Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: 1

Predicción de potenciales compradores en un Ecommerce.

Clasificación y Clusterización de visitantes web.

Autor: Andrea Giralt Castellano

Tutor: Santiago Rojo Muñoz

Profesor: Albert Solé Ribalta

Barcelona, 15 de enero de 2023

Créditos/Copyright

Una página con la especificación de créditos/copyright para el proyecto (ya sea aplicación por un lado y documentación por el otro, o unificadamente), así como la del uso de marcas, productos o servicios de terceros (incluidos códigos fuente). Si una persona diferente al autor colaboró en el proyecto, tiene que quedar explicitada su identidad y qué hizo.

A continuación se ejemplifica el caso más habitual, aunque se puede modificar por cualquier otra alternativa:

FICHA DEL TRABAJO FINAL

Título del trabajo:	Predicción de potenciales compradores en un Ecommerce. Clasificación y Clusterización de clientes
Nombre del autor:	Andrea Giralt Castellano
Nombre del colaborador/a docente:	Santiago Rojo Muñoz
Nombre del PRA:	Albert Solé Ribalta
Fecha de entrega (mm/aaaa):	01/2023
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Data Analysis y Big Data
Idioma del trabajo:	Español
Palabras clave	Ecommerce, Clusterización y Clasificación

Dedicatoria/Cita

A mi familia, por estar siempre. A mis amigos, en especial a David, por esa ayuda todos estos años.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción y objetivos

1.1. Introducción

En la actualidad, el sector de comercio electrónico o **ecommerce** está en pleno auge, impulsado en parte por la pandemia, se ha convertido en un método de compra relevante para gran parte de los consumidores. Muchas empresas emergentes han decidido vender directamente de forma online, mientras que las existentes han tenido que reinventarse y cambiar su forma de venta sustituyéndola o combinándola con la venta física.

El **ecommerce** se podría definir como una tienda virtual. Es decir, un método de compraventa que utiliza internet como medio para realizar transacciones entre vendedor y comprador. No sólo mediante una página web, sino también a través de las redes sociales. Este modelo de negocio no necesita de grandes infraestructuras, siendo una de las principales causas que ha propiciado su expansión.

Las nuevas tecnologías y las comunicaciones abren a las empresas un amplio abanico de posibilidades. Ofrece la oportunidad de llegar a un mayor número de consumidores sin importar el lugar en el que se encuentren. Además, la reducción de costes en cuanto al alquiler de un local físico, permite bajar el precio del producto y ser más competitivo. Según un estudio del INE [1], en junio de 2020, las ventas de comercio minorista por este canal fueron un 71.2 % superiores al mismo mes del año pasado.

Además, una de las mayores ventajas de cara al comercio online es la posibilidad de medir como los consumidores interaccionan con tu página web o tus redes sociales, pudiendo segmentar a tus posibles clientes. Esto puede llegar a suponer mayor efectividad de las acciones de comunicación y marketing mediante el lanzamiento de campañas especializadas para cada segmento.

[1] https://www.ine.es/

1.1.1. Motivación

Por tanto, este trabajo se ha escogido ya que se pretende analizar un mercado que está en su pleno esplendor el cual impacta directamente a toda la población mundial. Entender su funcionamiento y el impacto que tiene en la sociedad es relevante para su expansión.

Pensamos que el análisis detallado de los datos de un **ecommerce**, permitirá aclarar los patrones que determinan que un visitante de una web se convierta en cliente, frente a los que no, así como mejorar las comunicaciones que se realizan desde la empresa.

1.2. Objetivos

En este trabajo se aborda la creación de dos modelos con datos procedentes de un **ecommerce** con sede en Colombia, el cual se dedica a la venta de productos relacionados con la información de empresas del país como: Informes Comerciales y módulos de información detallada sobre Datos Financieros, Prensa, Administradores, Incidencias, Informes Sectoriales, Base de datos a medida, Productos de Marketing (mercadeo), Información de accionistas, Información de proveedores y clientes, etc.

Para la creación de los modelos es necesario previamente construir un dataset con el que entrenarlos. Para ello, se realizan diferentes análisis como la creación de nuevas variables a partir de los campos existentes, o como la realización de estudio de Missings, así como de la relevancia de las diferentes variables.

Por consiguiente, se dispone de diferentes repositorios con datos históricos relativos a:

- Características básicas del registro/lead: fecha registro, procedencia, marca de comprador, cuando compró, dominio email, tipo de persona.
- Movimientos/navegación en la web: consumos de productos Promocionales.
- Número de sesiones por fecha de conexión.
- Información sobre los productos que ha comprado.

Para la obtención de los dos modelos, se sigue dos metodologías diferentes:

- Modelo de aprendizaje supervisado con el objetivo de predecir los compradores potenciales. Como se sabe, no todos los visitantes a las webs llegan a comprar, identificar posibles patrones es crucial para poder enfocar mejor el marketing de cada empresa. Gracias a estos patrones, se pueden identificar desde errores en el proceso de compra a posibles traducciones incorrectas si la web está en varios idiomas.

- Modelo de aprendizaje no supervisado con el objetivo de clasificar en clusters a los visitantes de la web de venta online y establecer su nivel potencial de compra. Como hemos comentado, hoy en día es vital dar un servicio personalizado en lo que a marketing se refiere. Por ello, conocer una clasificación tanto de los compradores como de los no compradores de tu producto es crucial para poder enfocar al cluster correcto tus campañas y no desperdiciar, así, recursos innecesarios.

Para ello, se presentan los distintos modelos de aprendizaje tanto supervisado como no supervisado, detallándolos con varios ejemplos de aplicación.

1.2.1. Productos finales

En resumen, el objetivo principal de este trabajo es, por un lado, detectar los potenciales compradores de los visitantes web y por otro, segmentarlos para conocer su comportamiento dado un dataset previamente analizado. Obteniendo como entregables finales:

- Un archivo .pdf con la memoria.
- Un notebook de python con el código realizado en Google Colab.

Estos dos productos, se podrán consultar en el siguiente enlace de GitHub: *https://github.com/AndreaGiralt/TFM_agiralt001*.

1.3. Metodología

De cara a la correcta organización que permita la ejecución completa de este trabajo se ha optado por una organización semanal basada en metodologías ágiles [2]. El objetivo de seguir esta metodología es trabajar de una forma estructurada de tal forma que se divida el trabajo en varias fases principales:

- Recopilación y análisis documental: Esta fase consiste en la búsqueda, recopilación y análisis de datos bibliográficos sobre los diferentes algoritmos que se pueden utilizar para lograr los objetivos definidos. Así como, casos de uso similares a nuestro trabajo.
- Análisis de los datos: En una segunda fase, se realiza un estudio y análisis de los datos. Para ello, se obtienen nuevas variables a partir de los campos que ya se tienen, se realiza un estudio de Missings, se realiza un estudio de relevancia de variables, se convierten las variables continuas en categóricas y, por último, se lidia con problemas de desbalanceo del dataset.

[2] https://agilethought.com/blogs/scaling-data-science-use-CRISP-dm-agile/

- Entrenamiento de los modelos: Una vez obtenido el dataset final, en esta frase se entrenan los modelos. Para ello, se utilizan los modelos definidos en la primera fase.
- Análisis de resultados: En una última fase, se elabora una reflexión sobre los resultados obtenidos en los modelos seleccionados. Además se realiza una propuesta de continuación del trabajo a futuro.

Por otro lado, como metodología de minería de datos se hará uso de CRISP-DM[3] para realizar todo el proceso de análisis de datos, así como el despliegue final de los modelos utilizados en este trabajo.

1.3.1. Planificación

Dada la metofología definida, definimos una planificación para la realización del trabajo:

- En este primero gráfico de Gantt podemos ver la planificación de las primeras cuatro semanas:
  - Definición y planificación del trabajo, la cual se entrega el 10 de octubre.
  - El estado del arte, el cual se entrega el 23 de octubre, dividiéndose en dos partes:
    - Estudio y Análisis de modelos supervisados
    - Estudio y Análisis de modelos no supervisados

    Donde para ámbos casos se muestran ejemplos de aplicación afines al trabajo.

[3] https://www.sngular.com/es/data-science-CRISP-dm-metodologia/

Primeras cuatro semanas
2022
09
10
Week 1
Week 2
Week 3
Week 4
Definición y planificación del trabajo final
Estado del arte
Estudio y Análisis de modelos supervisados
Estudio y Análisis de modelos no supervisados

- En este segundo gráfico de Gantt, se puede visualizar la planificación de las siguientes nueve semanas, donde diseñamos e implementamos los modelos. Dividiéndose en cuatro tareas a realizar:
  - Creación del dataset.
  - Desarrollo de modelos supervisados.
  - Desarrollo de modelos no supervisados.
  - Análisis de resultados.

Siguientes nueve semanas
2022
10
11
12
Week 5
Week 6
Week 7
Week 8
Week 9
Week 10
Week 11
Week 12
Week 13
Creación del dataset
Implementación de modelos supervisados
Implementación de modelos no supervisados
Análisis de resultados

- Por último, en este último gráfico de Gantt podemos ver la planificación de las últimas seis semanas:
  - Redacción de la memoria, la cual se subdivide en dos entregas. La primera es para el 8 de enero, mientras que la final es para el 15.
  - Defensa pública, dividiéndose también, en dos partes. Una preparación de la defensa hasta el día 22 de enero y una presentación pública final el 3 de febrero.

Últimas seis semanas
2022
2023
12
01
02
Week 14
Week 15
Week 16
Week 17
Week 18
Week 19
Redacción de la memoria
Primera entrega
Entrega final
Defensa pública
Preparación
Defensa

Capítulo 2

El Ecommerce: Contexto y Análisis de estudios anteriores

2.1. Contexto

El rápido desarrollo y las mejoras de la tecnología han consentido la conexión a Internet a particulares y empresas de todo el mundo. La llegada de los teléfonos inteligentes, las supercomputadoras y los dispositivos digitales inteligentes han permitido mejorar aún más dicha conexión. Hoy en día, las personas de todos los ámbitos de edad, acceden e interactuan con frecuencia a Internet generando millones de datos, los cuales se analizan para determinar su valor comercial bajo los enfoques científicos e inteligentes de la ciencia de datos (DS) y el aprendizaje automático (ML).[3]

Los científicos de datos observan el impacto de los datos desencadenados por el usuario, generalmente generados debido a las interacciones humanas con la web. Uno de los accesos frecuentes a la web son las compras en línea, generando grandes cantidades de datos[4], necesitando después, un procesamiento de dichas transacciones.

Hoy en día, las compras en línea se están convirtiendo cada vez más en una parte rutinaria de nuestra vida diaria. También está impulsando a las empresas a recopilar datos más útiles para comprender a sus clientes a fin de crear productos que cumplan con sus requisitos y expectativas reales. Pero ¿cómo se originó el comercio electrónico o el **ecommerce**?

El comercio electrónico o **ecommerce**, se refiere a la transacción de bienes y servicios a través de comunicaciones electrónicas. Aunque el público, en general, se ha familiarizado con el **ecommerce** solo en la última década, existe desde hace más de 30 años. Esta nueva técnica de

[4]De acuerdo con el informe Data never Sleeps, elaborado por el sistema operativo basado en la nube Domo, cada día se crean en Internet más de 2,5 billones de bytes de datos, que van en aumento, especialmente en las plataformas de video.

realizar transacciones de diferentes bienes, ha transformado la forma en que las personas hacen negocios.

El **ecommerce** fue posible gracias al desarrollo del Intercambio Electrónico de Datos (EDI), lo que permite la transmisión estructurada de datos entre organizaciones por medios electrónicos. Sin embargo el enorme coste de conectarse a una red EDI y problemas técnicos limitaron su difusión y solo el 1 % de las empresas de Europa y Estados Unidos lo adoptaron en la década de los 90.

El inicio de Internet se remonta a la década de 1960, comenzó como una herramienta de investigación, cuando se estableció la Red informática de la Agencia de Proyectos de Investigación Avanzada (ARPANET), para la investigación en áreas de alta tecnología. Todavía a fines de la década de 1980, Internet aún mantenía su carácter no comercial. Fue el desarrollo de una interfaz gráfica de usuario (GUI) y la navegabilidad de la World Wide Web (WWW) lo que cambió la naturaleza del uso de Internet llegando a ser accesible para todo el mundo, originando, así, interés para el mundo de los negocios.

Existen dos tipos básicos de **ecommerce**: empresa a empresa (B2B) y empresa a consumidor (B2C). En B2B, las empresas realizan negocios con sus proveedores, distribuidores y otros socios a través de redes electrónicas. Mientras que en B2C, las empresas venden productos y servicios a los consumidores. A pesar de que el B2C es el más conocido por el público en general, el B2B es el que domina el comercio electrónico en términos de ingresos.[11]

En nuestro caso, los datos proceden de un **ecommerce** con sede en Colombia que se dedica a la venta de productos relacionados con la información de empresas del país: Informes Comerciales y módulos de información detallada sobre Datos Financieros, Prensa, Administradores, Incidencias, etc, Informes Sectoriales, Base de datos a medida, Productos de Marketing (mercadeo), Información de accionistas, Información de proveedores y clientes, etc.

En esta línea y a nivel mundial se compite con diferentes competidores por la cuota de mercado como son: DB, Experian, Equifax, Bureau Van Dijk, Transunion, CRIF, Informa. Este mercado pretende dar respuesta a los siguientes puntos[5]:

- Qué hacen las empresas, cuál es su rendimiento y las personas que les dirigen.
- Datos financieros, datos de personas jurídicas, actividad y noticias sobre fusiones y adquisiciones.
- Estructura corporativa y de propiedad.

Pero, ¿a quién puede interesar este tipo de productos? Hoy en día, las empresas tienden a ser cada vez más datadriven, por lo que poseer diferente información es vital para la superviviencia

[5]https://www.bvdinfo.com/es-es/

de la mayoría de las empresas. En este sentido, las empresas más interesadas en adquirir este tipo de productos son:

- Corporativo.
- Instituciones financieras.
- Sector público y organizaciones sin animo de lucro.
- Servicios profesionales y los Big4.
- Sector academico.

2.1.1. El ecommerce y la ética

Como se ha comentado, para un **ecommerce**, por tanto, el almacenamiento de datos y el acceso a los mismos son algunos de los desafíos de las últimas décadas ya que gracias a ellos se toman diferentes decisiones que afectan directamente a las estrategias de marketing de las distintas empresas. Pero, ¿se hace un uso eticamente responsable de los datos?

La ética en el uso de datos o de algoritmos de inteligencia artificial pretende estudiar si las conclusiones obtenidas son conformes a la ética o si por el contrario, están sesgadas. Es decir, pretende revisar si al crear los algoritmos se establecen ciertas inclinaciones que no son justas. Por ejemplo, en Estados Unidos se crearon un conjunto de algoritmos para analizar la posible reincidencia de las personas que habían estado en la cárcel solo teniendo en cuenta sus imágenes. La conclusión fue que la gran mayoría de personas de color tenían la probabilidad de reincidir más alta que el resto[6].

Al aplicar diferentes algoritmos de ML se tiene como objetivo encontrar patrones con la intención de generalizar conclusiones, y por tanto, se genera un sesgo hacia las minorías. Además, para el caso de un **ecommerce** también se pueden generar otro tipo de situaciones que pueden ser poco éticas como la falta de privacidad del usuario y la seguridad de sus datos o, incluso, puede llegar a crear adicción a las compras. Aun así, la aplicación de técnicas de ML también pueden llegar a crear experiencias positivas, como publicidad personalizada con el fin de recomendar productos afines a las necesidades de los usuarios o el acceso a precios dinámicos.[7]

En conclusión, en este trabajo se analizarán los datos de forma ética ya que no se tienen información de los usuarios, por lo que no es posible identificarlos, protegiendo así su privacidad. Además, no se tienen datos que pueden dar lugar a sesgos como el género, la religión o la edad, entre otros. Finalmente, nuestro mercado es el de la venta de información de empresas, el cual no da lugar a compras desmesuradas y por tanto, no tenemos usuarios con adicción a las compras.

[6] https://elpais.com/tecnologia/2021-11-26/los-algoritmos-que-calculan-quien-va-a-reincidir-discriminan-a-los-negros-y-no-es-facil-corregirlos.html

2.1.2. El ecommerce y la sostenibilidad

Por otro lado, es importante destacar que el sector del comercio electrónico es un puente entre la digitalización y la transición hacia una economía más sostenible. Además, es un sector en constante y rápida evolución, adaptándose a las nuevas necesidades del mercado.

Como hemos citado, gracias en parte a la pandemia, el **ecommerce** se ha expandido de forma rapida en todo el mundo. Donde antes diferentes familias debían coger el coche para hacer sus compras, ahora se puede gestionar con un único coche de reparto reduciendo así las emisiones de CO_2. A pesar de esto, todo depende del tipo de transporte seleccionado, es decir, la opción más rápida de transporte contamina más que si el usuario compra presencialmente en la tienda[7].

Aun así, de lo más crítico para un **ecommerce** a nivel de sostenibilidad es el embalaje. Un embalaje sin plásticos y reduciendo al máximo el uso del papel conseguirá ser más sostenible.

En consecuencia, nuestro **ecommerce** al vender productos relacionados con la información de empresas de Colombia, no tiene problemas a nivel de sostenibilidad. Ya que el producto puede ser compartido vía email, reduciendo al mínimo tanto las emisiones de CO_2 como cualquier consumo de papel o plástico.

2.2. Repaso de literatura en modelos de ML

Como hemos comentado, un **ecommerce** puede generar miles de interacciones diarias. Con la ayuda de técnicas de ML, podemos encontrar información valiosa que se puede utilizar para comprender mejor el negocio y a sus clientes.

2.2.1. Modelo de aprendizaje supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada después de haber analizado una serie de ejemplos etiquetados, más comunmente conocidos como datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados, las situaciones nunca vistas.

Hoy en día para las empresas que desean vender en línea, vender productos y servicios a nivel mundial es considerada una ventaja, pero, por otro lado, existe la desventaja de una mayor competencia. Esto hace que sea difícil predecir el comportamiento de compra de los consumidores.

[7] https://www.adslzone.net/2019/11/15/contamina-mas-comprar-internet-huella-carbono/

La estimación de compradores potenciales puede afectar a las empresas en sus políticas de suministro de materias primas y de marketing. Una buena predicción permitirá que la empresa tenga mayor éxito. Desde la década de 1980 hasta la actualidad, las empresas han utilizado la inteligencia artificial en campos como las finanzas, la publicidad y la estimación de ventas.

En nuestro caso, el objetivo es predecir compradores potenciales ya que no todos los visitantes a una web terminan comprando, identificar posibles patrones es crucial para asegurar el éxito de cada empresa. Pero, ¿qué algoritmo se ajusta más a nuestro objetivo? En esta sección vamos a repasar algunos ejemplos de algoritmos supervisados que se han utilizado a lo largo de la historia.

Uno de los métodos de análisis para estimar la compra en la inteligencia artificial son los estudios de Dutta[4] y Nam[9], los cuales se encuentran entre los primeros ejemplos en los que los pedidos online son estimados por redes neuronales.

Por otra parte, en el trabajo realizado por Vellido, et.[13] para estimar el comportamiento de compra online de los consumidores se recogieron cuatro factores. Estos son, la percepción del producto, la experiencia de compra, el servicio al cliente y el riesgo del consumidor. Se probó con dos métodos diferentes: la regresión logística y red neuronal. Como resultado, se encontró que las redes neuronales dan mejores resultados.

En 2015, Grażyna, et. [10] realizaron un trabajo el cual abordaba el problema de la clasificación de las sesiones de usuario en una tienda online en dos clases: sesiones de compra y sesiones de navegación. Se formuló el problema de predicción de sesiones de compra como un problema de clasificación supervisada con el fin de predecir si una sesión ocurriría una compra o no. El enfoque presentado utiliza la clasificación k-Nearest Neighbors (k-NN). Se construyó un clasificador k-NN y se verificó su eficiencia para diferentes tamaños de vecindarios. Un clasificador 11-NN fue el más efectivo tanto en términos de predicciones de sesión de compra como de predicciones generales, logrando una sensibilidad del 87,5 % y una precisión del 99,85 %.

En 2019, el Dr. İbrahim Topal[12] realizó un estudio, cuyo objetivo era crear una regla significativa al estimar el comportamiento de compra de los consumidores en línea con menos datos. Después de seleccionar la función Fisher Score en una base de datos, los datos de entrenamiento y prueba se determinaron con K fold y se creó una regla con Decision Tree. Como resultado, se puede sugerir que es posible determinar el comportamiento de compra de los consumidores en línea con gran precisión mediante el uso de una sola característica.

Viendo estos estudios realizados anteriormente, en nuestro caso se utilizan y se comparan cuantitativamente los resultados obtenidos de los siguientes dos algoritmos:

- Árbol de decisión: se trata de un algoritmo que realiza un análisis sistemático para obtener reglas y relaciones valiosas de un conjunto de datos que contiene una gran cantidad de registros y, a menudo, se usa en clasificación o predicción.

Hay dos métodos diferentes en los árboles de decisión: clasificación y regresión (CRT). Si la etiqueta de clase de los datos que se quiere predecir es categórica se utiliza el árbol de clasificación, por el contrario, si es un valor numérico continuo se utiliza CRT.

- k-Nearest Neighbors Method: es una técnica de aprendizaje supervisado que se usa a menudo en el reconocimiento de patrones para la clasificación, aunque también se puede usar para estimación y predicción. Un clasificador k-NN está basado en instancias y es conceptualmente simple. Trata de estimar la probabilidad a posteriori de que un elemento pertenezca a una clase a partir de la información proporcionada por el conjunto de datos etiquetados.

Tal y como se ha visto, los modelos utilizados en este trabajo son modelos de clasificación. Este tipo de modelos se caracterizan por ser explicables, es decir, no son modelos que den lugar a una 'black-box'. Cuando se utilizan este tipo de modelos, como resultado se obtiene una predicción de una variable objetivo, la cual se ha obtenido de un conjunto de datos previamente etiquetado. Por tanto, podemos observar si en el conjunto de datos existen posibles variables que puedan ser eticamente conflictivas. Actualmente, estas variables conflictivas tienden a eliminarse de los datasets o directamente a no ser recogidas.

Por otro lado, estos modelos son faciles de implementar, por lo que consumen pocos recursos. En consecuencia, también son más sostenibles y respetuosos con el medio ambiente. Es importante destacar que a más datos a procesar menos sostenible será cualquier algoritmo que se quiera utilizar.

2.2.2. Modelo de aprendizaje no supervisado

El aprendizaje no supervisado es un método de ML donde un modelo se ajusta a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori y no está etiquetado. Es decir, mientras que para el aprendizaje supervisado el objetivo es predecir una variable, para el no supervisado el objetivo es encontrar patrones en los datos.

La segmentación del mercado se basa en los datos recopilados sobre algunas características de los clientes. Por lo tanto, cuantos más datos recopilen las empresas, más precisa será la segmentación.

Desde el punto de vista de las ventas, el análisis de clusters permite definir, entre otros, nuevas estrategias de fijación de precios o descuentos especiales según el cluster al que el cliente pertenezca. Por tanto, se puede usar para mejorar la oferta y aumentar las ventas ayudando, así, a que el negocio crezca.

Por otro lado, desde el punto de vista del marketing, los clusters pueden utilizarse para analizar el comportamiento del cliente en función de sus características. Estas características

pueden incluir aspectos como la demografía del cliente, el comportamiento de compra y el consumo.

En nuestro caso, el objetivo es segmentar en clusters a los visitantes de la web de venta online y establecer su nivel potencial de compra. Esto permitirá enfocar al cluster correcto las campañas de marketing desarrolladas por la empresa y no desperdiciar recursos innecesarios[2]. Como hemos realizado en el punto anterior, en esta sección vamos a repasar algunos ejemplos de segmentación de clientes que se han analizado a lo largo de la historia.

Li, Zeying [8] propusieron un método en el que se tomó un supermercado minorista como objeto de investigación, donde se usaron métodos de minería de datos para obtener segmentos de clientes. Luego, las reglas de asociación obtenidas al utilizar el algoritmo Apriori se usaron para diferentes grupos de clientes y obtener reglas sobre las características de los mismos. La minería de datos se usó de manera eficiente para manejar la gran cantidad de datos históricos para encontrar información útil para las tiendas minoristas.

Wang, Zhenyu, Yi Zuo, Tieshan Li, CL Philip Chen y Katsutoshi Yada [15] analizaron la segmentación de clientes basada en un amplio sistema de aprendizaje que proporciona una visión alternativa del aprendizaje en una estructura profunda. En primer lugar, además del comportamiento de compra del cliente, también se incluyeron datos de identificación por radiofrecuencia, que pueden representar con precisión el comportamiento de los consumidores en la tienda. En segundo lugar, se utilizó el Sistema "Broad Learning System"para analizar la segmentación del consumidor. Este sistema es una de las mejores técnicas de ML y, además, es eficiente para tareas de segmentación. En tercer lugar, los datos de comportamiento del cliente utilizados se recopilaron de un supermercado de Japón. La segmentación de clientes se consideró como un problema de segmentación de etiquetas múltiples.

Kansal, Tushar, Suraj Bahuguna, Vishal Singh y Tanupriya Choudhury [5] realizaron la segmentación de clientes mediante K-means. Se desarrolló un programa Python y el programa se entrenó aplicando un escalador estándar en un conjunto de datos que tenía dos características de 200 muestras de entrenamiento tomadas de una tienda minorista local. Ambas características son el promedio de la cantidad de compras de los clientes y el promedio de la visita anual del cliente a la tienda. Al aplicar el agrupamiento, se formaron 5 segmentos de clúster etiquetados como Clientes descuidados, cuidadosos, estándar, objetivo y sensibles. Sin embargo, los autores obtuvieron dos nuevos grupos al aplicar el agrupamiento de cambio medio etiquetados como Compradores altos y visitantes frecuentes y Compradores altos y visitantes ocasionales.

En 2015, Yongyi Cheng, et.[16] realizaron un estudio en el cual analizaron 100 empresas de comercio electrónico de China. Este documento adopta un enfoque de minería de datos del método DBSCAN. En la fase de preprocesamiento de datos, adopta el análisis factorial para reducir la dimensionalidad. Mientras que en la fase de agrupación, se implementa un algoritmo

DBSCAN mejorado para procesar los datos de densidad desigual. Finalmente, se ofrecieron sugerencias a estas empresas basadas en los resultados del experimento.

El agoritmo de DBSCAN, también se ha utilizado en otros campos diferentes por ejemplo, en 2010, se adaptó el algoritmo P-DBSCAN para agrupar y analizar ubicaciones y eventos en base a imágenes recopiladas. Por otro lado, en 2011, se agrupó y analizó el contenido de una página web con información sobre salarios. En 2012, se realizó una segmentación de clientes de una empresa de telecomunicaciones. Por último, en 2014, se analizó la regularidad de viajar que tiene cada pasajero.

Viendo estos análisis previos, en este trabajo se utilizan y se comparan cuantitativamente los resultados obtenidos de los siguientes dos algoritmos:

- Algoritmo K-means[17]. Es uno de los algoritmos de agrupamiento más antiguos y más utilizados. El algoritmo de K-means comienza con la selección de k objetos como centros. Esta muestra utiliza la curva de codo para decidir el valor de k. Mediante un proceso de iteración, los objetos restantes se asignarán a sus centros más cercanos. Después de que todos los objetos se asignen a un centro, se reconsiderarán los centros del grupo.

  Este algoritmo es fácil de usar y de implementar, permitiendo ser utilizado por datasets de diferente índole, además, es robusto y altamente eficiente. Sin embargo, también tiene algunas limitaciones, es difícil de manejar con algunos conjuntos de datos complicados, como puede ser un dataset que tiene una alta dimensión. Además, es sensible a valores atípicos ya que los centros de los clusters se verán afectados significativamente por ellos.

  Por otro lado, para obtener un k correcto se utilizará el Silhouette score[6], esta puntuación mide cómo de lejos están los puntos de datos del clúster más cercano en comparación con su propio clúster en promedio. Una puntuación de Silhouette más alta es más deseable que una más baja.

- Algoritmo DBSCAN: es un algoritmo de agrupamiento espacial basado en la densidad. Puede descubrir grupos de forma arbitraria y manejar valores atípicos de manera efectiva. La idea básica es que el número de puntos de datos debe ser mayor que el número mínimo en una esfera de radio dado. Además, es un enfoque efectivo para resolver problemas de clúster para diferentes **ecommerce**.

  Este algoritmo no necesita de la especificación del número de clusters deseado como lo requiere k-means. DBSCAN puede encontrar clusters con formas geométricas arbitrarias. Además, tiene noción del ruido, y es robusto detectando outliers.

  Por otro lado, DBSCAN no es enteramente determinista: los puntos borde que son alcanzables desde más de un cluster pueden etiquetarse en cualquiera de estos. Afortunadamente, esta situación no es usual, y tiene un impacto pequeño sobre el cluster. Existe

una variación de este algoritmo que trata los puntos borde como ruido, y así logra un resultado completamente determinista. Además, la calidad de DBSCAN depende de la distancia utilizada, la más usada es la distancia euclidiana. DBSCAN no puede agrupar conjuntos de datos bien con grandes diferencias en las densidades.

Tal y como se ha visto, los modelos son modelos de segmentación. Este tipo de modelos se caracterizan por ser explicables e interpretables, es decir, como en el caso anterior, no existen ningún tipo de 'black-box' en estos modelos. Cuando se utilizan este tipo de modelos, como resultado se obtiene diferentes clusters, los cuales una vez analizados se puede determinar que variables han sido las utilizadas por el algoritmo para poder definirlos. Por tanto, podemos comprender lo que el algoritmo está realizando y evitar así medidas poco éticas y sesgadas.

Por otro lado, si se compara el K-means y el DBSCAN con una red neuronal a nivel de utilización de recursos, se puede ver que consumen muchos menos recursos ya que son algoritmos más sencillos de implementar. Por lo que también son más sostenibles y respetuosos con el medio ambiente.

2.2.3. Conjunto de datos desbalanceado

Por el contario, antes de implementar cualquier algoritmo se deberá lidiar con problemas de desbalanceo ya que como se ha comentado, existe un gran desbalanceo en los datos de nuestro **ecommerce**.

La presencia de datos desbalanceados es un problema común en el análisis y preprocesamiento de datos, y ocurre principalmente muy a menudo en los problemas de clasificación donde hay una clase predominante respecto a las demás. Esto puede afectar la calidad de nuestro modelo y su capacidad para predecir correctamente. Por ello es importante conocer los diferentes métodos para tratar con datos desbalanceados y así predecir mejor las clases minoritarias y poco representadas de nuestros datos[8].

Para hacer frente a este problema existen dos métodos[1]:

- Oversampling: estas técnicas tienen como finalidad replicar instancias en la clase minoritaria a partir de la información de los casos existentes o reales. Los tres métodos de oversampling más populares son: el Random Oversampling, Synthetic Minority Over-Sampling Technique (SMOTE) y Adaptative Synthetic Sampling (ADASYN). El riesgo asumido en estas técnicas es la de generar información que no está incluida realmente en los datos.

[8]https://datasciencepe.substack.com/p/como-manejar-el-desbalance-de-datos

- Undersampling: son técnicas que tienen como finalidad igualar las distribuciones desbalanceadas de datos eliminando instancias de la clase mayoritaria. Estamos, por tanto, ante un grupo de técnicas que operan sobre los casos de la clase mayoritaria aislando o respetando la distribución de casos de la clase minoritaria. En este grupo de técnicas, el riesgo asumido es la pérdida de información original vinculada a los datos de la clase mayoritaria.

En nuestro caso probaremos a utilizar ambos métodos con el fin de comparar los resultados de cada algoritmo y, así, poder obtener diferentes conclusiones.

2.3. Enfoque y método seguido

A partir de la revisión bibliográfica hemos podido entrar en contacto con diferentes modelos aplicados al **ecommerce**, lo que nos ha permitido seleccionar los que van a ser utilizados para el posterior apartado de análisis.

A continuación, realizaremos el mismo procedimiento para escoger la metodología adecuada en base al análisis de algunos estudios previamente realizados:

- Una metodología basada en KDD para clasificar la confianza en los sistemas de un **ecommerce**[14]. Este artículo propone una metodología basada en KDD para detectar el fraude en los sistemas de pago electrónico. Para evaluar esta metodología, se define el concepto de eficiencia económica y se aplica a un conjunto de datos reales de uno de los sistemas de pago electrónico más grandes de América Latina. Los resultados muestran un muy buen desempeño, proporcionando ganancias de hasta un 46,5 % en comparación con la estrategia empleada hasta la fecha por la empresa.

- Análisis del Comportamiento del Consumidor Online utilizando una metodología CRISP-DM[18]. Este trabajo realiza una introducción sobre aplicaciones de minería de datos en un **ecommerce** y finanzas utilizando la metodología CRISP-DM. Analiza cómo los problemas específicos de la aplicación pueden afectar el desarrollo de un proyecto de minería de datos.

En nuestro caso se utilizará una metodología CRISP-DM ya que es la más extendida en el sector. Esta metodología es iterativa y consiste en seis pasos, tal y como podemos observar en la imagen 2.1[9]:

[9]https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

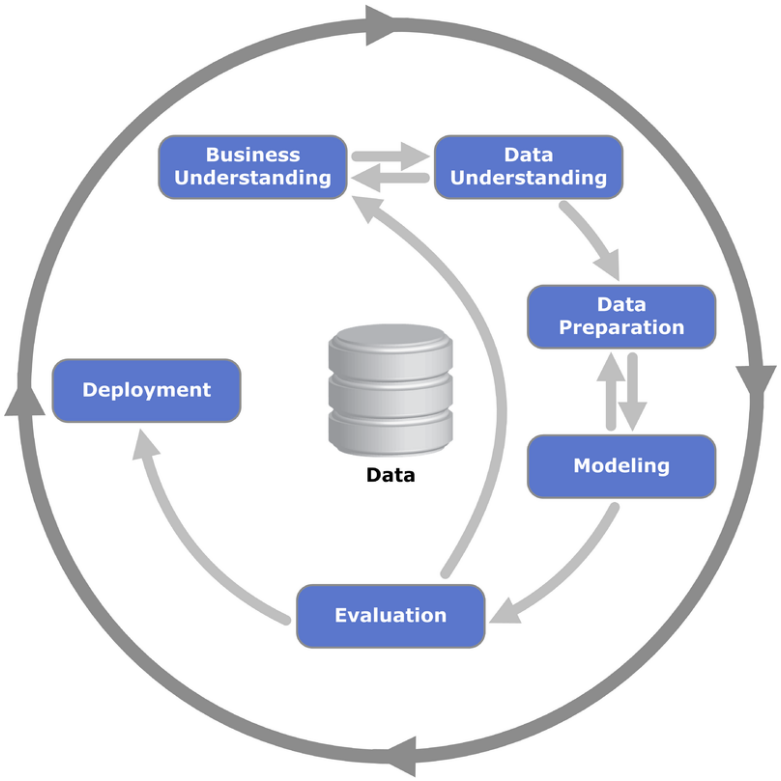

Figura 2.1: Proceso CRISP-DM

1. Comprensión del negocio: Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

2. Estudio y comprensión de los datos: La fase de entender los datos comienza identificando los problemas de calidad, descubriendo el conocimiento preliminar sobre los datos, y/o descubriendo posibles subconjuntos interesantes para formar hipótesis.

3. Análisis de los datos y selección de características: fase de preparación de datos con el objetivo de construir el conjunto final de datos para utilizarlo en el modelado. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos.

4. Modelado: se seleccionan y aplican las técnicas de modelado que sean pertinentes, y se calibran sus parámetros a valores óptimos.

5. Evaluación: En esta etapa se evaluan los modelos utilizados en la fase anterior. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

6. Despliegue: donde se proporciona la solución final.

Las fases 3 a 5 son iterativas entre sí. Del mismo modo adoptaremos un proceder Agile para las fases 4 y 5, con el fin de acelerar la generación y evaluación de los diferentes modelos.

Capítulo 3

Descripción y transformación del conjunto de datos

Tal y como se ha comentado en los capítulos previos, los datos proceden de un **ecommerce** con sede en Colombia que se dedica a la venta de productos relacionados con la información de empresas del país.

Los usuarios pueden llegar a la web del **ecommerce** desde diferentes canales como: directorios propios, webs de terceros, SEO, SEM... Dichos usuarios, tienen acceso a diferentes productos promocionales a cambio de registrarse mediante un formulario donde indican sus datos personales. Al registrarse los usuarios se convierten en "Registrados" o *"Leads"*.

Por tanto, una vez los usuarios se convierten en "Registrados" tienen acceso a productos promocionales que contienen información muy básica (Ficha de empresa promocional) sobre las empresas buscadas, dando posibilidad de consumir gratuitamente, durante 30 días, hasta cinco productos de información más detallada, llamados "Perfiles Promocionales de empresa".

El objetivo del **ecommerce** al proporcionar estos ficheros básicos es el de mostrar a los diferentes usuarios el nivel de contenido de los productos que se tienen.

Si el usuario Registrado está interesado en conocer un producto o empresa en detalle, tiene diferentes tipos de contratación:

- PPV: compra puntual de un producto.
- Bono: compra de un conjunto de productos a cambio de un pago anticipado.
- Suscripción: pagando una cantidad periódica permite el acceso y consumo de productos, limitado por el volumen de compra y por la fecha de caducidad de la suscripción.

Cuando se produce una de estas contrataciones el usuario "Registrado" pasa a ser llamado "Cliente".

3.1. Descripción del conjunto de datos

En este **ecommerce**, se disponen de cuatro diferentes conjuntos de datos, los cuales se relacionan entre ellos a través del campo IDUSUARIO. Es importante destacar, que para mantener la ética están completamente anonimizados y que además, proceden de diferentes plataformas.

3.1.1. USUARIOS

Este conjunto de datos recoge los datos del registro de los usuarios en un periodo de tiempo en concreto. Dispone de los siguientes campos:

- IDUSUARIO: Id único de usuario.
- TIPOUSUARIO: Tipo de usuario.
  - PJ: Persona Jurídica.
  - PF: Persona Física.
  - PX: Puede ser PJ pero no es seguro.
- FECHA_REGISTRO: Fecha de registro del usuario.
- CANAL_REGISTRO: Canal de registro del usuario.
  - 1: SEM.
  - 4: SEO.
  - 2, 3 y 7: directorios populares.
  - El resto son directorios especializados.
- IND_CLIENTE: Indicador de cliente, 1 si es cliente, 0 en caso contrario.
- FEC_CLIENTE: Fecha en la que el usuario se convierte en cliente.
- TIPOEMAIL: Dominio email del usuario.
- BONDAD_EMAIL: Bondad del email obtenida a través de campañas de emailing:
  - 20: Verde, es correcto.
  - 9: Naranja, ha dado un error temporal, pero se sigue enviando.
  - 1: SPAM.
  - 0: Rojo, inválido.

  - -10: Dominio inválido, inválido.
  - -20: No email.

- USU_TIPO. Recoge la ocupación del usuario, las cuales pueden ser:
  - SOCIEDAD COMERCIAL/INDUSTRIAL.
  - EMPRESARIO INDIVIDUAL.
  - ENTIDAD FINANCIERA O DE SEGUROS.
  - ENTIDAD SIN ANIMO DE LUCRO.
  - ORGANISMO ESTATAL.
  - HOLDING.
  - ENTIDAD EXTRANJERA.
  - SOCIEDAD NO COMERCIAL.
  - INDUSTRIA/COMERCIO.
- USU_TAMANIO: Tamaño de la compañía del usuario si TIPOUSUARIO=PJ:
  - GR: Grande.
  - MD: Mediana.
  - PQ: Pequeña.
  - MC: Micro.
  - SD: Sin definir.
- USU_CIIU: Código de Actividad CIIU si TIPOUSUARIO=PJ. Para más detalle consultar el siguiente pdf: https://www.dane.gov.co/files/sen/nomenclatura/ciiu/CIIURev3AC.pdf
- USU_ESTADO: Situación de la compañía si TIPOUSUARIO=PJ.
- USU_DEPARTAMENTO: Provincia de la sede del usuario si TIPOUSUARIO=PJ.

3.1.2. CONSUMOS

Este conjunto de datos recoge los datos de los consumos promocionales realizados por los usuarios. Dispone de los siguientes campos:

- IDCONSUMO: Identificador único del consumo.
- IDUSUARIO: Identificador del usuario.

- IDPRODUCTO: Identificador de producto consumido.
- DESCPRODUCTO: Descripción del producto consumido.
- FECHACONSUMO: Fecha del consumo.
- EMPCONSUL_ID: Id único de la empresa asociada al producto consumido.
- EMPCONSUL_CIIU: Código de actividad CIIU de la empresa asociada al producto consumido.
- EMPCONSUL_PROV: Departamento de la empresa asociada al producto consumido.
- EMPCONSUL_EST: Estado de la empresa asociada al producto consumido.

3.1.3. SESIONES

Este conjunto de datos recoge el número de sesiones abiertas en la web, agrupado por IDUSUARIO y fecha. Dispone de los siguientes campos:

- IDUSUARIO: Identificador del usuario.
- FECHA_SESION: Fecha sesión.
- SESIONS: Número de sesiones abiertas por el usuario en esa fecha.

3.1.4. VENTAS

Este conjunto de datos recoge los datos de las ventas realizadas al usuario. Dispone de los siguientes campos:

- IDVENTA: Identificador de registro de venta.
- IDUSUARIO: Identificador del usuario.
- FECHAVENTA: Fecha primera venta.
- TIPOVENTA: Tipo de la primera venta.
    - Venta puntual informe.
    - Venta puntual listado.
    - Suscripción.
    - Bono.

- IMPORTE: Importe de la primera venta.
- NUMVENTAS: Número de compras realizadas incluida la primera.
- IMPORTES: Suma importe de todas las ventas.
- VP Informe: 1 si se le ha vendido un informe en alguna ocasión, de lo contrario nulo.
- BONO: 1 si se le ha vendido un bono en alguna ocasión, de lo contrario nulo.
- SUSCRIPCION: 1 si se le ha vendido un suscripción en alguna ocasión, de lo contrario nulo.
- VP Listado: 1 si se le ha vendido un listado en alguna ocasión, de lo contrario nulo.

3.2. Tranformación del conjunto de datos

Con la tranformación de datos se persigue dos objetivos: Estructurar los datos de forma que los algoritmos analíticos que se van a utilizar sean capaces de extraer la máxima información de ellos, y crear variables nuevas a partir de la información que se tiene aportando gran valor para los modelos que desarrollaremos posteriormente.

3.2.1. USUARIOS

Antes de analizar cada variable, se ha revisado la proporción de clientes dentro de la tabla USUARIOS. Aquí, podemos apreciar que el dataset está muy desbalanceado, tenemos un total de 2.615 clientes frente a 140.121 de no clientes, es decir, en el dataset solo disponemos de un 1.83 % de clientes.

A continuación, se ha procedido a realizar un análisis detallado de cada campo de esta tabla:

- **TIPOUSUARIO**: se procede a eliminar del conjunto de datos los casos con la categoría 'PX', ya que no se puede determinar si es una persona jurídica o no. Se trata del 3.46 % de los datos, para los clientes supone un 5.35 % y para los no clientes un 3.42 %. Por otro lado, codificamos este campo creando la variable TIPO_USUARIO, donde si TIPOUSUARIO es igual a 'Persona física', entonces TIPO_USUARIO es 1, en caso contrario 0, y procedemos a eliminar el campo TIPOUSUARIO del conjunto de datos.
- **CANAL_REGISTRO**: se analiza que existen un 2.48 % de registros que tiene este campo vacío, es decir, un 0.73 % del total son clientes y un 2.51 % son no clientes. Por tanto, se decide eliminar estos registros, habiendo eliminado, teniendo en cuenta el campo

anterior, un total de 6.04 % de clientes y un 5.86 % de no clientes, manteniendo, por tanto, la proporción inicial.

Por otro lado, se decide agrupar los valores cuando CANAL_REGISTRO es igual a 2, 3 y 7, y cuando es diferente a 1, 2, 3, 4 y 7. Posteriormente se realiza un *One Hot Encoding*, obteniendo únicamente las variables DE (Directorios Especializados), DP (Directorios Populares) y SEM, y eliminando el resto, ya que estas variables ya guardan toda la información necesaria.

- **BONDAD_EMAIL**: se decide agrupar los valores cuando BONDAD_EMAIL es igual a 1, 0 y -10. Posteriormente, se realiza un *One Hot Encoding*, obteniendo finalmente las variables Invalido (cuando es igual a 1, 0 o -10), Naranja y Verde.

- **TIPOEMAIL**: se decide eliminar este campo ya que si no se dispone de correos electronicos, tal y como indica la variable observada anteriormente, esta variable no debería estar registrada y, por contra, lo está. Por tanto, se decide eliminar este campo ya que se considera que la calidad no es la esperada.

- **USU_TIPO**: se realiza un análisis donde se segmenta primero el dataset por Personas Físicas. Por tanto, en esta sección del dataset, este campo debería ser nulo o 'EMPRESARIO INDIVIDUAL', vemos que existen 9 registros que son iguales a 'ENTIDAD FINANCIERA O DE SEGUROS' y 2 a 'SOCIEDAD COMERCIAL/INDUSTRIAL'. Estos dos últimos casos, ninguno es cliente, por tanto procedemos a eliminarlos; mientras que de los 9 casos anteriores solo uno es cliente, los cuales también se proceden a eliminar. Por el contrario, si se segmenta por Personas Jurídicas, este campo no debería ser nulo, pero se encuentran 2775 casos donde lo son, los cuales se proceden a eliminar.

  Por otro lado, se realiza una agrupación si es diferente a nulo, a 'SOCIEDAD COMERCIAL/INDUSTRIAL' y a 'EMPRESARIO INDIVIDUAL'. A continuación, se crean las variables USUSCI (cuando el usuario es SOCIEDAD COMERCIAL/INDUSTRIAL), USUNulo y USUEI (cuando el usuario es EMPRESARIO INDIVIDUAL), y se procede a eliminar el resto de variables.

- **USU_TAMANIO**: se realiza un análisis donde se segmenta primero el dataset por Personas Físicas. Por tanto, en esta sección del dataset, este campo debería ser nulo, vemos que existen 344 Micro-Empresas y una Pequeña-Empresa. De estos 345 casos, tan solo 4 son clientes, por lo que procedemos a eliminarlos. Por otro lado, se han encontrado 44 valores que pertenecen a la categoría de SD, por lo que se convierten a nulos. Finalmente, se realiza un *One Hot Encoding*.

- **USU_CIIU**: se realiza un análisis donde se segmenta primero el dataset por Personas Físicas. Por tanto, en esta sección del dataset, este campo debería ser nulo, vemos que existen 44 casos en los que no. De estos 44 casos, tan solo 1 es cliente, por lo que procedemos a eliminarlos. Finalmente, se elimina este campo.

- **USU_ESTADO**: se obtienen dos variables nuevas de este campo, cuando es igual a 'ACTIVA' y cuando es igual a nulo, el resto de información se agruparía, y se elimina la variable original.

- **USU_DEPARTAMENTO**: se realiza un análisis donde se segmenta el dataset por Personas Jurídicas. Por tanto, en esta sección del dataset, este campo no debería ser nulo, pero vemos que existe un caso el cual no es cliente, por lo que procedemos a eliminarlo. Por otro lado, se crea la variable BOGOTA ya que la mayoría de registros se concentran en este departamento, donde es 1 si USU_DEPARTAMENTO es BOGOTA, 0 en caso contrario.

3.2.2. CONSUMOS

Tal y como hemos realizado en el caso anterior, se procede a realizar un análisis detallado de cada campo de la tabla de CONSUMOS:

- **DESCPRODUCTO**: A partir de este campo se crean dos variables, 'Ficha Básica' y 'Perfil Promocional', que recogen cuantos productos de perfil promocional o de ficha básica se consumen por ID_USUARIO.

- **EMPCONSUL_ID**: A partir de este campo se crea una variable, 'emp_dif', que recoge el número distinto de empresas consultadas por ID_USUARIO.

- **FECHACONSUMO**: A partir de este campo se crean seis variables, 'prim_consu', 'ult_consu', 'numconsultas', 'R_DaysCon', 'diasactivo' y 'difdays', que recogen el primer consumo realizado, el último, el número de consultas, el ratio de días por consulta, la diferencia entre la primera y última consulta y el número diferente de días consultados por ID_USUARIO, respectivamente.

A continuación, procedemos a añair esta información a la información anteriormente analizada utilizando el campo ID_USUARIO. Una vez realizado se obtiene una nueva variable, **'daysInactive'**, que es la diferencia entre la fecha de registro y la primera consulta.

3.2.3. SESIONES

- **ID_USUARIO**: A partir de este campo se obtiene el número de veces que está repetido, de esta forma se obtiene el número de sesiones totales por usuario.
- **FECHA_SESION**: A partir de este campo se crea una variable, 'numdias', que recoge el número distinto de días en el que se realizan diferentes sesiones por ID_USUARIO.

  Por otro lado, combinando esta nueva variable con la anterior, obtenemos la variable 'm_sesiones', que recoge de media, cuantas sesiones por día realiza cada usuario.

Como en el caso anterior, procedemos a añair esta información al dataset previamente obtenido del cruce entre las tablas CONSUMOS y USUARIOS, utilizando para ello, como en el caso anterior, el campo ID_USUARIO.

3.2.4. VENTAS

Se decide no tratar esta tabla ya que nuestro objetivo es detectar posibles compradores. Tal y como se ha definido anteriormente, esta tabla proporciona información una vez el usuario compra, por lo que no es de utilidad teniendo en cuenta nuestro objetivo final.

3.3. Dataset final

En definitiva, las variables que obtenemos de la combinación y transformación de los conjuntos de datos originales son:

- IND_CLIENTE: Indicador de cliente, 1 si es cliente, 0 en caso contrario.
- FEC_CLIENTE: Fecha en la que el usuario se convierte en cliente.
- ID_USUARIO: Id único de usuario.
- TIPO_USUARIO: tipo de usuario, 1 si es Persona Física, 0 si es Jurídica.
- DE: Canal de registro del usuario, 1 si lo hace mendiante directorios especializados, 0 en caso contrario.
- DP: Canal de registro del usuario, 1 si lo hace mendiante directorios populares, 0 en caso contrario.
- SEM: Canal de registro del usuario, 1 si lo hace mendiante SEM, 0 en caso contrario.

- Invalido: Bondad del email obtenida a través de campañas de emailing, 1 si es SPAM, Rojo o tiene el dominio inválido, 0 en caso contrario.
- Naranja: Bondad del email obtenida a través de campañas de emailing, 1 si es Naranja, 0 en caso contrario.
- Verde: Bondad del email obtenida a través de campañas de emailing, 1 si es Verde, 0 en caso contrario.
- USUSCI: Recoge la ocupación del usuario, 1 si es una SOCIEDAD COMERCIAL/INDUSTRIAL, 0 en caso contrario.
- USUNulo: Recoge la ocupación del usuario, 1 si no se tiene información de la ocupación, 0 en caso contrario.
- USUEI: Recoge la ocupación del usuario, 1 si es un EMPRESARIO INDIVIDUAL, 0 en caso contrario.
- TAM_GR: Tamaño de la compañía si se trata de una persona jurídica, 1 si es grande, 0 en caso contrario.
- TAM_MD: Tamaño de la compañía si se trata de una persona jurídica, 1 si es mediana, 0 en caso contrario.
- TAM_PQ: Tamaño de la compañía si se trata de una persona jurídica, 1 si es pequeña, 0 en caso contrario.
- TAM_MC: Tamaño de la compañía si se trata de una persona jurídica, 1 si es micro, 0 en caso contrario.
- TAM_SD: Tamaño de la compañía si se trata de una persona jurídica, 1 si está sin definir, 0 en caso contrario.
- ACTIVA: Situación de la compañía si se trata de una persona jurídica, 1 si es ACTIVA, 0 en caso contrario.
- ESTADOnull: Situación de la compañía si se trata de una persona jurídica, 1 si no está informado, 0 en caso contrario.
- DEPARTAMENTOnull: Provincia de la sede de la persona jurídica, 1 si no está informado, 0 en caso contrario.

- BOGOTA: Provincia de la sede de la persona jurídica, 1 si es BOGOTA, 0 en caso contrario.
- Ficha Básica: número de fichas básicas consumidas por ID_USUARIO.
- Perfil Promocional: número de perfiles promocionales consumidos por ID_USUARIO.
- emp_dif: número de empresas diferentes consultadas por ID_USUARIO.
- prim_consu: primer consumo realizado por ID_USUARIO.
- ult_consu: último consumo realizado por ID_USUARIO.
- consultas: número de consultas realizadas por ID_USUARIO.
- R_DaysCon: ratio de días por consulta por ID_USUARIO.
- diasactivo: diferencia entre primera y última consulta por ID_USUARIO.
- difdays: número diferente de días consultados por ID_USUARIO.
- FECREGISTRO: fecha de registro del usuario.
- daysInactive: diferencia entre la fecha de registro y de la primera consulta por ID_USUARIO.
- SESIONES: número de sesiones totales por usuario.
- numdias: número distinto de días en el que se realizan diferentes sesiones por ID_USUARIO.
- m_sesiones: recoge la media de sesiones por día que realiza cada usuario.

Después de realizar la tranformación y limpieza del conjunto de datos, se obtiene un total de 131.211 registros, de los cuales 128.806 son no clientes y 2.405 son clientes. Por lo que, se sigue mantiendo la misma proporción inicial de 1.83 % de clientes.

3.3.1. Colinealidad en el Dataset final

Para analizar la correlación entre las variables del conjunto de datos final, se seleccionan las variables que se utilizarán en los modelos supervisados y no supervisados que se presentarán en los siguientes capítulos. Las variables seleccionadas son las vistas en el capítulo anterior salvo, la variable objetivo IND_CLIENTE, las variables con formato fecha ya que no se utilizan en los modelos, FEC_CLIENTE, prim_consu, ult_consu y FECREGISTRO, y, por último, el identificador de cada usuario, ID_USUARIO, que tampoco se utiliza.

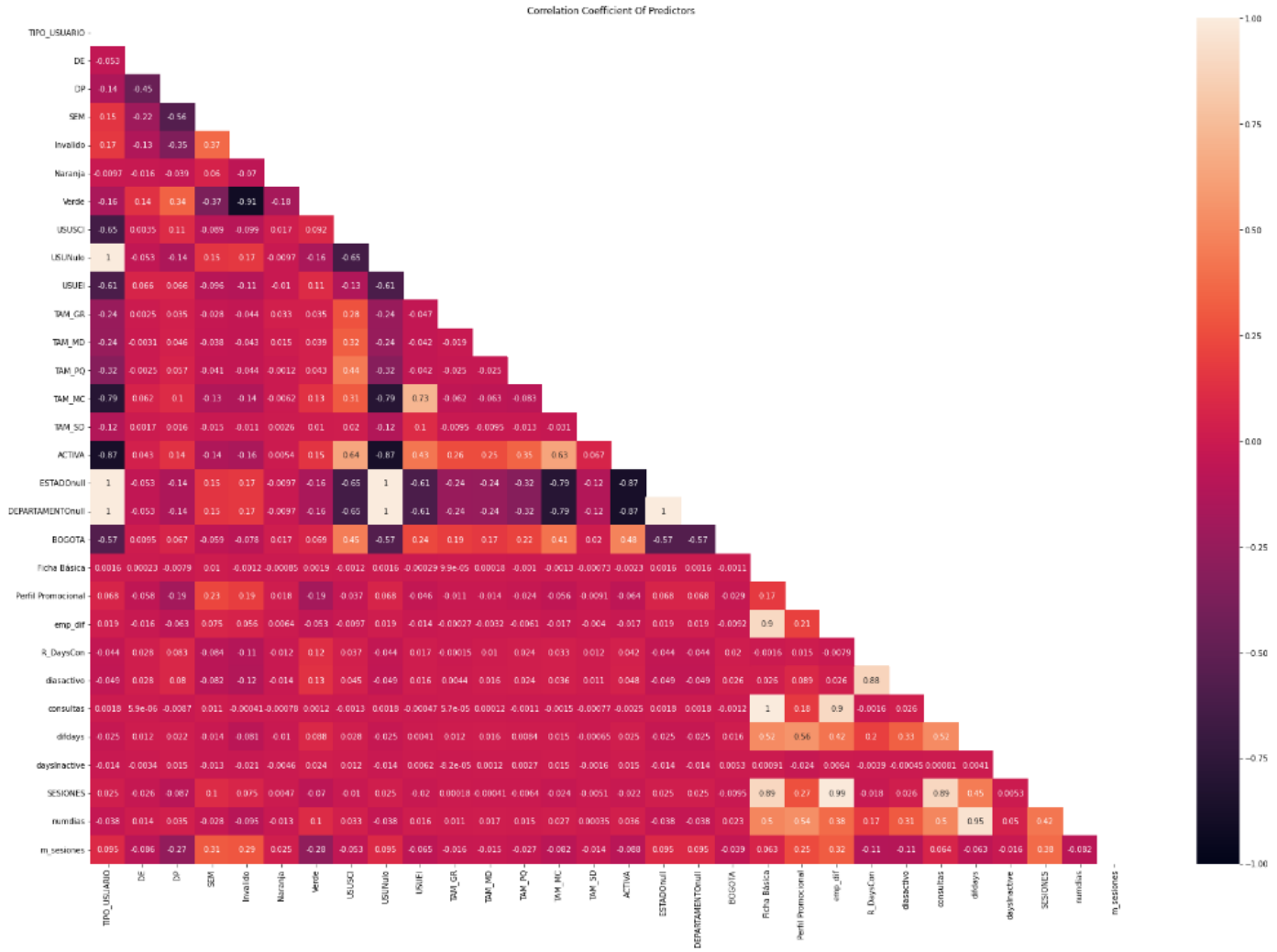

Figura 3.1: Colinealidad entre las variables

Para visualizar de forma rápida la colinealidad entre las variables, se realiza un gráfico de la matriz de correlación diagonal 3.1.

Gracias a este gráfico podemos identificar fuertes relaciones entre variables, como por ejemplo:

- emp_dif y Ficha básica
- ACTIVA y UsuNulo
- TAM_MC y TIPO_USUARIO
- SESIONES y consultas

Por lo tanto, este gráfico indica que se debe eliminar algunas variables para evitar repetir información. Para saber cuales eliminar sin eliminar demasiada información relevante, se calculan los valores VIF (Factor de Inflación de Varianza).

El factor de inflación de la varianza mide cuánto se ve afectada la varianza de una variable independiente por su interacción o correlación con las otras variables independientes. Por tanto, estos valores permiten una medición rápida de cuánto contribuye una variable al error estándar en la regresión. Cuando VIF es igual a uno, indica que las variables no están correlacionadas; si están entre uno y cinco, indican una correlación moderada; por último, si es mayor a cinco, indica una correlación alta.

Al obtener estos valores en nuestro conjunto de datos, se puede observar que existen 18 variables que superan el 5, de las cuales siete tienen un VIF de infinito, las cuales son: 'consultas', 'USUNulo', 'TIPO_USUARIO', 'Perfil Promocional', 'Ficha Básica', 'ESTADOnull' y 'DEPARTAMENTOnull'.

En consecuencia, se eliminarán una a una las variables que tengan el VIF más alto, hasta conseguir que todas tengan un valor por debajo de 5. Para realizar este proceso, se elimina una variable y se vuelve a obtener el VIF de todas las restantes. Este proceso se realiza así, ya que al eliminar una variable los VIF varían, por lo que así, se evita eliminar información relevante.

Por lo tanto, en nuestro caso se han realizado los siguientes pasos:

1. Se elimina la variable 'consultas'. Lo cual reduce a 16 las variables que tienen un VIF por encima de 5, de las cuales solo 4 tienen un VIF de infinito: 'USUNulo', 'TIPO_USUARIO', 'ESTADOnull' y 'DEPARTAMENTOnull'.

2. Se elimina la variable 'USUNulo'. Lo cual reduce a 15 las variables que tienen un VIF por encima de 5, de las cuales solo 3 tienen un VIF de infinito: 'TIPO_USUARIO', 'ESTADOnull' y 'DEPARTAMENTOnull'.

3. Se elimina la variable 'TIPO_USUARIO'. Lo cual reduce a 14 las variables que tienen un VIF por encima de 5, de las cuales solo 2 tienen un VIF de infinito: 'ESTADOnull' y 'DEPARTAMENTOnull'.

4. Se elimina la variable 'ESTADOnull'. Lo cual reduce a 13 las variables que tienen un VIF por encima de 5, de las cuales ninguna tienen un VIF de infinito. Donde la variable con un VIF mayor es 'SESIONES'.

5. Se elimina la variable 'SESIONES'. Lo cual reduce a 12 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'DEPARTAMENTOnull'.

6. Se elimina la variable 'DEPARTAMENTOnull'. Lo cual reduce a 11 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'difdays'.

7. Se elimina la variable 'difdays'. Lo cual reduce a 10 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'TAM_MC'.

8. Se elimina la variable 'TAM_MC'. Lo cual reduce a 6 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'emp_dif'.

9. Se elimina la variable 'emp_dif'. Lo cual reduce a 4 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'Verde'.

10. Se elimina la variable 'Verde'. Lo cual reduce a 2 las variables que tienen un VIF por encima de 5. Donde la variable con mayor VIF es 'numdias'.

11. Al eliminar la variable 'numdias', se consigue que el resto de variables tengan un VIF por debajo de 5, es decir, se consigue que el resto de variables tengan una correlación moderada.

Finalmente, las variables que se utilizarán en los modelos, sin tener en cuenta la variable objetivo IND_CLIENTE, serán:

- DE
- DP
- SEM
- Invalido
- Naranja

- USUSCI
- USUEI
- TAM_GR
- TAM_MD
- TAM_PQ
- TAM_SD
- ACTIVA
- BOGOTA
- Ficha Básica
- Perfil Promocional
- R_DaysCon
- diasactivo
- daysInactive
- m_sesiones

Capítulo 4

Resultados Modelos de aprendizaje Supervisado

Tal y como se ha comentado previamente, el aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Su objetivo es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada después de haber analizado una serie de ejemplos etiquetados, más comunmente conocidos como datos de entrenamiento. Para ello, el algoritmo generaliza a partir de los datos presentados las situaciones nunca vistas.

Antes de comenzar a utilizar cualquier tipo de algoritmo, se divide el conjunto de datos en train y test, donde obtenemos 497 clientes para el test y 1.908 para el train.

4.1. DecisionTree

Tal y como se vió en el capítulo 2, se aplica un árbol de decisión al conjunto de datos, donde para conocer el rendimiento del modelo, se utiliza el 'classification_report' de la librería sklearn, el cual nos devuelve las métricas principales. Tal y como podemos ver en la imagen 4.1.

Además, para poder comparar entre los modelos, también se obtiene el área bajo la curva ROC, obteniendo así el AUC, que para este caso nos da un AUC de 0.67.

Por otro lado, se revisan las variables más relevantes de este modelo. Las tres más importantes han sido 'm_sesiones', 'Ficha Básica' y 'Perfil Promocional', en orden descendente.

4.2. KNN

Como en el caso anterior, se aplica un k vecinos más próximos al conjunto de datos, donde también utilizamos el 'classification_report' de la librería sklearn. Obteniendo así, las principales

	precision	recall	f1-score	support
0	0.99	0.99	0.99	25746
1	0.42	0.35	0.38	497
accuracy			0.98	26243
macro avg	0.70	0.67	0.68	26243
weighted avg	0.98	0.98	0.98	26243

Figura 4.1: DecisionTree

	precision	recall	f1-score	support
0	0.98	1.00	0.99	25746
1	0.62	0.18	0.28	497
accuracy			0.98	26243
macro avg	0.80	0.59	0.63	26243
weighted avg	0.98	0.98	0.98	26243

Figura 4.2: KNN

métricas 4.2. En este caso, se obtiene un AUC de 0.59.

Para comparar los resultados de ambos modelos, además del AUC, se observa también el f1-score de la categoría clientes, ya que esta metrica se utiliza cuando el dataset está desbalanceado como es el caso. Por tanto, mientras que en el árbol de decisión tenemos un f1-score de 0.38, en el KNN tenemos un 0.28, además si observamos el AUC también podemos ver que los resultados son mejores en el DecisionTree. Aun así, ambos resultados no son concluyentes, por lo que, en los siguientes apartados, se prueban dos algoritmos más sotisficados como son el RandomForest y el XGBoost.

4.3. RandomForest

Un Random Forest es un conjunto de árboles de decisión combinados con bagging. Utilizando bagging permite que los distintos árboles vean distintas porciones del conjunto de datos. Es decir, ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y, por tanto, obtenemos una predicción que generaliza mejor.[10]

[10]https://www.iartificial.net/random-forest-bosque-aleatorio/

	precision	recall	f1-score	support
0	0.99	1.00	0.99	25746
1	0.58	0.25	0.35	497
accuracy			0.98	26243
macro avg	0.78	0.62	0.67	26243
weighted avg	0.98	0.98	0.98	26243

Figura 4.3: RandomForest

	precision	recall	f1-score	support
0	0.98	1.00	0.99	25746
1	0.79	0.15	0.25	497
accuracy			0.98	26243
macro avg	0.89	0.58	0.62	26243
weighted avg	0.98	0.98	0.98	26243

Figura 4.4: XGBoost

Al aplicar este algoritmo en nuestro conjunto de datos obtenemos las siguientes métricas 4.3. En este caso, se obtiene un AUC de 0.62.

Por otro lado, se revisan las variables más relevantes de este modelo. Las tres más importantes son las mismas que las obtenidas en el DecisionTree.

Si comparamos los resultados del RandomForest con los del DecisionTree, se aprecia que tanto f1-score como el AUC son mejores en el DecisionTree. Por tanto, parece ser que un modelo más simple se ajusta mejor a nuestro dataset.

4.4. XGBoost

Un XGBoost es un algoritmo de aprendizaje automático basado en un árbol de decision y utiliza un marco de potenciación de gradientes. Este algoritmo es, actualmente, de los más utilizados para problemas tabulares o estructurados.[11]

Al aplicar este algoritmo en nuestro conjunto de datos obtenemos las siguientes métricas 4.4. En este caso, se obtiene un AUC de 0.58.

Por otro lado, se revisan las variables más relevantes de este modelo. Las tres más importantes, en este, caso han sido 'Invalido', 'USUSCI' y 'ACTIVA', en orden descendente.

Si se compara los resultados del XGBoost con los del DecisionTree, se aprecia que tanto

[11] https://datascience.eu/es/programacion/xgboost-4/

	precision	recall	f1-score	support
0	0.99	0.95	0.97	25746
1	0.16	0.51	0.25	497
accuracy			0.94	26243
macro avg	0.58	0.73	0.61	26243
weighted avg	0.97	0.94	0.96	26243

Figura 4.5: Random Oversampling Decision Tree

f1-score como el AUC son mejores en el DecisionTree. Por tanto, podemos concluir que, viendo también el caso anterior, un modelo más simple se ajusta mejor a nuestro dataset.

En consecuencia, ya que el mejor modelo hasta ahora es el DecisionTree y los resultados no son todo lo buenos que se esperan, en los siguientes apartados se realizan dos técnicas de Oversampling y dos de Undersampling para aplicar más tarde los algoritmos KNN y DecisionTree, con el objetivo de mejorar estos primeros resultados.

4.5. Random Oversampling

Esta técnica consiste en duplicar aleatoriamente ejemplos en la clase minoritaria, en este caso, en la clase de clientes.

4.5.1. Random Oversampling DecisionTree

Se aplica el DecisionTree al nuevo conjunto de datos, es decir, tras aplicar el Random Oversampling al conjunto de datos original, obteniendo el mismo número de registros para cada clase. Se obtienen las siguientes principales métricas 4.5 y un AUC de 0.73.

Por otro lado, se revisan las variables más relevantes para este caso. Donde las tres más importantes son 'm_sesiones', 'Perfil Promocional' e 'Invalido', en orden descendente.

Si comparamos los resultados obtenidos con el DecisionTree, podemos ver que el AUC es mejor en el caso del DecisionTree con Oversampling, pero el f1-score lo es sin el Oversampling. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo, ya que al tener el dataset desbalanceado, el F1-score tiene en cuenta no solo la cantidad de errores de predicción que comete su modelo, sino que también analizan el tipo de errores que se cometen, por lo que es una métrica más fiable en esta clase de conjunto de datos.

	precision	recall	f1-score	support
0	0.99	0.94	0.97	25746
1	0.16	0.58	0.26	497
accuracy			0.94	26243
macro avg	0.58	0.76	0.61	26243
weighted avg	0.98	0.94	0.95	26243

Figura 4.6: Random Oversampling KNN

4.5.2. Random Oversampling KNN

Como en el caso anterior, se aplica el KNN al nuevo conjunto de datos balanceado. Se obtienen las siguientes principales métricas 4.6 y un AUC de 0.76.

Si comparamos los resultados obtenidos con el DecisionTree, podemos ver que el AUC es mejor en el caso del KNN con Oversampling, pero el f1-score es mejor para el DecisionTree. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

4.6. SMOTE Oversampling

Esta técnica en lugar de sobremuestrear aleatoriamente con reemplazo como en el caso anterior, SMOTE toma cada muestra minoritaria e introduce puntos de datos sintéticos que conectan la muestra minoritaria y sus vecinos más cercanos. Es importante destacar que los vecinos de los k vecinos más cercanos se eligen al azar.

4.6.1. SMOTE Oversampling DecisionTree

Se aplica el DecisionTree al nuevo conjunto de datos, es decir, tras aplicar el SMOTE Oversampling al conjunto de datos original. Se obtienen las siguientes principales métricas 4.7 y un AUC de 0.74.

Por otro lado, se revisan las variables más relevantes de este modelo. Siendo las tres más imptantes han sido las mismas que para el árbol de decisión con el Random Oversampling.

Si comparamos los resultados obtenidos con el DecisionTree, podemos ver que el AUC es mejor en el caso del DecisionTree con Oversampling, pero el f1-score sigue siendo mejor sin el Oversampling. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

	precision	recall	f1-score	support
0	0.99	0.96	0.98	25746
1	0.20	0.51	0.29	497
accuracy			0.95	26243
macro avg	0.59	0.74	0.63	26243
weighted avg	0.98	0.95	0.96	26243

Figura 4.7: SMOTE Oversampling DecisionTree

	precision	recall	f1-score	support
0	0.99	0.94	0.96	25746
1	0.16	0.63	0.26	497
accuracy			0.93	26243
macro avg	0.58	0.78	0.61	26243
weighted avg	0.98	0.93	0.95	26243

Figura 4.8: SMOTE Oversampling KNN

4.6.2. SMOTE Oversampling KNN

Como en el caso anterior, se aplica el KNN al nuevo conjunto de datos balanceado. Se obtienen las siguientes principales métricas 4.8 y un AUC de 0.79.

Para este caso, si comparamos los resultados obtenidos con el DecisionTree, podemos ver que el AUC es mejor en el caso del KNN con Oversampling, pero el f1-score es mejor para el DecisionTree. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

4.7. Random Undersampling

Esta técnica selecciona y elimina de forma aleatoria datos de la clase mayoritaria. Despues del sampling, la clase mayoritaria debe tener el mismo número de datos que la clase minoritaria.

4.7.1. Random Undersampling DecisionTree

Se aplica el DecisionTree al nuevo conjunto de datos, es decir, tras aplicar el Random Undersampling al conjunto de datos original, eliminando diferentes registros de los no clientes de forma aleatoria hasta igualar el número de registros de los clientes. Se obtienen las siguientes principales métricas 4.9 y un AUC de 0.86.

	precision	recall	f1-score	support
0	1.00	0.86	0.92	25746
1	0.11	0.86	0.19	497
accuracy			0.86	26243
macro avg	0.55	0.86	0.56	26243
weighted avg	0.98	0.86	0.91	26243

Figura 4.9: Random Undersampling DecisionTree

	precision	recall	f1-score	support
0	1.00	0.84	0.91	25746
1	0.09	0.85	0.17	497
accuracy			0.84	26243
macro avg	0.54	0.85	0.54	26243
weighted avg	0.98	0.84	0.90	26243

Figura 4.10: Random Undersampling KNN

Por otro lado, se revisan las variables más relevantes de este modelo. Las tres más importantes han sido 'm_sesiones', 'Ficha Básica' y 'Perfil Promocional', en orden descendente.

Por último, comparamos los resultados obtenidos con el DecisionTree. Podemos ver que el AUC es mejor en el caso del DecisionTree con Undersampling, pero el f1-score sigue siendo mejor sin el Undersampling. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

4.7.2. Random Undersampling KNN

Se aplica el KNN al nuevo conjunto de datos obtenido al aplicar el Random Undersampling, y se obtienen las siguientes principales métricas 4.10 y un AUC de 0.85.

Si comparamos los resultados obtenidos con el DecisionTree inicial, podemos ver que el AUC es mejor en este caso, pero el f1-score sigue siendo mejor sin el Undersampling. Por tanto, para este caso la técnica del Random Undersampling no ayuda a mejorar los resultados y siguen siendo mejores en el caso del DecisionTree sin ninguna técnica de balanceo.

	precision	recall	f1-score	support
0	0.98	0.19	0.32	25746
1	0.02	0.84	0.04	497
accuracy			0.21	26243
macro avg	0.50	0.52	0.18	26243
weighted avg	0.97	0.21	0.32	26243

Figura 4.11: NearMiss Undersampling DecisionTree

4.8. NearMiss Undersampling

NearMiss es una técnica de balanceo que ayuda a equilibrar un conjunto de datos desbalanceado. Para lograr esto, observa la distribución de clases y elimina aleatoriamente muestras de la clase más grande, en nuestro caso de la clase de no clientes. Cuando dos puntos que pertenecen a diferentes clases están muy cerca en la distribución, elimina el punto de datos de la clase más grande equilibrando la distribución.

4.8.1. NearMiss Undersampling DecisionTree

Se aplica el DecisionTree al nuevo conjunto de datos obtenido tras aplicar al conjunto de datos original la técnica de balanceado NearMiss Undersampling. Donde se obtienen las siguientes principales métricas 4.11 y un AUC de 0.52.

Por otro lado, se revisan las variables más relevantes de este modelo. Las tres más importantes han sido 'R_DaysCon', 'Ficha Básica' y 'm_sesiones', en orden descendente.

Si comparamos los resultados obtenidos con el DecisionTree original, podemos ver que tanto AUC como el f1-score son peores en este caso. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

4.8.2. NearMiss Undersampling KNN

Tal y como se ha realizado en el caso anterior, se aplica el KNN al nuevo conjunto de datos. Se obtienen las siguientes principales métricas 4.12 y un AUC de 0.44.

Si comparamos los resultados obtenidos con el DecisionTree original, podemos ver que, también en este caso, tanto AUC como el f1-score se obtienen peores resultados. Por tanto, los resultados siguen siendo mejores para el caso del DecisionTree sin ninguna técnica de balanceo.

Finalmente, tal y como se ha visto, no se ha conseguido mejorar los resultados del primer árbol de decisión, aun aplicando diferentes técnicas de balanceo. Es por ello que se realia un

	precision	recall	f1-score	support
0	0.97	0.21	0.34	25746
1	0.02	0.67	0.03	497
accuracy			0.22	26243
macro avg	0.49	0.44	0.19	26243
weighted avg	0.95	0.22	0.34	26243

Figura 4.12: NearMiss Undersampling KNN

	precision	recall	f1-score	support
0	0.98	1.00	0.99	25746
1	0.40	0.05	0.09	497
accuracy			0.98	26243
macro avg	0.69	0.52	0.54	26243
weighted avg	0.97	0.98	0.97	26243

Figura 4.13: PCA Random Forest

PCA, con el objetivo de ver si podemos mejorar estos resultados.

4.9. PCA

El Análisis de Componentes Principales es un método estadístico cuya utilidad radica en la reducción de la dimensionalidad del dataset con el que se trabaja. Esta técnica se utiliza cuando se quiere simplificar la base de datos, ya sea para disminuir el número de variables, o para comprender de una forma más sencilla un dataset.

Al aplicar el PCA al conjunto de datos inicial podemos ver que si se escogen 12 componentes ya se consigue explicar casi el 90 % de la varianza. A continuación, como en los casos anteriores, se obtienen las principales métricas 4.13 depués de aplicar un Random Forest al conjunto de datos obtenido depués del PCA y un AUC de 0.52.

Como se puede apreciar, el f1-score de la categoría minoritaria es muy bajo, por lo que se decide no continuar probando diferentes modelos con la reducción de dimensionalidad.

4.10. Conclusiones

En definitiva, se puede apreciar que sin aplicar ninguna técnica de balanceo se obtienen los siguientes resuntados:

- Decision Tree: Un AUC de 0.66 y un F1-score de 0.38.
- KNN: Un AUC de 0.59 y un F1-score de 0.28.
- Random Forest: Un AUC de 0.62 y un F1-score de 0.35.
- XGBoost: Un AUC de 0.58 y un F1-score de 0.25.

Por tanto, tal y como se ha comentado previamente, los mejores resultados si observamos el F1-score de la clase minoritaria, son los del Decision Tree, frente al XGBoost, que obtiene los peores resultados. En consecuencia, para este ecommerce, se puede apreciar que un modelo más simple se ajusta mejor que un modelo más complejo. Además, al observar el AUC, llegamos a la misma conclusión.

Por otro lado, si realizamos técnicas de Oversampling obtenemos los siguientes resultados:

- Decision Tree Random Oversampling: Un AUC de 0.73 y un F1-score de 0.25.
- KNN Random Oversampling: Un AUC de 0.76 y un F1-score de 0.26.
- Decision Tree SMOTE Oversampling: Un AUC de 0.74 y un F1-score de 0.29.
- KNN SMOTE Oversampling: Un AUC de 0.78 y un F1-score de 0.26.

Por tanto, si se observa el F1-score, se puede visualizar que los mejores resultados se obtienen cuando al Decision Tree se le aplica una técnica de SMOTE Oversampling. Por el contrario, si se observa el AUC, los mejores resultados se obtienen al aplicar la misma tecnica de Oversampling al modelo KNN. Finalmente, se decide que el mejor modelo al aplicar las técnicas de Oversampling es el de Decision Tree con SMOTE.

Si se compara este modelo, con el Decision Tree previo, se ve que los resultados de este último son mejores. Por lo que se concluye que las técnicas de Oversampling no sirven, en este caso, para mejorar los resultados.

Por último, si realizamos técnicas de Undersampling obtenemos los siguientes resultados:

- Decision Tree Random Undersampling: Un AUC de 0.86 y un F1 de 0.19.
- KNN Random Undersampling: Un AUC de 0.85 y un F1 de 0.17.
- Decision Tree NearMiss Undersampling: Un AUC de 0.52 y un F1 de 0.04.
- KNN NearMiss Undersampling: Un AUC de 0.44 y un F1 de 0.03.

Por tanto, si se observa tanto el F1-score como el AUC, se ve que los mejores resultados se obtienen cuando al Decision Tree se le aplica una técnica de Random Undersampling.

Si se compara estos resultados, con el Decision Tree sin aplicar ninguna técnica de balanceo, se aprecia que se obtienen mejores resultados en este último. Por lo que se concluye que las tecnicas de Undersampling tampoco ayudan a mejorar los resultados.

En conclusión, los mejores resultados son los obtenidos de aplicar un árbol de decisión al conjunto de datos original. Donde el orden de la relevancia de las variables para obtener estos resultados ha sido el siguiente:

1. m_sesiones
2. Ficha Básica
3. Perfil promocional
4. R_DaysCon
5. diasactivo
6. DP
7. BOGOTA
8. ACTIVA
9. SEM
10. Invalido
11. USUSCI
12. daysInactive
13. DE
14. TAM_PQ
15. TAM_MD
16. USUEI
17. TAM_GR
18. Naranja
19. TAM_SD

Capítulo 5

Resultados Modelos de aprendizaje No Supervisado

5.1. K-means

Antes de aplicar este algorítmo se debe obtener el número de clusters o grupos óptimo, es decir, se debe calcular la k óptima. Para ello, se utiliza el método del codo o, en inglés, *Elbow Method*. Por tanto, al aplicar este método se pueden visualizar el número de clusters óptimos.

En la gráfica 5.1 se puede apreciar que la k optima no está muy bien definida, podemos ver que una k igual a 5 podría ser la óptima, ya que en la gráfica la pendiente cambia a partir de este punto. Aun así, para contrastar esta hipótesis, se prueba con otra técnica: The silhouette value. Este método, mide que tan similar es un punto a su propio grupo (cohesión) en comparación con otros grupos (separación).

En esta gráfica 5.2 se puede visualizar que k debería ser 10. Aun así, se elige k=5 ya que si se

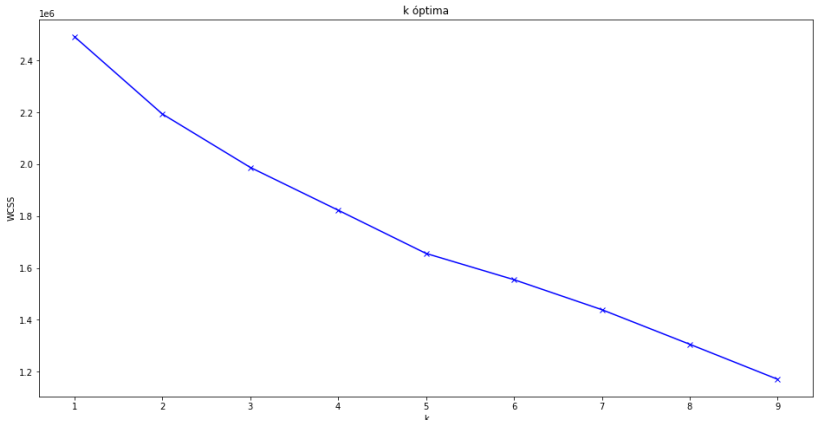

Figura 5.1: El método del codo

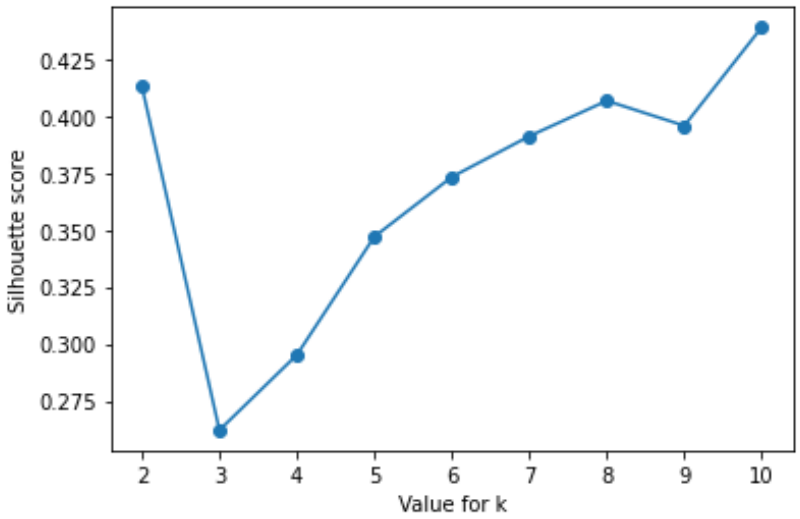

Figura 5.2: The silhouette value

elige una k demasiado grande, es más complicado explicar los clusters. Además, si se contempla la gráfica, se puede observar que a partir del 5 la pendiente se calma y la mejora no es tan notoria.

Una vez elegido el número de clusters óptimo, se aplica el algoritmo k-means al conjunto de datos, guardando el cluster al que pertenece cada registro en una nueva variable llamada 'Cluster'. Se puede apreciar que los usuarios se reparten de la siguiente forma:

- 46.43 % en el cluster 0.
- 11.23 % en el cluster 1.
- 26.16 % en el cluster 2.
- 3.03 % en el cluster 3.
- 13.16 % en el cluster 4.

A continuación, procederemos a analizar el impacto que tiene cada variable a cada grupo o cluster:

- En la tabla 5.1 analizamos la variable IND_CLIENTE, donde podemos ver que hay clusters como el 1 y el 3 que tienen una representación mayor de clientes, un 6.32 % y un 4.1 %, respectivamente, que la distribución original, que tiene un 1.83 %. Mientras que para el cluster 2 la proporción de clientes es menor, con un 0.83 %.

Cuadro 5.1: Distribución de los clientes

	Clientes	No clientes	Total usuarios
Cluster 0	817	60.104	60.921
Cluster 1	931	13.801	14.732
Cluster 2	286	34.035	34.321
Cluster 3	163	3.810	3.973
Cluster 4	208	17.056	17.264
Total	2.405	128.806	131.211

- En la tabla 5.2 analizamos la variable DE, donde podemos ver que la proporción de usuarios que se registran en la distribución original utilizando este método es de 14.92 %. Vemos que hay clusters como el 3 y 4 que tienen una representatividad mayor de usuarios que utilizan este canal con un 17.42 % y 100 %, respectivamente. En cambio, el resto de clusters, es decir, los cluster 0, 1 y 2 vemos que la proporción es más baja con un 0 %, 10.8 % y 0.1 %, respectivamente.

  Por tanto, vemos que mientras en el cluster 4, todos son usuarios que realizan el registro utilizando este canal, el cluster 0, no lo hace ninguno.

Cuadro 5.2: Distribución de los usuarios que se registran utilizando directorios especializados

	No DE	DE	Total usuarios
Cluster 0	60.921	0	60.921
Cluster 1	13.141	1.591	14.732
Cluster 2	34.288	33	34.321
Cluster 3	3.281	692	3.973
Cluster 4	0	17.264	17.264
Total	111.631	19.580	131.211

- En la tabla 5.3 analizamos la variable DP, donde podemos ver que la proporción de usuarios que se registran mediante este método en la distribución original es de 53.93 %. Además, vemos que hay en todos los clusters, salvo en el 2 y 4, proporcionalmente, más usuarios que se registran utilizando este método. Mientras que, por otro lado, en el cluster 4 no lo hace ninguno y en el cluster 2 solo el 0.79 %.

- En la tabla 5.4 analizamos la variable SEM, donde podemos ver que la proporción de usuarios que se registran utilizando este método en la distribución original es de 21.23 %. Por el contrario, vemos que solo el cluster 2 es mayor, en proporción, comparado con la distribución original, tenemos un 76 % de los clientes que se registran mediante SEM. El

Cuadro 5.3: Distribución de los usuarios que se registran utilizando directorios populares

	No DP	DP	Total usuarios
Cluster 0	4.190	56.731	60.921
Cluster 1	3.978	10.754	14.732
Cluster 2	34.049	272	34.321
Cluster 3	967	3.006	3.973
Cluster 4	17.264	0	17.264
Total	60.448	70.763	131.211

resto de clusters, lo hace en una proporción menor, siendo en el cluster 0 y 4, el 0 % de los usuarios.

Cuadro 5.4: Distribución de los usuarios que se registran utilizando SEM

	No SEM	SEM	Total usuarios
Cluster 0	60.917	4	60.921
Cluster 1	13.151	1.581	14.732
Cluster 2	8.235	26.086	34.321
Cluster 3	3.786	187	3.973
Cluster 4	17.264	0	17.264
Total	103.353	27.858	131.211

- En la tabla 5.5 analizamos la variable Invalido, donde podemos ver que la proporción de usuarios con el email inválido en la distribución original es de 26.35 %. Por otro lado, vemos que tan solo el cluster 2 es mayor, en proporción comparado con la distribución original, tenemos un 67.33 % de los clientes que tienen un email inválido. El resto de clusters, lo hace en una proporción menor.

Cuadro 5.5: Distribución de los usuarios que tienen el email inválido

	No Inválido	Inválido	Total usuarios
Cluster 0	53.859	7.062	60.921
Cluster 1	12.702	2.030	14.732
Cluster 2	11.212	23.109	34.321
Cluster 3	3.893	80	3.973
Cluster 4	14.967	2.297	17.264
Total	96.633	34.578	131.211

- En la tabla 5.6 analizamos la variable Naranja, donde podemos ver que la proporción de usuarios con este tipo de email en la distribución original es de 1.34 %. Además, vemos que tan solo los clusters 1 y 2 son mayores en proporción, comparado con la distribución

original, tenemos un 2.02 % y un 2.26 % de los clientes que tienen un email Naranja, respectivamente. El resto de clusters, lo hace en una proporción menor.

Cuadro 5.6: Distribución de los usuarios que tienen el email Naranja

	No Naranja	Naranja	Total usuarios
Cluster 0	60.380	541	60.921
Cluster 1	14.434	298	14.732
Cluster 2	33.547	774	34.321
Cluster 3	3.954	19	3.973
Cluster 4	17.140	124	17.264
Total	129.455	1.756	131.211

- En la tabla 5.7 analizamos la variable USUSCI, donde podemos ver que la proporción de usuarios en la distribución original es de 11.89 %. Por otro lado, vemos que tan solo los clusters 1 y 3 son mayores en proporción, comparando con la distribución original, tenemos un 7.77 % y un 15.18 % de los usuarios que son una sociedad comercial/industrial, respectivamente. El resto de clusters, lo hace en una proporción menor.

Cuadro 5.7: Distribución de los usuarios que son una SOCIEDAD COMERCIAL/INDUSTRIAL

	No USUSCI	USUSCI	Total usuarios
Cluster 0	60.584	337	60.921
Cluster 1	1.144	13.588	14.732
Cluster 2	34.037	284	34.321
Cluster 3	3.370	603	3.973
Cluster 4	16.476	788	17.264
Total	115.611	15.600	131.211

- En la tabla 5.8 analizamos la variable USUEI, donde podemos ver que la proporción de usuarios en la distribución original es de 10.62 %. Además, vemos que los clusters 0, 3 y 4 son mayores en proporción, comparado con la distribución original, tenemos un 14.52 %, 14.02 % y un 16.85 % de los usuarios que son empresarios individuales, respectivamente. El resto de clusters, lo hace en una proporción menor.

- En la tabla 5.9 analizamos la variable TAM_GR, donde podemos ver que la proporción de usuarios en la distribución original es de 1.85 %. Por otro lado, vemos que tan solo el cluster 1 es mayor en proporción, comparado con la distribución original, tenemos un 15.73 % de los usuarios que son empresas grandes. El resto de clusters, lo hace en una proporción menor, destacando que los clusters 0 y 2 son próximos al 0 %.

Cuadro 5.8: Distribución de los usuarios que son empresarios individuales

	No USUEI	USUEI	Total usuarios
Cluster 0	52.077	8.844	60.921
Cluster 1	14.590	142	14.732
Cluster 2	32.836	1.485	34.321
Cluster 3	3.416	557	3.973
Cluster 4	14.355	2.909	17.264
Total	117.274	13.937	131.211

Cuadro 5.9: Distribución de los usuarios que son empresas grandes

	No TAM_GR	TAM_GR	Total usuarios
Cluster 0	60.909	12	60.921
Cluster 1	12.415	2.317	14.732
Cluster 2	34.311	10	34.321
Cluster 3	3.912	61	3.973
Cluster 4	17.231	33	17.264
Total	128.778	2.433	131.211

- En la tabla 5.10 analizamos la variable TAM_MD, donde podemos ver que la proporción de usuarios en la distribución original es de 1.88 %. También, vemos que tan solo el cluster 1 es mayor en proporción, comparado con la distribución original, tenemos un 15.98 % de los usuarios que son empresas medianas. El resto de clusters, lo hace en una proporción menor, destacando que los clusters 0 y 2 son próximos al 0 %.

Cuadro 5.10: Distribución de los usuarios que son empresas medianas

	No TAM_MD	TAM_MD	Total usuarios
Cluster 0	60.905	16	60.921
Cluster 1	12.378	2.354	14.732
Cluster 2	34.316	5	34.321
Cluster 3	3.904	69	3.973
Cluster 4	17.237	27	17.264
Total	128.740	2.471	131.211

- En la tabla 5.11 analizamos la variable TAM_PQ, donde podemos ver que la proporción de usuarios en la distribución original es de 3.26 %. Además, vemos que tan solo el cluster 1 es mayor en proporción, comparado con la distribución original, tenemos un 27.6 % de los usuarios que son empresas pequeñas. El resto de clusters, lo hace en una proporción menor, destacando que los clusters 0 y 2 son próximos al 0 %.

Cuadro 5.11: Distribución de los usuarios que son empresas pequeñas

	No TAM_PQ	TAM_PQ	Total usuarios
Cluster 0	60.908	13	60.921
Cluster 1	10.666	4.066	14.732
Cluster 2	34.307	14	34.321
Cluster 3	3.820	153	3.973
Cluster 4	17.231	33	17.264
Total	126.932	4.279	131.211

- En la tabla 5.12 analizamos la variable TAM_SD, donde podemos ver que la proporción de usuarios en la distribución original es de 0.47 %. Además, vemos que tan solo el cluster 2 es menor en proporción, comparado con la distribución original, con un 0.23 % de los usuarios. El resto de clusters, lo hace en una proporción mayor.

Cuadro 5.12: Distribución de los usuarios que son empresas sin definir

	No TAM_SD	TAM_SD	Total usuarios
Cluster 0	60.587	334	60.921
Cluster 1	14.653	79	14.732
Cluster 2	34.241	80	34.321
Cluster 3	3.937	36	3.973
Cluster 4	17.173	91	17.264
Total	130.591	620	131.211

- En la tabla 5.13 analizamos la variable ACTIVA, donde podemos ver que la proporción de empresas activas en la distribución original es de 19.63 %. Vemos que los clusters 1 y 3 son mayores en proporción, comparado con la distribución original, con un 93.37 % y un 26.6 % de los usuarios, respectivamente. El resto de clusters, lo hace en una proporción menor. Destacar, por tanto, que en el cluster 1, casi todos los usuaios son compañías activas.

Cuadro 5.13: Distribución de las empresas activas

	No ACTIVA	ACTIVA	Total usuarios
Cluster 0	53.928	6.993	60.921
Cluster 1	977	13.755	14.732
Cluster 2	33.268	1.053	34.321
Cluster 3	2.956	1.017	3.973
Cluster 4	14.323	2.941	17.264
Total	105.452	25.759	131.211

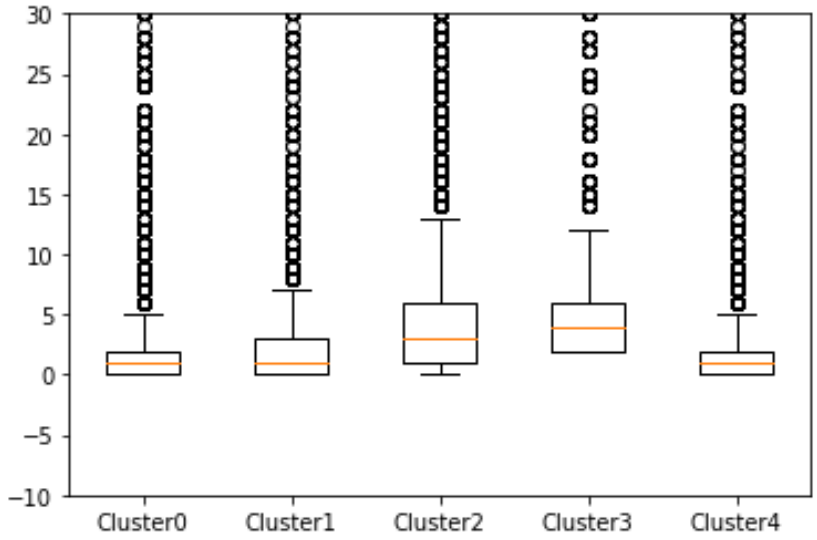

Figura 5.3: Box-plot de Ficha Básica por cluster

- En la tabla 5.14 analizamos la variable BOGOTA, la proporción de usuarios en la distribución original es de 9.43 %. Vemos que los clusters 1 y 3 son mayores en proporción, comparado con la distribución original, con un 49.71 % y un 10.75 % de los usuarios, respectivamente. El resto de clusters, lo hace en una proporción menor.

Cuadro 5.14: Distribución de los usuarios con sede en Bogotá

	No BOGOTA	BOGOTA	Total usuarios
Cluster 0	58.046	2.875	60.921
Cluster 1	7.409	7.323	14.732
Cluster 2	33.483	838	34.321
Cluster 3	3.546	427	3.973
Cluster 4	16.351	913	17.264
Total	118.835	12.376	131.211

- En la imagen 5.3 podemos visualizar la distribución de la variable 'Ficha Básica' por cada cluster, donde podemos visualizar que los clusters 0, 1 y 4 consumen menos en media, y que los clusters 2 y 3 lo hacen más. Además, remarcar que el cluster 2 es el que tiene outliers más altos.

- En la imagen 5.4 podemos visualizar la distribución de la variable 'Perfil Promocional' por cada cluster, donde podemos visualizar que los clusters 0, 1 y 4 consumen menos en media, y que los clusters 2 y 3 lo hacen más. Además, remarcar que el cluster 2 es el que tiene outliers más altos.

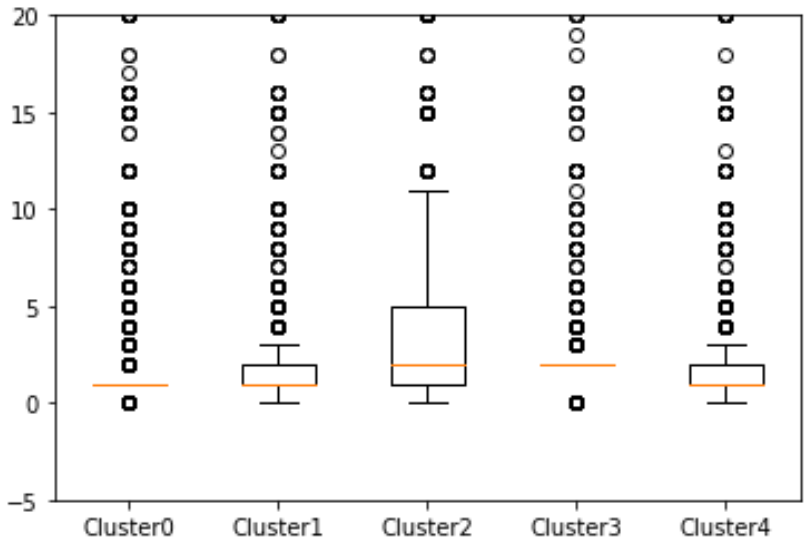

Figura 5.4: Box-plot de Perfil Promocional por cluster

- En la imagen 5.5 podemos visualizar la distribución de la variable 'R_DaysCon' por cada cluster, donde podemos visualizar que cuanto más grande es este ratio, más probabilidad de pertenecer al clsuter 3.

- En la imagen 5.6 podemos visualizar la distribución de la variable 'diasactivo' por cada cluster, donde como podemos ver, como en el caso anterior, cuanto más grande este ratio, más probabilidad de que esté en el cluster 3.

- Al realizar lo mismo para la variable 'daysInactive', vemos que el box-plot no nos es de utilidad, es por ello que se utiliza la función describe para obtener los principales estadísticos de esta variable segmentada por clusters, donde como conclusión podemos decir que el cluster 1 es el grupo con mayor diferencia entre la fecha de registro y la primera consulta, y por el contrario, el cluster 2 es el que tiene menor diferencia.

- Por último, en la imagen 5.7 podemos visualizar la distribución de la variable 'm_sesiones' por cada cluster, donde el cluster 3 consume menos en media, y que el cluster 2 lo hace más.

Después de analizar cada variable detalladamente observando el impacto que tiene cada una a los clusters, se puede concluir que cada cluster tiene las siguientes características:

- Cluster 0.
    - Los usuarios no se registran mediante directorios especializados, ni mediante SEM.

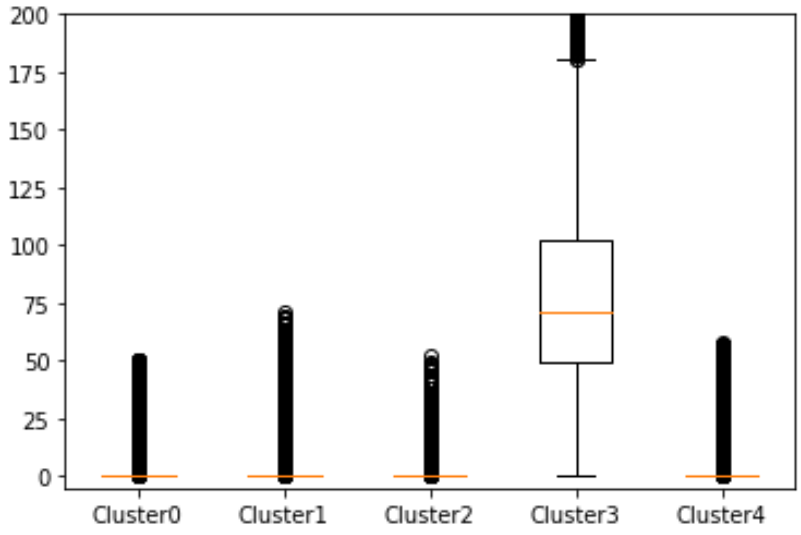

Figura 5.5: Box-plot de R_DaysCon por cluster

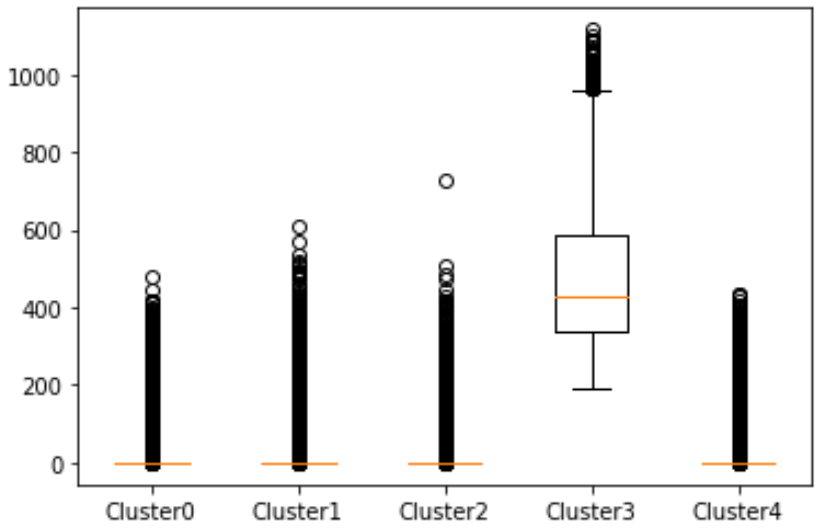

Figura 5.6: Box-plot de diasactivo por cluster

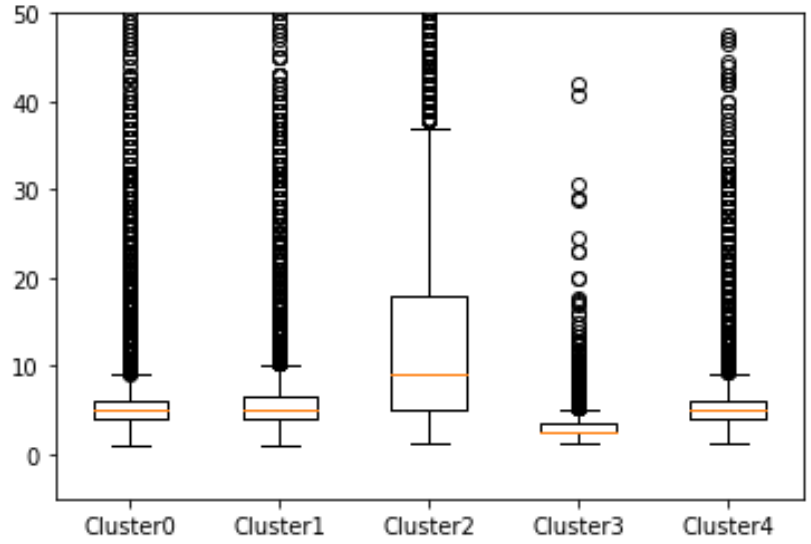

Figura 5.7: Box-plot de m_sesiones por cluster

- Son personas físicas.
- Son usuarios con un alto interés, ya que cuando se registran hacen rápidamente la primera consulta.

- Cluster 1.
  - Proporcionalmente, es el cluster con mayor número de clientes.
  - Más del 25 % son empresas pequeñas.
  - La mitad son empresas activas.
  - Son los usuarios que más tardan en hacer su primera consulta desde el registro, esto puede deberse a que tienen plena confianza en la marca y de que podrán encontrar la información que necesitan.

- Cluster 2.
  - No hay clientes.
  - Los usuarios no se registran mediante directorios especializados, ni directorios populares, sino que mayoritariamente lo hacen a través de SEM.
  - Son personas físicas.
  - Consultan más los productos.
  - Son los que más sesiones diarias hacen.

- Cluster 3.
  - Consultan más los productos.
  - Pasan muchos días entre la primera y última consulta, por lo que tardan más en hacerse clientes.
  - Son los que menos sesiones diarias hacen.
- Cluster 4.
  - Son usuarios que se registran mediante directorios especializados.
  - Guarda el mayor porcentaje, en proporción, de empresarios individuales.

5.1.1. Conclusiones

Por tanto, se puede concluir que el cluster menos recomendado para nuestro ecommerce es el 2, ya que son usuarios que consumen muchos productos de forma gratuita pero no terminan haciendose clientes nunca, es decir, no terminan comprando nunca.

Por el contrario, se ve que el cluster 1 es el más interesante para negocio, ya que, proporcionalmente, es el que más clientes guarda. Al no realizar la consulta inmediatamente después de la fecha de registro, muestra que son clientes que saben que la empresa les puede ayudar a conseguir los documetos que requieran, por lo que muestran confianza con la marca.

Finalmente, se puede ver que el cluster 0 son los que tienen mayor potencial en convertirse en clientes, ya que muestran mayor interés al ser los que menos tardan en realizar la primera consulta. Por tanto, a este grupo de usuarios se podría plantear realizar una acción de marketing con el objetivo de aumentar la conversión de clientes.

5.2. DBSCAN

Se ha intentado ejecutar este algoritmo sobre el conjunto de datos, pero finalmente, no se han podido obtener resultados por falta de RAM.

Capítulo 6

Propuestas de mejora

En este trabajo se han analizado los diferentes conjuntos de datasets, se han realizado diferentes técnicas de balanceado, se han utilizado diferentes modelos supervisados con el objetivo de predecir posibles compradores de un ecommerce, y se ha realizado un clustering del conjunto de datos obteniendo conclusiones relevantes para el futuro de la empresa.

Aun así, se han apreciado diferentes mejoras a realizar para este ecommerce:

1. Como se ha comentado, el objetivo de este análisis era el de crear clusters de usuarios y el de predecir compradores. Es por eso por lo que no se ha utilizado la tabla de VENTAS, pero un análisis que se podría hacer es analizar la recurrencia de compra de los clientes (usuarios que ya han comprado) y ver su comportamiento. Así, la empresa podría determinar si se están fidelizando a los clientes, o por el contrario, se están perdiendo.

2. Gracias al código CIIU que se dispone en el conjunto de datos, se podría realiza un grafo, para determinar como se relacionan las empresas entre si y ver, así, el interés que tienen las empresas respecto a otras. Esto podría ser útil para nuestro ecommerce a la hora de lanzar campañas de diferentes productos.

3. Realizar un modelo para predecir una probabilidad de compra teniendo en cuenta el momento de vida del usuario. Es decir, tener un modelo que devuelva, según la accion que acaba de hacer el posible cliente, una probabilidad de compra. Por ejemplo, si un usuario se registra un determinado día y se vuelve a conectar al cabo de dos y consulta diferentes productos, ver como la probabilidad de compra va aumentando a medida que va realizando acciones.

4. Utilizar un ordenador con más RAM que nos permita implementar el algoritmo DBSCAN, ya que como se ha visto en el capítulo dos, en diferentes análisis se han obtenido buenos resultados con este algoritmo.

Bibliografía

[1] Joaquín García Abad. Comparativa de técnicas de balanceo de datos. Aplicación a un caso real para la predicción de fuga de clientes. Master's thesis, Universidad de Oviedo, Oviedo, España, 2021.

[2] Vani Ashok; Rahul R Kamath; Adithya RK; Supreeth Singh; Ajay Bhati. Customer segmentation in ecommerce. *Journal of Emerging Technologies and Innovative Research*, 8:908–913, July 2021.

[3] Hussain Saleem; Khalid Bin Muhammad; Altaf Hussain Nizamani; Samina Saleem; Jamshed Butt. Data science and machine learning approach to improve e-commerce sales performance on social web. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 12:401–424, April 2021.

[4] S.; Wong W. Y. Dutta, S.; Shekhar. Decision support in non-conservative domains: Generalization with neural networks. 11:527–544, June 1994.

[5] Tushar Kansal, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. Customer segmentation using k-means clustering. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 135–139, 2018.

[6] Panayu Keelawat. E-Commerce Customer Clustering. Master's thesis, University of California San Diego, California, US, 2015.

[7] Thi Mai Le and Shu-Yi Liaw. Effects of pros and cons of applying big data analytics to consumers' responses in an e-commerce context. *Sustainability*, 9(5), 2017.

[8] Zeying Li. Research on customer segmentation in retailing based on clustering model. 06 2011.

[9] T. Nam, K.; Schaefer. Forecasting international airline passenger traffic using neural networks. *The Logistics and Transportation Review*, 31:239–252, Novembre 1995.

[10] Grażyna Suchacka;Magdalena Skolimowska-Kulig; Aneta Potempa. A k-nearest neighbors method for classifying user sessions in e-commerce scenario. *Journal of Telecommunications and information technology*, 3:64–69, 2015.

[11] Yan Tian. *History of E-Commerce*. IGI Global, 2007.

[12] Ibrahim Topal. Estimation of online purchasing intention using decision tree. *Journal of Management and Economics Research*, 17:269–280, Decembre 2019.

[13] P. J. G.; Meehan K. Vellido, A.; Lisboa. Quantitative characterization and prediction of online purchasing behavior: A latent variable approach. *International Journal of Electronic Commerce*, 4:83–104, Decembre 2015.

[14] Jose Felipe Junior; Adriano C. M. Pereira; Wagner Meira Jr.; Adriano Veloso. A kdd-based methodology to rank trust in e-commerce systems. *Conference: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 1, Novembre 2013.

[15] Zhenyu Wang, Yi Zuo, Tieshan Li, C. L. Philip Chen, and Katsutoshi Yada. Analysis of customer segmentation based on broad learning system. In *2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 75–80, 2019.

[16] Yongyi Cheng; Yumian Yang; Jianhua Jiang; GaoChao Xu. Cluster analysis of e-commerce sites with data mining approach. *International Journal of Database Theory and Application*, 8:343–354, August 2015.

[17] Bohan Zhao. Research on using market segmentation to do recommendation in e-commerce. In *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, pages 3017–3022. Atlantis Press, 2022.

[18] Carlos Soares; Yonghong Peng; Jun Meng; Takashi Washio; Zhi-Hua Zhou. Applications of data mining in e-business finance: Introduction. *Conference: Proceedings of the 2008 conference on Applications of Data Mining in E-Business and Finance*, 177:1–9, June 2008.