eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Grado: Administración y Dirección de Empresas/
Sarriko

TRABAJO FIN DE GRADO

Desarrollo e Investigación sobre Big Data para una Pyme

Julen Díaz Adrados

Directora:

Dra. Marta Regúlez Castillo

30 de mayo de 2024

Agradecimientos

Quiero mostrar mi más sincero agradecimiento a Marta Regúlez Castillo por su total dedicación durante estos últimos meses el que me ha ayudado a poder estructurar y organizar lo mejor posible este trabajo. No podría haber tenido mejor tutora que ella.

También a mi hermano Daniel Díaz Adrados que brindó un apoyo fundamental a la hora de diseñar la arquitectura del sistema y desarrollo del mismo. Sin su ayuda todo hubiera sido más difícil.

Resumen

En este trabajo se explica, desde una perspectiva teórica y práctica, cómo se digitaliza y se desarrolla un ecosistema Big Data para una pyme. La transformación a la era digital suele ser costosa y difícil de aplicar para las pymes, por ello, este trabajo pretende ser una guía para aquellas pequeñas empresas que quieran dar el salto a lo digital. Se comenzará explicando que es el Big Data y la situación actual en España y Europa para así más adelante poder realizar una simulación de cómo se realizaría la transformación digital y todo el desarrollo de un ecosistema Big Data.

Palabras Clave

Big Data, digitalización, cloud, pyme, ETL, Wordpress, escalabilidad, CRM, python, Power BI, CRISP-DM

Abstract

This present work shows the explanatory and practical concept of business digitalization for an SME and how a Big Data ecosystem is developed. The transformation to the digital era is usually expensive and difficult to apply for SMEs, therefore this work will be a guide for those small companies that want to make the leap to digital. We will begin by explaining what big data is and what the current panorama is like in Spain and Europe so that later we can carry out a simulation of how to carry out the digital transformation and all the development of a Big Data ecosystem.

Key Words

Big Data, digitalización, cloud, pyme, ETL, Wordpress, escalabilidad, CRM, python, Power BI, CRISP-DM

Índice de contenido

Índice de tablas

Índice de figuras

Introducción

Hoy en día es fundamental para una empresa tener presencia en internet ya que se ha convertido en una herramienta imprescindible para interactuar con el cliente y obtener datos para la toma de decisiones, donde el consumidor es a la vez usuario y proveedor de información útil para la empresa. La gestión de este proceso implica un coste muy alto para muchas empresas al no contar con personal especializado en ello o una capacidad financiera fuerte como para poder subcontratar a otra empresa. Por este motivo, muchos negocios no llegan a digitalizarse. El 68,17 % de las empresas con menos de 10 empleados y el 21,51 % de las que tienen entre 10-50 trabajadores todavía no disponen de página web empresarial (Instituto Nacional de Estadística, 2023).

Por otro lado, tenemos el denominado "Big Data". Esta tecnología trata de gestionar y analizar grandes volúmenes de datos para extraer información valiosa y ayudar en la toma de decisiones empresariales. Hasta la fecha son muy pocas las empresas que han implementado un sistema de Big Data. La Unión Europea ha marcado como objetivo que el 75% de las empresas cuenten con un sistema de Big Data para 2030 y España se ha marcado como objetivo que para 2025 el 25% de las empresas cuenten con este sistema (Observatorio Nacional de Tecnología y Sociedad, 2023). Actualmente en España, el 20,8% de las medianas empresas y el 11,9% de las pequeñas empresas cuentan con un sistema de Big data (Instituto Nacional de Estadística, 2023).

Este Trabajo de Fin de Grado trata sobre la digitalización e implementación de sistemas de Big Data en pymes (pequeñas y medianas empresas). Con este estudio se quiere mostrar, por medio de una empresa ficticia o simulación, el proceso y el coste de digitalizar un negocio, así como la creación de un sistema de Big Data y análisis de datos. Este proceso, aportaría a la empresa una mayor eficiencia operativa, mejora en la toma de decisiones y mayor facilidad a la hora de identificar nuevas oportunidades. Todos los servicios que se utilicen para desarrollar la investigación serán de coste gratuito o de coste muy bajo, con el fin de que cualquier empresa pueda realizarla.

Los objetivos marcados para este trabajo son comprender mejor lo que significa la digitalización empresarial, el Big Data y cómo se utilizan los datos para dar un valor añadido a la empresa. Mediante la simulación se van a utilizar herramientas y plataformas digitales que serán muy útiles para las empresas.

En la realización de este trabajo se ha decidido utilizar un método experimental que implica una simulación de una transformación digital y la implementación de un sistema Big Data. Para ello se ha simulado un negocio digital que vende móviles modulares eco-friendly. A lo largo del estudio, a la vez que se realiza esta simulación se va a investigar y explicar todos los procesos involucrados. Aunque el tema principal de este trabajo no sea realizar un negocio próspero se intentará que sea lo más realista posible, teniendo en cuenta las limitaciones existentes.

El trabajo se estructura de la siguiente manera. En la primera sección se contextualiza el Big Data, en la que se definirá su concepto y se comprenderá cuales son las necesidades de realizar este trabajo. La segunda sección introduce la digitalización empresarial y cuales son los pasos a seguir para realizarla. Posteriormente, en la tercera sección, se explicará una metodología para llevar a cabo el sistema Big Data. Seguido, en la cuarta sección, se intentará cuantificar los costes de las herramientas digitales utilizadas. Por último, se mostrarán las conclusiones y futuras líneas de trabajo. Al final del trabajo se recoge el apéndice y los anexos donde se recogen las aplicaciones utilizadas en la digitalización, un glosario de términos y los programas en código Python.

1. Contextualización del Big Data

1.1 Definición de Big Data

¿Qué es el Big Data? Para contestar a esta pregunta, lo primero que se necesita es comprender cuales son los orígenes y las diferentes definiciones que se le da al Big Data (macrodatos o inteligencia de datos en español). Sin embargo, hoy en día el concepto de "Big Data" es confuso y tiene orígenes inciertos (Gandomi y Haider, 2015). Diebold (2012) cree que en la década de 1990 se originó el término durante algunas conversaciones en Silicon Graphics Inc. (SGI), en las que John Mashey, Doctor en Ciencias de la Computación, fue esencial para su concepción. Actualmente, con el auge del Big Data, el origen se le puede atribuir a IBM y otras empresas tecnológicas líderes que invirtieron en la creación de un nicho de mercado de análisis de datos.

Ahora bien, ¿Qué es el Big Data? En esencia, se trata de una cantidad masiva y compleja de información que resulta prácticamente imposible de usar sin herramientas informáticas (Ruiz, 2022). Cuando hablamos de Big Data, en realidad nos estamos refiriendo a un proceso de análisis de datos con el fin de extraer información valiosa para nuestro negocio.

La Fundación TechAmerica define Big Data de la siguiente manera: "Big Data es un término que describe grandes volúmenes de datos variables, complejos y de alta velocidad que requieren técnicas y tecnologías avanzadas para permitir la captura, almacenamiento, distribución, gestión y análisis de la información" (TechAmerica Foundation's Federal Big Data Commission, 2012).

Otra manera de definir el Big Data nació de IBM y Gartner que plantearon el Big Data como un modelo de tres dimensiones conocido como "Las tres V" en referencia a Volumen, Velocidad y Variedad (Gartner, s.f.).

La primera de ellas, el *Volumen*, hace referencia a la cantidad de datos que existen y se generan constantemente, ya sean dentro de la empresas o fuera como datos de ventas, compras, inversiones, datos de las redes sociales, datos que se generan mediante sensores, cualquier dato que pueda ser de utilidad para la empresa. En 2020, se ha

alcanzado la cifra de 25 mil millones de *endpoints*[1] y de ellos se ha calculado un volumen de 40 mil millones de GB (UNIR, 2020), un volumen demasiado grande para poder almacenar de una manera tradicional (hasta hace unos años, lo más común era almacenar los datos en discos duros propios de la empresa) ya que las empresas no cuentan con una infraestructura tan grande.

La segunda de ellas, la *Velocidad*, hace referencia a la rapidez con la que se obtienen, se procesan y se almacenan los datos. Actualmente es fundamental la velocidad con la que se transmiten los datos, por ello los sistemas deben de ser capaces de acceder, almacenar y procesar estos flujos de datos a una velocidad casi instantánea y en tiempo real. Wal-Mart, por ejemplo, procesa más de un millón de transacciones por hora (Cukier, 2010).

La última es la *Variedad*, en los sistemas Big Data se tratan todo tipo de datos, estos pueden ser de diferentes orígenes o fuentes (móviles, sensores, archivos, base de datos, webs, redes sociales…), diferentes tipos de formato como por ejemplo numéricos, imágenes, textos. *Variedad* también hace referencia a los diferentes tipos de estructura de datos, que pueden ser estructurados, semiestructurados o no estructurados. La clasificación más importante es esta última, ya que de ella dependerá como se debe almacenar ese tipo de datos.

Antes de la llegada del Big Data, los datos estructurados se organizaban en bases de datos relacionales[2] (por ejemplo Microsoft Access) o hojas de cálculo, que constituyen el 5% de todos los datos existentes (Cukier, 2010). Los datos estructurados son más fáciles de procesar y almacenar que los no estructurados, estos últimos no se podían almacenar de una manera eficiente ya que no siguen ningún tipo de patrón o clasificación (OCI Oracle Cloud Infrastructure, s.f.). Con la llegada del Big Data, se realizaron innovaciones tecnológicas que lograron que los datos de tipo no estructurados o semiestructurados pudieran ser almacenados (Datas Lakes). Los datos no-estructurados son los de formato videos, texto o audios, los cuales requieren un procesamiento previo para poder obtener significado de ellos. Los datos no estructurados, especialmente los datos en formato de vídeo, son el mayor componente del Big Data (Gandomi y Haider, 2015). Los datos semiestructurados son por ejemplo PDFs, archivos que contienen información pero no está

[1]Endpoints significa cualquier punto que sea la parte final de una red (Móviles, PC, sensores…), dispositivos de los cuales se puede obtener datos.
[2]Las bases de datos relacionales se basan en el modelo relacional, una forma intuitiva y directa de representar datos en tablas (OCI Oracle Cloud Infrastructure, s.f.).

clasificada de ninguna manera por lo cual para poder almacenarla en una base de datos se necesita un proceso parecido al de los datos no estructurados.

Si se quiere utilizar eficazmente el Big Data, se requiere utilizar Data Lakes (Lago de datos). Un Data Lake es una estrategia de almacenamiento donde se almacenan conjuntamente cualquier tipo de dato, ya sea datos estructurados, semi-estructurados y no estructurados. En lugar de realizar un esquema y requisitos de datos para su almacenamiento, se utilizan herramientas para asignar identificadores únicos a los elementos de datos, de modo que solo se consulte un subconjunto de datos relevantes para analizar. Esto quiere decir que se pueden almacenar todos los datos sin un diseño cuidadoso y sin saber el porqué exacto de su almacenamiento (BasuMallick, 2022). Una vez se consiga tratar, limpiar y analizar los datos, se podrán incluir en Warehouses (Base de datos) en donde se conservarán de una manera estructurada.

Un sistema Big Data debe de poder dar solución a estas tres variables, almacenando y procesando de información, siendo capaz de tratar diferentes tipos de datos a tiempo real. En la actualidad se han definido otras dimensiones que deben tenerse en consideración, además de las anteriormente citadas, tales como Veracidad, Volatilidad, Valor, Viabilidad y Visualización. En cualquier caso, el Big Data debe servir para proporcionar un valor añadido a los datos mediante una manipulación y almacenaje correcto de los mismos.

En conclusión, hay diferentes definiciones, por lo que es difícil saber cuál es la definición exacta de Big Data. Por ello, aún está por desarrollarse una comprensión coherente del concepto y su nomenclatura. Lo que queda claro es que es un proceso en el que se logra procesar una inmensa cantidad de datos mediante el uso de herramientas que trabajan conjuntamente y permiten captar, almacenar y gestionar los datos para la toma de decisiones.

1.2. Cómo funciona el Big Data

Para entender cómo funciona un sistema Big Data, en esta sección se van a explicar tres procesos clave: Integración, Gestión y Análisis (OCI Oracle Cloud Infrastructure *, s.f.)*.

1.2.1. Integración

Como hemos visto anteriormente, un sistema Big data concentra datos de numerosas fuentes y aplicaciones distintas. La integración es la encargada de combinar la información de todas las fuentes de datos. Para lograrlo, se utilizan mecanismos de integración de datos, como "extraer, transformar y cargar" (extract, transform, load (ETL[3])) o también el mecanismo de "extraer, cargar y transformar" (extract, load, transform(ELT)), este último se utiliza para los datos no estructurados. Estos mecanismos son ideales para abordar el reto de la integración de datos ya que se logra extraer los datos de todas las fuentes que se necesiten, se les da el tratamiento necesario para poder ser almacenados.

La idea es integrar todos los datos de las diferentes fuentes en un data warehouse (almacén o repositorio de datos) que puede ser local o externo o incluso una combinación de ambos. Es importante que la integración sea capaz de mapear[4] los registros de los datos incluidos, esto quiere decir que sea capaz de que si en una fuente de datos hay campos con la etiqueta de "nombre" y en otra "nom" sea capaz de almacenar esos registros de una manera conjunta. La integración es fundamental para el Big Data, sin datos no se puede realizar análisis, por ello se tiene que asegurar que se cuente con todos los datos necesarios.

Para una correcta integración, se necesita incorporar los datos, procesarlos y formatearlos, y deben estar disponibles de tal forma que los analistas de datos puedan empezar a utilizarlos para lo que necesiten. En la sección 3, se explicará más en profundidad cómo tratar los datos y qué metodología se va a utilizar.

[3]Los ETLs son técnicas de extracción de datos que se explicarán más adelante.
[4]Proceso en el que hay que hacer coincidir campos de datos de una fuente con campos de datos de otra fuente. Este proceso ayuda a garantizar la correcta transferencia de datos sin perder coherencia.

1.2.2. Gestión

El Big data requiere una gran capacidad de almacenamiento. Esta necesidad puede ser solventada o bien mediante una infraestructura local o mediante el almacenamiento en la nube. Si se requiere de un almacenamiento masivo de datos la opción preferible es el almacenamiento en nube. Este servicio informático lo ofrecen diferentes empresas como son Google, Microsoft o Amazon. El almacenamiento en la nube (Cloud) es uno de los muchos servicios que ofrece Cloud Computing. En 2020, los datos almacenados a nivel mundial alcanzaron los 6,7 zetabytes (Roa, 2021). A consecuencia del incremento de la cantidad de datos que se crean o almacenan cada año, las empresas requieren de mayor capacidad de almacenamiento.

El almacenamiento en nube permite almacenar los datos de la forma que se desee, incorporando los requisitos de procesamiento de la preferencia de cada usuario y el procesamiento necesarios de los conjuntos de datos para on-demand[5], esto quiere decir que ofrecen al usuario la posibilidad de acceder a contenidos multimedia en el momento exacto en que lo desee. La nube está aumentando progresivamente su demanda porque sirve de apoyo para otras tecnologías actuales como el Big Data y proporciona utilidad y accesibilidad a los usuarios, es decir, que accedan a una utilidad cuando lo necesiten y desde donde sea que lo necesiten (Buyya et al., 2009).

Dependiendo de los requerimientos de la empresa hay varias opciones de contratación de Cloud Computing. Según Torres i Viñals (2012) hay tres niveles: *IaaS* con los servicios básicos de infraestructura, *PaaS* que incluye tanto el sistema operativo como los softwares intermedios para el desarrollo de softwares y finalmente *SaaS* que ofrece un servicio totalmente funcional con todos los servicios posibles. No hay un nivel mejor que otro, siempre dependerá de las necesidades de cada empresa. La figura 1.1 ilustra las diferencias de servicios que se ofrecen entre los tres niveles. A continuación se explica brevemente cada uno de ellos ilustrado con algún ejemplo:

a) Software como servicio (SaaS): En este caso, la empresa cloud ofrece la infraestructura hardware y los entornos necesarios para ejecutar la aplicación desde un determinado portal o interfaz web. Esto quiere decir que el servicio que se presta es de una “aplicación”, “herramienta” o “plataforma” ya totalmente funcional la cual no puede ser modificada como por ejemplo Gmail, Hotmail o Drive. En Drive tenemos un conglomerado de aplicaciones que se pueden usar como lo es las Hojas

[5]Actividad económica creada por las empresas de tecnología que satisfacen la demanda de los consumidores a través de la provisión inmediata de bienes y servicios (Escobar, 2016).

de Cálculo de Google, Documentos Google, Diapositivas Google… Algunas de estas aplicaciones se utilizarán para el desarrollo de este trabajo, sobre todo la Hoja de Cálculo de Google.

b) Plataforma como servicio (PaaS): Las empresas cloud se encargan de ofrecer la infraestructura, las herramientas de desarrollo y middleware (lógica de intercambio de información entre aplicaciones), gracias a esto, las empresas clientes se les facilita la construcción y puesta en funcionamiento sus aplicaciones y el desarrollo de estas. En otras palabras, PaaS presenta al usuario un lugar virtual donde poder desarrollar, ejecutar y gestionar sus aplicaciones sin la necesidad de preocuparse de construir y mantener la infraestructura asociada con el proceso de desarrollar y lanzar la aplicación. Ejemplos de PaaS son AWS Elastic, Beanstalk o Heroku.

c) Infraestructura como servicio (IaaS): Tal como indica su nombre, ofrece servicios de infraestructura como pueden ser computación y almacenamiento, de tal manera que se puede disponer de recursos informáticos como la memoria de almacenamiento, RAM, procesadores o equipamientos de red. Gracias a este servicio, las empresas no necesitan invertir en infraestructura informática, en vez de eso, alquilan los recursos de hardware y pagan por su utilización. Esto quiere decir que las empresas pueden ir variando el consumo de los recursos en función de sus necesidades y sus capacidades, a esto se le conoce como elasticidad de la infraestructura (Torres i Viñals, 2012). Esto es una gran ventaja actual para las empresas que no puedan invertir en inmovilizado tecnológico muy costoso. En la actualidad existen empresas con un gran despliegue de infraestructura de servidores como lo son AWS, Microsoft Azure y Google Cloud que son los líderes del sector.

Más adelante, en la sección de la creación del sistema de Big Data, se explicará en detalle cuál de estos tres servicios se utiliza para cada etapa del trabajo y, en concreto, para qué se usa dentro del ecosistema.

Figura 1.1: Diferencias de servicios en la nube

SERVICIOS QUE OFRECE CADA TIPO DE CLOUD	IaaS	Paas	Saas
APLICACIONES	✗	✗	✓
DATOS	✗	✗	✓
SOFTWARE INTERMEDIO (PLATAFORMAS)	✗	✓	✓
SISTEMA OPERATIVO	✗	✓	✓
SERVIDORES	✓	✓	✓
REDES	✓	✓	✓

Fuente: Elaboración propia. Adaptado de Red Hat, Diferencias entre IaaS, PaaS y SaaS

1.2.3 Análisis

La acción de analizar se realizará una vez se tenga la información almacenada y limpiada correctamente. Los profesionales que se encargan de este paso son los analistas de negocios. Su labor consiste en analizar la información ya integrada utilizando diferentes herramientas como puede ser Power BI (ver apéndice Tabla A.6). Esta aplicación permite realizar análisis visuales con mayor claridad mediante reportes dinámicos. Esta fase debe conseguir que la información sea intuitiva de utilizar para que cualquier directivo de la empresa pueda entender la información y poder tomar las decisiones necesarias. Depende que tipo de análisis se use, se conseguirán respuestas a diferentes preguntas.

Entre los tipos de modelos de análisis de datos se encuentran los siguientes: prescriptivos, predictivos, diagnósticos y descriptivos (Losada, 2022). Para cada negocio, se implementan los modelos más adecuados para ayudar a lograr los objetivos aportando información relevante y una mejor toma de decisiones. A continuación se da una breve explicación de cada uno de estos modelos excepto el análisis predictivo que se explicara más en detalle:

Análisis prescriptivo: Se utiliza para recomendar qué acciones o decisiones realizar basándose en los datos disponibles y el objetivo empresarial. Este tipo de análisis no sólo predice el futuro sino también orienta en las pautas que hay que realizar para lograr los resultados deseados. Los modelos utilizados necesitan datos actuales y algoritmos avanzados (Gibson, 2018).

Análisis descriptivo: Este análisis sirve para describir los datos existentes en la base de datos de una manera concisa y significativa, dando una razón de por qué la empresa se encuentra en esa situación. Este tipo de análisis se suele realizar para comprender mejor la situación de la empresa o del proyecto, intentando apreciar posibles patrones que influyen considerablemente en el resultado. Para ello se suelen utilizar tablas, gráficas y estadísticos descriptivos como la moda, la mediana, etc.

Análisis de tipo diagnóstico: También conocidos como causales, este tipo de análisis sirve para identificar las causas de algún problema o anomalía en los datos. Los modelos utilizados, ayudan a comprender cuál ha sido el motivo de no conseguir los resultados esperados. Se utilizan pruebas de hipótesis y análisis de la varianza para poder identificar las variables que han causado efecto en los resultados.

Análisis predictivos: Estos análisis utilizan datos históricos y actuales para intentar predecir el futuro. Se necesita utilizar algoritmos de aprendizaje supervisado para identificar los patrones de los datos históricos y así poder hacer predicciones sobre datos nuevos. Las técnicas tradicionales de análisis predictivo se basan principalmente en métodos estadísticos, por ejemplo se utilizan técnicas de regresión lineal, árboles de decisión, redes neuronales o regresión logística.

Con la llegada del Big Data, fue necesario el desarrollo de nuevos métodos estadísticos por tres razones principales (Gandomi y Haider, 2015). En primer lugar, el Big Data cuenta con unas características que dificulta la utilización de técnicas convencionales: heterogeneidad, acumulación de ruido, correlaciones espurias (relación matemática en la que dos o más eventos o variables están asociados pero no relacionados causalmente) y endogeneidad incidental (Fan, Han, y Liu, 2014).

En segundo lugar, los métodos estadísticos convencionales se basan en una pequeña muestra de la población y el resultado se compara con el azar para examinar la importancia de una relación particular. Luego, la conclusión obtenida se generaliza a toda la

población. En el caso del Big Data, las muestras son masivas y representan a la mayoría de la población, por ello, la noción de significancia estadística no es tan relevante.

Por último, en tercer lugar, muchos métodos convencionales para muestras pequeñas no se pueden adaptar a la enorme cantidad de datos que se utilizan en el Big Data. Más adelante se entrará más en detalle de las nuevas técnicas que se utilizan para este tipo de análisis.

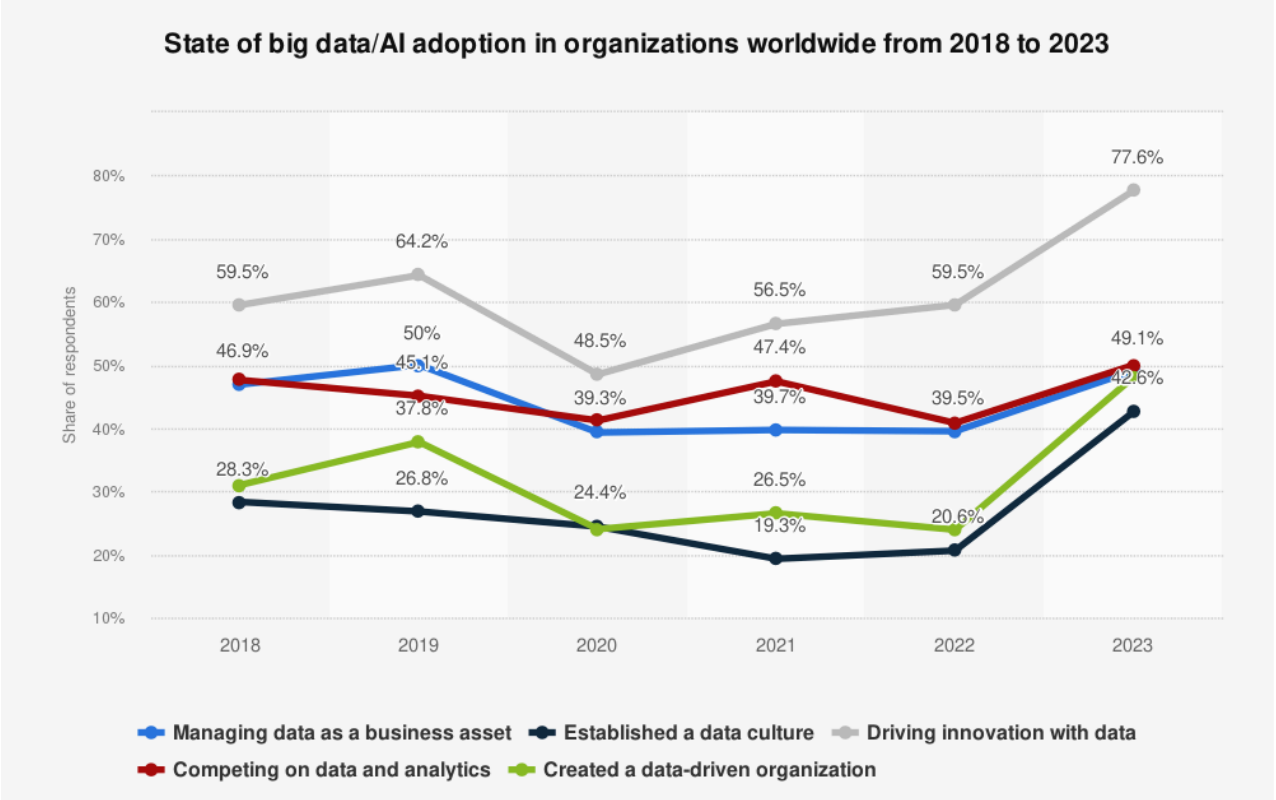
En resumen, de estos tres procesos, un sistema de Big Data tiene que dar una solución eficaz para el procesamiento global de los datos. Se debe dar respuestas a cómo se van a conseguir, almacenar y analizar los datos.

1.3 Utilización del Big data en la actualidad

En la actualidad, las grandes empresas ya han desarrollado o están desarrollando este tipo de sistemas, ya que entienden la importancia que tiene el saber interpretar correctamente los datos. El Big Data da respuestas a preguntas que ni siquiera sabemos que existen, por ello si se obtiene esa información se va a lograr impulsar los diferentes negocios. Para 2027, se espera un volumen mundial de ingresos de 103 mil millones de dólares según Statista (SiliconANGLE, 2018).

A nivel mundial la evolución de la utilización del Big Data ha sido muy positiva, durante los últimos años las empresas se han concienciado de la importancia que tienen los datos en la actualidad. En la figura 1.2, se puede observar la proporción de empresas que han implementado el Big Data o IA (Inteligencia Artificial) dentro de su organización.

Figura 1.2: Estado de la adopción de Big Data/IA en organizaciones de todo el mundo de 2018 a 2023



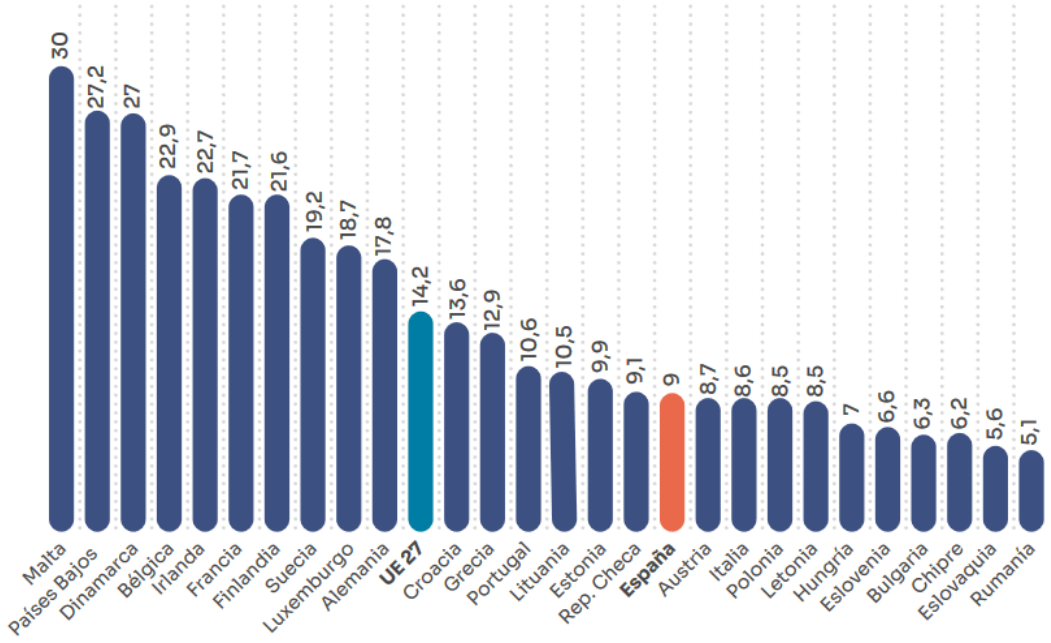
Fuente: Statista

Nota: Esta gráfica contiene el porcentaje de respuestas de compañías de todo el mundo. Directores de datos, directores de datos y análisis o jefes de datos, análisis o inteligencia artificial de más de 100 organizaciones de Fortune 1000 y organizaciones líderes a nivel mundial

En términos generales de la encuesta, los datos son positivos. Los puntos a destacar son la línea gris y verde. La línea gris indica las empresas que han impulsado la innovación con datos, en 2023 en 77% de las empresas ha afirmado haber utilizado los datos como herramienta de innovación. La línea verde que son las empresas que siguen una organización basada en los datos, se puede observar un crecimiento drástico en 2023 llegando al 42.6% de las organizaciones que afirmaron haberse convertido en empresas basadas en datos (NewVantage Partners, y Wavestone, 2023).

Como ya se ha comentado anteriormente, el desarrollo de estos sistemas en empresas pymes en España es muy bajo y más comparado con el resto de Europa, ocupando un puesto bajo en la utilización del Big Data. Como muestra la Figura 1.3, el porcentaje de empresas españolas que utilizan Big Data es de un 9%, situándose por detrás de la Unión Europea (UE 27) con un 14,2% (Observatorio Nacional de Tecnología y Sociedad, 2023).

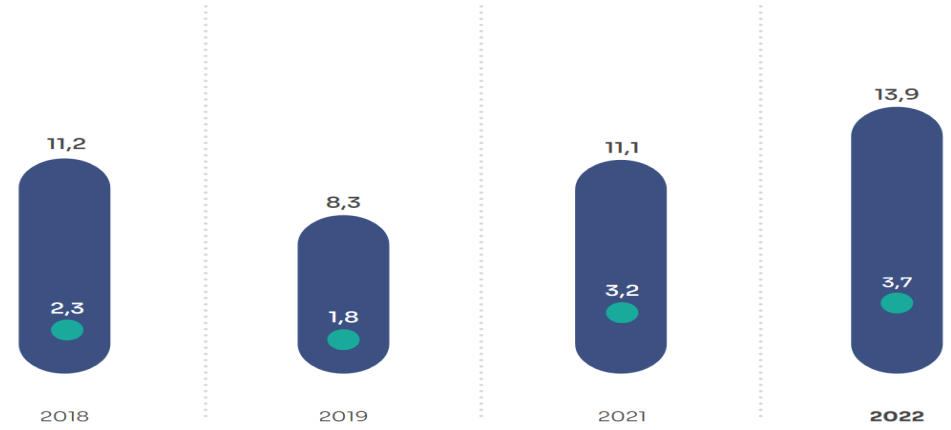
Figura 1.3: Porcentaje de empresas que utilizan Big Data en la UE (2020)



Fuente: Observatorio Nacional de Tecnología y Sociedad (2023).

Como se puede observar en la Figura 1.4, que distingue entre empresas con más de 10 trabajadores y microempresas (menos de 10 trabajadores, color turquesa), desde el 2018 al 2019 se observó un descenso en el uso del Big Data por parte de las empresas españolas pasando del 11,2% al 8,3%. Pero a partir del 2019 se muestra una tendencia ascendente, llegando hasta el 13,9%. (Observatorio Nacional de Tecnología y Sociedad, 2023).

Figura 1.4: Porcentaje de empresas que analizan Big Data en España (2018 - 2022)



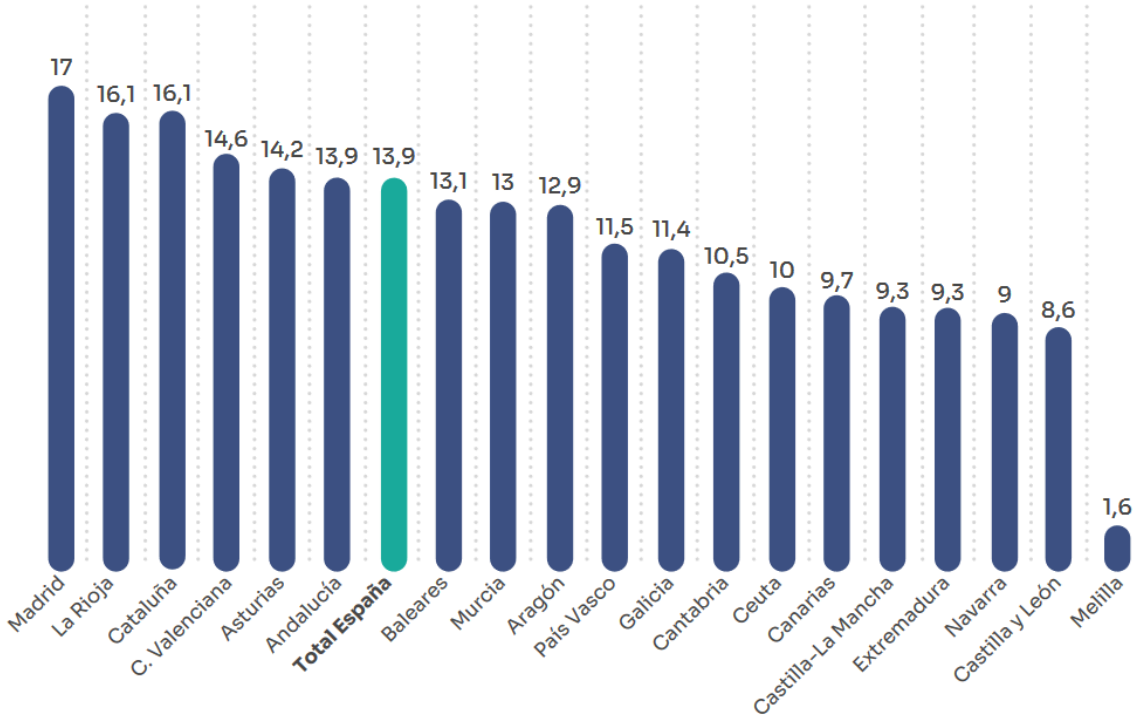
Fuente: Observatorio Nacional de Tecnología y Sociedad (2023).

Nota: En azul oscuro se representan las empresas con 10 o más trabajadores y en turquesa las microempresas con menos de 10.

Si analizamos el porcentaje de empresas que utilizan Big Data por comunidad autónoma, como se puede apreciar en la Figura 1.5, la comunidad de Madrid es la que más destaca por el uso del Big Data con un 17%. La comunidad autónoma del País Vasco se sitúa con un 11,5% por debajo de la media española que es de un 13,9% (Observatorio Nacional de Tecnología y Sociedad, 2023).

El gobierno de España, con la financiación de la Unión Europea, ha puesto en marcha el programa de ayudas "Kit Digital" (Acelera pyme, s.f.) que tiene por objeto la concesión de ayudas a pymes y microempresas para la adopción de soluciones de digitalización disponibles en el mercado. Se pueden percibir ayudas de hasta 12.000 euros. Por ello es un buen momento para unirse a la era digital.

Figura 1.5: Porcentaje de empresas que utilizan Big Data por comunidad autónoma (2022)



Fuente: (Observatorio Nacional de Tecnología y Sociedad, 2023)

Una vez contextualizado lo que es el Big Data y su papel dentro de la empresa, se puede dar paso a la siguiente sección, en la cual se explicará cómo las empresas pueden integrar estas tecnologías en sus procesos. Esta sección estará centrada en la digitalización empresarial, un paso crucial en la evolución de las organizaciones modernas. Se analizará cómo la adopción de tecnologías digitales no solo facilita la implementación de Big Data, sino que también transforma integralmente los modelos de negocio, las estructuras organizativas y las estrategias de mercado.

2. Digitalización Empresarial

Digitalizar una empresa implica el procedimiento mediante el cual ésta incorpora herramientas, tecnología y sistemas digitales con el propósito de ofrecer un valor superior a sus clientes, así como generar nuevas experiencias, soluciones y modelos de negocio (Ramírez, 2023). La definición que le dan Fresnadillo y López (2018) es “La digitalización es aquel proceso mediante el cual una empresa crea un perfil digital para tener presencia en Internet, siendo este perfil la plataforma de su actividad productiva o fiel reflejo de la tienda física”(p. 5).

El proceso de digitalizar una empresa, aporta numerosas ventajas. Por un lado se consigue automatizar procesos y optimizar los recursos necesarios gracias a la utilización de diversas herramientas tecnológicas facilitando y mejorando la eficiencia de los trabajadores.

Por otro lado, con la digitalización se abren nuevos canales de distribución, como los e-commerce o los marketplaces. La distribución en una empresa digital suele ser omnicanal, es decir, al cliente se le ofrecen varios canales entre los que puede elegir por donde realizar la compra, incluso empezar la compra en un canal y terminarla en otro. Pero con la digitalización no solo se consiguen nuevos canales de ventas sino que también de comunicación, llegando al público objetivo por múltiples formas como redes sociales, emails o apps.

En el caso de este trabajo se van a realizar las acciones necesarias para poder crear un conglomerado de herramientas y plataformas para habilitar un sistema Big Data. Para comenzar se va a explicar cómo se realiza una creación de una página web y una implementación de un plataforma de gestión de las relaciones con los clientes.

2.1 Creación de página web

En esta sección se aborda el tema de cómo crear la página web para una empresa que quiera vender productos en internet (E-commerce)[6]. Puede ser el caso en el que una empresa solo desee una página web para fines de comunicación pero como el caso del

[6] El Comercio electrónico es el proceso en el que tanto venta, pedido y pago de productos o servicios se realiza por internet.

E-commerce es el más complejo se va desarrollar este último. Este proceso es la parte principal de la transformación digital, en la que una organización integra tecnología digital en el core del negocio de la empresa. En el momento que se tenga la página web totalmente funcional, se podrán realizar compras, reservas, petición de información o cualquier servicio de la empresa desde esta página web.

A continuación, se procede a explicar los pasos realizados en la creación de la página web y cómo tenerla totalmente funcional para una empresa real.

2.1.1 Servidor de la página web y dominio

A la hora de crear una página web, lo primero de todo hay que contratar un servidor donde se aloje nuestra página web. Para el estudio he escogido a la empresa Hostinger (ver Apéndice Tabla A.1), conocida por ser una opción económica y popular de proveedor de alojamiento web. El coste de uso de su servidor ronda los 2,5 euros mensuales e incluye un dominio de páginas web gratuito.

Una vez contratado el servicio, elegimos un dominio para la página web. El nombre elegido para la empresa estuvo condicionado por los dominios libres que habían, ya que no se puede usar un dominio que otra empresa ya esté utilizando, como por ejemplo “Iphone.com”, por lo que hay que elegir un dominio disponible. Para este trabajo, después de barajar las distintas posibilidades y que tuviera que ver con móviles eco-friendly, la elección final fue greentechphone.com

2.1.2 Herramientas para la creación de páginas web comerciales

A la hora de elaborar una página web desde cero, hay varias opciones de creadores de páginas web. Para saber cuál utilizar es necesario conocer las necesidades de tu organización y dependiendo de ello se escoge un creador u otro. Por un lado están las aplicaciones que no requieren muchos conocimientos técnicos para su funcionamiento como “Weebly”, “Shopify” o “Wix” los cuales son fáciles de utilizar e intuitivos. Si no se desea una escalabilidad o una gestión personal del sitio web son una buena opción.

Por otro lado, la plataforma Wordpress (ver apéndice Tabla A.1) permite al usuario crear y desarrollar sitios web sin la necesidad de un gran nivel de programación (Código abierto), sencilla, adaptable y profesional. Wordpress es mundialmente conocida y utilizada por muchas empresas en la actualidad, superando el 15% de la cuota de mercado en el sector E-commerce (Impulsa Ecommerce, 2024). Esta plataforma cuenta con la posibilidad de instalar diferentes plugins[7] y lo mejor de todo es que es de uso gratuito, el único coste que tiene es la necesidad de un servidor. Para realizar el trabajo he optado por utilizar esta plataforma, al ser actualmente el creador de páginas web de mayor escalabilidad y adaptabilidad.

También he utilizado los siguientes plugins: WooCommerce, Monster Insight, Yoast SEO, GTM4WP y Blocksy (ver apéndice Tabla A.1). Los más importantes son WooCommerce y GTM4WP.

2.1.2 Desarrollo de la página web

Para comenzar a desarrollar la página web desde Wordpress, se necesitará utilizar WooCommerce (ver apéndice Tabla A.1), un plugin de comercio electrónico para Wordpress que permite crear tiendas digitales. Este plugin, permite usar plantillas prediseñadas para incorporar a la página web y de esta forma no se necesitará programar mediante código la mayor parte de la página. Por ello para el caso de este estudio, se ha utilizado una de las plantillas para facilitar y agilizar la creación de la página web , solo se tuvo que editar alguna sección e incluir algunos subapartados.

La plantilla utilizada contaba con una página principal y 4 subcategorías de Productos, Blogs, Carrito y Contacto. Una vez implementada la plantilla, se modificó la página principal incluyendo diferentes imágenes de móviles[8], editando texto e incorporando diferentes enlaces dentro de la página gracias al plugin de Blocksy el cual me permite editar el sitio web sin la necesidad de tener muchos conocimientos de programación. Este plugin lo he utilizado para la edición tanto de la página principal como blogs y productos del sitio web. El resultado final de la página principal se puede observar en la figura 2.3.

Una vez confeccionada la página principal, se procede a elaborar la plantilla de la página de los productos. Para este trabajo, se va a simular que se venden diferentes productos que contaría una empresa de venta al consumidor de teléfonos móviles. En

[7]Un plugin es una aplicación que permite extender las funciones de otra aplicación o programa sin tener que modificar el código.
[8]Todas las imágenes han sido creadas con la ayuda de la inteligencia artificial de Microsoft Bing.

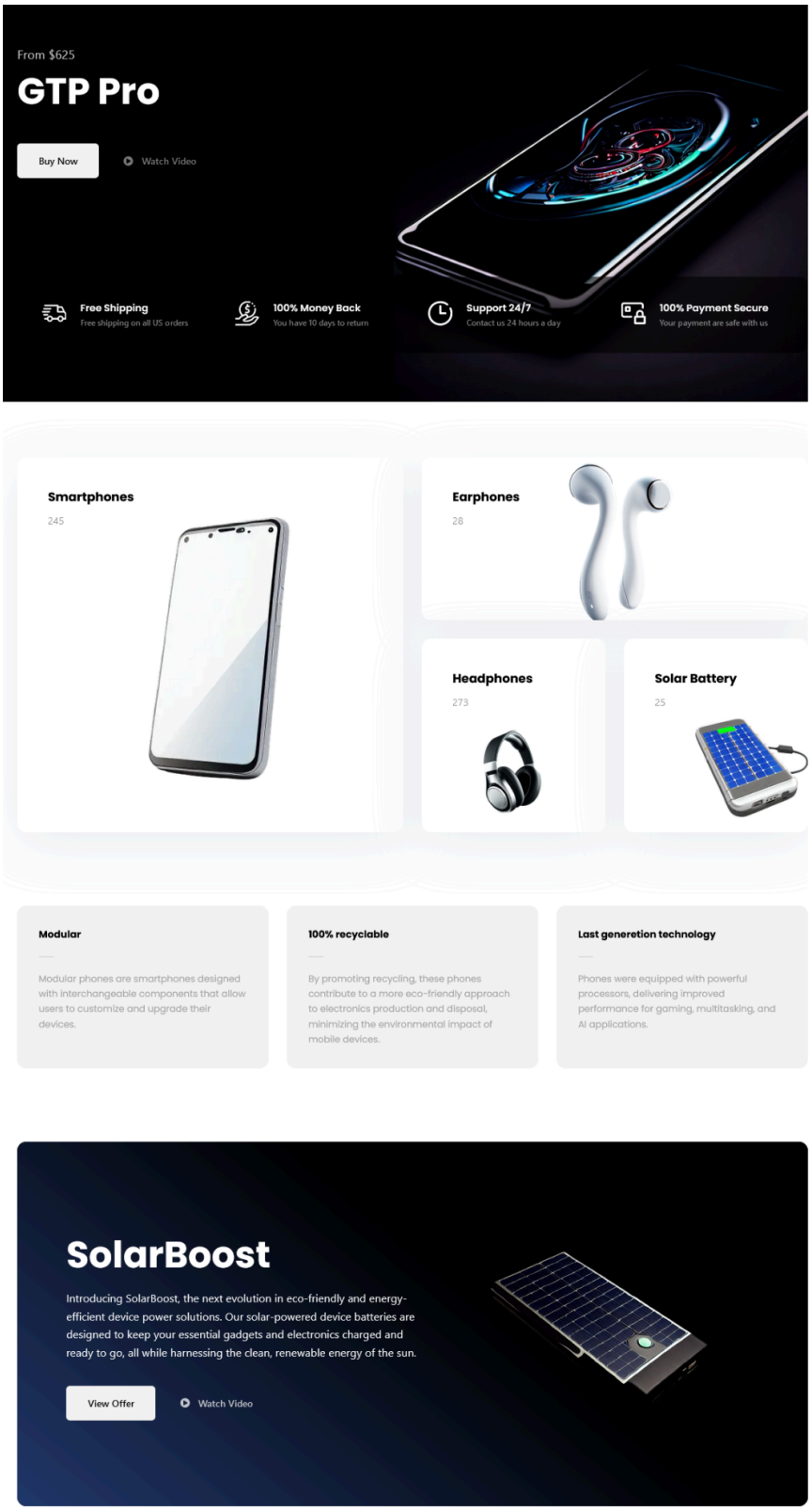
primer lugar se desarrolla la idea de dos móviles ecofriendly “GTP Pro” y “GTP Lite”. A continuación se incluyen diferentes complementos: Auriculares “EarBliss”, cascos “EchoBass”, fundas de móvil “GreenGuard”, cargadores solares ”SolarBoost”. Se pueden añadir tantos productos como se requiera , introducir nuevos productos dentro de la página web es muy intuitivo y sencillo gracias a la interfaz de Wordpress. Cada uno de ellos tendrá su propio stock, el cual variará de forma automática cuando se realice una compra o alguna acción que haga variar el stock.

El diseño en el desarrollo de una página web, es un factor que contribuye significativamente al éxito de un sitio web. El diseño de la interfaz del usuario (UI) debe ser intuitivo y fácil de usar. Debe contener diseños visuales que faciliten la navegación por la web mediante botones y menús siempre priorizando que sea funcional. En cuanto al diseño de los productos, se ha elaborado de tal manera que imite lo máximo posible una página web de una empresa real, configurando sus características e incluyendo diferentes categorías, atributos y etiquetas. Para la descripción de los productos, hay que intentar ser lo más preciso posible y utilizar un vocabulario fácil de entender. Proporciona detalles específicos sobre el producto, incluyendo dimensiones, materiales, colores disponibles, y cualquier característica técnica relevante.

Se pueden utilizar diferentes herramientas para facilitar y optimizar los textos como puede ser Chat GPT (ver apéndice Tabla A.6), la cual es muy útil si se le da un uso adecuado. Esta aplicación puede ser de gran ayuda a la hora de desarrollar textos descriptivos ya que puede realizar textos pidiendo que respete el SEO (Search Engine Optimization)[9] de la página web. Esta aplicación hay que utilizarla como soporte y por ello no debe reemplazar el trabajo humano.

[9]Se refiere al posicionamiento en los motores de búsqueda realizando acciones de optimización de las palabras claves, orientado a conseguir un mejor posicionamiento de los sitios web.

Figura 2.3: Página principal de la página web Greentech Phone



Fuente: Elaboración: Propia

Por último, se ha incluido una página que incluye tres blogs que se han incorporado en una sección de la web. Estos blogs abordaran temas relacionados con el producto: móviles modulares, competencia y dispositivos eco-friendly. Todos los blogs tienen un control de SEO mediante el plugin de Yoast SEO (ver en el apéndice tabla A.1) para saber la calidad del contenido y guiar mejor su elaboración. Esto sirve para poder posicionar mejor la página web. Es fundamental que todo el contenido de la página web siga una estrategia de SEO para que cuando los usuarios busquen productos que vende la empresa, la página web salga en las primeras opciones. No hay información que aclare cómo es el funcionamiento del algoritmo de Google, pero se entiende que cuanto mejor esté desarrollada la página web mayor valor le dará este algoritmo y la posicionará mejor.

2.1.3 Pasarela de pago online

Una parte importante de la tienda virtual es la experiencia de compra y en ella influye la pasarela de pago que se utilice. Una pasarela de pago (payment gateway) es un servicio que autoriza los pagos con tarjetas de débito o crédito y también garantiza una transacción segura, encriptando la información financiera del cliente antes de transferirla a la cuenta de nuestro negocio (Vargas, 2021).

A continuación, se explicará cómo se incorpora el sistema de pago online. Woocommerce facilita un sistema básico el cual permite pagar mediante tarjetas de crédito o débito como Visa, MasterCard, etc. Todo ello mediante Woo Payments[10] que por el servicio de pasarela de pago se debe pagar una comisión de 1,26% del valor de la venta +0,23€ por cada transacción. Para que el comprador tenga más opciones se ha introducido la opción de pago mediante Stripe (ver apéndice Tabla A.1), un software de procesamiento de pagos e interfaces de programación de aplicaciones para sitios web de comercio electrónico y aplicaciones móviles. Stripe se lleva una comisión del 1,4% del valor de la venta + 0,25€ por transacción que se realice mediante su software de pago, muy parecida al sistema de pagos de Paypal.

[10]Woopayments es una extensión de Woocommerce la cual facilita métodos de pagos

2.2 Customer Relationship Management (CRM)

La gestión de las relaciones con los clientes o CRM es una herramienta que permite que haya un conocimiento estratégico de los clientes y sus preferencias, así como un manejo eficiente de la información sobre los mismos dentro de la organización (Montoya Agudelo y Boyero Saavedra, 2013). Esta herramienta digital ayuda a tener una base de datos más completa sobre los clientes, prospectos[11] y leads[12], optimizar las interacciones con ellos a lo largo de todo el ciclo de vida, mejorar la eficiencia de los procesos comerciales y de marketing, aumentar la retención de clientes y, en última instancia, impulsar el crecimiento y la rentabilidad del negocio.

Actualmente, las empresas líderes en el mercado de CRM son SalesForce, Microsoft Dynamics 365, Oracle CRM y Hubspot. Para una empresa pyme, lo mejor será contratar un servicio de un coste reducido. Para el trabajo se ha utilizado Hubspot (ver apéndice Tabla A.4), al contar con una versión gratuita y luego un pago mensual relativamente barato (sobre 20 euros), que incluye algunas herramientas adicionales así como la posibilidad de añadir contactos de marketing con lo que poder realizar estrategias de marketing. La versión gratuita es más bien una versión demo para probar como funciona la plataforma, para el trabajo con la versión gratuita fue suficiente.

Entre las funciones que se pueden usar se encuentra el “mailing” que es una estrategia de marketing que consiste en enviar correos electrónicos con un contenido que destaca por su diseño, para promocionar los productos y servicios de la marca (Silva, 2022). Para ello es fundamental haber conseguido previamente los contactos respetando en todo momento la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. Antes de poder hacer una campaña de mailing, los usuarios deben consentir que se les mande propaganda al correo, sino será considerado spam. Esto se puede lograr de diversas formas, por ejemplo cuando realicen alguna compra dar la posibilidad a que se suscriban al Newsletter u ofrecer descuentos si se suscriben, etc.

Una vez se cuente con una base de datos de clientes, se podrá proceder a enviar periódicamente información y promociones mediante el correo electrónico. Esta estrategia nos permitirá conseguir datos interesantes, como por ejemplo cuántas personas han abierto

[11]Un prospecto es una persona que aunque no haya dejado información sobre él es un posible cliente potencial.
[12]Un lead es un cliente potencial que ha realizado alguna acción indicando que está interesado en la empresa.

el correo, cuántos han pedido más información o incluso cuántos han usado el enlace para realizar alguna compra.

Cuando se tenga contratado el servicio con Hubspot, hay que conectar con esta plataforma con WordPress. Para ello existe un plugin dentro de Wordpress que facilita esta tarea. Una vez instalado el plugin, se procederá a vincular las dos aplicaciones para poder acceder a las principales funcionalidades de Hubspot desde Wordpress, de esta manera se conseguirán nuevas herramientas como crear base de datos de clientes, crear formularios o poner un chat en vivo en la página web. Por último, solo quedaría importar los contactos y configurar el CRM a nuestro gusto.

Mencionar que hay otras herramientas muy útiles que no se van a utilizar en este trabajo, como por ejemplo en el área de marketing, existen plataformas de publicidad en línea. Google Ads o Meta Ads (ver apéndice Tabla A.3) ayudan a poder promocionar la marca. Estas herramientas van encaminadas a estrategias de marketing de negocios, concretamente a SEM[13] (Search Engine Marketing). Según Cabrera (2021) para realizar una estrategia básica en Google Ads se necesita invertir mínimo 5€ diarios.

Por otro lado, se encuentran los sistemas de software llamados ERP (Enterprise resource planning) que le ayudan a administrar todos los recursos empresariales, respaldando la automatización y los procesos en finanzas, recursos humanos, fabricación, cadena de suministro, servicios y adquisiciones. El objetivo de implementar un sistema ERP es reducir los costos, un aumento de la productividad, planificar y realizar la automatización de sus procesos, así como la integración completa del negocio (Díaz, P. et al., 2005).

En resumen, los servicios y herramientas que se han explicado son los necesarios para poder operar de una manera básica en el entorno digital. Cada empresa dependiendo del sector y sus necesidades utilizará diferentes tipos de herramientas más específicas para cada caso.

[13]Estrategia de marketing digital utilizada para aumentar la visibilidad de un sitio web en los resultados de los motores de búsqueda principalmente a través de publicidad pagada.

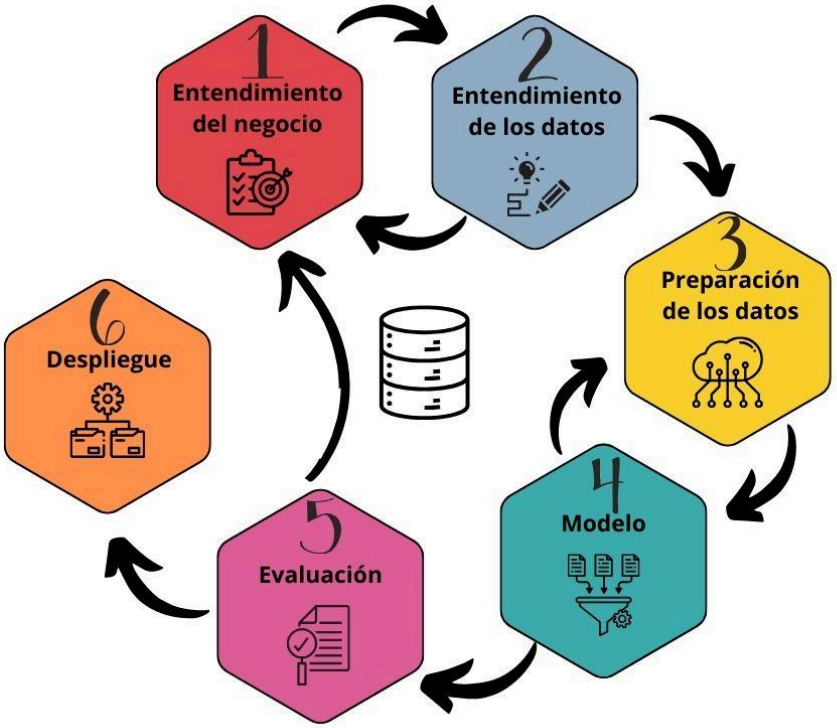
3. Creación de un sistema Big Data

Una vez realizada la digitalización empresarial se podrá dar paso a la creación de un sistema de Big Data. En esta sección se explicará todo el proceso que conlleva la integración, gestión y análisis de datos. Para ello, este trabajo se basará en una metodología estándar en la minería de datos llamada CRISP-DM (CRoss-Industry Standard Process for Data Mining). Esta metodología se basa en encontrar patrones e información útil a partir de una gran cantidad de datos. Fue desarrollada por un equipo creado a partir de varias empresas para dar solución a la minería de datos (Chapman, P. et al. ,1999). CRISP-DM sigue un ciclo de seis fases con dependencias entre ellas.

La figura 3.1 muestra como es el proceso de esta metodología, hay 6 fases en total y la secuencia de las fases no es totalmente estricta, lo que quiere decir que se puede avanzar y retroceder entre fases si es necesario. Las dos primeras fases tratan de entender el negocio y los datos necesarios para dar respuesta a las necesidades empresariales. Una vez comprendido qué es lo que se necesita, se da paso a integrar los datos. La tercera fase, es la encargada de crear y utilizar modelos de análisis para poder conseguir diferente información que pueda ser útil para su análisis. Por último, se analiza la información obtenida y se crean los diferentes informes para los diferentes departamentos de la empresa.

La metodología CRISP-DM es iterativa, esto quiere decir que las fases pueden repetirse varias veces a medida que se van ajustando los modelos en función a los resultados y a los conocimientos nuevos. A continuación se va a explicar más en detalle cada una de las fases junto con algún ejemplo de cómo se han implementado en el trabajo.

Figura 3.1: Esquema del ciclo CRISP-DM estándar



Fuente: Elaboración propia. Adaptado de Wikipedia

3.1.Entendimiento del Negocio

El entendimiento del negocio es la primera fase CRISP-DM y es fundamental para poder desarrollar correctamente todo el sistema Big Data. El primer paso es entender cuál es el propósito de la empresa y todo su entorno que la rodea. Para ello hay que analizar el sector de la empresa y se pueden usar planes estratégicos recientes o documentos internos de la empresa para poder comprender mejor el negocio. En el caso de este trabajo, se analizará el sector de la telefonía móvil y más específicamente el de smartphones ecológicos. En esta simulación no se ha realizado un estudio minucioso del negocio ya que no es el objetivo de este trabajo, por lo que se ha tenido en cuenta únicamente la competencia y el crecimiento del sector. La competencia directa son empresas que venden móviles ecofriendly como pueden ser Fairphone, Shiftphone y Oppo. La competencia indirecta son los que dominan el sector de la telefonía móvil tales como Iphone, Samsung, Xiaomi, Huawei y Nokia. También hay que saber cual es el objetivo de la empresa y qué es lo que quiere lograr para más adelante poder obtener información necesaria para ayudar a conseguir esos objetivos. Puede ser que solo se necesite realizar un proyecto de un departamento específico dentro de la empresa y eso dependerá de las necesidades de cada una.

Una vez conocidas las necesidades de la empresa se da paso a la comprensión de los datos. Esto no quiere decir que no vaya a ser necesario volver a esta fase para realizar algún ajuste o si no se ha tenido en cuenta alguna variable importante en el contexto empresarial.

3.2. Entendimiento y preparación de los datos

En estas dos fases, el objetivo es adquirir diferentes fuentes de datos que podrían interesar a la empresa tanto internos como externos y, por otro lado, integrar esos registros en una base de datos, propia de la empresa en el caso de que sean datos estructurados o en un Data Lake si son datos no estructurados. Los datos deben de ser verídicos, por ello es importante que las fuentes de donde se extraen los datos sean fiables. Por otro lado, deben de ser representativos, esto quiere decir que se necesita una cantidad de registros lo suficientemente grande como para poder analizarlos correctamente teniendo en cuenta toda la diversidad de datos posibles. Para este trabajo, esta fase se ha dividido en 2 grupos: los datos internos y externos. Para cada grupo la obtención será diferente.

Por un lado, los datos internos son lo que la empresa posee o que se pueden obtener de fuentes propias como lo son Wordpress o Hubspot. Entre los datos que se pueden obtener se destacan las visitas por día a la página web, cuánto han durado las búsquedas de media, desde qué país han buscado la página web, las ventas online, la variación de stock, los pedidos, etc.

Por otro lado, los datos externos son los que se han encontrado, capturados o proporcionados, desde fuera de la empresa ya sea mediante distintas aplicaciones o plataformas, estudios externos, bases de datos de terceros, como por ejemplo Google trends, X o Instagram.

3.2.1 Datos internos

La extracción de los datos de la página web se ha realizado mediante una conexión entre Wordpress y Google analytics. Por otro lado, para los datos propios del negocio se ha utilizado un proceso ETL (Extract, Transform and Load) que se explicará seguidamente.

3.2.1.1 Datos de la página web

Para la extracción de los datos de la página web, se debe crear una cuenta en Google Analytics (GA4) (ver apéndice Tabla A.2), crear un contenedor de la propiedad de la página web y configurar toda la plataforma para el uso de la empresa.

Una vez configurado Google Analytics, debemos vincular la cuenta a Wordpress. Este paso se puede realizar de dos formas (B., 2017): manualmente, incluyendo la conexión en el código del wordpress o mediante el uso de un plugin. Para realizarlo manualmente, GTM (Google Tag Manager) facilita un fragmento de código de seguimiento que hay que incluir en el código de programación de la página web, concretamente encabezado del tema (header.php) en la sección <*head*> de la página web. Este código lo facilita Google Analytics para la incorporación en Wordpress, en el listado 3.1 se puede observar el código. Es importante donde pone "ID de Google Analytics" incluir el ID propio, por temas de seguridad no se ha detallado cuál es el de este trabajo.

Listado 3.1: Código de seguimiento de Google Analytics para Wordpress

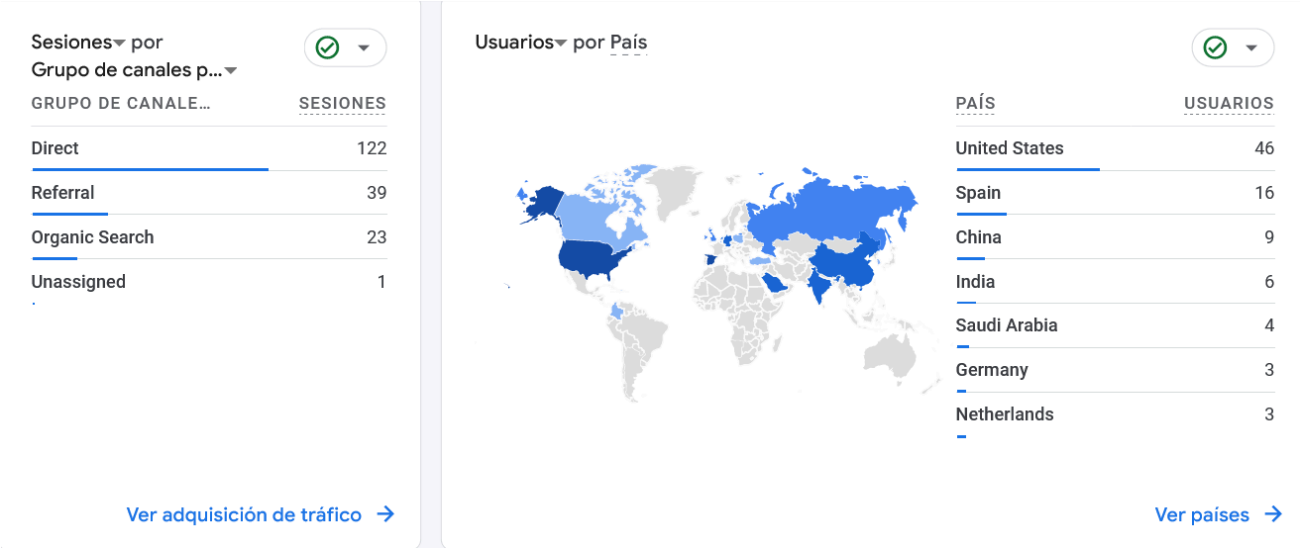
```
 "add_action('wp_head','my_analytics', 20);
function my_analytics() {
?>
<!-- Global site tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=(ID de Google
Analytics )"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());
  gtag('config', 'UA de GTM');
</script>
<?php
}".
```

Fuente: Elaboración propia facilitado por Google Analytics

La otra forma de vincular las dos plataformas es más sencilla. Para ello se necesita instalar un plugin llamado "Monster Insights"(ver apéndice Tabla A.1). Este plugin, mediante la ID de Tag Manager (ver apéndice Tabla A.2) permite conectar las plataformas. Este ID hay que activarlo para su funcionamiento y funciona mediante etiquetas (tags) que permiten agregar, actualizar y administrar fragmentos de código en sus sitios web sin tener que modificar directamente el código fuente del sitio. Únicamente hace falta agregar esta etiqueta a WordPress, para ello hay que ir a Ajustes, seleccionar Google Tag Manager y pegar en la sección ID de Google Tag Manager.

Gracias a esta conexión, se podrá extraer toda la información deseada que se pueda captar desde la página web como por ejemplo saber cuántas personas han accedido en el día a la página web, cuantos clicks en el carrito de compra, desde dónde han entrado al sitio web o de dónde son los usuarios. Esta información nos la facilita Google Analytics proporcionando información mediante paneles informativos o gráficas. La información que se extrae de esta manera suele ser muy útil para el departamento de marketing. Esta información muchas veces no hace falta ni que pase por modelos de análisis, estando ya lista para ser analizada.

Figura 3.2.1: Paneles de Google Analytics del tráfico web de Greentech Phone



Fuente: Elaboración propia a través de Google Analytics

La figura 3.2.1 muestra mediante el mapa del mundo de qué países son los usuarios que han buscado la página web. Estados Unidos, España y China han sido los países donde más han buscado la página web. Se ha configurado este informe para que se pueda utilizar en los modelos de análisis de datos que sería la siguiente fase o pasar directamente a la evaluación y visualización (fases 5 y 6 de la Figura 3.1).

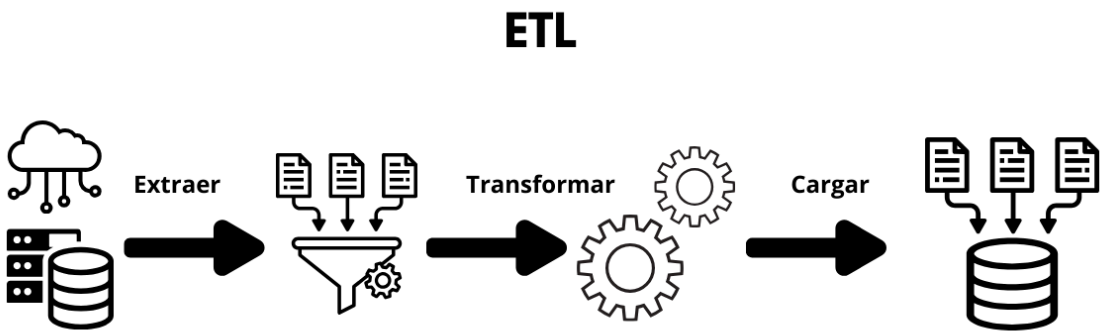
3.2.1.2 ETL de datos de Wordpress

Para la integración de los datos propios de la actividad empresarial o información generada dentro del negocio como pueden ser las ventas, las reservas, los stocks de

producto, los clientes, etc. tendremos que extraer los datos de la base de datos de Wordpress donde se guarda automáticamente cualquier evento. Para ello vamos a utilizar un proceso llamado ETL (Extract, Transform and Load). El nombre de este proceso hace referencia a las tres pasos para poder integrar los datos que deseamos al data warehouse[14] propio de la empresa, el proceso se puede observar en la figura 3.2.2. Los datos se obtienen (extraer) de varias fuentes que deseemos, se convierten (transformar) en un formato apto para su almacenaje y se almacenan (cargar) en el apartado del data warehouse que deseamos.

Para este trabajo se han podido obtener datos de los comentarios en los productos, del stock de los productos y de los usuarios, pero al ser una simulación no se han podido extraer datos tales como las ventas, las reservas o las devoluciones.

Figura 3.2.2: Proceso de ETL



Fuente: Elaboración: Propia

Para entender mejor cómo es el proceso ETL, se va a explicar como se ha usado en los datos de Wordpress (productos, usuarios y comentarios en los productos).

El primer paso consiste en programar las órdenes de extracción de los datos. Para ello es necesario utilizar un lenguaje de programación, en este caso Python (ver apéndice Tabla A.2). Todo el código de programación se encuentra en el Anexo 1. La primera orden que tenemos que dar es entrar en la base de datos de nuestro Wordpress, siendo necesario proveer las credenciales del usuario y contraseña para poder realizarlo, sino cualquier

[14]Almacén electrónico donde una empresa mantiene una gran cantidad de información.

persona podría extraer información de cualquier base de datos. Una vez conectado a la base de datos, se ejecuta la orden de extracción de datos. En el caso de los datos de los productos, al no existir compras no hay variaciones. Por ello, se ha aleatorizado el número de mercancías cada día para poder observar cómo cambia el número de stock en la base de datos, generando una transformación en el dato del stock. La programación del stock de cada producto se ha considerado que sea como mínimo de 500 mas un número elegido al azar entre 0 y 1500.

El segundo paso es la utilización de Docker (ver apéndice Tabla A.2). Este software permite utilizar la ETL desde cualquier dispositivo o cualquier lugar. Para ello, hay que convertir todo el código de programación en una imagen mediante Docker (conocido también como contenedor) que son paquetes de software que incluyen todos los elementos necesarios para ejecutar los productos propios en cualquier entorno. Este software pedirá unos parámetros y requerimientos para su ejecución (se pueden observar los parámetros en el listado 3.2 y los requerimientos en el listado 3.3). El primer parámetro es el más importante ya que pide una API[15] para poder conectarse al código de programación de la ETL, en este caso es "tiangolo/uvicorn-gunicorn-fastapi". Los requerimientos son los servicios que se necesitan para procesar la orden de programación, es decir, para poder ejecutar un código. Si no se cuenta con esos requerimientos no se podrá ejecutar.

Una vez creada la imagen, habrá que subirla a Google Cloud (ver apéndice tabla A.2). Una de las razones para ello, es el peso de estas imágenes. En este caso, el archivo pesa 1,5 Gb. por lo que se necesita usar el servicio de Google Cloud llamado Container Registry[16] en donde vamos a almacenar todas las ETLs. El segundo de los servicios que necesitaremos es Cloud Run[17]. Gracias a este servicio se va a poder ejecutar desde cualquier sitio las imágenes creadas y subidas en Container Registry. Por último, se va a utilizar Cloud Scheduler[18] para programar la ejecución diaria de todas las imágenes que se encuentren en Google Cloud.

[15]API es un mecanismo que permite que dos aplicaciones se comuniquen entre ellas. Como pueden ser dos páginas webs diferentes.
[16]Cloud Container Registry es un servicio integrado en el ecosistema de la nube de Google que se encarga del almacenamiento, administración y protección de las imágenes de contenedores privados, como los de la plataforma de Docker(Team, 2022).
[17]Cloud Cloud Run es una plataforma de procesamiento administrada que te permite ejecutar contenedores directamente sobre la infraestructura escalable de Google. *(Cloud Run*, s.f).
[18]Cloud Scheduler permite programar una orden para operaciones de infraestructura de nube en horarios definidos o a intervalos regulares.

Listado 3.2: Código de configuración para la creación de imagen Docker para desplegar el código de la ETL en la nube.

```
FROM tiangolo/uvicorn-gunicorn-fastapi
COPY requirements.txt .
RUN pip install -r requirements.txt
RUN mkdir -p app
COPY ./app app
EXPOSE 8080
CMD ["uvicorn", "app.main:app", "--host", "0.0.0.0", "--port", "8080"]
```

Fuente: Elaboración propia

Listado 3.3: Código de programación requerimiento de Docker

```
pandas==2.1.0
mysql-connector-python==8.1.0
gspread==5.11.2
gspread-dataframe==3.3.1
google-api-core==2.11.1
google-api-python-client==2.100.0
google-auth==2.23.0
google-auth-httplib2==0.1.1
google-auth-oauthlib==1.1.0
googleapis-common-protos==1.60.0
PyDrive==1.3.1
fastapi==0.103.1
```

Fuente: Elaboración propia

La programación de la frecuencia de este servicio se especifica con el formato Cron de Unix. "Cron", en palabras simples, es un usuario programador de trabajos basado en el tiempo para ejecutar un trabajo específico periódicamente en horas, fechas o intervalos fijos (Soam, 2018). En el caso del trabajo se escogió la "*55 23 * * *" que quiere decir que todos los días a las 23 horas y 55 minutos se ejecute la imagen, es este caso la ETL de los datos de Wordpress.

Los tres servicios utilizados son servicios de tipo *IaaS* ya que únicamente se necesita infraestructura para poder almacenar, ejecutar, y programar. Tener las ETLs en la nube nos permite automatizar la orden de extracción y recogida de datos sin la necesidad de realizarla manualmente todos los días. Este servicio se paga por uso, pero es un coste insignificante para una empresa, rondando los 50 céntimos al mes.

Por último, solo faltaría cargar los datos en la base de datos que se desee. En el caso de este trabajo, el almacenaje de los datos se va a realizar en una hoja de cálculo de

Google Sheets (ver apéndice Tabla A.2) que proporciona Google Drive, ya que los datos no van a ser de unas dimensiones muy grandes y con los 15 Gb que ofrece Google gratuitamente es suficiente. En un caso real es mejor utilizar el almacenamiento de Cloud computing conocido como Bigquery, un servicio de Google como *PaaS* que permite almacenar y organizar los datos con una escalabilidad y disponibilidad casi ilimitada. Los costes de la utilización, como ya se ha mencionado anteriormente, son de pago por uso, por lo que cuanto más se necesite más se pagará por ello. En el caso de una pyme, estos costes serán relativamente bajos, aunque siempre dependerá de la cantidad de datos que necesite la empresa. En el Anexo 1 se muestra el código de programación para realizar esta carga de datos en la base de datos de Google Sheets. Se han creado tres documentos diferentes para cada una de las fuentes de datos: Productos, Usuarios y Comentarios. Una vez comprobado que los datos se guardan correctamente en la base de datos se podrá automatizar su extracción diaria.

Mencionar que no es totalmente necesario utilizar Google Cloud para la integración de estos datos. Las ETLs se pueden realizar manualmente, pero para poder automatizar este proceso lo mejor es utilizar alguna herramienta que lo permita y en el caso de este trabajo se ha utilizado Google Cloud.

3.2.2 Datos externos

Los datos externos a la empresa son algo más difíciles de obtener. Cada vez se está limitando más la extracción de datos debido al valor que tienen para los diferentes negocios. Por ejemplo, en el año 2021 la plataforma X (conocida anteriormente como Twitter) permitía extraer al mes unos 100.000 tweets. En la actualidad, esta cantidad se ha reducido a 1.000 unidades para que sea gratuito. Si se desea extraer un número mayor de tweets se debe realizar un pago mensual pactado con la compañía propietaria de la plataforma. Esto se debe a que con el aumento de los data scraping (raspadores de datos), se llegó a saturar considerablemente los servidores de Twitter y perjudicar la experiencia de los usuarios de la red social (Nast, 2023). Otras compañías grandes como Amazon o Google también se han percatado de la importancia de los datos e intentan cada vez más limitar su obtención gratuita.

En este trabajo se ha extraído información del mercado bursátil de las grandes empresas de teléfonos móviles (Nokia, Apple, Samsung y Xiaomi). Gracias a estos datos se podrá analizar cuales son los motivos de que el valor bursátil de la compañía aumente o

descienda, para ello es importante obtener otros datos de estas empresas, como noticias, ventas, cuota de mercado…

Por otro lado, mediante la página de Google Trends se ha recogido información sobre las diferentes búsquedas que realizan los usuarios de Google sobre los móviles, información útil para saber qué es lo que interesa a los usuarios cada día. Por último, se ha utilizado Amazon para saber la opinión sobre los móviles de la competencia con el objetivo de saber qué es lo que les gusta y lo que no les gusta de cada uno de ellos y así tener información sobre lo que realmente buscan nuestros potenciales clientes.

Para la obtención de estos datos se ha utilizado el mismo proceso que para los datos internos del negocio (es decir un proceso ETL). Por ello, no se va a entrar en profundidad en cómo se realiza el proceso, solamente matizar que al ser información externa se utilizan otras técnicas. Para Amazon y el mercado bursátil se ha utilizado un técnica llamada Web Scraping[19]. Esta técnica sirve para extraer información del sitio web a la que hayamos accedido. Esta información siempre es pública por lo que lo único que se está haciendo es guardar esa información que está a nuestro alcance. En los anexos 2 y 3 se encuentran los códigos de las dos ETLs de estas dos fuentes de información.

En el caso de Google Trends (ver apéndice Tabla A.3) se realiza del mismo modo con una ETL pero en este caso se va a utilizar un mecanismo llamado API para solicitar la extracción de la información que deseemos. Puede ocurrir que, como se ha mencionado anteriormente, para cursar esta solicitud se necesite realizar algún tipo de pago a la empresa en cuestión por la extracción de esa información, pero en este caso es totalmente gratuita. En el Anexo 4, se encuentra todo el código de programación de esta ETL de Google Trends. Esta ha sido la ETL más completa de todas, ya que se han obtenido 4 tipos de datos diferentes. El primero de todos ha sido las búsquedas en Google de la competencia que hay de empresa de móviles, tanto directa e indirecta. Todos los días, al final del día se ha ejecutado la ETL para poder tener una línea temporal de estas búsquedas. En el segundo, se ha pedido los datos de los países donde se busca el término “Eco Phone” y cuantas veces se han realizado estas búsquedas. Esta información es muy útil para saber en qué zona geográfica se encuentran los usuarios potenciales. Por último, se ha considerado la extracción de las búsquedas más parecidas a “Eco Phone”. Google Trends permite obtener información muy útil para las empresas.

[19]Simular una navegación humana por la web para la extracción de información del sitio web.

En este trabajo, se han utilizado estas tres fuentes de información externa explicadas anteriormente, pero se pueden incluir todas las que sean necesarias, siempre teniendo en cuenta las limitaciones actuales y valorando cuales son necesarias. Por ejemplo, se pueden conseguir datos de publicidad de Google Ads y Meta Ads. Estas fuentes están sobre todo destinadas para el departamento de Marketing.

Una vez concluida la preparación de las fuentes necesarias para la construcción de la base de datos, se necesitarán unos días de recolección de datos a fin de obtener una muestra lo bastante representativa como para que se pueda comenzar a utilizar algún modelo de análisis de estos datos.

3.3. Modelo de análisis de datos

Esta fase es la encargada de procesar los datos mediante el uso de modelos analíticos. Un modelado de datos es un tipo de lenguaje orientado a explicar la relación que tienen los datos entre sí, con el objetivo de poder dar un significado a los mismos. Se necesitará utilizar diferentes técnicas dependiendo del tipo de datos que sean, no es lo mismo utilizar datos tipo texto que de audio o vídeo. Como ya se ha comentado anteriormente, los datos pueden ser estructurados, semiestructurados o no estructurados.

En las últimas décadas ha evolucionado drásticamente el procesamiento computacional, lo que ha conllevado mejoras exponenciales en los entrenamientos de modelos basados en Machine Learning (ML) (aprendizaje automático), una tecnología que sirve de apoyo en un sistema Big Data. Ahora bien, ¿qué es el Machine Learning? ML es una rama dentro del campo de la Inteligencia Artificial que mediante una colección de algoritmos se dota a los ordenadores de la capacidad de identificar patrones de los datos y elaborar un análisis. La idea es desarrollar modelos analíticos que aprendan de los datos minimizando la participación humana. En ML, los algoritmos que se utilizan pueden ser de dos tipos “supervisados” o “no supervisados” (Lee, 2019).

En el aprendizaje supervisado, los algoritmos trabajan con datos “etiquetados” a partir de un conjunto de datos históricos, intentando encontrar una función que dadas las variables de entrada, les asigne la etiqueta de salida adecuada. En otras palabras, el algoritmo se entrena con un “histórico” de datos y así “aprende” a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida (Simeone, 2018). Este tipo de aprendizaje es el más común en ML. Este aprendizaje se utiliza para en los casos

que queramos realizar una regresión o una clasificación. Se pueden considerar los siguientes algoritmos:

- Algoritmos de clasificación: Naive Bayes, Máquinas de Vectores de Soporte (Support Vector Machines, SVM), Regresión Logística, Árboles de Decisión, Bosques Aleatorios y Redes Neuronales (Larrosa, C. , 2023).
- Algoritmos de regresión: Regresión Lineal, Regresión no Lineal, Máquinas de Vectores de Soporte (SVM), Árboles de Decisión, Bosques Aleatorios y Redes Neuronales (Gonzalez, 2023).

El aprendizaje no supervisado se utiliza en datos sin etiquetas y el objetivo es encontrar relaciones en los datos. Este aprendizaje es de carácter exploratorio, su función es poder describir la estructura de los datos a partir de unos datos de entrada (Santos, 2021). Este aprendizaje se utiliza para la agrupación de los datos (clustering). En este trabajo no se va a entrar más en detalle en los algoritmos no supervisados.

A continuación se van a mencionar una variedad de técnicas comunes en ML para diferentes tipos de datos, hay más tipos de análisis pero en este trabajo se van a explicar el análisis de texto, el análisis de audio, el análisis de vídeo y por último el análisis predictivo (Gandomi y Haider, 2015). Este último va a ser el que más se desarrolle.

a) Análisis de texto: Las técnicas de este tipo analizan cualquier clase de texto, blogs, encuestas, redes sociales, noticias o cualquier dato textual. El análisis de texto mediante la combinación del análisis estadístico, la lingüística computacional y el aprendizaje automático, crea una técnica de análisis avanzada llamada Procesamiento de Lenguaje Natural (Natural Language Processing, NLP). Esta técnica se utiliza para diferentes fines como traducir automáticamente textos (ejemplo de ello es Google Translate), crear agentes virtuales y chatbots (Siri de Apple y Alexa de Amazon), análisis de sentimiento o resumen de textos (IBM, s.f.).

b) Análisis de Audio: En la actualidad, es común ver cómo este tipo de análisis se aplica en los centros de atención al cliente y la atención sanitaria. El análisis del audio, también conocido como análisis del habla, utiliza habitualmente dos enfoques de reconocimiento de voz: el enfoque basado en transcripciones conocido como Reconocimiento Continuo de Voz de Amplio Vocabulario (LVCSR) y el enfoque basado en fonética.(Gandomi y Haider, 2015).

c) Análisis de Vídeo: El análisis de vídeo, también conocido como análisis de contenido de vídeo (Video Content Analysis VCA), implica una variedad de técnicas para monitorear, analizar y extraer información significativa de las transmisiones de vídeo. Este tipo de análisis ha experimentado un crecimiento significativo en los últimos años, por el auge de las cámaras de circuito cerrado de televisión (CCTV) y por la popularidad de sitios web para compartir vídeos. Los sectores en los que más se utiliza este tipo de análisis son el de seguridad y vigilancia y el comercio minorista. En el sector de la seguridad y vigilancia sirve para realizar funciones de vigilancia de manera más eficiente, gracias a estos análisis se pueden detectar infracciones de zonas restringidas o identificar objetos sospechosos. En el sector minorista, sirve para conseguir información relevante del comportamiento de compra de los diferentes colectivos.

d) Análisis Predictivo: Como ya se ha explicado en la sección 1.2.3, dadas las características del Big Data se necesitan utilizar técnicas avanzadas. Entre las técnicas de Machine Learning, se encuentran las denominadas “Ensemble learning” (Aprendizaje por conjuntos). Esta técnica es de aprendizaje supervisado y es de las más conocidas para dar solución a los retos del Big Data. Ensemble learning reúne un conjunto de métodos utilizando múltiples algoritmos de aprendizaje para obtener un rendimiento predictivo más eficiente que si se realiza cada método por separado. Su objetivo es aprovechar las fortalezas y compensar las debilidades de cada modelo. El fin de esta técnica es combinar los resultados de los diversos modelos para crear una predicción más precisa, en la actualidad esta técnica ha conseguido ser una de las herramientas más valiosas en las áreas del Big Data y Machine Learning (Singh, 2023).

Entre las técnicas Ensemble learning se pueden diferenciar 2 grupos según su complejidad, el primero grupo son las técnicas de conjunto simples (Simple Ensemble Techniques) y el segundo grupo son las técnicas avanzadas de conjunto (Advanced Ensemble techniques)(Singh, 2023).

Entre las técnicas simples se pueden destacar tres:

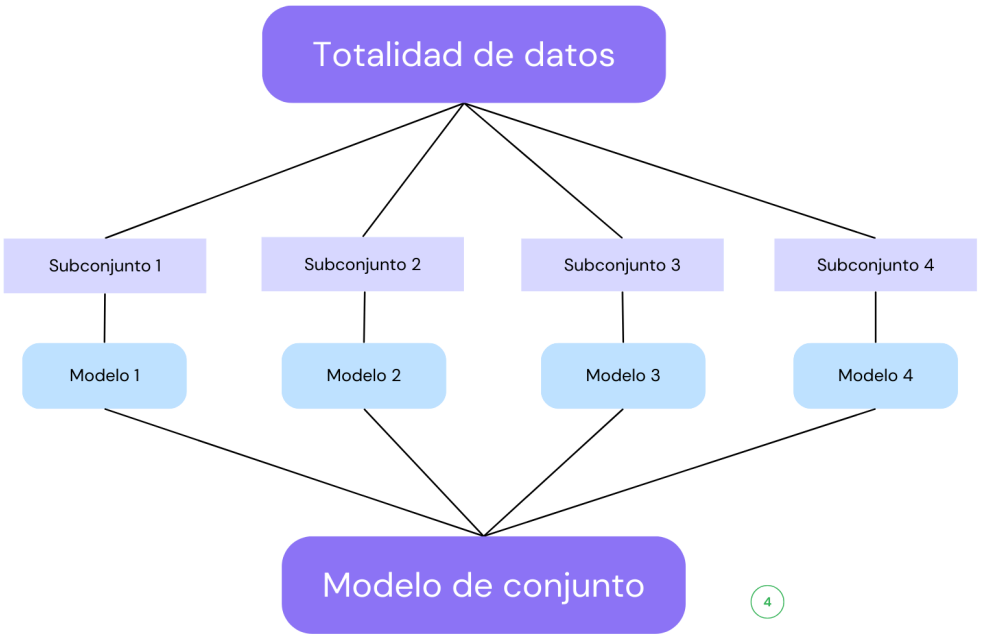
- Max voting (Votación por mayoría): Esta técnica trata de elegir la predicción que más se repite entre todos los modelos utilizados para realizar una predicción, para ello todos los modelos deberán utilizar los mismos datos.
- Averaging (promediar): Esta técnica trata de tomar un promedio de las predicciones de todos los modelos utilizados para hacer una predicción final.

- Weighted Averaging (Promedio ponderado): Esta técnica es parecida a la anterior, la diferencia es que a todos los modelos utilizados se les asignan diferentes pesos que definen la importancia de cada modelo para la predicción.

Entre las técnicas avanzadas también se encuentran varias técnicas de algoritmos muy utilizadas, pero en este trabajo solo se van a explicar las técnicas de Bagging (embolsado) y Boosting (Impulsando) los dos son meta-algoritmos. La idea básica de las dos técnicas es reducir la varianza en relación al ECM (Fernández et al., 2021), aunque en el Boosting también se han desarrollado métodos para problemas de regresión.

La técnica de Bagging desarrollada por Breiman (Breiman, 1996), utiliza un método de muestreo en el que se crean subconjuntos de observaciones a partir del conjunto de datos original. El tamaño total de todos los subconjuntos tiene que ser el mismo que el conjunto original. Cada subconjunto utiliza uno de los modelos, gracias a esto los modelos funcionan en paralelo y son independientes entre sí. El objetivo de utilizar esta técnica es combinar los resultados de los múltiples modelos para obtener un resultado general.. En la figura 3.3.1 se observa cual es el proceso de la técnica de Bagging.

Figura 3.3.1: Proceso de la técnica de Bagging

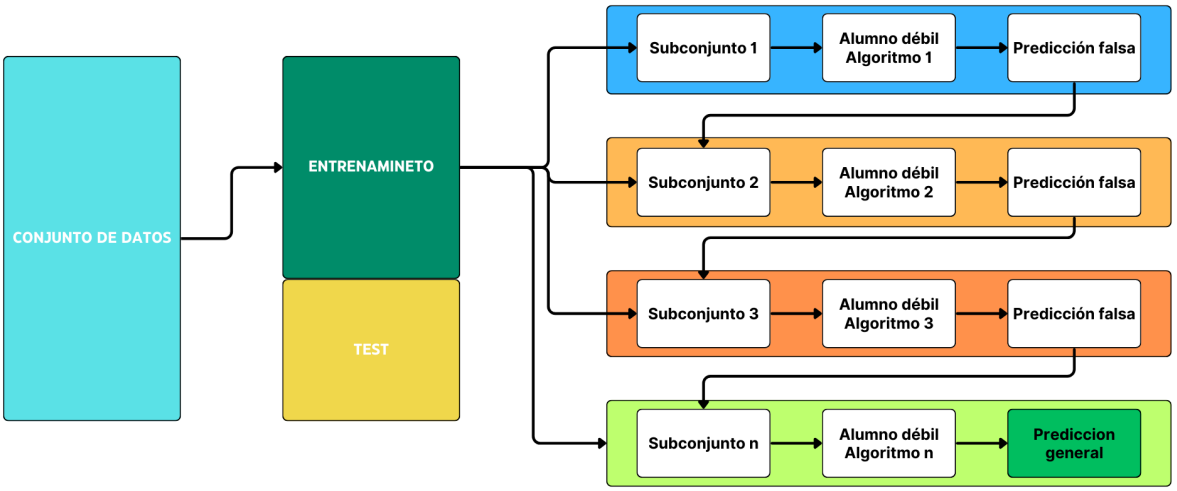


Fuente: Elaboración propia

Por otro lado, está la técnica Boosting que utiliza un proceso secuencial y como en la técnica Bagging, los datos se separan por subconjuntos. Esta técnica fue desarrollada por Kearns y Valiant (1994) para los problemas de clasificación (Kearns y Valiant, 1994), tardarían unos años hasta que se desarrollarían también metodologías para implementarla en problemas de regresión, exactamente hasta el año 2001, en el que se creó el método llamado "*gradient boosting machine*" (Friedman, 2001).

La técnica Boosting lo que intenta conseguir es que cada modelo mejore el rendimiento del conjunto de todos los modelos. Esta técnica utiliza a los modelos como aprendices. Los primeros modelos serán estudiantes débiles (modelos de aprendizaje automático que funcionan mal) pero con el paso de la secuencia, estos modelos se irán convirtiendo en estudiantes fuertes que son capaces de realizar predicciones de mayor confianza, reduciendo la varianza. (Eslamijam, 2022). La idea principal es "impulsar" el desempeño de varios alumnos débiles, cada uno de los cuales se desempeña ligeramente mejor que un esquema de toma de decisiones aleatoria, en un algoritmo conjunto sólido y preciso (Mendonça et al., 2024). Una vez desarrollado el primer modelo débil (utilizando cualquier modelo, como árbol de decisión, regresión logística, etc.), se evalúa su desempeño en todo el conjunto de datos. La importancia reside en los puntos mal clasificados o valorados, porque el siguiente modelo aprenderá en base a esos errores y no en base a los puntos bien clasificados o valorados.

Figura 3.3.2: Proceso de la técnica de Boosting



Fuente: Elaboración propia adaptado de (Ali, 2023)

En la figura 3.3.2 se puede observar cómo funciona la técnica de Boosting. Una parte del conjunto de los datos no se ha utilizado para el entrenamiento y se reserva para posteriormente realizar el test para el modelo. Selecciona al azar el 80% de los datos como muestra de entrenamiento y el 20% restante como muestra de test (Fernández, Costa y Oviedo, 2021). En la fase 4 de la Evaluación en la metodología CRISP-DM se explicará más detalladamente.

En el caso de este trabajo se ha decidido utilizar un modelo ya entrenado, concretamente un modelo de Procesamiento de Lenguaje Natural (NLP). Específicamente se utilizará para los comentarios de los productos de la competencia en Amazon que son datos de tipo no estructurado. Se ha utilizado un modelo ya entrenado con una base de millones de datos con el fin de que pueda ser capaz de analizar los datos que se le pidan más adelante aunque en su entrenamiento no los haya utilizado. A este tipo de modelo, se le conoce con el nombre de *Zero-shot classification,* un paradigma en el que un modelo puede clasificar ejemplos nuevos que pertenecen a clases que no estaban presentes en los datos de entrenamiento. Zero-shot implica predecir una clase que el modelo no haya visto durante su entrenamiento. Este enfoque puede considerarse como una instancia de aprendizaje por transferencia, que implica utilizar un modelo entrenado para una tarea en un contexto diferente al de su entrenamiento original. Esta técnica resulta especialmente útil en situaciones donde la disponibilidad de datos etiquetados es limitada.

En *Zero-shot classification*, proporcionamos al modelo una indicación y una secuencia de texto que describe la tarea que queremos que realice, todo en lenguaje natural (lenguaje humano). En este tipo de clasificación, no se incluye ningún ejemplo específico de la tarea deseada. Esto se diferencia de la clasificación de uno (One Shot) o pocos disparos (Few shot), donde se proporcionan uno o varios ejemplos al modelo para que resuelva la tarea en cuestión (Hugging Face, s.f.)

Estos modelos se pueden utilizar sin coste gracias a la comunidad de análisis de datos. Los miembros de esta comunidad suben sus modelos a Hugging Face, una plataforma donde la comunidad desarrolla modelos de machine learning, aportan datos y aplicaciones para poder aprender y ayudar a los demás. Estos modelos son públicos para que cualquier persona los pueda descargar y utilizar. En este caso, se ha optado por el modelo de Lik Xun Yaun llamado “distilbert-base-multilingual-cased-sentiments-student” por ser idóneo para los datos del trabajo, además de ser el más popular y utilizado entre los modelos NLP en los últimos meses. En el mes de febrero de 2024 se ha descargado más

de 11 millones de veces para su uso (Xun Yuang, 2023). Este modelo contiene más de 135 millones de parámetros[20] y con más de 590.000 datos para su entrenamiento previo.

Para poder usar este modelo, habrá que utilizar una vez más Docker (ver apéndice Tabla A.2) y mediante Google Colab ejecutar el modelo NLP para que analice los datos de los comentarios que se encuentran en nuestra base de datos. En el Anexo 5 se muestra el código de programación utilizado.

Este modelo es capaz de entender varios idiomas diferentes, pero para homogeneizar el análisis a un idioma se traducen todos los comentarios a inglés. Para ello se debe hacer una llamada en el código de programación a Google Translate, para que cuando se vayan cargando los datos se traduzcan automáticamente al inglés.

El siguiente paso es proporcionar y autorizar la entrada en la base de datos propia para poder extraer y analizar los datos que deseamos. Este paso se realiza mediante un token específico que funciona como contraseña para poder acceder a la base de datos. Se ha omitido el token utilizado por temas de seguridad y privacidad debido a que sino cualquier persona podría acceder a la base de datos.

Finalmente, sólo quedaría que el modelo analice los comentarios de Amazon tanto los positivos como los negativos y guarde la información de cada uno de ellos en diferentes apartados. Este modelo nos devuelve una información tanto en gráficos de barra y correlación como en formato Tag Cloud[21] (Nube de palabras). Estas representaciones se podrán usar para analizar esa información obtenida en el último paso pero antes de ello, hay que realizar otra fase llamada Evaluación. En esta fase se podrá ver si la información obtenida era la que se necesitaba o si los resultados son acordes con los mínimos exigidos.

Algunos de los resultados finales de este modelo son los que aparecen en la Figura 3.3.3 y la Figura 3.3.4. La primera figura es una nube de palabras de los comentarios negativos y la segunda es una tabla que relaciona las valoraciones de los comentarios positivos.

[20]Un parámetro es una variable propia del modelo que se obtiene de forma automática a partir de los datos usando el algoritmo de entrenamiento. Aparte de los parámetros, también existen los "hiperparámetros", que son variables externas al modelo que se deben definir manualmente al momento de programar el algoritmo de entrenamiento. (Sotaquirá, 2023).
[21]Tag Cloud es una representación visual de datos de texto que a menudo se utiliza para representar metadatos de palabras clave. Las etiquetas suelen ser palabras individuales y la importancia de cada etiqueta se muestra con el tamaño o el color de la fuente.(Wikipedia contributors, 2023)

La nube de palabras contiene palabras de diferentes tamaños, cuanto más grande es el tamaño de las palabras quiere decir que más se repite esa palabra dentro del análisis realizado de los comentarios negativos. En la figura 3.3.3, se pueden observar palabras con un connotación claramente negativa como “old”, “problem”, “issue”. Pero hay otras palabras que a priori no se entiende que sean negativas, pero el modelo lo ha percibido como algo malo, como por ejemplo la palabra “Samsung” que se repite mucho en los comentarios negativos y el modelo lo ha destacado como un negativo. Esto podría significar muchas cosas como, por ejemplo, que los compradores de móviles tienden a ver de mala calidad los productos de Samsung o, por lo contrario, podría ser que lo usan como comparativa para criticar a otros competidores. Otra palabra a destacar sería “camera” esto puede ser debido a que los clientes consideran muy importante la calidad de la cámara del móvil y por ello si no cumple con sus perspectivas valorarán negativamente el producto. Gracias a este análisis se puede saber a que dan importancia los clientes y realizar estrategias en función de los resultados obtenidos.

En este trabajo al ser una simulación no se dispone de una cantidad de comentarios lo suficientemente grande para que el modelo sea totalmente representativo. Para ello se necesitan miles de comentarios, pero gracias a este modelo se puede tener una idea de cuales son las palabras más comunes en el caso del sector de los móviles.

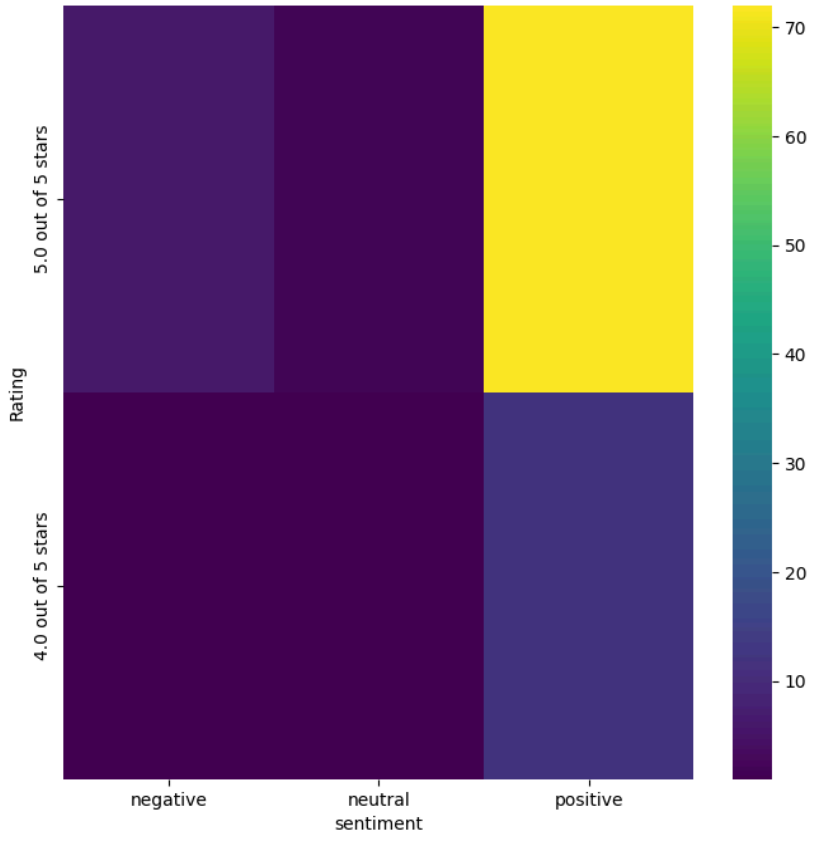
Figura 3.3.3: Resultado Wordcloud del Modelo de NLP de los comentarios negativos

Fuente: Elaboración propia mediante Google Colab

En la figura 3.3.4 se representan los comentarios positivos (de 4 y 5 estrellas) mediante un mapa de calor. Un mapa de calor es una disposición de rectángulos, cada uno de ellos coloreado de un color, desde tonos oscuros a tonos más claros.. Cuanto más se acerque al color amarillo mayor número de comentarios residirán en ese rectángulo. El eje X se divide en la variable de si el comentario es positivo, negativo o neutral y el eje Y se divide entre comentarios con 4 o 5 estrellas. Donde mayor comentarios residen en el rectángulo de comentarios de 5 estrellas y con sentimiento positivo y los segundo que más son los de 4 estrellas y con sentimiento positivo. Este tipo de tabla se utiliza para comprobar si el modelo realmente está funcionando correctamente mediante una clasificación de los datos. En este caso tiene sentido que la mayoría de los comentarios positivos tengan una connotación positiva, destacar que algunos de los comentarios de 5 estrellas han sido clasificados con una connotación negativa, por ello sale coloreado con un morado no tan oscuro como los comentarios de 4 estrellas.

Se pueden realizar otros tipos de gráficas o representaciones con este modelo, pero para este trabajo se han escogido esto dos para mostrar cómo funciona. En el caso de una empresa real, ésta podrá usar todos los que necesite para realizar su análisis.

Figura 3.3.4: Tabla de los Modelo de NLP de los comentarios Positivos



Fuente: Elaboración propia mediante Google Colab

3.4. Evaluación

En esta fase no se va a entrar en profundidad dado que no es el objetivo del trabajo. Aún así, nos ha parecido importante explicar en qué consiste esta etapa. Esta fase trata de evaluar si los modelos han funcionado correctamente así como si se han tenido en cuenta todas las variables necesarias para que el resultado sea el más acertado posible. Esto implica tener que evaluar el modelo construido en la fase anterior, revisar el proceso y el proyecto en general. Es importante asegurarse de que el modelo cumpla con los objetivos del proyecto y sea preciso con los resultados. Si alguno de los resultados no fuera el esperado se tendrá que regresar a la primera fase, como muestra la figura 3.1.1, para analizar las variables relevantes que no se han tenido en cuenta a la hora de desarrollar el modelo.

Para la arquitectura del sistema hay que utilizar controles de calidad entre las etapas para verificar que el procesamiento de los datos es exitoso (PowerData, 2017). Estos controles deben incluir testeos, proceso en el que se ejecuta un programa con el objetivo de encontrar errores. El primer paso sería la comprobación de los datos, asegurar que todas las fuentes de datos se han insertado correctamente al sistema. Se suelen utilizar softwares para comprobar que los datos de origen y los introducidos en el sistema son los mismos. Por cada ETL, se realiza un test para verificar que todos los datos que queremos extraer estén almacenados y ordenados como se requería. En esta revisión suelen concurrir la mayoría de los fallos, por ello hay que intentar dar una solución eficaz. Los errores más comunes son la duplicidad de algunos datos y sobreescribir celdas de datos ya utilizadas.

En la fase 4 (Modelos), dependiendo del modelo que se utilice, se podrá saber si no se ha tenido en cuenta alguna variable importante. En los modelos predictivos se utiliza una métrica de error para evaluar la calidad de los modelos. Las métricas más comunes son por un lado, el error cuadrático medio (ECM) que mide la media de los errores al cuadrado entre las predicciones y los valores reales. Por otro lado, la raíz del error cuadrático medio (RECM) que se utiliza para interpretar la magnitud del error en la misma unidad que los valores de la variable objetivo. Cuanto mayor sea el error más lejos se está de tener un modelo idóneo para la predicción, gracias a estas métricas se podrá saber si algún parámetro no se ha tenido en cuenta o si algún parámetro no tiene suficiente relevancia en el modelo.

Posteriormente, hay que comprobar el rendimiento del sistema verificando cómo de rápido se realiza cada extracción y las transformaciones de cada ETL o sistema de carga de cualquier proceso. Por último, si la evaluación se considera apta en los diferentes modelos utilizados y que concuerdan con los objetivos marcados, se procederá a planificar cómo presentar los resultados.

3.5. Despliegue y análisis de los datos obtenidos

En la etapa final de la metodología CRISP-DM es la encargada de visualizar los datos y distribuir estos informes a los diferentes departamentos de la empresa que necesiten esa información. Para ello, se va a utilizar Microsoft Power BI (ver apéndice Tabla A.5), una plataforma intuitiva que cualquier persona sin conocimientos técnicos puede utilizar. Esta plataforma permite desarrollar informes interactivos mezclando diferentes fuentes de datos para poder visualizar información enriquecedora para la empresa. Los informes pueden estar formados por los paneles que se deseen analizar con tablas, gráficas circulares, gráficas de barras, gráficas de líneas, etc.

Para mostrar mejor cómo funciona esta herramienta de Microsoft, se van a realizar dos diferentes tipos de informes. Por un lado, se va a utilizar un archivo que incluye más de dos millones de filas de datos de ventas de la empresa ficticia Contoso Inc (*Download ContosoSalesForPowerBI from Official Microsoft Download Center*, s.f.). Dado que con nuestra simulación en este trabajo no se pueden obtener datos de ventas de una empresa, se ha optado por utilizar esta base de datos relacional que ofrece Microsoft gratuitamente para aprender a utilizar PowerBI.

Esta base de datos usa un modelo E-R (entidad-relación) que permite representar interrelaciones y propiedades de cada entidad. Este tipo de base de datos cuenta con diferentes tablas que están relacionadas entre sí mediante un atributo clave que permite identificar cada registro de una tabla dentro de la base de datos . Estas tablas pueden ser de dos tipos: Fact (Hechos) o Dim (Dimensiones) (Loshin, 2018). Las tablas de tipo Fact contienen eventos que pueden medirse como pueden ser las ventas, las visitas o los beneficios. En cambio, las tablas de tipo Dim son tablas que sirven para segmentar los eventos como por ejemplo los clientes, los proveedores, las ciudades o la competencia. Cada tabla Dim está relacionada mediante un código con las tablas Fact con las que tenga relación. Este sistema de relacionar tablas es el mismo que en otras plataformas de gestión de base de datos como puede ser Access.

El lenguaje para operar dentro de PowerBI es el mismo que el de las fórmulas de Excel. A este lenguaje se le llama Dax (Data Analysis Expressions), un lenguaje específico para análisis de datos creado por Microsoft en el año 2010 (López, 2018). En el caso de este trabajo se han utilizado diversas medidas para obtener datos interesantes para analizar. Se han creado informes diferentes (Ingresos, Devoluciones y Tiendas). Todos los informes son paneles dinámicos, lo que quiere decir que si, por ejemplo, se desea conocer información de un producto en concreto, con seleccionarlo en una de los elementos del panel, todas las demás gráficas y tablas representarán los datos relacionados con ese producto. El resultado de estos informes se encuentra en el Anexo 6.

Comenzando por el informe de las devoluciones (ver figura A6.1), se han representado dos tablas sobre el ratio de devolución y precio unitario de los productos, un filtro por continentes y por último dos gráficas, una de ellas una gráfica de barras por ingresos generados por cada categorías de productos y la otra una gráfica temporal de cómo han ido evolucionando los ingresos totales en los años.

En segundo lugar se encuentra el informe de ingresos (ver figuras A6.2) que consta de dos tablas y una serie de gráficos. En una de las tablas se muestra el total de ingresos por cada categoría de producto y en la otra los ingresos de cada mes comparando si la variación es positiva o negativa con respecto al mismo mes del año anterior. También se han realizado diferentes gráficas sobre los ingresos por continentes, por años, por canal y total de ingresos en categorías.

El último informe (ver Figura A6.3) contiene información referente a las tiendas físicas siguiendo una estructura muy parecida a la de los dos informes anteriores. Por ejemplo, una de las tablas indica cuántas tiendas hay por cada ciudad y en una de las gráficas se muestra cuántos ingresos totales han generado en cada una de las ciudades.

Por otro lado, se ha utilizado la información obtenida vía Google Analytics y la ETL de los datos internos del Wordpress para realizar otro informe sobre la página web. Al no contar con muchos datos para el análisis, solamente se han podido utilizar los datos del stock y las visitas de la página web. En este caso, se ha utilizado un mapa del mundo con diferentes colores para saber la cantidad y de dónde son los usuarios que entran en la página web. También se han realizado diferentes gráficas como por ejemplo el stock medio para cada uno de los productos o la captación de usuarios por cada ciudad. El resultado se

muestra en en la figura A7.1 del Anexo 7. Los datos que contiene este anexo no son del todo representativos al no ser datos de una página web habilitada para la venta.

Estos han sido algunos ejemplos para ilustrar cómo funciona la plataforma de Power BI. Hay muchos más datos recabados mediante las ETLs que también deben ser utilizados para los informes. Para el desarrollo de estos informes, lo mejor es contar con personal de los diferentes departamentos de la empresa para crear un equipo multidisciplinar que pueda facilitar el despliegue de estos informes dinámicos.

Una vez explicada la fase de despliegue, la metodología de CRISP-DM habrá finalizado, lo que quiere decir que se ha logrado crear un sistema Big Data totalmente funcional. Esto no quiere decir que no haya que hacer labores de mantenimiento y mejoras ya que siempre saldrán nuevas fuentes de datos enriquecedoras para la empresa. Por ello, siempre hay que ir actualizando según las necesidades que vayan surgiendo. También mencionar, que existen otros softwares para el desarrollo de un sistema de Big Data, como por ejemplo Apache Hadoop o Apache Spark[22] mundialmente conocidos por su uso para el desarrollo de estos sistemas. Pero para el caso de una pyme puede ser menos adecuado su uso ya que, estos softwares son utilizados para operar con cantidades gigantescas de datos y requieren de una capacidad financiera muy elevada. Por ello, para este trabajo se han utilizado otras herramientas para el procesamiento de los datos. Estos dos softwares están implementados dentro de Google Cloud mediante Dataproc[23], así que en un futuro se podría implementar siempre y cuando sean totalmente necesarios. Hay que ser prudentes con el uso de todas las herramientas de Cloud Computing ya que su uso puede ser muy costoso.

4. Costes de mantenimiento del sistema y reflexión sobre ODS

En cuanto a los costes de mantenimiento del sistema, los gastos para este trabajo han rondado los 20 euros al mes y 10 euros iniciales. En el caso de una empresa real serán más elevados dependiendo de la cantidad de datos que se necesiten procesar o al sector al

[22]Son marcos de código abierto que se utilizan para administrar y procesar grandes volúmenes de datos para su análisis.

[23]Es un servicio de Google Cloud para administrar Apache Hadoop, Apache Spark, Apache Pig y Apache Hive.

que pertenezca la empresa. En la Tabla 4.1 se presenta una aproximación de los gastos generados en este trabajo para el proceso de digitalización y creación del sistema Big Data por cada una de las aplicaciones utilizadas. Claramente, estos costes serán más elevados en un caso real sobre todo por la densidad de los datos.

Tabla 4.1. Costes aproximado de las herramientas digitales

Herramienta	Servicio	Coste
Hostinger	Host de la página Web	36€ anuales
Hostinger	Dominio de la página Web	10€ anuales
Hubspot	Gestión de las relaciones con los clientes	240€ anuales
Google Cloud	Container Registry	120€ anuales
Google Cloud	Cloud Run	600€ anuales
Google Cloud	Cloud Scheduler	24€ anuales
Google Cloud	Cloud SQL	912€ anuales
Google Ads	Mercadotecnia en motores de búsqueda	1800€ anuales
Meta Ads	Mercadotecnia en motores de búsqueda	600€ anuales
Total		4.342€ anuales

Fuente: Elaboración propia

Como ya se ha mencionado anteriormente, dependiendo de las necesidades de uso de Google Cloud, el coste varía. El cálculo de este coste lo facilita Google Cloud mediante una aplicación llamada “calculadora de precios”. Esta aplicación permite saber una aproximación de los gastos mensuales por la utilización de cada uno de los servicios.

En el cálculo de los costes se ha incluido el servicio de Google SQL, un servicio de Google Cloud que permite almacenar base de datos relacionales (en el trabajo se ha utilizado la hoja de cálculo de Google Drive pero en el caso de una empresa real es más eficiente utilizar este servicio). Por último, se han incluído los costes de las herramientas de publicidad como Google Ads y Meta Ads que al igual que Google SQL aunque no se hayan

utilizado, son fundamentales para las empresas si quieren operar en el mundo digital. En estos dos últimos, se recomienda invertir entre el 5% y el 10% de los ingresos en publicidad pero siempre dependerá de la empresa.

Reflexión sobre el trabajo y los Objetivos de Desarrollo Sostenible

El ODS 9 se centra en construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible, y fomentar la innovación. En este contexto, mi proyecto subraya la importancia de proporcionar a las pymes las herramientas necesarias para integrar tecnologías avanzadas como el Big Data. Este tipo de tecnología no solo ayuda a las empresas a prosperar, sino que también contribuye al desarrollo económico inclusivo y sostenible. Al mejorar la infraestructura digital y promover la adopción de tecnologías innovadoras, se facilita la creación de industrias más resilientes y capaces de adaptarse a los cambios del mercado. Una investigación reciente de la OCDE, muestra como el ritmo de los beneficios de la productividad digital global ha estado por debajo del crecimiento económico general. Una de las razones que se pueden destacar de la investigación es la brecha en la digitalización entre las grandes empresas y el resto del ecosistema industrial (*OECD*, s.f.).

Un aspecto crucial que he abordado en mi Trabajo de Fin de Grado es la necesidad de apoyo y formación para las pymes en la adopción de estas tecnologías. Sin un conocimiento adecuado y una infraestructura de apoyo, muchas pequeñas y medianas empresas pueden quedar rezagadas, ampliando la brecha digital. Por ello, es fundamental que las políticas públicas y las iniciativas privadas se alineen para proporcionar los recursos y la capacitación necesarios. Este trabajo espero que sirva como inspiración y guía para todas aquellas empresas que lo necesiten.

En resumen, mi TFG destaca cómo la digitalización y la implementación de sistemas de Big Data pueden ser motores clave para el desarrollo sostenible de las pymes, alineándose con los objetivos del ODS 9. Este proyecto no solo ha aumentado mi comprensión sobre la integración tecnológica en el ámbito empresarial, sino que también ha reforzado mi compromiso con la promoción de prácticas sostenibles.

5. Conclusiones

En este Trabajo de Fin de Grado he estudiado y desarrollado un ecosistema Big Data, una tecnología fundamental en el mundo actual de la información, adaptado específicamente para una pequeña y mediana empresa,. A lo largo de este proyecto, se ha explicado en detalle la arquitectura, las tecnologías y las metodologías necesarias para construir un sistema Big Data eficaz a la hora de dar solución al desafío de grandes volúmenes de datos. A pesar de los recursos limitados con los que cuentan las pymes, la implementación de soluciones de Big Data puede ser factible y altamente beneficiosa. Destacar la importancia que tiene la correcta recolección, almacenamiento y procesamiento de grandes volúmenes de datos, así como su análisis avanzado para obtener conocimientos valiosos que puede ayudar a tomar mejores decisiones. Sin embargo, es importante reconocer que el desarrollo y la implementación de un sistema Big Data no es una tarea sencilla, sobre todo en la capacitación del personal y la garantía de seguridad y privacidad de los datos, estos aspectos deben abordarse cuidadosamente para garantizar el éxito a largo plazo de la iniciativa.

En resumen, el trabajo ha proporcionado una visión general de un sistema Big Data y las herramientas necesarias para su desarrollo, destacando los grandes desafíos que han ido surgiendo en su elaboración con los que he logrado adquirir nuevos conocimientos a la par de ir solventando cada uno de ellos. Por último, hacer hincapié en los avances de la era digital y la necesidad de adaptarse a ella. El papel del Big Data seguirá siendo fundamental en la generación de conocimiento y en la toma de decisiones, lo que hace primordial la necesidad de seguir explorando y desarrollando nuevas soluciones en este campo en constante evolución. En base a las limitaciones encontradas a la hora de realizar este trabajo, considero que habría que realizar diferentes trabajos futuros para poder obtener unos resultados concretos. Por un lado, se pueden implementar las técnicas Ensemble Learning para desarrollar sistemas de Big Data más eficientes. Por otro lado, este trabajo no está especializado en ningún sector concreto, por ello sería interesante analizar cómo afecta el Big Data en diferentes sectores profesionales. Para ello, habría que colaborar con una empresa para poder analizar datos reales.

Referencias

¿Qué es ETL? (s.f.). Www.sas.com. Recuperado el 7 de marzo de 2024 de https://www.sas.com/es_es/insights/data-management/what-is-etl.html#:~:text=ETL%20es%20un%20tipo%20de

Acelera pyme (s.f.) Kit Digital. Recuperado el 22 de octubre de 2024 de https://www.acelerapyme.gob.es/kit-digital

Ali, S. (2023, 7 de diciembre). *Bagging vs. Boosting in machine learning*. Educative. https://www.educative.io/blog/bagging-vs-boosting-in-machine-learning

B., G. (2017, 2 de mayo). *Cómo agregar Google Analytics a WordPress - Guía para principiantes*. Tutoriales Hostinger. https://www.hostinger.es/tutoriales/agregar-google-analytics-wordpress

BasuMallick, C. (2022, 26 de agosto). *What is a Data Lake? Definition, Architecture, Tools, and Applications*. Spiceworks. https://www.spiceworks.com/tech/cloud/articles/what-is-data-lakes/

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1007/bf00058655

Buyya,R., Yeo, C. S., Venugopal, S., Broberg, J. y Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the5th utility. *Future Generation Computer Systems, 25 (6), 599-616.* https://doi.org/10.1016/j.future.2008.12.001

Cabrera, A. (2021, 15 de marzo). *¿Cuánto invertir en Google Ads? La cantidad recomendada*. Atecnis. https://www.atecnis.com/cuanto-invertir-en-google-ads/

Chapman, P. , Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R.(1999). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Cukier, K. (2010, 27 de febrero). *Data, data everywhere*. The Economist. https://www.economist.com/special-report/2010/02/27/data-data-everywhere

Díaz, A., Gonzales, J. C., y Ruiz, M. E. (2005). Implantación de un sistema ERP en una organización. *RISI*, *2*(3), 30-37. https://gc.scalahed.com/recursos/files/r161r/w24108w/S12_04.pdf

Diebold, F. X. (2012). *A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline, Second Version*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843

Download ContosoSalesForPowerBI from Official Microsoft Download Center. (s.f.). Microsoft Store - Download Center. Recuperado el 25 de marzo de 2024 de https://www.microsoft.com/en-us/download/details.aspx?id=46801

Escobar, A. (2016, julio). *ECONOMÍA ON DEMAND*. Linkedin. https://www.linkedin.com/pulse/econom%C3%ADa-demand-angeles-escobar/

Eslamijam, M. (2022, 16 de julio). *What is boosting in machine learning?*. Tech Talks. https://bdtechtalks.com/2022/07/16/what-is-boosting-in-machine-learning/

Fan, J., Han, F., y Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Fernández, R., Costa, J., y Oviedo, M. (2021). 3 Bagging y boosting. En *Aprendizaje Estadístico*. https://rubenfcasal.github.io/aprendizaje_estadistico/

Fernández, V., Leyton, J., y González, A. (2010). *Cloud Computing.* http://www.profesores.elo.utfsm.cl/~agv/elo322/1s10/project/reports/cloudcomputing-10s01.pdf

Fresnadillo, S., y López, B. (2018). Marketing Digital: *la digitalización de empresas y sus efectos* (Universidad de Córdoba, Ed.). helvia.uco.es. https://helvia.uco.es/bitstream/handle/10396/17641/raydem_2_4.pdf?sequence=1&isAllowed=y

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5) 1189-1232. https://doi.org/10.1214/aos/1013203451

Gandomi, A., y Haider, M. (2015). Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, *35*(2), 137-144. https://www.sciencedirect.com/science/article/pii/S0268401214001066

Gartner Glossary. (s.f). *Big Data*. Gartner. https://www.gartner.com/en/information-technology/glossary/big-data

Gibson, P. (s.f.). *Types of Data Analysis*. Data Tutorials. Data Analysis; Chartio. Recuperado el 12 de mazo de 2024, de https://chartio.com/learn/data-analytics/types-of-data-analysis/

Google Cloud (s.f.) ¿Qué es Cloud Run? Documentación de Google Cloud Recuperado el 25 de marzo de 2024, de https://cloud.google.com/run/docs/overview/what-is-cloud-run?hl=es-419

Grupo PowerData. (2011). *Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.* Powerdata.es. https://www.powerdata.es/big-data

Haya, P. (2021, 29 de noviembre). *La metodología CRISP-DM en ciencia de datos - IIC*. Instituto de Ingeniería Del Conocimiento. https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

Hugging Face (s.f.) What is Zero-Shot Classification? Recuperado el 23 de febrero de 2024, de https://huggingface.co/tasks/zero-shot-classification

IBM. (s.f.). *¿Qué es el procesamiento del lenguaje natural (NLP)?.* https://www.ibm.com/es-es/topics/natural-language-processing

Impulsa ecommerce (2023, 6 de septiembre).*Ranking de los Mejores Creadores de Páginas Web (2024)*. https://impulsaecommerce.com/creadores-de-paginas-web/

Instituto Nacional de Estadística (2023, 25 de octubre). *Encuesta sobre el uso de TIC y del comercio electrónico en las empresas, Año 2022* – Primer trimestre 2023[Nota de prensa]. Recuperado de https://www.ine.es/prensa/tic_e_2022_2023.pdf

Iouf, P. S., Boly, A.. y Ndiaye, S. (2018). Variety of data in the ETL processes in the cloud: State of the art. *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*. https://doi.org/10.1109/icird.2018.8376308

Javier. (s.f.). *Google Ads; ¿Qué precio, tarifas y cuánto cuesta la gestión de publicidad?*. Xplora. Recuperado el 11 de febrero de 2024 de https://www.xplora.eu/precio-google-ads/

Kearns, M., y Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM*, *41*(1), 67-95. https://doi.org/10.1145/174644.174647

Larrosa, C. (2023, 2 de febrero). *Aprendizaje supervisado en Machine Learning.* | *Algoritmos, ejemplos*. Datarmony. https://www.datarmony.com/blog/aprendizaje-supervisado-algoritmos-ejemplos/

Lee, W. M. (2019). *Python Machine Learning*. John Wiley & Sons.

López, N.(2018, 13 de noviembre). *Introducción al lenguaje DAX*. Dataxbi. https://www.dataxbi.com/blog/2018/11/13/introduccion-lenguaje-dax/

Losada, C. (2022, 27 de septiembre). *4 tipos de análisis de datos que debes conocer*. Making Science. https://www.makingscience.es/blog/4-tipos-de-analisis-de-datos-que-debes-conocer/

Loshin, D. (2018, 27 de marzo). *Tablas de dimensión vs tablas de hechos: ¿Cuál es la diferencia?* ComputerWeekly. https://www.computerweekly.com/es/consejo/Tablas-de-dimension-vs-tablas-de-hechos-Cual-es-la-diferencia

Lualouro. (2022, 22 de noviembre). *Blocksy — Tema para WordPress gratis, completo y fácil de usar*. Lúa Louro. https://lualouro.com/blocksy/#Que_es_Blocksy

Mendonça, M. o. K., Netto, S. L., Diniz, P. S. R., y Theodoridis, S. (2024). Machine learning. En *Elsevier eBooks* (pp. 869-959). https://doi.org/10.1016/b978-0-32-391772-8.00019-3

Montoya Agudelo, C. A., y Boyero Saavedra, M. R. (2013). *El CRM como herramienta para el servicio al cliente en la organización. Visión de Futuro, 17(1).* http://www.scielo.org.ar/scielo.php?pid=S1668-87082013000100005&script=sci_arttext

Nast, C. (2023, 14 de julio). *Demanda de X Corp., de Musk, culpa al "data scraping" por sobrecargar servidores de Twitter*. WIRED. https://es.wired.com/articulos/demanda-de-x-corp-culpa-al-data-scraping-por-sobrecargar-servidores-de-twitter

NewVantage Partners, y Wavestone. (2023, 25 de diciembre). State of big data/AI adoption in organizations worldwide from 2018 to 2023 [Graph]. In Statista. Recuperado el 30 de mayo, 2024, de https://www.statista.com/statistics/742993/worldwide-survey-corporate-disruptive-technology-adoption/

Observatorio Nacional de Tecnología y Sociedad. (2023). Uso de inteligencia artificial y big data en las empresas españolas. Secretaria de Estado de Digitalización e Inteligencia Artificial. Red.es. Obtenido de: https://www.ontsi.es/sites/ontsi/files/2023-02/Br%C3%BAjula_IA_Big_data_2023.pdf

OCI Oracle Cloud Infrastructure *(s.f.) ¿Qué es Big Data? Recuperado el 21 de diciembre de 2023, de* https://www.oracle.com/es/big-data/what-is-big-data/

OCI Oracle Cloud Infrastructure (s.f.) ¿Qué es una base de datos relacional (sistema de gestión de bases de datos relacionales)? Recuperado el 26 de marzo de 2024, de https://www.oracle.com/es/database/what-is-a-relational-database/

OECD (s.f.).*Digitalisation and productivity.* Www.oecd.org.
https://www.oecd.org/economy/growth/digitalisation-productivity-and-inclusiveness/

PowerData (s.f.). *Transformación digital. Qué es y su importancia y relación con los datos.* En la biblioteca Digital de Powerdata.es. Recuperado el 28 de febrero de 2024, de https://www.powerdata.es/transformacion-digital

PowerData *(2017, 27 de febrero). Principales métodos para data testing de Big Data.* Recuperado el 15 de marzo de 2024, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/principales-metodos-para-data-testing-de-big-data

Ramírez, L. (2023). *Guía completa para digitalizar una empresa. Thinking for Innovation.* https://www.iebschool.com/blog/como-digitalizar-una-empresa-en-10-pasos-tecnologia/

Ray, S. (2024, 19 de marzo). *Top 10 Machine Learning Algorithms to Use in 2024*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Redhat (2023, 20 de enero) ¿Qué es Docker? Recuperado el 10 de abril https://www.redhat.com/es/topics/containers/what-is-docker

Redhat (2023, 4 de agosto) Diferencias entre IaaS, PaaS y SaaS. Recuperado el 28 de noviembre de 2023 de https://www.redhat.com/es/topics/cloud-computing/iaas-vs-paas-vs-saas

Repsol (2023, 2 de octubre) El reto de la transformación digital. Recuperado el 20 de marzo de 2024, de https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/digitalizacion-de-empresas/index.cshtml

Roa, M. M. (2021, 22 octubre). El big bang del big data. *Statista Daily Data*. https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/

Roberto, C. (2022, 31 de mayo). *Crisp-DM: las 6 etapas de la metodología del futuro. Blog MBA Esalq USP.* https://blog.mbauspesalq.com/es/2022/05/31/crisp-dm-las-6-etapas-de-la-metodologia-del-futuro/

Ruiz, P. (2022, 5 de abril). *¿Por qué las pequeñas empresas necesitan el big data?.* Recuperado el 28 de febrero de 2024, de https://www.holded.com/es/blog/big-data-empresas

Santos, P. R. de los. (2021, 2 de diciembre). *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. Telefónica Tech. Recuperado el 18 de marzo de 2024, de https://telefonicatech.com/blog/que-algoritmo-elegir-en-ml-aprendizaje

SiliconANGLE. (2018, 9 de marzo). Volumen de ingresos del sector de big data a nivel mundial de 2016 y 2027, por área de negocio (en miles de millones de dólares) [Gráfica]. En Statista. Recuperado el 27 de mayo de 2024, de https://es.statista.com/estadisticas/601155/ingresos-del-sector-big-data/

Simeone, O. (2018). A Very Brief Introduction to Machine Learning With Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking*, *4*(4), 648–664. https://doi.org/10.1109/tccn.2018.2881442

Singh, A. (2023, 22 de noviembre). *A Comprehensive Guide to Ensemble Learning (with Python codes)*. Analytics Vidhya. Recuperado el 27 de marzo de 2024, de https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

Silva, L. (2022). *Qué es el mailing, cómo hacerlo y ejemplos exitosos*. Blog.hubspot.es.Recuperado el 15 de noviembre de 2023, de https://blog.hubspot.es/marketing/que-es-mailing

Soam, T. (2018, 26 de noviembre). *Cron Expression: Tutorial*. Medium. https://medium.com/@tushar0618/cron-expression-tutorial-721d85e4c2a7

Sotaquirá, M. (2023, 2 de junio). *Parámetros e hiperparámetros en el Machine Learning*. Codificando Bits. Recuperado el 9 de marzo de 2024, de https://www.codificandobits.com/blog/parametros-hiperparametros-machine-learning/

Statistics Explained. (2016). Europa.eu. https://ec.europa.eu/eurostat/statistics-explained/index.php?

Team, K. (2022, 17 de mayo). *¿Qué es Google Container Registry? | KeepCoding Bootcamps*. Keepcoding.io. Recuperado el 10 de marzo de 2024, de https://keepcoding.io/blog/que-es-google-container-registry/

TechAmerica Foundation's Federal Big Data Commission. (2012).*Demystifying Big Data: A Practical Guide To Transforming The Business of Government* https://breakinggov.sites.breakingmedia.com/wp-content/uploads/sites/4/2012/10/TechAmericaBigDataReport.pdf

Torres i Viñals, J. (2012) "Del Cloud Computing al Big Data: visión introductoria para jóvenes emprendedores," Editor: Universitat Oberta de Catalunya (UOC). Biblioteca Virtual Campus Europeo, consulta 6 de marzo de 2024, de https://bibliotecavirtual.campuseuropeo.es/document/16.

UNIR. (2020). *Las 3 V del Big Data y el procesamiento de datos.* Recuperado el 28 de febrero de 2024, de. https://www.unir.net/ingenieria/revista/3-v-big-data/

Vargas, D. (2021, 30 de junio). *Qué es una pasarela de pago: beneficios y las 7 mejores del mercado*. Tutoriales Hostinger. https://www.hostinger.es/tutoriales/pasarela-de-pago

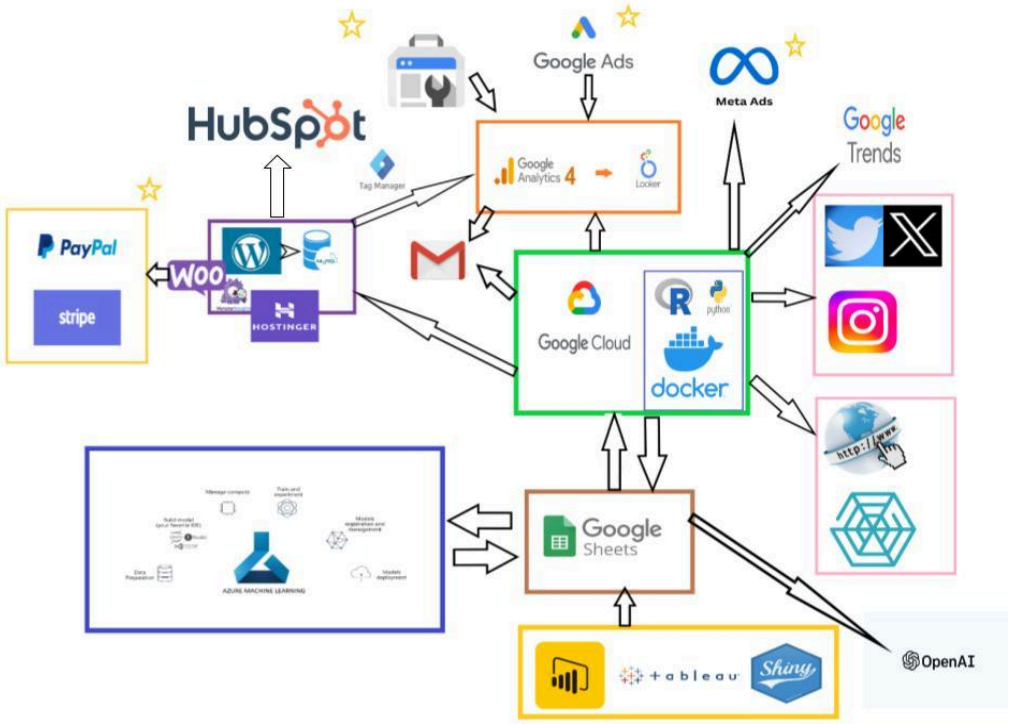
Xun Yuang, L. (2023, 23 de mayo).*distilbert-base-multilingual-cased-sentiments-student · Hugging Face*. https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

Wikipedia contributors. (2023, 12 de diciembre). *Tag cloud*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Tag_cloud&oldid=1189486436

Apéndice: Aplicaciones para la digitalización y desarrollo del ecosistema Big Data

En este apéndice se explican todas las aplicaciones o plataformas utilizadas para la realización del trabajo. Se han mencionado algunas aplicaciones que no se han utilizado pero se aconseja utilizar en un caso real para llevar a cabo una digitalización empresarial. El apéndice está organizado por áreas, cada área engloba las aplicaciones usadas en un proceso dentro del ecosistema del Big Data. En la figura A1 se muestra la arquitectura de las aplicaciones para el desarrollo del trabajo y como están conectadas entre ellas. La dirección de las flechas indica de donde se extraen los datos. Cada color hace referencia a un área del trabajo. El color morado al desarrollo página web, color verde a las aplicaciones para los procesos de extracción de datos, color marrón a la base de datos, color naranja a las aplicaciones para extraer los datos de la página web, color rosa a las fuentes externas de información, color azul al Learning machine de Azure para la creación los modelos de análisis de datos[24] y por ultimo el color amarillo al software de visualización de datos interactivo.

Figura A1: Arquitectura de aplicaciones para la digitalización y sistema Big Data



Fuente: Elaboración Propia

[24] Esta plataforma finalmente no se ha utilizado en este trabajo pero es de gran utilidad para generar modelos de análisis propios.

Tabla A1: Desarrollo de una página web

Icono	Nombre	Definición y Uso
	Wordpress	Plataforma de gestión de contenidos de sitios web y en la que se puede publicar y editar contenido web. Es mundialmente utilizada y conocida por su interfaz de usuario intuitiva y fácil de usar.
	Hostinger	Servidor para alojar la página web en la red y tener un dominio propio. Este servicio es vital para poder existir en la red, ya que toda página web debe de estar bajo un servidor y Hostinger es una de los mayores proveedores de alojamiento web en la actualidad. Ofrecen servicios de hosting compartido, hosting en la nube, VPS (Servidores Virtuales Privados) y alojamiento de correo electrónico.
	MonsterInsights	Plugin de Wordpress el cual permite realizar análisis estadísticas de los datos recogidos del sitio web. Facilita la integración y el uso de Google Analytics en sitios web de WordPress.
	Woocommerce	Plugin de Wordpress para convertir un blog en una tienda web en la que se pueden vender tus productos de una manera sencilla, muy destacada entre las demás posibilidades ya que su precio es muy barato por lo que es muy usada por pequeñas empresas.
	Yoast SEO	Plugin de WordPress con el que se puede analizar diferentes configuraciones para mejorar el posicionamiento de nuestra página web. En el trabajo se utiliza para el desarrollo de los blogs y las diferentes plantillas de la página web para poder desarrollarlas respetando el SEO (ver Glosario).

Fuente: Elaboración propia

Icono	Nombre	Definición y Uso
	GTM4WP	Software de código abierto que facilita colocar fragmentos de código para modificar las etiquetas <head> y <body> de la página web. Este plugin es necesario para realizar la conexión entre Google Analytics y Wordpress.
	Blocksy	Tema para WordPress que ha sido creado pensado para ser utilizado junto con el maquetador de WordPress, Gutenberg (lualouro, 2022). Un tema es un conjunto de carpetas con archivos de plantillas, hojas de estilo, lenguaje de programación PHP, que hacen que tu sitio web tenga un determinado aspecto.
	Stripe	Sistema de pago online seguro el cual permite realizar pagos sin la necesidad de realizar transferencias bancarias. Por cada transacción se lleva un porcentaje de comisión.

Fuente: Elaboración propia

Tabla A2: Ecosistema aplicaciones para la extracción y carga de datos

Logo	Nombre	Definición y Uso
	Google analytics	Herramienta de análisis web desarrollada por Google que permite a los propietarios de sitios web y a los especialistas en marketing rastrear y analizar el rendimiento de sus sitios web. Esta plataforma proporciona una amplia gama de datos e información sobre cómo los usuarios interactúan con un sitio web, lo que ayuda a los propietarios de sitios y a los profesionales de marketing a comprender mejor el comportamiento de los visitantes y a tomar decisiones informadas para mejorar la experiencia del usuario y la eficacia de su estrategia en línea.
	Looker Studio	Herramienta online que sirve para convertir los datos en informes o paneles de información. Esta herramienta se utiliza para conseguir gráficamente datos de usuario de la página web.
	Google Tag Manager	Herramienta de administración de etiquetas en línea desarrollada por Google. Permite a los propietarios de sitios web y a los profesionales de marketing agregar, actualizar y administrar de manera más eficiente las etiquetas de seguimiento y fragmentos de código en sus sitios web sin tener que modificar directamente el código fuente del sitio.
	Gmail	Utilizaremos como correo electrónico gmail ya que lo necesitaremos para poder vincular las demás aplicaciones y para eso necesitamos una cuenta google. También lo utilizaremos para que nos lleguen diferentes tipos de alertas.

Fuente: Elaboración propia

Logo	Nombre	Definición y Uso

	Google Cloud	En el caso del trabajo se optó por utilizar Google Cloud para automatizar la recogida de datos y el almacenaje de estos como base de datos. Esta plataforma será la encargada de mantener en la nube todas las ETL y procesos necesarios para la extracción de datos. Los servicios utilizados son Container Registry, Cloud Run, Cloud Scheduler. Todos ellos son servicios de IaaS.
	Python	Python es un lenguaje de programación de alto nivel, interpretado y generalista. Es conocido por su sintaxis legible y su facilidad de uso. Se utiliza en una amplia variedad de aplicaciones, desde desarrollo web hasta análisis de datos, inteligencia artificial y más. En el caso de trabajo se utiliza como lenguaje de programación de las ETLs y los modelos de análisis de datos.
	Docker	Se trata de un proyecto de código abierto que automatiza el despliegue de aplicaciones dentro de contenedores de software, en el trabajo se utilizará para la extracción de datos. El propósito de los contenedores es ejecutar varios procesos por separado para que se pueda aprovechar mejor la infraestructura sin afectar a la seguridad. Esto permite dividir el código de programación en secciones como se ha realizado en el trabajo. Esta herramienta de contenedores, proporciona un modelo de implementación basado en imágenes permitiendo compartir fácilmente un conjunto de servicios. Es por esta razón por la que se utiliza Docker, pudiendo ejecutar los procesos desde cualquier ordenador.
	Google Sheets	Programa de hojas de cálculo, que se utilizará como almacenaje de los datos que se extraigan. Esta aplicación está disponible en la plataforma Google Drive.

Tabla A3: Herramientas para obtener datos externos

Logo	Nombre	Definición y uso
	Google Ads	Plataforma de publicidad en línea desarrollada por Google que permite a las empresas promocionar sus productos, servicios o contenido a través de anuncios en los resultados de búsqueda de Google o en cualquier plataforma de Google (como YouTube o sitios web de socios). Esta plataforma utiliza un modelo de publicidad de pago por clic (PPC), lo que significa que los anunciantes sólo pagan cuando un usuario hace clic en su anuncio o realiza una acción específica, como una compra o una descarga de la aplicación. También cuenta con herramientas de seguimientos y de análisis. Se pueden buscar las palabras clave (keywords) que más nos puedan interesar. Estas keywords se pueden utilizar para saber qué es lo que más se busca por los internautas y utilizarla para un análisis de datos. Esta plataforma se puede utilizar como una fuente externa de datos.
	Google Trends	Esta herramienta sirve para mostrar los términos de búsqueda más populares en las últimas horas. Una herramienta muy útil para saber qué es lo que buscan los usuarios y una buena fuente de información externa que se puede extraer vía API (ver Glosario). Se ha utilizado para saber cuantas veces y de donde se buscan los móviles eco-friendly.

Fuente: Elaboración propia

Tabla A4: CRM(*Customer Relationship Management)*

Logo	Nombre	Definición y uso
HubSpot	Hubspot	Plataforma para la gestión basada en relaciones con los clientes en la cual se combinan herramientas de inbound marketing, ventas y servicio al cliente. Ayudará a la empresa a atraer clientes potenciales, gestionar relaciones con clientes, automatizar procesos de ventas y ofrecer un servicio al cliente eficiente.

Fuente: Elaboración propia

Tabla A5: Visualizar la información

Logo	Nombre	Definición y uso
	Power BI	Esta aplicación sirve para dar una solución de análisis empresarial, que permite unir diferentes fuentes de datos, analizarlos y presentar de una manera visual a través de informes y paneles. La encargada de visualizar en paneles los datos obtenidos. Pertenece al ecosistema de Microsoft.

Fuente: Elaboración propia

Tabla A6: Inteligencia artificial

Logo	Nombre	Definición y uso
	OpenAI	Es una organización líder en investigación en inteligencia artificial conocida en el desarrollo de modelos de lenguaje altamente avanzados. Su objetivo es promover el avance de la IA de manera ética y segura, al tiempo que brinda acceso a herramientas y recursos para desarrolladores y la comunidad en general. Utilizado para el desarrollo de la página web, en la creación de blogs y descripciones.

Fuente: Elaboración propia

Glosario

Denominación en Inglés	Denominación en Castellano	Significado
Application Programming Interface (API)	Interfaz de Programación de Aplicaciones	Una pieza de código que permite a diferentes aplicaciones comunicarse entre sí y compartir información y funcionalidades.
Big Data	Macrodatos	Conjuntos de datos tan grandes y complejos que precisan de aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente.
Bagging	Embolsado	Meta-algoritmo diseñado para mejorar la estabilidad y precisión de los algoritmos de aprendizaje automático utilizados en la clasificación y regresión estadística. También reduce la variación y ayuda a evitar el sobreajuste.
Boosting	Impulsando	Un metaalgoritmo conjunto para reducir principalmente el sesgo, la variación en el aprendizaje supervisado y una familia de algoritmos de aprendizaje automático que convierten a los estudiantes débiles en fuertes.
Business Intelligence (BI)	Inteligencia de Negocio	Conjunto de estrategias, aplicaciones, datos, productos, tecnologías y arquitectura técnicas, los cuales están enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.
Cloud Computing	Computación en la nube	Es el uso de una red de servidores remotos conectados a internet para almacenar, administrar y procesar datos, servidores, bases de datos, redes y software.
Container	Contenedor	Los contenedores son básicamente un entorno de computación usualmente en la nube, totalmente funcional y portátil. Individualmente, cada contenedor simula una aplicación de software diferente y ejecuta procesos aislados.

Fuente: Elaboración propia con la ayuda de Wikipedia

Denominación en	Denominación en	Significado

Inglés	Castellano	
Customer Relationship Management (CRM)	Gestión de Relación con los Clientes	Sistemas informáticos de apoyo a la gestión de las relaciones con los clientes, a la venta y al marketing. Un software para gestionar las ventas y los clientes de la empresa.
Data Scraping	Raspador de dato	Transferencia de datos entre programas para lograr utilizar estructuras de datos adecuadas para el procesamiento automatizado por computadoras.
Data warehouse	Almacén de datos	Es donde se almacena una colección de datos orientada a un determinado ámbito de la empresa, estos datos son integrados, no volátiles y variables en el tiempo, que ayudan a la toma de decisiones.
Dim	Dimensión	Un tipo de tablas que sirven para segmentar los eventos de una base de datos.
E-Commerce	Comercio electrónico	Este comercio está constituido por transacciones comerciales que se realizan a través de Internet.
Endpoint	Punto final	Cualquier punto que sea la parte final de una red (Móviles, PC, sensores…)
Ensemble learning	Aprendizaje por conjuntos	Es un conjunto de métodos que utilizan múltiples algoritmos de aprendizaje para obtener un rendimiento predictivo mejor que el que podría obtenerse con cualquiera de los algoritmos de aprendizaje constituyentes por sí solos.
ETL	Extract, Transform and Load	Es un proceso de tres fases en el que los datos se extraen de una fuente de entrada, se transforman (incluida la limpieza) y se cargan en un contenedor de datos de salida.
Enterprise resource planning (ERP)	Planificación de recursos empresariales	Sistemas de gestión de información que automatizan muchas de las prácticas de negocio asociadas con los aspectos operativos o productivos de una empresa.
Fact	Hechos	Un tipo de tablas que contienen eventos que pueden medirse como pueden ser ventas, visitas o beneficios.

Fuente: Elaboración propia con la ayuda de Wikipedia

Denominación en Inglés	Denominación en Castellano	Significado
Infrastructure as a Service (IaaS)	Infraestructura como un servicio	Un modelo de computación en la nube por medio del cual el proveedor de servicios en la nube suministra recursos de computación tales como el almacenamiento, la red, los servidores y la virtualización.
Identity document (ID)	Identificador	Símbolos léxicos que nombran entidades. Nombrar las entidades hace posible referirse a las mismas, lo cual es esencial para cualquier tipo de procesamiento simbólico.
Lead	Cliente potencial	Un cliente que reúne las características de un cliente ideal para una empresa vendedora, pero que no ha expresado todavía interés en sus productos o servicios.
Learning machine	Aprendizaje automático	Subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.
Large-Vocabulary Continuous Speech Recognition (LVCSR)	Reconocimiento Continuo de Voz de Amplio Vocabulario	Sistema de reconocimiento de voz el cual es capaz de dividir en fonemas que coinciden con palabras y frases en un diccionario para producir una transcripción de texto completo.
Mailing	Correo directo	Consiste en enviar información publicitaria por correo electrónico, más concretamente un folleto publicitario.
Mean square error (MSE)	Error cuadrático medio (ECM)	Es un estimador que mide el promedio de los errores al cuadrado, dicho de otra forma la diferencia que hay entre el estimador y lo que se estima.
Natural language processing (NLP)	Procesamiento de Lenguaje Natural	Mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural.
On-demand	Bajo demanda	Actividad económica creada por las empresas de tecnología que satisfacen la demanda de los consumidores a través de la provisión inmediata de bienes y servicios.

Fuente: Elaboración propia con la ayuda de Wikipedia

Denominación en Inglés	Denominación en Castellano	Significado
One Shot or Few shot classification	Clasificación de un o pocos disparos	Se utiliza en los modelos de aprendizaje automático, tiene como objetivo clasificar objetos a partir de uno o unos pocos ejemplos.
Platform as a Service (PaaS)	Plataforma como servicio	Un modelo de computación en la nube donde se ofrece la infraestructura y todo lo necesario para construir nuevas aplicaciones y servicios, lo que facilita la construcción y puesta en funcionamiento de las aplicaciones.
Payment gateway	Pasarela de pago	Servicio de un proveedor de servicios de aplicación de comercio electrónico, con el que se autorizan pagos a negocios electrónicos, ventas en línea al detalle.
Plugin	Complemento informático	Es una aplicación que permite extender las funciones de otra aplicación o programa sin tener que modificar el código.
Prospects	Prospecto	Una persona que aunque no haya dejado información sobre él es un posible cliente potencial.
Root Mean Squared Error (RMSE)	Raíz del error cuadrático medio (RECM)	Una medida de uso frecuente de las diferencias entre los valores de población predichos por un modelo y los valores observados.
Software as a Service (SaaS)	Software como servicio	Un modelo de computación en la nube por medio del cual el proveedor ofrece tanto la infraestructura hardware y los entornos de ejecución necesarios como los productos software que interaccionamos con el usuario desde un determinado portal o interfaz a través de Internet.
Search Engine Optimization (SEO)	Optimización de motores de búsqueda	Es un conjunto de acciones orientadas a mejorar el posicionamiento de un sitio web en los motores de búsqueda.
Search Engine Marketing (SEM)	Marketing de motores de búsqueda	Consiste en aumentar la visibilidad y posicionamiento de un negocio o página web en las páginas de resultados de motores de búsqueda mediante anuncios.

Fuente: Elaboración propia con la ayuda de Wikipedia

Denominación en	Denominación en	Significado

Inglés	Castellano	
Tag	Etiqueta	Una palabra clave asignada a un dato almacenado en un repositorio.
Tag Cloud	Nube de palabras	Representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia.
Video content analysis (VCA)	Análisis de contenido de vídeo	Capacidad de analizar automáticamente vídeo para detectar y determinar eventos temporales y espaciales.
Web Scraping	Raspado web	Es el proceso de extracción de datos de forma automática de una página web.
Zero-shot classification	Clasificación de Tiro Cero	Paradigma en el que un modelo puede clasificar ejemplos nuevos que pertenecen a clases que no estaban en los datos de entrenamiento.

Fuente: Tabla de elaboración propia y fuentes de información de Wikipedia

Anexos

Anexo 1: ETL Wordpress

La primera de todas las ETLs creadas para este trabajo es sobre los datos de la página web que están soportados en Wordpress. Para ello el primer paso es acreditar las credenciales de Google Drive para poder acceder a la base de datos donde queramos guardar los datos y también conectarse a la base de datos propia de wordpress que es de tipo Mysql, se pedirán el nombre del usuario, contraseña y el host. Por temas de seguridad se han omitido en el código los datos dichos anteriormente. Una vez conectadas las bases de datos solo quedaría traspasar los datos de una base a otra especificando qué datos se quieren obtener. En este caso los datos de productos, usuarios y comentarios. Hay que pedir que se actualicen los datos para que no se generen duplicidades. A continuación se muestra el código en Python.

```
scopes =
['https://www.googleapis.com/auth/spreadsheets','https://www.googleapis.com
/auth/drive']
credentials = Credentials.from_service_account_file('Token de drive',
scopes=scopes)
gc = gspread.authorize(credentials)
gauth = GoogleAuth()
drive = GoogleDrive(gauth)
#Poner los datos de autentificación
connection = mysql.connector.connect(
    user = 'USUARIO DE LA BASE DE DATOS',
    password = 'CONTRASEÑA',
    host = 'DIRECCIÓN',
    database = 'NOMBRE DE LA BASE DE DATOS')
#obtener datos de productos
query = "SELECT * FROM wp_wc_product_meta_lookup"
products = pd.read_sql(query, connection)
#obtener datos de usuarios
query = "SELECT * FROM wp_users"
users = pd.read_sql(query, connection)
#obtener datos de comentarios
query = "SELECT * FROM wp_posts"
posts = pd.read_sql(query, connection)
#Randomizar el valor de stock
cursor = connection.cursor()

# Definir la actualización de la consulta
update_query = """
UPDATE wp_wc_product_meta_lookup
SET stock_quantity = CAST(500 + (1500 * RAND()) AS INTEGER)
"""
# Actualizar la consulta
cursor.execute(update_query)
# Realizar cambios
```

```
connection.commit()
connection.close()
#-----------Import Data Google Sheets-----------

#Productos

spreadsheet = gc.open('Products')

worksheet = spreadsheet.worksheet('Products')

worksheet.update([products.columns.values.tolist()] +
products.values.tolist())

#Append data just to mantain the historic

df_values = products.values.tolist()

spreadsheet.values_append('Products', {'valueInputOption': 'RAW'},
{'values': df_values})

#Usuarios

spreadsheet = gc.open('Users')

worksheet = spreadsheet.worksheet('Users')

worksheet.update([users.columns.values.tolist()] + users.values.tolist())

#Comentarios

spreadsheet = gc.open('Posts')

worksheet = spreadsheet.worksheet('Posts')

worksheet.update([posts.columns.values.tolist()] + posts.values.tolist())

      print("Done")
```

Anexo 2: ETL de los comentarios de productos de Amazon

Para esta ETL de datos externos, se ha programado la orden de ir extrayendo la información de los productos que se desee de amazon, la lista de productos que se desea analizar se tiene que encontrar en una base de datos ya que los datos deben de estar totalmente estructurados. En el caso de este trabajo, se ha utilizado Google Sheet. Para ello se ha facilitado el nombre del producto y sus URL de venta en Amazon. Se han obtenido datos de las siguientes marcas y versiones de móviles: Samsung Galaxy S23 Ultra, SAMSUNG Galaxy S23, Apple iPhone 12 Pro Max, Apple iPhone 11 Pro Max, OnePlus Open, SAMSUNG Galaxy A25, OnePlus 8 Pro, Google Pixel 7a, Xiaomi Redmi Note 12, Xiaomi Poco X5, Xiaomi Redmi 12C, Google Pixel 7a, Google Pixel 4, Google Pixel 6a, Google Pixel 6, Google Pixel 8, Motorola Moto G Stylus, Motorola Edge | 2021, OnePlus Nord N30, SAMSUNG Galaxy Z Fold 5, SAMSUNG Galaxy A54, Samsung Galaxy S20 y Xiaomi Redmi 12.

Lo primero que se tiene que hacer es acreditar los datos de la base de datos propia y pedir la entrada a “amazon.com” para poder extraer los datos. Una vez el servidor de amazon haya aceptado la entrada, se comienza a realizar las peticiones de información. La primera petición es extraer todo el HTML de la página y después bajo órdenes se especifica cuales son los datos que se quieren obtener. Una vez determinado cuales son los datos que se desean extraer, se facilita una base de datos donde almacenar los registros. Para ello se ha utilizado hojas de cálculo de Google Drive.

A continuación se muestra cual es el código de programación en lenguaje Python para este caso.

```
scopes =
['https://www.googleapis.com/auth/spreadsheets','https://www.googleapis.com
/auth/drive']
credentials = Credentials.from_service_account_file(#Tokens de tu base de
datos, scopes=scopes)
gc = gspread.authorize(credentials)
gauth = GoogleAuth()
drive = GoogleDrive(gauth)
headers = {
    "authority": "www.amazon.com",
    "pragma": "no-cache",
    "cache-control": "no-cache",
    "dnt": "1",
    "upgrade-insecure-requests": "1",
    "user-agent": "Mozilla/5.0 (X11; CrOS x86_64 8172.45.0)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.64 Safari/537.36",
    "accept":
"text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apn
g,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9",
    "sec-fetch-site": "none",
    "sec-fetch-mode": "navigate",
    "sec-fetch-dest": "document",
    "accept-language": "en-GB,en-US;q=0.9,en;q=0.8",
}

def get_page_html(page_url: str) -> str:
    resp = requests.get(page_url, headers=headers)
    return resp.text


def get_reviews_from_html(page_html: str) -> BeautifulSoup:
    soup = BeautifulSoup(page_html, "lxml")
    reviews = soup.find_all("div", {"class": "a-section celwidget"})
    return reviews


def get_review_date(soup_object: BeautifulSoup):
  try:
    date_string = soup_object.find("span", {"class":
"review-date"}).get_text()
    return date_string
  except AttributeError as e:
        print("Error occurred while extracting DATE:", e)
        return "N/A"
```

```
def get_review_text(soup_object: BeautifulSoup) -> str:
    try:
      review_text = soup_object.find("span", {"class": "a-size-base
review-text review-text-content"}).get_text()
      return review_text.strip()
    except AttributeError as e:
        print("Error occurred while extracting STARS:", e)
        return "N/A"

def get_review_header(soup_object: BeautifulSoup) -> str:
  try:
    review_header = soup_object.find("a",{"class": "a-size-base
a-link-normal review-title a-color-base review-title-content
a-text-bold"}).get_text()
    return review_header.strip()
  except AttributeError as e:
        print("Error occurred while extracting title:", e)
        return "N/A"

def get_number_stars(soup_object: BeautifulSoup) -> str:
  try:
    stars = soup_object.find("span", {"class": "a-icon-alt"}).get_text()
    return stars.strip()
  except AttributeError as e:
        print("Error occurred while extracting STARS:", e)
        return "N/A"


def get_product_name(soup_object: BeautifulSoup) -> str:
  try:
    product = soup_object.find("a", {"class": "a-size-mini a-link-normal
a-color-secondary"}).get_text()
    return product.strip()
  except AttributeError as e:
        print("Error occurred while extracting product name:", e)
        return "N/A"  #se puede poner N/A o cualquier expresión que
signifique sin dato.

def orchestrate_data_gathering(single_review: BeautifulSoup) -> dict:
    return {
        "review_text": get_review_text(single_review),
        "review_date": get_review_date(single_review),
        "review_title": get_review_header(single_review),
        "review_stars": get_number_stars(single_review),
        "review_flavor": get_product_name(single_review),
    }

def row_has_max_length(row):
    return all(len(str(cell)) <= 50000 for cell in row)

if __name__ == '__main__':

    spreadsheet = gc.open('ASINS')
    worksheet = spreadsheet.worksheet('ASINS')

#Comentarios positivos recientes
    URLS = worksheet.col_values(2)
    all_results = []
    for u in URLS:
        print(u)
```

```
        time.sleep(3)
        html = get_page_html(u)
        reviews = get_reviews_from_html(html)
        for rev in reviews:

            data = orchestrate_data_gathering(rev)
            print(data)
            all_results.append(data)

    out = pd.DataFrame.from_records(all_results)

    out= out[out.apply(row_has_max_length, axis=1)]
    df_values=out.values.tolist()
    spreadsheet = gc.open('ConsumerReviews')
    spreadsheet.values_append('Positives', {'valueInputOption': 'RAW'},
{'values': df_values})

#Comentarios negativos recientes
    URLS = worksheet.col_values(3)
    all_results = []
    for u in URLS:
        print(u)
        time.sleep(3)
        html = get_page_html(u)
        reviews = get_reviews_from_html(html)
        for rev in reviews:

            data = orchestrate_data_gathering(rev)
            print(data)
            all_results.append(data)

    out = pd.DataFrame.from_records(all_results)
    out= out[out.apply(row_has_max_length, axis=1)]
    df_values=out.values.tolist()
    spreadsheet = gc.open('ConsumerReviews')
    spreadsheet.values_append('Negatives', {'valueInputOption': 'RAW'},
{'values': df_values})
```

Anexo 3: ETL de valores bursátiles de compañías de Teléfonos Móviles

Los pasos para esta ETL son muy parecidos a la del anexo 2. En este caso, al querer generar un histórico de cómo ha ido evolucionando el valor bursátil de la competencia se deberá pedir la generación de un almacenamiento histórico, para ello se van a recoger la fecha de cada valor. A continuación se muestra como es el código de programación en Python.

```
scopes =
['https://www.googleapis.com/auth/spreadsheets','https://www.googleapis.com
/auth/drive']
credentials = Credentials.from_service_account_file(#Tokens de tu base de
datos, scopes=scopes)
gc = gspread.authorize(credentials)
gauth = GoogleAuth()
drive = GoogleDrive(gauth)
symbols = ['AAPL', '005930.KS', 'XIACF','NOK']  #Apple, Samsung, Xiaomi,
Nokia.

try:
    # Create an empty DataFrame to store historical data
    df = pd.DataFrame()

    # Fetch historical data for each symbol and append to the DataFrame
    for symbol in symbols:
        ticker = yf.Ticker(symbol)
        historical_data = ticker.history(period="max")
        # Rename the columns with the symbol to distinguish data
        historical_data.columns = [f"{symbol}_{col}" for col in
historical_data.columns]
        df = pd.concat([df, historical_data], axis=1)

    # Display the DataFrame with historical data for multiple symbols
    #print(df)

except Exception as e:
    print(f'An error occurred: {str(e)}')

    #----------- Limpieza de datos-----------
def row_has_max_length(row):
    return all(len(str(cell)) <= 50000 for cell in row)

# Iterate through all columns in the DataFrame
for col in df.columns:
    if df[col].dtype == 'float64':  # Check if the column contains float
values
        df[col] = df[col].astype(str)  # Convert float values to strings

df= df[df.apply(row_has_max_length, axis=1)]

#-----------Importar a la base de datos de Google Sheets-----------
spreadsheet = gc.open('MarketCap')
worksheet = spreadsheet.worksheet('CAP')
worksheet.update([df.columns.values.tolist()] + df.values.tolist())
```

Anexo 4: ETL datos Google Trends

Este código recoge todo el proceso de la ETL de los datos de Google Trends. En este proceso a diferencia de los otros, se ha utilizado una llamada API a Google Trends para pedir la extracción de la información deseada. En el caso de este trabajo se ha pedido

información sobre "Eco phone", "Eco phone case" más concretamente información de la cantidad de personas que han buscado los dos términos así como palabras que tengan relación con estos términos.

```
!pip install pytrends
import pandas as pd
import gspread
from gspread_dataframe import set_with_dataframe
from google.oauth2.service_account import Credentials
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
import datetime
from pytrends.request import TrendReq
scopes =
['https://www.googleapis.com/auth/spreadsheets','https://www.googleapis.com/a
uth/drive']
credentials = Credentials.from_service_account_file('token de credenciales',
scopes=scopes)
gc = gspread.authorize(credentials)
gauth = GoogleAuth()
drive = GoogleDrive(gauth)
#-----------Retrieve data from Google Trends-----------
#-----------1.1 Interest overtime No Direct Competitors-----------
#import matplotlib.pyplot as plt

# Create a pytrends object
pytrends = TrendReq(hl='en-US', tz=360)

# Define the keyword you want to get interest over time for
keywords = ["Iphone","Samsung","Xiaomi","Huawei","Nokia"]

# Build the payload with the specified keyword
pytrends.build_payload(keywords, cat=0, timeframe='all', geo='', gprop='')

# Get interest over time
interest_over_time1 = pytrends.interest_over_time()
interest_over_time1.reset_index(inplace=True)
#-----------1.1 Interest overtime Direct Competitors-----------

# Define the keyword you want to get interest over time for
keywords = ["Oppo","Fairphone","Shiftphone"]

# Build the payload with the specified keyword
```

```
pytrends.build_payload(keywords, cat=0, timeframe='all', geo='', gprop='')

# Get interest over time
interest_over_time2 = pytrends.interest_over_time()
interest_over_time2.reset_index(inplace=True)

#-----------2. Related search queries-----------

# List of keywords you want to retrieve related queries for
keywords = ["Eco phone", "Eco phone case"]  # Replace with your keywords

# Create a PyTrends object
pytrends = TrendReq()
# Initialize empty lists to store data
top_queries_list = []
rising_queries_list = []

# Iterate through the list of keywords and retrieve related queries
for keyword in keywords:
    # Build the payload for the current keyword
    pytrends.build_payload([keyword], cat=0, timeframe='now 1-d', geo='',
gprop='')

    # Get related queries for the current keyword
    related_queries = pytrends.related_queries()

    # Extract the related queries data and create DataFrames
    top_queries = pd.DataFrame(related_queries[keyword]['top'])
    rising_queries = pd.DataFrame(related_queries[keyword]['rising'])

    # Add the keyword as a column to identify it
    top_queries['Keyword'] = keyword
    rising_queries['Keyword'] = keyword

    # Append the DataFrames to the respective lists
    top_queries_list.append(top_queries)
    rising_queries_list.append(rising_queries)

# Concatenate the lists of DataFrames into single DataFrames
top_queries_df = pd.concat(top_queries_list, ignore_index=True)
rising_queries_df = pd.concat(rising_queries_list, ignore_index=True)


#-----------3. Top 10 COUNTRIES-----------
```

```
# Single keyword you want to retrieve data for
keyword = "Eco Phone"  # Replace with your keyword

# Create a PyTrends object
pytrends = TrendReq(hl='en-US', tz=360)

# Build the payload for the keyword
pytrends.build_payload([keyword], cat=0, timeframe='now 1-d', geo='',
gprop='')

# Get interest data by region (city or region level)
interest_by_region = pytrends.interest_by_region(resolution='COUNTRY')  #
Change resolution as needed

# Sum the interest values for each region and sort by interest value in
descending order
top_countries = interest_by_region.sum(axis=1).sort_values(ascending=False)

# Check if there are at least 10 countries available
if len(top_countries) >= 10:
    top_countries = top_countries.head(10)

# Create a DataFrame with the top countries for the keyword
top_countries_df = pd.DataFrame({keyword: top_countries})

# Reset the index to move the region information from the index to columns
top_countries_df.reset_index(inplace=True)


#-----------DATA CLEAN-----------

#function to eliminate the rows with more than 50000 characters. (Google
sheets cell max)
def row_has_max_length(row):
    return all(len(str(cell)) <= 50000 for cell in row)


#Get current time
ct = datetime.datetime.now()

#Insert timestamps
top_queries_df['exportDate'] = ct
rising_queries_df['exportDate'] = ct
top_countries_df['exportDate'] = ct
```

```
#IOTDIRECT
interest_over_time1['date'] = (interest_over_time1['date'] -
pd.Timestamp("1970-01-01")) // pd.Timedelta(seconds=1)

# Iterate through all columns in the DataFrame
for col in interest_over_time1.columns:
    if interest_over_time1[col].dtype == 'float64':  # Check if the column
contains float values
        interest_over_time1[col] = interest_over_time1[col].astype(str)  #
Convert float values to strings

interest_over_time1=
interest_over_time1[interest_over_time1.apply(row_has_max_length, axis=1)]

#IOTDIRECT
interest_over_time2['date'] = (interest_over_time2['date'] -
pd.Timestamp("1970-01-01")) // pd.Timedelta(seconds=1)
# Iterate through all columns in the DataFrame
for col in interest_over_time2.columns:
    if interest_over_time2[col].dtype == 'float64':  # Check if the column
contains float values
        interest_over_time2[col] = interest_over_time2[col].astype(str)  #
Convert float values to strings

interest_over_time2=
interest_over_time2[interest_over_time2.apply(row_has_max_length, axis=1)]

#TOP Queries
top_queries_df['exportDate'] = (top_queries_df['exportDate'] -
pd.Timestamp("1970-01-01")) // pd.Timedelta(seconds=1)

# Iterate through all columns in the DataFrame
for col in top_queries_df.columns:
    if top_queries_df[col].dtype == 'float64':  # Check if the column
contains float values
        top_queries_df[col] = top_queries_df[col].astype(str)  # Convert
float values to strings

top_queries_df= top_queries_df[top_queries_df.apply(row_has_max_length,
axis=1)]

#Rising Queries
rising_queries_df['exportDate'] = (rising_queries_df['exportDate'] -
pd.Timestamp("1970-01-01")) // pd.Timedelta(seconds=1)
```

```
# Iterate through all columns in the DataFrame
for col in rising_queries_df.columns:
    if rising_queries_df[col].dtype == 'float64':  # Check if the column
contains float values
        rising_queries_df[col] = rising_queries_df[col].astype(str)  #
Convert float values to strings

rising_queries_df=
rising_queries_df[rising_queries_df.apply(row_has_max_length, axis=1)]

#top countries
top_countries_df['exportDate'] = (top_countries_df['exportDate'] -
pd.Timestamp("1970-01-01")) // pd.Timedelta(seconds=1)

# Iterate through all columns in the DataFrame
for col in top_countries_df.columns:
    if top_countries_df[col].dtype == 'float64':  # Check if the column
contains float values
        top_countries_df[col] = top_countries_df[col].astype(str)  # Convert
float values to strings

top_countries_df= top_countries_df[top_countries_df.apply(row_has_max_length,
axis=1)]

#-----------Import Data Google Sheets-----------

#Open Spreadsheet
spreadsheet = gc.open('Google Trends Interest')

try:
    #IOTNODIRECT
    worksheet = spreadsheet.worksheet('IOTNODIRECT')
    worksheet.update([interest_over_time1.columns.values.tolist()] +
interest_over_time1.values.tolist())
except:
    print(f"An API error occurred:INTEREST OVERTIME NO DIRECT")

try:
    #IOTDIRECT
    worksheet = spreadsheet.worksheet('IOTDIRECT')
    worksheet.update([interest_over_time2.columns.values.tolist()] +
interest_over_time2.values.tolist())
except:
    print(f"An API error occurred: INTEREST OVERTIME DIRECT")
```

```
#Open Spreadsheet
spreadsheet = gc.open('Google trends Queries')

try:
    #Top queries
    #Append data just to mantain the historic
    df_values = top_queries_df.values.tolist()
    spreadsheet.values_append('TopQueries', {'valueInputOption': 'RAW'},
{'values': df_values})
except:
    print(f"An API error occurred: TOP QUERIES")

try:
    #Rising queries
    #Append data just to mantain the historic
    df_values = rising_queries_df.values.tolist()
    spreadsheet.values_append('RisingQueries', {'valueInputOption': 'RAW'},
{'values': df_values})
except:
    print(f"An API error occurred: RISING QUERIES")

try:
    df_values = top_countries_df.values.tolist()
    spreadsheet.values_append('COUNTRY', {'valueInputOption': 'RAW'},
{'values': df_values})
except:
    print(f"An API error occurred: TOP 10 Country")
```

Anexo 5: Modelo de análisis de datos NLP comentarios Amazon.

Para realizar el modelo de análisis de datos NLP lo primero que se debe hacer es instalar la librería Python necesarias para este modelo, esto se realiza mediante una llamada api a esas librerías con el código de programación que contiene funciones que se van a necesitar (esto se realizada para no tener que escribir código que se puede reusar de otras librerías que contienen los fragmentos de código que se necesita). Por otro lado, se necesita realizar una llamada API a la herramienta Google Translator para poder traducir los textos dentro del modelo, para que así el resultado que nos dé sea totalmente es español. A continuación, se muestra el código necesario para instalar todas las librerías, las

llamadas API y realizar la conexión con Drive para poder extraer los datos con los que se van a trabajar.

```
!pip install deep_translator
import gspread
from gspread_dataframe import set_with_dataframe
from google.oauth2.service_account import Credentials
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
import pandas as pd
from deep_translator import GoogleTranslator
from transformers import pipeline
from wordcloud import WordCloud
from wordcloud import STOPWORDS
import matplotlib.pyplot as plt
import seaborn as sns

scopes                                                                  =
['https://www.googleapis.com/auth/spreadsheets','https://www.googleapis.com
/auth/drive']

credentials = Credentials.from_service_account_file(#Tokens de tu base de
datos, scopes=scopes)

gc = gspread.authorize(credentials)

gauth = GoogleAuth()

drive = GoogleDrive(gauth)
```

El segundo paso será indicar al modelo cuáles son comentarios positivos y cuáles negativos. En este trabajo se ha optado porque esta información esté en dos hojas de cálculo separadas, una con los positivos y otra con los negativos. A continuación se muestra el código de programación.

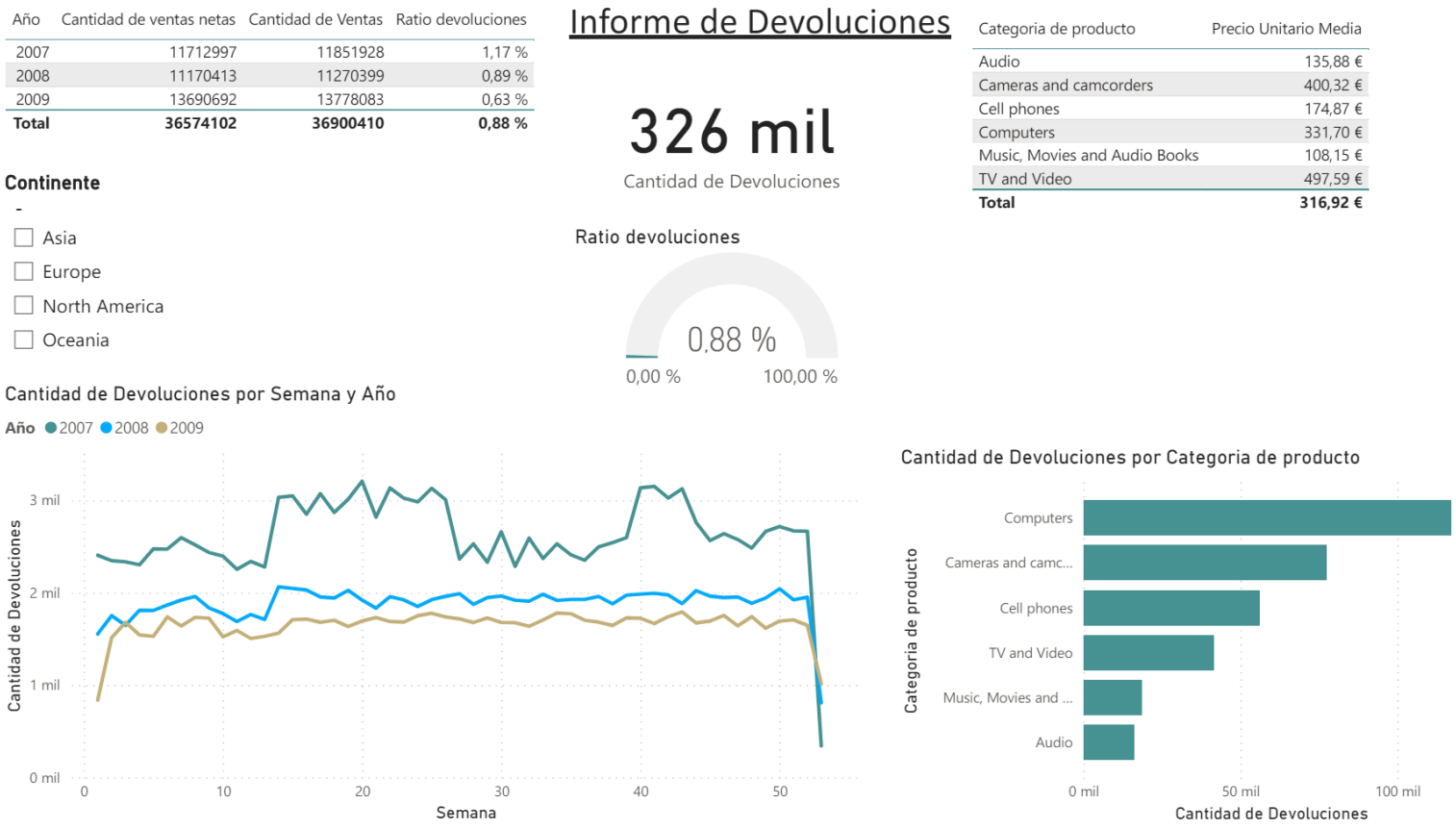
```
#Positive
spreadsheet = gc.open('ConsumerReviews')
worksheet = spreadsheet.worksheet('Positives')
data=worksheet.get_all_values()
positives=pd.DataFrame(data[1:], columns=data[0])
#Negative
spreadsheet = gc.open('ConsumerReviews')
worksheet = spreadsheet.worksheet('Negatives')
data=worksheet.get_all_values()
negatives=pd.DataFrame(data[1:], columns=data[0])
positives=positives.drop_duplicates()
negatives=negatives.drop_duplicates()
data1 = positives['Translated_Text'].tolist()
data2 = negatives['Translated_Text'].tolist()
```

Por último, solo quedaría hacer la llamada al modelo para que realice un análisis de sentimiento de los comentarios. Y nos devuelva los datos en gráficas u otras modalidades que deseemos. En el caso de este trabajo se ha pedido una gráfica de barras para que muestre cuántas estrellas tiene cada comentario, la correlación entre los diferentes comentarios y una nube de palabras.

```
sentiment_pipeline                                                    =
pipeline(model="lxyuan/distilbert-base-multilingual-cased-sentiments-studen
t")
reviewsPositive=[]
for review in data1:
   try:
     sentiment = sentiment_pipeline(review)
     reviewsPositive.append({'Translated_Text': review, 'sentiment':
sentiment[0]['label'],'score': sentiment[0]['score']})
   except:
     pass
reviewsNegative=[]
for review in data2:
   try:
     sentiment = sentiment_pipeline(review)
     reviewsNegative.append({'Translated_Text': review, 'sentiment':
sentiment[0]['label'],'score': sentiment[0]['score']})
   except:
     print('pass')
    pass
data_positive = pd.merge(pd.DataFrame(reviewsPositive), positives,
on='Translated_Text')
data_negative = pd.merge(pd.DataFrame(reviewsNegative), negatives,
on='Translated_Text')
```

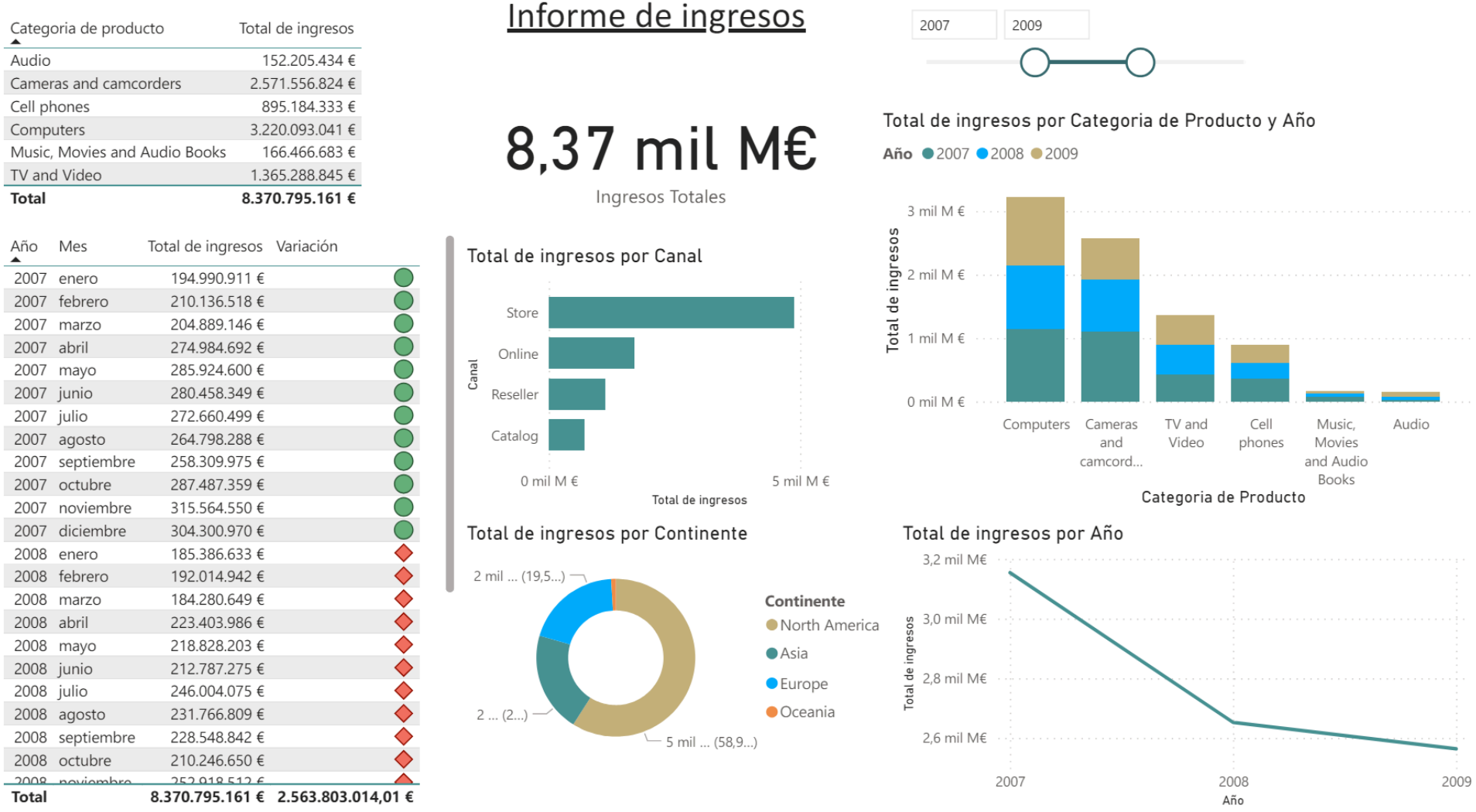
Anexo 6: Informe Power BI sobre los datos de *CostosoSales*

Figura A6.1: Informe de devoluciones



Fuente: Elaboración propio mediante Power BI

Figura A6.2: Informe de ingresos

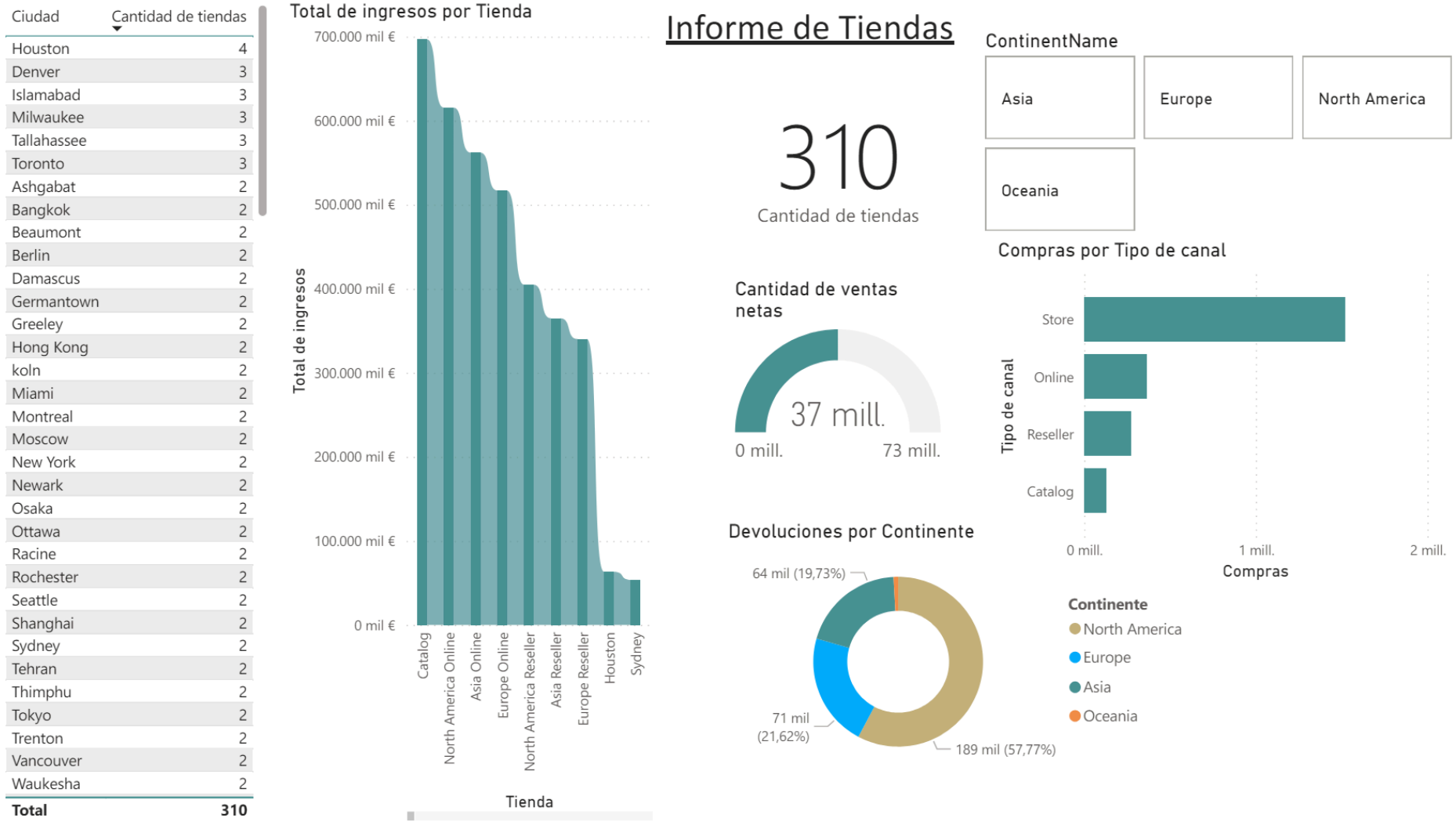


Categoria de producto	Total de ingresos
Audio	152.205.434 €
Cameras and camcorders	2.571.556.824 €
Cell phones	895.184.333 €
Computers	3.220.093.041 €
Music, Movies and Audio Books	166.466.683 €
TV and Video	1.365.288.845 €
Total	**8.370.795.161 €**

Año	Mes	Total de ingresos	Variación
2007	enero	194.990.911 €	
2007	febrero	210.136.518 €	
2007	marzo	204.889.146 €	
2007	abril	274.984.692 €	
2007	mayo	285.924.600 €	
2007	junio	280.458.349 €	
2007	julio	272.660.499 €	
2007	agosto	264.798.288 €	
2007	septiembre	258.309.975 €	
2007	octubre	287.487.359 €	
2007	noviembre	315.564.550 €	
2007	diciembre	304.300.970 €	
2008	enero	185.386.633 €	
2008	febrero	192.014.942 €	
2008	marzo	184.280.649 €	
2008	abril	223.403.986 €	
2008	mayo	218.828.203 €	
2008	junio	212.787.275 €	
2008	julio	246.004.075 €	
2008	agosto	231.766.809 €	
2008	septiembre	228.548.842 €	
2008	octubre	210.246.650 €	
2008	noviembre	252.918.512 €	
Total		**8.370.795.161 €**	**2.563.803.014,01 €**

Fuente: Elaboración propia mediante Power BI

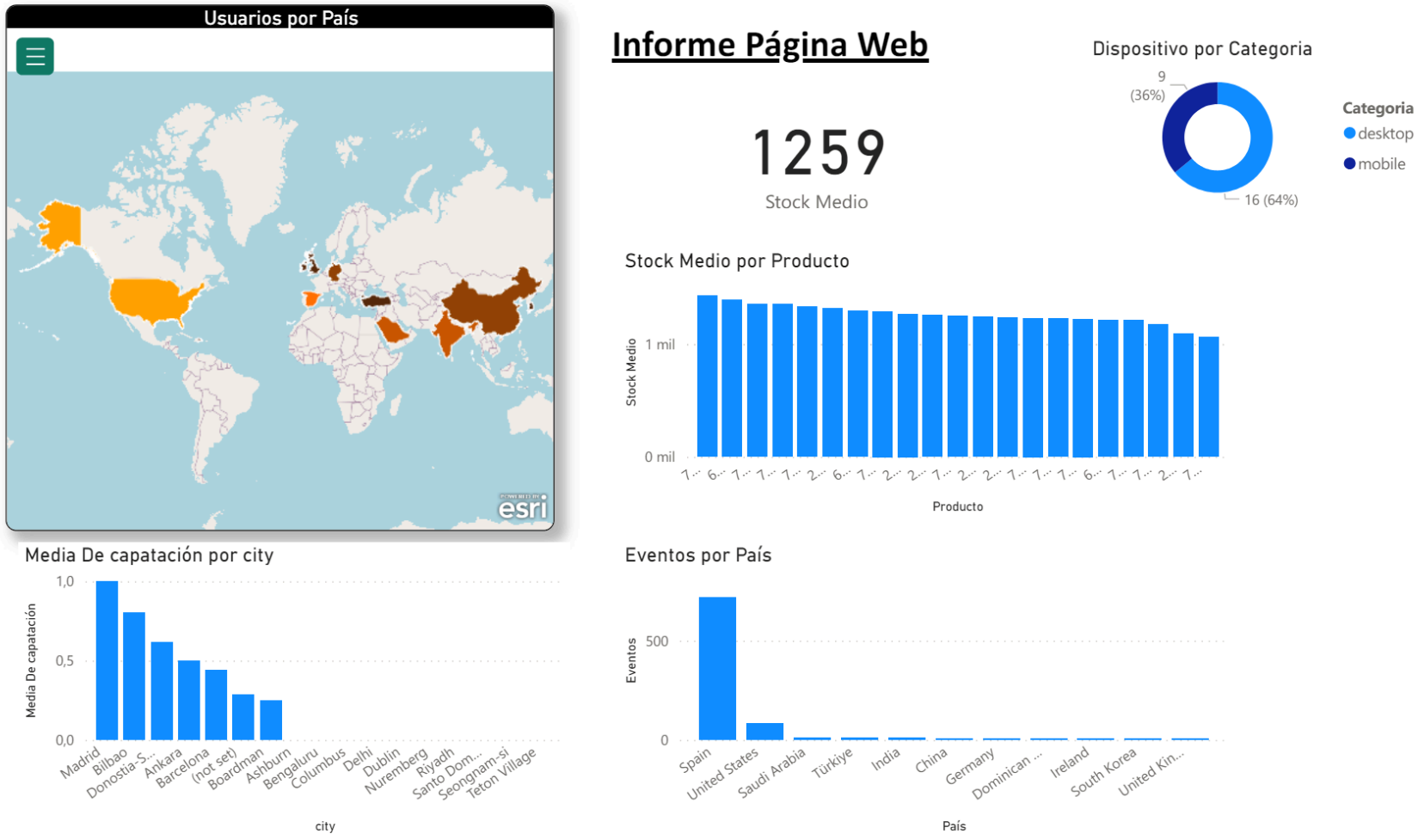
Figura A6.3: Informe de tiendas



Fuente: Elaboración propia mediante Power BI

Anexo 7: Informe Power BI sobre la página web

Figura A7.1: Informe de la página web



Fuente: Elaboración propia mediante Power BI