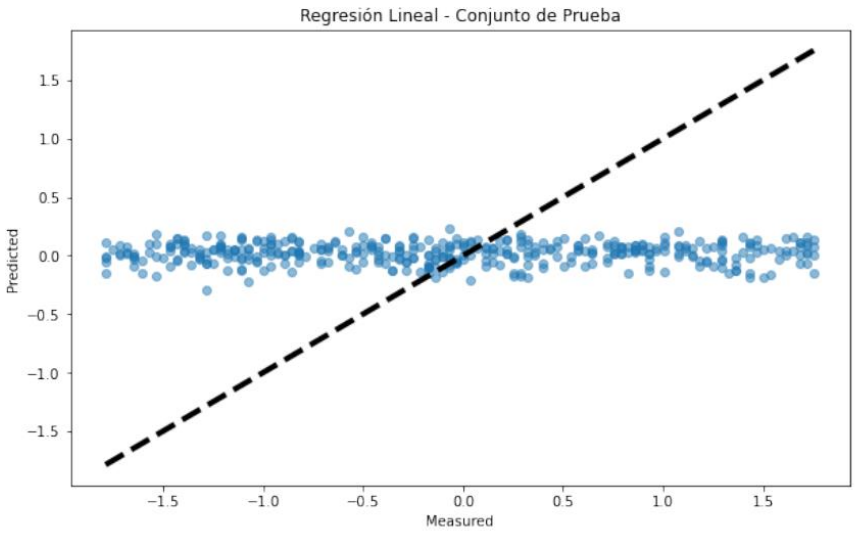


Fuente: Elaboración propia realizada en Python

[23] Véase anexo 22
[24] Véase anexo 23

En este gráfico, Figura, los puntos representan las predicciones del modelo versus los valores reales. La línea discontinua por su parte representa el lugar donde estarían los puntos si las predicciones fueran perfectas. Se observa que los puntos están distribuidos horizontalmente a lo largo de la línea, lo cual es indicativo de que no hay una gran variación en las predicciones conforme cambian los valores observados. No hay un patrón claro que muestre que el modelo está prediciendo mejor para ciertos rangos de valores. Asimismo, el hecho de que los puntos estén tan dispersos y no sigan la línea discontinua indica que el modelo no tiene una precisión alta, esto es, las predicciones no coinciden estrechamente con los valores reales.

Figura 10: Visualización modelo de regresión lineal con los datos de prueba



Fuente: Elaboración propia realizada en Python

En este otro gráfico, Figura, similar al gráfico de entrenamiento, muestra una comparación entre los valores reales y los predichos por el modelo, pero para datos que el modelo no ha visto durante el entrenamiento. Al igual que en el conjunto de entrenamiento, los puntos están bastante dispersos y no siguen la línea discontinua, lo que sugiere que el modelo no predice con precisión en el conjunto de prueba. Por último, la dispersión de los puntos es similar en ambos conjuntos, lo que indica que el modelo tiene una consistencia en cómo predice, pero el nivel de error parece ser igualmente alto en ambos.

1.3. Interpretación y presentación de los datos

La exploración inicial de los datos a través de una matriz de correlación y de histogramas indicó una distribución variada de las variables, pero con una falta de correlaciones significativas con el 'Spending Score (0-100)'. Del mismo modo, los gráficos de dispersión no muestran relaciones lineales fuertes entre el 'Spending Score (0-100)' y las variables independientes. Estos resultados son consistentes con los bajos valores de R^2 resultantes del modelo de regresión lineal, lo que sugiere que estas variables, al menos en su forma actual y sin considerar interacciones o efectos no lineales, tienen una capacidad limitada para explicar las variaciones en la puntuación de gasto.

La consistencia en el patrón de dispersión entre los conjuntos de entrenamiento y prueba sugiere que no hay sobreajuste, lo cual es positivo. Sin embargo, el modelo claramente no está capturando bien la variabilidad de los datos, como se refleja asimismo en el bajo R^2 calculado anteriormente en el conjunto de entrenamiento, y negativo en el conjunto de prueba, así como en MSE de entrenamiento y prueba resultante cercano a 1. Esto implica que el modelo es incapaz de explicar una cantidad significativa de la varianza en la variable objetivo, demostrando que el modelo no predice el 'Spending Score (0-100)' con precisión.

Los p-valores resultantes sugieren que la mayoría de las variables independientes no son estadísticamente significativas en la predicción del 'Spending Score (0-100)'. Sin embargo, la edad mostró un p-valor al borde de la significancia, indicando una posible, aunque pequeña, influencia negativa en el 'Spending Score (0-100)'.

Por último, no hay un patrón claro de errores sistemáticos; es decir, no parece que el modelo esté prediciendo consistentemente valores demasiado altos o bajos en relación con el valor real, lo que sería indicado por una agrupación de puntos sistemáticamente por encima o por debajo de la línea discontinua.

Si bien el modelo no ha logrado capturar la relación entre las variables independientes y el 'Spending Score (0-100)', si el modelo hubiera logrado un alto grado de precisión en la predicción del 'Spending Score (0-100)', esta técnica de aprendizaje automático habría ayudado a optimizar las estrategias de marketing a través de las aplicaciones prácticas de este modelo como:

- Segmentación de Clientes. Los resultados podrían haber informado una segmentación de mercado más precisa, identificando grupos de clientes con patrones de gasto similares y adaptando las comunicaciones y ofertas a estos segmentos con publicidad dirigida.

- Personalización de Ofertas. La predicción precisa del 'Spending Score (0-100)' habría permitido personalizar ofertas, recomendaciones y promociones según la probabilidad de gasto de cada cliente.

- Predicción de Comportamiento de Compra. Se podrían prever cambios en el comportamiento de compra basados en cambios anticipados en las variables clave, como 'Annual Income ($)' o 'Family Size', para ajustar las tácticas de marketing proactivamente.

- Optimización de Presupuesto de Marketing. Asignando recursos de manera más efectiva hacia aquellos segmentos de clientes más propensos a realizar compras, basado en un 'Spending Score (0-100)' predicho.

- Desarrollo de Producto. El entendimiento profundo de los factores que conducen a un mayor 'Spending Score (0-100)' podría influir en la estrategia de desarrollo de productos, centrándose en características que alineen con las preferencias de los clientes de alto gasto.

2. RANDOM FOREST

Para intentar mejorar la precisión del modelo en la predicción del 'Spending Score (1-100)', se explora la aplicación de un modelo de random forest sobre el conjunto de datos. Como la baja efectividad del modelo lineal puede deberse a características no lineales en los datos que no se ajustan bien a los supuestos de la regresión lineal, se opta por el uso de esta técnica que es capaz de modelar interacciones más complejas entre las variables. Asimismo, este modelo es menos susceptible al sobreajuste y puede manejar automáticamente las variables no numéricas sin necesidad de codificación previa.

Como se ha mencionado en apartados anteriores, el algoritmo random forest está basado en un conjunto de árboles de decisión. El random forest combina múltiples árboles

de decisión para conseguir una mayor capacidad predictiva, siendo uno de los algoritmos más utilizados en problemas de clasificación, pudiendo también aplicarse en regresiones o series de tiempo. Su gran ventaja respecto a la regresión lineal es que nos permite capturar relaciones no lineales a la vez que nos aporta facilidad en la interpretación de los resultados.

Siguiendo con la misma base de datos 'Customers', se ha llevado a cabo la realización de un segundo caso práctico en el lenguaje de programación Python a través de la técnica de random forest. En este sentido, vamos a intentar clasificar el 'Spending Score (1-100)', nuestra variable 'target', en base al resto de las variables. Para el análisis descriptivo de los datos, se remite al apartado 1.1. del capítulo IV.

2.1. Análisis prescriptivo de los datos

El análisis comienza con la importación de bibliotecas clave para el manejo de datos, la visualización, el modelado estadístico y la preparación de datos[25]. A continuación, se cargan los datos desde un archivo de tipo CSV[26], el cual contiene diferentes variables de los consumidores, expuestas en la Figura 1.

En segundo lugar, se ha procedido a la limpieza y análisis de datos. Por un lado, se elimina la columna 'CustomerID' ya que no aporta información predictiva relevante para el modelo[27]. A continuación, se verifica si hay valores faltantes en el conjunto de datos[28]. Como resultado, Figura 11, nos encontramos que en la variable 'Profession' existen 35 valores faltantes. Para el tratamiento de estos datos, se decide rellenar los valores faltantes en variable 'Profession' con el valor más repetido en esa columna[29], el cual es 'Artist'. Este método de tratamiento se empleó siguiendo el mismo criterio utilizado en secciones anteriores en el modelo de regresión lineal.

[25] Véase anexo 24
[26] Véase anexo 25
[27] Véase anexo 26
[28] Véase anexo 27
[29] Véase anexo 28

Figura 11: Tabla descriptiva de los registros faltantes de las variables de 'Customers'

```
Gender                     0
Age                        0
Annual Income ($)          0
Spending Score (1-100)     0
Profession                35
Work Experience            0
Family Size                0
dtype: int64
```

Fuente: Elaboración propia realizada en Python

Una vez se han tratado los valores faltantes, se identifica si existen valores atípicos o *outliers* en las variables. Para identificar *outliers*, se utiliza la técnica del rango intercuartílico (IQR), explicada en secciones anteriores[30]. La ejecución de la sentencia nos muestra que existen valores atípicos en la columna 'Work Experience' con un valor de 17 años en 5 de los registros del conjunto de datos, Figura 3, valor inusualmente alto en comparación con los demás datos. Antes de decidir cómo manejar estos *outliers*, visualizamos los valores atípicos con un boxplot[31] y evaluamos si estos casos deben ser considerados errores o datos legítimos.

La visualización del boxplot, Figura 4, muestra que la mayoría de los valores en la columna 'Work Experience' están bastante agrupados. Dado que estos valores son pocos y están muy separados del resto, podemos considerar la opción de "capar" estos valores, como realizado en el apartado de la regresión lineal.

En tercer lugar, se ha procedido al procesamiento de los datos. Para entender el tipo de procesamiento que podría ser necesario, se comprueban los tipos de datos de cada columna[32]. Como mostrado en la Figura 12, encontramos que dos de las variables, 'Gender' y 'Profession', son variables categóricas.

[30] Véase anexo 29
[31] Véase anexo 30
[32] Véase anexo 31

Figura 12: Tabla descriptiva del tipo de datos de las variables de 'Customers'

```
CustomerID                int64
Gender                   object
Age                       int64
Annual Income ($)         int64
Spending Score (1-100)    int64
Profession               object
Work Experience           int64
Family Size               int64
dtype: object
```

Fuente: Elaboración propia realizada en Python

Procedemos por tanto a codificar las variables categóricas[33] y a preparar el resto del conjunto de datos para el modelado. Se utiliza LabelEncoder para transformar las variables categóricas del conjunto de datos en formatos numéricos que puedan ser procesados por modelos de 'machine learning'.

En relación con las variables numéricas, no se realiza el escalado de las mismas ya que debido a la naturaleza de random forest, el modelo realiza divisiones en los datos utilizando umbrales en las características, lo que significa que la escala de las variables no afecta al proceso de entramiento de la misma manera que puede afectar a algoritmos basados en distancias o gradientes

A continuación, se categoriza la variable 'Spending Score (0-100)' en diferentes grupos dividiendo las puntuaciones en 3 rangos: 'Low', 'Medium' y 'High'[34]. Los límites de estas categorías están definidos con los valores [0, 33, 66, 100]. Esto significa que:

- Los puntajes de 0 a 33 se categorizarán como 'Low' (Bajo).
- Los puntajes de 34 a 66 se categorizarán como 'Medium' (Medio).
- Los puntajes de 67 a 100 se categorizarán como 'High' (Alto).

Del mismo modo, se transforman las categorías 'Low', 'Medium' y 'High' en valores numéricos utilizando la función 'LabelEncoder' de scikit-learn[35] para que puedan ser

[33] Véase anexo 32
[34] Véase anexo 33
[35] Véase anexo 34

procesados por algoritmos de aprendizaje automático. En este caso, se codifican las categorías como 0, 1 y 2, Figura 13 , respectivamente.

Figura 13: Tabla descriptiva de las variables de 'Customers' posterior categorización de la variable 'Spending Score (1-100)'

	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size	Spending Score Category
0	1	19	15000	39	5	1	4	2
1	1	21	35000	81	2	3	3	0
2	0	20	86000	6	2	1	1	1
3	0	23	59000	77	7	0	2	0
4	0	31	38000	40	3	2	6	2
...	...	...	...	...	...	...	...	...
1995	0	71	184387	40	0	8	7	2
1996	0	91	73158	32	1	7	7	1
1997	1	87	90961	14	5	9	2	1
1998	1	77	182109	4	4	7	2	1
1999	1	90	110610	52	3	5	2	2

Fuente: Elaboración propia realizada en Python

Una vez procesados los datos, se preparan los datos para el modelo de random forest. En primer lugar, se prepara el conjunto de características (*features*) 'X' y la variable objetivo ('target') 'Y', que en este caso es 'Spending Score (0-100)'[36]. Posteriormente, se divide el conjunto en datos de entrenamiento y prueba[37]. A través de la función 'train_test_split', se particionan los datos aleatoriamente, asignando el 30% de los datos al conjunto de prueba y el restante al conjunto de entramiento.

Seguidamente, se crea un clasificador de bosque aleatorio (RandomForestClassifier) con 100 árboles de decisión (n_estimators=100). Después, se entrena el clasificador con el método '.fit()' usando los conjuntos de entrenamiento 'X_train' y 'y_train'[38]. Finalmente, se utiliza el clasificador entrenado para hacer predicciones sobre el conjunto de prueba 'X_test'[39].

En relación con la evaluación del modelo de clasificación, se calculan dos métricas de evaluación. En primer lugar, mediante la función 'accuracy_score' se compara las

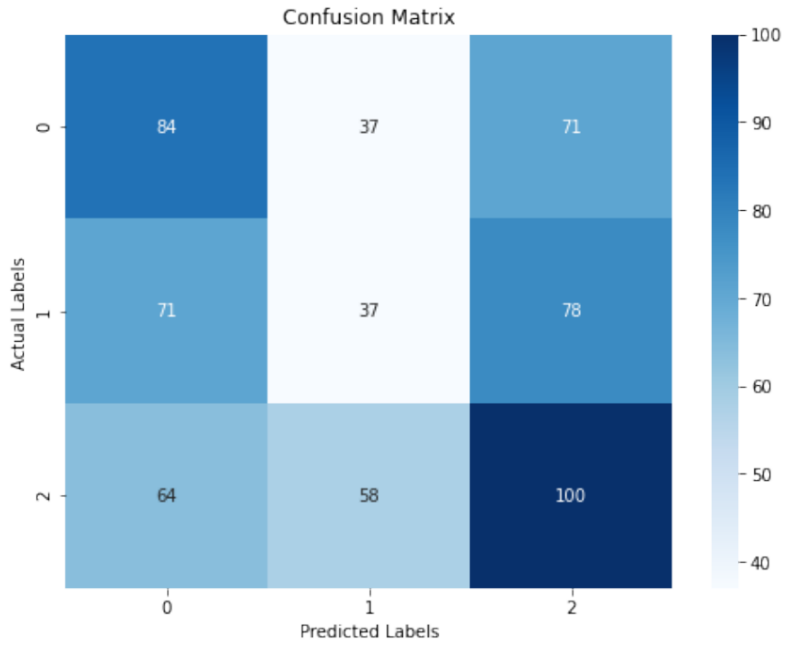
[36] Véase anexo 35
[37] Véase anexo 36
[38] Véase anexo 37
[39] Véase anexo 38

predicciones del modelo (y_pred) con los valores verdaderos (y_test) para calcular la precisión del modelo, que es la proporción de predicciones correctas en relación con todas las predicciones hechas[40]. En este caso, el resultado de la precisión es 36.83%.

Por otro lado, a través de la realización y visualización[41] de la matriz de confusión, Figura 14, se describe el rendimiento del modelo de clasificación. Esta matriz proporciona una visión de los tipos de errores que el modelo está cometiendo, comparando las etiquetas verdaderas con las predicciones del modelo. En una matriz de confusión típica encontramos:

- Verdaderos positivos (TP): El modelo predijo correctamente la clase positiva.
- Verdaderos negativos (TN): El modelo predijo correctamente la clase negativa.
- Falsos positivos (FP): El modelo predijo incorrectamente la clase positiva.
- Falsos negativos (FN): El modelo predijo incorrectamente la clase negativa.

Figura 14: Matriz de confusión 'Spending Score (1-100)'



Fuente: Elaboración propia realizada en Python

[40] Véase anexo 39
[41] Véase anexo 40

La matriz de confusión que se genera corresponde a un modelo de clasificación con tres categorías, que en nuestro caso representan los niveles de 'Spending Score (0-100)' categorizados como 'Low', 'Medium' y 'High', respectivamente. Las filas representan las clases reales de los datos de prueba, y las columnas representan las clases predichas por el modelo. En la matriz de confusión, los valores diagonales (84, 37, 100) representan las predicciones correctas (verdaderos positivos) para cada clase:

- 84 observaciones que eran realmente de la clase 'Low', el modelo también las predijo como 'Low'.
- 37 observaciones que eran realmente de la clase 'Medium', el modelo también las predijo como 'Medium'.
- 100 observaciones que eran realmente de la clase 'High', el modelo también las predijo como 'High'.

El resto de los valores en la matriz representan las clasificaciones incorrectas. Por ejemplo, 64 observaciones que eran realmente 'High', fueron predichas como 'Low', o 71 observaciones que eran realmente 'Low', fueron predichas como 'High'.

Una vez evaluado el modelo, procederemos a visualizar la importancia de las características (*features*) en el modelo de clasificación de random forest para entender qué tan influyentes son estas características en las predicciones del modelo. Para empezar, creamos el atributo del clasificador del random forest entrenado para que nos proporcione los valores que representan la importancia de cada característica en la decisión del modelo[42]. A continuación, creamos un DataFrame 'features_df' para visualizar los valores de estas importancias ordenándolo descendentemente para que las características más importantes aparezcan primero[43], Figura 15. Finalmente, creamos un gráfico de barras para visualizar las importancias, Figura 16[44].

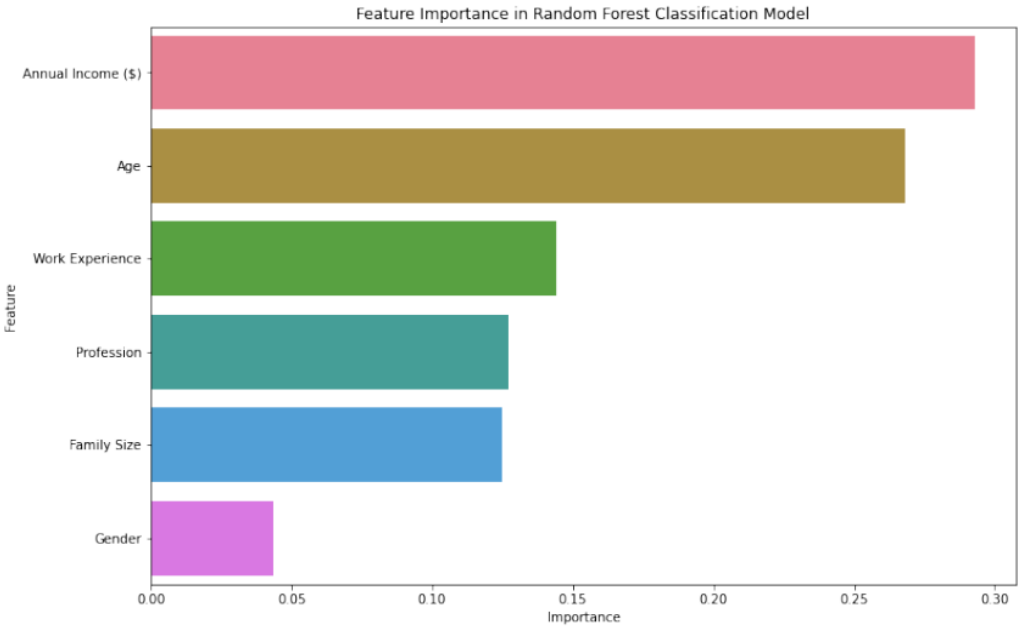
[42] Véase anexo 41
[43] Véase anexo 42
[44] Véase anexo 43

Figura 15: Tabla descriptiva de los valores de las importancias de las variables *features* de 'Customers'

	Feature	Importance
2	Annual Income ($)	0.292868
1	Age	0.268196
4	Work Experience	0.143905
3	Profession	0.126994
5	Family Size	0.124670
0	Gender	0.043367

Fuente: Elaboración propia realizada en Python

Figura 16 : Gráfico de barras de las importancias de las variables *features* de 'Customers'



Fuente: Elaboración propia realizada en Python

La visualización de la importancia de las características indica que el 'Annual Income ($)' y la variable 'Age' son los factores más significativos que el modelo de random forest está utilizando para realizar las clasificaciones de 'Spending Score (0-100)'. 'Work

Experience' y 'Profession' también contribuyen significativamente, pero en una magnitud menor, mientras que 'Family Size' y ‘Gender’ son las variables que tienen menor impacto.

Por último, se utiliza el modelo entrenado para estimar la categoría de la variable ‘Spending Score (0-100)’ para un cliente hipotético[45]. Es decir, el modelo intentará clasificar el ‘Spending Score (0-100)’ del cliente en las categorías que han sido previamente codificadas como 0 para 'Low', 1 para 'Medium', y 2 para 'High'. Este cliente hipotético no es parte del conjunto de datos original sobre el cual se entrenó el modelo, sino que es una nueva instancia para probar cómo el modelo se comporta con nuevos datos.

Al realizar pruebas con clientes hipotéticos podemos observar que el modelo ha aprendido y es capaz de predecir el ‘Spending Score (0-100)’ basándose en las características conocidas. No obstante, en este caso la precisión general del modelo, 36,83%, indica que el modelo no está realizando predicciones precisas en la mayoría de los casos. Esto es, en más del 60% de los casos el modelo no logra clasificar correctamente el ‘Spending Score (0-100)’ de los clientes según las categorías establecidas.

2.2. Interpretación y presentación de los resultados

En el modelo realizado, los resultados muestran un nivel de *accuracy* (precisión) del modelo random forest con un valor de 36.84%. Esta precisión podría ser un indicativo de varios problemas potenciales en el modelo, como características no informativas, falta de suficientes datos de entrenamiento, necesidad de ajustes en los parámetros del modelo, o incluso la posibilidad de que la relación entre las variables de entrada y el "Spending Score" no sea adecuadamente capturada por un modelo de clasificación de random forest.

El modelo random forest mostró una capacidad limitada para clasificar de manera precisa el ‘Spending Score (0-100)’ basado en las variables proporcionadas, con una precisión que sugiere una necesidad de mejorar la capacidad predictiva del modelo. No obstante, basado en la naturaleza de la variable ‘Spending Score (0-100)’, métrica cuantitativa que pretende reflejar la propensión o disposición de un cliente a gastar dinero

[45] Véase anexo 44

en los productos o servicios de una empresa, podemos utilizar la categorización de la variable 'Spending Score (0-100)' para identificar grupos específicos de consumidores:

- Bajo (0-33): Este grupo representa a los consumidores más cautelosos o con menor capacidad o disposición a gastar. Pueden ser clientes que solo compran los productos más esenciales o que buscan ofertas y descuentos antes de realizar una compra.
- Medio (34-66): Los consumidores en este grupo son moderadamente activos en sus gastos. Este grupo puede incluir a clientes regulares que realizan compras con cierta frecuencia y que son sensibles a la relación calidad-precio.
- Alto (67-100): Este segmento agrupa a los clientes más valiosos en términos de ingresos generados para la empresa. Son consumidores que probablemente adquieran nuevos lanzamientos, no se detengan ante precios altos y a menudo elijan productos premium.

La importancia de las características reveló que 'Annual Income ($)' y 'Age' son las variables más influyentes en la clasificación del 'Spending Score (0-100)', seguidas por 'Work Experience' y 'Profession', mientras que 'Family Size' y 'Gender' tuvieron menor impacto. Las variables más predictivas 'Annual Income ($)' y 'Age', pueden proporcionar información valiosa para la segmentación del mercado y el desarrollo de estrategias de marketing.

En relación con las estrategias de marketing sugeridas, dado que el 'Annual Income ($)' es un fuerte predictor del 'Spending Score (0-100)', se podrían desarrollar campañas de marketing dirigidas a clientes con ingresos más altos, ofreciendo productos premium o servicios exclusivos que puedan captar su interés. Asimismo, considerando que 'Age' también es una variable influyente, ajustar el contenido de marketing para adaptarse a diferentes grupos de edad podría mejorar la efectividad. Por ejemplo, se podrían personalizar campañas de tecnología y moda para los más jóvenes, y campañas de salud y bienestar para las personas mayores.

CAPÍTULO V. CONCLUSIONES

1. CONCLUSIONES Y LIMITACIONES DEL ESTUDIO

De acuerdo con el análisis teórico y práctico llevado a cabo durante la investigación, y en base al objetivo principal de la misma, se evidencia la integración de herramientas predictivas de inteligencia artificial en el campo del marketing. A través de la implementación de esta tecnología, se ha producido una transformación en lo que respecta a las estrategias empresariales. Mediante el análisis predictivo y el aprendizaje automático, las empresas son capaces de anticipar las necesidades y comportamientos de los consumidores, optimizando así sus estrategias de marketing dirigido y la toma de decisiones empresariales.

En concreto, el análisis predictivo se erige como una herramienta valiosa en términos de comprensión y anticipación del comportamiento del consumidor. Esta capacidad, permite a las empresas no solo mejorar la precisión de sus campañas publicitarias, sino tambien ajustar sus ofertas de servicio, resultando en una mejora significativa de la eficacia operativa y la satisfacción del cliente.

Por otra parte, se ha comprobado como el 'machine learning', especialmente a través de técnicas de aprendizaje supervisado y no supervisado, facilita una segmentación de mercado más precisa y, consecuentemente, una personalización profunda de las estrategias de marketing. Los dos casos de éxito presentados han demostrado la aplicabilidad práctica y la eficacia real de las teorías discutidas en los capítulos teóricos de la investigación sobre la convergencia de la inteligencia artificial, y en concreto el análisis predictivo, con el entorno empresarial. Los mismos ilustran cómo las soluciones personalizadas, impulsadas por la inteligencia artificial, pueden conducir a un mayor *engagement* y satisfacción del cliente, lo cual es crucial en mercados altamente competitivos.

Además, si bien no se han logrado conseguir resultados precisos a través de la realización de la aplicación práctica de los modelos de regresión lineal y random forest, el capítulo relativo al caso práctico ha demostrado la utilidad de las técnicas del análisis predictivo en la optimización de estrategias de marketing, facilitando una mayor personalización y efectividad en las campañas dirigidas. Los modelos aplicados

mostraron cómo, mediante el uso de datos y análisis avanzado, se pueden transformar grandes volúmenes de información en acciones concretas que potencialmente mejoran la relación entre la marca y el consumidor.

Por otro lado, aunque la implementación de IA en marketing conlleva desafíos éticos, como la gestión de la privacidad de datos, así como técnicos, como el alto costo de la tecnología, los beneficios superan significativamente a los perjuicios. De esta manera, la inteligencia artificial se revela como una herramienta indispensable para el futuro del marketing dirigido, marcando un avance estratégico, a la vez que proporcionando una ventaja competitiva en el mercado digital actual.

Por último, a pesar de los *insights* obtenidos, este estudio enfrenta varias limitaciones. Principalmente, los modelos predictivos utilizados están limitados por la calidad y la cantidad de datos disponibles. Cualquier limitación en estos aspectos puede afectar la precisión y la aplicabilidad de los resultados obtenidos, afectando y restringiendo la generalización de los resultados. Asimismo, el estudio se ha centrado en aplicaciones dentro de un contexto empresarial específico, lo que limita su aplicabilidad en otros sectores menos digitalizados o en mercados emergentes donde la adopción de tecnología puede ser desigual. Además, la implementación de la inteligencia artificial en marketing requiere una infraestructura tecnológica avanzada y habilidades especializadas. Estas exigencias pueden representar barreras significativas para aquellas organizaciones que cuentan con recursos limitados.

2. FUTURAS LÍNEAS DE INVESTIGACIÓN

La investigación actual ha sentado una base en la comprensión de cómo la inteligencia artificial puede ser aplicada efectivamente en el marketing dirigido a través del análisis predictivo. Sin embargo, la continua evolución de la tecnología y las cambiantes dinámicas del mercado generan diversas vías para investigaciones futuras.

Mientras que el presente estudio se centró en aplicaciones específicas en ciertos contextos empresariales, sería beneficioso expandir la investigación a una variedad de industrias, incluyendo aquellas menos digitalizadas o en mercados emergentes. Esto ayudaría a comprender la adaptabilidad y los desafíos de la inteligencia artificial en

diferentes entornos de mercado, y podría ofrecer *insights* sobre cómo diferentes sectores pueden aprovechar las herramientas predictivas de la IA para sus estrategias de marketing.

Por otro lado, sería conveniente examinar el uso de técnicas más avanzadas de aprendizaje automático en el marketing dirigido. Estos estudios podrían comparar la eficacia de diferentes algoritmos y modelos en términos de precisión de predicción, capacidad de adaptación a nuevos datos y eficiencia en diferentes escalas de datos.

Integrar información más detallada sobre el perfil del cliente, como datos demográficos ampliados, psicográficos, y el contexto socioeconómico, podría ayudar a personalizar aún más las estrategias de marketing. Estos datos permitirían una segmentación más precisa y campañas de marketing más personalizadas basadas en características individuales del consumidor.

Explorar cómo la IA puede interactuar y potenciarse con otras tecnologías emergentes, como la realidad aumentada (AR), la realidad virtual (VR) y el Internet de las Cosas (IoT), abriría nuevas dimensiones en el marketing dirigido. Estos estudios podrían centrarse en la creación de experiencias de cliente inmersivas y personalizadas a través de la integración de múltiples tecnologías.

Dado el creciente interés y la preocupación por la ética en el uso de IA, futuras investigaciones podrían centrarse en cómo las prácticas de marketing predictivo afectan la percepción de los consumidores sobre la privacidad y la ética. Esto podría incluir estudios que exploren la transparencia, el consentimiento del consumidor y la percepción de manipulación, así como el desarrollo de modelos de IA que prioricen la ética y la privacidad.

Por último, investigar el impacto económico a largo plazo de implementar la IA en el marketing dirigido podría ofrecer *insights* sobre la rentabilidad y sostenibilidad de estas inversiones tecnológicas. Estudios que cuantifiquen el retorno de inversión (ROI) y examinen la correlación entre la adopción de IA y el crecimiento económico de las empresas proporcionarían datos valiosos para justificar futuras inversiones en IA.

DECLARACIÓN DE USO DE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL GENERATIVA EN TRABAJOS FIN DE GRADO

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Alejandra García Boullosa, estudiante de E-3 Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis predictivo para el marketing dirigido: un enfoque basado en la inteligencia artificial", declaro que he utilizado la herramienta de Inteligencia artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 24 de abril de 2024

Firma: Alejandra García Boullosa

BIBLIOGRAFÍA

Abeliuk, A., y Gutiérrez, C. (2021). *Historia y evolución de la inteligencia artificial*. Revista Bits de Ciencia, Vol. 21, pp. 14-21. Recuperado de: https://revistasdex.uchile.cl/index.php/bits/article/view/2767

Amazon Web Services (s.f.). *What is Amazon Personalize? Amazon Personalize: Developer Guide,* p. 1. Recuperado de: https://docs.aws.amazon.com/pdfs/personalize/latest/dg/personalize-dg.pdf#what-is-personalize

Artun, O., & Levin, D. (2015). *Predictive marketing: Easy ways every marketer can use customer analytics and big data*. John Wiley & Sons, Inc, p. 3-4. ISBN 978-1-119-03736-1.

Bender, A., y Mazza, N. (2017). *Análisis Predictivo: Difusión e Impacto en Áreas de la Sociedad*. STS, Simposio Argentino sobre Tecnología y Sociedad, p. 2. DOI:oai:sedici.unlp.edu.ar:10915/64515

Chesñevar, C. I., y Estevez, E. C. (2018). *El comercio electrónico en la era de los bots,* p. 130-131. Recuperado de: https://ri.conicet.gov.ar/bitstream/handle/11336/94908/CONICET_Digital_Nro.3de882eb-3f30-4ac2-92db-e9e1a95d66e0_A.pdf?sequence=2&isAllowed=y

Duarte, V., Zuniga-Jara, S., & Contreras, S. (2022). *'Machine learning' and marketing: A systematic literature review*. IEEE Access, Vol. 10, pp. 93273-93288. Recuperado de: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9869838

Frankenfield, J. (2022). *Undesrtanding 'machine learning': Uses, Example*. Recuperado de: https://www.investopedia.com/terms/m/machine-learning.asp

Gaikwad, K & Gautam, R. (2023). *Artificial Intelligence and its Application in Today's Marketing Context.* International Journal for Multidisciplinary Research, Vol.5, Issue 5. DOI: 10.36948/ijfmr.2023.v05i05.6887

García, A. J. C., Vera, M. C. Q., Vargas, M. T. P., & Luzardo, J. S. Z. (2024). *La inteligencia artificial como herramienta en la segmentación de mercado*. Ciencia y Desarrollo, Vol. 27(1), p. 195. DOI: http://dx.doi.org/10.21503/cyd.v27i1.2556

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in Artificial Intelligence: A Global Study.* The University of Queensland and KPMG Australia, p. 8. DOI: 10.14264/00d3c94

IBM. (2023). *IBM Global AI Adoption Index – Enterprise Report*. pp. 36-46 Recuperado de: https://es.newsroom.ibm.com/announcements?item=122807

Mackay-Castro, R., Muñoz-Feraud, I., Medrano-Freire, E., Mackay-Véliz, R. (2023). *La inteligencia artificial como nueva alternativa para el marketing*. 593 Digital Publisher CEIT, Vol. 8(6), p. 666. DOI: doi.org/10.33386/593dp.2023.6.2099

McCarthy, J. (2007). *What is artificial intelligence?* Stanford University, Computer Science Department. Recuperado de: https://www-formal.stanford.edu/jmc/whatisai.pdf

Monferrer, D. (2013). *Fundamentos de marketing*. Publicacions de la Universitat Jaume I, p. 57. ISBN: 978-84-695-7093-7. Recuperado de: https://repositori.uji.es/xmlui/bitstream/handle/10234/49394/s74.pdf

Murillo-Andrade, A. D., y Vizuete-Muñoz, J. M. (2024). *El Impacto de la IA en el Marketing de Contenidos dentro del Contexto del Marketing 5.0*. Revista de investigación SIGMA, Vol. 11(01), p. 75. DOI: https://doi.org/10.24133/yz85g716

Nadler, A., & McGuigan, L. (2018). *An impulse to exploit: The behavioral turn in datadriven marketing*. Critical Studies in Media Communication, 35(2), pp. 151–165. Recuperado de: https://doi.org/10.1080/15295036.2017.1387279

Observatorio Nacional de Tecnología y Sociedad. 2023. *Uso de inteligencia artificial y big data en las empresas españolas.* 2023. Madrid: Ministerio de Asuntos Económicos y Transformación Digital, Secretaria General Técnica. Recuperado de: https://www.ontsi.es/sites/ontsi/files/2023-03/brujula-uso-IA-big-data-2023.pdf

Parlamento Europeo. (2021). *Proyecto de Resolución Legislativa del Parlamento Europeo sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia artificial) y se modifican determinados actos legislativos de la Unión (COM(2021)0206-C9-0146/2021-2021/0106(COD)).* Recuperado de: https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_ES.html

Prasad, B., & Ghosal, I. (2022). Forecasting buying intention through artificial neural network: an algorithmic solution on direct-to-consumer brands. *FIIB Business Review*, *11*(4), pp. 405-421. DOI: https://doi.org/10.1177/23197145211046126

Rajiv. M & Chilimbi, T. (2024). *«Amazon Rufus» AI experience comes to the Amazon Shopping app*. US About Amazon. Recuperado de: https://www.aboutamazon.com/news/retail/amazon-rufus

Randstad Research. (2024). *IA y Mercado de trabajo en España. Una aproximación cuantitativa y cualitativa a los efectos futuros de una tecnología revolucionaria*, p. 30. Recuperado de: https://eal.economistas.es/wp-content/uploads/sites/11/2024/02/RANDSTAD-RESEARCH-Informe-IA.pdf

Real Academia Española. (s.f.). *Diccionario de la lengua española*. Cultura. Recuperado de: https://dle.rae.es/inteligencia

Reddy, B. R., Rohith, S., Vamshi, V. S., & Dheeraj, O. (2023). *Segmenting Clients using 'machine learning' to get the Best Leads in an E-Commerce Business*. International Journal of Scientific Research in Science and Technology, pp. 184-191. DOI: https://doi.org/10.32628/ijsrst52310220

Rivera-Montaño, S. (2023). *Impacto de la inteligencia artificial (IA) en la efectividad de las estrategias de marketing personalizado*. Revista Científica Anfibios, Vol. 6(2), pp. 70-81. Doi: https://orcid.org/0009-0004-2975-5965

Rosário, A. T., y Dias, J. C. (2023). How has data-driven marketing evolved: Challenges and opportunities with emerging technologies. *International Journal of Information Management Data Insights*, *Vol. 3*(2), 100203. Recuperado de: https://doi.org/10.1016/j.jjimei.2023.100203

Rouhiainen, L. (2018). *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro*. Alienta Editorial. Recuperado de: https://planetadelibrosec0.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

Russell, S., & Norvig, P. (2004) *Inteligencia artificial un enfoque moderno*. [Traducido al español de Juan Manuel Corchado Rodríguez]. 2ª ed. Madrid, España: Pearson Prentice Hall. pp. 21-32. Recuperado de: https://luismejias21.files.wordpress.com/2017/09/inteligencia-artificial-un-enfoque-moderno-stuart-j-russell.pdf

Sunikka, A., & Bragge, J. (2012). *Applying text-mining to personalization and customization research literature–Who, what and where?* Expert Systems with Applications, Vol. 39(11), pp. 10049–10058. Recuperado de: https://doi.org/10.1016/j.eswa.2012.02.042

Sylvester Walusala, W., Rimiru, R., & Otieno, C. (2017). *A hybrid 'machine learning' approach for credit scoring using PCA and logistic regression*. International Journal of Computer (IJC), Vol. 27(1), pp. 88. Recuperado de: https://core.ac.uk/download/pdf/229656025.pdf

Wachsmuth, I. (2000). The concept of intelligence in AI. In Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Vol. 3, pp. 43-55. Dordrecht: Springer Netherlands. Recuperado de: https://link.springer.com/chapter/10.1007/978-94-010-0870-9_5

World Economic Forum. (2023). Future of Jobs Report. p. 24. Recuperado de: https://www.weforum.org/publications/the-future-of-jobs-report-2023/

Zulaikha, S., Mohamed, H., Kurniawati, M., Rusgianto, S., & Rusmita, S. A. (2020). Customer predictive analytics using artificial intelligence. *The Singapore Economic Review*, pp. 1-12. Recuperado de: https://doi.org/10.1142/S0217590820480021

ANEXOS

Base de datos: https://www.kaggle.com/datasets/datascientistanna/customers-dataset/data

- Anexo 1

```
import pandas as pd

csv_data=pd.read_csv(r'C:\Users\AlejandraGarcía\OneDrive\ICADE\5º\Conjuntodatos_Customers.csv', delimiter=';')
```

- Anexo 2

```
csv_data = csv_data.drop('CustomerID', axis=1)
```

- Anexo 3

```
csv_data.info()
```

- Anexo 4

```
most_common_profession = csv_data['Profession'].mode()[0]

csv_data['Profession'].fillna(most_common_profession, inplace=True)

csv_data['Profession'].isnull().sum(), most_common_profession
```

- Anexo 5

```
Q1 = csv_data.quantile(0.25)

Q3 = csv_data.quantile(0.75)

IQR = Q3 - Q1 outliers = ((csv_data < (Q1 - 1.5 * IQR)) | (csv_data > (Q3 + 1.5 * IQR))).sum() outliers

work_experience_outliers = csv_data[(csv_data['Work Experience'] < (Q1['Work Experience'] - 1.5 * IQR['Work Experience'])) |
```

(csv_data['Work Experience'] > (Q3['Work Experience'] + 1.5 * IQR['Work Experience']))]

work_experience_outliers

- Anexo 6

import seaborn as sns import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

sns.boxplot(csv_data['Work Experience'])

plt.title('Distribution of Work Experience')

plt.show()

- Anexo 7

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

categorical_features = ['Gender', 'Profession']
categorical_transformer = OneHotEncoder(drop='first')

- Anexo 8

numeric_features = ['Age', 'Annual Income ($)', 'Spending Score (1-100)', 'WorkExperience', 'Family Size'] numeric_transformer = StandardScaler()

- Anexo 9

preprocessor = ColumnTransformer(transformers=[('num', numeric_transformer, numeric_features), ('cat', categorical_transformer, categorical_features)])
data_preprocessed = preprocessor.fit_transform(csv_data)

- Anexo 10

```
import seaborn as sns
import matplotlib.pyplot as plt

correlation_matrix = csv_data[numeric_features].corr()

plt.figure(figsize=(10, 8)) sns.heatmap(correlation_matrix, annot=True,
cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Matriz de Correlación de Variables Numéricas')
plt.show()
```

- Anexo 11

```
plt.figure(figsize=(15, 10))
for i, column in enumerate(['Age', 'Annual Income ($)', 'Spending Score (1-100)', 'Work
Experience', 'Family Size'], 1):
        plt.subplot(2, 3, i)
        sns.histplot(csv_data[column], kde=True)
        plt.title(f'Distribution of {column}')
plt.tight_layout()
plt.show()
```

- Anexo 12

```
plt.figure(figsize=(15, 10)) for i, column in enumerate(numeric_features, 1):
        plt.subplot(2, 3, i)
        sns.scatterplot(x=csv_data[column], y=csv_data['Spending Score (1-100)'])
        plt.title(f'Spending Score vs {column}')
plt.tight_layout()
plt.show()
```

- Anexo 13

```
import statsmodels.api as s

columns_transformed = (numeric_features +
```

```
list(preprocessor.named_transformers_['cat']. get_feature_names(categorical_features)))
data_transformed = pd.DataFrame(data_preprocessed, columns=columns_transformed)
```

- Anexo 14

```
X = data_transformed.drop('Spending Score (1-100)', axis=1)
y = data_transformed['Spending Score (1-100)']
```

- Anexo 15

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.8, test_size = 0.2,
random_state = 100)
```

- Anexo 16

```
from sklearn.linear_model import LinearRegression
model = LinearRegression() model.fit(X_train, y_train)
```

- Anexo 17

```
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)
```

- Anexo 18

```
from sklearn.metrics import mean_squared_error, r2_score
train_mse = mean_squared_error(y_train, y_pred_train)
test_mse = mean_squared_error(y_test, y_pred_test)
train_r2 = r2_score(y_train, y_pred_train)
test_r2 = r2_score(y_test, y_pred_test)
```

- Anexo 19

```
X_const = sm.add_constant(X)
```

- Anexo 20

```
model = sm.OLS(y, X_const).fit()
```

- Anexo 21

```
results_ = {
        "MSE Train": train_mse,
        "MSE Test": test_mse,
        "R² Train": train_r2,
        "R² Test": test_r2,
        "Model Summary": model.summary()
}
results_
```

- Anexo 22

```
def plot_regression_results(X, y, y_pred, title):
        plt.figure(figsize=(10, 6))
        plt.scatter(y, y_pred, alpha=0.5)
        plt.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
        plt.xlabel('Measured')
        plt.ylabel('Predicted')
        plt.title(title)
        plt.show()

plot_regression_results(X_train, y_train, y_pred_train, "Regresión Lineal - Conjunto de Entrenamiento")
```

- Anexo 23

```
plot_regression_results(X_test, y_test, y_pred_test, "Regresión Lineal - Conjunto de Prueba")
```

- Anexo 24

```
import numpy as np

import pandas as pd
```

```
# Gráficos

import matplotlib.pyplot as plt

# Preprocesado y modelado

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn import metrics

from sklearn.metrics import mean_squared_error, r2_score, confusion_matrix,
accuracy_score

from math import sqrt
```

- Anexo 25

```
csv_data=pd.read_csv(r'C:\Users\AlejandraGarcía\OneDrive\ICADE\5º\Conjuntodatos_
Customers.csv', delimiter=';')
```

- Anexo 26

```
csv_data = csv_data.drop('CustomerID', axis=1)
```

- Anexo 27

```
missing_values = csv_data.isnull().sum()
missing_values
```

- Anexo 28

```
most_common_profession = csv_data['Profession'].mode()[0]

csv_data['Profession'].fillna(most_common_profession, inplace=True)

csv_data['Profession'].isnull().sum(), most_common_profession
```

- Anexo 29

```
def identify_outliers(data_column):

Q1 = np.percentile(data_column, 25)

Q3 = np.percentile(data_column, 75)

IQR = Q3 - Q1

outlier_step = 1.5 * IQR

outliers = data_column[(data_column < Q1 - outlier_step) | (data_column > Q3 + outlier_step)]

return outliers


numeric_columns = csv_data.select_dtypes(include=[np.number]).columns

outliers_visual = {column: identify_outliers(csv_data[column]) for column in numeric_columns}

outliers_visual
```

- Anexo 30

```
import seaborn as sns

plt.figure(figsize=(10, 6))

sns.boxplot(csv_data['Work Experience'])

plt.title('Distribution of Work Experience')

plt.show()
```

- Anexo 31

```
csv_data.dtypes
```

- Anexo 32

```
from sklearn.preprocessing import LabelEncoder

le_gender = LabelEncoder()

le_profession = LabelEncoder()

csv_data['Gender'] = le_gender.fit_transform(csv_data['Gender'])

csv_data['Profession'] = le_profession.fit_transform(csv_data['Profession'].astype(str))
```

- Anexo 33

```
bins = [0, 33, 66, 100]

labels = ['Low', 'Medium', 'High']

csv_data['Spending Score Category'] = pd.cut(csv_data['Spending Score (1-100)'],
bins=bins, labels=labels, include_lowest=True)
```

- Anexo 34

```
le_spending_score = LabelEncoder()

csv_data['Spending Score Category'] = le_spending_score.
fit_transform(csv_data['Spending Score Category'])
```

- Anexo 35

```
X = csv_data.drop(['Spending Score (1-100)', 'Spending Score Category'], axis=1)

y = csv_data['Spending Score Category']
```

- Anexo 36

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,random_state=42)
```

- Anexo 37

```
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)

rf_classifier.fit(X_train, y_train)
```

- Anexo 38

```
y_pred = rf_classifier.predict(X_test)
```

- Anexo 39

```
accuracy = accuracy_score(y_test, y_pred)*100

conf_matrix = confusion_matrix(y_test, y_pred)

accuracy

conf_matrix
```

- Anexo 40

```
import seaborn as sns

plt.figure(figsize=(8, 6))

sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=True)

plt.title('Confusion Matrix')

plt.xlabel('Predicted Labels')

plt.ylabel('Actual Labels')

plt.show()
```

- Anexo 41

```
feature_importances = rf_classifier.feature_importances_
```

- Anexo 42

```
features_df = pd.DataFrame({
'Feature': X.columns,
'Importance': feature_importances
}).sort_values(by='Importance', ascending=False)
features_df
```

- Anexo 43

```
plt.figure(figsize=(12, 8))
sns.barplot(x='Importance', y='Feature', data=features_df, palette='husl')
plt.title('Feature Importance in Random Forest Classification Model')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.show()
```

- Anexo 44

```
#caso hipotético 1
hypothetical_customer = pd.DataFrame({
'Gender': [le_gender.transform(['Male'])[0]], # Codificar género
'Age': [30],
'Annual Income ($)': [50000],
'Profession': [le_profession.transform(['Engineer'])[0]], # Codificar profesión
```

```
'Work Experience': [5],
'Family Size': [3]
})
predicted_spending_score = rf_classifier.predict(hypothetical_customer)
predicted_spending_score[0]
#output caso hipotético 1
0
hypothetical_customer = pd.DataFrame({
'Gender': [le_gender.transform(['Male'])[0]], # Codificar género
'Age': [20],
'Annual Income ($)': [30000],
'Profession': [le_profession.transform(['Artist'])[0]], # Codificar␣
↪profesión
'Work Experience': [2],
'Family Size': [1]
})
predicted_spending_score = rf_classifier.predict(hypothetical_customer)
predicted_spending_score[0]
#output caso hipotético 2
2
```