Implementación de soluciones en la nube para el análisis de datos públicos a través de modelos de inteligencia artificial

-

Implementation of cloud solutions for public data analysis through artificial intelligence models

Trabajo de Fin de Master
Curso 2024–2025

Autor
Cristian Molina Muñoz

Director
Jose Luis Vazquez-Poletti
Rubén Fuentes-Fernández

Máster en Ingeniería Informática
Facultad de Informática
Universidad Complutense de Madrid

Implementación de soluciones en la nube para el análisis de datos públicos a través de modelos de inteligencia artificial

-

Implementation of cloud solutions for public data analysis through artificial intelligence models

Trabajo de Fin de máster en Ingeniería Informática

Autor
Cristian Molina Muñoz

Director
Jose Luis Vazquez-Poletti
Rubén Fuentes-Fernández

Convocatoria: *Septiembre 2025*
Calificación:

Máster en Ingeniería Informática
Facultad de Informática
Universidad Complutense de Madrid

25 de agosto de 2025

Autorización de difusión

los abajo firmantes, matriculados en el Master en Ingeniería en Informática de la Facultad de Informática, autorizan a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a sus autores el presente Trabajo Fin de Master: "Implementación de soluciones en la nube para el análisis de datos públicos a través de modelos de inteligencia artificial", realizado durante el curso académico 2024-2025 bajo la dirección de Jose Luis Vazquez-Poletti y Rubén Fuentes-Fernández, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Cristian Molina Muñoz

25 de agosto de 2025

Dedicatoria

A mis padres, por hacer posible todo esto. Por su esfuerzo

Agradecimientos

Muchas gracias a todos los profesores y compañeros que nos han acompañado en este viaje y de los que tanto hemos aprendido

Resumen

[TODO 250 palabras sobre datos, cloud y IA] [Se redacta en pasado y no debe incluir abreviaturas, referencias a figuras o tablas ni citas bibliográficas. Tampoco se debe incluir información que no aparezca en el proyecto.]

Los ficheros de GitHub se encuentran en el siguiente repositorio:

https://github.com/crismo04/TFM-cloud-soliutions-to-public-data/

Palabras clave

Tratamiento de datos, Cloud, Big data, inteligencia Artificial, [TODO mas sobre clouds, se mencionan en orden alfabético]

Abstract

[TODO three paragraphs on data, cloud and AI].

The GitHub files can be found in the following repository:

https://github.com/crismo04/TFM-cloud-soliutions-to-public-data/

Keywords

Data processing, Cloud, Big data, Artificial intelligence, TODO something else about clouds.

Índice

Índice de figuras

Capítulo 1

Introducción

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
— Alan Turing

1.1. Motivación

Empezaremos por el principio, definiendo que son los tres principales elementos de este proyecto [TODO]

El alcance de este proyecto es, por un lado [TODO]

1.2. Plan de trabajo

Una vez definido el alcance, me gustaría destacar las cinco fases en las que se ha dividido el proyecto, que se han ido iterando para la creación de varios prototipos funcionales:

1. **Fase de investigación:** Búsqueda de información a cerca de diferentes fuentes publicas de datos, tecnologías en la nube y modelos o herramientas de IA que nos ayuden a tratar, filtrar y entender todos los datos públicos recopilados.
2. **Fase de análisis de requisitos:** [TODO]
3. **Fase de implementación:** [TODO]
4. **Fase de pruebas:** [TODO]

5. **Memoria:** Elaboración de este documento, plasmando las fases anteriores en texto y especificando el desarrollo del proyecto y los resultados del mismo.

Chapter 1

Introduction

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
— Alan Turing

1.1. Motivation

We will start at the beginning by defining what the three main elements of this project are, [TODO]

The scope of this project is, on the one hand [TODO]

1.2. Work plan

Having defined the scope, I would like to highlight the five phases in which the project has been partitioned, which have been iterated for the creation of several functional prototypes:

1. **Research phase.** Search for information about different public data sources, cloud technologies and AI models or tools that help us to treat, filter and understand all the public data collected.
2. **Requirements analysis phase:** [TODO].
3. **Implementation phase:** [TODO]
4. **Testing phase:** [TODO]

5. **Memory phase:** [TODO] The elaboration of this document, translating the previous phases into text and specifying the development of the project and its results.

Capítulo 2

Estado de la Cuestión

En este apartado expondremos el estado actual de los puntos principales de nuestro proyecto según las investigaciones realizadas, así como los trabajos o artículos relacionados con los temas a tratar. Estos son, entre otros, trabajos relacionados con los principales proveedores Cloud y su comparación, trabajos que traten con grandes volúmenes de datos públicos, o trabajos que utilicen diferentes IAs para el tratamiento de datos y la obtención de conclusiones a partir de estos.

Aunque también usaremos datos, estudios y aplicaciones de otras partes del globo, nos intentaremos centrar en datos del territorio español, ya que la cantidad de datos de toda la web es inmensa y de esta manera acotaremos el alcance del proyecto y contribuiremos a aprovechar datos que no han sido tan explotados y explorados como pueden ser los datos abiertos de Google (Google, 2025) o Amazon (Amazon, 2025) .

2.1. Datos

Llevamos mucho tiempo escuchando que vivimos en la era de la información o de los dato, desde la invención del transistor en 1947 (Wikipedia, 2025), pasando la primera vez que se acuño el termino 'Big data'y 'web 2.0én 2005 (Press, 2013), así como su rápido crecimiento y adopción en todas las esferas (Brown et al., 2011), hasta el presente, donde los datos y su tratamiento a través de múltiples herramientas, incluyendo la recientemente omnipresente Inteligencia Artificial, llegaran a generan ,según proyecciones, la asombrosa cifra de 149 Zettabytes de datos en 2024, ¡¡23 ceros en bytes!! (Taylor, 2025) & (Pangarkar, 2025).

En España, los datos también muestran un aumento significativo, según los datos de telecomunicaciones del CNMC, los cuales se han analizado con este mismo proyecto (CNMC, 2025), el uso de datos generales en las principales empresas es de 0.092 Zettabytes de datos en 2024. Esto es solo un 0.06 % del volumen global, lo cual no cuadra del todo con otras estimaciones (Insights + Analytics, 2024), que por volumen de mercado sitúan a España en un 0.9 % del volumen global, lo cual se puede explicar debido a que el CNMC solo toma en cuenta datos de las principales empresas de telecomunicaciones.

Aun con estas discrepancias en cuanto a números, lo que esta claro es que el mercado de los datos no para de crecer año tras año y cada vez resulta mas difícil separar la información relevante del ruido, evitando la ínfoxicaciónó sobrecarga informativa (Cornella, 2000). En este escenario, tecnologías como la computación en la nube e inteligencia artificial pueden ser claves para encontrar los patrones o llegar a conclusiones.

Mencionar también brevemente que los 'datosño suelen aparecer en formatos consistentes, y para este trabajo se han tratado diferentes formatos: CSV, JSON, bases de datos diversas, excel, APIs, etc. (Khan y Alam, 2019). Esto es así porque queríamos que las fuentes de datos fueran heterogéneas y no excluir datos por que su extracción o tratamiento fueran complejos, ya que este es el caso para la mayoría de aplicaciones en el mundo real. Esto se explicara mas en detalle en el apartado de Materiales y métodos 3.

Lo primero para la realización de este proyecto, es la obtención de datos públicos. Esto presenta tres grandes complicaciones a tener en cuenta:

La primera es que **no todos los datos que deberían ser públicos lo son**, y cuando lo son, su acceso y tratamiento es complicado, ya sea a conciencia o por indolencia. Según la OECD (OECD, 2023), sólo el 48 % de los conjuntos de datos de gran valor están disponibles como datos abiertos en los países estudiados, datos que bajan al 30 % cuando se trata de datos financieros. y estudios de otras partes del mundo también avalan esta reticencia a la correcta apertura de información publica (De la Torre y Núñez, 2023), (Jorge y Herrera, 2023).

La segunda causa es **la regulación**, el tratamiento de datos en Europa debe seguir la RGPD de 2016 y las regulaciones propias de cada estado (Ramos-Simón, 2017) & (Union Europea, 2016), así como la mas reciente Ley de Gobernanza de Datos (Union Europea, 2023) & (Julián Valero Torrijos, 2022). Para cumplir con estas normativas, en este proyecto nos centraremos en el uso de datos oficiales abiertos, evitando técnicas como el 'scraping'que pueden estar sujetas a controversia a la vista de estas regulaciones. También se verificaran las licencias de todos los datos y modelos utilizados para asegurarnos de que no incumplimos ninguna de las regulaciones existentes.

En cuanto a datos de otros países fuera de la Unión Europea, tenemos pa-

noramas diversos los cuales vale la pena mencionar, desde una regulación mas laxa en Estados Unidos, hasta un control estricto en países como China. Estos datos no se utilizaran en este trabajo por temas de alcance, ya que se prefiere dar prioridad a fuentes de datos nacionales, pero las herramientas desarrolladas serian aplicables a estos mismos datos cumpliendo sus normativas.

En Estados Unidos, el panorama es sobretodo abierto, pero fragmentado. Cuentan con regulaciones sectoriales, como la 'Health Insurance Portability and Accountability Act'(HIPAA) para datos médicos (Congreso de los Estados Unidos de America, 1996) y regulaciones estatales como la 'California Consumer Privacy Act'(CCPA) (Estado de California, 2018) para proteger derechos individuales. También existe una legislación nacional que promueve los datos abiertos, la ÓPEN Government Data Act'(2019) (Congreso de los Estados Unidos de America, 2019), que establece que los datos gubernamentales deben ser abiertos y utilizables.

Por su parte, China ha establecido un marco regulatorio estricto con leyes como la 'Personal Information Protection Law'(PIPL) (Standing Committee of the National People's Congress, 2021b), que habla de principios de consentimiento y derechos del individuo, y la 'Data Security Law'(DSL) (Standing Committee of the National People's Congress, 2021a), que prioriza la seguridad nacional y el control sobre los datos generados en el país.

Por ultimo, la tercera causa es **la tecnología**, como ya hemos hablado, los datos pueden estar en formatos diferentes, y la cantidad de herramientas para su tratamiento va en aumento, y hay que tener en cuenta también la integración, el procesamiento escalable a la cantidad de datos en aumento y el gobierno de los flujos de datos y modelos en un entorno 'cloud'que está en evolución constante. Por ello, en este trabajo se ha optado por emplear herramientas ampliamente extendidas, soportadas, y principalmente abiertas, así como intentar hacer del conjunto de ellas lo mas amplio posible, para estudiar y comparar un amplio abanico de soluciones.

2.1.1. Trabajos anteriores

A parte de todas las referencias ya incluidas en esta sección, me gustaría destacar todo el trabajo de Jaime Gómez-Obregón para liberar y hacer accesibles los datos de España (Gómez-Obregón, 2025a), con acciones como publicar las subvenciones a las empresas en España a través del portal ministerial y hacerlas accesibles (Gómez-Obregón, 2025c), o estudios sobre donaciones sospechosas de corrupción (Gómez-Obregón, 2025b). Todo este trabajo ha guiado también a este proyecto hacia un uso ético de los datos.

2.1.2. Conjuntos de datos

A nivel institucional, Europa tiene su propio portal para acceder a datos públicos (Union Europea, 2025), y a nivel nacional, el Instituto Nacional de Estadística (INE) y la Agencia Tributaria han sido actores clave en la liberación de datos abiertos y el fomento de su reutilización para la investigación e innovación, impulsando proyectos como el Portal de Transparencia del Gobierno de España y las iniciativas de datos abiertos de comunidades autónomas y ayuntamientos (Gobierno de España, 2025); (Ayuntamiento de Madrid, 2025) & (Registradores de España, 2025), los cual se esfuerzan por hacer públicos datos de alto valor [A]. También destacar iniciativas que fomentan su transparencia, como InfoParticipa (Universitat Autònoma de Barcelona, 2025) o iniciativas privadas para la recolección de datos públicos (Esri España, 2025). Por ultimo, también añadir a la interminable lista de datos públicos disponibles iniciativas individuales como .Awesome public datasets"(Awesome data, 2025) que se dedica a recopilar fuentes de datos fiables (aunque en este caso principalmente de Estados Unidos), o iniciativas como UniversiDATA (UniversiDATA, 2025).

Todos estos portales y aplicaciones son de gran importancia y constituyen la base material sobre la que se sustentan trabajos como el presente.

2.2. Nubes

2.2.1. Principales Proveedores de Nube y sus Capas Gratuitas

A continuación detallaremos las pruebas gratuitas de los principales proveedores de servicios en la nube, información crucial para la selección tecnológica de este proyecto. Solo se listarán sus principales servicios, ya que la lista total es muy extensa (R.I.Pienaar, 2025).

Google Cloud Platform

- **App Engine**: 28 horas/día de ejecución 'frontend', 9 horas/día de ejecución 'backend'.
- **Cloud Firestore**: 1GB almacenamiento, 50.000 lecturas, 20.000 escrituras, 20.000 borrados por día.
- **Compute Engine**: 1 e2-micro no susceptible de interrupción, 30GB disco duro, 5GB de instantáneas, con regiones restringidas.

- **Cloud Storage**: 5GB, 1GB de tráfico de salida de red.
- **Cloud Shell**: Terminal Linux basado en web con 5GB de almacenamiento persistente. Límite de 60 horas/semana.
- **Cloud Pub/Sub**: 10GB de mensajes por mes.
- **Cloud Functions**: 2 millones de invocaciones por mes.
- **Cloud Run**: 2M de peticiones por mes, 360.000 GB/segundos de memoria, 180.000 segundos de CPU virtual.
- **Google Kubernetes Engine**: Sin tarifa de gestión de clústeres para un clúster zonal.
- **BigQuery**: 1 TB de consultas por mes, 10 GB de almacenamiento.
- **Cloud Build**: 120 minutos de construcción por día.
- **Cloud Source Repositories**: Hasta 5 usuarios, 50 GB de almacenamiento, 50 GB de tráfico de salida.
- **Google Colab**: Entorno gratuito de desarrollo con 'Jupyter Notebooks'.
- **Lista completa**: `https://cloud.google.com/free`

Amazon Web Services

- **CloudFront**: 1TB de tráfico de salida por mes y 2M invocaciones de funciones.
- **CloudWatch**: 10 métricas personalizadas y 10 alarmas.
- **CodeBuild**: 100min de tiempo de ejecución por mes.
- **CodeCommit**: 5 usuarios activos, 50GB almacenamiento, 10000 peticiones por mes.
- **CodePipeline**: 1 pipeline activo por mes.
- **DynamoDB**: 25GB base de datos NoSQL.
- **EC2**: 750 horas/mes de t2.micro o t3.micro, 12 meses.
- **EBS**: 30GB por mes de SSD propósito general o magnético, 12 meses.
- **Elastic Load Balancing**: 750 horas por mes, 12 meses.
- **RDS**: 750 horas/mes de db.t2.micro, 20GB almacenamiento SSD, 12 meses.

- **S3**: 5GB almacenamiento estándar, 20K peticiones Get, 2K peticiones Put, 12 meses.
- **Glacier**: 10GB almacenamiento a largo plazo.
- **Lambda**: 1 millón de peticiones por mes.
- **SNS**: 1 millón de publicaciones por mes.
- **SES**: 3.000 mensajes por mes, 12 meses.
- **SQS**: 1 millón de peticiones de colas de mensajería.
- **Lista completa**: `https://aws.amazon.com/free/`

Microsoft Azure

- **Virtual Machines**: 1 B1S Linux VM, 1 B1S Windows VM, 12 meses.
- **App Service**: 10 aplicaciones web, móviles o de API, con 60 minutos CPU/día.
- **Functions**: 1 millón de peticiones por mes.
- **DevTest Labs**: Entornos de desarrollo y pruebas.
- **Active Directory**: 500.000 objetos.
- **Azure DevOps**: 5 usuarios activos, repositorios Git privados ilimitados.
- **Azure Pipelines**: 10 trabajos paralelos con minutos ilimitados para código abierto.
- **Microsoft IoT Hub**: 8.000 mensajes por día.
- **Load Balancer**: 1 IP pública con balanceo de carga gratuita.
- **Notification Hubs**: 1 millón de notificaciones 'push'.
- **Ancho de banda**: 15GB de entrada y 5GB de salida por mes, 12 meses.
- **Cosmos DB**: 25GB almacenamiento y 1000 unidades de solicitud de rendimiento
- **Static Web Apps**: Aplicaciones estáticas con SSL, autenticación y dominios personalizados
- **Storage**: 5GB almacenamiento de archivos o 'blobs'con redundancia local, 12 meses.
- **Cognitive Services**: APIs de IA/ML con transacciones limitadas.

- **Cognitive Search**: Búsqueda basada en IA, para 10.000 documentos.
- **Azure Kubernetes Service**: Servicio Kubernetes gestionado, gestión de clústeres.
- **Event Grid**: 100K operaciones/mes.
- **Lista completa**: `https://azure.microsoft.com/free/`

Oracle Cloud

- **Compute**: 2 máquinas virtuales AMD con 1/8 OCPU y 1 GB memoria cada una.
- **Block Volume**: 2 volúmenes, 200 GB total para computación.
- **Object Storage**: 10 GB.
- **Load Balancer**: 1 instancia con 10 Mbps.
- **Databases**: 2 bases de datos, 20 GB cada una.
- **Monitoring**: 500 millones de puntos de ingesta de datos, 1 millardo de recuperación.
- **Ancho de banda**: 10 TB de tráfico de salida por mes, velocidad limitada a 50 Mbps.
- **IP Pública**: 2 IPv4 para máquinas virtuales, 1 IPv4 para balanceador de carga.
- **Notifications**: 1 millón de opciones de entrega por mes, 1000 emails enviados por mes.
- **Lista completa**: `https://www.oracle.com/cloud/free/`

IBM Cloud

- **Cloudant database**: 1 GB de almacenamiento de datos.
- **Db2 database**: 100MB de almacenamiento de datos.
- **API Connect**: 50.000 llamadas API por mes.
- **Availability Monitoring**: 3 millones de puntos de datos por mes.
- **Log Analysis**: 500MB de registros diarios.
- **Lista completa**: `https://www.ibm.com/cloud/free/`

Cloudflare

- **Application Services**: DNS, Protección DDoS, CDN con SSL, Firewall de aplicaciones web.
- **Zero Trust & SASE**: Hasta 50 usuarios, 24 horas de registro de actividad.
- **Cloudflare Tunnel**: Exponer puertos HTTP locales a través de túnel.
- **Workers**: Desplegar código sin servidor - 100k peticiones diarias.
- **Workers KV**: 100k lecturas diarias, 1000 escrituras diarias, 1 GB datos almacenados.
- **R2**: 10 GB por mes, 1 millón operaciones por mes.
- **D1**: 5 millones de filas leídas por día, 100k filas escritas por día, 1 GB almacenamiento.
- **Pages**: Desplegar aplicaciones web - 500 despliegues mensuales, 100 dominios personalizados.
- **Queues**: 1 millón de operaciones por mes.
- **TURN**: 1TB de tráfico saliente por mes.
- **Lista completa**: `https://www.cloudflare.com/plans/free/`

También, aunque no son nubes propiamente dichas, hemos querido añadir en esta sección otras herramientas que tienen interés para el proyecto:

Hugging Face Spaces

- **Tipo**: Plataforma para desplegar, compartir y descubrir modelos de Aprendizaje Automático (MLOps). Esencial para proyectos de IA. Permite desplegar demostraciones de modelos con interfaz web de forma sencilla.
- **Capa Gratuita - CPU**:
  - 2 CPUs virtuales por espacio.
  - 16 GB de RAM.
  - Espacio de almacenamiento: 50 GB (para modelos, datos y código).
  - Ancho de banda: 100 MB/hora para CPUs.

    - **Apagado automático**: Los espacios se suspenden tras 48 horas de inactividad para ahorrar recursos, reactivándose con la siguiente visita.
- **Capa Gratuita - GPU (T4)**:
    - Acceso a una GPU NVIDIA T4 por espacio.
    - 16 GB de RAM.
    - Espacio de almacenamiento: 50 GB.
    - Ancho de banda: 30 MB/hora para GPUs.
    - Uso: Hasta 30 horas de uso de GPU por mes, pero sujeto a disponibilidad.
    - **Apagado automático**: Las GPU se apagan automáticamente tras 1 hora de inactividad.
- **Enfoque**: Despliegue, demostración y compartición de modelos de IA. Integración nativa con el Hub de modelos y conjuntos de datos.
- **URL**: `https://huggingface.co/spaces`

Kaggle Kernels/Notebooks

- **Tipo**: Entorno de ejecución para cuadernos 'Jupyterén la nube. Proporciona acceso gratuito a aceleradores hardware de gama alta, eliminando la barrera de entrada para entrenar modelos complejos.
- **Capa Gratuita - Sesiones de Ejecución**:
    - **GPU (NVIDIA Tesla P100)**: Hasta 30 horas por semana (4.3h/día aprox.).
    - **TPU (v3)**: Hasta 20 horas por semana (2.8h/día aprox.).
    - **CPU**: 20 horas de tiempo total por semana, sin límite de sesiones concurrentes.
- **Límites por Sesión**:
    - **Tiempo máximo de ejecución**: 12 horas por sesión. Tras este tiempo, el kernel se detiene automáticamente.
    - **Internet**: Los cuadernos deben tener la opción de Internet activada manualmente para acceder a datos externos o instalar librerías.
    - **Almacenamiento Volátil**: 20 GB de espacio temporal de disco. Los datos no persisten entre sesiones, aunque se puede usar el

sistema de conjuntos de datos de Kaggle para almacenamiento persistente.

- **Enfoque**: Análisis exploratorio de datos, competiciones de ML y, crucialmente, **entrenamiento de modelos** que requieran GPU/TPU.
- **URL**: `https://www.kaggle.com/code`

2.2.2. Trabajos anteriores

2.3. Inteligencia Artificial

[TODO]

Tener en cuenta también la nueva normativa que la Unión europea ha establecido con el Reglamento de Inteligencia Artificial (Union Europea, 2024), el cual se ha tenido de base para el uso de IA en este proyecto, intentando aplicar buenas practicas al uso de las mismas, así como documentar las fuentes de datos, métodos de anonimización y posibles sesgos.

2.3.1. Trabajos anteriores

Capítulo 3

Materiales y métodos

[TODO, importante a tener en cuenta: - detallarse cada paso que se ha dado para llegar a los resultados describiendo, en orden lógico y expresado con claridad, los materiales y recursos empleados. - No avanzar resultados y redactarse en pasado

]

En este capítulo vamos a describir el proceso que se ha seguido en la realización del trabajo, las distintas tecnologías, lenguajes de programación y herramientas, así como las que se ha valorado pero descartado. También se definirán los métodos de desarrollo, aplicaciones e incluso modelo de trabajo.

3.1. Materiales

[TODO, herramientas, programas y material utilizado, incluyendo por ejemplo los tipos de IA]

3.1.1. Lenguajes

PYTHON

Python es un lenguaje de programación interpretado y centrado en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional.

SQL

SQL es un lenguaje de dominio específico utilizado en programación, diseñado para administrar, y recuperar información de sistemas de gestión de bases de datos relacionales. Es un sistema que facilita el tratamiento de datos, así como la separación de estos datos del programa principal, permitiendo tener más modularidad. Utilizamos SQL para almacenar información, así como para extraer esta misma información, tratarla y almacenarla ya tratada en la base de datos.

LaTeX

LaTeX es un sistema de composición tipográfica de alta calidad que incluye funcionalidades diseñadas para la producción de documentación técnica y científica. Es el estándar de facto para la comunicación y publicación de documentos científicos, el cual nos ha permitido desarrollar una memoria profesional y facilitar el diseño sin tener que preocuparnos por la forma cada vez que añadíamos cambios. Hemos usado LaTeX para desarrollar este documento en la aplicación de TeXstudio y el compilador MikteX.

3.1.1.1. Lenguajes descartados

[TODO]

3.1.2. Herramientas

Visual Studio Code

Visual Studio Code es un editor de código fuente desarrollado por Microsoft para Windows, Linux y MacOS. Incluye soporte para la depuración, control integrado de Git, resaltado de sintaxis, finalización inteligente de código, fragmentos y refactorización de código.

Utilizamos Visual Studio Code como entorno de desarrollo software, ya que comparándolo con otras alternativas que nos brindan en la carrera, lo creemos bastante más útil, sobre todo para la programación en Python o React.

Lo que nos ha hecho decantarnos por él por encima del resto, es la gran comunidad que tiene detrás, la cual cuenta con un gran número de tutoriales y extensiones que nos facilitan mucho la programación y la integración con otras aplicaciones como Github. También destacar su intérprete, para probar pequeños fragmentos de código, lo cual nos ha ahorrado tiempo en depuración de errores.

Github

GitHub es una forja para alojar proyectos utilizando el sistema de control de versiones Git. Se utiliza principalmente para la creación de código fuente de programas de ordenador.

Utilizamos GitHub como sistema de control de versiones y repositorio de código por su tremenda utilidad para comunicarnos y trabajar en paralelo, lo cual ha sido una necesidad en los tiempos de pandemia en los que este proyecto se ha realizado. Esto nos ha asegurado no perder nada de progreso y llevar un control del avance del proyecto en todo momento. Además nuestros tutores nos facilitaron un repositorio privado en el que nos asegurábamos la seguridad del código, por lo que su uso era casi una obligación frente a otras alternativas.

TeXstudio y MiKTeX

TeXstudio es un editor de LaTeX de código abierto y Multiplataforma con una interfaz similar a Texmaker. TeXstudio es un IDE de LaTeX que proporciona un soporte moderno de escritura, como la corrección ortográfica interactiva, plegado de código y resaltado de sintaxis, por lo que lo hemos considerado ideal para la elaboración de este documento. Mientras que MiKTeX es el gestor de paquetes integrado, que instala los paquetes que hacen falta para el correcto funcionamiento de TeXstudio y para la creación de este documento.

3.1.3. Herramientas descartadas

[TODO]

3.2. Metodos

Para llegar a nuestro objetivo de diseño, hemos dividido la implementación en diferentes módulos:

- **búsqueda y almacenamiento de datos.**
- **Tratamiento básico de los datos.**
- **Estudio con modelos de IA en diferentes nubes**
- **Comparación y estudio de resultados.**

3.2.1. Utilización de la solución

Capítulo 4

Resultados

[TODO, importante a tener en cuenta: Aquí se recogen los nuevos conocimientos que el proyecto aporta al conocimiento científico, redactarse en pasado. utilizando recursos gráficos.]

Capítulo 5

Manual de usuario y casos de uso

Capítulo 6

Conclusiones y Trabajo Futuro

[TODO, importante a tener en cuenta: Señalar los principios y relaciones que indican los resultados (qué es lo que se ha sacado en claro con la investigación, futuras implicaciones que se pueden extraer, etc.). · Relacionar los resultados con otros trabajos publicados. · Hay que mencionar también las excepciones, faltas de correlación o aspectos no resueltos. · Indicar futuras líneas de trabajo]

Trabajo futuro

Chapter 6

Conclusions and Future Work

[TODO]

Apéndice A

Definiciones y acornimos

A.0.1. Definiciones

El gobierno de España define los **datos de alto valor** como "documentos cuya reutilización está asociada a considerables beneficios para la sociedad, el medio ambiente y la economía, en particular debido a su idoneidad para la creación de servicios de valor añadido, aplicaciones y puestos de trabajo nuevos, dignos y de calidad, y al número de beneficiarios potenciales de los servicios de valor añadido y aplicaciones basados en tales conjuntos de datos" Esta definición nos ofrece varias pistas sobre la manera en la que se prevé que se identifiquen esos conjuntos de datos de alto valor a través de una serie de indicadores que incluirían:

- Su potencial para generar beneficios sociales o medioambientales significativos.
- Su potencial para generar beneficios económicos y nuevos ingresos.
- Su potencial para generar servicios innovadores.
- Su potencial en cuanto a número de usuarios beneficiados, con atención particular a las PYMEs.
- Su capacidad para ser combinados con otros conjuntos de datos

A.0.2. Acronimos

AI Inteligencia Artificial (Artificial Intelligence)

API Interfaz de Programación de Aplicaciones (Application Programming Interface)

AWS Amazon Web Services

CCPA Ley de Privacidad del Consumidor de California (California Consumer Privacy Act)

CDN Red de Distribución de Contenidos (Content Delivery Network)

CNMC Comisión Nacional de los Mercados y la Competencia

CPU Unidad Central de Procesamiento (Central Processing Unit)

CSV Valores Separados por Comas (Comma-Separated Values)

D1 D1 Database (Base de datos de Cloudflare)

DDoS Ataque de Denegación de Servicio Distribuido (Distributed Denial of Service)

DNS Sistema de Nombres de Dominio (Domain Name System)

DSL Ley de Seguridad de Datos (Data Security Law) - China

EEUU Estados Unidos

EU Unión Europea

GCP Google Cloud Platform

GPU Unidad de Procesamiento Gráfico (Graphics Processing Unit)

HIPAA Ley de Portabilidad y Responsabilidad de Seguros de Salud (Health Insurance Portability and Accountability Act)

IBM International Business Machines

INE Instituto Nacional de Estadística

JSON Notación de Objetos de JavaScript (JavaScript Object Notation)

KV Key-Value (Almacenamiento Clave-Valor)

LGD Ley de Gobernanza de Datos

ML Aprendizaje Automático (Machine Learning)

MLOps Operaciones de Aprendizaje Automático (Machine Learning Operations)

NLP Procesamiento del Lenguaje Natural (Natural Language Processing)

OECD Organización para la Cooperación y el Desarrollo Económicos

OGDA Ley de Datos Abiertos del Gobierno (OPEN Government Data Act)

OCI Oracle Cloud Infrastructure

PIPL Ley de Protección de Información Personal (Personal Information Protection Law) - China

R2 R2 Storage (Almacenamiento de Cloudflare)

RAM Memoria de Acceso Aleatorio (Random Access Memory)

RGPD Reglamento General de Protección de Datos

SASE Acceso Seguro al Borde del Servicio (Secure Access Service Edge)

SSL Capa de Conexión Segura (Secure Sockets Layer)

SSD Disco de Estado Sólido (Solid State Drive)

TPU Unidad de Procesamiento Tensorial (Tensor Processing Unit)

TURN Traversal Using Relays around NAT

UE Unión Europea

Apéndice B

Manual de usuario y casos de uso

Bibliografía

Amazon. Datos abiertos en aws. Disponible en https://aws.amazon.com/es/opendata/.

Awesome data. Awesome public datasets. Disponible en https://github.com/awesomedata/awesome-public-datasets.

Ayuntamiento de Madrid. Portal de datos abiertos del ayuntamiento de madrid. 2025.

Brown, B., Chui, M. y Manyika, J. Are you ready for the era of 'big data'. *McKinsey Quarterly*, vol. 4(1), páginas 24–35, 2011.

CNMC. Telecomunicaciones anual datos generales cnmc. Disponible en https://data.cnmc.es/telecomunicaciones-y-sector-audiovisual/datos-anuales/datos-generales/telecomunicaciones-anual.

Congreso de los Estados Unidos de America. Health insurance portability and accountability act of 1996 (hipaa). 1996.

Congreso de los Estados Unidos de America. Foundations for evidence-based policymaking act of 2018, title ii: Open government data act. 2019.

Cornella, A. Conferencia - cómo sobrevivir a la infoxicación. 2000.

Esri España. Portal de datos abiertos de esri españa. 2025.

Estado de California. California consumer privacy act (ccpa). 2018.

Gómez-Obregón, J. Jaime gómez-obregón. Disponible en https://github.com/awesomedata/awesome-public-datasets.

Gómez-Obregón, J. La donación - jaime gómez-obregón. Disponible en https://github.com/JaimeObregon/subvenciones/tree/main/files.

Gómez-Obregón, J. Subvenciones - jaime gómez-obregón. Disponible en https://github.com/JaimeObregon/subvenciones/tree/main/files.

Gobierno de España. Portal de datos abiertos. 2025.

Google. Data commons. Disponible en https://docs.datacommons.org/custom_dc/.

Insights + Analytics, E. Datos para espaÑa y mundiales de investigaciÓn de mercados 2023. Disponible en https://ia-espana.org/wp-content/uploads/2024/10/NdpdatosEncuentro_16102024.pdf.

Jorge, R. y Herrera, R. Aumenta la negativa a abrir datos públicos: Informa inai tendencia de 4t a opacidad. llama presidenta a ciudadanos a iniciar defensa de ente autónomo. 2023. Copyright - Copyright Editora El Sol, S.A. de C.V. Mar 24, 2023; Última actualización - 2023-03-24.

Julián Valero Torrijos, R. M. G. *DATOS ABIERTOS Y REUTILIZACIÓN DE LA INFORMACIÓN DEL SECTOR PÚBLICO*. CRC Press, 2022. ISBN 978-84-1369-269-2.

Khan, S. y Alam, M. File formats for big data storage systems. *International Journal of Engineering and Advanced Technology (IJEAT) Volume-9 Issue-1* , 2019.

OECD. 2023 oecd open, useful and re-usable data óurdataíndex: Results and key findings. *OECD Public Governance Policy Papers, No. 43, OECD Publishing, Paris*, 2023.

Pangarkar, T. Big data statistics 2025 by patterns in the dimensions. Disponible en https://scoop.market.us/big-data-statistics/.

Press, G. A very short history of big data. Disponible en https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.

Ramos-Simón, L. F. El uso de las licencias libres en los datos públicos abiertos. *Revista Espanola de Documentacion Cientifica*, vol. 40(3), páginas 1–16, 2017. Copyright - Copyright Consejo Superior de Investigaciones Científicas Jul/Sep 2017; Última actualización - 2017-10-04.

Registradores de España. Portal de datos abiertos de los registradores de españa. 2025.

R.I.Pienaar. Free for devs. Disponible en https://github.com/ripienaar/free-for-dev.

Standing Committee of the National People's Congress. Data security law of the people's republic of china. 2021a.

Standing Committee of the National People's Congress. Personal information protection law of the people's republic of china. 2021b.

Taylor, P. Big data - statistics and facts. Disponible en https://www.statista.com/topics/1464/big-data/#topicOverview.

De la Torre, S. y Núñez, S. Transparencia en la administración pública municipal del ecuador. *Estudios de la Gestión*, (14), páginas 53–73, 2023. Copyright - Copyright null 2023; Última actualización - 2024-12-12; SubjectsTermNotLitGenreText - Ecuador.

Union Europea. Data protection under gdpr. Disponible en https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm.

Union Europea. Explicación de la ley de gobernanza de datos. Disponible en https://digital-strategy.ec.europa.eu/es/policies/data-governance-act-explained.

Union Europea. Reglamento de inteligencia artificial. Disponible en https://eur-lex.europa.eu/eli/reg/2024/1689.

Union Europea. Portal de datos abiertos. 2025.

UniversiDATA. Universidata. Disponible en https://www.universidata.es/.

Universitat Autònoma de Barcelona. infoparticipa. 2025.

Wikipedia. Information age. Disponible en https://en.wikipedia.org/wiki/Information_Age.