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

20 de septiembre de 2025

Autorización de difusión

El abajo firmante, matriculado en el Master en Ingeniería en Informática de la Facultad de Informática, autoriza 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

20 de septiembre de 2025

Dedicatoria

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

Agradecimientos

Agradecer a todas las personas que han aportado su granito de arena a la persona que soy, y por extensión, a este mismo trabajo. Sobre todo a profesores y compañeros de estudio y trabajo, de los que tanto he aprendido.

Resumen

La computación en la nube, la Inteligencia artificial y el tratamiento de grandes volúmenes de datos se revelaron como tecnologías profundamente disruptivas en el panorama tecnológico de los últimos años. En este trabajo se estudió el estado de la cuestión de estas tecnologías en profundidad, analizando su utilidad intrínseca y explorando las distintas metodologías, técnicas y servicios existentes, aportando también una implementación práctica para examinar su capacidad de aportar un beneficio público tangible.

Se examinó la disponibilidad de conjuntos de datos públicos a nivel europeo como base para el proyecto, cumpliendo con los nuevos marcos legales que este impone, A partir de ello, se elaboró una metodología replicable que describió el recorrido de los datos desde su origen hasta la generación de valor público, haciendo uso de la computación en la nube y las capacidades avanzadas de los algoritmos de Aprendizaje automático, así como las tecnologias de Auto-Machine Learning que las grandes nubes publicas proporcionan. Esto permitió identificar las configuraciones más eficientes y óptimas para la construcción de estas soluciones, comparando con el resto de opciones disponibles. De igual manera, permitió desarrollar un enfoque general del estado de las tres tecnologías estudiadas en la actualidad.

Los ficheros de GitHub se encuentran en el siguiente repositorio: `https://github.com/crismo04/TFM-cloud-soliutions-to-public-data/`

Palabras clave

Big Data, Computación en la nube, Inteligencia Artificial, Machine Learning, Open Data, Tratamiento de datos, Valor público

Abstract

Cloud computing, artificial intelligence, and big data processing emerged as profoundly disruptive technologies in the technological landscape of recent years. In this work, the state of the art of these technologies was studied in depth, analyzing their intrinsic utility and exploring the different methodologies, techniques, and services available, while also providing a practical implementation to examine their ability to deliver tangible public benefit.

The availability of public datasets at the European level was examined as the foundation of the project, in compliance with the new legal frameworks imposed in this area. Based on this, a replicable methodology was developed that described the journey of data from its origin to the generation of public value, making use of cloud computing and the advanced capabilities of machine learning algorithms, as well as the Auto Machine Learning technologies offered by major public cloud providers. This made it possible to identify the most efficient and optimal configurations for building such solutions, in comparison with other available options. Likewise, it enabled the development of a general overview of the current state of the three technologies under study.

The project files can be found in the following repository: `https://github.com/crismo04/TFM-cloud-soliutions-to-public-data/`

Keywords

Artificial Intelligence, Big Data, Cloud Computing, Data Processing, Machine Learning, Open Data, Public Value

Índice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

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

Este proyecto se basa en tres tecnologías: La computación en la nube, la cual se entiende como el suministro de servicios informáticos (servidores, almacenamiento, bases de datos, redes, software, análisis y más) a través de internet, permitiendo un acceso flexible y escalable a recursos sin la necesidad de poseer ni gestionar la infraestructura física; la inteligencia artificial, que abarca el desarrollo de sistemas que demuestran la capacidad de aprender, adaptarse, razonar y resolver problemas complejos, así como de percibir y comprender su entorno (virtual o físico), a menudo a través del análisis y la inferencia a partir de grandes volúmenes de datos (más específicamente, en este proyecto se emplearán las metodologías del aprendizaje automático, una rama de la IA centrada en la mejora del rendimiento en tareas específicas a través de algoritmos); Finalmente, los grandes volúmenes de datos se refieren a colecciones masivas y heterogéneas de información generada o recopilada por entidades gubernamentales u organizaciones, accesible al público.

1.1. Motivación

La motivación de este proyecto surge debido al enorme auge que han tenido las tres tecnologías estudiadas, así como de su capacidad para crear sinergias y aportar valor público. En el marco de la Unión Europea, se están tomando medidas para la liberalización de grandes volúmenes de datos públicos y

promoviendo su uso por entidades públicas, así como poniendo el foco en la inteligencia artificial y su potencial transformador. La sociedad es cada vez más consciente del potencial de modelos del lenguaje, pero aun falta camino para que estos sean capaces de generar beneficios sociales de la manera más automatizada posible. Este trabajo pretende centrarse en eso, en estudiar cual es la mejor manera de que los distintos actores (individuales, públicos y privados) pueda generar valor a través de estas tecnologías.

1.2. Objetivos y alcance

El objetivo y alcance de este proyecto es triple:

- Realizar un análisis exhaustivo del panorama actual de la computación en la nube, la inteligencia artificial y el estado de los datos públicos, identificando las metodologías, técnicas y servicios más relevantes.
- Desarrollar una metodología replicable que permita el uso de estos datos, desde su origen, hasta la generación de valor. Afinando esta metodología a través de la puesta en práctica de la misma.
- Aplicar la metodología propuesta a diversos conjuntos de datos disponibles a nivel europeo y nacional, evaluando diferentes modelos y configuraciones para identificar las soluciones más eficientes y efectivas para generar valor público a partir de la información analizada.

1.3. Plan de trabajo

Una vez definido el alcance, destacaremos las seis fases en las que se ha dividido el proyecto, que se han ido iterando en un esquema ágil para la creación de varias versiones funcionales:

1. **Fase de investigación académica:** Búsqueda de estudios o trabajos acerca del estado actual de las tres tecnologías y las últimas innovaciones para establecer el marco teórico y contextualizar las bases del proyecto.
2. **Fase de investigación técnica:** Búsqueda de información acerca de diferentes fuentes públicas 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.
3. **Fase de análisis de requisitos:** A partir del conocimiento adquirido, se diseñó la arquitectura de la solución, se seleccionaron los conjuntos

de datos públicos objetivo y se planificó la metodología concreta a seguir, estableciendo las métricas para evaluar el éxito del proyecto.

4. **Fase de implementación:** Fase central para materializar la solución. Esta aplica la metodología definida para llevar a la práctica e integrar todos los elementos: la configuración del entorno en la nube, la adquisición y limpieza de los datasets seleccionados, el entrenamiento de los modelos de aprendizaje automático y su despliegue y el uso del resto de tecnologías.

5. **Fase de pruebas y evaluación:** Valoración de los prototipos desarrollados y resultados obtenidos, así como del rendimiento de los modelos y la calidad del valor público generado, teniendo en cuenta el aspecto ético de los mismos.

6. **Fase de documentación:** Etapa transversal al resto, documentando los pasos seguidos y hallazgos en la elaboración de este documento, plasmando también los resultados obtenidos, las conclusiones y las posibles líneas de trabajo futuras.

1.4. Estructura de esta memoria

Toda esta memoria se ha construido con LaTeX [3.2.2] y ayuda de la plantilla TeXiS, El resto de la memoria se estructurará por capítulos de esta manera:

Capítulo 2: Estado de la cuestión, donde se plasmaran las conclusiones de las primeras dos fases de investigación: estudios, trabajos, marcos, tecnologías, conjuntos de datos y demás herramientas encontradas. También estudiará el panorama actual en la sociedad con respecto a estas.

Capítulo 3: Materiales y métodos, donde se plasmaran las siguientes dos fases de análisis e implementación, concretando las tecnologías y conjuntos de datos utilizados finalmente en el proyecto, al igual que la metodología y pasos que se han seguido durante el proyecto para obtener resultados.

Capítulo 4: Resultados y trabajo futuro En este último capitulo (que también estará disponible en ingles), se concluirá con los resultados obtenidos en el proyecto, así como las líneas de investigación que inevitablemente quedan abiertas debido a la extensión de los temas tratados.

Anexo A "Definiciones y acrónimos": Para evitar la excesiva longitud de ciertos apartados, **algunas definiciones se han movido a este apartado, apareciendo en el texto de la siguiente manera: [Definición 1]**. Aquí también se podrán encontrar los Acrónimos que aparecen durante todo el trabajo [A.2].

Aclarar que esta memoria utilizará las palabras en español o sus anglicismos correspondientes indistintamente, debido a su popularización y uso en la rama de la computación. Esto es palabras tales como "machine learning" o "aprendizaje automático", "dataset" o "conjunto de datos", "cloud" o "nube", etc.

Chapter 1

Introduction

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

This project is based on three main technologies: Cloud computing, which is understood as the provision of computing services (servers, storage, databases, networking, software, analytics, and more) over the internet, allowing flexible and scalable access to resources without the need to own or manage physical infrastructure; artificial intelligence, which encompasses the development of systems that demonstrate the ability to learn, adapt, reason, and solve complex problems, as well as perceive and understand their environment (either virtual or physical), often through analysis and inference from large volumes of data (More specifically, this project will employ machine learning methodologies, a branch of AI focused on improving performance in specific tasks through algorithms); finally, public big data refers to massive and heterogeneous collections of information generated or collected by governmental or organizational entities, publicly accessible.

1.1. Motivation

The motivation for this project arises from the enormous growth of the three studied technologies, as well as their capacity to create synergies and deliver public value. Within the framework of the European Union, measures are being taken to liberalize large volumes of public data and promote their use by public entities, while also focusing on artificial intelligence and its

transformative potential. Society is increasingly aware of the potential of language models, but there is still a long way to go before these can generate social benefits in the most automated way possible. This work aims to focus precisely on that: studying the best way for different actors (individuals, public and private entities) to generate value using these technologies.

1.2. Objectives and Scope

The objective and scope of this project are triple:

- Conduct a comprehensive analysis of the current landscape of cloud computing, artificial intelligence, and the state of public data, identifying the most relevant methodologies, techniques, and services.
- Develop a replicable methodology that allows the use of these data, from their origin to the generation of value, refining this methodology through practical implementation.
- Apply the proposed methodology to various datasets available at the European and national levels, evaluating different models and configurations to identify the most efficient and effective solutions for generating public value from the analyzed information.

1.3. Work Plan

Once the scope has been defined, we will highlight the six phases into which the project has been divided, which have been iterated in an agile framework for the creation of several functional releases:

1. **Academic research phase:** Search for studies or works about the current state of the three technologies and the latest innovations to establish the theoretical framework and contextualize the foundations of the project.
2. **Technical research phase:** Search for information about different public data sources, cloud technologies, and AI models or tools that help process, filter, and understand all the collected public data.
3. **Requirements analysis phase:** Based on the knowledge acquired, the solution architecture was designed, target public datasets were selected, and a concrete methodology was planned, establishing metrics to evaluate the project's success.

4. **Implementation phase:** Central phase to materialize the solution. This applies the defined methodology to practically integrate all elements: cloud environment configuration, acquisition and cleaning of selected datasets, training of machine learning models, their deployment, and the use of the other technologies.
5. **Testing and evaluation phase:** Assessment of developed prototypes and obtained results, as well as model performance and the quality of the public value generated, taking ethical aspects into account.
6. **Documentation phase:** Cross-cutting stage documenting the steps followed and findings in the preparation of this document, also reflecting the obtained results, conclusions, and future lines of work.

1.4. Structure of this Thesis

This thesis has been built using LaTeX [3.2.2] with the help of the TeXiS template. The rest of the thesis is structured in chapters as follows:

Chapter 2: State of the Art, which will present the conclusions of the first two research phases: studies, works, frameworks, technologies, datasets, and other tools found. It will also study the current landscape in society regarding these topics.

Chapter 3: Materials and Methods, which will cover the next two phases of analysis and implementation, specifying the technologies and datasets finally used in the project, as well as the methodology and steps followed to obtain results.

Chapter 4: Results and Future Work In this last chapter (which will also be available in English), the results obtained in the project will be concluded, as well as the research lines that inevitably remain open due to the breadth of the topics covered.

Appendix A "Definitions and Acronyms": To avoid excessive length in certain sections, **some definitions have been moved to this appendix, appearing in the text as follows: [Definition 1]**. Here, the acronyms used throughout the work can also be found [A.2].

Capítulo 2

Estado de la Cuestión

"Somos una generación frontera. La única que ha conocido la vida antes y después de la hiperconectividad y los dispositivos móviles. [...] La última que pudo abarcar toda la tecnología de su tiempo."
— Jaime Gómez-Obregón

En este apartado expondremos el estado actual de los puntos principales de este proyecto, así como los trabajos o artículos relacionados con los temas a tratar: trabajos relacionados con los principales proveedores Cloud y su comparación, trabajos que traten con grandes volúmenes de datos públicos o que estudien los datos públicos, o trabajos que utilicen diferentes IAs para el tratamiento de datos y la obtención de conclusiones a partir de estos. No es una tarea fácil, ya que los artículos relacionados en "Google Scholar" se cuentan por millones al buscar "data", "Artificial Intelligence" o "cloud computing", por lo que, aunque también se estudiaran conjuntos de datos, artículos y aplicaciones de otras partes del globo, la parte practica del proyecto se intentara centrar en datos del territorio español, 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).

También dividiremos esta sección en los tres elementos que componen el trabajo, y estudiaremos las posibilidades y estado de la cuestión de cada uno.

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 acuñó en 2005 el termino "web 2.0"y "Big data" (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, llegarán a generan, según proyecciones, la asombrosa cifra de 149 Zettabytes de datos en 2024, ¡un 1 seguido de 23 ceros en bytes! (Taylor, 2025) & (Pangarkar, 2025) (numero impresionante a pesar de la naturaleza especulativa de estas proyecciones). Esta evolución no ha sido lineal ni uniforme, sino que ha estado marcada por distintos enfoques, motivaciones y metodologías en todo el mundo en lo que se denominan las tres olas de datos abiertos [Definición 5] , diferentes etapas evolutivas por las que ha transitado el movimiento de apertura de datos

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, o a que los datos tienen mas valor de mercado que en otras regiones. Aún con estas discrepancias en cuanto a números, lo que está claro es que el mercado de los datos no para de crecer año tras año y cada vez resulta más difícil separar la información relevante del ruido, evitando la "infoxicación" o 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 "datos" no 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 porque su extracción o tratamiento fueran complejos, ya que este es el caso para la mayoría de aplicaciones en el mundo real. Esto se explicará más en detalle en el Capítulo **??**: Materiales y métodos

2.1.1. Obtención de datos públicos

Lo primero para la realización de este proyecto es la obtención de datos públicos, o datos abiertos [Definición 4]. Esto presenta tres grandes complicaciones a tener en cuenta:

La primera es que, aunque existe un consenso creciente sobre la importancia de la apertura de datos, la realidad muestra que **muchos datos de alto valor aún no son accesibles, lo son de forma limitada o su uso es complejo**, 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 pública (Soledad De la Torre, 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 más 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 verificarán 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 panoramas diversos los cuales vale la pena mencionar, desde una regulación más laxa en Estados Unidos, hasta un control estricto en países como China. Estos datos no se utilizarán en este trabajo por temas de alcance, ya que se prefiere dar prioridad a fuentes de datos nacionales, pero las herramientas desarrolladas serían aplicables a estos mismos datos cumpliendo sus normativas.

En Estados Unidos, el panorama es sobre todo 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 "OPEN 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 Commit-

tee 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 último, 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 más amplio posible, para estudiar y comparar un amplio abanico de soluciones.

2.1.2. Trabajos anteriores y relacionados

A parte de todas las referencias ya incluidas en esta sección, nos 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.

Sobre "Big data" y datos públicos, han surgido trabajos en España desde sus inicios (Ferrer-Sapena A., 2011) [Definición 2] desde diversos campos como las Ciencias de la Información, y uno de los más completos que he podido encontrar en nuestro territorio es (Herrera Capriz, 2024), un reciente y extenso trabajo sobre los datos abiertos en España donde, partiendo de una extensa experiencia en la administración pública, la autora busca combinar dos campos con demandas complementarias bajo el marco teórico de la "Teoría de la Ventana" [Definición 6] y estudios anteriores de valor público (Meynhardt, 2009):

- Los Estudios de **Valor Público** [Definición 3]: Carecen de una extensa evidencia empírica sólida.
- La **transparencia y los Datos Abiertos** [Definición 4]: Carecen de medición del valor final que generan para los ciudadanos.

La autora trata de medir el valor real que la transparencia y los datos abiertos generan para los ciudadanos, más allá de su mera publicación. Para ello propone un marco metodológico que permite cuantificar el valor de los da-

tos abiertos a través de la percepción de los usuarios. Este enfoque consigue identificar que las dimensiones utilitaria y hedonista (relacionadas con la funcionalidad y la experiencia de usuario) reciben puntuaciones altas, mientras que las dimensiones político-social y moral-ética (relacionadas con la generación de comunidad, equidad y trato justo) lastran el valor potencial, y detectando también que determinantes clave como la frecuencia de uso y el tipo de datos (geoespacial, movilidad, turismo) son factores condicionantes para maximizar el valor.

Destaco la gran labor de investigación sobre datos abiertos del trabajo, que ha sido clave como base para la realización de este mismo proyecto y la utilidad de los resultados, que influenciará en la utilización de los datos públicos de este proyecto.

En cuanto a trabajos más práctico-tecnológicos, hay muchos de ellos donde destacar, la Unión Europea en sus estudios de casos de uso sobre datos públicos, tiene más de 600 casos estudiados, 150 con impacto significativo, 30 participaron en el estudio del volumen 1 (Giulia Carsaniga, 2022) y finalmente 13 en el volumen dos del mismo (Nijssen, 2022), de los cuales nos gustaría destacar 3 españoles y uno Francés:

- **UniversiDATA:** Un portal que integra seis universidades españolas para el análisis avanzado y dinámico de datos abiertos con el objetivo de crear resultados interactivos y en tiempo real, facilitando el uso compartido de recursos y mejorando la comprensión de datos abiertos (UniversiDATA, 2025). El equipo también fomentaba el uso de sus datos con diferentes análisis propios (UniversiDATA, 2020) o el lanzamiento de eventos centrados en datos (o "Datathones" [Definición 7] (UniversiDATA, 2024) de los cuales hablaremos a continuación). También resuelven dudas de usuarios e investigadores en los conjuntos de datos o análisis a través de comentarios, fomentando aún más la comprensión de los datos

- **Tangible Data:** Transforma datos espaciales digitales en esculturas físicas accesibles.

- **Planttes:** Aplicación que informa sobre la floración de plantas y su impacto en las alergias al polen, combinando datos abiertos con aportaciones de usuarios, fomentando la concienciación, información y educación sobre alergias (Concepción De Linares, 2025b).

- **Open Food Facts:** Aplicación que informa sobre detalles de productos de supermercado, queriendo nombrarla por la enorme cantidad de datos que ha conseguido recopilar de usuarios de todo el mundo y lo intuitiva que es a la hora de usar toda esta cantidad de datos (más de 1 millón de productos). (Concepción De Linares, 2025a).

Como ya hemos comentado otra fuente importante de proyectos relacionados con datos serían los "Datathones" [Definición 7], de los cuales pueden salir decenas de proyectos relacionados con datos abiertos en muy poco tiempo y que seria inabarcable mencionar en este proyecto debido a los más de 20 Datathones diferentes encontrados y los múltiples proyectos que hay en cada uno, pero si que nos gustaría mencionar iniciativas como la del gobierno de España (Gobierno de España, 2025b) y (Gobierno de España, 2025a) con más de cuarenta eventos sobre datos a fecha de publicación de este trabajo.

Por último, también mencionaremos trabajos académicos de compañeros que han implementado soluciones con datos públicos, que aunque son algo antiguos siguen aportando valor:

- **"Auditoría y metodología de implantación de open data para smart cities"** (Melendrez Moreto, 2016): Donde el autor hace un análisis extensivo de los datos abiertos, de los indices y métricas para evaluar el valor de estos datos y de herramientas como "CKAN" para la gestión de los datos. Audita diversas fuentes de datos nacionales dejando todas dentro del umbral de datos abiertos según AENOR.
- **"Uso de geolocalización y de fuentes de datos abiertas para la creación de servicios turísticos por la ciudad de Madrid"** (LLamocca Portela, 2016): El cual utiliza datos abiertos de geolocalización en Madrid para buscar sitios cercanos en una app móvil.
- **"Integración y visualización de datos abiertos medioambientales"** (Arellano Bruno, 2019): También hace un análisis extensivo de los datos abiertos, de las definiciones para evaluar estos datos y de herramientas como "CKAN". Además, comenta iniciativas de limpieza de datos interesantes como "Data Tamer" o "Data Wrangler". Finalmente, crea una aplicación para el uso de datos medioambientales en tiempo real.

2.1.3. Conjuntos de datos

Para embarcarse en el tema de los conjuntos de datos, primero tenemos que saber qué tipos de datos hay en el mundo real (Sarker, 2021) dependiendo de formato y forma, estos pueden ser:

- **Estructurados:** Datos con un formato definido, siguiendo un modelo de datos estándar como tablas o bases de datos relacionales (fechas, direcciones, geolocalización).
- **No estructurados:** Datos sin un formato o estructura predefinida, mas difíciles de procesar y analizar, principalmente texto y multimedia

(correos electrónicos, blogs, PDFs, audio, videos, etc.).

- **Semi-estructurados (Semi-structured):** Datos que no se almacenan en bases de datos relacionales, pero tienen cierta organización (HTML, XML, JSON o bases de datos NoSQL).
- **Metadatos (Metadata):** Datos sobre otros datos, que describen información relevante para dar significado a los datos (autor de un documento, tamaño de archivo, fecha de creación palabras clave, etc.).

Teniendo esto en cuenta, para el desarrollo de este proyecto donde pretendemos obtener valor de los mismos de la forma mas automática y escalable, se elegirán principalmente datos estructurados, semiestructurados y los metadatos que proporcionen los portales. También, recalcando la prioridad de este proyecto en conjuntos de datos cercanos, me gustaría destacar algunos de los conjuntos de datos e iniciativas mas interesantes encontrados durante el estudio:

- **Internacionales:** Iniciativas individuales como "Awesome public datasets" (Awesome data, 2025), que recopila fuentes de datos fiables (aunque principalmente de Estados Unidos).
- **Europeos:** Europa cuenta con su propio portal para acceder a datos públicos (Union Europea, 2025b) con datasets sobre población, educación y ciencia.
- **Españoles:**
  - 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 investigación e innovación.
  - Proyectos como el Portal de Transparencia del Gobierno de España y las iniciativas de datos abiertos de comunidades autónomas y ayuntamientos (Gobierno de España, 2025c); (Ayuntamiento de Madrid, 2025) & (Registradores de España, 2025), los cuales se esfuerzan por hacer públicos datos de "alto valor" [Definición 1].
  - Famosas iniciativas como "Kaggle" cuentan con datos de todo tipo sobre España, muchos recopilados de otras fuentes publicas ya nombradas, pero otros interesantes (Kaggle, s-f.b).
  - Iniciativas que fomentan la transparencia, como InfoParticipa (Universitat Autònoma de Barcelona, 2025).
  - Iniciativas privadas para la recolección de datos públicos (Esri España, 2025).
  - Iniciativas ya mencionadas 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. Los conjuntos de datos escogidos se detallan en el apartado Sección 3.2.1: Materiales y datos.

2.1.4. Gobierno de los datos

Por último, y teniendo claro todo explicado sobre los datos, destacaremos la importancia de la gobernanza [Definición 8] de los mismos como uno de los desafíos fundamentales al manejarlos (Theodorakopoulos Leonidas, 2024). En el marco europeo llevamos años promulgando mecanismos para aumentar la confianza para un mayor y mejor intercambio de datos (European Parliament, 2022). El reglamento presenta tres vías principales: la reutilización segura de datos protegidos del sector público, la intermediación neutral de datos y la cesión altruista de datos para el interés general. Estos mecanismos operativos materializan los principios de la gobernanza: calidad, seguridad, interoperabilidad y confianza.

Para la gestión de datos en este proyecto, adoptaremos un marco de gobernanza de datos basado en el modelo de tres capas propuesto por la OECD (Estratégica, Táctica y Operativa) (OECD, 2019), integrado con el resto de normativas europeas vigentes, (como "Data Governance Act" (European Parliament, 2022) o el Reglamento General de Protección de Datos (Union Europea, 2016)).La aplicación de este marco se articulara en la Sección 3.1.2.1: Métodos de Gobernanza.

2.2. Nubes

La capacidad real de extraer valor de los volúmenes masivos de datos abiertos detallados en la sección anterior está intrínsecamente ligada a la disponibilidad de recursos computacionales potentes, escalables y económicamente accesibles. Y aunque se lleva años hablando de soluciones como las nubes distribuidas [Definición 9], el paradigma de la computación en la nube gestionada (o publica, que no abierta) (*cloud computing*), con planes gratuitos, es muy relevante para democratizar este acceso, permitiendo a investigadores, startups e instituciones públicas superar las limitaciones del hardware local.

Estas nubes gestionadas nos proporcionan unidades de procesamiento gráfico (GPUs) y tensorial (TPUs) bajo demanda, además de un ecosistema completo de servicios gestionados diseñados específicamente para el ciclo de vida completo de los datos y la IA. También nos brindan mecanismos de seguridad que junto a la implementación diligente por parte del usuario (modelo de responsabilidad compartida) aporta la seguridad necesaria para el proyecto. Para este trabajo se ha optado por esta solución. Otra opción posible sería la nube híbrida, que combina el control de una infraestructura privada ("On premise") con la escalabilidad y economía de la nube pública. Esto es ideal para cargas con datos sensibles, pero complicaría en enfoque de este proyecto. Para ayudarnos mejor a valorar todas las opciones, vamos a listar la oferta gratuita de algunas de las nubes estudiadas (Microsoft, s.f.a), (Lisdorf, 2021).

En el panorama nacional, Eurostat muestra que el 35.8 % de las empresas españolas usaban tecnologías cloud en 2024, tendencia que va en aumento (pronosticando un 17 % más para los próximos años: Market report analytics (2024)), pero que se encuentra por debajo en comparación con Europa, donde se usa en más del 45 % de las empresas (Eurostat, s.f.). de cara al futuro, mientras la estrategia europea el "edge computing" frente a la cloud privada para ganar autonomía tecnológica y soberanía de datos (European Commission, s.f.a), el enfoque del Plan de Digitalización español (Gobierno de España, 2021b) no se aleja de la nube pública, sino que aboga por un modelo híbrido soberano usando NubeSARA (plataforma híbrida que, por sus precios (Eurostat, s.f.), e información encontrada (Gobierno de España, s.f.), (PreparaTIC, s.f.) no vamos a utilizar en este estudio).

Matiz sobre "edge computing": Aunque es especialmente importante en el uso de aplicaciones de IA [Tabla A.2], debido a la importancia de la privacidad en escenarios con datos tratados automáticamente, pero al ser esto una investigación y no una tecnología dirigida a un usuario final (el cual requeriría velocidad y privacidad), se optara por ejecutar los modelos en la nube en vez en edge, aunque se reconoce su potencial para aplicaciones practicas.

También hay alternativas para usar cloud en el territorio europeo y español, como pueden ser (clouding.io, s.f.), (GmbH, s.f.) o (Gigas, s.f.), en las cuales se podría desplegar maquinas virtuales para la ejecución de modelos y tratamiento de datos, pero que al no tener infraestructura especializada en Inteligencia artificial y carecer en su mayoría de capa gratuita, se han excluido del estudio. También seria interesante utilizar herramientas europeas como BDTI (European Commission, s.f.b), (Gobierno de España, 2021a), la cual brinda infraestructura gratuita, pero solo a petición de organismos públicos para proyectos como este, o SIMPL (European Commission, 2024), un framework tecnológico, un middleware para construir sobre diferentes proveedores cloud y edge. Teléfonica Tech también vende una especie de nube pública basada en VMware, pero principalmente son servicios gestionados multi-cloud (Telefónica, s.f.).Por último, nombrar dos herramientas europeas que si pueden resultar utiles, y que analizaremos en la próxima sección, "OVHcloud" (proveedor de cloud francés) y "OpenNebula" (plataforma española open-source para gestionar clouds).

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. (R.I.Pienaar, 2025).

Google Cloud Platform

Ecosistema de servicios en la nube Google con infraestructura escalable, herramientas de análisis y soluciones de inteligencia artificial gestionadas que cubre todo el ciclo de vida de los datos y aplicaciones. A parte de las aplicaciones listadas y muchas mas que se pueden encontrar en su **Lista completa**: `https://cloud.google.com/free`, Google también ofrece 300€ para exceder estos limites los primeros 3 meses de prueba, lo cual puede ayudar enormemente a proyectos de tamaño medio.

Servicios específicos:

- **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".

Amazon Web Services

Plataforma de cloud computing más usada a nivel empresarial, con una enorme cantidad de servicios, desde cómputo básico hasta servicios de IA, machine learning, IoT, etc. Tiene una capa gratuita de 12 meses, aquí se puede consultar la **Lista completa** de servicios: `https://aws.amazon.com/free/`

Servicios específicos:

- **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.

Microsoft Azure

Ecosistema cloud de Microsoft, muy integrado con todas sus herramientas empresariales y de desarrollo, como la suite de DevOps, copilot y más soluciones de IA, enfoque en el lenguaje .NET. También tiene una capa gratuita de 12 meses,Aquí se puede consultar la **Lista completa** de servicios: `https://azure.microsoft.com/free/`

Servicios específicos:

- **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.

Oracle Cloud

Nube especializada en bases de datos de alto rendimiento (Oracle Database), aplicaciones Java, y soluciones de analytics. Ofrece una capa gratuita con recursos que no expiran, aquí se puede consultar la **Lista completa** de servicios: `https://www.oracle.com/cloud/free/`

Servicios específicos:

- **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.

IBM Cloud

Plataforma centrada en la transformación digital de grandes empresas con necesidades híbridas y multicloud, aunque también con una capa gratuita. Se puede consultar la **Lista completa** de servicios aquí: `https://www.ibm.com/cloud/free/`

Servicios específicos:

- **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.

Cloudflare

Plataforma especializada en rendimiento web, seguridad y confiabilidad. Aunque no es una nube generalista, sino más bien una red global que acelera y protege sitios web, APIs y aplicaciones mediante su CDN, DNS, servicios de seguridad, etc. De todas formas tambien tiene capa gratuita, la **Lista completa** se encuentra en: `https://www.cloudflare.com/plans/free/`

- **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.

OVHcloud

Proveedor de cloud francés con un fuerte compromiso con la soberanía de los datos y el RGPD. Ofrece una gama completa de servicios de infraestructura (IaaS) y plataforma (PaaS) desde sus centros de datos.Ofreces 200€ en creditos para probar el servicio durante un mes, aunque tiene una **Lista completa** de servicios: `https://www.ovhcloud.com/en/public-cloud/prices/`, los disponibles en el plan gratuito son los siguientes (OVHCloud, s.f.):

- **Despliegue de un e-commerce**: con 2 servidores B2-7, 1 base de datos MySQL, 1 IP adicional, 1 Balanceador de Carga y 10 GB de Almacenamiento de Objetos.
- **Prueba de Kubernetes y escalado**: 3 servidores B2-15 durante 1 mes y 12 horas de picos de tráfico en 10 servidores C2-30.
- **Desarrollo y Entrenamiento de IA**: 1 TB de Almacenamiento de Objetos, 35 horas de IA Notebook (AI1-1-GPU) y 5 horas de entrenamiento de IA en 4 nodos AI1-1-GPU.

2.2.2. Otras herramientas interesantes

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 "Jupyter" en 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`

Open Data Editor

Herramienta de código abierto `https://okfn.org/en/projects/open-data-editor/` diseñada para la gestión y publicación de datos abiertos. Desarrollada por la "Open Knowledge Foundation", facilita la creación, validación y limpieza de conjuntos de datos en formatos abiertos, con un enfoque en la usabilidad para usuarios no técnicos, para garantizar la calidad y accesibilidad de los datos públicos. Funcionalidades clave:

- Creación y edición tabular de datos.
- Validación de esquemas y meta datos.
- Integración con plataformas de datos abiertos (CKAN, S3, etc.).
- Exportación a formatos estandarizados (CSV, JSON, XLSX).

RapidMiner Studio

Aunque sin ser una cloud, tiene relevancia por ser una plataforma de software para el ciclo de vida completo de la ciencia de datos y el aprendizaje automático (ML), abarcando desde la preparación de datos hasta la creación y despliegue de modelos, centrándose en la automatización y la facilidad de uso mediante una interfaz gráfica de usuario con modelo de "cajas" (university, 2024).

- **Tipo:** Plataforma integral de Data Science y Machine Learning (desde ETL hasta MLOps básicos).
- **Capa Gratuita - RapidMiner Studio Free:**
    - **Funcionalidades:** Acceso a algoritmos de ML (clasificación, regresión, clustering, etc.), herramientas de preparación de datos,

evaluación y visualización. Incluye capacidades de AutoML para la selección de modelos y optimización de hiperparámetros.

- **Limitaciones:** Generalmente restringido en el número de filas de datos procesables en memoria y el número de operadores por proceso, adecuado para aprendizaje y proyectos de tamaño pequeño a mediano.

- **URL:** `https://altair.com/altair-rapidminer`

Nubes descentralizadas

Herramientas como (Golem Network, s.f.), (Akash, s.f.) ó (Render Network, s.f.) podrían ser útiles en proyectos con exceso de potencia de computación, ya que se podría alquilar esta a cambio de tokens que mas tarde se podría usar para las tareas intensivas cuando fuera necesario [Definición 9].

OpenNebula

Plataforma de código abierto con origen en España y EE.UU `https://opennebula.io/`, la cual se centra en virtualización de centros de datos y gestión de nubes privadas, híbridas y públicas. Permite construir y gestionar infraestructuras IaaS (Infraestructura como Servicio) para cloud sobre la infraestructura de tecnología existente. Ofrece funcionalidades para el aprovisionamiento automático, es independiente del proveedor y soporta diversas interfaces de nube, proporcionando flexibilidad, control y soberanía de datos (Kumar et al., 2014), (Vogel et al., 2016). Sin embargo, pese a ser una opción interesante, es posible que no se tenga en cuenta para este proyecto, que prioriza el acceso inmediato a hardware y servicios de IA gestionados sin coste inicial de infraestructura ni tiempo.

2.2.3. Trabajos anteriores y relacionados

La literatura referente a la computación en la nube es muy extensa, a parte de toda la literatura ya citada hasta ahora, citare algunos ejemplos mas, como el libro **"Cloud Computing Technology"** (Huawei Technologies Co., Ltd., 2023), aunque su traducción al ingles no es excelente, agrupa todos los conceptos del panorama cloud actual, así como los elementos que todas las nubes comparten entre si, y da un panorama de la situación cloud en china, lo cual amplia los horizontes del conocimiento cloud. Otro libro: (Fowdur, 2021), cuyo capitulo 2 tiene interesantes definiciones sobre cloud. También

el trabajo de (Nigro, 2022), que estudia las oportunidades, desafíos y antecedentes de la computación en la nube, o el trabajo de (Bommala et al., 2024), que tiene un enfoque muy interesante, denominando al ecosistema "cloud verse" y enfocándose en las innovaciones de los últimos años (infraestructuras de nube híbrida, modelos de computación sin servidor o "serverless", "edge computing", integración de IA, etc.) con un interesante enfoque en seguridad y cumplimiento de normativas, la integración de "blockchain" y el énfasis en la computación en la nube "verdez sostenible.

Aunque no unicamente dirigido a cloud, también vale la pena destacar portales o iniciativas que buscan opciones de código abierto o de fácil acceso a las tecnologías sobre las que trata el articulo y las cuales han sido de mucha ayuda, como la iniciativa de la Unión Europea (Union Europea, 2025a) o compilaciones como (R.I.Pienaar, 2025).

Por último, querría destacar trabajos de compañeros como **"Optimización de infraestructuras de Cloud Computing basadas en máquinas virtuales"** (Sánchez de Paz, 2023), el cual hace una excelente labor de investigación de todo lo relacionado con la computación en la nube para predecir el consumo futuro de recursos en máquinas virtuales de Azure (con sus datos públicos), y así mejorar la eficiencia y la gestión de la infraestructura.

2.3. Inteligencia Artificial

La Inteligencia Artificial (IA), y mas específicamente el aprendizaje automático (ML), esta emergiendo como el paradigma tecnológico más transformador de la década y uno de los mas importante de la historia [Definición 12]. Aunque desde sus inicios ha pasado por algunos "inviernos" (Cheok y Zhang, 2023), el lanzamiento de "ChatGPT" en 2022 catapulto la fama de esta tecnología [Figura A.3]. Estudios del ONTSI (Observatorio Nacional de Tecnología y Sociedad, 2025) revelan que en 2024, el 11,4 % de las empresas españolas de 10 o más empleados usa IA, un dato ligeramente inferior a la media de la UE (13,5 %), aunque por encima de potencias como EE.UU. (5,7 % según OECD, o 7,3 % del total de empresas según BTOS). En cuanto a la población general, en 2024 un 73,8 % de los españoles tenían conocimiento de la existencia de la IA generativa, y un 56,8 % de la utilizó durante el año.

Aunque esta fama viene dada por los modelos generativos basados en "transformers" y modelos generativos (Vaswani et al., 2017), la aplicación de esta tecnología aplicación va mas allá de esto, el núcleo de esta revolución reside en el conjunto de herramientas algorítmicas para el análisis predictivo y la extracción de patrones en datos masivos, lo cual incluye los modelos generativos, pero no exclusivamente. En este proyecto se pretende utilizar ese potencial a través de plataformas cloud como Google Vertex AI, el uso de diferentes modelos de clasificación y clusterización o de detección de anomalías, con el propósito de intentar transformar el vasto volumen de datos públicos, a menudo infrautilizados, en "insights" o conclusiones de valor publico.

También aclarar que, por acotar el alcance del proyecto, este trabajo no destacará avances en IA que traten datos multimedia, como vídeos, audios y imágenes, aunque reconocemos el increíble avance que han tenido tecnologías de visión por computador o creación de multimedia y si utilidad en LLMs multimodales, los cuales si entran en el "scope" del proyecto. Con esto en mente, en las siguientes secciones vamos a ver como aplicar el aprendizaje automático al proyecto, revisando los modelos tradicionales, tecnologías que proporcionan las nubes para usarlos, como el AutoML. Igual que para los datos se han tenido en cuenta ideas de gobernanza de datos, para la aplicación de los modelos se tendrán en cuenta los principios de las Operaciones de aprendizaje automático (MLOps) [Definición 14].

2.3.1. Modelos de ML

Vamos a estudiar los modelos de aprendizaje automático pertinentes para el análisis predictivo y la extracción de patrones en datos. Aunque se pueden

clasificar de varias maneras, clasificaremos los modelos por:

- Tipos de aprendizaje según la disponibilidad y calidad de etiquetas (supervisado, no supervisado ,por refuerzo, etc.) [Definición 13]
- La naturaleza de los datos (tabulares, no estructurados, etc.) [2.1.3]

Debido a la utilización principal de datos estructurados y semiestructurados, se priorizaran los modelos adaptados mejor a este tipo de datos (Fowdur, 2021).

Para la elección del modelo, seguiremos el siguiente esquema (Brownlee, 2024b):

Paso uno: ¿Se necesita predecir algo?

- **Si (Aprendizaje Supervisado):**
  - Si la predicción es sobre **categorías** → *Tarea de Clasificación.*
  - Si la predicción es sobre **datos temporales** → *Predicción de Series Temporales.*
  - Si la predicción es sobre **otras características** → *Tarea de Regresión.*
- **No (Aprendizaje No Supervisado):**
  - Si el objetivo es **encontrar grupos** → *Clusterización.*
  - Si el objetivo es **encontrar anomalías** → *Detección de Anomalías.*

Paso dos: ¿Qué tipo de datos tenemos?

- **Simple:** Datos Estructurados → Algoritmos ML Simples:
  - **Árboles de Decisión o k-Nearest Neighbors** (para clasificación / regresión)
  - **K-means** (para clusterización)
- **Intermedio:** Múltiples Características / Imágenes de Baja Resolución → Métodos de Ensemble:
  - **Random Forests, XGBoost** (para clasificación / regresión)
  - **DBSCAN** (para clusterización)
- **Complejo:** Imágenes / Texto / Audio → Algoritmos mas complejos:
  - Redes Neuronales Profundas
  - Grandes modelos de Lenguaje (LLM)

Para tomar mejor las decisiones, se han estudiado estos modelos para comprobar su valor dependiendo del tipo de datos [Definición 14j]. Pero debido a que los modelos elegidos para cada conjunto de datos depende del tipo de los mismos (e incluso aun sabiendo el tipo de datos es dificil saber si ciento modelo funcionara mejor que otro), estos se detallaran para cada conjunto de datos en el apartado Sección 3.2.1: Materiales y datos.

2.3.2. ML en la nube

Todos los modelos que se han comentado, particularmente los más complejos, tienen algo en común, son bastante exigentes en el consumo de procesamiento, tanto de CPU, como sobretodo de GPU o TPU (Sze et al., 2017) Esta barrera de entrada hardware ha sido tradicionalmente un impedimento para investigadores, estudiantes y pequeñas organizaciones para ejecutar estos modelos, y en este punto es donde puede ser muy útil el paradigma de la computación en la nube y, sobre todo, los grandes proveedores con sus capas gratuitas, las cuales no solo proporcionan la infraestructura base para poder crear estos modelos, sino que además tienen herramientas especializadas para ello, facilitando la implementación de un ciclo de vida completo de Machine Learning (MLOps).

En este proyecto utilizaremos varias herramientas. En cuanto a los datos, los almacenaremos dependiendo de la nube utilizada para su procesamiento, pues cada una de las nubes puede comunicarse con un numero limitado y especifico de herramientas de almacenamiento. Los modelos entrenados se almacenaran también en la solución de almacenamiento de la propia nube (AWS S3, Azure Storage, etc.) cuando se requiera su uso, pero para entrenamiento o inferencia, a parte de las plataformas de MLOps comentadas, también merece la pena destacar opciones como "Kaggle notebooks" o "HuggingFace Spaces" que detallaremos en los siguientes apartados.

2.3.2.1. Plataformas MLOps en la Nube

Los principales proveedores de nube ofrecen plataformas completas diseñadas para gestionar el ciclo de vida de los modelos de ML (Böer, 2023):

- **Google Vertex AI:** Integra herramientas para todo el ciclo de vida del ML: preparación de datos, entrenamiento, despliegue y monitoreo en su GUI. Abstrae gran parte de la gestión de infraestructura y ofrece comunicacion con las principales herramientas de GCP como Bigquery. Además, ofrece una gran selección de modelos preentrenados de ML para diversos problemas, incluyendo modelos de Hugging Face, Stable Diffusion, etc. (Google, s-f.a) .

- **AWS SageMaker:** Proporciona un conjunto completo de herramientas y servicios para construir, entrenar y desplegar modelos de ML a escala, incluyendo notebooks gestionados, algoritmos y entrenamiento distribuido. Soporta una amplia gama de fuentes de datos como S3, Athena, Redshift, Snowflake y Databricks(Amazon Web Services, s-f.).
- **Azure Machine Learning:** Ofrece un entorno para el desarrollo y despliegue con herramientas para la preparación de datos, entrenamiento de modelos, gestión de experimentos y automatización de flujos de trabajo de ML, ademas de ofrecer modelos de Hugging Face y OpenAI. Sus fuentes de datos principales incluyen Azure Blob Storage, Azure File Share y Azure Data Lake, (microsoft, s-f.).
- **OVHcloud (AI Notebook y AI Training):** Ofrece sus servicios .ªI Notebookz .ªI Training", con infraestructura gestionada, incluyendo acceso a GPUs, para el desarrollo y entrenamiento de modelos. No tiene un entorno tan potente como el resto de clouds, pero ofrece una capa gratuita generosa y localizada en Europa, (ovhcloud, s-f.).

2.3.2.2. Otras herramientas

Además de las plataformas comentadas, existen herramientas especializadas de gran utilidad y con capas gratuitas que ya hemos comentado en [la sección de cloud 2.2.2], pero merece la pena concretar:

- **Kaggle Kernels/Notebooks:** Es ideal para análisis exploratorio de datos, competiciones de ML y entrenamiento de modelos intensivos (Kaggle, s-f.a) por el acceso a sus GPUs.
- **Google Colab:** Alternativa excelente para el entrenamiento de modelos complejos, funcionando como un entorno de desarrollo gratuito con Jupyter Notebooks que ofrece GPUs y TPUs, especialmente popular por su facilidad de uso y su integración con el ecosistema de Google y GCP (Google, s-f.b).
- **Hugging Face Spaces:** Plataforma dedicada al despliegue, compartición y descubrimiento de modelos de ML ya creados, muy útil para pruebas de concepto, permitir la prueba de modelos o utilizar los que la comunidad ha creado (huggingface, s-f.).

También vale la pena destacar los grandes modelos generativos, los cuales presentan una capacidad de adaptación enorme, y aunque no están diseñados para la ingesta de grandes volúmenes de datos, el aumento de las ventanas de contexto (numero de tokens a los que el modelo "presta atención" (Vaswani et al., 2017)) ha echo que tratar de utilizarlos para el tratamiento de

datos tabulares no sensibles sea un caso que vale la pena considerar, siempre teniendo en cuenta estas consideraciones éticas y de privacidad.

Teniendo en cuenta la ventana de contexto de los modelos (LLM Stats, s.f.) y que un token son aproximadamente cuatro caracteres (Microsoft, s.f.b) se ha elaborado una tabla con la cantidad de datos tabulares que los principales modelos de LLM podrían asumir, considerando dos tipos de datos tabulares,:

- Uno pequeño, con un par de columnas en forma de ID y etiqueta (15 tokens por fila)
- Uno mediano, con varias columnas o una columna de texto con aproximadamente 65 palabras (90 tokens por fila)

Tabla 2.1: tabla con ventanas de contexto de diferentes LLM

Modelo	**Tokens ventana**	**Filas, 15 tok/fila**	**Filas, 90 tok/fila**
Llama 4 Scout	10.000.000	666.666	111.111
Gemini 1.5 Pro	2.097.152	139.810	23.301
GPT-4.1	1.047.576	69.838	11.639
Nova Pro	300.000	20.000	3.333
Grok-4	256.000	17.066	2.844
Claude 3.5 Sonnet	200.000	13.333	2.222
DeepSeek-V3 / R1	128.000	8.533	1.422

Fuente: Elaboración propia con datos de LLM Stats (s.f.).

Con estos datos se puede observar que para ciertos tipos de datos tabulares, hay modelos que están empezando a poder tratar un numero considerado de filas, y ya hay estudios que lo están revisando (Fang et al., 2024).

2.3.2.3. AutoML

Como colofón a las estrategias de construcción de modelos, el Aprendizaje Automático Automatizado (AutoML) representa un conjunto de técnicas y herramientas que simplifican y aceleran el proceso de desarrollo del ML, reduciendo la complejidad de la implementación. Lo consiguen automatizando una, varias o todas las siguientes tareas:

- **Preprocesamiento de datos:** Limpieza y preparación de datos antes de entrenar el modelo.
- **Ingeniería de características:** Creación y selección automática de "features" relevantes.

- **Selección de modelo:** Evaluación automática de varios modelos para escoger el más adecuado.
- **Ajuste de hiperparámetros:** Optimización automática de los parámetros del modelo.
- **Evaluación del modelo:** Cálculo de métricas y generación de informes de desempeño.
- **Ensamblado y "stacking":** Combinación de múltiples modelos para mejorar la precisión.
- **Reportes y documentación:** Seguimiento automático de todo el proceso de ML.
- **Despliegue y producción:** Implementación del modelo entrenado en un entorno productivo.

Aunque esta tecnología es muy útil para tratar grandes volúmenes de datos de forma rápida, también es muy costosa en cuanto a computación. por lo que no todas las nubes ofrecen sus herramientas propias, de todas formas las listaremos brevemente (Böer, 2023):

- **Google Vertex AI AutoML**
- **AWS SageMaker Autopilot**
- **Microsoft Azure AutoML**

Otra opción seria utilizar herramientas especializadas como "AutoKeras" (Jin et al., 2023) o "AutoSklearn" (Feurer et al., 2020) en la infraestructura que la nube proporciona. Estas librerías implementan soluciones interesantes de AutoML para clasificación.

2.3.3. IA, normativa y ética

También hay que tener en cuenta la nueva normativa que la Unión Europea ha establecido con el Reglamento de Inteligencia Artificial (Union Europea, 2024), así como el marco en el que este proyecto se engloba. Este reglamento se ha usado de base para el uso de IA en este proyecto, y aunque la mayor parte de las consideraciones éticas desarrolladas en los últimos años, se centran en modelos de procesamiento del lenguaje natural, (como los estudios conducidos por DeepMind) (Gabriel et al., 2024). En Este aspecto, el presente trabajo se compromete con la "Declaración Responsable sobre Autoría y Uso Ético de Herramientas de Inteligencia Artificial (IA) de la Universidad Complutense de Madrid" adjunta en el repositorio de este mismo proyecto.

Pero además se compromete a seguir los siete estándares definidos por el reglamento de la Unión Europea en principios éticos de de la Unión Europea: Sección 3.1.2.2.

Para terminar, anteriormente se ha hablado sobre la obtención de "insights" automáticos: Es sabido que la IA acelera la identificación de patrones, pero para poder efectuar una interpretación estratégica se necesita el contexto de política pública y la validación humana experta. El principio HITL "humano en el bucle" (Mosqueira-Rey et al., 2023) es un requisito metodológico y ético para convertir patrones en decisiones, por lo que sera un requisito que seguiremos en este proyecto.

2.3.4. Trabajos anteriores y relacionados

Aunque ya se han nombrado varios anteriormente y en las [Definiciones 12], querría añadir algunas referencias mas, para empezar, el libro (Fowdur, 2021), cuyo capitulo 3 hace un profundo análisis sobre diferentes modelos de inteligencia artificial. También el libro (Brownlee, 2016) y todo el trabajo del autor en general, el cual ha sido de gran ayuda para afianzar los conceptos de esta sección.

También nombrar un par de trabajos muy interesantes sobre AutoML, que al abordar este tema, tratan todas las fases del ciclo completo de los algoritmos de aprendizaje automático. (Salehin et al., 2024) desglosa cada etapa del flujo de trabajo y explica cómo AutoML automatiza cada una. Destaca la búsqueda de arquitecturas neuronales (NAS), que automatiza el proceso de diseño de la arquitectura de una red neuronal, como motor del crecimiento del AutoML. Por otro lado, (Barbudo et al., 2023) se centra mas en el campo de investigación de AutoML, con una taxonomía que clasifica todo el campo en fases, tareas y técnicas.

Por último, mencionar trabajos académicos de compañeros con ideas similares a la tratada en este estudio y que han ayudado a guiar el mismo:

- **"Análisis de datos de la ciudad de Madrid"** (Santos, 2023): Donde se realiza un análisis del impacto de la pandemia en la movilidad peatonal de Madrid con datos públicos del ayuntamiento (2019-2021). El autor utiliza el modelo ARIMA (Autoregresivo) para las series temporales, o K-means para clusterización de datos.
- **"Madrid, isla de calor"** (Meneses Vicente y García Ruíz, 2023): También analizando datos del Ayuntamiento de Madrid, esta vez del clima y la contaminación (2021). Los autores utilizaron clusterización, mayormente K-Means, para comprobar hipotesis sobre el clima en la ciudad.

Capítulo 3

Materiales y métodos

En este capítulo describimos el proceso que se ha seguido en la realización del trabajo, las distintas tecnologías, lenguajes de programación, conjuntos de datos y herramientas utilizados e incluso algunos de los valorados en el "capítulo 2: Estado de la cuestión", pero descartados, así como los motivos para ello. También se han definido los métodos de desarrollo y modelo de trabajo.

3.1. Métodos

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

- **Búsqueda y almacenamiento de datos:** Se analizaron los datos públicos disponibles, se recopilaron metadatos sobre estos y se seleccionó un método de descarga y almacenamiento en la nube, ya fuera con un "script" que moviera los datos a la nube o con almacenamiento local en el caso de volúmenes de datos suficientemente pequeños.

- **Tratamiento básico de los datos:** Una vez almacenados, se procedió a su estudio y análisis para comprobar la consistencia y validez de los mismos, intentando comprobar duplicados, datos nulos, vacíos o valores atípicos que pudieran no tener sentido e interferir con el rendimiento de los modelos. Las modificaciones realizadas a los datos también se documentaron para poder ser replicadas, y los datos se almacenaron una vez procesados en el mismo sistema de almacenamiento que los

originales.

- **Estudio con modelos de IA en diferentes nubes:** Una vez los datos estuvieron tratados, se pasó a la fase de extracción de conclusiones. Se establecieron unas hipótesis a confirmar o se dejó a los modelos extraer sus propios clústeres de datos. En esta fase también se seleccionó el modelo más adecuado dependiendo del tipo de datos y de las conclusiones a las que se quiso llegar.

- **Comparación y estudio de resultados:** Tras procesar los datos y analizarlos, se pasó a la revisión de las hipótesis y conclusiones a las que los modelos llegaron, para comprobar cuáles fueron los resultados de valor que se pudieron obtener de estos datos o si las hipótesis establecidas se cumplieron.

- **Revisión, automatización y generalización del proceso:** Una vez finalizado el proceso, se realizó un estudio de las partes del mismo que se podían automatizar para siguientes iteraciones. En este paso se analizaron todos los anteriores en búsqueda de mejoras, y también se documentó el número de recursos utilizados en la nube, para tener controlados los costes del proceso y buscar puntos de optimización.

3.1.1. Utilización de la solución

En este capitulo vamos a ver los detalles de la solución para distintos conjuntos de datos.

Para la utilización de la solución, hemos seguido las buenas practicas del tratamiento de los datos, así como las buenas practicas de MLOps y las buenas practicas recomendadas por cada una de las nubes utilizadas. Estos conjuntos de buenas practicas se detallaran mas adelante en este mismo capitulo.

[TODO]

3.1.2. Métodos utilizados: Datos

Para tratar los datos a través del método principal de desarrollo, hemos seguido principios de gobernanza de datos:

3.1.2.1. Aplicación de la gobernanza de datos como método

Como ya hemos definido en la Sección 2.1.4: Estudios de datos y gobernanza, La gobernanza de datos en este proyecto se implementó adoptando este modelo de tres capas (Estratégica, Táctica y Operativa o de entrega). El objetivo fue garantizar que el proceso de análisis, adquisición y almacenamiento de datos públicos y la obtención de valor mediante IA, se ha realizado de forma ética, segura, y en pleno cumplimiento del marco regulatorio actual. En cuanto a las tres capas, aunque cobran mayor importancia en proyectos grandes con múltiples equipos y no en trabajos de una sola persona, se han adaptado a este trabajo por la estructura metodológica que proponen y la utilidad en cuanto a la gestión de datos y procesos:

1. Capa Estratégica: Liderazgo y Visión

En esta capa se definieron los objetivos generales y principios de la gobernanza de datos en el proyecto.

- **Visión:** Convertir los datos abiertos en generadores de conocimiento mediante técnicas de inteligencia artificial y tecnologías cloud.
- **Seguridad y soberanía:** Aunque se utilizan servicios de nubes públicas por su acceso gratuito, capas de seguridad y capacidades en IA, se configuraron para operar dentro de la UE.
- **Transparencia y reproducibilidad:** Todo el proceso (origen de datos, transformaciones, código, y resultados de modelos) se documentó en esta mismo memoria para garantizar la transparencia y la posibilidad de auditar o reproducir el análisis, de acuerdo a los principios de la Unión Europea: Sección 3.1.2.2.

2. Capa Táctica: Capacidades de Implementación y Marco Normativo

Esta capa detalla cómo se implementa la estrategia a través de políticas, procesos, directrices, etc:

- **Uso del dato:** Se priorizaron datos públicos abiertos de administraciones españolas y de la Unión Europea, prestando especial atención a las licencias para asegurar la legalidad de su reutilización y evitando el uso de datos sensibles. Respecto a datos sensibles se aplicó un principio de precaución: cualquier conjunto de datos con riesgo de contener información sensible fue filtrado, descartado o anonimizado.

- **Gestión de accesos y credenciales:** Como único usuario, se gestionan las credenciales de acceso a los servicios cloud con el máximo nivel de seguridad, evitando su filtración a repositorios públicos o terceros.
- **Competencias y coordinación:** Todas las funciones fueron asumidas por un único investigador, esto centralizó la toma de decisiones y facilitó el cumplimiento normativo y la trazabilidad de todo el proceso. De todas formas se utilizan herramientas como "Git" o "Trello" para auto-organizarse.
- **Selección de proveedores y servicios:** Para la selección de plataformas cloud se evaluó la capacidad para proporcionar entornos de procesamiento seguro, y la localización de sus centros de datos para asegurar el cumplimiento normativo. En cuanto a la IA, también se revisaron sesgos en los datos de entrenamiento.

3. Capa Operativa: Infraestructura, Integración del Ciclo de Valor y Arquitectura

Esta última capa corresponde a la implementación práctica de la estrategia, la gestión diaria del ciclo de valor de los datos para integrarlo con la infraestructura técnica.

- **Infraestructura:** Se emplearon servicios en la nube principalmente para el almacenamiento, procesamiento y análisis de los datos. Los entornos se han configurado con mecanismos de seguridad estándar. Aunque una de las ideas de este proyecto es el tratamiento de datos con el menor número de recursos posibles, también se usaron dispositivos on-premise (computador personal) para la ingesta de datos y posterior almacenamiento en cloud cuando esto facilitó el proceso, aunque se priorizaron tecnologías en la nube.
- **Arquitectura de datos y ciclo de valor:** Se diseñó un flujo simple de trabajo centrado en cloud y basado en la ingesta de datos abiertos de fuentes oficiales que cubrió todo el ciclo de vida del dato:
  - **Adquisición:** Se descargaron conjuntos de datos abiertos, registrando metadatos sobre origen, licencia, calidad, formato y condiciones de uso.
  - **Almacenamiento y gestión:** Se organizaron en buckets con estructura clara siguiendo la arquitectura de medalla [Definición 11]. Este enfoque facilitó la exploración, el modelado y la generación de valor con herramientas nativas como BigQuery.
  - **Procesamiento y transformación:** Se realizó la limpieza, ano-

nimización y feature engineering en entornos gestionados como Dataflow o Vertex AI Workbench. Se mantuvo un registro de los experimentos realizados (hipótesis, parámetros, versiones de modelos) para asegurar reproducibilidad y transparencia.

- **Uso/compartición:** Se utilizaron diversas técnicas de IA para la identificación de patrones en los datos y se han publicado los resultados bajo licencias abiertas, priorizando la transparencia.

- **Optimización y sostenibilidad:** Se monitorizó el uso de recursos en las diferentes nubes para mantener el proyecto dentro del coste cero, también se optimizaron las configuraciones de los servicios para asegurar la eficiencia tanto económica como ecológica del proyecto.

3.1.2.2. Principios éticos

Destacar también que en el proyecto, se siguieron rigurosamente los siete principios éticos de la IA definidos por el reglamento de la Unión Europea (Union Europea, 2024):

- **Acción y supervisión humanas:** En este trabajo se supervisó tanto cada etapa del desarrollo, como los resultados. Se descartaron aquellos que no cumplieran con los criterios éticos establecidos.

- **Solidez técnica y seguridad:** Se veló por la mayor excelencia técnica y se definieron principios de seguridad tanto para los datos como para los modelos.

- **Gestión de la privacidad y de los datos:** Se ha abordado cumpliendo estrictamente con los principios del RGPD como se indica en la sección de datos.

- **Transparencia:** Se documentaron los procesos, algoritmos utilizados y decisiones tomadas durante el desarrollo para asegurar que el proceso fuera comprensible y reproducible. También se documentaron los casos en los que se recurrió a la Inteligencia Artificial, indicando claramente las razones para esto.

- **Diversidad, no discriminación y equidad:** Se comprobaron los sesgos de los datos y algoritmos utilizados, decisiones tomadas y resultados obtenidos, con el objetivo de identificar y mitigar potenciales discriminaciones o sesgos.

- **Bienestar social y ambiental:** El motivo último de este trabajo es el procesamiento de datos públicos para mejorar el valor de los mismos y, mediante su aplicación, el bienestar social en general.

- **Rendición de cuentas:** Se mantuvo un registro detallado de las decisiones de diseño, preprocesamiento de datos y selección de modelos,para permitir una auditoría clara y una rendición de cuentas transparente.

3.1.3. Métodos utilizados: Aprendizaje automático

3.1.4. Métodos utilizados: Aprendizaje automático

3.2. Materiales

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

3.2.1. Conjuntos de Datos

Los conjuntos de datos utilizados tienen sus propios metadatos asociados en el Github de este proyecto, pero en esta seccion se listaran brevemente los datos utilizados.

3.2.2. Materiales para el desarollo

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. [TODO] uso en ia]

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.

bash Script

La comunicación con las nubes de AWS y GCP, se ha realizado principalmente con la ejecución de scripts bash en su GUI con los comandos que estas mismas facilitan, tambien se han realizado scripts especificos apra ciertas tareas.

3.2.3. 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 entre muchas otras funciones.

Utilizamos Visual Studio Code como entorno de desarrollo software por la gran comunidad que tiene detrás, la cual mantiene extensiones y tutoriales al día, lo que nos facilita mucho la programación y la integración con otras aplicaciones. También destacar su intérprete, para probar pequeños fragmentos de código, lo cual nos ha ahorrado tiempo en depuración de errores.

GitHub

GitHub es una plataforma para alojar proyectos utilizando el sistema de control de versiones Git, que se utiliza principalmente para la creación, almacenamiento y control de código fuente.

[TODO]

TeXstudio y MiKTeX

TeXstudio es un editor de LaTeX de código abierto y multiplataforma con una interfaz amigable, es un IDE que proporciona un soporte moderno de escritura, como la corrección ortográfica interactiva, plegado de código y resaltado de sintaxis, por lo que se ha 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 compilación y estructuración de este documento.

3.2.4. Herramientas descartadas

[TODO]

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

Añadir al estudio un coste ecológico de las tecnologías.

Contactar con responsables de algunos de los estudios citados y conducir una encuesta a los usuarios para comprender la utilidad de los datos.

Contactar con las nubes privadas europeas y españolas, ya que aunque es normal que en su mayoría no ofrezcan planes gratuitos para evitar su abuso, es posible que contactando como entidad investigadora dieran acceso a las cloud para poder comprobar su utilidad en este estudio. También con las iniciativas como BDTI de la Unión Europea (European Commission, s.f.b) para probar proyectos como este.

Construir interfaz de usuario para visualizar los datos de una manera mas optima, ya sea con Firebase, o una instancia de Superset.

Chapter 6

Conclusions and Future Work

[TODO]

Apéndice A

Definiciones y acrónimos

"Saber dónde encontrar la información y cómo usarla. Ese es el secreto del éxito"
— Albert Einstein

-

Dedicaremos este apéndice a la explicación de conceptos en más extensión, ya sea conceptos más generales o la explicación del significado de los acrónimos de esta memoria.

A.1. Definiciones

Separaremos este apartado, de la misma manera que se ha hecho en otras partes de la memoria, en los tres grandes elementos que conforman esta memoria, definiciones referentes a datos, a nubes y a Inteligencia artificial. En cada uno de ellos enumeraremos las definiciones que se han ido incrustando en la memoria de la manera: [Definición 1].

La enumeración de las definiciones es independiente del apartado en el que se encuentra

A.1.1. Referente a datos

1. 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.

2. El **open government** o gobierno abierto es una forma de comunicación abierta, permanente y bidireccional entre la administración y los ciudadanos, basada en la transparencia por parte de la administración y la participación y colaboración con la sociedad civil y las empresas. Teniendo como punto clave el movimiento open data o datos abiertos, esta estructura y formatos abiertos permiten que los datos puedan reutilizarse proporcionando nuevos servicios a ciudadanos y empresas. En Europa sus orígenes se sitúan en la Directiva 2003/98/CE del Parlamento y del Consejo Europeos sobre el acceso y la reutilización de la información del sector público. (Ferrer-Sapena A., 2011).

3. El **Valor Público** se puede definir de muchas maneras y depende de la perspectiva de muchos autores:

   - Para **Mark Moore**, su creador, consiste en conocer y satisfacer los deseos de la gente, un valor que lo público debe crear de forma análoga a como el sector privado crea valor económico. (Moore, 1995)
   - **Bozeman** lo define desde una perspectiva ciudadana como el consenso sobre los derechos y obligaciones de los ciudadanos, así como

los principios sobre los que debe basarse el gobierno (BOZEMAN, 2007). A menudo se refiere a "valores públicos", en plural, para destacar su diversidad, tema que seria llevado mas en profundidad por **Talbot**, que sugiere que a veces estos son contradictorios entre sí, reflejando la combinación de las diversas y conflictivas preferencias del público (Talbot, 2011).

- **Benington** lo vincula directamente con la "esfera pública", argumentando que el valor público no es solo lo que el público valora individualmente, sino también aquello que agrega valor a este espacio colectivo. (Benington, 2009)
- Finalmente, **Timo Meynhardt** lo conceptualiza como un fenómeno relacional que surge de las percepciones. El valor público se crea en la relación entre el individuo y la sociedad, y depende de cómo las acciones de las organizaciones públicas impactan en la satisfacción de las necesidades básicas de las personas: morales, sociales, utilitarias y hedonistas (Meynhardt, 2009).

En un intento de resumirlo, el valor publico surge de las evaluaciones y percepciones que los individuos y colectivos realizan sobre cómo las acciones, servicios o políticas de las organizaciones públicas (y otras entidades) impactan en la satisfacción de sus diversas necesidades básicas dentro de un marco relacional que involucra a la esfera pública. Valor creado para y por la sociedad.

4. Los **Datos Abiertos** se refieren a conjuntos de datos digitales que se publican bajo una filosofía de apertura, garantizando y facilitando el libre acceso, uso, modificación, reutilización y redistribución por parte de cualquier persona o entidad, en cualquier momento, lugar y con cualquier finalidad. Una parte especifica y importante para este trabajo son los Datos Abiertos de Gobierno, aquellos datos que se originan, producen, encargan o publican los gobiernos u organismos públicos en el ejercicio de sus funciones. Estos datos buscan, como fin último, fomentar la transparencia, la creación de valor público, la colaboración intersectorial y la resolución de problemas.

   La materialización de esta filosofía de apertura se concreta en requisitos técnicos y jurídicos específicos, cuya interpretación puede variar ligeramente entre las entidades que los definen. Desde el Grupo de Trabajo sobre Datos Abiertos "Open Knowledge Foundation" (OKF), "El conocimiento está abierto si alguien tiene la libertad de acceder a él, usarlo, modificarlo y compartirlo, sujeto, como máximo, a medidas que preserven su procedencia y su apertura" (Open Knowledge Foundation, 2025). El Portal Europeo de Datos y el "Open Data Charter",

por su parte, enfatizan las condiciones de acceso y las libertades de uso, incluyendo la gratuidad y la ausencia de limitaciones, detallando la necesidad de características técnicas y jurídicas para que los datos sean libremente reutilizables y redistribuibles (Portal Europeo de Datos, 2025), (Open Data Charter, s.f.). Todo esto subraya la complejidad y la multifuncionalidad de los Datos abiertos como catalizador para la innovación y el desarrollo socioeconómico, con implicaciones legales y técnicas que deben ser gestionadas cuidadosamente para maximizar su potencial.

5. Las **Tres olas del "Open Data"** representan las diferentes etapas evolutivas por las que ha transitado el movimiento de apertura de datos. La **Primera Ola** (1990s-2000s) se fundamentó principalmente en Estados Unidos, dirigido a periodistas, abogados y activistas que solicitaban datos específicos bajo el modelo de "derecho a saber", enfrentando riesgos de secretismo gubernamental y requiriendo auditores de información. La **Segunda Ola** (2000s-2010s) evolucionó hacia la apertura por defecto con alcance internacional, expandiendo su audiencia a agencias gubernamentales, empresas tecnológicas y organizaciones comunitarias,pero generando desafíos de privacidad que impulsaron la creación de portales de datos abiertos y responsables. La **Tercera Ola** (2010s-presente) representa la madurez del movimiento con colaboración intersectorial y flujos transfronterizos, dirigiéndose a ONGs, instituciones académicas, pequeñas empresas y gobiernos, estableciendo marcos de responsabilidad en materia de datos. Esta evolución se visualiza en la tabla comparativa [Tabla A.1].

6. La **Teoría de la Ventana** (Matheus y Janssen, 2020) es un marco conceptual que analiza la transparencia generada por los Datos Abiertos de Gobierno, concibiéndola como una "ventana" que el gobierno abre para que el público vea su funcionamiento interno. Postulando que la transparencia es una construcción diversa y continua, cuyo objetivo principal es facilitar la transferencia de información entre el gobierno y sus públicos. Su materialización está influenciada por factores que la facilitan o la impiden, clasificados en:

   - Características de los datos.
   - Características del sistema.
   - Características de la organización.
   - Características del uso individual.

   Esto genera consecuencias (intencionadas o no) como la rendición de

Tabla A.1: Resumen de las principales características de las denominadas olas de datos abiertos.

	Primera ola	**Segunda ola**	**Tercera ola emergente**
Concepto	Libertad de información	Datos públicos abiertos	Reutilización de datos públicos y privados
Propuesta de valor	Transparencia	Transparencia y resolución de problemas	▪ Elaboración de políticas basadas en pruebas ▪ Innovación e iniciativa empresarial
Método	Datos a petición (derecho a saber)	Abierto por defecto (derecho a compartir)	Publicar con propósito
Enfoque	Enfoque de impulso	Enfoque de atracción	Asociaciones (colaboraciones con datos)
Énfasis geográfico	Nacional	Internacional y nacional	▪ Subnacional y local ▪ Flujo transfronterizo de datos con fines específicos
Audiencia / Demanda	▪ Periodistas ▪ Abogados y activistas ▪ Tecnólogos cívicos y "data geeks"	▪ Agencias gubernamentales ▪ Empresas, start-up tecnológicas ▪ Organizaciones comunitarias	▪ ONG, derechos humanos y justicia social ▪ Instituciones académicas ▪ Pequeñas empresas y start-ups ▪ Gobierno
Riesgos y políticas	Secretismos y ofuscación	Privacidad – Efecto mosaico, información demográfica identificable (DII)	Marco de responsabilidad y derechos en materia de datos
Respuestas institucionales	Auditores de información	▪ Responsable de datos ▪ Portales de datos abiertos	▪ Director de datos ▪ Intermediarios

Fuente: Traducción propia de Verhulst et al. (2020).

cuentas, la participación cívica, el aumento de la eficiencia o la afectación a la privacidad.

7. Un **"Datathon"** (Anslow et al., 2016) es un evento colaborativo o competitivo intensivo centrado en datos, y derivado de los términos "data" y "maratón", donde equipos de expertos y otros individuos se reúnen para analizar grandes volúmenes de estos datos con el fin de encontrar soluciones innovadoras a problemas específicos. Esto va desde desarrollar aplicaciones para sacar partido a estos datos, hasta la optimización de procesos o la creación de modelos predictivos, las posibilidades son amplias.

8. **"Data Governance "** (Gobierno del Dato) se define como un marco integral de políticas, estrategias, estándares, roles y procesos que rigen toda la gestión del ciclo de vida completo de los datos (desde su creación y recopilación hasta su almacenamiento, procesamiento, uso, compartición, archivado y ,finalmente, eliminación) con el objetivo de garantizar su calidad, integridad, seguridad, accesibilidad, interoperabilidad y confiabilidad (Herrera Capriz (2024), pg 241). La OECD propone que un modelo de "Data Governance" exitoso integra tres capas: una capa estratégica (liderazgo y visión); una capa táctica (capacidades de implementación y normativa: comités, formación, directrices, etc.); y una capa de entrega o operativa (de infraestructura y arquitectura: estándares, catálogos de datos, ciclo de valor, etc.) (OECD, 2019).

   También vale la pena mencionar que hay mas metodologías que hablan de las buenas practicas del Gobierno del dato, como es el caso de DAMA (de Datos Abiertos del Gobierno de España, 2021), que propone un marco de referencia (DAMA-DMBOK) donde el gobierno de datos coordina e integra otras diez áreas de conocimiento: la arquitectura de datos, modelado y diseño de datos, seguridad de datos, calidad de datos, gestión de metadatos y almacenamiento de datos y operaciones.

Figura A.1: Capas de gobierno del dato

Fuente: (OECD, 2019) traducida.

A.1.2. Referente a Cloud

9. La **"Computación distribuida"** / **"Nube distribuida"** es un paradigma que aprovecha la capacidad de cálculo de miles o millones de dispositivos conectados a internet, creando una red de computación masiva, paralela y descentralizada. Los usuarios solo necesitan instalar un cliente de software que, cuando su dispositivo no está en uso, se encarga de descargar un pequeño fragmento de datos, procesarlo y devolver el resultado a un servidor central que coordina todas las tareas y ensambla los resultados finales. Este es un enfoque especialmente adecuado para problemas altamente paralelizables. Uno de los primeros productos en llevar la idea a cabo fue BOINC (Anderson, 2004) para ayudar al cómputo de proyectos científicos. Actualmente, nubes como AWS o GCP se han apropiado del termino Nube distribuida, teniendo centros de datos en diferentes localizaciones para distribuir su trabajo a localizaciones mas cercanas o llevar la propia infraestructura y servicios de la nube fuera de sus data centres (Google Cloud, s.f.). Esta idea evolucionó hasta la llamada "Computación Voluntaria" (Nov et al., 2010), donde dispositivos personales de voluntarios se usaban para este fin (ahora BOINC, cuyo software aún está disponible seria considerado computación voluntaria), y también se puede considerar que este paradigma se usa de forma malintencionada con varios casos famosos de ataques DDoS distribuidos (radwere, s.f.), (Danysoft, s.f.)

que usan dispositivos infectados para conseguir su propósito.

También han surgido modelos con estas ideas paradigmas como el "Dew computing" (Ray, 2018) donde los dispositivos personales se usan como almacenamiento, o plataformas de "fog computing" (que añade dispositivos intermedios en el calculo centralizado) o "edge computing" (Que ejecuta directamente en dispositivos finales) como SONM (SONM, s.f.) o otros proyectos (Uriarte y De Nicola, 2018) que surgieron con el auge de las "Blockchains" y las usaban para poner en contacto usuarios que quisieran proporcionar potencia de cálculo con quienes querían usarla. En la actualidad, hay nubes que siguen activas como (Golem Network, s.f.), (Akash, s.f.) ó (Render Network, s.f.) y que permiten comprar y vender potencia de cómputo.

En resumen, todo lo englobado a la computación distribuida que hemos comentado se puede ver en esta tabla:

Modelo	**Descripción**	**Ejemplo destacado**
Nube distribuida	Infraestructura descentralizada, extendiendo servicios hacia el edge o centros locales.	Gestión centralizada con despliegue en edge/localización (AWS CloudFront, Google Cloud CDN).
Computación voluntaria	Uso de recursos ociosos de dispositivos personales para cómputo distribuido voluntario.	BOINC, SETI@home, HTCondor, Techila Grid.
"Dew Computing"	Combina almacenamiento local y en la nube, sincronizando datos y permitiendo disponibilidad offline.	Dropbox.
"Fog Computing"	Procesamiento intermedio entre dispositivos y la nube, reduciendo latencia y uso de ancho de banda.	Aplicaciones IoT y casos de baja latencia / Vehículos autónomos.
"Edge Computing"	Procesamiento en dispositivos finales, priorizando latencia y seguridad.	Procesamiento de vídeo en tiempo real, reconocimiento facial en dispositivos móviles.
"BotNets"	Redes de dispositivos comprometidos que realizan tareas maliciosas de forma distribuida, usando el mismo principio de cómputo compartido pero con fines ilícitos.	Ataques DDoS (Mydoom), spam masivo, minería de criptomonedas no autorizada.

Tabla A.2: Comparación de modelos relacionados con la computación distribuida.

10. La **clasificación de la Computación en Nube** típicamente se puede dividir según dos dimensiones principales (Huawei Technologies Co., Ltd., 2023):

    Por modelo de operación:

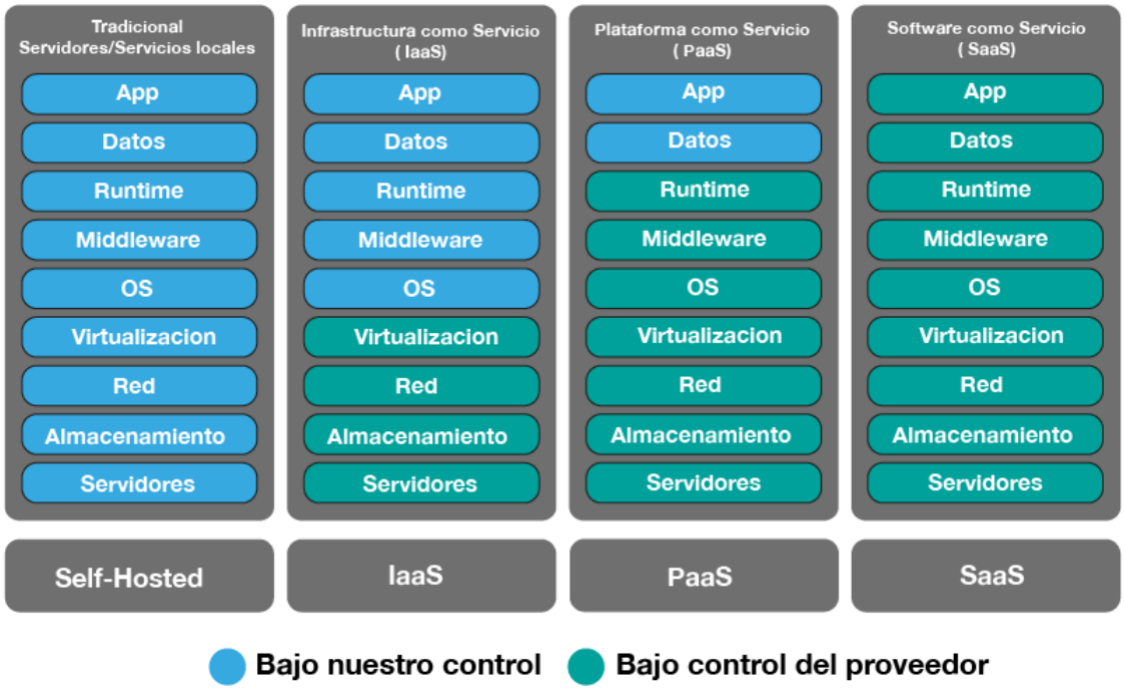
    - **Nube Pública:** Infraestructura compartida y gestionada por proveedores externos (ej. AWS, Azure, Google Cloud), la cual ofrece acceso universal mediante pago por uso.
    - **Nube Privada:** Infraestructura exclusiva para una organización, gestionada interna o externamente, y que tiene mayor control y seguridad.
    - **Nube Comunitaria:** Compartida por varias organizaciones con intereses comunes (ej. instituciones académicas), y que presenta un equilibrio entre control y costes.
    - **Nube Híbrida:** Combina nubes públicas y privadas, permitiendo mover cargas de trabajo según necesidades de seguridad, coste o escalabilidad.
    - **Nube Industrial:** Especializada en sectores específicos (sanidad, automoción), con componentes optimizados para casos de uso particulares.

    Por modelo de servicio:

    - **IaaS (Infraestructura como Servicio):** Proporciona recursos fundamentales (máquinas virtuales, almacenamiento, redes). Ej: Amazon EC2, Google Compute Engine.
    - **PaaS (Plataforma como Servicio):** Ofrece entornos de desarrollo y ejecución para aplicaciones. Ej: Google App Engine, Microsoft Azure App Services.
    - **SaaS (Software como Servicio):** Software completo gestionado por el proveedor y accesible vía web. Ej: Gmail, Salesforce, Office 365.

    En la Figura A.2 se pueden observar una tabla que simplifica todos estos modelos de cloud por servicio.

Figura A.2: Distintos tipos de cloud



Fuente: (Albert Barron, 2015).

11. La **"Arquitectura en medalla"** es un patrón de diseño utilizado para gestionar el ciclo de vida de los datos en plataformas de datos, esta organiza y transforma datos en capas sucesivas: Bronce, Plata y Oro, y cada capa representa un nivel creciente de calidad y utilidad de los datos. Permite estructurar los datos para facilitar su ingestión, limpieza, enriquecimiento, análisis y consumo empresarial (Microsoft, s.f.c). Aunque lo hemos definido este nombre, quizás al lector le sea más familiar otra forma de procesar datos en capas (Bobrov, 2025), ya que la idea no es nueva en la ingeniería de datos y muchas otras metodologías introducen conceptos similares:

    - **Arquitectura Clásica de Data Warehouse (Inmon):**
        - Staging Area (Etapa de Carga) ≈ Capa Bronze
        - CIF (Corporate Information Warehouse) ≈ Capa Silver
        - Data Marts (Almacenamiento de Industria) ≈ Capa Gold
    - **"Data Vault":**
        - Staging Layer para procesamiento de flujos crudos ≈ Capa

Bronze

- Raw Vault (donde se almacenan los datos originales) ≈ Capa Silver
- Business Vault o Data Marts (transforman datos en métricas) ≈ Capa Gold

- **Data Mesh (Concepto de Propiedad de Datos Distribuida):**
  - Raw Domain Data ≈ Capa Bronze
  - Aggregated Domain Data ≈ Capa Silver
  - Consumer-Oriented Products ≈ Capa Gold
- **Patrón "Write-Audit-Publish":**
  - Write (recolección de datos) ≈ Capa Bronze
  - Audit (limpieza y procesamiento de datos) ≈ Capa Silver
  - Publish (preparación para el uso) ≈ Capa Gold

A.1.3. Referente a Inteligencia Artificial

12. La **"Historia de la Inteligencia Artificial"** se remonta a hace casi un siglo, con la ideas de matemáticos como Alan Turing, que sentó las bases de la computación y propuso el primer test para evaluar inteligencia artificial (Cheok y Zhang, 2023). A partir de ese punto, no han parado de surgir avances en métodos matemáticos y algoritmos impulsando la búsqueda e implementación de modelos prácticos. Hitos como el desarrollo del perceptrón (Rosenblatt, 1958), un modelo esencial para las redes neuronales; las cadenas de Markov, usadas como modelos estadísticos (Hidden Markov Model (HMM)); o el desarrollo del "backpropagation" que sentó las bases para que el entrenamiento de redes neuronales profundas fuera efectivo (Werbos, 1974).

    Los avances, junto con el aumento de la computación siguiendo la ley de Moore, permitió el surgimiento y aumento de las redes neuronales recurrentes (Hopfield, 1984), así como su uso para aplicaciones de procesamiento del lenguaje natural (Bengio et al., 2003). Otro punto de inflexión lo puso Google con su articulo "Attention Is All You Need" (Vaswani et al., 2017), donde definía que lo único que se necesitaba para procesar secuencias de forma eficaz era un mecanismo de atención, eliminando la necesidad de recurrencia y convoluciones, lo que permitió entrenar modelos más potentes sobre conjuntos de datos masivos, sentando las bases técnicas para los Grandes Modelos de Lenguaje (LLMs) como BERT o la serie GPT (Transformadores Pre-entrenados Generativos).

    Tomando esta tecnología de los transformers, "OpenAI" lanzo diferentes versiones de su modelo GPT hasta culminar con una una interfaz conversacional basada en su modelo GPT-3,: ChatGPT. Esta interfaz se convirtió en una revolución, impulsando un interés público sin precedentes y consolidado a la IA como una herramienta indispensable. En los últimos años han seguido surgiendo modelos con tecnologías como la multimodalidad, donde los agentes combinan el procesamiento de texto escrito con otras fuentes como audio y visión en tiempo real, o el "Mixture of experts (MoE)" (Jiang et al., 2024) donde el modelo utiliza una combinación de múltiples redes neuronales (expertos), activando selectivamente un subconjunto de ellas para procesar cada entrada. Esta explosión en la presencia de la IA en la vida cotidiana, también ha impulsado, sobre todo en Europa, el desarrollo de marcos regulatorios para garantizar un uso responsable y transparente de estas tecnologías (Union Europea, 2024).

    Matiz: Si bien tecnologías de visión por computador o generación de multimedia son pilares fundamentales de la IA moderna y cruciales

para los LLMs multimodales, es necesario acotar el alcance de este trabajo. Por lo tanto, esta definicion se centra en los avances relacionados con el procesamiento de datos estructurados y texto, dejando fuera un análisis detallado del tratamiento específico de vídeo, audio e imagen, a pesar de reconocer su enorme relevancia e impacto.

Para facilitar la visualización de esta definición, resumiremos los hitos que consideramos importantes en orden cronológico:

Tabla A.3: **Linea de tiempo de la IA**

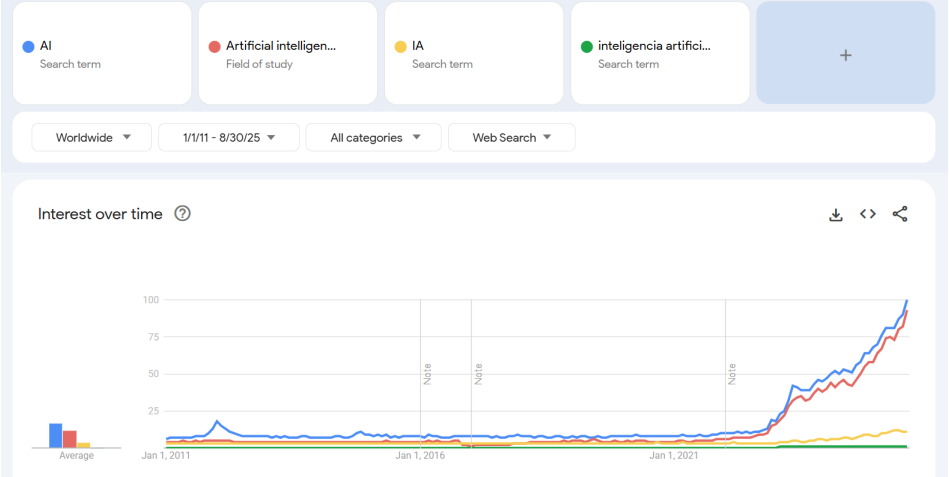
:

1936	**Turing machines** - Marco teórico para computación
1950	**Turing test** - Primera medida práctica de inteligencia máquina
1958	**Perceptrón** - Modelo base de las redes neuronales
1964	**ELIZA** - Primer chatbot conversacional
1966	**Hidden Markov Model** - Modelo estadístico temprano
1967	**K-means** - Algoritmo popular de clustering
1974	**Backpropagation** - Algoritmo fundamental para ANN multicapa
1982	**RNN** - Redes neuronales recurrentes
1997	**LSTM** - RNN con memoria a largo plazo
	Deep Blue vence a Kasparov
2003	**NPLM** Aprendizaje de .^{em}beddings"de palabras, base de LLM
2014	**GAN** - Generación de datos nuevos
2016	**AlphaGo** - Derroto al campeón del mundo
2017	**Transformers** - Arquitectura basada en atención
2018	**BERT**, **GPT-1** - Modelos de lenguaje grandes (LLM)
2020	**GPT-3** - aumento de las capacidades y parámetros
2021	**DALL-E** - Imágenes desde texto
2022	**ChatGPT** - Interfaz conversacional, aumenta el interés publico
2023	**GPT-4** - Modelo multimodal
	LLaMA Modelo Open-Source
2024	**GPT-4o** - Modelo multimodal con voz y visión en tiempo real
	EU AI Act - Regulación integral sobre IA
2025	**DeepSeek** - LLM de bajo costo
	IA Agéntica y especializada - agentes autónomos, tareas específicas

Fuente: Adaptación y actualización de (Cheok y Zhang, 2023), complementada con información de (Sapkota et al., 2025), (bit2brain, 2020), (Campbell et al., 2002) y las referencias mencionadas en el texto.

Como frase final, y espero que esperanzadora, incluir que a pesar de que esta tecnología ha pasado por algunas épocas de menor interés publico, los datos de Web parecen mostrar que esta vez la tecnología viene para quedarse [Figura A.3]. Ya que a pesar de que un ultimo estudio del MIT anuncia que el 95 % de los proyectos con IA reciben retorno cero (Challapally et al., 2025), diversos analistas sostienen que esto no implica una burbuja vacía, sino más bien una fase inicial en la que muchos pilotos fracasaran por problemas de integración o organizativos y no por fallos tecnológicos. Al igual que ocurrió con la burbuja de las dot.com, de la cual emergieron gigantes como Amazon y Google, se espera que también en IA sobrevivan los proyectos con verdadero valor (AI Explained, 2025).

Figura A.3: Tendencias de búsqueda de IA



Fuente: (Google Trends, 2025), *Destacar el pico de 2011 por el lanzamiento de Siri (Apple) y la tendencia desde 2022 por la salida de chatGPT.*

13. Los tipos de **Técnicas de ML según aprendizaje** es la forma de clasificar estos modelos dependiendo de la forma en la que los sistemas aprenden patrones, o del grado de supervisión que reciben. Se clasifican en cuatro tipos principales (Sarker, 2021):

    - **Aprendizaje Supervisado:** utiliza datos etiquetados para aprender una función que mapea entradas a salidas. Esta orientada a tareas concretas, y sus aplicaciones más comunes incluyen clasificación y regresión, como predecir la cantidad de ventas de un producto.

    - **Aprendizaje No Supervisado:** analiza datos no etiquetados,

buscando estructuras, patrones o relaciones ocultas. Es un enfoque orientado al descubrimiento y exploración, y se aplica típicamente en clusterización, reducción de dimensionalidad o detección de anomalías.

- **Aprendizaje Semi-Supervisado:** combina datos etiquetados y no etiquetados, aprovechando la abundancia de datos sin etiqueta para mejorar el rendimiento del modelo más allá de lo que se lograría solo con datos etiquetados. Es útil en contextos donde las etiquetas son costosas o escasas, como en detección de fraude o clasificación de texto.
- **Aprendizaje por Refuerzo:** se basa en la interacción de un agente con un entorno, aprendiendo mediante recompensas o penalizaciones para optimizar su comportamiento. Se aplica en tareas de automatización compleja y optimización, como robótica o conducción autónoma.

14. **MLOps** u operaciones de aprendizaje automático, son un conjunto de prácticas y herramientas que busca automatizar y gestionar el ciclo de vida completo de un proyecto de machine learning, desde el desarrollo del modelo hasta su despliegue, monitoreo y mantenimiento en producción, basándose en principios de DevOps pero adaptados a las particularidades de los sistemas de ML (Kreuzberger et al., 2023). Los nueve principios son los siguientes:

    a) **Automatización CI/CD:** Integración, entrega y despliegue continuos que automatizan el desarrollo, prueba y puesta en producción de los modelos, proporcionando información rápida sobre el éxito o fracaso de cada paso.

    b) **Orquestación del flujo de trabajo (workflow):** Coordinación y gestión de las tareas de un "pipeline" o proceso de ML, estableciendo el orden de ejecución y dependencias, típicamente mediante DAGs.

    c) **Reproducibilidad:** Capacidad fundamental de replicar el experimento de ML y obtener los mismos resultados.

    d) **Control de versiones:** Implica la gestión y el seguimiento riguroso de los cambios en los datos, modelos y código (indispensable para la trazabilidad y el cumplimiento normativo).

    e) **Colaboración:** Promueve el trabajo conjunto y una cultura comunicativa y abierta entre los diferentes roles del proyecto para minimizar aislamiento de datos e información.

f) **Entrenamiento y evaluación continuos:** Implica el retrenamiento periódico y automático de los modelos con nuevos datos, incluyendo una evaluación constante de su calidad para asegurar que sigan siendo relevantes, precisos y óptimos en cuanto a rendimiento.

g) **Seguimiento y registro de metadatos:** Documentación detallada de la información relevante de cada ejecución del flujo para asegurar trazabilidad: parámetros utilizados, métricas, código, tiempo de ejecución y los datos utilizados.

h) **Monitoreo continuo:** Observación periódica de los datos, modelos, código, infraestructura y rendimiento del modelo desplegado, con el objetivo de detectar anomalías, errores o bajadas en la calidad.

i) **Bucles de retroalimentación:** Integrar los hallazgos y las lecciones aprendidas del monitoreo y la evaluación de calidad en el proceso de desarrollo.

j) **Modelos de Machine learning**. Estas definiciones se basan en el curso de (Brownlee, s.f), el libro del mismo autor (Brownlee, 2024a) y en los libros (Fowdur, 2021) y (Sarker, 2021), así como en la asignatura "desarrollo de aplicaciones y servicios inteligentes" impartida por la Universidad Complutense de Madrid. Aunque para estas definiciones **no** nos centraremos en como efectuar la selección de hiperparámetros (parámetros externos al modelo que lo configuran), nombrar que el rendimiento y complejidad de estos modelos depende de la correcta elección de esta característica. También, por alcance, detallaremos solo los modelos que se han elegido para estudiar en este de entre la gran variedad de ellos que existe (Wikipedia, s.f), estos son los siguientes:

1) **Árboles de Decisión (Decision Trees):** Son un método de aprendizaje **supervisado** no paramétrico utilizado para **clasificación y regresión**. Consiguen su objetivo dividiendo recursivamente el conjunto de datos de entrenamiento en subconjuntos más pequeños, evaluando todas las posibles divisiones en cada característica de entrada para encontrar el punto de división óptimo que minimice la función de coste. El proceso continúa hasta alcanzar un criterio de detención, obteniendo los nodos terminales que contienen la predicción final (la clase más común o el promedio de valores).

2) **k-Nearest Neighbors (k-NN):** Es un algoritmo de apren-

dizaje **supervisado** de .ªprendizaje perezoso"(no ejecuta hasta que no se le pida una predicción), utilizado para **clasificación y regresión**. Consigue su objetivo para nuevos datos calculando la distancia entre este y todos los ejemplos del conjunto de entrenamiento. Luego, selecciona los k ejemplos más parecidos (vecinos). Para clasificación, la predicción es la clase más común entre estos k vecinos; para regresión, es el promedio de sus valores de salida.

3) **K-means:** Es un algoritmo de aprendizaje **no supervisado** utilizado para la **clusterización**. Consigue su objetivo al identificar k centroides en el conjunto de datos y luego asigna cada punto de datos al centroide más cercano durante varias iteraciones, estos centroides se actualizan continuamente moviéndose al promedio de todos los puntos asignados a su cluster, y despues de esto los puntos se reasignan, con el fin de minimizar la suma de las distancias cuadradas entre cada punto de datos y su centroide asignado.

4) **Random Forests (Bosques Aleatorios):** Es un método de ensamble de aprendizaje supervisado utilizado para clasificación y regresión. Consigue su objetivo construyendo un "bosque"de múltiples árboles de decisión individuales. Cada árbol se entrena en una submuestra aleatoria del conjunto de datos con reemplazo (bagging), y en cada punto de división se considera solo un subconjunto aleatorio de características. Las predicciones finales se obtienen combinando las predicciones de todos los árboles del bosque: por votación mayoritaria para clasificación o por promedio para regresión, lo que reduce la varianza y mejora la precisión.

5) **XGBoost (Extreme Gradient Boosting):** Es una implementación optimizada y escalable de algoritmos de "gradient boosting", un método de ensamble de aprendizaje **supervisado para clasificación y regresión**. Consigue su objetivo construyendo una serie de modelos individuales de forma secuencial (normalmente árboles de decisión), donde cada nuevo modelo se entrena para corregir los errores residuales de los árboles previos. Para ello, minimiza una "función de pérdida" que cuantifica el error del modelo empleando aproximaciones de segundo orden. Esto significa que, computa tanto la pendiente de la función de pérdida (como en el descenso de gradiente básico) como su curvatura para encontrar el mínimo, lo que resulta en una convergencia más rápida y una mayor exactitud. Además, aplica técnicas de regularización

avanzadas (L1 y L2) para prevenir el sobre-ajuste y mejorar la generalización. La regularización L1 (Lasso) penaliza la suma de los valores absolutos de los coeficientes, pudiendo llevar algunos a cero y realizando selección de características; la regularización L2 (Ridge) penaliza la suma de los cuadrados de los coeficientes, reduciéndolos sin eliminarlos.

6) **DBSCAN:** Es un algoritmo de aprendizaje **no supervisado para clusterización** basado en "densidad". Consigue su objetivo identificando regiones densas de puntos de datos separadas por regiones de menor densidad, clasificando los puntos como centrales, frontera o de ruido, y forma cada clúster a partir de puntos que están densamente conectados. No requiere que se especifique el número de clústeres de antemano y es capaz de descubrir clústeres de formas arbitrarias, además de identificar valores extremos.

7) **Redes Neuronales Profundas (Deep Neural Networks - DNN):** Brevemente, son modelos de aprendizaje **supervisado** que incluyen arquitecturas como los perceptrones multicapa (MLP). Consiguen su objetivo aprendiendo representaciones jerárquicas de los datos a través de múltiples capas ocultas de procesamiento (capas ocultas). El proceso de entrenamiento implica la propagación hacia adelante de la entrada para calcular una salida, seguida de la retropropagación ("Backpropagation") del error para ajustar iterativamente los pesos de las conexiones, minimizando la diferencia entre la salida predicha y la esperada. Son aptas para **clasificación y regresión en datos complejos**.

8) **Grandes Modelos de Lenguaje (Large Language Models - LLM):** Son un tipo avanzado de Redes Neuronales Profundas, frecuentemente basadas en la arquitectura de "transformers". Consiguen su objetivo mediante un pre-entrenamiento masivo en vastos corpus de texto y código, permitiendo aprender patrones complejos de lenguaje, gramática, semántica y contextualización. Su mecanismo de atención les permite ponderar la importancia de diferentes partes de la entrada, facilitando una comprensión profunda y la generación de texto coherente y relevante. Se utilizan para una amplia gama de tareas de procesamiento de lenguaje natural (PNL).

k)

A.2. Acrónimos

AI Inteligencia Artificial (Artificial Intelligence)

AI Act Ley de Inteligencia Artificial (Artificial Intelligence Act)

ANN Redes Neuronales Artificiales (Artificial Neural Network)

API Interfaz de Programación de Aplicaciones (Application Programming Interface)

ARIMA Modelo Autorregresivo Integrado de Medias Móviles (Autoregressive Integrated Moving Average)

AWS Amazon Web Services

BDTI Big Data Test Infrastructure

BERT Representaciones de Codificador Bidireccional de Transformadores (Bidirectional Encoder Representations from Transformers)

BOINC Berkeley Open Infrastructure for Network Computing

BTOS Encuesta sobre Tendencias y Panorama Empresarial (Business Trends and Outlook Survey)

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

CDN Red de Distribución de Contenidos (Content Delivery Network)

CI/CD Integración Continua/Entrega Continua (Continuous Integration/-Continuous Delivery)

CIF Corporate Information Warehouse

CKAN Comprehensive Knowledge Archive Network

CNN Red Neuronal Convolucional (Convolutional Neural 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)

DAGs Grafos Acíclicos Dirigidos (Directed Acyclic Graphs)

DAMA Asociación de Gestión de Datos (Data Management Association)

DBSCAN Clustering Espacial Basado en Densidad de Aplicaciones con Ruido (Density-Based Spatial Clustering of Applications with Noise)

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

DevOps Development and Operations

DNN Redes Neuronales Profundas (Deep Neural Networks)

DNS Sistema de Nombres de Dominio (Domain Name System)

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

EC2 Elastic Compute Cloud (Amazon EC2)

EEUU Estados Unidos

ETL Extract, Transform, Load (Extraer, Transformar, Cargar)

EU European Union (Unión Europea)

GAN Red Generativa Antagónica (Generative Adversarial Network)

GCP Google Cloud Platform

GPT Transformador Pre-entrenado Generativo (Generative Pre-trained Transformer)

GPU Unidad de Procesamiento Gráfico (Graphics Processing Unit)

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

HITL Human in the loop (humano en los procesos (de IA))

HMM Modelo Oculto de Markov (Hidden Markov Model)

HTTP Hypertext Transfer Protocol

HTML Lenguaje de Marcado de Hipertexto (HyperText Markup Language)

IA Inteligencia Artificial (Artificial Intelligence)

IaaS Infraestructura como Servicio (Infrastructure as a Service)

IBM International Business Machines (empresa)

IDE Entorno de Desarrollo Integrado (Integrated Development Environment)

INE Instituto Nacional de Estadística

IoT Internet de las Cosas (Internet of Things)

IP Internet Protocol

IPv4 Internet Protocol version 4

JSON Notación de Objetos de JavaScript (JavaScript Object Notation)

k-NN k-Nearest Neighbors (k-Vecinos Más Cercanos). Algoritmo de aprendizaje automático.

KV Key-Value (Almacenamiento Clave-Valor)

LGD Ley de Gobernanza de Datos

LLM Modelo de Lenguaje a Gran Escala (Large Language Model)

LSTM Memoria a Largo-Corto Plazo (Long Short-Term Memory)

MIT Massachusetts Institute of Technology (Instituto de Tecnología de Massachusetts)

ML Aprendizaje Automático (Machine Learning)

MLOps Operaciones de Aprendizaje Automático (Machine Learning Operations)

MLP Perceptrón Multicapa (Multi-layer Perceptron)

MoE Mezcla de Expertos (Mixture of Experts)

NAS Búsqueda de Arquitecturas Neuronales (Neural Architecture Search)

NLP Procesamiento del Lenguaje Natural (Natural Language Processing)

NoSQL Not only SQL

NPLM Modelo de Lenguaje Neuronal Probabilístico (Neural Probabilistic Language Model)

OECD Organización para la Cooperación y el Desarrollo Económicos (Organisation for Economic Co-operation and Development)

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

OKF Open Knowledge Foundation

ONG Organización No Gubernamental

ONTSI Observatorio Nacional de Tecnología y Sociedad

PaaS Plataforma como Servicio (Platform as a Service)

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

PLN Procesamiento del Lenguaje Natural (Natural Language Processing)

PYMEs Pequeñas y Medianas Empresas

R2 R2 Storage (Almacenamiento de Cloudflare)

RAM Memoria de Acceso Aleatorio (Random Access Memory)

RDS Servicio de Bases de Datos Relacionales (Relational Database Service)

RGPD Reglamento General de Protección de Datos

RIA Reglamento de Inteligencia Artificial

RL Aprendizaje por Refuerzo (Reinforcement Learning)

RNN Redes Neuronales Recurrentes (Recurrent Neural Network)

S3 Simple Storage Service (Amazon S3)

SaaS Software como Servicio (Software as a Service)

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

SES Simple Email Service (Amazon SES)

SIMPL Smart Middleware Platform for Cloud-to-Edge (EU)

SNS Simple Notification Service (Amazon SNS)

SQS Simple Queue Service (Amazon SQS)

SQL Lenguaje de Consulta Estructurado (Structured Query Language)

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 (Network Address Translation)

UE Unión Europea

VM Máquina Virtual (Virtual Machine)

XML Lenguaje de Marcado Extensible (eXtensible Markup Language)

XLSX Excel Open XML Spreadsheet

Bibliografía

AI Explained. An ai bubble? what altman actually said, the facts and nano banana. 2025.

Akash. The decentralized cloud built for ai's next frontier. s.f.

Albert Barron, t. p. C. M. Pizza as a service. 2015.

Amazon. Datos abiertos en aws. 2025.

Amazon Web Services. Amazon sagemaker. s-f.

Anderson, D. Boinc: a system for public-resource computing and storage. 2004.

Anslow, C., Brosz, J., Maurer, F. y Boyes, M. Datathons: An experience report of data hackathons for data science education. página 615–620, 2016.

Arellano Bruno, J. B. Uso de geolocalización y de fuentes de datos abiertas para la creación de servicios turísticos por la ciudad de madrid. 2019.

Awesome data. Awesome public datasets. 2025.

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

Barbudo, R., Ventura, S. y Romero, J. R. Eight years of automl: categorisation, review and trends. *Knowledge and Information Systems*, vol. 65(12), páginas 5097–5149, 2023.

Bengio, Y., Ducharme, R., Vincent, P. y Jauvin, C. A neural probabilistic language model. *Journal of Machine Learning Research*, vol. 3(Feb), páginas 1137–1155, 2003.

Benington, J. Creating the public in order to create public value? *International Journal of Public Administration*, vol. 32(3-4), 2009.

Böer, J. Self-service ai for everyone? a comparison of automl services. 2023.

bit2brain. Historia de la IA Inteligencia Artificial. 2020.

Bobrov, K. Data engineering: Now with 30 2025.

Bommala, H. et al. Cloud verse: mapping the new frontiers of cloud computing. vol. 392, página 01081, 2024.

BOZEMAN, B. *Public Values and Public Interest: Counterbalancing Economic Individualism*. Georgetown University Press, 2007. ISBN 9781589011779.

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.

Brownlee, J. *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Machine Learning Mastery, 2016.

Brownlee, J. *Machine Learning Algorithms From Scratch*. Machine Learning Mastery, 2024a.

Brownlee, J. A practical guide to choosing the right algorithm for your problem: From regression to neural networks. 2024b.

Brownlee, J. Machine learning mastery. s.f.

Campbell, M., Hoane Jr., A. J. y Hsu, F.-h. Deep Blue. *Artificial Intelligence*, vol. 134(1-2), páginas 57–83, 2002.

Challapally, A., Pease, C., Raskar, R. y Chari, P. State of ai in business 2025: The genai divide. 2025. Preliminary findings from AI Implementation Research, Research Period: January–June 2025.

Cheok, A. y Zhang, E. From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. 2023.

clouding.io. clouding.io, infraestructura cloud a tu servicio. s.f.

CNMC. Telecomunicaciones anual datos generales cnmc. 2025.

Concepción De Linares, J. B. Open food facts. 2025a.

Concepción De Linares, J. B. Planttes. 2025b.

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.

Danysoft. El ataque de los botnets. s.f.

de Datos Abiertos del Gobierno de España, P. El potencial uso de la metodología de dama en la gestión de los datos abiertos. 2021.

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

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

European Commission. Simpl: Cloud-to-edge federations empowering eu data spaces. 2024.

European Commission. Cloud computing. s.f.a.

European Commission. From hype to action using the big data test infrastructure (bdti). s.f.b.

European Parliament, C. Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance and amending regulation (eu) 2018/1724 (data governance act). Official Journal of the European Union, L 152, pp. 1-44, 2022. Accessed: 28 August 2025.

Eurostat. Cloud computing services by size class of enterprise. s.f.

Fang, X., Xu, W., Tan, F. A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S. y Faloutsos, C. Large language models (llms) on tabular data: Prediction, generation, and understanding – a survey. *arXiv preprint arXiv:2402.17944*, 2024. Version 4, last revised 21 June 2024.

Ferrer-Sapena A., A.-B. R., Peset F. Acceso a los datos públicos y su reutilización: Open data y open government. *El Profesional de la Información*, páginas 260–269, 2011.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M. y Hutter, F. Auto-sklearn 2.0: Hands-free automl via meta-learning. *arXiv:2007.04074 [cs.LG]*, 2020.

Fowdur, T. P. *Real-Time Cloud Computing and Machine Learning Applications*. Computer Science, Technology and Applications. Nova Science Publishers, Incorporated, New York, 2021. ISBN 978-1536198133. Description based upon print version of record.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., de Haas, S., Ibitoye, Y., Dafoe, A., Goldberg, B., Krier, S., Reese, A., Witherspoon, S., Hawkins, W., Rauh, M., Wallace, D., Franklin, M., Goldstein, J. A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Morris, M. R., King, H., y Arcas, B. A., Isaac, W. y Manyika, J. The ethics of advanced ai assistants. 2024.

Gigas. Gigas - cloud hosting solutions. s.f.

Giulia Carsaniga, D. R., Jochem Dogger. The use case observatory a 3-year monitoring of 30 reuse cases to understand the economic, governmental, social and environmental impact of open data volume i. *Publications Office of the European Union*, 2022.

GmbH, N. Nextcloud - your own private cloud. s.f.

Gómez-Obregón, J. Jaime gómez-obregón. 2025a.

Gómez-Obregón, J. La donación - jaime gómez-obregón. 2025b.

Gómez-Obregón, J. Subvenciones - jaime gómez-obregón. 2025c.

Gobierno de España. Big data test infrastructure: Un entorno gratuito para que las aa.pp experimenten con sus datos abiertos. 2021a.

Gobierno de España. Plan de digitalización de las administraciones públicas 2021-2025. 2021b.

Gobierno de España. Eventos datathon gobierno de españa. 2025a.

Gobierno de España. Eventos datos abiertos gobierno de españa. 2025b.

Gobierno de España. Portal de datos abiertos. 2025c.

Gobierno de España. Nubesara. s.f.

Golem Network. Golem network. s.f.

Google. Data commons. 2025.

Google. Google vertex ai platform. s-f.a.

Google. Welcome to colab! s-f.b.

Google Cloud. Amplía la infraestructura y la ia de google cloud on-premise. s.f.

Google Trends. Google trends artificial intelligence. 2025. Se puede observar un pico en 2011, posiblemente por la salida de Siri, y la tendencia creciente desde 2022 debido a chatGPT. Las cifras representan el interés de búsqueda en relación con el punto más alto del gráfico para la región y el momento determinados. Un valor de 100 es el pico de popularidad del término. Un valor de 50 significa que el término es la mitad de popular. Una puntuación de 0 significa que no había datos suficientes para este término.

Herrera Capriz, M. E. El valor intangible de la transparencia: un análisis de los datos abiertos de españa en el marco de la corriente "valor público". 2024. Tesis inédita de la Universidad Complutense de Madrid, Facultad de Ciencias de la Información, leída el 08/05/2024.

Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Natl Acad Sci U S A*, vol. 81(10), páginas 3088–3092, 1984.

Huawei Technologies Co., Ltd. *Cloud Computing Technology*. Springer and Posts & Telecom Press, Singapore and Beijing, 2023. ISBN 978-981-19-3026-3. Open access book.

huggingface. huggingface spaces. s-f.

Insights + Analytics, E. Datos para espaÑa y mundiales de investigaciÓn de mercados 2023. 2024.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Le Scao, T., Gervet, T., Lavril, T., Wang, T., Lacroix, T. y El Sayed, W. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jin, H., Chollet, F., Song, Q. y Hu, X. Autokeras: An automl library for deep learning. *Journal of Machine Learning Research*, vol. 24(6), páginas 1–6, 2023.

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.

Kaggle. Kaggle code notebook. s-f.a.

Kaggle. Kaggle datasets. s-f.b.

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.

Kreuzberger, D., Kühl, N. y Hirschl, S. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, vol. 11, páginas 31866–31879, 2023.

Kumar, R., Adwani, L., Kumawat, S. y Jangir, S. K. Opennebula: Open source IaaS cloud computing software platforms. 2014.

Lisdorf, A. *Cloud Computing Basics: A Non-Technical Introduction*. Apress, 2021. ISBN 978-1-4842-6921-3. Accessed: 2025-08-27.

LLamocca Portela, P. Integración y visualización de datos abiertos medioambientales. 2016. Máster en Ingeniería Informática, curso 2015-2016.

LLM Stats. Llm leaderboard. s.f.

Market report analytics. Spain cloud computing market market's growth blueprint. 2024.

Matheus, R. y Janssen, M. A systematic literature study to unravel transparency enabled by open government data: The window theory. *Public Performance & Management Review*, vol. 43(3), páginas 503–534, 2020.

Melendrez Moreto, I. Auditoría y metodología de implantación de open data para smart cities. 2016. Máster en Ingeniería Informática, curso 2015-2016, a destacar: gran compilacion de datasets, conceptos interesantes como cinco estrellas del Open Data, manual de CKAN.

Meneses Vicente, G. y García Ruíz, J. F. Madrid, isla de calor. 2023. Trabajo de Fin de Master en Ingeniería Informática.

Meynhardt, T. Public value inside: What is public value creation? *International Journal of Public Administration*, vol. 32(3-4), páginas 192–219, 2009.

microsoft. microsoft azure machine learning. s-f.

Microsoft. Cloud computing terminology. s.f.a.

Microsoft. ¿qué es la tokenización? s.f.b.

Microsoft. What is the medallion lakehouse architecture? s.f.c.

Moore, M. Creating public value: Strategic management in government. 1995.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. y Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, vol. 56(4), páginas 3005–3054, 2023. ISSN 1573-7462.

Nigro, H. Cloud computing: Retos y oportunidades. *Rev. Ingeniería, Matemáticas y Ciencias de la Información*, vol. 9(18), páginas 11–16, 2022.

Nijssen, D. The use case observatory a 3-year monitoring of 30 reuse cases to understand the economic, governmental, social and environmental impact of open data volume ii. *Publications Office of the European Union*, 2022.

Nov, O., Anderson, D. y Arazy, O. Volunteer computing: A model of the factors determining contribution to community-based scientific research. *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, páginas 741–750, 2010.

Observatorio Nacional de Tecnología y Sociedad. Indicadores de uso de inteligencia artificial en las empresas españolas. edición 2025 - datos 2024. 2025. NIPO: 230250053, DOI: 10.30923/230250053.

OECD. *The Path to Becoming a Data-Driven Public Sector*. OECD Digital Government Studies. OECD Publishing, 2019.

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

Open Data Charter. ¿qué son los datos abiertos? s.f.

Open Knowledge Foundation. Open definition. 2025.

ovhcloud. Ovhcloud ai training. s-f.

OVHCloud. Public cloud free trial ovhcloud. s.f.

Pangarkar, T. Big data statistics 2025 by patterns in the dimensions. 2025.

Sánchez de Paz, M. Optimización de infraestructuras de cloud computing basadas en máquinas virtuales. 2023. Trabajo de Fin de Máster en Ingeniería Informática, Curso 2022/2023.

Portal Europeo de Datos. ¿qué son los datos abiertos? 2025.

PreparaTIC. Servicio nubesara. casos prÁcticos y ejemplos de aplicaciÓn. s.f.

Press, G. A very short history of big data. 2013.

radwere. What is mydoom malware? s.f.

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.

Ray, P. P. An introduction to dew computing: Definition, concept and implications. *IEEE Access*, vol. 6, páginas 723–737, 2018.

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

Render Network. The distributed gpu render network. s.f.

R.I.Pienaar. Free for devs. 2025.

Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65(6), páginas 386–408, 1958.

Salehin, I., Islam, M. S., Saha, P., Noman, S., Tuni, A., Hasan, M. M. y Baten, M. A. Automl: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, vol. 2(1), páginas 52–81, 2024. ISSN 2949-7159.

Santos, J. S. Análisis de datos de la ciudad de madrid. 2023. Trabajo de Fin de Grado en Ingeniería Informática.

Sapkota, R., Roumeliotis, K. I. y Karkee, M. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *Information Fusion*, vol. 126, 2025. ISSN 1566-2535.

Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, vol. 2(3), página 160, 2021.

SONM. Sonm whitepaper. s.f.

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.

Sze, V., Chen, Y.-H., Emer, J., Suleiman, A. y Zhang, Z. Hardware for machine learning: Challenges and opportunities. 2017.

Talbot, C. Paradoxes and prospects of public value. *Public Money & Management*, vol. 31(1), 2011.

Taylor, P. Big data - statistics and facts. 2025.

Telefónica. Telefónica tech cloud platform. s.f.

Theodorakopoulos Leonidas, S. Y., Theodoropoulou Alexandra. A state-of-the-art review in big data management engineering: Real-life case studies, challenges, and future research directions. *Eng*, vol. 5(3), página 1266, 2024. Copyright - © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Última actualización - 2024-09-27.

Soledad De la Torre, S. N. 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. 2016.

Union Europea. Explicación de la ley de gobernanza de datos. 2023.

Union Europea. Reglamento de inteligencia artificial. 2024.

Union Europea. European alternatives for popular services. 2025a.

Union Europea. Portal de datos abiertos. 2025b.

UniversiDATA. Análisis de desplazamientos interurbanos en estudiantes. 2020.

UniversiDATA. Ii datathon universidata. 2024.

UniversiDATA. Universidata. 2025.

Universitat Autònoma de Barcelona. infoparticipa. 2025.

university, T. Data mining without coding. 2024.

Uriarte, R. B. y De Nicola, R. Blockchain-based decentralised cloud/-fog solutions: Challenges, opportunities and standards. *IEEE Communications Standards Magazine*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. Attention is all you need. páginas 5998–6008, 2017.

Verhulst, S., Young, A., Zahuranec, A., Calderon, A., Gee, M. y Aaronson, S. The emergence of a third wave of open data: How to accelerate the re-use of data for public interest purposes while ensuring data rights and community flourishing. *International Journal of Public Administration*, página 8, 2020.

Vogel, A., Griebler, D., Maron[1], C., Schepke, C. y Fernandes, L. Private iaas clouds: A comparative analysis of opennebula, cloudstack and openstack. páginas 672–679, 2016.

Werbos, P. Beyond regression: New tools for prediction and analysis in the behavioral science. thesis (ph. d.). appl. math. harvard university. 1974.

Wikipedia. Information age. 2025.

Wikipedia. listo of machine learning methods. s.f.