

Analítica de Clientes y Predicción de Demanda mediante Modelos de Machine Learning en el Sector Retail



Abel Mora Vázquez

Nombre del Programa
Grado Ciencia de Datos Aplicada

Nombre Tutor/a de TF
Antonio Gutiérrez Blanco

Profesor/a responsable de la asignatura
Susana Acedo Nadal

Fecha Entrega
07 de enero de 2025

Universitat Oberta
de Catalunya



Universitat Oberta
de Catalunya

uoc.edu



Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada

[3.0 España de Creative Commons](#)

FICHA DEL TRABAJO FINAL

Título del trabajo:	Analítica de Clientes y Predicción de Demanda mediante Modelos de Machine Learning en el Sector Retail
Nombre del autor:	Abel Mora Vázquez
Nombre del consultor/a:	Antonio Gutiérrez Blanco
Nombre del PRA:	Susana Acedo Nadal
Fecha de entrega (mm/aaaa):	01/2025
Titulación o programa:	<i>Grado de Ciencia de Datos Aplicada</i>
Área del Trabajo Final:	Analítica de clientes
Idioma del trabajo:	Castellano
Palabras clave	Analítica de clientes, Aprendizaje automático, Predicción demanda
Resumen del Trabajo	<p>El trabajo tiene como finalidad la práctica de diferentes áreas estudiadas en el grado aplicadas a un entorno empresarial. En la actualidad los datos que una empresa dispone a causa de su proceso de negocio como por ejemplo, ventas, compras, stock, datos de clientes tienen mucho valor ya que se pueden obtener insights relevantes del estado de la empresa, áreas de mejora y los tipos de clientes pudiendo estudiar su distribución geográfica, características de clientes, conocer los hábitos de compra, y sus gustos o preferencias permitiendo con ello poder realizar predicciones de venta por artículos o predicciones de venta en función del mes o semana del año. El contenido del trabajo se desarrolla en una empresa del sector retail realizando análisis exploratorio, procesamiento de lenguaje natural, analítica de clientes mediante clusterización junto con reglas de asociación y predicción de la demanda. Debido a la importancia y el valor que tienen estos datos para las empresas los datos que se utilizan son datos artificiales. Los resultados obtenidos reflejan las limitaciones de los datos sintéticos los cuales en algunos casos se han debido de imputar.</p> <p>Las conclusiones del trabajo son que la ciencia de datos se puede aplicar en varios campos del sector retail. Se puede definir la estrategia comercial de la empresa en función del tipo de cliente, así como descubrir nuevos tipos de cliente para tratar de fidelizarlos, optimizar el stock, las horas del personal, y los procedimientos logísticos para asegurar disponer de suficiente oferta para cubrir la demanda esperada.</p>

Abstract

The purpose of the work is to practice different areas studied in the degree applied to a business environment. Currently, the data that a company has due to its business process, such as sales, purchases, stock, and customer data, are very valuable as they can provide relevant insights into the state of the company, areas for improvement, and the types of customers by studying their geographical distribution, customer characteristics, purchasing habits, and their tastes or preferences, thereby allowing for sales predictions by item or sales predictions based on the month or week of the year. The content of the work is developed in a retail sector company performing exploratory analysis, natural language processing, customer analytics through clustering along with association rules, and demand forecasting. Due to the importance and value of these data for companies, the data used are artificial. The results obtained reflect the limitations of synthetic data which in some cases had to be imputed. The conclusions of the work are that data science can be applied in various fields of the retail sector. It is possible to define the company's commercial strategy based on the type of customer, as well as to discover new types of customers in an attempt to retain them, optimize stock, staff hours, and logistical procedures to ensure there is enough supply to meet the expected demand.

Tabla de contenido

1.	Introducción	1
1.1	Contexto y justificación del Trabajo	1
1.2	Objetivos del Trabajo.....	2
1.3	Impacto en sostenibilidad, ético-social y de diversidad.....	3
1.3.1	Dimensión sostenibilidad.	3
1.3.2	Dimensión comportamiento ético y de responsabilidad social.....	4
1.3.3	Dimensión diversidad, género y derechos humanos.....	4
1.4	Enfoque y método seguido	4
1.5	Planificación del Trabajo	5
1.6	Breve sumario de productos obtenidos.....	7
1.7	Breve descripción de los otros capítulos de la memoria	7
2.	Materiales y métodos	8
2.1	Obtención de datos.....	8
2.2	Análisis exploratorio y limpieza de datos	8
2.3	Análisis de la ética y legalidad de los datos	9
2.4	Minería de textos Procesamiento de lenguaje Natural.....	12
2.5	Analítica de Clientes.....	15
2.6	Reglas de Asociación	17
2.7	Análisis y predicción de demanda	19
3.	Resultados obtenidos.....	20
3.1	Obtención de datos	20
3.2	Análisis Exploratorio y limpieza de datos.....	24
3.3	Análisis de la ética y legalidad de los datos	29
3.4	Minería de textos Procesado de lenguaje Natural	30
3.5	Analítica de clientes.....	33
3.5.1	Análisis de clientes por generación, ingresos y educación.....	33
3.5.2	Análisis Perspectiva de género	39
3.5.3	Segmentación de clientes	40
3.6	Reglas de asociación	46
3.7	Predicción de demanda	49
3.7.1	PROPHET	54
4	Conclusiones y trabajos futuros	61
5	Glosario	64
6	Bibliografía	64
7	Anexos.....	73

Lista de figuras

<i>Figura 1 Tareas Diagrama de Gantt</i>	5
<i>Figura 2 Diagrama de Gantt.....</i>	6
<i>Figura 3 Guía ODI [8]</i>	11
<i>Figura 4: Gráfico de solicitudes de datos a empresas.</i>	22
<i>Figura 5 Diagrama Relación datasets</i>	29
<i>Figura 6 Sumatorio de productos por categoría</i>	32
<i>Figura 7 Distribución de porcentaje de categorías</i>	32
<i>Figura 8 Distribución de generaciones.....</i>	34
<i>Figura 9 Distribución de salarios</i>	36
<i>Figura 10 Distribución de educación</i>	37
<i>Figura 11 Distribución de salario por generación.....</i>	38
<i>Figura 12 Distribución de educación por generación</i>	38
<i>Figura 13 Distribución de generaciones por genero</i>	39
<i>Figura 14 Distribución de educación por género.....</i>	39
<i>Figura 15 Distribución de salario por género.....</i>	40
<i>Figura 16: Primeras filas customer_info.....</i>	41
<i>Figura 17: Número de clústers elbow</i>	41
<i>Figura 18: Número de clusters silhouette.</i>	42
<i>Figura 19: Distribución de clusters de clientes.....</i>	45
<i>Figura 20: Reglas de asociación</i>	47
<i>Figura 21: Reglas de asociación ordenadas</i>	47
<i>Figura 22: Distribución de ventas USA por día de la semana</i>	48
<i>Figura 23: Distribución de ventas USA por día de la semana y mes</i>	51
<i>Figura 24: Registro de ventas USA por día.</i>	52
<i>Figura 25: Sumatorio días de la semana sin ventas en tiendas de USA</i>	52
<i>Figura 26: Registro de ventas y días con venta por tipo de tienda USA</i>	53
<i>Figura 27: Métricas modelos Prophet.....</i>	54
<i>Figura 28: Predicción de ventas en supermercados de USA modelo Prophet</i>	55
<i>Figura 29: Frecuencia de ventas en supermercados periodo vacacional.</i>	56
<i>Figura 30: Frecuencia de ventas en supermercados semanal</i>	56
<i>Figura 31: Frecuencia de ventas en supermercados mensual y trimestral.....</i>	57
<i>Figura 32: Frecuencia de ventas en supermercados anual</i>	57
<i>Figura 33: Métricas modelos entrenados Prophet.....</i>	56
<i>Figura 34: Métricas modelos crossvalidation Prophet</i>	59
<i>Figura 35: Comparación modelos Prophet.....</i>	59

Lista de tablas

Tabla 1 Tabla de Gestión de riesgos.....	6
Tabla 2 Solicitud de datos a empresas en España	21
Tabla 3 Solicitud de datos a empresas en Europa y Reino Unido.....	21
Tabla 4 Solicitud de datos a empresas en Estados Unidos de América.....	22
Tabla 5 df_calendar	24
Tabla 6 df_customers	25
Tabla 7 df_products.....	26
Tabla 8 df_regions	26
Tabla 9 df_returns.....	27
Tabla 10 df_stores	27
Tabla 11 df_trans97	28
Tabla 12 df_trans98	28
Tabla 13 Clientes que no han realizado compras por generación	35
Tabla 14: Comparación de clusters de clientes	43
Tabla 15: Comparación de mejores modelos.....	60

1. Introducción

1.1 Contexto y justificación del Trabajo

La ciencia de datos nos permite extraer conocimiento de datos que disponemos. En las empresas existen muchos datos que pueden aportar valor si se recopilan y tratan de una manera eficiente. Para el sector retail existen preguntas muy interesantes como:

- ¿Por qué un cliente compra un producto?
- ¿Por qué los clientes compran un producto en una empresa o en otra?
- ¿Qué demanda existe del consumidor objetivo?
- ¿Cómo obtener nuevos clientes y mantener los que existen?
- ¿Se puede ofrecer el producto necesario que satisface la demanda?
- ¿Existen productos que influyan en la compra de otros productos?
- ¿El cliente está satisfecho con los productos y el servicio?

Este proyecto se centra en la analítica de clientes y la predicción de demanda, explorando cómo las empresas pueden utilizar los datos para mejorar la satisfacción del cliente. El contexto del trabajo final de grado es extraer datos relevantes del comportamiento de los clientes en el sector retail. El interés en este contexto surge de la experiencia laboral en el sector y en el auge que se está experimentando en la recogida de datos en el sector mediante cookies, aplicaciones, cuentas de usuario o tarjetas de fidelización. La escalabilidad es muy alta ya que pasamos del pequeño comercio donde se conoce al cliente y sus preferencias a poder trabajar los datos en bruto y detectar áreas de mejora en las ventas, optimizar ofertas para que diferentes

tipos de clientes acudan al establecimiento en un día determinado, así como tener el stock suficiente para satisfacer la demanda. La temática elegida es el análisis del comportamiento de los clientes realizando segmentación de los clientes para obtener perfiles de clientes en función de sus hábitos de compra. Posteriormente obtener reglas de asociación que permitan conocer conjuntos de productos frecuentes en las transacciones, crear un modelo de recomendación de productos basados en los datos de clientes con comportamientos de compra similares y realizar un análisis de la venta en los establecimientos mediante series temporales y predecir la venta en semanas o meses siguientes.

1.2 Objetivos del Trabajo

Los objetivos de este trabajo son los siguientes:

- Poner en práctica los conocimientos adquiridos durante el grado y obtener una visión del sector retail que actualmente no tengo dado mi puesto actual.
- Obtener información específica del comportamiento de los clientes en diferentes tiendas e intentar descubrir posibles estrategias para la mejora de resultados tanto económicos como de satisfacción de los clientes.
- Realizar segmentación de los clientes y mediante reglas de asociación realizar un modelo de recomendación de productos basados en los datos de clientes con comportamientos de compra similares.
- Realizar un análisis de la venta en los establecimientos y mediante series temporales poder predecir la venta en meses o años siguientes.
- Realizar un panel de control que pueda ser utilizado por diferentes departamentos como marketing, CRM, logística, compras que ofrezca información relevante visualmente.

1.3 Impacto en sostenibilidad, ético-social y de diversidad

Para analizar el impacto del trabajo en estos campos se tiene como referencia la competencia de compromiso ético y global (CCEG) que se alinean con los Objetivos de Desarrollo Sostenible (ODS) en inglés (SDG). La definición por parte de las Naciones Unidas es: *“Los Objetivos de Desarrollo Sostenible son un llamado a la acción por parte de todos los países (pobres, ricos y de ingresos medios) para promover la prosperidad y al mismo tiempo proteger el planeta. Reconocen que poner fin a la pobreza debe ir de la mano de estrategias que generen crecimiento económico y aborden una variedad de necesidades sociales, incluidas la educación, la salud, la protección social y las oportunidades laborales, al tiempo que se aborda el cambio climático y la protección ambiental.”* [1]. Son por tanto unos propósitos que favorecen a la sociedad y al planeta.

Se analiza el impacto en las 3 dimensiones sostenibilidad, ética y diversidad.

1.3.1 Dimensión sostenibilidad.

El trabajo final de grado tiene un impacto negativo medioambiental y en huella ecológica. Para el desarrollo del trabajo se ha de utilizar unos dispositivos y acceder a servidores que consumen energía, así como una vez entregado se hará uso de dispositivos de almacenamiento sea físico o en la nube que también tienen un impacto negativo. El impacto negativo no se puede cuantificar, pero sí que se ha tratado de mitigar haciendo un uso responsable de los medios utilizados para la ejecución tanto físicos como tecnologías, tratando de optimizar los procesos para que tarden el mínimo tiempo posible y con ello el menos uso de energía que puede contribuir a la contaminación. La entrega digital del trabajo contribuye a reducir su impacto medioambiental ya que no es necesaria la impresión en hojas de papel ni el uso de tinta.

1.3.2 Dimensión comportamiento ético y de responsabilidad social.

El trabajo tiene un impacto en comportamiento ético y de responsabilidad social. Para mitigarlo se ha realizado siguiendo las normativas europea y española de protección de datos y siguiendo una guía de comportamiento ético tratando de no tener un impacto negativo. En cuanto a la propiedad intelectual se siguen las normativas y los usos de licencia que han establecido los autores de las referencias utilizadas. Los resultados obtenidos de este trabajo no ponen en riesgo o empeoran algún tipo de trabajo ya que se trata de un trabajo académico no existiendo ninguna relación comercial con ninguna empresa.

1.3.3 Dimensión diversidad, género y derechos humanos.

El trabajo tiene un impacto en la diversidad de género y derechos humanos negativo. Se tratan datos de personas por lo que existe la necesidad de tomar medidas que mitiguen este impacto. Las medidas tomadas son el tratamiento de los datos siguiendo las normativas europea y española de protección de datos, así como realizar un apartado que trata la perspectiva de género.

1.4 Enfoque y método seguido

El enfoque de este trabajo es práctico ya que los datos pueden no ser replicables en otras empresas y no se pretende realizar un artículo científico u obtener un nuevo conocimiento que pueda ser relevante para la sociedad. El método a seguir es la realización de tareas de la planificación inicial junto con la planificación de PEC de la asignatura. Se realiza análisis exploratorio de datos, procesamiento de lenguaje natural, analítica de clientes mediante análisis socio demográficos, clustering, reglas de asociación y predicción de demanda mediante modelos de series temporales.

1.5 Planificación del Trabajo

Para la planificación de este trabajo se sigue la planificación inicial realizada en la asignatura Contextualización y Diseño de Trabajo Final de Grado. A esta planificación se le hicieron modificaciones para incluir las PEC de la asignatura y modificaciones sustanciales en cuanto a las tareas ya que se han incluido tareas que no estaban planificadas inicialmente. También se ha creado una tabla de riesgos con las medidas a tomar ante ciertos riesgos. Los hitos entre PEC son las tareas cumplidas hasta la fecha de la PEC. Para realizar la planificación se ha utilizado el programa *GanttPro*. [2].

	Nombre de tarea	Fecha de inicio	Fecha final	Estimación	Duración	Estado
		23/09/2024	06/01/2025	300h	321.00	
1	▣ Trabajo Final de Grado	23/09/2024	06/01/2025	300h	321.00	
1.1	Obtención de datos	23/09/2024	27/09/2024	20h	20.00	● Terminado
1.2	PEC 1	30/09/2024	30/09/2024		10.00	● Terminado
1.3	Análisis Exploratorio y limpieza	30/09/2024	04/10/2024	20h	24.00	● Terminado
1.4	Ética y legalidad de los datos	07/10/2024	11/10/2024	20h	15.00	● Terminado
1.5	Validación de dataset para el trabajo final de grado	14/10/2024	14/10/2024		1.00	● Terminado
1.6	PEC 2	29/10/2024	29/10/2024		5.00	● Terminado
1.7	Procesamiento de Lenguaje Natural	29/10/2024	07/11/2024	0	24.00	● Terminado
1.8	Analítica de clientes	14/10/2024	21/11/2024	20h	24.00	● Terminado
1.9	Reglas de asociación	29/10/2024	03/12/2024	20h	24.00	● Terminado
1.10	PEC 3	03/12/2024	03/12/2024		5.00	● Terminado
1.11	Estudio de demanda y datos históricos	04/12/2024	09/12/2024	20h	14.00	● Terminado
1.12	Predicción demanda y series temporales	10/12/2024	20/12/2024	40h	30.00	● Terminado
1.13	Gráficas de visualización	23/12/2024	25/12/2024	25h	15.00	● Terminado
1.14	Panel interactivo	30/12/2024	02/01/2025	35h	20.00	● Terminado
1.15	Desarrollo memoria final y presentación	21/11/2024	06/01/2025	80h	90.00	● Terminado

Figura 1 Tareas Diagrama de Gantt

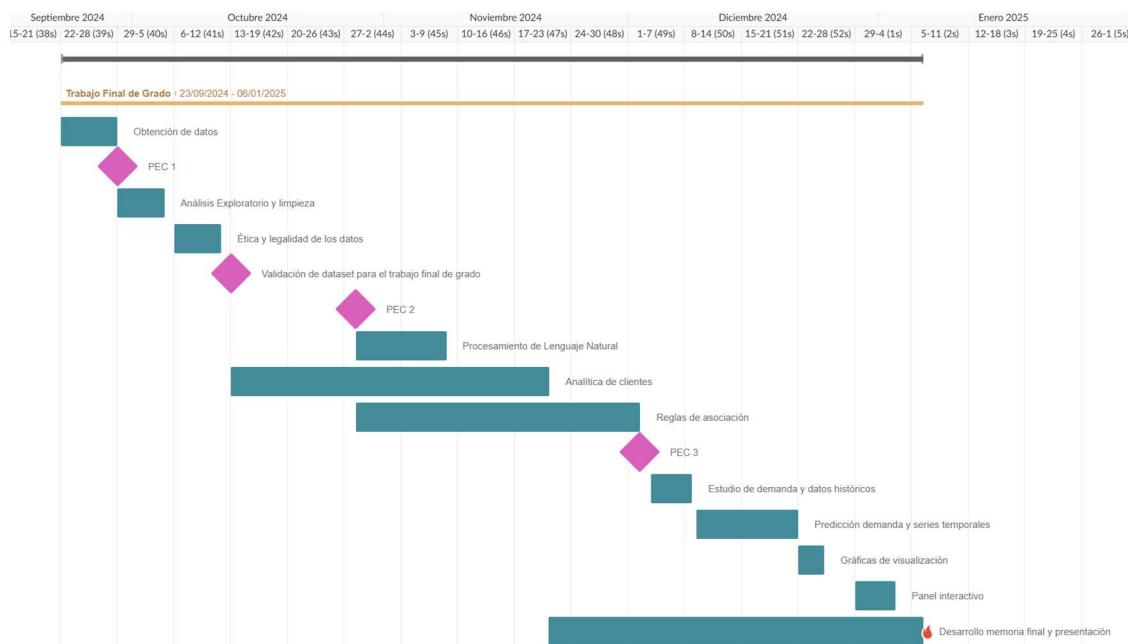


Figura 2 Diagrama de Gantt

Tabla de Gestión de Riesgos					
RIESGOS INTERNOS	MEDIDAS A TOMAR	RIESGOS EXTERNOS	MEDIDAS A TOMAR	RIESGOS TÉCNICOS	MEDIDAS A TOMAR
DATOS INSUFICIENTES	Consensuar con el tutor que medidas se pueden tomar para enfocar el trabajo en otra línea de investigación	MODIFICACIÓN SITUACIÓN LABORAL	Actualización de la planificación de las tareas.	FUNCIONAMIENTO INCORRECTO O ROTURA DE EQUIPOS	Copias de seguridad en discos externos o en la nube. Reparación o sustitución de los equipos o solicitud de préstamo de equipo
CUMPLIMIENTO NORMATIVO	Tratamiento de los datos con técnicas de privacidad de datos o solicitud de consentimiento	MODIFICACIÓN SITUACIÓN SALUD PERSONAL O FAMILIAR	Actualización de la planificación de las tareas y/o solicitud de aplazamiento.	TECNOLOGÍAS QUE NO FUNCIONAN	Actualización de software o sustitución del software. Aprendizaje de uso del nuevo software/técnica.

Tabla 1 Tabla de Gestión de riesgos

Para el trabajo se ha requerido de material informático como ordenador con tarjeta gráfica dedicada para la ejecución de los modelos. Cámara para la captura del vídeo, conexión a internet y recursos de aprendizaje.

1.6 Breve sumario de productos obtenidos

Se han obtenido los siguientes productos

- Conjunto de datos
- Modelo de procesamiento de lenguaje natural
- Modelo de segmentación de clientes
- Modelo de reglas de asociación
- Modelo de predicción de demanda
- Documento Jupyter Notebook link en el anexo.
- Memoria del trabajo
- Presentación del trabajo en slides
- Presentación del trabajo en vídeo

1.7 Breve descripción de los otros capítulos de la memoria

El contenido de los siguientes capítulos es el siguiente

- Capítulo 2 Materiales y métodos: Una descripción de la metodología utilizada
- Capítulo 3 Resultados: Descripción de las tareas y el resultado obtenido.
- Capítulo 4 Conclusiones y trabajos futuros: Análisis crítico del trabajo y posibilidades de desarrollo
- Capítulo 5 Bibliografía: Cita de los recursos utilizados para el trabajo
- Capítulo 6 Glosario: Listado y descripción de las palabras más utilizadas
- Capítulo 7 Anexos: Documentos adicionales

2. Materiales y métodos

2.1 Obtención de datos

Es necesario obtener un conjunto de datos útil que permita la consecución de los objetivos marcados. Los datos han de reunir las siguientes características:

- Datos de clientes
- Datos de productos
- Datos de establecimientos
- Datos de transacciones
- Datos de tiempo

Es fundamental que los datos estén bajo una licencia de uso que permita su manipulación y publicación y seguir escrupulosamente los requisitos de la licencia en cuanto a la referencia a los autores.

2.2 Análisis exploratorio y limpieza de datos

Para trabajar con los datos obtenidos se ha de seguir un proceso de análisis de los datos que permita conocer en qué formato se encuentran, adaptarlos para el uso y visualizar los datos de una manera gráfica para poder descubrir patrones y distribuciones en los datos. A este proceso se le llama Exploratory Data Analysis (EDA). Según Shirly “*El EDA Analysis o análisis exploratorio de datos es una técnica estadística que apunta a revelar estructuras subyacentes, identificar patrones o anomalías y cualquier indicio de relaciones clave que existan en un conjunto de datos o data set. El objetivo del EDA no es confirmar hipótesis, sino que se centra en generar preguntas y sus posibles direcciones para las investigaciones futuras. Para entenderlo mejor: el EDA en el Data Science es el arte de hacer preguntas más que el de buscar respuestas específicas.*” [3]

Durante el proceso de EDA se realiza la limpieza de datos que consiste en modificar, sustituir o eliminar datos para el uso posterior de los mismos. La limpieza de datos se ha de realizar tras analizar las consecuencias que implica cada técnica ya que se podría modificar la estructura general de los datos y obtener resultados erróneos.

2.3 Análisis de la ética y legalidad de los datos

Los datos han de cumplir la normativa vigente de protección de datos. El diccionario panhispánico jurídico define la protección de datos así: “*1. Adm. Conjunto de medidas para garantizar y proteger los datos de carácter personal (cualquier información concerniente a personas físicas identificadas o identificables) registrados en soporte físico, que los haga susceptibles de tratamiento, y a toda modalidad de uso posterior de estos datos por los sectores público y privado, a los efectos de garantizar y proteger las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar.*

Tales medidas se basan en los principios de calidad de los datos, el derecho de información en la recogida de datos, el consentimiento del afectado, los datos especialmente protegidos, los datos relativos a la salud, la seguridad de los datos, el deber de secreto, la limitación a la comunicación y el acceso a los datos por parte de terceros; así como en los derechos de las personas a la impugnación de valoraciones, a la consulta del Registro General de Protección de Datos, a la oposición, acceso, rectificación o cancelación de sus datos, a la tutela de tales derechos y a la indemnización. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE; Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales, de España. Ley 25326

Protección de datos personales, de Argentina, art.1. Constitución Nacional, de Argentina, art.43, párr. 3º.” [4]

Los datos que se protegen son de carácter personal, la web de la unión europea define los datos personales de la siguiente manera: “*Los datos personales son cualquier información relacionada con una persona identificada o identifiable, también denominada "el interesado". Ejemplos de datos personales:*

- *nombre y apellidos*
- *dirección*
- *número de documento de identidad/pasaporte*
- *ingresos*
- *perfil cultural*
- *dirección de protocolo internet (IP)*
- *datos en poder de hospitales o médicos (que identifican únicamente a una persona con fines sanitarios).*

Categorías especiales de datos

No se pueden tratar datos personales sobre:

- *origen racial o étnico*
- *orientación sexual*
- *opiniones políticas*
- *convicciones religiosas o filosóficas*
- *afiliación sindical*
- *datos genéticos, biométricos o sanitarios, salvo en casos específicos (por ejemplo, cuando se da un consentimiento explícito o cuando el tratamiento es necesario por razones de interés público esencial, sobre la base del Derecho nacional o de la UE)*
- *condenas e infracciones penales, a menos que lo autorice el Derecho nacional o de la UE” [5]*

Que los datos cumplan con los requerimientos legales no implica que se estén tratando éticamente. Un ejemplo puede ser realizar una encuesta y no recoger datos de personas debido a su nacionalidad, sexo, raza o ideología. Gartner define la ética de los datos como: “*un sistema de valores y principios morales relacionados con la recopilación, el uso y el intercambio responsable de datos. Las violaciones a la ética de los datos van desde abiertas y públicas hasta sutiles y secretas, como algoritmos que sugieren tasas de interés más altas para los solicitantes de hipotecas de minorías o líneas de crédito más bajas para las mujeres solicitantes de tarjetas de crédito*”.) [6]

Por tanto, no parece sencillo gestionar los datos de manera ética desde su creación o recopilación hasta cumplir el ciclo de vida de los datos, sin embargo, existen algunos recursos que aportan soporte en la gestión ética de los datos. La página del ministerio para la transformación digital [7] muestra varias formas en la que tratar los datos de manera ética y responsable y propone un recurso del Open Data Institute [8] que incluye un curso y una guía de buenas prácticas en la gestión ética de los datos.



Figura 3 Guía ODI [8]

2.4 Minería de textos Procesamiento de lenguaje Natural

La Minería de textos (Text Mining) es una parte de la Minería de datos en la que se analiza un texto generalmente sin estructura y se convierte en un dato o texto estructurado que nos aporta valor. DataScientest lo define de la siguiente manera *“El Text Mining, o análisis de textos, consiste en transformar un texto no estructurado en datos estructurados para proceder posteriormente al análisis. Esa práctica se basa en la tecnología de “Natural Language Processing” (procesamiento natural del lenguaje), que permite que las máquinas comprendan y traten el lenguaje humano de manera automática”.* [9]

Existen muchas definiciones sobre lo que es el procesamiento de lenguaje natural, a continuación, podemos ver como lo definen IBM y AWS. Para IBM: *“El procesamiento del lenguaje natural (PLN) es un subcampo de la informática y la inteligencia artificial (IA) que utiliza el machine learning para permitir que los ordenadores entiendan y se comuniquen con el lenguaje humano. El PLN permite a los ordenadores y dispositivos digitales reconocer, comprender y generar texto y voz combinando la lingüística computacional (el modelado del lenguaje humano basado en reglas) junto con el modelado estadístico, el machine learning y el deep learning.”* [10]

Según Amazon Web Services (AWS): “El procesamiento de lenguaje natural (NLP) es una tecnología de machine learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano.” [11]

El PLN está muy de moda en la actualidad ya que es una de las tecnologías que utilizan los chats de inteligencia artificial como ChatGPT, CoPilot, Gemini entre otros. El PLN permite realizar diferentes tareas en función del campo de uso, SAS destaca entre ellas las siguientes tareas: [12] -

- *Categorización de contenido. Un resumen del documento basado en la lingüística, incluyendo búsqueda e indexación, alertas de contenido y detección de duplicación.*
- *Clasificación. La clasificación basada en BERT se utiliza para captar el contexto y el significado de las palabras de un texto con el fin de mejorar la precisión en comparación con los modelos tradicionales.*
- *Análisis de corpus. Comprender el corpus y la estructura de los documentos mediante estadísticas de salida para tareas como el muestreo eficaz, la preparación de datos como entrada para modelos posteriores y la elaboración de estrategias de modelización.*
- *Extracción contextual. Extraiga automáticamente información estructurada de fuentes basadas en texto.*
- *Análisis de sentimiento. Identificación del estado de ánimo u opiniones subjetivas en grandes cantidades de texto, incluyendo minería de sentimiento y opiniones promedio.*
- *Conversión de voz a texto y de texto a voz. Transformación de órdenes vocales en texto escrito y viceversa.*
- *Resumen de documentos. Generación automática de sinopsis de grandes cuerpos de texto y detección de lenguas representadas en corpus multilingües (documentos).*
- *Traducción basada en máquina. Traducción automática de texto o habla de un idioma a otro..-*

El PLN se puede realizar de manera supervisada donde se entrena con un conjunto de datos etiquetado y el modelo aprende a realizar la clasificación. El PLN no supervisado se alimenta de entradas no etiquetadas y produce una salida mediante modelos

estadísticos como por ejemplo la función autocompletar al escribir un mensaje. El proceso para llevar a cabo el PLN es el siguiente:

- **Preprocesamiento del texto**
- **Representación del texto**
- **Modelo de aprendizaje**
- **Evaluación**

Para el preprocesamiento de textos es necesario realizar una tokenización la cual consiste en dividir el texto en partes pequeñas (tokens) como pueden ser palabras o signos de puntuación. Posteriormente se normaliza el texto a minúsculas y se eliminan acentos o caracteres especiales según el idioma y se realiza la lematización (stemming) que consiste en capturar el lexema de la palabra. El último paso consiste en la eliminación de stopwords, las cuales son palabras frecuentes según el idioma, como artículos y preposiciones. La representación del texto consiste en transformar el texto preprocesado en forma vectorial para que las computadoras puedan procesarlo. En función de la tarea a realizar se procesará el texto en diferentes tipos de vectores:

- **Bag of words (BoW)**: Los vectores cuentan la frecuencia de las palabras
- **Term Frequency – Inverse Document Frequency (TF-IDF)**: Ponderación de la importancia de las palabras en el texto y en el corpus.
- **Word Embeddings**: Vectores semánticos que recogen relaciones de palabras como macho – gato, hembra -gata.
- **Contextual Embeddings**: Vectores dinámicos según el contexto de la palabra en una frase como ganar una copa (trofeo) y tomar una copa de vino (recipiente).

Los modelos de aprendizaje se pueden ejecutar mediante algoritmos como la

regresión logística, árboles de decisión SVM o mediante redes neuronales donde destacan RNN, LTSM, CNN GRU como modelos para procesar el texto de manera ordenada, CNN que procesa secuencias de texto para clasificar oraciones y Transformers que son modelos para traducción, clasificación y resumen. Estos modelos de aprendizaje consumen muchos recursos computacionales.

2.5 Analítica de Clientes

Para cualquier empresa resulta fundamental recopilar datos de los clientes ya sean datos demográficos, transacciones, datos de consumo, opiniones o valoraciones que haya realizado. Las empresas, cumpliendo la normativa vigente, pueden decidir qué datos recopilan y de qué manera. Para ello es necesario un estudio de los datos que se pueden recoger del cliente, los medios por los cuales se van a recoger, el formato y el tratamiento que se va a dar a los datos. Según los datos, las empresas pueden establecer estrategias de marketing, predicción de demanda, actualización de servicios o productos. La analítica de clientes permite a las empresas tomar decisiones informadas y tratar de diferenciarse de la competencia. Existen diferentes herramientas para recopilar información de los clientes, IBM [13] sugiere los siguientes:

- *Cookies*
- *Paneles de control de CRM*
- *Correo electrónico*
- *Redes sociales*
- *Encuestas*
- *Sitios web*

Una vez recopilados los datos se deben de tratar legal y éticamente previo a que se puedan aplicar técnicas de aprendizaje automático (machine learning) para obtener información sobre los comportamientos de compra de diferentes grupos de clientes. [14]

El término aprendizaje automático o machine learning procede de Arthur Samuel (1901 – 1990) quien durante su trabajo en IBM dedicó su tiempo a desarrollar un programa Samuel Checkers el cual, era un juego de damas que trataba de ganar a un humano. Debido a los problemas de memoria, Samuel introdujo un sistema de puntuación que permitía calcular las probabilidades de ganar en ciertas posiciones. Cada vez que se jugaba una nueva partida, el sistema mejoraba y aprendía nuevas probabilidades de victoria. [15]. El machine learning puede ser ejecutado de 3 maneras diferentes [16]:

- Aprendizaje supervisado: Los datos se dividen en conjunto de entrenamiento y prueba y se evalúa el rendimiento.
- Aprendizaje no supervisado: Los datos se usan al completo siendo el algoritmo quien detecta diferentes grupos o patrones que no se aprecian.
- Aprendizaje semi supervisado: se utilizan datos de entrenamiento con etiquetas y sin etiquetas

También existe el aprendizaje por refuerzo, el cual AWS lo califica como una técnica de machine learning [17] “*El aprendizaje por refuerzo (RL) es una técnica de machine learning (ML) que entrena al software para que tome decisiones y logre los mejores resultados. Imita el proceso de aprendizaje por ensayo y error que los humanos utilizan para lograr sus objetivos. Las acciones de software que trabajan para alcanzar su objetivo se refuerzan, mientras que las que se apartan del objetivo se ignoran*”. En la analítica de clientes se puede obtener información sobre el comportamiento de compra de los clientes, agruparlos por consumos similares, analizar que supone el cambio de precio en un producto para los clientes, analizar el ciclo de vida del cliente, es decir, el

gasto total que se espera de un cliente en la empresa. Se va a desarrollar en este trabajo la agrupación de clientes por hábito de compra para obtener perfiles de clientes que permita a los departamentos de marketing, compras y ventas tomar medidas de captación, y fidelización mediante campañas de marketing, ofertas, o disponibilidad de producto. Se utiliza el modelo no supervisado Kmeans. Su funcionamiento según explica scikit-learn [18] “ *El algoritmo **KMeans** agrupa los datos intentando separar las muestras en n grupos de varianza igual, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo (ver más abajo). Este algoritmo requiere que se especifique el número de grupos. Se adapta bien a grandes cantidades de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.*

El algoritmo k-means divide un conjunto de N muestras X en K cúmulos disjuntos C, cada uno descrito por la media μ_j de las muestras del conglomerado. Las medias se denominan comúnmente “centroídes” del conglomerado; tenga en cuenta que, en general no son puntos de X, aunque vivan en el mismo espacio.

*El algoritmo K-means tiene como objetivo elegir centroídes que minimicen la **inercia** o el **criterio de suma de cuadrados dentro del grupo del grupo**”*

$$\sum_{i=0}^n \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right)$$

2.6 Reglas de Asociación

Para un negocio conocer los datos de ventas, compras, gastos, los tipos de cliente que tiene, sus hábitos de compra, es muy importante como ya se ha mencionado, pero una empresa ha de decidir como ofrece sus productos, es decir como pone a disposición del cliente los productos o servicios. Para ello, las reglas de asociación analizan las transacciones de los clientes y permiten hallar combinaciones de productos o servicios en función de la probabilidad de que aparezcan juntos. En un supermercado, tendría

sentido para un cliente por ejemplo que la pasta y la salsa estuvieran cerca ya que, si cocina pasta, es probable que le ponga alguna salsa, o en el caso de querer hacer un pastel, que los huevos y la harina estuvieran cerca. Las reglas de asociación permiten encontrar concurrencias frecuentes partiendo de un antecedente y un consecuente.

Lidgi González en un artículo de aprendeia lo explica así: [19]

“Un antecedente es un elemento que se encuentra dentro de los datos. Un consecuente es un elemento que se encuentra en combinación con el antecedente. Las reglas de asociación se crean buscando en los datos patrones frecuentes de “if-then” y utilizando los criterios de apoyo y confianza para identificar las relaciones más importantes. El apoyo es una indicación de la frecuencia con que los elementos aparecen en los datos. La confianza indica el número de veces que las afirmaciones del tipo “if-then” se consideran verdaderas. Se puede utilizar una tercera métrica, llamada fit para comparar la confianza con la confianza esperada, o cuántas veces se espera que una afirmación del tipo “if-then” se considera cierta.”. Las reglas de asociación se pueden obtener mediante diferentes algoritmos los cuales usan diferentes técnicas:

- **Algoritmo A priori:** Itera conjuntos. **Estructura horizontal.** Los subconjuntos de un conjunto frecuente han de ser frecuentes.
- **Algoritmo Eclat:** Intersección de listas. **Estructura vertical.** Se considera frecuente en función del tamaño de la intersección.
- **Algoritmo FP-Growth:** **Árbol de patrones.** Comprime los datos y extrae patrones frecuentes del árbol creado.

La selección del algoritmo depende de la estructura de los datos tanto en cantidad como en profundidad es decir se seleccionará el algoritmo según la cantidad de registros de transacciones o cantidad de productos diferentes. También deben de tenerse en cuenta los recursos computacionales disponibles.

2.7 Análisis y predicción de demanda

Para el realizar la predicción de demanda se han de utilizar las series temporales. Una serie temporal es un conjunto de datos medido en intervalo de tiempo regular y ordenado. Las observaciones pasadas tienen influencia en las observaciones posteriores y futuras. Las series temporales permiten analizar tendencias y estacionalidades en datos a lo largo del tiempo. Se utilizan por ejemplo para analizar ventas, producción, valores de acciones o producto (oro, petróleo), análisis meteorológicos o consumo de electricidad. Existen muchos tipos de modelos de análisis temporales como Prophet, Arima, Sarima o redes neuronales. [20]. Para analizar las series temporales una de las técnicas es el uso de redes neuronales que son muy efectivas. Sin embargo. Las redes neurales pueden tener problemas con datos estacionarios y su ejecución implica muchos recursos de computación, así como tiempo para su calibración. Existen modelos como Prophet que captan muy bien la estacionalidad y es robusto a la falta de datos El modelo de predicción de series temporales Prophet es descrito en su página web así [21] “*Prophet es un procedimiento para pronosticar datos de series temporales basado en un modelo aditivo en el que las tendencias no lineales se ajustan a la estacionalidad anual, semanal y diaria, además de los efectos de los días festivos. Funciona mejor con series temporales que tienen fuertes efectos estacionales y varias temporadas de datos históricos. Prophet es robusto ante datos faltantes y cambios en la tendencia, y normalmente maneja bien los valores atípicos.*” Otra manera de analizar las series temporales es mediante modelos ARIMA (Autoregressive Moving Average), o SARIMA (Seasonal Arima) para datos estacionales. Para Pablo Herrera [22] “Los modelos ARIMA funcionan bien cuando se disponen de series temporales largas (más de 40 puntos al menos) y el patrón de comportamiento es estable o consistente durante el tiempo. Requieren que la serie sea estacionaria, lo

cual quiere decir que no deben tener una tendencia ni tampoco una variabilidad entre picos elevada". Define también los 3 términos del modelo ARIMA."

- **AR:** término autoregresivo. Establece la relación entre un valor determinado y otro anterior (laggado).
- **I:** integrado. Se refiere a la capacidad de diferenciar la serie para eliminar la tendencia y la variabilidad creciente o decreciente.
- **MA:** término *moving average*. Representa el error del modelo como combinación de términos de error anteriores."

Por tanto, el método a seguir para las series temporales depende del tipo y de la calidad de los datos que se vayan a analizar, del coste computacional y de recursos que implica la ejecución del modelo y los resultados parciales que se obtengan. Para el trabajo se ha creado modelos de Prophet y de SARIMA obteniendo mejores resultados en Prophet.

3. Resultados obtenidos

3.1 Obtención de datos

Para la ejecución del trabajo final de grado es necesario disponer de datos de clientes, productos, transacciones, tipo de establecimiento donde se venden y además han de tener progresión en el tiempo, es decir han de ser datos continuados en un periodo de tiempo suficiente para poder conseguir los objetivos planteados. El primer paso para la obtención de datos ha sido contactar con empresas del sector retail informando de la finalidad de este trabajo final de grado y solicitando al departamento de protección de datos un conjunto de datos que cumpliera las características necesarias. Se ha enviado correos en idioma español e inglés. Se ha contactado con las siguientes empresas:

Solicitud de datos a empresas en España		
EMPRESA	CONTACTO	RESPUESTA
MERCADONA	dpo@mercadona.es	SIN RESPUESTA
CARREFOUR	derechosprotecciondedatos@carrefour.com	RECHAZADA
ALDI	protecciondatos@aldi.es	SIN RESPUESTA
EROSKI	dpo@eroski.es	SIN RESPUESTA
DIA	dpo.es@diagroup.com	SIN RESPUESTA
ALCAMPO	dpoare@alcampo.es	SIN RESPUESTA
BON PREU	seguretatlopd@bonpreu.cat	RECHAZADA
ALIMERKA	rgpd@alimerka.es	SIN RESPUESTA
MUELLER	protecciondedatos@muller.es	SIN RESPUESTA
CONSUM	delegadoprotecciondatosconsum@consum.es	RECHAZADA
TRANSGOURMET	rgpd@transgourmet.es	SIN RESPUESTA
MAKRO	dpo@makro.es	SIN RESPUESTA
AHORRAMAS	protecciondedatos@ahorramas.com	RECHAZADA
HIPERCENTRO	rgpd@hipercentro.net	SIN RESPUESTA
EL CORTE INGLES	delegado.protecciondatos@elcorteingles.es	SIN RESPUESTA
PCCOMPONENTES	lop@pcccomponentes.com	SIN RESPUESTA
MEDIAMARKT	dpo@mediamarkt.es	RECHAZADA
AMAZON	eu-privacy@amazon.es	SIN RESPUESTA
MOVISTAR	DPO_Movistar@telefonica.com	RECHAZADA
ORANGE	orangeproteccion.datos@es.orange.com	RECHAZADA
MASMOVIL	dpo@masmovil.com	SIN RESPUESTA
DAZN	datasubjectsrights@dazn.com	SIN RESPUESTA
GLOVO	gdpr@glovoapp.com	SIN RESPUESTA

Tabla 2 Solicitud de datos a empresas en España

Solicitud de datos a empresas en Europa y Reino Unido		
EMPRESA	CONTACTO	RESPUESTA
JUST EAT	dpo@citypantry.com	RECHAZADA
MORRISONS	dataprotection@morrisonsplc.co.uk	RECHAZADA
ASDA	dataprotection@asda.co.uk	SIN RESPUESTA
SAINSBURY	privacy@sainsburys.co.uk	RECHAZADA
NTT DATA	data.protection.office@nttdata.com	SIN RESPUESTA

Tabla 3 Solicitud de datos a empresas en Europa y Reino Unido

Solicitud de datos a empresas en Estados Unidos de America

EMPRESA	CONTACTO	RESPUESTA
KROGER	KrogerPrivacyOffice@kroger.com	SIN RESPUESTA
NORDSTROM	privacy@nordstrom.com	SIN RESPUESTA
TESCO	dpo@tesco.com	SIN RESPUESTA
WALMART	consumerprivacy@wal-mart.com	RECHAZADA
COSTCO	privacy@costco.com	SIN RESPUESTA

Tabla 4 Solicitud de datos a empresas en Estados Unidos de America

Se obtienen de la solicitud de datos a empresas los siguientes resultados:

- Se han realizado 23 solicitudes de datos en España obteniendo 7 respuestas rechazando la solicitud y 16 solicitudes sin respuesta.
- Se han realizado 5 solicitudes a empresas en Europa y Reino Unido obteniendo 3 respuestas rechazando la solicitud y 2 solicitudes sin respuesta.
- Se han realizado 5 solicitudes de datos en América obteniendo 1 respuesta rechazando la solicitud y 4 solicitudes sin respuesta.

Solicitudes de datos a empresas

Figura 4: Gráfico de solicitudes de datos a empresas.

En resumen, se han realizado 33 solicitudes de datos obteniendo 11 respuestas rechazando la solicitud lo que implica el 33,3% de las solicitudes y no obteniendo respuesta en 22 que implica el 66,6% de las solicitudes. Las consultas han sido realizadas directamente al correo del departamento de protección de datos de cada empresa, por lo que puede existir un filtro de correo spam que puede explicar el alto porcentaje de no respuesta obtenido. Analizando las respuestas obtenidas, todas han sido rechazadas, lo cual era un resultado esperado ya que las empresas ponen en valor los datos internos y son reacias a compartir datos debido a la información que se puede extraer de ellos y que son objetivo de este trabajo. Además, se ha de tener en cuenta la protección de datos de los diferentes estados y las cláusulas que han aceptado los clientes en materia de protección de datos. Al no poder obtener los datos de empresas se ha realizado una búsqueda de datasets publicados seleccionando el dataset que mejor se adapta al propósito del trabajo. Se ha obtenido un dataset sintético de un repositorio de Github [23] de los datos de Maven una empresa ficticia del sector retail que tiene presencia en diferentes países. Este dataset es muy utilizado para crear visualizaciones y para su uso mediante código SQL. Se encuentra en formato CSV. El dataset en realidad es un conjunto de datasets que se pasan a explicar brevemente:

1. df_calendar: Dataset con las fechas de compra
2. df_customers: Dataset con los datos de clientes vinculados a tarjeta de fidelización.
3. df_products: Dataset con los datos descriptivos de los productos
4. df_regions: Dataset con los datos de distritos y regiones
5. df_returns: Dataset con los datos de las devoluciones realizadas por los clientes
6. df_stores: Dataset con los datos descriptivos de las tiendas
7. df_trans97: Dataset con las transacciones realizadas en el año 97
8. df_trans98: Dataset con las transacciones realizadas en el año 98

Una vez descritos individualmente se realizan técnicas de EDA y se empieza a trabajar en la unión de los diferentes datasets para obtener insights relevantes.

3.2 Análisis Exploratorio y limpieza de datos

El trabajo se realiza en el programa Python mediante JupyterNotebook. Python es uno de los lenguajes de programación que actualmente son más usados para la ciencia de datos ya que permite mediante la carga de librerías realizar cálculos matemáticos y análisis de datos. JupyterNotebook es una aplicación web de código abierto que permite crear y compartir código en diferentes lenguajes de programación. Databricks define JupyterNotebook así [24] “*Un Jupyter Notebook es una aplicación web de código abierto que permite a los científicos de datos crear y compartir documentos que incluyen código en vivo, ecuaciones y otros recursos multimedia. ¿Para qué se utilizan los Jupyter Notebooks? Los cuadernos Jupyter se utilizan para todo tipo de tareas de ciencia de datos, como análisis exploratorio de datos (EDA), limpieza y transformación de datos, visualización de datos, modelado estadístico, aprendizaje automático y aprendizaje profundo*

. Una vez cargados los datos en el JupyterNotebook se analizan los datasets de manera individual antes de la creación de datasets específicos para el desarrollo del trabajo. Se desglosan los resultados del análisis de los dataset y las medidas adoptadas.

Nombre dataset	Número de filas			Número de columnas
df_calendar	730			1
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
date	object	730	0	Fecha transacción
Acciones llevadas a cabo				
Se convierte la columna date de tipo object a tipo fecha date.time				

Tabla 5 df_calendar

Nombre dataset	Número de filas			Número de columnas
df_customers	10281			20
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
Customer_id	Int64	10281	0	Número id del cliente
customer_acct_num	Int64	10281	0	Número de cuenta del cliente
first_name	object	2300	0	Nombre
last_name	object	5574	1	Apellido
customer_address	object	10264	0	Dirección
customer_city	object	108	0	Ciudad
customer_state_province	object	13	0	Provincia
customer_postal_code	Int64	9841	0	Código postal
customer_country	object	3	0	País
birthdate	object	8149	0	Fecha de nacimiento
marital_status	object	2	0	Estado civil M = Married, S= Single
yearly_income	object	8	0	Rango ingreso anual, miles dólares
gender	object	2	0	Género M=Male, F=Female
total_children	Int64	6	0	Número de hijos de 0 a 5
num_children_at_home	Int64	6	0	Número de hijos en casa de 0 a 5
education	object	5	0	Nivel de educación
acct_open_date	object	1648	0	Fecha de inicio cuenta
member_card	object	4	0	Normal, Bronze, Silver, Golden
occupation	object	5	0	Tipo de ocupación
homeowner	object	2	0	Propietario vivienda Y=Yes, N=No
Acciones llevadas a cabo				
Se sustituye el valor nulo de la columna last_name por Unknown para no perder el registro. No se quiere perder el registro ya que el nombre y apellido no es relevante para el estudio y se anonimizarán los datos. Se convierten las columnas birthdate y acct_open_date a formato datetime				

Tabla 6 df_customers

Nombre dataset	Número de filas			Número de columnas
df_products	1560			9
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
product_id	Int64	1560	0	Número id del producto
product_brand	object	111	0	Marca del producto
product_name	object	1560	0	Nombre del producto
product_sku	Int64	1560	0	SKU del producto
product_retail_price	float64	315	0	Precio de venta al público
product_cost	float64	173	0	Coste del producto
product_weight	float64	376	0	Peso del producto
recyclable	float64	1	687	Indica si el producto es recicitable
low_fat	float64	1	1008	Indica si producto es bajo en grasas
Acciones llevadas a cabo				
Existen valores nulos en recyclable y en low_fat sin embargo solo existe un valor para cada uno de ellos. Se trata de una variable categórica que toma como valores 1.0 o Nan. Se transforman los datos para que 1.0 sea 1 y nan sea 0. El valor 1 indica Si y el valor 0 indica No.				

Tabla 7 df_products

Nombre dataset	Número de filas			Número de columnas
df_regions	109			3
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
region_id	Int64	109	0	Número id de la región
sales_district	object	22	0	Distrito de venta
sales_region	object	7	0	Región de venta
Acciones llevadas a cabo				
Ninguna				

Tabla 8 df_regions

Nombre dataset	Número de filas			Número de columnas
df_returns	7087			4
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
return_date	object	639	0	Fecha de la devolución
product_id	Int64	1539	0	Número id del producto
store_id	Int64	20	0	Número id de la tienda
quantity	Int64	2	0	Cantidad de unidades
Acciones llevadas a cabo				
Se convierte la columna return_date a formato datetime				

Tabla 9 df_returns

Nombre dataset	Número de filas			Número de columnas
df_stores	24			13
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
store_id	int64	24	0	Número id de la tienda
region_id	int64	23	0	Número id de la región
store_type	object	5	0	Tipo de tienda
store_name	object	24	1	Nombre de la tienda
store_street_address	object	24	0	Dirección de la tienda
store_city	object	23	0	Ciudad
store_state	object	10	0	Estado
store_country	object	3	0	País
store_phone	object	24	0	Número de teléfono de la tienda
first_opened_date	object	24	0	Fecha de apertura
last_remodel_date	object	24	0	Fecha de última remodelación
total_sqft	int64	24	0	Metros cuadrados totales
grocery_sqft	int64	24	0	Metros cuadrados de sala de ventas
Acciones llevadas a cabo				
Se convierten first_opened_date y last_remodel_date a formato datetime				

Tabla 10 df_stores

Nombre dataset	Número de filas			Número de columnas
df_trans97	86837			6
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
transaction_date	object	323	0	Fecha de la transacción
stock_date	object	370	0	Fecha de entrada de stock
product_id	int64	1559	0	Nombre del producto
customer_id	int64	5581	0	Número id del cliente
store_id	int64	13	0	Número id de la tienda
quantity	int64	6	0	Cantidad de unidades
Acciones llevadas a cabo				
Se convierten las columnas transaction_date y stock_date a formato datetime				

Tabla 11 df_trans97

Nombre dataset	Número de filas			Número de columnas
df_trans98	182883			6
Nombre columna	Tipo de dato	Valores únicos	Valores nulos	Descripción
transaction_date	object	350	0	Fecha de la transacción
stock_date	object	370	0	Fecha de entrada de stock
product_id	int64	1559	0	Nombre del producto
customer_id	int64	8060	0	Número id del cliente
store_id	int64	24	0	Número id de la tienda
quantity	int64	6	0	Cantidad de unidades
Acciones llevadas a cabo				
Se convierten las columnas transaction_date y stock_date a formato datetime				

Tabla 12 df_trans98

Una vez que se han analizado los datos de cada dataset y se ha verificado que no existen nulos, se crea un único dataset mediante merge, que unirá los datasets de transacciones, el dataset de productos, el dataset de clientes, el dataset de tiendas, el dataset de calendario, el dataset de devoluciones y el dataset de regiones.

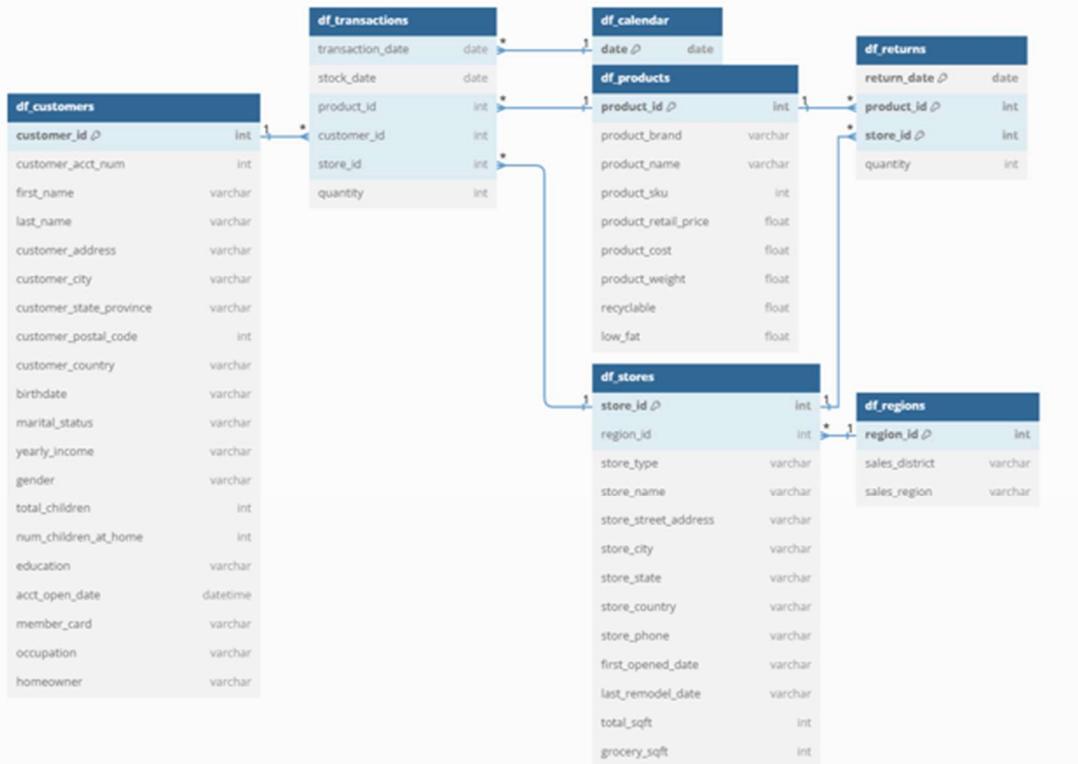


Figura 5 Diagrama Relación datasets

De este dataset unificado se irán seleccionando columnas de interés para trabajar tanto la analítica de clientes como la predicción de demanda. En este dataset unificado, se crean las columnas año, mes semana y día ya que servirá para analizar los datos por esas medidas de tiempo. En el dataset unificado se clasifican los ingresos de los clientes en tres categorías: Low Income, Mid Income y Hight Income.

3.3 Análisis de la ética y legalidad de los datos

Al trabajar con los datasets, aunque sean datos sintéticos, se decide aplicar las prácticas para cumplir con la normativa europea de protección de datos [25] la cual está implementada en España mediante la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD) [26]. Es una tarea que se realiza para poner en práctica los conocimientos adquiridos durante el grado y que se debe de

realizar en el ámbito laboral. Se ha realizado supresión de datos personales que pudieran identificar a los clientes como nombre y dirección. Otra técnica que se podría haber utilizado es la anonimización pero en este trabajo no se va a tratar con datos individuales de clientes por lo que no son necesarios más allá del identificador de cliente. Los datos se han tratado de manera ética siguiendo pasos de la guía de ODI. Se ha verificado que no existe discriminación por preferencia sexual, etnia o raza, así como, que no existe discriminación alguna en cuanto a su recogida ya que son sintéticos.

3.4 Minería de textos Procesado de lenguaje Natural

Como paso previo a la analítica de clientes, se realiza la obtención de las categorías de los productos analizando la descripción del producto mediante procesamiento de lenguaje natural PLN. Con esta acción se obtendrán las categorías a las que pertenece cada producto por ejemplo carnes, pescados, frutas, bebidas... con el fin de poder analizar que tipos de productos compran los clientes por familias o categorías. En este punto se ha realizado una investigación de las categorías que pueden representar a productos para el caso de cadena de tiendas del sector retail de alimentación [27] y de estas categorías una vez analizado el dataset se seleccionan las que más describen a los productos. Se describen los pasos seguidos para el procesamiento de lenguaje natural. En el preprocesamiento del texto se realiza la limpieza y normalización de texto que corresponde a la columna product_name. A continuación se realiza una clasificación por palabras clave creando diccionarios para la asignación de categorías mediante reglas (rule-based). Para el procesado del texto y su vectorización se utiliza el modelo en_core_web_trf de spacy basado en Transformers que tokeniza las descripciones por nombre, pronombre y adjetivo y se obtienen las palabras clave. Se realiza la clasificación mediante el modelo zero-shot classification de Hugging Face el cual permite la clasificación sin entrenar el modelo (zero-shot). El modelo recibe las palabras

tokenizadas y las categorías seleccionadas y asigna una categoría a cada palabra. Finalmente se analizan los resultados y se vuelve a procesar la categoría de artículos congelados con el modelo ya que dentro de esta categoría se pueden volver a clasificar los productos en otras categorías como verduras congeladas, carne congelada, frutas congeladas. En resumen, se han realizado los siguientes pasos:

- Obtención de categorías de Openfoodfacts mediante request.
- **Preprocesamiento de datos:** Limpieza y normalización.
- **Análisis lingüístico** con spaCy donde se obtienen los tokens y su etiqueta POS.
- **Clasificación semántica** con modelo Zero-Shot de Hugging face mediante BART (Bidirectional and Auto-Regresive Transformer) el cual clasifica el texto en las diferentes categorías seleccionadas.
- **Clasificación manual** o ajuste fino, mediante palabras clave que el modelo no ha sabido clasificar correctamente.

Se han aplicado las técnicas a la columna product_name pasando por cada nombre y categorizando cada uno de ellos. Se ha utilizado los recursos de la tarjeta gráfica para optimizar el tiempo de clasificación. Con todo ello se obtienen las categorías o familias de los productos que previamente no se disponían. Se crea una columna llamada category. Se han seleccionado 45 categorías de productos se puede observar en la figura 5 un sumatorio de la cantidad de productos que incluye cada categoría. Se observa que las 6 categorías con más productos son Non Food (artículos no comestibles) 157, Vegetables (verduras) 110, Snacks 105, Dairy (productos lácteos) 90, Meat (productos cárnicos) 90 y Fruits (frutas) 90.. Las categorías con menos artículos son Honeys (mieles) 5, Baby Goods (productos de bebé) 5, Fish (pescados). En cuanto a su distribución Figura 6, se obtiene que la categoría Non Food representa el 10.06%, Vegetables 7.05%, Snacks 6.73%, Dairy Meat y Fruits 5.77% cada una.

category	Coffees & Teas	20	
Non Food	157	Oils	20
Vegetables	110	Candies	20
Snacks	105	Frozen Pizzas	20
Dairy	90	Salads	20
Meat	90	Pastas	18
Fruits	90	Frozen Breakfast	15
Personal Care	85	Jellies	15
Canned Goods	75	Sauces & Condiments	15
Pastries	55	Preserves	15
Drinks	50	Frozen Meat	15
Spreads	45	Frozen Fries	15
Soups	45	Vegan Goods	10
Cleaning Goods	38	Sugar	10
Ice Cream & Desserts	35	Beer	10
Cereals	31	Tomatoes	10
Wines	30	Rices	6
Nuts & Seeds	25	Honeys	5
Juices	25	Baby Goods	5
Magazine	25	Frozen Fondue	5
Frozen Vegetables	25	Fish	5
Breads	25	Salt	5
Eggs	20	Spices & Seasonings	5

Figura 6 Sumatorio de productos por categoría

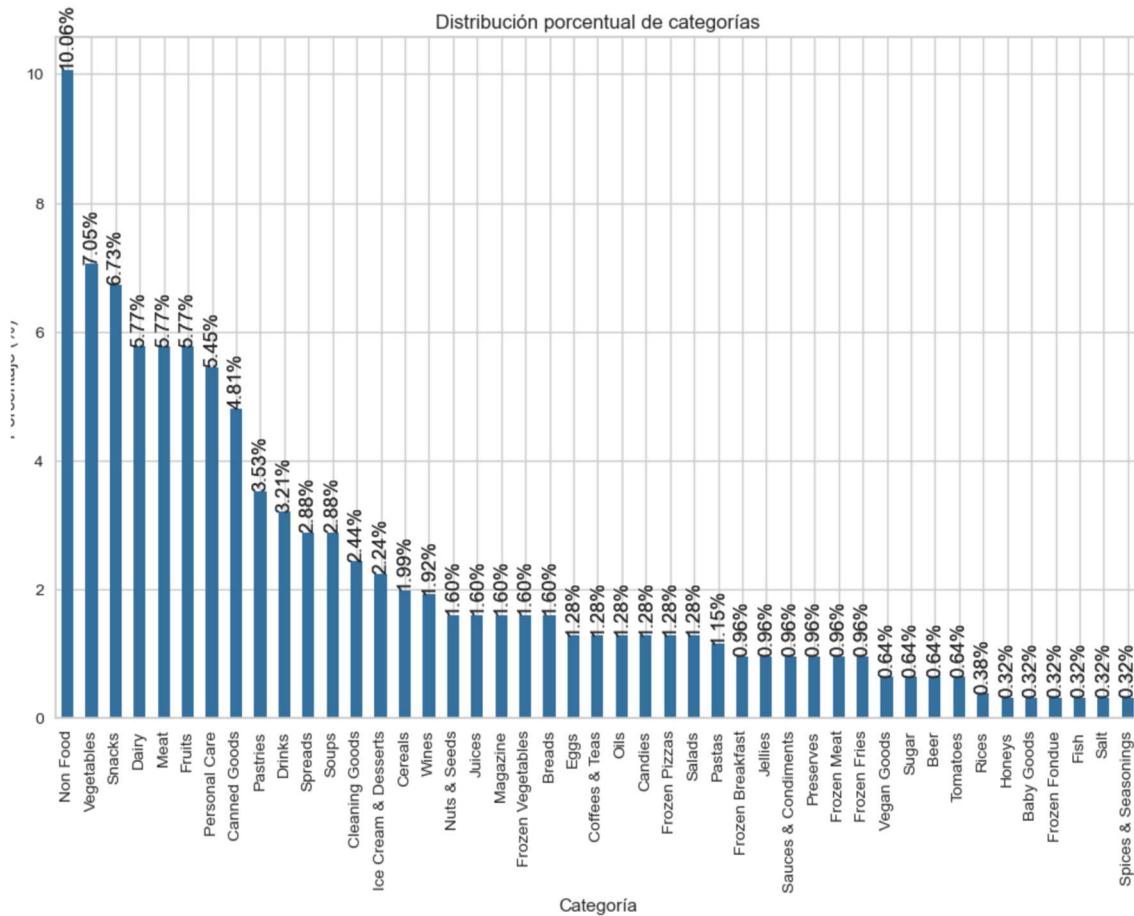


Figura 7 Distribución de porcentaje de categorías

3.5 Analítica de clientes

La analítica de clientes puede ser aplicada para diferentes usos como segmentación, sistemas de recomendación, valor del ciclo de vida del cliente, análisis de sentimientos, detección de fraude. La analítica de clientes realizada está basada en la segmentación para categorizar a los clientes en grupos en base a la edad, generación a la que pertenecen y sus hábitos de compra. Se realiza una categorización de ciertas variables descriptivas de los clientes para poder realizar posteriormente una segmentación.

3.5.1 Análisis de clientes por generación, ingresos y educación.

Se obtiene el dato edad para cada cliente ya no se dispone. Se calcula mediante las columnas fecha de nacimiento y fecha de transacción. La edad se obtiene realizando la resta de la fecha de la última transacción y la fecha de nacimiento ya que las acciones posteriores de los departamentos de marketing o CRM se basarán en la edad en el momento de la campaña. Se ha clasificado las edades de los clientes en las categorías:

- Menor: menor de 18 años en España
- Joven: Entre 18 y 30 años
- adulto joven: Entre 31 y 44 años
- adulto medio: Entre 45 y 60 años
- senior: Más de 60 años

Se ha decidido realizar una clasificación de la generación a la que pertenecen los clientes ya que diferentes generaciones pueden tener hábitos de compra o características diferentes a los grupos de edad. Se ha clasificado la edad en base a la generación a la que pertenecen para poder determinar las campañas en base a la generación siendo las generaciones obtenidas las siguientes:

- Millenials nacidos entre 1981 y 1996
- Generación X: nacidos entre 1965 y 1980
- Baby Boomers: nacidos entre 1946 y 1964
- Silent Generation: nacidos antes de 1946

A continuación, se presentan las distribuciones de los clientes por cada categoría:

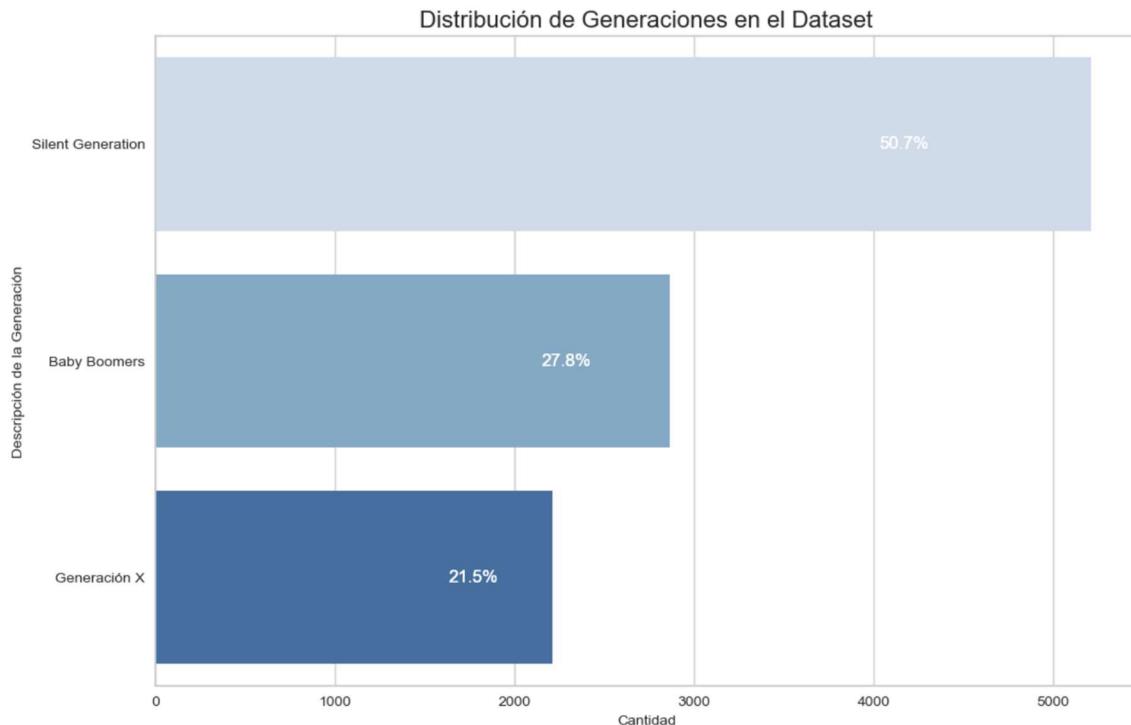


Figura 8 Distribución de generaciones

Si se analiza la distribución de los clientes por generación se obtiene que la mayoría de pertenecen a la Silent Generation 5208 clientes que representan el 50.65%, el segundo grupo baby boomers 2863 clientes 27.85% y el tercero la generación X 2210 clientes 21.5%. No se obtienen usuarios Millenials ya que la fecha de nacimiento del cliente mas joven es 1980. El código utilizado se deja con la categoría de Millenial para futuras actualizaciones de datos. Se analizan los datos de valores nulos en edad, estos indican que no se ha obtenido ninguna edad ya que no ha existido ninguna transacción durante los años 1997 y 1998. Se obtienen 1439 clientes un 13.99 por ciento del total de clientes que disponen de la tarjeta de fidelización, pero no han realizado compras.

Clientes que no han realizado compras por generación		
GENERACIÓN	NÚMERO DE CLIENTES	% RESPECTO A GENERACIÓN
Silent Generation	745	14.3%
Baby Boomers	387	13.5%
Generación X	307	13.8%

Tabla 13 Clientes que no han realizado compras por generación

Entre las causas que pueden explicar este motivo pueden ser las siguientes:

- No se ha identificado con la tarjeta de fidelización durante la compra.
- Se ha trasladado a una localidad donde no haya una tienda.
- No tiene interés de comprar en la cadena de supermercados y por lo tanto se ha perdido el cliente.
- Ha fallecido.

Estos son insights relevantes para el departamento de CRM y marketing ya que pueden actualizar bases de datos de clientes en caso de inoperatividad de la cuenta. Se proponen las siguientes acciones:

- Clientes pertenecientes a Silent Generation: se podría entender que no tienen las facultades para ir a comprar o han fallecido por lo que se podría anular la cuenta en caso de inoperatividad durante un periodo de 2 años.
- Clientes pertenecientes a Baby Boomers: iniciar una campaña de recordatorio de pasar la tarjeta de identificación durante la compra incentivando el uso mediante descuentos.
- Clientes pertenecientes a Generación X: iniciar una campaña de marketing mediante redes sociales y correo para que realicen las compras en los establecimientos.

Categorización de clientes por salario.

Se definen los criterios de salario:

- Salario bajo (entre 10K - 30K dólares),
- Salario medio (entre 30K - 50K dólares)
- Salario alto (entre 50K - +150K dólares).

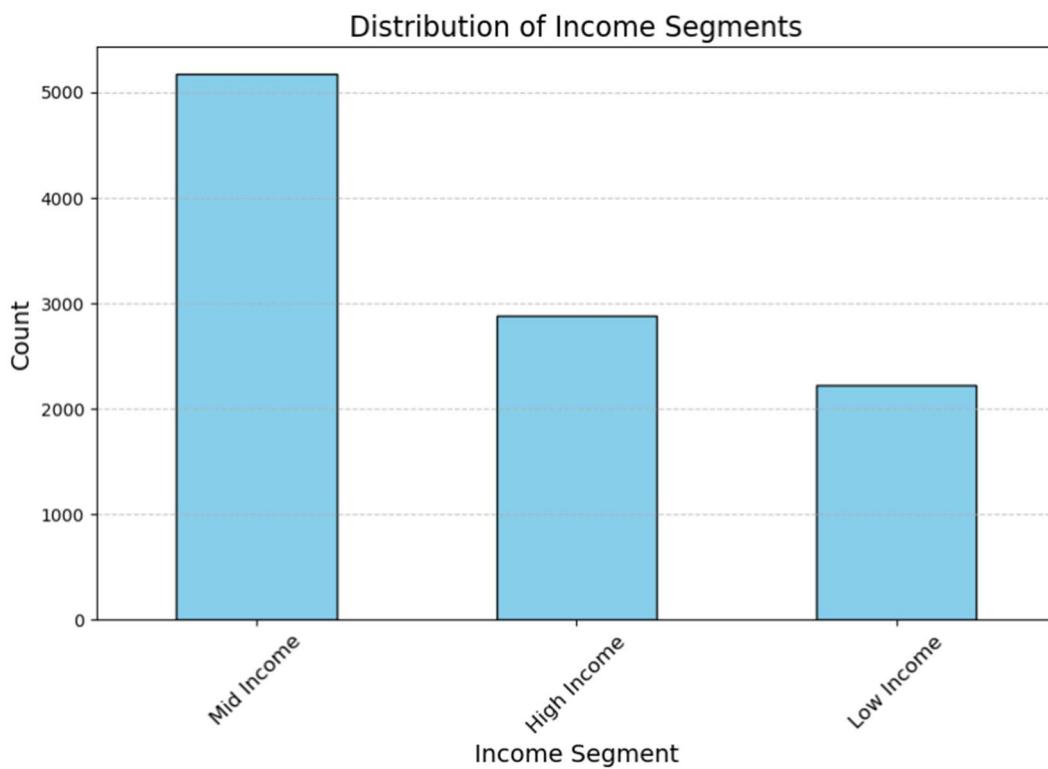


Figura 9 Distribución de salarios

Se analiza la distribución de los rangos de salario de los clientes y se observa que la mayor parte tienen ingresos medios seguidos de altos ingresos y en menor cuantía los ingresos bajos. La distribución de los clientes por salario es muy similar a la distribución de clientes por edad. Puede ser un indicativo de que los datos sintéticos tienen un patrón de configuración.

Categorización de clientes por educación

Se definen los criterios de educación:

- educación baja Partial High School,
- educación media High School Degree y Partial College
- educación alta Bachelors Degree y Graduate Degree

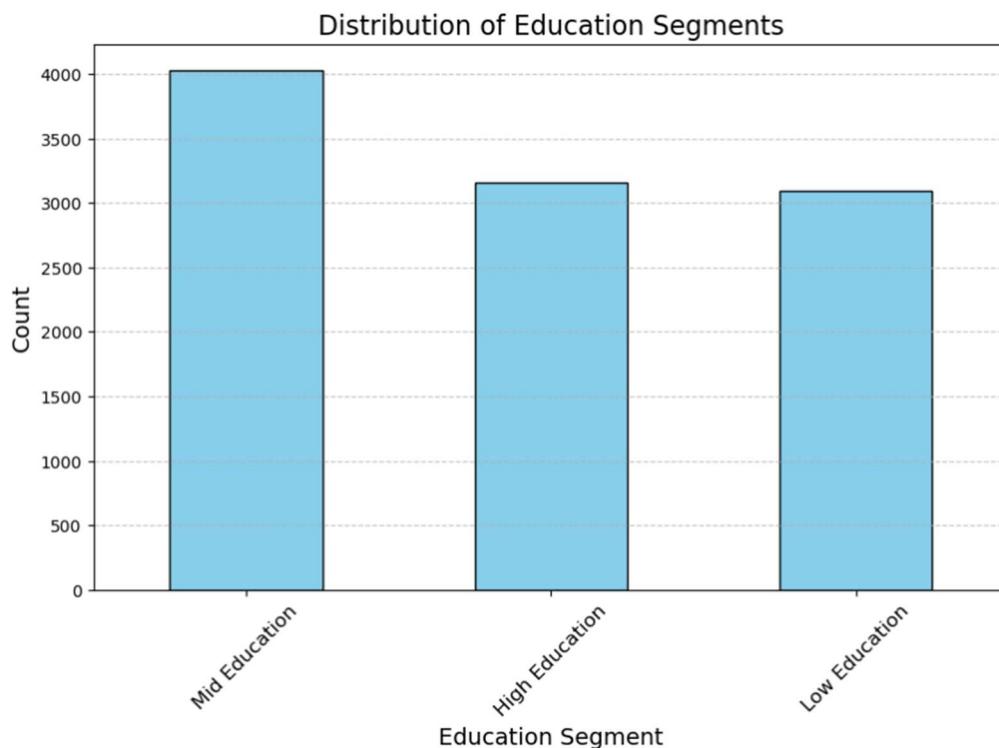


Figura 10 Distribución de educación

Se observa que los clientes con educación media son la mayoría y el número de clientes con educación alta y educación baja son similares. Al combinar los datos de salarios y educación por generación se observa la distribución de salarios por grupo generacional y la distribución de educación por grupo generacional.

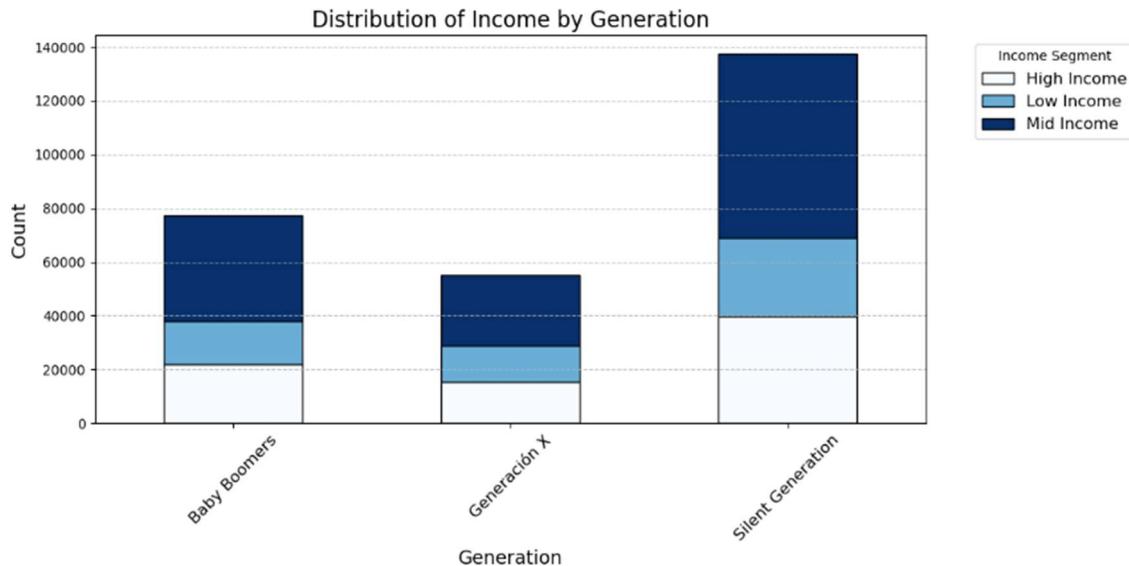


Figura 11 Distribución de salario por generación

Los datos al ser sintéticos no reflejan una gran diferencia entre grupos generacionales ya que debería de existir una mayor diferencia en cuanto a las nuevas generaciones y la silent generation debido a edades de jubilación, así como a la inflación en los salarios.

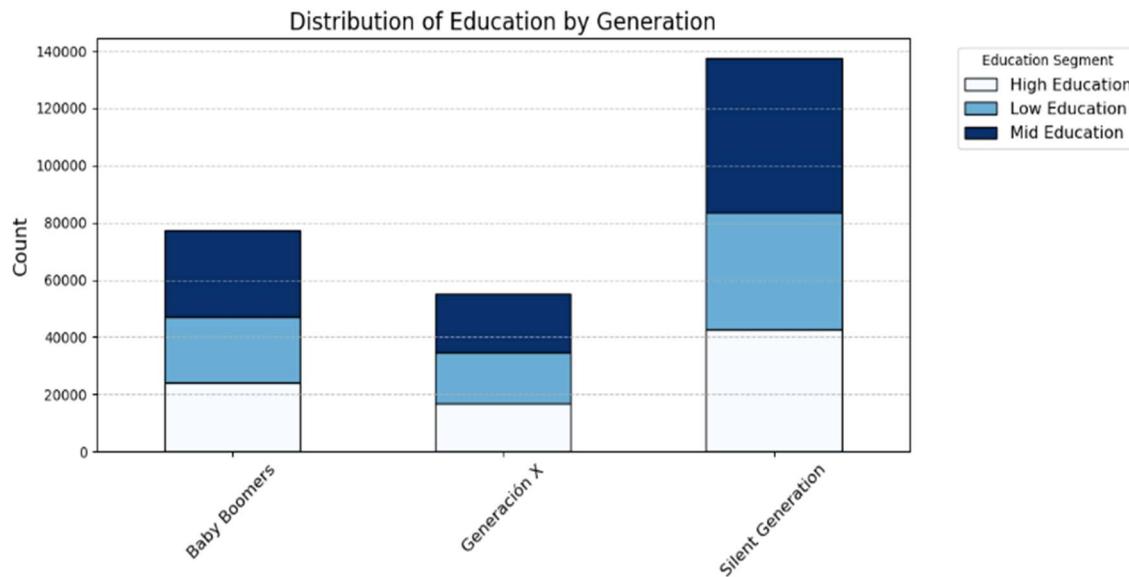


Figura 12 Distribución de educación por generación

De igual manera sucede con la educación donde la Silent Generation debería de reflejar la dificultad de acceso a los estudios respecto a las generaciones más jóvenes.

3.5.2 Análisis Perspectiva de género

Se analiza el conjunto de datos para obtener datos relevantes en cuanto a la perspectiva de género. Para ello se analiza la distribución de hombres y mujeres en las diferentes generaciones, así como en niveles de educación y rango salarial.

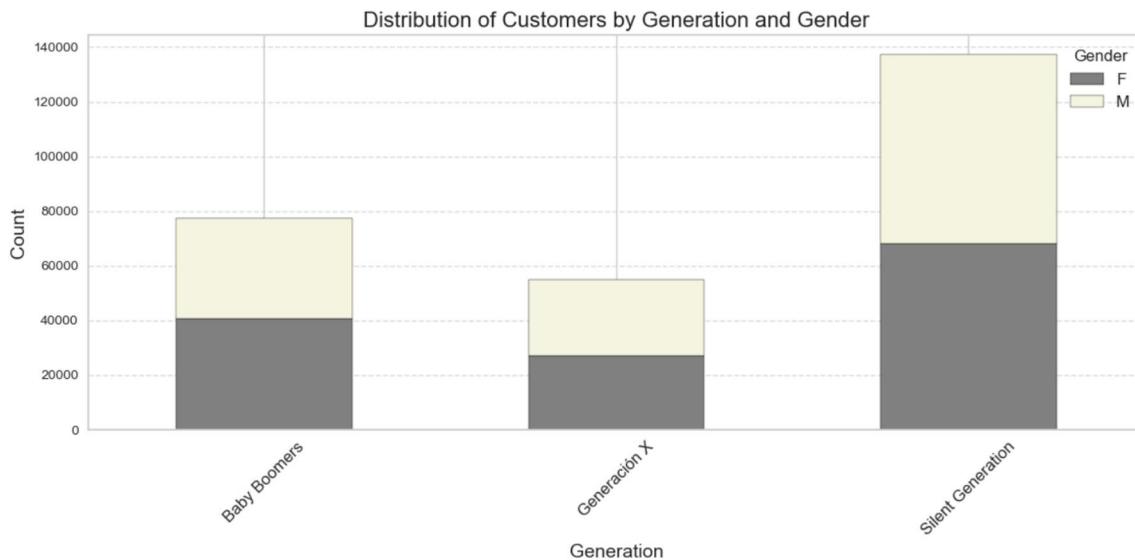


Figura 13 Distribución de generaciones por género

La distribución por géneros en las generaciones es similar para cada generación.

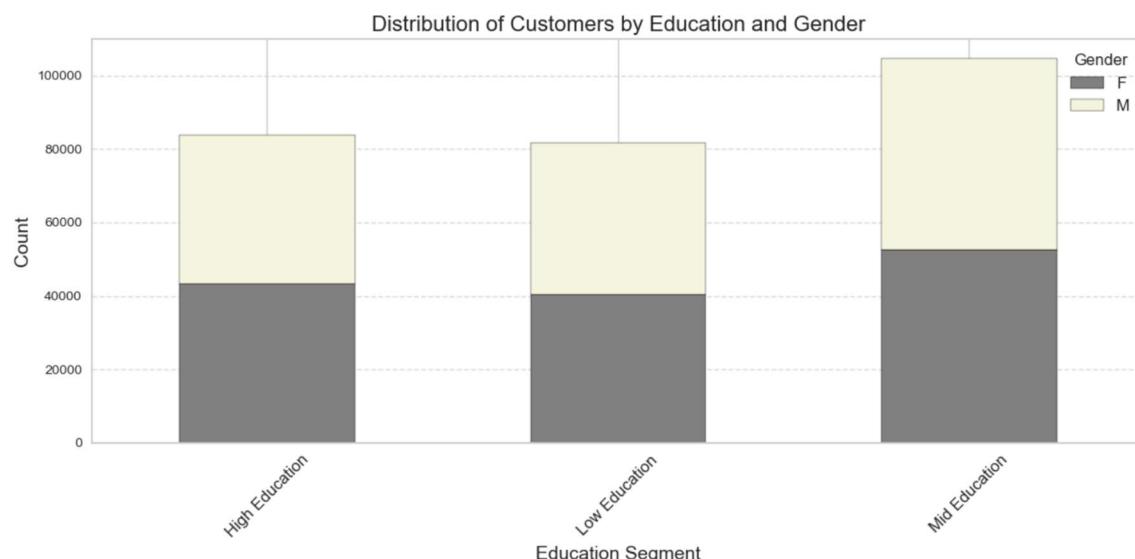


Figura 14 Distribución de educación por género

La distribución de hombres y mujeres en cuanto a educación refleja paridad en los tres niveles de educación bajo medio y alto.

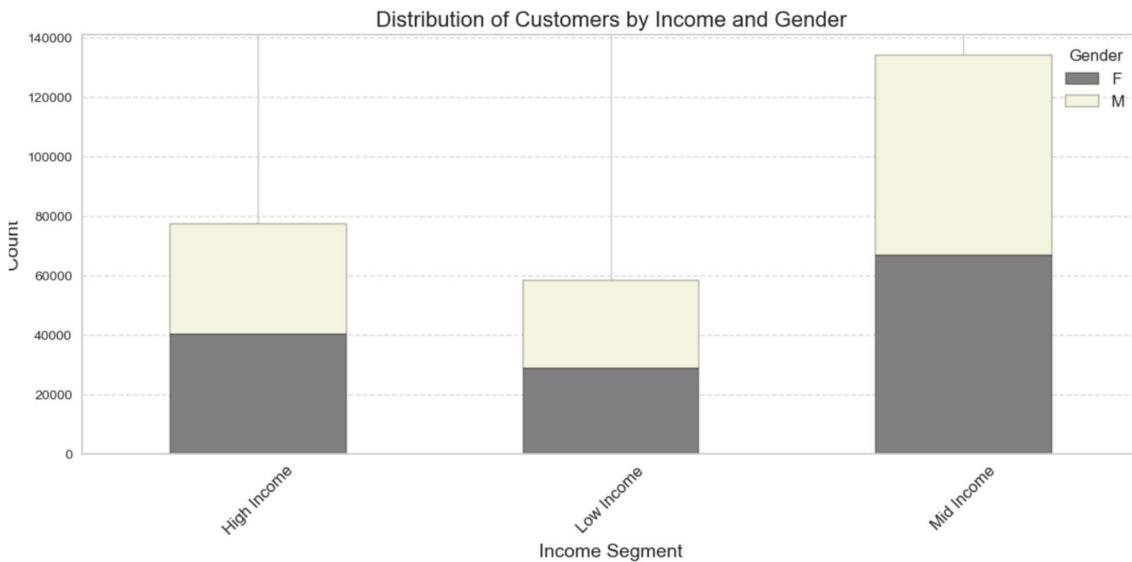


Figura 15 Distribución de salario por género

La distribución de hombres y mujeres en cuanto a salarios refleja paridad en los 3 niveles de ingresos. La perspectiva de género en el conjunto de datos refleja paridad entre ambos sexos en todos los campos analizados. Si bien esto sería lo ideal en cualquier ámbito en el que se analice la perspectiva de género, se ha de considerar que los datos debido a ser sintéticos no reflejan diferencia alguna entre ambos sexos.

3.5.3 Segmentación de clientes

Se pretende realizar una segmentación de clientes que permita clasificar en grupos los comportamientos de compra y características de cada cliente. Se ha calculado el gasto promedio mensual y el número promedio de artículos en compras de cada cliente y se ha combinado estos datos con los datos que se consideran relevantes del dataset principal, así como con la moda de las categorías para cada cliente por lo que obtenemos para cada cliente que categoría de producto es la que más compra.

	customer_id	marital_status	num_children_at_home	age_group	generation	income_segment	education_segment	avg_monthly_spent	avg_monthly_products
0	3449	0	0	1	2	1	1	31.92	15.33
1	7859	1	0	2	2	1	1	19.79	11.00
2	106	1	0	1	2	1	2	33.54	13.40
3	8248	1	0	4	0	0	0	33.17	17.57
4	1906	0	1	2	1	1	2	20.09	11.12
5	5802	1	0	4	0	1	0	27.45	13.71
6	4825	0	2	3	0	2	1	17.95	9.25
7	5262	1	0	4	0	2	2	43.03	20.00
8	2571	0	1	3	0	2	1	31.36	14.33
9	2200	1	0	2	1	1	2	26.83	12.67

Figura 16: Primeras filas customer_info

Las columnas tienen variables categóricas y variables numéricas. Las variables numéricas se han de escalar ya que al tener valores más altos tendrían más peso a la hora de clasificar. El algoritmo seleccionado para la segmentación de los grupos de clientes es Kmeans. Para ello se ha decidido utilizar el algoritmo Kmeans. El modelo Kmeans necesita recibir el número de grupos o clusters “k” que debe calcular. Para ello existen técnicas como elbow y silhouette, que indican en función del parámetro que número de clusters es el óptimo. Se analizan las métricas de Silhouette y Elbow para decidir el número de clusters.

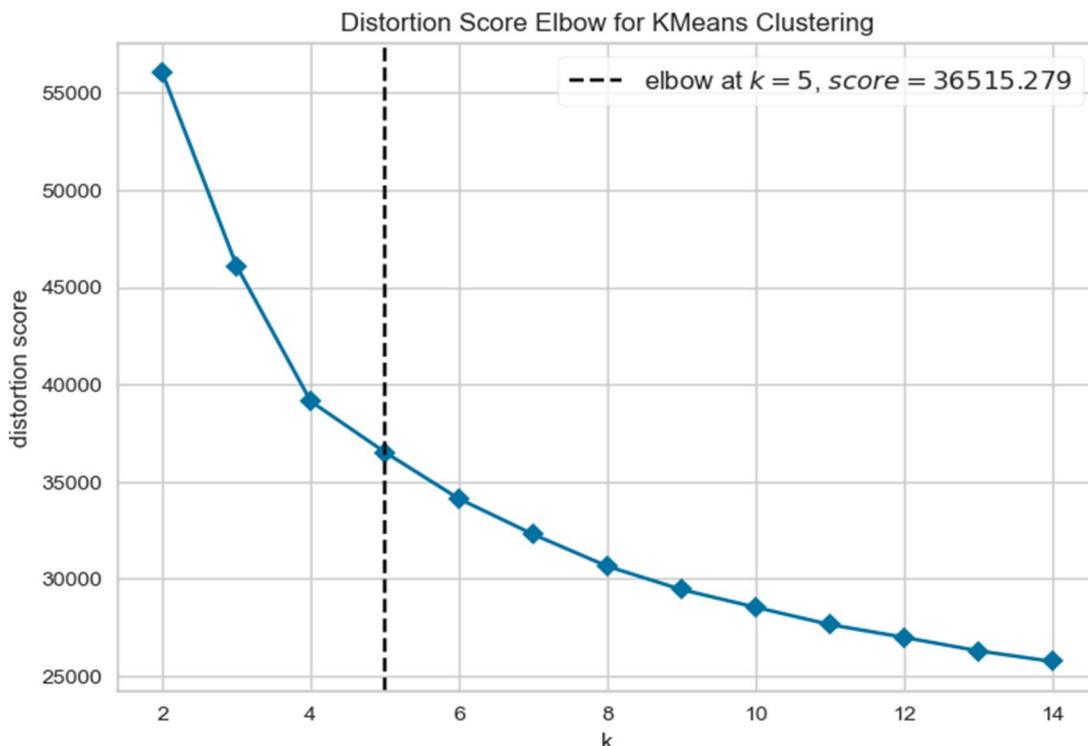


Figura 17: Número de clusters elbow.

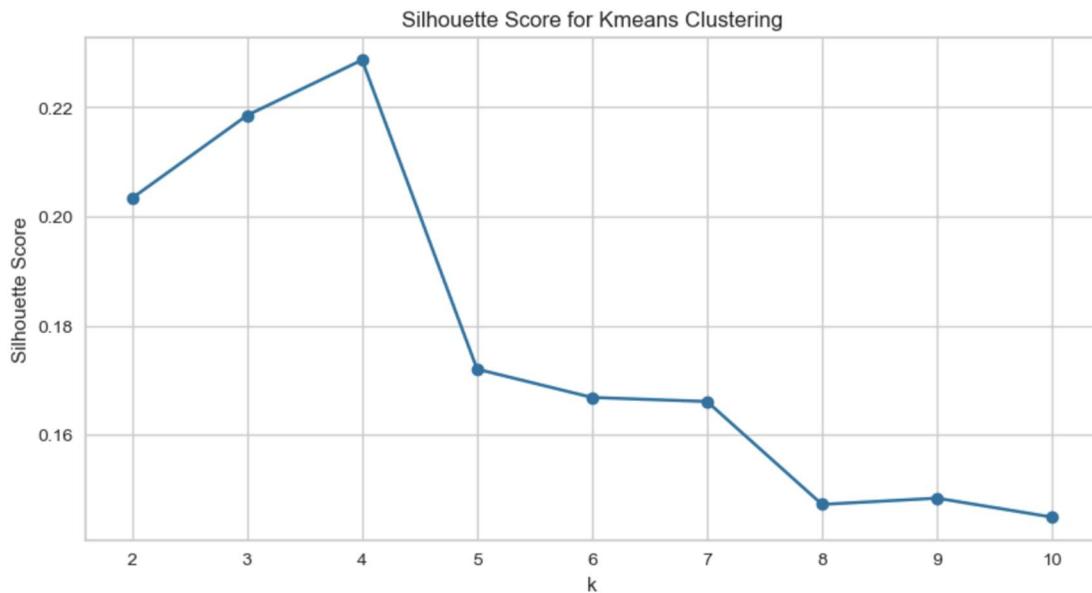


Figura 18: Número de clusters silhouette

Con el método elbow se obtienen que 5 clusters sería lo óptimo, aunque en 4 clusters ya se aprecia un cambio en la tendencia. Para el método silhouette se obtienen 4 clusters. Se decide realizar el algoritmo para $k = 4$ clusters lo que significa que clasificará a los clientes dentro de 4 grupos. Las variantes escogidas para la clasificación son:

- Marital_status: Está casadx o es solterx.
- Num_children_at_home: Número de hijxs en casa.
- Age_group: Grupo de edad.
- Generation: Generación a la que pertenece.
- Income_segment: Nivel de ingresos.
- Education_segment: Nivel de educación.
- Avg_monthly_spent: Gasto promedio mensual.
- Avg_mothly_products: Promedio de productos comprados.

Se ha añadido la categoría de producto que más compra cada clientx mediante el cálculo de la moda en la columna category. Se obtienen los siguientes resultados.

Comparación de clusters					
Nombre columna	Descripción	Cluster 0	Cluster 1	Cluster 2	Cluster 3
marital_status	0 Married 1 Single	0.640795	0.622354	0.000000	0.534682
num_children_at_home	De 0 a 5	0.188874	0.293353	3.085018	0.622832
age_group	0 Minor 1 Young 2 Young Adult 3 Mid Adult 4 Senior	3.642914	1.509469	2.951249	2.994220
generation	0 SilentGeneration 1 BabyBoomers 2 Generation X	0.160530	1.585221	0.596908	0.582370
income_segment	0 Low Income 1 Mid Income 2 High Income	1.067550	1.057557	1.046968	1.108382
education_segment	0 Low Education 1 Mid Education 2 High Education	1.021457	1.006313	0.974435	1.041908
avg_monthly_spent	Media gasto	29.553399	30.202499	31.956046	72.704812
avg_monthly_products	Media productos	13.967889	14.308782	15.129905	33.913237

Tabla 14: Comparación de clusters de clientes.

En la tabla de reflejan los resultados obtenidos más característicos de la clasificación de los clusters. Para cada cluster hay un valor que representa en qué lugar se encuentra dentro del rango de valores de cada variable descriptiva. A continuación, se procede a explicar cada cluster.

- **Cluster 0: Adultxs mayores de 45 y seniors**

En este cluster existen más solteros que casados 0.64, predominan los clientes sin hijos en casa 0.18 , El grupo de edad es Adulto cerca de senior pertenece a la generación Silent Generation, con ingresos medios y educación media. Compran pocos productos 13.96 artículos de media destacando, Non Food 0.23 Y Dairy 0.10. Grupo de personas en edad previa a la jubilación y jubilados divorciados o viudos con compras pequeñas.

- **Cluster 1: Jóvenes Solterxs**

En este cluster existen más solterxs que casadxs 0.62, predominan sin hijos en casa 0.29 el grupo de edad es joven hacia adulto joven, la generación es baby boomer y presencia de Generación X, con ingresos medios y educación media con compras pequeñas 14.30 artículos. Predominan artículos de Non Food 0.24 y Canned Goods 0.10. Jóvenes solterxs con compras pequeñas.

- **Cluster 2: Familias numerosas**

Clientes casados 0.00 con hijos en casa 3.08, el grupo de edad es adulto joven en su tramo mas alto (cercanos a los 45) que pertenecen a la generación Silent generation predominando Baby Boomers. Nivel de ingresos medios y educación baja rozando media. Realizan compras medianas 15.12 artículos destacan Non Food 0.25 y Dairy 0.10. Refleja a un grupo de personas que son familia numerosa con compras medias.

- **Cluster 3: Adultxs edad media 45 años**

Predominan los solteros, aunque hay casi paridad 0.53 con un o ningún hijo en casa 0.62. El grupo de edad se sitúa en adultos jóvenes al límite de adulto medio 2.99 pertenecen a la generación Silent Generation. Tienen ingresos más altos 1.10 y más estudios 1.04, realizan compras grandes 33.91 donde predominan Non Food 0.46 y Vegetables (0.09). Adultos edad media alrededor de 45 años compras grandes.

Estos clusters de clientes son muy importantes para obtener una visión de los tipos de cliente que compra en las tiendas. Se analiza la distribución de los clusters en el conjunto de clientes.

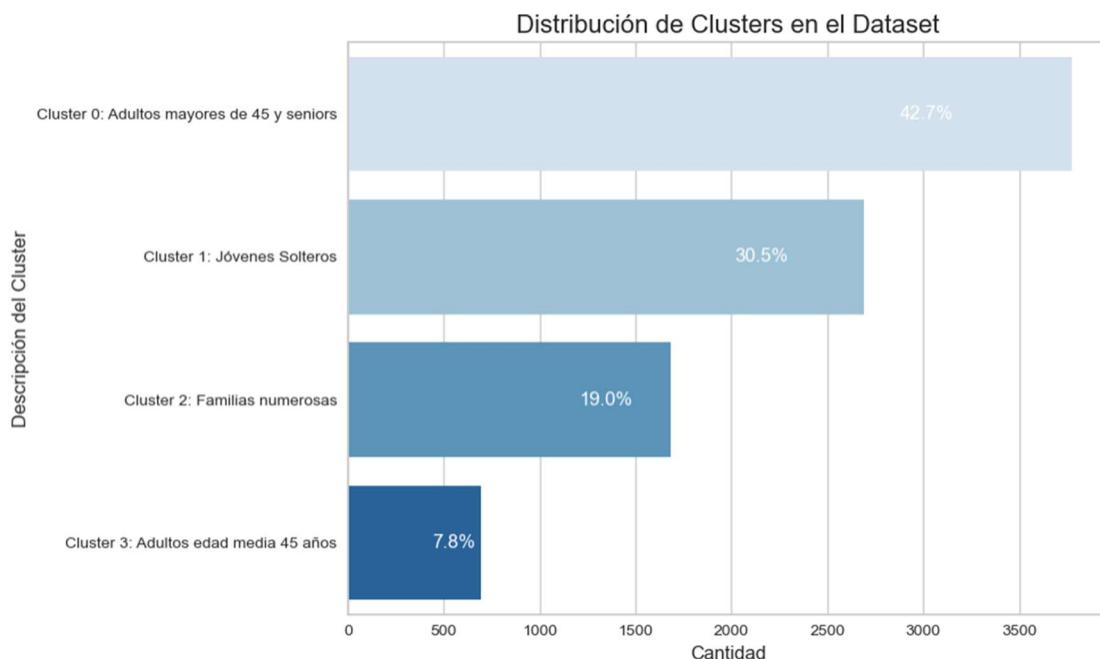


Figura 19: Distribución de clusters de clientes.

Se observa que el 42.7 por ciento de los clientes son adultos mayores de 45 años. Este cluster es el que menos artículos compra. El porcentaje de jóvenes solteros es elevado 30.5 por ciento respecto a las familias numerosas 19 por ciento. El porcentaje de adultos cercanos a 45 años que representa el cluster 3 es solo del 7.8%. Este cluster es el que más dinero gasta y más productos compra.

Con estos datos, es claro que se ha de establecer unas medidas para la captación de nuevos clientes con las características de los clientes pertenecientes al cluster 3. Posibles medidas serían campaña de marketing, ofertas de productos de interés artículos de Non Food y verduras. Al realizar compras grandes, y tener poder adquisitivo, ofrecer cupón de descuento o producto gratis al superar cierta cantidad gastada al mes.

3.5.4 Reglas de asociación

Las reglas de asociación permiten obtener una relación de productos frecuentes en las transacciones, lo que permite poder realizar ofertas de productos o reubicar productos en el establecimiento con el fin de aumentar las ventas. Para realizar las reglas de asociación se crea un dataset en el que se obtiene una de los productos comprados en cada transacción. A continuación se crea una lista con todas las listas de artículos. Los datos de los que se disponen no permiten obtener reglas de asociación con los nombres de los productos ya que no se obtienen reglas con unas métricas aceptables. Se decide realizar las reglas de asociación con las categorías de los productos ya que permitirá conocer que categorías compra en combinación los clientes y así poder realizar ofertas. Se dispone de 58381 transacciones y 45 categorías de productos. Para las reglas de asociación se ha escogido el algoritmo apriori. Se analiza cuantas reglas de asociación se obtendrían con el algoritmo apriori para diferentes valores de soporte mínimo.

- Con min_support de 0.03 se obtienen 91 reglas de asociación
- Con min_support de 0.02 se obtienen 163 reglas de asociación.
- Con min_support de 0.01 se obtienen 404 reglas de asociación.

Cuanto más alto es el support más frecuente es la asociación. En este caso, se va a analizar las combinaciones de categorías de productos más frecuentes que se pueden obtener. Otro tipo estudio sería analizar combinaciones menos frecuentes para obtener relaciones de productos que se puedan explotar en un futuro. Estableciendo el soporte mínimo en 0.03 se obtienen las reglas de reglas de asociación.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Dairy)	(Canned Goods)	0.232593	0.196948	0.050787	0.218352	1.108680	0.004978	1.027383	0.127737
1	(Canned Goods)	(Dairy)	0.196948	0.232593	0.050787	0.257871	1.108680	0.004978	1.034062	0.122067
2	(Fruits)	(Canned Goods)	0.233432	0.196948	0.046162	0.197755	1.004097	0.000188	1.001006	0.005323
3	(Canned Goods)	(Fruits)	0.196948	0.233432	0.046162	0.234389	1.004097	0.000188	1.001249	0.005082
4	(Meat)	(Canned Goods)	0.235402	0.196948	0.047327	0.201048	1.020819	0.000965	1.005132	0.026673
5	(Canned Goods)	(Meat)	0.196948	0.235402	0.047327	0.240303	1.020819	0.000965	1.006451	0.025396
6	(Non Food)	(Canned Goods)	0.365684	0.196948	0.074065	0.202539	1.028389	0.002045	1.007011	0.043520
7	(Canned Goods)	(Non Food)	0.196948	0.365684	0.074065	0.376065	1.028389	0.002045	1.016639	0.034375
8	(Pastries)	(Canned Goods)	0.151625	0.196948	0.034275	0.226051	1.147770	0.004413	1.037603	0.151755
9	(Canned Goods)	(Pastries)	0.196948	0.151625	0.034275	0.174030	1.147770	0.004413	1.027126	0.160320
10	(Snacks)	(Canned Goods)	0.264196	0.196948	0.05532	0.210192	1.067248	0.003499	1.016769	0.085635

Figura 20: Reglas de asociación

Para analizarlas, se pueden filtrar y ordenar por las métricas.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
8	(Pastries)	(Canned Goods)	0.151625	0.196948	0.034275	0.226051	1.14777	0.004413	1.037603	0.151755
9	(Canned Goods)	(Pastries)	0.196948	0.151625	0.034275	0.174030	1.14777	0.004413	1.027126	0.160320
30	(Spreads)	(Dairy)	0.126171	0.232593	0.033127	0.262558	1.12883	0.003781	1.040634	0.130606
31	(Dairy)	(Spreads)	0.232593	0.126171	0.033127	0.142426	1.12883	0.003781	1.018954	0.148718
1	(Canned Goods)	(Dairy)	0.196948	0.232593	0.050787	0.257871	1.10868	0.004978	1.034062	0.122067

Figura 21: Reglas de asociación ordenadas

A continuación se explican las métricas mas relevantes

- Antecedents: lo que compran primero.
- Consequents: Productos que se compran después de los antecedentes.
- Antecedent Support: Frecuencia de aparición relativa en las transacciones del antecedente
- Consequent Support: Frecuencia de aparición relativa en las transacciones del consecuente.
- Support: Proporción de transacciones que contienen los antecedentes y los consecuentes.
- Confidence: Probabilidad de que, dado que se compraron los antecedentes, también se compren los consecuentes.
- Lift: Probabilidad que se compren los consecuentes si los antecedentes ya se compraron, comparado con comprarlos por separado. Se mide el valor que sobrepasa de 1.

Si analizamos la primera regla de la Flgura 23 obtenemos:

Antecedent Pastries, Consequent Canned Goods

Pastries aparece en el 15 por ciento de las transacciones y canned goods en el 19 por ciento. En un 3 .43 por ciento de las transacciones se han comprado Pastries y Canned Goods. Existe un 22.6 por ciento de probabilidad que un cliente que compra Pastries, tambien compre canned goods. Aumenta un 14.78 por ciento la probabilidad que alguien compre canned good si previamente ha comprado Pastries.

Antecedent Spreads, Consequent Dairy

Los productos Spreads(mantequillas, mermeladas) aparecen en el 12.6 por ciento de las transacciones y los productos Dairy(lácteos) aparecen en el 23.26 por ciento de las transacciones. Juntos aparecen en el 3.31 por ciento de las transacciones lo que supone aparecer en 1932 transacciones. Si un cliente compra Spreads existe un 26.26 por ciento de probabilidad que compre también Dairy. Aumenta un 12.88 por ciento la probabilidad de comprar Dairy si antes se ha comprado Spreads.

Los clientes que compran Spreads(mantequillas, mermeladas)

Se ha filtrado y ordenado por support y confidence y lift con diferentes parámetros y no se hayan reglas de asociación muy relevantes. Esto es debido a los datos de transacciones que se disponen. Como resultado de las dos reglas analizadas se obtienen los siguientes **insights**:

- **Regla 1 Pastries, Canned Goods** Si se desea aumentar la venta de productos enlatados debido por ejemplo a un cambio de formato, se podría realizar una oferta de productos de pastries que son más económicos lo que aumentaría la probabilidad de que los clientes comprasen productos enlatados.

- **Regla 2 Spreads, Dairy** Si se desea aumentar el volumen de ventas de productos lácteos una buena solución sería que estuviese ubicado en el mismo pasillo que los productos de desayuno untables Spreads. De igual manera los productos lácteos refrigerados deben de estar cerca de los untables de desayuno refrigerados.

Los resultados con el algoritmo a priori para las categorías de productos son reglas de asociación con muy poca representación en el conjunto de transacciones. Por tanto, aún resultará menos frecuente la aparición de productos individuales combinados. Como se ha descrito anteriormente la principal diferencia entre el algoritmo a priori y el Eclat es el tipo búsqueda. Apriori realiza la búsqueda de elementos en horizontal mientras que el Eclat lo realiza en vertical. El algoritmo a priori es un método exhaustivo de búsqueda de elementos en el dataset completo mientras que el Eclat optimiza recursos buscando la intersección en conjuntos de items. El algoritmo a priori es muy válido para datasets que no sean muy grandes. El algoritmo Eclat es muy válido para datasets muy grandes sin embargo consume muchos recursos de memoria. Tras realizar pruebas de rendimiento se ha utilizado el algoritmo a Apriori.

3.6 Predicción de demanda

Para la predicción de la demanda se hace uso de las series temporales. Un aspecto fundamental de las series temporales es disponer del mayor número de registros posibles para poder realizar una buena predicción. Se analizan las fechas de las transacciones y se agrupan por país para saber que rango de fechas disponemos.

- País: Canadá Fecha inicio: 1998-01 Fecha fin: 1998-12 Total de meses: 12
- País: México Fecha inicio: 1998-01 Fecha fin: 1998-12 Total de meses: 12
- País: USA Fecha inicio: 1997-01 Fecha fin: 1998-12 Total de meses: 24

Se decide utilizar el dataset filtrado por país USA ya que es el único del que se disponen datos de los dos años. Canadá y México se descartan ya que el modelo de predicción con 12 meses no es suficiente para poder recoger las frecuencias. Tras realizar algunas pruebas, se decide que la frecuencia de los datos sea diaria ya que tanto con los datos agrupados de ventas mensuales como semanales no eran suficientes para obtener los resultados esperados. Se asigna en una columna el nombre del día de la semana para poder detectar los fines de semana o días de cierre.

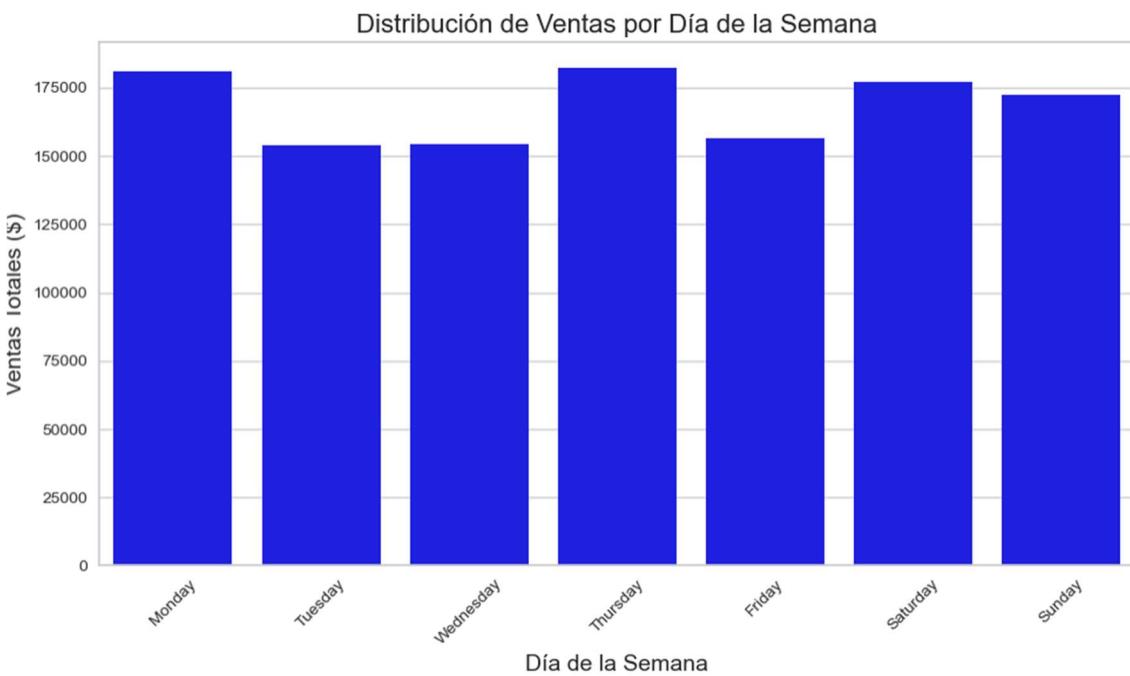


Figura 22: Distribución de ventas USA por día de la semana.

En la figura 24 se muestra un histograma de la venta total distribuida en los días de la semana. Se observa que no existe un día de cierre programado cíclico ya que las ventas están distribuidas a lo largo de la semana de manera similar. Se aprecia que lunes y jueves son los días con más venta y martes y miércoles los que menos.

Se realiza un histograma de las ventas totales por día de la semana y mes para ver si existen diferencias a lo largo de los meses del año.

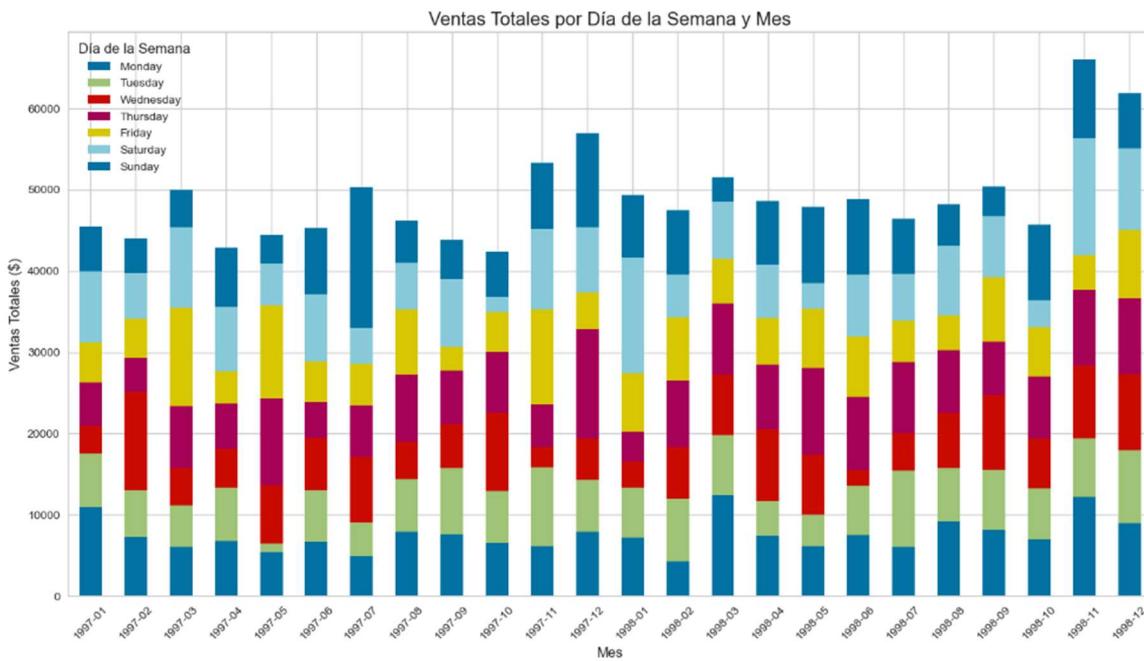


Figura 23: Distribución de ventas USA por día de la semana y mes.

En el histograma de la figura 25 aparece una columna por cada mes en el periodo que va de 1997 a 1998 es decir 24 columnas. Dentro de cada columna aparecen 7 colores uno por cada día de la semana. Se aprecia que en varios meses hay tramos de color que son muy pequeños por lo que existen pocas ventas. Por ejemplo, en enero de 1997 el color rojo que representa miércoles es muy reducido. Otro ejemplo es el mes de mayo de 1997 donde el color verde que representa el martes es mucho más reducido que en el resto de los meses. Se realiza un análisis de este suceso y se comprueba en las ventas por día que existen grandes diferencias en el dinero gastado por los clientes el sábado día 4 de enero y el miércoles 8 de enero respecto al resto de días. Se crea una tabla donde se realiza un sumatorio de los días de la semana en los que no hay registro de venta por cada tienda. De esta manera se muestra de manera gráfica el número de días faltantes de datos por cada tienda.

store_country	Transaction Date	total_daily_spent	Day of Week
386	USA 1997-01-01	706.34	Wednesday
387	USA 1997-01-02	1304.53	Thursday
388	USA 1997-01-03	1294.12	Friday
389	USA 1997-01-04	42.87	Saturday
390	USA 1997-01-05	1987.19	Sunday
391	USA 1997-01-06	2162.34	Monday
392	USA 1997-01-07	2696.61	Tuesday
393	USA 1997-01-08	69.30	Wednesday
394	USA 1997-01-09	1090.65	Thursday
395	USA 1997-01-10	1022.08	Friday

Figura 24: Registro de ventas USA por día.

Day of Week	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Store ID							
2	6	0	3	2	2	4	1
3	0	0	4	4	0	1	5
6	0	2	4	2	2	2	4
7	2	4	5	2	1	3	2
11	1	2	2	0	2	3	2
13	1	1	5	3	2	5	2
14	4	1	3	2	3	3	3
15	0	1	3	2	4	3	2
16	2	2	2	2	0	2	3
17	0	2	3	0	2	3	2
22	1	1	1	2	2	3	3
23	3	2	1	3	2	1	3
24	1	4	1	5	2	1	1

Figura 25: Sumatorio días de la semana sin ventas en tiendas de USA.

En la Figura 25 se observa como en la tienda 2 existen 6 lunes, 3 miércoles, 2 jueves, 2 viernes, 4 sábados y 1 domingo que no se ha registrado ninguna venta. Se observa que, en el resto de las tiendas, también aparecen días sin ventas. Estos datos se han

de tratar de alguna manera ya que esos días sin ventas aleatorios afectan a la predicción de demanda. Se decide realizar una exploración de los días con venta por tipo de tienda.

	store_type	total_sales	total_products	average_sales_per_day	average_products_per_day	days_with_data
0	Deluxe Supermarket	328154.64	155134	1367.311000	646.391667	240
1	Gourmet Supermarket	98114.78	46129	817.623167	384.408333	120
2	Mid-Size Grocery	47576.78	22377	396.473167	186.475000	120
3	Small Grocery	28800.92	13598	80.002556	37.772222	360
4	Supermarket	675391.36	318694	938.043556	442.630556	720

Figura 26: Registro de ventas y días con venta por tipo de tienda USA.

Se observa en la Figura 28 que en la columna days_with_data aparece el número de días con ventas registradas. Se disponen de dos años de registro por lo que en total son $365 * 2 = 730$ días. El tipo de tienda que más datos dispone es Supermarket que tiene 720 días de registro de ventas. Debido a esta situación se ha de realizar la predicción de la demanda de los supermercados de USA descartando el resto de los tipos de tienda debido a la falta de datos. Una vez agrupados los días por fecha se obtiene que tenemos un registro de 487 días faltan 243 días para completar los 730 que supondrán 2 años. Se decide completar los días que no hay registro para ello se va a realizar la imputación de datos mediante knearest neighbours (KNN) que calcula la distancia euclíadiana entre un número k de vecinos cercanos. Se selecciona un número bajo 3 para tratar de captar las ventas más similares cercanas. Se crea un dataset con los datos filtrados de tipo de tienda supermarket y se unifica con un dataset de fechas comprendidas entre enero de 1997 y diciembre de 1998 donde se obtienen los días sin venta ya que aparecen como valores nulos en ventas. Ya que en el dataset original no aparece venta en el día 01 de enero de 1997, se decide eliminar el registro de 01 de enero de 1998.

3.7.1 PROPHET

Se realiza análisis de serie temporal mediante Prophet. Se han realizado varios modelos modificando los parámetros de estacionalidad, añadiendo períodos de días festivos y agregando regresores. También se ha realizado modelos con una división de los datos de entrenamiento y de validación para compararlos a los resultados de los modelos con todos los datos. Por último, se ha realizado cross validation para evaluar la capacidad del modelo por períodos de tiempo y horizontes diferentes. Las métricas que se han utilizado para la evaluación del modelo son las siguientes:

MSE (Mean Squared Error) mide lo distante que está la predicción de los datos reales

RMSE (Root Mean Squared Error): Mide el error en unidades de los datos.

MAE (Mean Absolute Error): mide el tamaño promedio del error.

MAPE (Mean Absolute Percentage Error) Media del porcentaje de error absoluto.

MDAPE (Median Absolute Percentage Error): Mediana del porcentaje de error absoluto, menos sensible a valores atípicos que el MAPE.

SMAPE (Symmetric Mean Absolute Percentage Error): Como el MAPE pero más equilibrado (simétrico) entre predicciones y valores reales.

Coverage indica el porcentaje de predicciones que están dentro del intervalo de confianza. Los resultados de las métricas de los modelos son los siguientes.

	RMSE	MAE	MAPE	MdAPE	sMAPE	Coverage
Modelo 1	591.481098	432.213587	33.828926	25.933646	30.269278	0.874828
Modelo 2	570.391208	419.824226	32.647045	25.916953	29.362570	0.873453
Modelo 3	568.949359	418.025094	32.452927	25.947805	29.244514	0.877579

Figura 27: Métricas modelos Prophet.

En la tabla de modelo sin separación de entrenamiento ni cross validation se observan los resultados de 3 modelos. Las métricas en cuanto a cobertura son similares, aunque esta métrica puede llevar a engaño en caso de que el modelo esté sobreajustado o esté capturando un intervalo de confianza es decir un rango de posibles valores muy amplio. Donde se aprecian diferencias es en el RMSE y el MAE siendo el Modelo 3 el que

valores más bajos tiene. El MAPE también es más bajo para el Modelo 3 por lo que se obtiene que el modelo que menor error tiene es el Modelo 3. El modelo 1 se ha ejecutado sin ninguna configuración adicional. El modelo 2 se ha ejecutado añadiendo la estacionalidades diaria, semanal, mensual, trimestral y anual. También se le ha añadido el regresor día de la semana. El modelo 3 ha sido ejecutado con los mismos parámetros que el Modelo 2 añadiendo el rango de fechas de días festivos. Se ha obtenido un mejor resultado en el modelo que tiene más parámetros para capturar la estacionalidad. En los resultados el RMSE indica que se puede equivocar en 568.94 unidades en la predicción. Puede parecer mucho, pero los datos de los que se disponen ya reflejan esta diferencia existiendo días que se ha facturado menos de 100 dólares. Ante estas situaciones anómalas, es de esperar que existan errores de esta magnitud. El MAPE indica que el modelo se equivoca el 32.45 por ciento de las veces por lo que no es nada bueno ya que, si se pretende predecir la venta, este rango es muy elevado. En la figura 28 se observa en azul los datos de ventas, en rojo intenso los valores predichos por el modelo y en rojo tenue los intervalos de confianza donde el modelo estima que puede predecir el rango de venta. El modelo intenta capturar la tendencia siendo mejor cercano a 1999. La muestra una tendencia positiva en cuanto a las ventas.

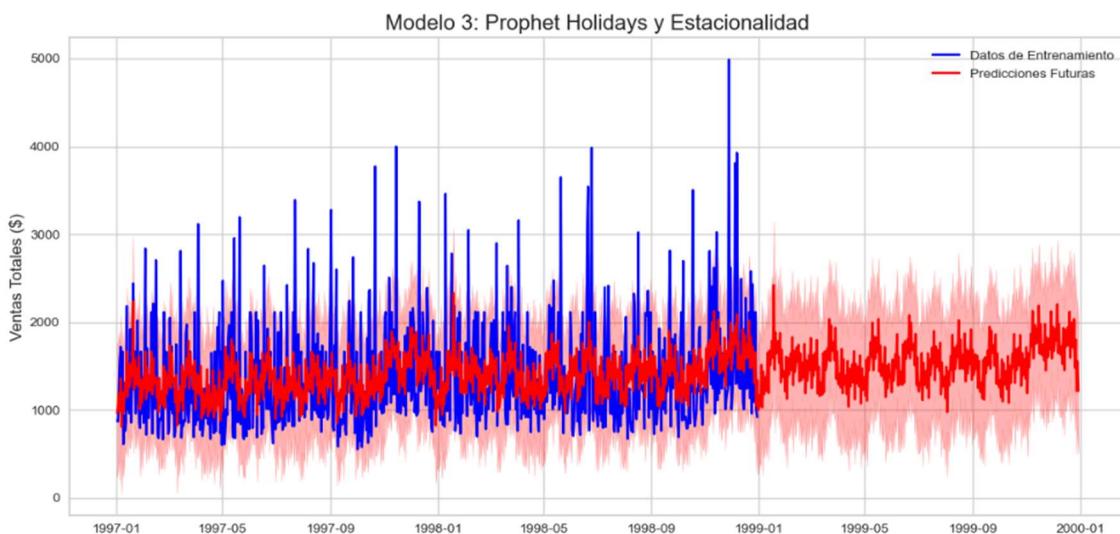


Figura 28: Predicción de ventas en supermercados de USA modelo Prophet

Se analiza la estacionalidad capturada por el modelo en las gráficas

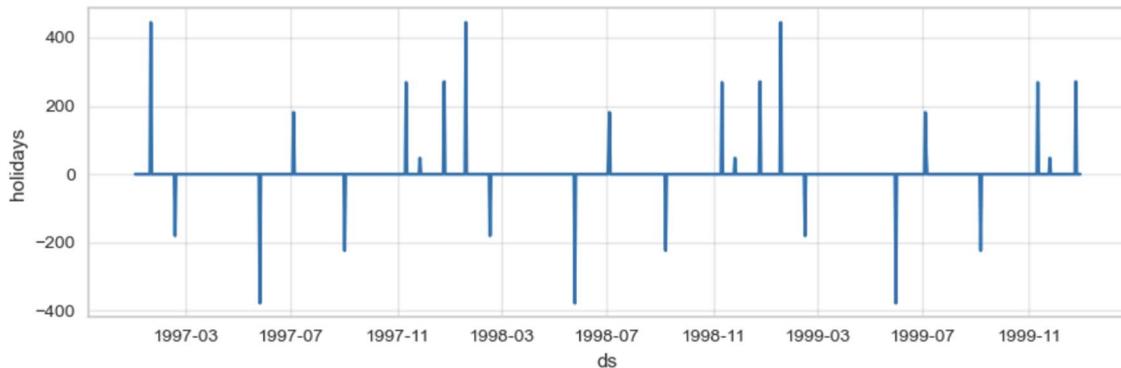


Figura 29: Frecuencia de ventas en periodo vacacional

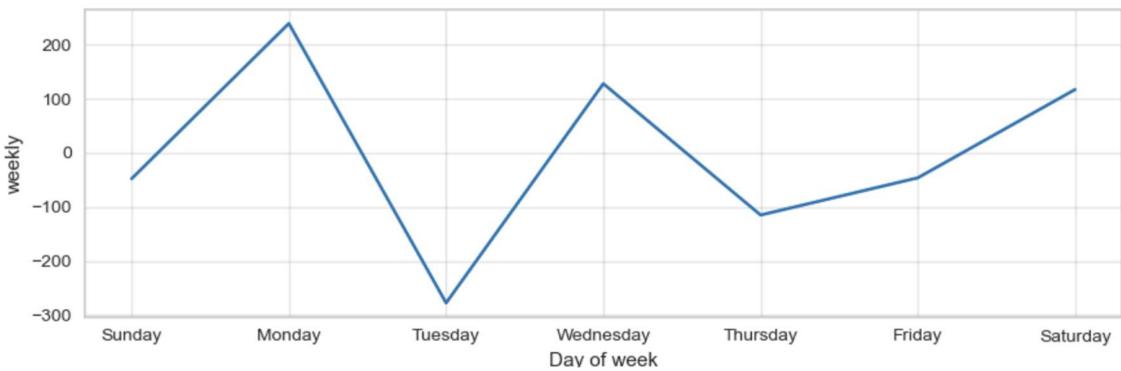


Figura 30: Frecuencia de ventas en supermercados semanal

En la figura 29 se observa la representación de la estacionalidad de los días festivos añadidos al modelo mediante el calendario de días festivos. Se observa como varias estacionalidades están representadas en negativo por lo que estos días el modelo no ha captado el patrón de ventas. Al ser los datos sintéticos es probable que no hayan reflejado de manera significativa los festivos. En cuando al análisis semanal se observa una caída en el día martes que representa precisamente la gran diferencia entre unos días y otros respecto a la venta. Los datos no proceden de una venta completa sino de un sesgo de la venta que representan los clientes con tarjeta de socio. Esa caída tan abrupta puede indicar que existen ofertas los lunes y por ende los martes los clientes con tarjeta de fidelización no acuden tanto a comprar. Se observa

claramente como martes y jueves son los días con menos frecuencia de compras viernes y sábado existe un aumento de ventas y se mantiene la venta el domingo

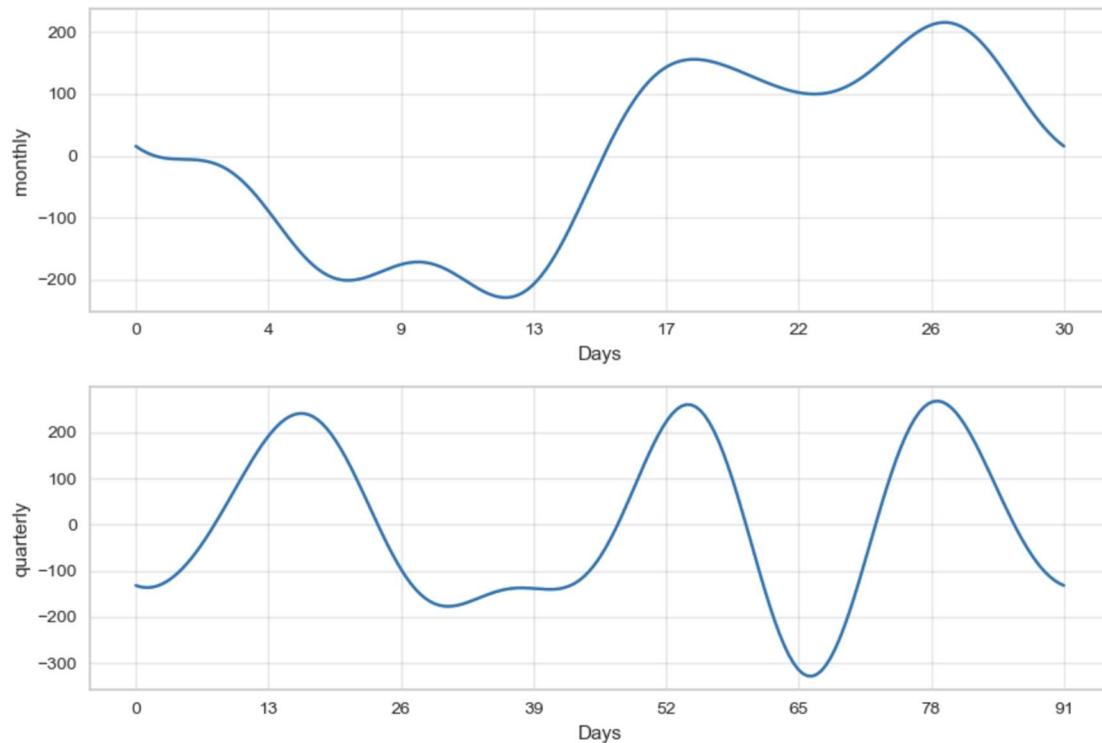


Figura 31: Frecuencia de ventas en supermercados mensual y trimestral

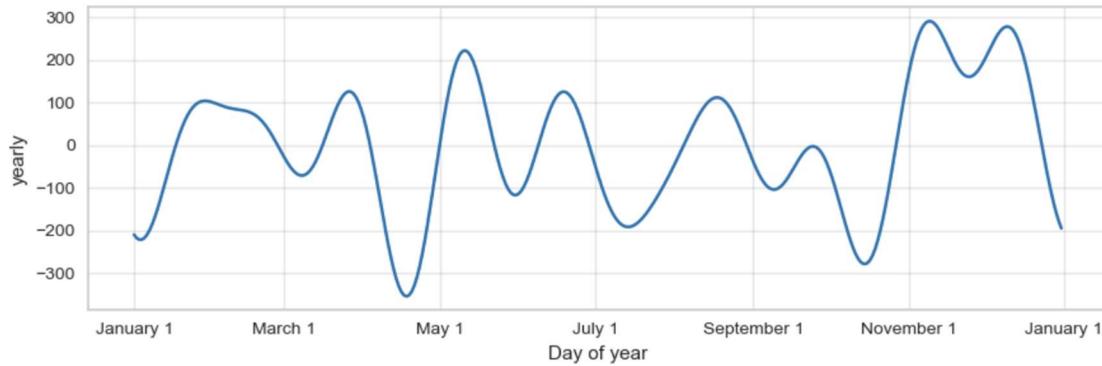


Figura 32: Frecuencia de ventas en supermercados anual

En cuanto a la frecuencia mensual en la figura 31 se observa cómo hay una caída en las ventas en la segunda semana del mes y una subida muy grande en la tercera semana que podría representar ciclos de cobro de pensiones o salarios a mitad de mes. En la frecuencia trimestral se observa de igual manera las crecidas en ese

periodo de medio mes, En cuanto a la frecuencia anual de la figura 30 se observa como el mes de abril es el que menos venta tiene. Las ventas también caen en julio y octubre. Los periodos de mas venta coinciden con fechas a final de mes destacadas como día de acción de gracias, Black Friday, y navidad. El análisis de esta predicción es que no es buena ya que no permite captar las ventas de manera correcta en el periodo de validación y los errores debido a los datos son demasiado grandes. En estos resultados influyen de manera muy significativa los siguientes hechos:

- Los datos de ventas **están sesgados** por clientes con tarjeta de socio
- Los datos de venta han sido reducidos a un país y a un tipo de establecimiento (supermercados).
- Se ha realizado **imputación de datos faltantes** para completar los datos de dos años de registros
- Existen **diferencias significativas** de ventas entre días de la semana que no corresponden a una frecuencia normal.
- Se pierde la posible estacionalidad de las ventas por temporada alta o baja de los lugares más turísticos sea en verano o en invierno.

Se han obtenido las métricas de los modelos con las mismas configuraciones de parámetros, pero esta vez, entrenándolo con datos hasta una fecha. Se ha decidido marcar como fecha de corte de los datos 1998-09-01. De esta manera el modelo recibe más datos de entrenamiento de inicio. Los resultados mejoran al otro modelo en cuanto al MAPE sin embargo el error RMSE es mucho mayor.

	RMSE	MAE	MAPE	MdAPE	sMAPE	Coverage
Modelo 1 train	675.788128	459.902279	31.617817	24.601911	28.904009	0.842975
Modelo 2 train	630.434644	450.586023	30.827171	26.154118	28.383200	0.859504
Modelo 3 train	624.220472	442.050446	30.152656	25.927703	27.853596	0.876033

Figura 33: Métricas modelos entrenados Prophet

Se ha realizado cross validation para los modelos con mejores resultados obteniendo las siguientes métricas donde H es Horizon y P period.

	Modelo	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	Modelo 1 H30,P7	16 days 12:00:00	397206.534937	627.404430	449.569471	0.339478	0.264625	0.301400	0.834325
1	Modelo 1 H30,P1	16 days 12:00:00	395010.663150	628.481957	448.440571	0.339458	0.259412	0.301247	0.832869
2	Modelo 1 H7,P1	4 days 00:00:00	399154.402289	631.786224	453.256933	0.339765	0.263618	0.303111	0.835072
3	Modelo 3 H30,P1	16 days 12:00:00	420683.398706	648.588004	468.231941	0.353322	0.272321	0.313390	0.798982
4	Modelo 3 H30,P7	16 days 12:00:00	422023.275785	646.860154	468.990952	0.351937	0.275512	0.313671	0.802331

Figura 34: Métricas modelos crossvalidation Prophet

Los resultados de los modelos son peores, ya que no es capaz de captar tendencias o estacionalidades largas y debido a la imputación de datos realizada ya que al haber configurado periodos cortos se ve influido por los valores imputados. El mejor modelo es el 3 que incluye regresores, estacionalidades y el calendario de vacaciones. Se comparan los modelos 3 sin entrenamiento y con entrenamiento mediante la selección de la fecha de corte.

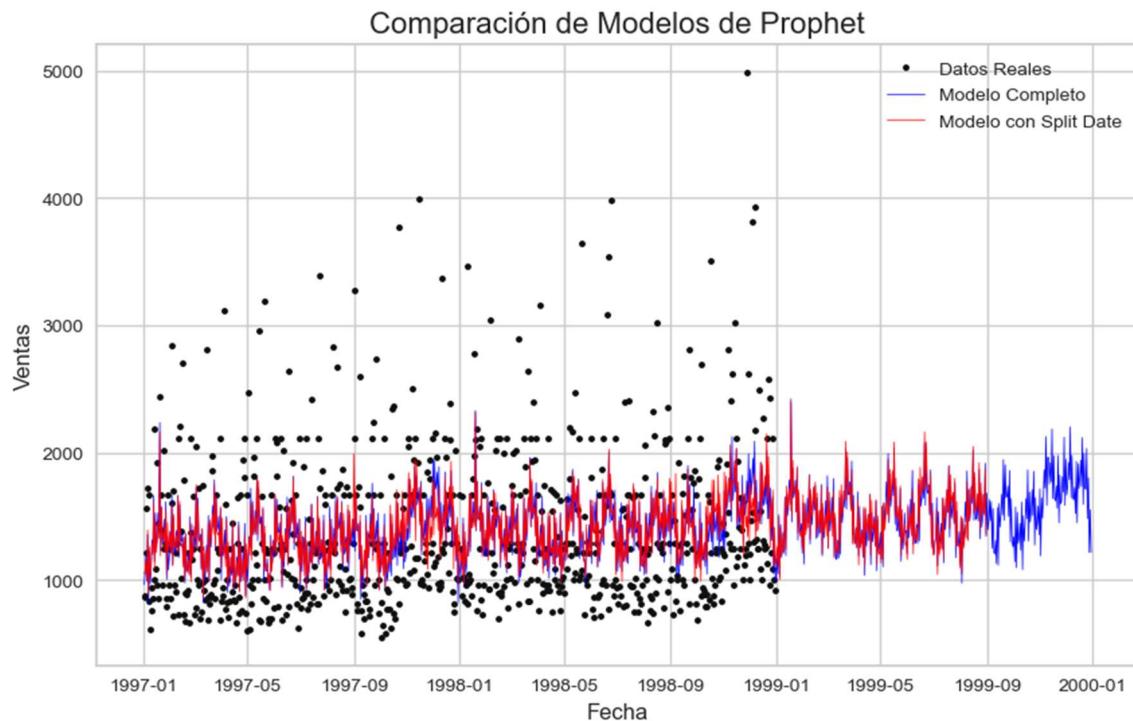


Figura 35: Comparación modelos Prophet

Comparación de mejores modelos		
Métricas	Modelo 3	Modelo 3 Entrenado
RMSE	568.949359	624.220472
MAE	418.025094	442.050446
MAPE	32.452927	30.152656
MdAPE	25.947805	25.927703
Smape	29.244514	27.853596
Coverage	0.885832	0.876033

Tabla 15: Comparación de mejores modelos.

Se observa que el modelo 3 tiene un menor RMSE y MAE y los intervalos de confianza recogen mejor los valores reales. Esto es debido a que se entrena y valida durante todo el proceso. El modelo 3 entrenado tiene mejor MAPE y SMAPE, pero esto es debido a que desde la fecha marcada como fin de entrenamiento hasta que se acaban los datos solo hay 180 días de registros. De los dos mejores modelos se decide que el modelo 3 sin entrenamiento es mejor, aunque, no es bueno debido a los niveles de error altos y a la poca cobertura.

4 Conclusiones y trabajos futuros

Los resultados no son los esperados debido a la falta de datos suficientes. En el proyecto se planteaba realizar un sistema de recomendación de productos que no se ha podido llevar a cabo debido a la falta de registro de transacciones que han influido también en la calidad de los resultados de las reglas de asociación y en la predicción de demanda.

Para la predicción de demanda se pretendía realizar dos partes una enfocada en las ventas y otras por los productos para poder obtener la cantidad de producto necesaria para satisfacer la demanda esperada. Los datos son sintéticos y enfocados en clientes que están fidelizados con una tarjeta por lo que los datos tienen un sesgo que no permite analizar las ventas reales de la empresa, así como su proyección futura. Este trabajo trata de exponer la relevancia que tienen los datos recopilados de una empresa para el análisis del pasado, presente y futuro de la organización. Se ha reflejado en los resultados obtenidos la importancia de la recogida de datos inicial, cumpliendo con las normativas de regulación de privacidad, y el impacto ético-social, de sostenibilidad y diversidad. También se ha expuesto la necesidad del tratamiento de los datos una vez almacenados, para su modificación, actualización o supresión. El análisis exploratorio ha permitido de lo que inicialmente eran datos en tablas, obtener insights relevantes para la empresa como los tipos de cliente, su distribución por generación y grupo de edad y su comportamiento en los hábitos de compra para tomar decisiones estratégicas en cuanto a procedimientos de ventas, marketing y CRM. También se ha reflejado que los datos históricos permiten el análisis futuro de la tendencia de consumo lo cual permite tomar medidas de abastecimiento de productos para cumplir con la demanda esperada, así como con análisis posteriores en conjunto con datos socio demográficos de la población permitirán destacar zonas objetivo para realizar una expansión comercial. Se tenía como objetivo realizar una visualización o panel de control el cual no se ha podido llevar a cabo debido a las tareas adicionales realizadas en análisis

exploratorio de datos y en el tratamiento de los datos faltantes. Sin embargo, debido a la falta de datos se ha obtenido un conocimiento profundo en EDA y en que tipos de datos se necesitan y su estructura para el uso de modelos de machine learning en el sector retail.

La planificación inicial era muy optimista ya que se esperaba tener los datos suficientes para poder realizar todo el proyecto. Se ha debido de rehacer tareas planificadas debido a la falta de datos, volviendo a hacer las preguntas de que se quiere obtener y que se puede obtener con los datos que se disponen. Debido al tiempo limitado de este proyecto se ha decidido descartar la creación de un panel de control que reflejara los datos para diferentes departamentos. Los impactos ético-sociales se han visto afectados debido a la falta de datos ya que no se ha podido analizar los hábitos de consumo por diferentes países u obtener perfiles de cliente por país. Se ha obtenido un análisis de la perspectiva de género y distribución de los clientes por edades, generación y situación económica y se ha tratado de mantener el género neutro durante el trabajo incluyendo mensaje inclusivo en la descripción de los perfiles de clientes (clientxs). Una solución posible a la falta de datos es la creación de datos nuevos tanto en cantidad mediante nuevas recogidas o datos o creación de datos sintéticos, la cual no se ha podido llevar a cabo por la limitación de tiempo, como en profundidad obteniendo nuevos datos mediante la combinación de datos existentes como el cálculo de las edades, generaciones, tickets de compra, categorías de productos que se han llevado a cabo así como cruzar otro tipo de datos como sociodemográficos de la población por localidades, tasas de paro o delincuencia que permitieran abrir nuevas líneas de análisis. Los datos que se han dispuesto son de los años 1997 y 1998, donde se refleja las dificultades que existían para la recogida de datos. Actualmente, se recogen datos mediante cookies, aplicaciones, donde el cliente acepta las condiciones de uso a cambio de un servicio facilitando así la recogida de datos.

Las líneas de trabajo futuro son las siguientes.

- Creación de un panel de control por departamentos que muestre los datos obtenidos y hacerse nuevas preguntas para establecer las necesidades de obtención de datos como por ejemplo nivel de satisfacción del cliente en diferentes ámbitos, servicio de atención, disponibilidad, estado de la tienda, calidad del producto.
- Realizar clusterización de las tiendas aportando datos sociodemográficos como número de habitantes de la población, tasas de paro, criminalidad, desempleo, renta per cápita, servicios, clima.
- Analítica de establecimientos para desarrollo de plan de remodelación modelando el impacto de la remodelación optimizado el coste-beneficio y modelo de clasificación que indique si la tienda se tiene que remodelar.

Resulta Análisis de necesidades de los modelos para definir y mejorar las estrategias en la recogida de datos que podría implementar un departamento de científicos de datos, controlling o analytics. el conocimiento obtenido de este trabajo refleja las necesidades que requieren los modelos y permite conocer de antemano que datos se deben de recoger para cada análisis. También refleja que se pueden tomar decisiones erróneas en función del nivel de profundidad que se pueda alcanzar en el análisis exploratorio de los datos y la importancia de verificar cada proceso. Para ello sería crucial introducir este departamento en el modelado de proceso de negocio ya que permite obtener una trazabilidad de las acciones llevadas a cabo.

5 Glosario

Cluster: Grupo de datos o elementos similares

Dataset: Conjunto organizado de datos

EDA (Análisis Exploratorio de Datos): Proceso de analizar datasets para resumir sus características principales, a menudo visualmente.

Insight: Conocimiento o entendimiento

Minería de textos: Proceso de transformar texto no estructurado en formato estructurado.

Reglas de asociación: Relación entre un antecedente y un consecuente en base a métricas

6 Bibliografía

- [1] United Nations (s.f.). *Sustainable Development GOALS*. Recuperado 21 de diciembre de 2024 de: <https://www.un.org/sustainabledevelopment/>
- [2] GanttPRO. (s. f.). Recuperado el 21 de diciembre de 2024 de: <https://app.ganttpro.com/#/project/1718524373015/gantt>
- [3] Nowak, S. [Shirly], (2024). ¿Qué es EDA (Exploratory Data Analysis) en Data Science?. Recuperado el 20 de diciembre de 2024 de: <https://nuclio.school/blog/eda-exploratory-data-analysis/>

- [4] Diccionario panhispánico del español jurídico (2023). *Protección de datos*. Recuperado el 20 de diciembre de 2024 de: <https://dpej.rae.es/lema/protecci%C3%B3n-de-datos>
- [5] Your Europe (s. f.). *Reglamento general de protección de datos*. Recuperado el 21 de diciembre de 2024 de: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm#inline-nav-3
- [6] Starita, L. [Laura], (2020). *3 Ways to Embrace Proactive Data Ethics*. Gartner. Recuperado el 21 de diciembre de 2024 de: <https://www.gartner.com/smarterwithgartner/3-ways-to-embrace-proactive-data-ethics>
- [7] datos.gob.es (2017). *La ética en la gestión de los datos*. Gobierno de España. Recuperado el 21 de diciembre de 2024 de: <https://datos.gob.es/es/noticia/la-etica-en-la-gestion-de-los-datos>
- [8] Open data institute (2021). *The Data Ethics Canvas*. Recuperado el 21 de diciembre de 2024 de: https://theodi.org/in_sights/tools/the-data-ethics-canvas-2021/
- [9] Data Scientest (2024). *Text Mining o minería de textos: definición, técnicas, casos de uso*. Recuperado el 21 de diciembre de 2024 de: <https://datascientest.com/es/text-mining-o-mineria-de-textos-definicion-tecnicas-casos-de-uso>
- [10] Stryker, C. [Cole] & Holdsworth, J. [Jim], (2024). *¿Qué es el PLN?*. IBM. Recuperado el 20 de diciembre de 2024 de: <https://www.ibm.com/es-es/topics/natural-language-processing>
- [11] AWS (2024). *¿Qué es el procesamiento de lenguaje natural?*. Recuperado el 20 de diciembre de 2024 de: <https://aws.amazon.com/what-is/nlp/>

- [12] SAS (2023). *Procesamiento del lenguaje natural (NLP). ¿Qué es el procesamiento de lenguaje natural?*. Recuperado el 27 de diciembre de 2024, de https://www.sas.com/es_es/insights/analytics/what-is-natural-language-processing-nlp.html
- [13] O'Brien, K. [Keith] & Downie, A. [Amanda]. (2024). *¿Qué es el análisis de clientes?*. IBM. Recuperado 27 de diciembre de 2024 de: <https://www.ibm.com/es-es/think/topics/customer-analytics>
- [14] Llorens, M. [Mónica]. (s.f.). *Customer Analytics*. Deloitte España. Recuperado el 27 de diciembre de 2024 de: <https://www.deloitte.com/es/es/services/consulting/services/customer-analytics.html>
- [15] IBM (s.f.). *Los juegos que ayudaron a la IA a evolucionar*. Recuperado el 28 de diciembre de 2024 de: <https://www.ibm.com/history/early-games>
- [16] Repsol Global (s.f.). *¿Qué es el machine learning y que usos tiene?*. Recuperado el 28 de diciembre de 2024 de: https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/machine-learning/index.cshtml?gad_source=1&gclid=Cj0KCQiA4L67BhDUARIsADWrI7G3csgS4vDcjGshddS7EPAJLBQOLTSj50GYSJoArA8e9EkIhhQ_Q6QaAuzxEALw_wcB
- [17] AWS (s.f.). *¿Qué es el Aprendizaje mediante refuerzo?*. Recuperado el 28 de diciembre de 2024 de: <https://aws.amazon.com/es/what-is/reinforcement-learning/>
- [18] scikit-learn developers (2024). *2.3. Clustering .1.6.0 documentation*. Recuperado el 29 de diciembre de 2024 de: <https://scikit-learn.org/stable/modules/clustering.html#k-means>

- [19] González, L. [Ligdi]. (2021). *Reglas de Asociación*. Aprende IA. Recuperado el 28 de diciembre de 2024 de: <https://aprendeia.com/reglas-de-asociacion/>
- [20] Munar, P. [Pere]. (2023). *Data Science: predicciones de series temporales con machine learning*. Cyberclick. Recuperado el 3 de enero de 2025 de: <https://www.cyberclick.es/numerical-blog/data-science-predicciones-de-series-temporales-con-machine-learning>
- [21] Facebook Open Source (2023). *Prophet. Forecasting at scale*. Recuperado el 8 de octubre de 2024 de: <https://facebook.github.io/prophet/>
- [22] Herrera, P. [Pablo]. (2018). *Predicción del consumo eléctrico (IV): una aproximación con time-series ARIMA*. Recuperado el 3 de enero de 2025 de: <https://www.energchisquared.com/post/predicci%C3%B3n-del-consumo-el%C3%A9ctrico-iv-una-aproximaci%C3%B3n-con-time-series-arima/>
- [23] Babakr, A. [Abdulaziz]. (2022). *Maven-Market*. Public. GitHub. Recuperado el 30 de septiembre de 2024 de: <https://github.com/abdulazizbabakr/Maven-Market/tree/main>
- [24] databricks (2025). *Jupyter Notebook Guide*. Recuperado el 21 de diciembre de 2024 de: <https://www.databricks.com/glossary/jupyter-notebook>
- [25] Jefatura de Estado. *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)*. DOUE núm. 119 (2016). <https://www.boe.es/buscar/doc.php?id=DOUE-L-2016-80807> [26] BOE-A-2018-

[26] Jefatura del Estado. *Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales*. BOE núm. 294 (2018).
<https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>

[27] Gigandet, S. [Stéphane], (2024). OpenGoodFacts. Recuperado el 20 de diciembre de 2024
de:
<https://es.openfoodfacts.org/>

2.3. *Clustering — scikit-learn 1.6.0 documentation.* (s. f.). Recuperado 29 de diciembre de 2024, de <https://scikit-learn.org/stable/modules/clustering.html#k-means>

3 Ways to Embrace Proactive Data Ethics. (s. f.). Recuperado 21 de diciembre de 2024, de <https://www.gartner.com/smarterwithgartner/3-ways-to-embrace-proactive-data-ethics>

BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. (s. f.). Recuperado 21 de diciembre de 2024, de <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>

Customer Analytics | Deloitte España. (s. f.). Recuperado 27 de diciembre de 2024, de <https://www.deloitte.com/es/es/services/consulting/services/customer-analytics.html>

Data Science: predicciones de series temporales con machine learning. (s. f.). Recuperado 3 de enero de 2025, de <https://www.cyberclick.es/numerical-blog/data-science-predicciones-de-series-temporales-con-machine-learning>

Definición de protección de datos - Diccionario panhispánico del español jurídico - RAE. (s. f.). Recuperado 20 de diciembre de 2024, de <https://dpej.rae.es/lema/protecci%C3%B3n-de-datos>

GanttPRO. (s. f.). Recuperado 21 de diciembre de 2024, de <https://app.ganttpro.com/#/project/1718524373015/gantt>

GitHub - abdulazizbabakr/Maven-Market. (s. f.). Recuperado 21 de diciembre de 2024, de <https://github.com/abdulazizbabakr/Maven-Market/tree/main>

Home - United Nations Sustainable Development. (s. f.). Recuperado 21 de diciembre de 2024, de <https://www.un.org/sustainabledevelopment/>

Jupyter Notebook Guide | Databricks. (s. f.). Recuperado 21 de diciembre de 2024, de <https://www.databricks.com/glossary/jupyter-notebook>

La ética en la gestión de los datos | datos.gob.es. (s. f.). Recuperado 21 de diciembre de 2024, de <https://datos.gob.es/es/noticia/la-etica-en-la-gestion-de-los-datos>

Los juegos que ayudaron a la IA a evolucionar | IBM. (s. f.). Recuperado 28 de diciembre

de 2024, de <https://www.ibm.com/history/early-games>

Predicción del consumo eléctrico (IV): una aproximación con time-series ARIMA ·

Energía Chi-cuadrado. (s. f.). Recuperado 3 de enero de 2025, de
<https://www.energychisquared.com/post/predicci%C3%B3n-del-consumo-electrico-iv-una-aproximaci%C3%B3n-con-time-series-arima/>

Prophet | Forecasting at scale. (s. f.). Recuperado 8 de octubre de 2024, de

<https://facebook.github.io/prophet/>

Protección de Datos conforme al reglamento RGPD - Your Europe. (s. f.). Recuperado

21 de diciembre de 2024, de https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm#inline-nav-3

¿Qué es EDA (Exploratory Data Analysis) en Data Science? - NDS. (s. f.). Recuperado

20 de diciembre de 2024, de <https://nuclio.school/blog/eda-exploratory-data-analysis/>

¿Qué es el análisis de clientes? | IBM. (s. f.). Recuperado 27 de diciembre de 2024, de

<https://www.ibm.com/es-es/think/topics/customer-analytics>

¿Qué es el Aprendizaje mediante refuerzo? - Explicación del Aprendizaje mediante

refuerzo - AWS. (s. f.). Recuperado 28 de diciembre de 2024, de
<https://aws.amazon.com/es/what-is/reinforcement-learning/>

¿Qué es el machine learning y que usos tiene? | Repsol. (s. f.). Recuperado 28 de

diciembre de 2024, de https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/machine-learning/index.cshtml?gad_source=1&gclid=Cj0KCQiA4L67BhDUARIsADWrI7G3csgS4vDcjGshddS7EPAJLBQOLTSj50GYSJoArA8e9EkIhhQ_Q6QaAuzxEALw_wcB

¿Qué es el PLN (procesamiento del lenguaje natural)? | IBM. (s. f.). Recuperado 20 de diciembre de 2024, de <https://www.ibm.com/es-es/topics/natural-language-processing>

¿Qué es el procesamiento de lenguaje natural? - Explicación del procesamiento de lenguaje natural - AWS. (s. f.). Recuperado 20 de diciembre de 2024, de <https://aws.amazon.com/what-is/nlp/>

Qué es y por qué es importante | SAS. (s. f.). Recuperado 27 de diciembre de 2024, de https://www.sas.com/es_es/insights/analytics/what-is-natural-language-processing-nlp.html

Reglamento - 2016/679 - EN - GDPR - EUR-Lex. (s. f.). Recuperado 21 de diciembre de 2024, de <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679>

Reglas de Asociación - Aprende IA. (s. f.). Recuperado 28 de diciembre de 2024, de <https://aprendeia.com/reglas-de-asociacion/>

Text Mining o minería de textos: definición, técnicas, casos de uso. (s. f.). Recuperado 21 de diciembre de 2024, de <https://datascientest.com/es/text-mining-o-mineria-de-textos-definicion-tecnicas-casos-de-uso>

The Data Ethics Canvas | The ODI. (s. f.). Recuperado 21 de diciembre de 2024, de <https://theodi.org/insights/tools/the-data-ethics-canvas-2021/>

¿Qué es el PLN (procesamiento del lenguaje natural)? | IBM. (n.d.). Retrieved December 20, 2024, from <https://www.ibm.com/es-es/topics/natural-language-processing>

¿Qué es el procesamiento de lenguaje natural? - Explicación del procesamiento de lenguaje natural - AWS. (n.d.). Retrieved December 20, 2024, from <https://aws.amazon.com/what-is/nlp/>

Yellowbrick

Yellowbrick: Machine Learning Visualization

<https://www.scikit-yb.org/en/latest/>

<https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html>

upgraded threadpoolctl from version 2.2.0 to version 3.1.0 and this solved the issue

<https://stackoverflow.com/questions/71352354/sklearn-kmeans-is-not-working-as-i-only-get-nontype-object-has-no-attribute>

7 Anexos

Jupyter Notebook TFG