Universidad de Alcalá

Escuela Politécnica Superior

Grado en Ingeniería Informática



**Trabajo Fin de Grado**

Plataforma basada en Python para transformación e integración de datos abiertos (open data) sobre servicios de análisis y visualización

**Autor:** Marcos Navarro Juan

**Tutor:** Manuel De Buenaga Rodríguez

2023

|  |
| --- |
| UNIVERSIDAD DE ALCALÁ  Escuela Politécnica Superior  **Grado en Ingeniería Informática**  Trabajo Fin de Grado  Plataforma basada en Python para transformación e integración de datos abiertos (open data) sobre servicios de análisis y visualización  **Autor:** Marcos Navarro Juan  **Tutor/es:** Manuel De Buenaga Rodríguez  **TRIBUNAL:**  **Presidente:** << Nombre y Apellidos >>  **Vocal 1º:** << Nombre y Apellidos >>      **Vocal 2º:** << Nombre y Apellidos >>    **FECHA**: << Fecha de depósito >> |

*A mi familia y amigos*

Agradecimientos.

Son muchas las personas a las que tengo que agradecer que hoy sea mucho mejor persona que cuando empecé este camino, cuando empecé esta bonita travesía de estudiar el grado ingeniería informática.

En primer lugar, quiero dar las gracias a mis padres y a mis dos hermanos, que gracias a su apoyo he podido crecer tanto profesionalmente como persona.

En segundo lugar, ha sido de vital importancia rodearme de mi círculo de amistades, los cuales me han dado apoyo tanto en las mejores situaciones como en la peores. Especialmente a mis compañeros de la universidad, con los cuales hemos sacado multitud de proyectos y exámenes adelante.

Quiero destacar el curso que realice con la beca Erasmus, el año que pase en Varsovia, Polonia. El cual, fue fundamental para entender quién soy a día de hoy y esto es gracias a la Politechnika Warszawska y a las amistades inolvidables que forje en ese increíble viaje.

Por último, quiero agradecer a quienes han hecho posible este recorrido, agradecer a la propia universidad de Alcalá y a los diferentes profesores que me han ido acompañando en cada una de las asignaturas y por supuesto a mi tutor de TFG Manuel de Buenaga.

Resumen.

Abstract.

Resumen Extendido.

Contenido

[Agradecimientos. 4](#_Toc141982130)

[Resumen. 5](#_Toc141982131)

[Abstract. 6](#_Toc141982132)

[Resumen Extendido. 7](#_Toc141982133)

[1. Introducción. 14](#_Toc141982134)

[1.1. Motivación 14](#_Toc141982135)

[1.2. Objetivos 15](#_Toc141982136)

[1.3. Línea de Trabajo. 15](#_Toc141982137)

[1.4. Estructura de la Memoria. 17](#_Toc141982138)

[2. Estado del Arte. 19](#_Toc141982139)

[2.1. Open Data 19](#_Toc141982140)

[2.2. Análisis de datos 21](#_Toc141982141)

[2.3 Proceso ETL. 22](#_Toc141982142)

[2.4 Python. 23](#_Toc141982143)

[2.5 Visualización en PowerBi. 25](#_Toc141982144)

[3. Herramientas a Utilizar. 26](#_Toc141982145)

[4. Implementación. 27](#_Toc141982146)

[4.1. Tipos de Fuentes de Datos y Proyectos de Investigación. 27](#_Toc141982147)

[4.1.1. Smacite. 27](#_Toc141982148)

[4.1.2. EULOGH. 29](#_Toc141982149)

[4.2. Fuentes de Datos Abiertas. 30](#_Toc141982150)

[4.2.1. Incendios producidos en España entre el 2006 y el 2015. 32](#_Toc141982151)

[4.2.2. Concesión de Nacionalidad Española entre el 2010 y el 2019. 33](#_Toc141982152)

[4.2.3. Capacidad Asistencial durante la Covid-19. 34](#_Toc141982153)

[4.2.4. Catálogo de Parques Municipales en la Ciudad de Madrid. 35](#_Toc141982154)

[4.2.5. Siniestralidad en la Carreteras en la Ciudad de Madrid. 37](#_Toc141982155)

[4.2.6. Población por Provincias de España 1996-2021. 38](#_Toc141982156)

[4.2.7. Población por Países en la unión europea 2001-2022. 39](#_Toc141982157)

[4.2.7. Catálogo del Bosque Urbano de la Ciudad de Madrid. 40](#_Toc141982158)

[4.2.8. Calificación de las diferentes fuentes de datos abiertas. 42](#_Toc141982159)

[4.3. Proceso ETL en las Fuentes de Datos. 43](#_Toc141982160)

[4.3.1. Proceso ETL en Catálogo de Parques Municipales en la ciudad de Madrid. 43](#_Toc141982161)

[4.3.2. Proceso ETL en Capacidad Asistencial durante la Covid-19. 46](#_Toc141982162)

[4.3.3. Proceso ETL en Incendios producidos en España entre el 2006 y el 2015. 49](#_Toc141982163)

[4.3.4. Proceso ETL en Concesión de Nacionalidad Española entre el 2010 y el 2019. 50](#_Toc141982164)

[4.3.5. Proceso ETL en Catálogo del Bosque Urbano de la Ciudad de Madrid. 52](#_Toc141982165)

[4.3.6. Proceso ETL en Población por Provincias de España 1996-2021. 56](#_Toc141982166)

[4.4. Comparativa entre las librerías ETL. 57](#_Toc141982167)

[4.5. Modelos Estadísticos. 59](#_Toc141982168)

[4.5.1. Análisis Descriptivo. 59](#_Toc141982169)

[7. Bibliografía. 60](#_Toc141982170)

Figuras

[Figura 1. Diagrama de Metodología *CRISP-DM* 17](#_Toc141968247)

[Figura 2. Tabla de puntaciones de los países miembros en el año 2022. 20](#_Toc141968248)

[Figura 3. Tabla de puntaciones de la aportación de fuentes de datos abiertas a nivel europeo. 21](#_Toc141968249)

[Figura 4. Esquema del funcionamiento de un proceso ETL. 23](#_Toc141968250)

[Figura 5. Evolución de los lenguajes de Programación en función del tiempo. 24](#_Toc141968251)

[Figura 6. Mapa de las universidades relacionadas con Eulogh. 29](#_Toc141968252)

[Figura 7. Campos que aborda el proyecto Eulogh. 30](#_Toc141968253)

[Figura 8. Ejemplo de tabla extraída de la fuente de datos “Incendios producidos en España entre el 2006-2015”. 33](#_Toc141968254)

[Figura 9. Ejemplo de tabla extraída de la fuente de datos “Concesión de Nacionalidad entre el 2010 y el 2019”. 34](#_Toc141968255)

[Figura 10. Ejemplo del histórico extraído de la fuente de datos “Capacidad Asistencial durante el Covid-19”. 35](#_Toc141968256)

[Figura 11. Ejemplo de cómo se expone la información en la fuente de datos “Catalogo de Parques Municipales de Madrid”. 37](#_Toc141968257)

[Figura 12. Ejemplo de la composición de la fuente de datos “Siniestralidad en las Carreteras en la Ciudad de Madrid”. 38](#_Toc141968258)

[Figura 13. Ejemplo de la fuente de datos “Población por Provincias de España”. 39](#_Toc141968259)

[Figura 14. Ejemplo de la fuente de datos “Población por Países de la unión Europa”. 40](#_Toc141968260)

[Figura 15. Ejemplo de la exposición de la información de la fuente de datos “Valor del Bosque Urbano de Madrid”, de la página 51 del PDF. 41](#_Toc141968261)

[Figura 16. Parte del data warehouse una vez realizado el proceso ETL. 46](#_Toc141968262)

[Figura 17. Resultado del Proceso ETL de la fuente de datos “Capacidad Asistencial durante la Covid-19”. 48](#_Toc141968263)

[Figura 18. Tabla extraída de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”, página 26. 50](#_Toc141968264)

[Figura 19. Parte del data warehouse una vez realizado el proceso ETL. 50](#_Toc141968265)

[Figura 20. Parte del data warehouse una vez realizado el proceso ETL de la fuente de datos “Concesiones de Nacionalidad Española entre el 2010 y el 2019”. 52](#_Toc141968266)

[Figura 21. Parte de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”. 55](#_Toc141968267)

[Figura 22. Parte de la visualización del data warehouse de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”. 55](#_Toc141968268)

[Figura 23. Parte de la visualización del data warehouse de la fuente de datos “Población por Provincias de España 1996-2021”. 56](#_Toc141968269)

Tablas

[Tabla 1. Calificación de las diferentes fuentes de datos seleccionadas. 42](#_Toc141968275)

[Tabla 2. Comparación entre las diferentes librerías para los procesos ETL. 58](#_Toc141968276)

[Tabla 3. Comparación entre las librerías para la extracción de tablas. 58](#_Toc141968277)

Código.

[Código 1. Función get\_text del proceso de extracción de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”. 44](#_Toc141968288)

[Código 2. Función clean del proceso de transformación de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”. 45](#_Toc141968289)

[Código 3. Función export\_xlsx del proceso de carga de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”. 45](#_Toc141968290)

[Código 4. Proceso de extracción de la fuente de datos “Capacidad Asistencial durante la Covid-19”. 47](#_Toc141968291)

[Código 5. Proceso de carga de la fuente de datos “Capacidad Asistencial durante la Covid-19”. 48](#_Toc141968292)

[Código 6. Proceso de extracción y transformación de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”. 49](#_Toc141968293)

[Código 7. Proceso de Carga de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”. 49](#_Toc141968294)

[Código 8. Proceso de ETL de la fuente de datos “Concesiones de Nacionalidad Española entre el 2010 y el 2019”. 51](#_Toc141968295)

[Código 9. Proceso de extracción de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”. 53](#_Toc141968296)

[Código 10. Proceso de transformación de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”. 54](#_Toc141968297)

[Código 11. Proceso de carga de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”. 54](#_Toc141968298)

[Código 12. Proceso ETL de la fuente de datos “Población por Provincias de España 1996-2021”. 56](#_Toc141968299)

Acrónimos y Abreviaturas.

*Km*

*Ud*

*M^2*

Capítulo 1

1. Introducción.

En este primer capítulo del TFG, en primer lugar, se va a exponer diferentes argumentos del porque se ha realizado este proyecto, una serie de objetivos a cumplir, se definirán unas líneas de trabajo a seguir y finalmente se describirá brevemente la estructura del documento.

* 1. Motivación

La información, determinar qué información es la relevante y cual no lo es y exponerla de una forma adecuada en la que se pueda interpretar, tiene un increíble poder en nuestra sociedad. A lo largo de nuestra historia como seres humanos, la información siempre ha jugado un papel determinante y es en la actualidad donde más información recibidos y recabamos. Por ello, todo tipo de gobiernos, empresas, y más organismos invierten tanto tiempo y dinero en poder entender y descifrar la mayor cantidad de información posible y actuar en consecuencia.

En la actual, el papel de internet juega un papel vital. Gracias a internet y la era de las tecnologías en la que vivimos recabar esta información o datos es un proceso que se produce en cada momento que estamos conectados. No importa en que dispositivo o aplicación o en qué lugar estemos conectados que permanente se está enviando información relevante.

Al visualizar la información que se considera de valor utilizando las diferentes fuentes de datos, provoca en una mejor toma de decisiones. Al tener una plataforma que facilite la transformación y la integración de datos abiertos, los diferentes agentes realizaran un análisis más consistente de la situación, por tanto, aumentará la calidad de las decisiones, que esta soportadas por evidencias solidas.

Los datos abiertos contienen una gran cantidad de información de gran valor que ayuda a comprender mejor diversos fenómenos, esta información esta presentada por diferentes organismos públicos o privados que garantizan la consistencia del dato. El acceso a estos datos abiertos es una manera libre, por lo que, provoca un impulso en la innovación y una colaboración ciudadana.

El concepto de Open Data es un concepto que nació en el siglo XXI el cual tiene una tendencia evolutiva exponencial, donde son cada vez más los organismos públicos o privados que ponen a disposición de aquel que lo desee información sobre infinidad de cuestiones.

La utilización del lenguaje Python y su amplia gama de bibliotecas que se van incrementando de forma exponencial supone una eficiencia en el procesamiento de datos y en el análisis de datos. Utilizando e investigando las bibliotecas de Python puede permitir la automatización de tareas tediosas y repetitivas, ahorrando tiempo y recursos en el manejo de grandes volúmenes de datos

Debido a este conjunto de ideas y de la importancia que siempre ha tenido y más en la actualidad el extraer información de valor sobre un conjunto de datos nace la motivación para la creación de este proyecto. Además, el proyecto en cuestión afectaría a infinidad de campos, donde algunos de estos suponen de una importancia notable para la sociedad. Por ejemplo, se puede destacar el campo de la medicina, el de la salud y medioambiente, campos financieros, deportivos.

* 1. Objetivos

El objetivo del proyecto es trabajar sobre un conjunto de fuentes de datos que cumplan la característica de ser open data y que pertenezcan a diferentes campos de la sociedad actual. Algunas de estas fuentes de datos estarán relacionadas con proyectos de investigación en los que ha participado el departamento (Eugloh o Smacite). Se aplicarán nuevas funcionalidades a través de Python para la extracción, transformación y carga de los datos (ETL) y para el análisis a los diferentes conjuntos de datos. El resultado de este análisis se presentará de la manera más visual apoyándose en el software de Microsoft PowerBi. Este objetivo principal está compuesto por subobjetivos más concretos. Estos subjetivos son los siguientes:

* La utilización de fuentes de datos abiertas (Open Data), cuyos dominios están relacionados con las ciudades inteligentes (Smart Cities) y con la salud y el medioambiente. Las administraciones públicas proporcionan una serie de bases de datos abiertas de múltiples campos de información. Algunos de estos campos pueden ser de información relacionada con accidentes de tráfico, la distribución de mobiliario urbano, valores climatológicos, … Donde cualquier tipo de agente individual u organismo puede acceder a esta información y realizar proyectos que fomentan la innovación y el desarrollo.
* Relación al proyecto internacional “Eugloh” *[1],* realizado por universidades europeas, incluida la Universidad de Alcalá de Henares. Eugloh es un proyecto relacionado con la salud global y Smart Cities.
* Relación con el proyecto internacional “Smacite” *[2],* el cual es un proyecto financiado por la unión europea en donde trabajan diferentes organizaciones, una de ellas es la universidad de Universidad de Alcalá de Henares. Es un proyecto relacionado con el concepto de ciudad inteligente y con todo lo representa.
* Realización de procesos de extracción, transformación y carga de datos a través de Python para las fuentes de datos abiertas seleccionadas, para que el dato sea consistente y heterogéneo a la hora de aplicar las diferentes técnicas de análisis. Donde incluso las fuentes de datos utilizadas tengan diferente formato, pero sean validas a la hora de integrarse en la plataforma.
* La implementación de técnicas avanzadas basadas en analítica de datos, desarrolladas en Python, sobre las bases de datos abiertas anteriormente descritas y su integración en PowerBi. A partir de las soluciones obtenidas en las diferentes técnicas de análisis se aporta una valoración de los resultados que aportan funcionalidad.
* Visualizar la información recabada tras la implementación de técnicas avanzadas basadas en analítica de datos, de una forma intuitiva y compresible para el usuario. Tambien se visualizará las soluciones obtenidas y la interpretación de valor de cada una de ellas.
  1. Línea de Trabajo.

Para el desarrollo del trabajo se utiliza la metodología de *CRISP-DM*. Esta metodología consiste en dividir el trabajo en diferentes fases, en concreto, el proyecto se divide en 7 fases de *CRISP-DM*, que a su vez se divide en subtareas:

1. Definición de objetivos y alcance del proyecto. Se definen los objetivos que se quieren alcanzar y se determina si son viables o no, en función del tiempo y de los recursos. También se define el alcance del proyecto teniendo en cuenta las funcionalidades especificas a desarrollar.
2. Investigación y selección de fuentes de datos abiertos. Se investigan y se seleccionan las fuentes de datos abiertas a utilizar en el proyecto. La investigación se realiza en diferentes paginas gubernamentales donde se exponen las diferentes bases de datos y sus características. Se seleccionan las fuentes de datos en función de la temática (en relación con “Eulogh”, ciudades inteligentes, salud y medioambiente …) y teniendo en cuenta la cantidad de información y la consistencia de los datos.
3. Implementación de la extracción y transformación de datos. Se realizan los diferentes programas a través de Python para realizar las operaciones ETL para cada una de las fuentes de datos.
4. Implementación de las técnicas avanzadas basadas en analítica de datos. Se realizan las diferentes técnicas de analítica de datos para las diferentes fuentes de datos seleccionadas. Dependiendo de los objetivos esperados de cada fuente de datos se realiza una técnica de u otra. Se realiza el análisis de datos utilizando Python y diferentes bibliotecas que provee el lenguaje.
5. Desarrollo de las herramientas de visualización en PowerBi. En función de los resultados obtenidos en el paso anterior y de las fuentes de datos se utiliza el software PowerBi para visualizar diferentes aspectos relevantes al proyecto.
6. Pruebas y evaluación de la plataforma. A lo largo de todo el proyecto se realizan las pruebas y las evaluaciones correspondientes a cada etapa del proyecto.
7. Documentación. A medida que se va a realizando cada una de las etapas, se va explicando en el documento los pasos realizados y a aportación teórica necesaria para comprender de una forma sencilla el funcionamiento del proyecto.
8. Despliegue. Una vez realizada cada una de las etapas se prepara la plataforma para su despliegue y su complemento funcionamiento.

La línea de trabajo tiene el siguiente flujo de estados:

Diagrama

Descripción generada automáticamente

Figura . Diagrama de Metodología *CRISP-DM*

Una ventaja notoria de utilizar *CRISP-DM* *[33]*, es que permite moverse entre las distintas fases, cuando el proyecto lo requiere. Como se puede observar en la Figura 1, a lo largo del proyecto se realiza de forma paralela la documentación y también teniendo en todo momento acceso a las diferentes fuentes de datos.

Para facilitar la viabilidad se fija un hito, en mitad del tiempo destinado al desarrollo, donde se ha avanzado de forma de significado en todas las fases del proyecto. Completando todas las fases restantes.

* 1. Estructura de la Memoria.

La estructura de la memoria está compuesta por capítulos, la estructura es la siguiente:

* Capítulo 1. Se realiza una breve introducción del proyecto, se explica las motivaciones que se han producido para llevar a cabo el proyecto, se define un flujo de trabajo y se explica cómo va a ser la estructura de la memoria.
* Capítulo 2. Se explican teóricamente los conceptos necesarios para entender el proyecto. Los conceptos que destacan son los de: Open Data, ETL, Análisis de datos, Python y Visualización con PowerBy.
* Capítulo 3. Se explican las herramientas que se han llevado a cabo para realizar el proyecto y el porqué de la utilización de cada una.
* Capítulo 4. Se expone la implementación llevada a cabo. Se expone cada una de las fuentes de datos escogidas y el porqué de la elección, se explica todo lo correspondiente con el proceso de ETL, todo el proceso del análisis de datos y de la visualización de resultados y la estructura de la plataforma.
* Capítulo 5. Se exponen todos los resultados obtenidos y se explica detalladamente que significado y que información se pueden obtener.
* Capítulo 6. Se comentan las conclusiones obtenidas al finalizar el proyecto, la posible escalabilidad del mismo y temas de consideración como posibles aplicaciones, mejoras y análisis.
* Capítulo 7. Se indican todas las referencias utilizadas para la realización del proyecto.

Capítulo 2

1. Estado del Arte.

A continuación, se explican los conceptos de mayor relevancia del proyecto y las razones por las que estos conceptos son utilizados en el proyecto. Para la compresión del proyecto es necesario tener claro una serie de conceptos, entender cómo funcionan y cuáles son sus puntos fuertes y débiles. Por ello, en este punto se presentan los siguientes conceptos desde una perspectiva teórica y se explican las diferentes ventajas por las que se ha decidido que forman parte del proyecto.

* 1. Open Data

El concepto de “Open Data” se basa en la premisa proporcionar al público ciertos conjuntos de dato de una manera libre y sin restricciones excesivas. Existen diferentes organizaciones que generan o poseen datos que ponen a disposición del público en general.

El objetivo del concepto es de promover la transferencia y la apertura de la información, haciéndola accesible a todo el que lo desee. Al liberar la información, se busca fomentar la participación ciudadana para colaborar y desarrollar soluciones innovadoras en todo tipo de ámbitos. Tambien puede provocar un aumento de la calidad de la información presentada, ya que, al permitir que los ciudadanos e investigadores analicen y examinen la información pueden detectar errores o mejorías.

Las fuentes de datos se pueden estructurar en dos categorías principales, fuentes datos estructuradas, no estructuradas y semiestructuradas, durante el proyecto se trata con los dos primeros tipos de fuentes de datos.

Las fuentes de datos estructuradas contienen información organizada en un formato uniforme y predefinido, los datos forman parte de un esquema fijo con tablas, columnas y filas. Ejemplos de este tipo de fuente son las bases de datos relacionales, archivos CSV y XML y bases de datos SQL.

Su principal característica es que son fácilmente legibles por computadores y se pueden almacenar, consular y analizar utilizando lenguajes de consulta estructurados.

Las fuentes de datos no estructuradas contienen la información en un formato sin una estructura predefinida y uniforme. La información se presenta de una forma libre sin seguir ningún tipo de esquema fijo. Algunos ejemplos de este formato es la información presentada en documentos de Word o archivos PDF, publicaciones de redes sociales, imágenes y videos.

La principal característica es que son más difíciles de analizar y procesos, ya que no siguen ningún patrón definido lo que dificulta la extracción. Para su tratamiento se requiere el uso de técnicas más avanzadas, como pueden ser el procesamiento de lenguaje natural (NLP), reconocimiento de patrones y aprendizaje automático.

Las fuentes de datos semiestructuradas tienen cierto grado de estructura, pero no cumplen los suficientes requisitos de un formato predefinido. Un ejemplo de este formato son los JSON, que proporciona una estructura flexible para almacenar y transferir datos.

Las fuentes de datos que se utilizan en este proyecto son todas de fuentes de datos abierto, ya que, nos brindan las siguientes ventajas en función a los objetivos definidos del proyecto:

* Libre acceso sin costo alguno. Debido a que la característica principal de las fuentes de datos abiertas es que tienen un acceso libre y sin restricciones provoca que el uso de las bases de datos de diferentes ámbitos no tenga coste alguno. Lo cual, es una ventaja significativa a tener en cuenta en todos tipos de proyecto y más cuando son de índole académica.
* Acceso a fuentes de datos de diversos indoles. Al utilizar open data, se tiene un acceso a una amplia variedad de bases de datos de diferentes dominios. Organizaciones como puede ser el gobierno de España o la unión europea ponen a disposición multitud de fuentes de datos abiertas de diferentes categorías como demografía, empleo, medio ambiente, economía, …
* Calidad y confiabilidad de los datos. Como detrás de estas fuentes de datos abiertas se encuentran instituciones de reconocidas provoca un mayor grado de confiabilidad y calidad en comparación con otras fuentes de datos no verificadas. Lo que permite trabajar sabiendo que la información con la se trabaja es precisa y confiable, lo cual, es un aspecto fundamental para el análisis y visualizaciones confiables.
* Innovación. Al utilizar open data se fomenta la innovación y el desarrollo de soluciones creativas. Gracias estos conjuntos de datos se pueden descubrir nuevas conclusiones, patrones y tendencias que pueden impulsar la creación de herramientas analíticas y visuales únicas.
* Enfoque en problemas sociales. El contenido de las fuentes de datos abiertas se corresponde con temas sociales, por lo que, se pueden abordar problemas sociales o deficiencias del sistema y proporcionarle una solución de calidad que tenga un impacto real en la calidad de vida.

Cabe indicar que España se encuentra entre los lideres del fomento del Open Data a nivel europeo. Todos los gobiernos de los países pertenecientes a la unión europea tienen un sistema donde proporcionan fuentes de datos abiertas que abarcan todos los ámbitos de la vida social. El “*Open Data Maturity Report [3]*” recoge el desempeño de los países europeos en términos de aportaciones y lo califica con un sistema de puntación de cuatro dimensiones (política, social, impacto y calidad). En los últimos años, los países miembros de la unión europea han mejorado notablemente sus aportaciones, por tanto, sus puntuaciones, en el 2015 la puntuación media era del 44% y en el 2022 se ha incrementado hasta el 81%. En la *Figura 2* se puede observar las puntuaciones de cada país miembro de la unión europea en función de la política, social, impacto y calidad.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figura . Tabla de puntaciones de los países miembros en el año 2022.

España se encuentra en la tercera posición con más aportaciones y se encuentra por encima de la media europea como se puede observar en la *Figura 3*.

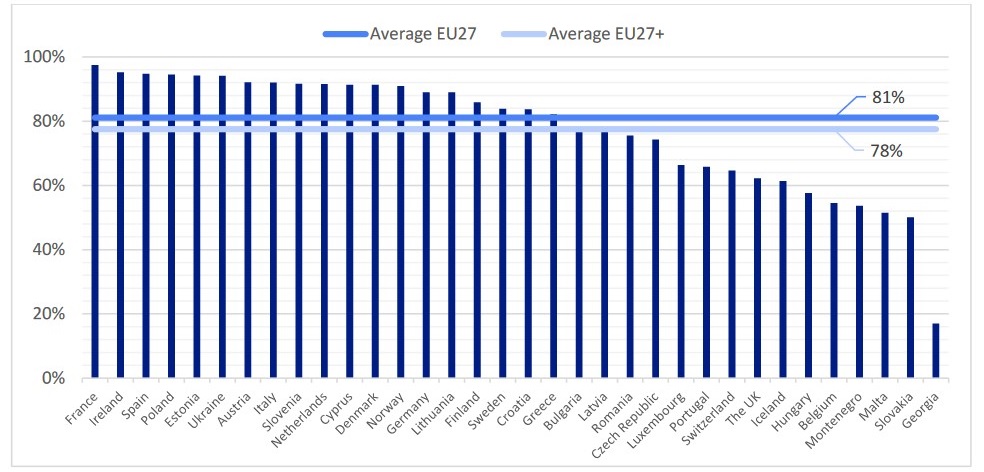


Figura . Tabla de puntaciones de la aportación de fuentes de datos abiertas a nivel europeo.

Como se puede observar la utilización de fuentes de datos abiertas es una práctica que ha ido aumentado de forma notable y para el futuro se prevé que evolucione siguiendo la misma tendencia exponencial.

2.2. Análisis de datos

El análisis de datos es el proceso por el cual se examinan, se limpian, se transforman e interpretan una serie de conjuntos de datos. Tiene como objetivo encontrar patrones, relaciones para extraer conocimientos que aporten valor para tomar decisiones con información relevante.

Este concepto es fundamental a la hora de tomar decisiones relevantes. Ya que un correcto análisis de datos otorga cierta información de valor, como pueden ser tendencias, patrones o relaciones, que ayudan a comprender situaciones, evaluar opciones y respaldar la toma de decisiones. La información obtenida puede servir para optimizar procesos, al analizar los datos de diferentes procesos, sirve para identificar ineficiencias, cuellos de botella o áreas de mejora, lo cual puede aumentar la eficiencia y reducir costos.

También la información obtenida a partir del análisis de datos sirve para identificar oportunidad y riesgos. Se pueden detectar oportunidades de negocio, nichos de mercado o tendencias emergentes, pero a su vez, también se puede detectar riegos potenciales, como anomalías o desviaciones significativas en los datos.

Existen diferentes tipos de análisis de datos que se pueden realizar sobre un conjunto de datos, cada uno con enfoques y técnicas diferentes para conseguir una serie de objetivos. A continuación, se describen los más significativos:

* Análisis Descriptivo. Consiste en resumir y describir los datos utilizando medidas de estadística básica como pueden ser promedios, medianas, desviaciones. Tiene como objetivo proporcionar una compresión inicial de los datos.
* Análisis Exploratorio. Consiste en descubrir patrones y relaciones ocultas en los datos a partir de técnicas gráficas y procesos estadísticos más avanzados como pueden ser análisis de regresión, análisis de conglomerados (clustering) y análisis de componentes principales (PCA). Tiene como objetivo ayudar a generar una hipótesis sobre los datos y orientar investigaciones futuras.
* Análisis Predictivo. Consiste en identificar patrones históricos y tendencias para posteriormente realizar una predicción que se asemeje con la realidad, toda predicción tiene un error característico. Utiliza técnicas estadísticas y algoritmos de aprendizaje automático (Machine Learning) para construir modelos predictivos.

Cabe destacar que cuantos más datos se estén procesando y mejor sea el algoritmo o el proceso con el que se esté trabajando, más patrones históricos y tendencias serán identificadas, por lo que, significa que se obtendrán predicciones más reales y con menos error. Este tipo de análisis de datos es fundamental para ámbitos como la salud, finanzas, marketing o logística.

* Análisis Prescriptivo. Tiene como objetivo proporcionar recomendaciones y soluciones optimas basadas en datos a través de combinar técnicas de análisis predictivo con optimización matemática con restricciones y objetivo específicos. Se utilizar para argumentar toma de decisiones y optimización de recursos.
* Análisis de texto y Minería de Datos: Se aplica a conjunto de datos no estructurados, como pueden ser textos, reseñas, comentarios en redes sociales. Combina las técnicas de procesamiento del leguaje natural (NLP) y la extracción de grandes conjuntos de datos. Tiene como objetivo entender patrones de comportamientos, de pensamiento u opiniones.

En el proyecto se realizan diferentes tipos de análisis de datos, desde el que nos proporcionar soluciones más simples hasta el que nos muestra resultados más eficientes, con el objetivo de abarcar un marco amplio de resultados. El tipo de análisis que predomina en el proyecto es el de análisis predictivo y la utilización de algoritmos de aprendizaje automático o Machine Learning.

El proyecto estará compuesto de tres modelos estadísticos, donde cada uno está enfocado a una fuente de datos abierta diferentes. Dependiendo de cómo este compuesta la fuente de datos abierta y de los objetivos sobre el análisis se utiliza una técnica de análisis u otra.

Los tipos de análisis realizados son, el análisis descriptivo, el análisis exploratorio donde destaca el análisis de regresión y por último el análisis predictivo con diferentes técnicas de Machine Learning.

Cabe destacar que los procesos de extracción, transformación y carga (ETL) como los procesos de extracción, carga y transformación (ELT) forman parte del concepto de análisis de datos. El proceso ETL tiene una gran importancia en el proyecto.

2.3 Proceso ETL.

El proceso ETL (Extract, Transform y Load) es una metodología utilizada en el análisis de datos para la integración y la preparación de datos de diferentes fuentes de datos o data warehouse. Estos procesos se utilizan para garantizar la calidad, coherencia y disponibilidad de los datos antes de su análisis.

El proceso de ETL está compuesto por diferentes fases, a continuación, se explica cada una ellas:

1. Extracción. En esta fase se extraen los datos de las fuentes de datos, que en el caso del proyecto serán fuentes de datos abiertas. Dependiendo de cómo este presentada la fuente de datos se utiliza unas técnicas de extracción u otras, por ejemplo, la extracción se puede realizar en fuentes de datos SQL, o de ficheros Excel o incluso de documentos PDF.

Se utiliza Python para la realización de las técnicas de extracción, ya que, Python ofrece infinidad de bibliotecas y herramientas que facilitan la extracción de datos, como puede ser Pandas *[4]* o Beatiful Soup *[5]*.

1. Transformación. Una vez que ya se han extraído todos los datos, los datos se someten a una serie de procesos de limpieza, normalización, enriquecimiento y consolidación para garantizar su calidad y coherencia, que será fundamental a la hora de realizar el análisis de datos y la visualización de la información pertinente. Las operaciones más comunes que se suelen realizar en esta fase son operaciones como la eliminación de duplicados, la conversión de formatos, la agregación de datos y la creación de nuevas variables. Para realizar la transformación también se utiliza Python, debido a sus amplias bibliotecas que facilitan la transformación de la información, algunas de estas son Pandas *[4]*, NumPy *[6]* o SciPy.
2. Carga. Una vez que ya se ha producido la fase de extracción y la de transformación se vuelven a cargar estos datos en data warehouse. Un data warehouse es un sistema de almacenamiento de datos que esta creado para un posterior análisis y generación de informes.

El objetivo de esta fase es cargar los datos ya transformados en una estructura de almacenamiento diseñada específicamente para su análisis, para su visualización o su acceso. A través de Python se producirá a la carga de datos en los diferentes sistemas data warehouse.

A continuación, se presenta un esquema del funcionamiento global de un proceso ETL:

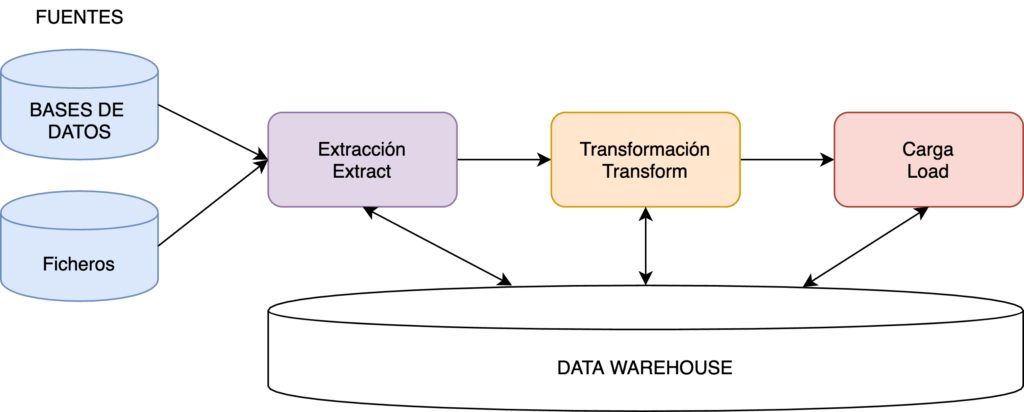


Figura . Esquema del funcionamiento de un proceso ETL.

2.4 Python.

Los de procesos de ETL de las diferentes fuentes de datos abiertas, tanto los diferentes tipos de análisis aplicados a las fuentes de datos se realiza a través del lenguaje de programación Python.

Python es un lenguaje de programación de alto nivel y de propósito general. Se caracteriza por ofrecer una amplia gama de bibliotecas y herramientas que hacen más sencillo multitud de procesos. Relacionados con el proyecto, los procesos de manipulación, procesamiento y análisis de datos.

Python destaca por su evolución en los últimos años, ya que, ha experimentado un crecimiento significativo comparado con los otros lenguajes de programación. La utilización de Python frente a otros lenguajes se debe a las ventajas que propone.

En la *Figura 5* se presenta la evolución en función del tiempo, desde antes del año 2006 hasta el 2022, de los lenguajes de programación más utilizados globalmente. La información esta extraída de “*PYPL [7]*”, donde realizan la gráfica analizando la frecuencia con la que se buscan tutoriales sobre los lenguajes de programación en Google. Como se observa Python se convierte en la más utilizada actualmente con un incide de búsquedas del 27,43% según “*PYPL”*.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura . Evolución de los lenguajes de Programación en función del tiempo.

Se escoge este lenguaje de programación y no otro, debido a las ventajas notables que proporciona para un proyecto de este tipo, las cuales son las siguientes:

* Facilidad de uso. Python desataca por tener una síntesis clara y legible, lo cual, a la hora de manipular y transformar los diferentes datos es fundamental tener un código fácilmente comprensible.
* Amplio Ecosistema de Bibliotecas. Python tiene un amplio ecosistema de bibliotecas, que no para de incrementar. Estas bibliotecas son fundamentales a la hora de proporcionar herramientas para realizar tareas de extracción, transformación y carga (ETL) y las diferentes técnicas de análisis de datos.
* Compatibilidad con las fuentes de datos abiertos. Python puede trabajar con diferentes formatos de datos, lo cual es fundamental para el manejo de datos abiertos. No importa que los datos abiertos se encuentren almacenados en archivos de diferentes tipos como pueden ser CSV, JSON, XML que Python ofrece bibliotecas y herramientas que facilitan el proceso de extracción.
* Integración de ostras tecnologías. Python se integra de forma sencilla con otras tecnologías y software. Se puede combinar con bases de datos, servicios web y otras herramientas. En concreto, es compatible con el software PowerBi, por lo que, se puede realizar la visualización de la información de una forma concisa.
* Soporte de la comunidad. Cuenta con una comunidad inmensa de documentación y de desarrolladores, por lo que, facilita el aprendizaje y desarrollo de la plataforma.

2.5 Visualización en PowerBi.

PowerBi es una herramienta de Business Intelligence desarrollada por Microsoft, la cual permite transformar y visualizar datos en informes interactivos o en paneles de control dinámicos.

Se utiliza el software PowerBi para llevar a cabo la visualización de los resultados obtenidos a través de Python y las fuentes de datos abiertas utilizadas.

Capítulo 3

3. Herramientas a Utilizar.

Para lograr los objetivos del proyecto es necesario la utilización de diferentes tecnologías y herramientas a lo largo del proceso. Cada herramienta está relacionada con las diferentes fases del proyecto. Las fases del proyecto se pueden visualizar en la figura 1.

Para la fase de ETL de las diferentes fuentes de datos y posterior análisis de datos se utiliza Python *[8],* la versión 3.7, como lenguaje de programación y Visual Studio Code *[9]* como entorno de desarrollo. Durante el análisis de datos se van utilizando las siguientes librerías:

* Pandas.
* Fitz.
* XlsxWriter.
* Numpy.
* Datatime.
* Camelot.

Para el almacenamiento de datos durante el análisis de datos se utiliza un almacén de datos (Data Warehouse) utilizando la aplicación Excel *[10]* desarrollada por Microsoft y que forma parte de la suite de Microsoft Office.

Para la fase de visualización se utiliza el software PoweBi *[11].*

Para la fase de documentación, se utiliza la aplicación de procesamiento de texto Microsoft Word *[12]*, que forma parte de la suite de Microsoft Office.

Se utiliza el software en línea LucidChart *[13]* para la realización de diferentes tipos de diagramas, que posteriormente son implementados en Microsoft Word como figuras.

Para alojar, gestionar el proyecto se utiliza la plataforma en la nube GitHub *[14]*, que es una plataforma de desarrollo colaborativa para desarrolladores que permite alojar, gestionar y colaborar con proyectos almacenados y sus posibles versiones del proyecto. El proyecto esta almacenado en un repositorio donde se van gestionando las posibles versiones y las actualizaciones, el repositorio es el siguiente:

GitHub [15]

Capítulo 4

4. Implementación.

En el capítulo 4 se desarrolla el sistema capaz de realizar todos los procesos. En primer lugar, se definen los ámbitos de los que se van a obtener las diversas fuentes datos abiertas de la plataforma, acto seguido se definen que fuentes de datos en concreto se van a utilizar y el porqué de su elección.

Se realiza una investigación sobre las posibles herramientas ETL que Python ofrece, considerando que herramientas funcionan de una manera más optima dependiendo de cómo este constituida la fuente de datos. Dependiendo del punto anterior se aplican unas herramientas ETL u otras a las fuentes de datos, alojando en un data warehouse en el resultado del proceso.

Una vez obtenidas las fuentes de datos abiertas, en las que se han realizado el proceso de limpia y transformación se realiza el análisis a las fuentes de datos escogidas. Se realizan 3 tipos de análisis, análisis descriptivo, análisis exploratorio y análisis predictivo.

Ya con el análisis de datos realizado se visualizan tanto las fuentes de datos abiertas seleccionadas como los resultados del análisis utilizando diferentes técnicas de visualización apoyándose en el software PowerBi.

4.1. Tipos de Fuentes de Datos y Proyectos de Investigación.

A continuación, se presenta los diferentes campos en los que se realiza la investigación de diferentes fuentes de datos abiertas, para una posterior selección. Los diferentes ámbitos son “*Smacite*” y “*Eulogh*”

4.1.1. Smacite.

Smacite *[2]* es un proyecto internacional financiado por la unión europea que tiene como objetivo fundamental abordar los problemas causados por las grandes ciudades en la actualidad. El proyecto está diseñado en función de una serie de objetivos, los cuales derivan de un objetivo común divido en tres partes:

* Abordar el déficit de competencias de los técnicos e ingenieros proporcionándoles formación y educación.
* Mejorar la accesibilidad de los profesionales de las Smart Cities a formaciones de alta calidad a través de herramientas educativas adecuadas y del desarrollo de recursos de aprendizaje.
* Unificar los resultados obtenidos en el aprendizaje y asi fomentar la movilidad en los países de la unión europea a los técnicos e ingenieros de Smart Cities.

Multitud de organizaciones aportan información de valor al proyecto en cuestión, pueden ir de un nivel mas educativo como son universidades u organizaciones propias de la unión europea que están dirigidas a nivel de empresa. Las universidades relacionadas con el proyecto son:

* La universidad de Patras en Grecia.
* La propia universidad de Alcala, España.
* La universidad autónoma de Madrid, España.
* La universidad de West Attica de Atenas, Grecia.
* La politecnica Ikastegia Txorierri de Bizkaia, España.

Para poder entender el proyecto de “Smacite” y entender sus objetivos es necesario saber que es el concepto de ciudad inteligente.

Las ciudades inteligentes utilizan tecnologías TIC para mejorar la calidad de vida de sus habitantes y hacer un uso responsable de los recursos. Consiste en aplicar soluciones tecnológicas y digitales para solventar o mejorar la eficiencia, sostenibilidad, conectividad y calidad de servicios públicos.

Una de las características fundamentales de las Smart cities y parte fundamental del proyecto es que estas ciudades se basan en la recopilación de datos en tiempo real provenientes de diferentes tipos de sensores, cámaras y recopilatorios de datos. Estos datos son una parte fundamental a la hora de tomar decisiones informadas, para la optimización del funcionamiento de la ciudad y de la mejora de la calidad de vida del ciudadano.

La relación entre el concepto de Open Data y las ciudades inteligentes se produce cuando los datos abiertos se utilizan para rellenar los sistemas de información y análisis. Al recabar información y presentarla de una manera accesible los gobiernos y otras organizaciones permiten que diferentes desarrolladores, investigadores y ciudadanos utilicen estos datos para realizar análisis y crear aplicaciones.

La recopilación de información en tiempo real y básicamente el concepto de Smart Cities abarca multitud de temas, como pueden ser demografía, educación, medioambiente, salud, empleo, economía, …

Existen multitud de ciudades en la actualidad que se pueden denominar ciudades inteligentes, incluso en nuestro país el concepto de ciudad inteligente ya abarca casi todas las ciudades, pero destaca por encima del resto la ciudad de Barcelona, ya que ocupa el 3 puesto del ranking de ciudades inteligentes del mundo según el estudio realizado por [Juniper Research](https://www.juniperresearch.com/press/worlds-no-1-smart-city-for-2022-shanghai) en 2022, siendo la primera en el ranking la ciudad de Shanghái.

Para el proyecto el concepto de las ciudades inteligentes tiene una gran importancia, ya que nos permite, investigar y estudiar infinidad de fuentes de datos abiertas sobre una multitud de campos diferentes, los cuales pueden tener objetivos muy dispares. Gracias a estas ciudades, como las que hay en España o en toda la unión Europa se han obtenido las fuentes de datos abiertas que han hecho posible la creación de este proyecto.

Las fuentes de datos abiertas que se han empleado en el proyecto se han extraído de tres organizaciones diferentes, las cuales dos de ellas se fundamental en el concepto de ciudades inteligentes. Las dos organizaciones son:

* Por el gobierno de España.
* Por la unión europea, que son proporcionados por todos los países pertenecientes.
* Por el ayuntamiento de la ciudad de Madrid.
* Por el instituto nacional de estadística (INE).

4.1.2. EULOGH.

Eulogh *[1]* tiene el significado de “Alianza Universitaria Europea para la Salud Global”. Se caracteriza por ser un proyecto financiando por la Unión Europea que tiene como objetivo la creación un gran conjunto de datos que se puedan utilizar para analizar y estudiar para una posterior aplicación en el mundo real que sirva para la optimización o la finalización de problemas. Esta gestionado por un consorcio de 9 universidades europeas, en las que pertenece la propia universidad de Alcalá, de ahí el interés de contribuir al proyecto de Eulogh. Las demás universidades que participan en el proyecto son las siguientes:

* La universidad de Paris-Saclay, de Francia
* La universidad de Lund, de Suecia.
* La universidad de Szeged, de Hungría.
* La universidad de Oporto, de Portugal.
* La universidad de Múnich, de Alemania.
* La universidad del artico, de Noruega.
* La universidad de Novi Sad, de Serbia.
* La universidad de Hamburgo, de Alemania.

Gráfico

Descripción generada automáticamente

Figura . Mapa de las universidades relacionadas con Eulogh.

El proyecto de Eulogh tiene una serie de objetivos, donde algunos de ellos presentan similitudes al proyecto realizado. Estos objetivos que destacan son los siguientes:

1. Proporcionar a los expertos en el mundo de la salud un conocimiento y unas herramientas extras que ayuden u optimicen su labor de cara al ciudadano.
2. Contribuir al espacio europeo de educación, al espacio europeo de investigación y al espacio europea de salud.
3. Actuar como fuerza propulsora a la hora de resolver diferentes aspectos de la salud global, que van desde la salud pública, las enfermedades emergentes, el cambio climático y los peligros ambientales.

Los dominios que abarca este proyecto tienen una gran similitud entre los dominios que abarcan las Smart cities. El concepto de Smart cities del punto anterior y el proyecto de Eulogh están altamente relacionados, ya que, las fuentes de datos proporcionadas que se utilizan en el proyecto están suministradas por las diferentes ciudades inteligentes de la unión europea.

Los campos que abarca el proyecto de Eulogh contribuyen a una mejor compresión del desafío de social de la salud global. Estos campos se pueden ver representados en la *Figura 7*, la imagen de la *Figura 7* se ha obtenido de la página oficial del Eulogh.

Escala de tiempo

Descripción generada automáticamente con confianza media

Figura . Campos que aborda el proyecto Eulogh.

Como se puede observar los campos que aborda el proyecto de Eulogh están estrechamente relacionados con la salud global. Tambien implica la relación de la tecnología con la salud, la ingeniería, la inteligencia artificial, el Big Data y la robótica con herramientas que se utilizan para lograr sus objetivos. Tal como pasa en el proyecto que se está desarrollando.

Teniendo en cuenta los objetivos y los campos que abarca Eulogh, mi proyecto personal puede aportar gran valor y funcionalidad. Ya que los proyectos proponen objetivos similares y los campos que se abordan son los mismos. En los siguientes puntos se explica cada una de las fuentes de datos abiertas que se han escogido y sus razones, pero una de las razones de algunas de ellas por la que se han seleccionado es para abarcar los mismos temas de salud global de Eulogh.

4.2. Fuentes de Datos Abiertas.

Se seleccionan diferentes fuentes de datos abiertas, a través de las diferentes páginas web que proporcionan las organizaciones gubernamentales. Se ha realizado una investigación de multitud fuentes de datos y se han seleccionado unas pocas para la realización del proyecto, la selección se ha realizado en función del cumplimiento de una serie de propiedades, las propiedades son las siguientes:

* En función del tamaño. Para poder trabajar con un conjunto de datos debe de tener un tamaño considerable, para que a la hora de realizar el análisis se obtengan resultados de valor.
* En función de la calidad de los datos. La base de datos debe de tener datos que se ajusten a la realidad y que se estén almacenados siguiendo el formato que con el cual esta presentado en la fuente de datos. Como las fuentes de datos se obtienen a través de instituciones gubernamentales suelen presentar una alta calidad en los datos.
* En función del tipo archivo de almacenamiento. Dependiendo con qué tipo de archivo estén almacenados los datos tendrán una serie de ventajas o desventajas suponiendo una mayor facilidad de acceso de e integración con el lenguaje de programación Python, el cual va a ser el encargado de realizar la transformación y posterior análisis.
* En función de la relación con el proyecto Eulogh. Como el proyecto que se está realizando tiene como objetivo aportar funcionalidad al proyecto europeo de Eulogh, es necesario que las fuentes de datos seleccionadas estén relacionadas con los temas que se tratan en Eulogh, que estén relacionadas con todo lo que representa la salud global.
* En función del grado de interés. Un aspecto que también es importante para seleccionar cada una de las fuentes de datos abiertas, es el interés de conocimiento que producen, tanto para el autor del proyecto como para sus lectores. Durante la investigación se buscan fuentes de datos que sean atractivas.

Tambien es necesario seleccionar las fuentes de datos en función de la facilidad de acceso a los datos, en función del costo de los datos y en función de la licencia de los datos. Pero estas tres características se solventan al ser fuentes de datos abiertas, ya que, una de las principales características del “Open Data” es que son fácilmente accesibles, no tienen ningún tipo de costo y tiene una licencia libre donde se pueden usar los datos para cualquier propósito.

Siguiendo estas propiedades se seleccionan las siguientes fuentes de datos abiertas, teniendo en cuenta en que formato se almacenan la información:

1. El histórico de incendios producidos en España entre los años 2006 hasta 2015. En formato PDF.
2. El histórico de concesiones de nacionalidad española, considerando su nacionalidad de origen y el sexo (masculino o femenino) entre los años 2010 y 2019. En formato xlsx.
3. La capacidad asistencial en España durante la crisis del Covid-19 en España. En formato CSV.
4. El catálogo de parques municipales en la ciudad de Madrid. En formato PDF.
5. La siniestralidad producida en las carreteras de la ciudad de Madrid, durante el año anterior (enero del 2022 hasta diciembre del 2022). En formato xlsx.
6. El histórico de la población de las provincias de España entre los años 1996 hasta 2021 y también el histórico en función del sexo (masculino o femenino). En formato xlsx.
7. El histórico de la población en los países pertenecientes a la unión europea entre los años 2001 hasta el 2022. En formato xlsx.
8. Catalogo del bosque urbano de la ciudad de Madrid. En formato PDF.

A continuación, se explican cada una de las fuentes de datos seleccionadas, como están constituidas, el porqué de su selección y la complejidad que alberga realizar un análisis.

4.2.1. Incendios producidos en España entre el 2006 y el 2015.

Esta fuente de datos está suministrada por el gobierno de España, más concretamente por el ministerio de agricultura, pesca y alimentación. Consiste en un archivo PDF en el que en su interior se encuentran diferentes tablas con información relevante, por ejemplo, en el Cuadro 2.1 del PDF, podemos encontrar una tabla con el número de siniestros y superficies afectadas del 1968 al 2015. En el PDF se explican multitud de aspectos relacionados con los incendios en España. Este tipo de fuente de dato pertenece a la categoría de biodiversidad y el enlace a la fuente de datos es el siguiente *[16]*:

[Incendios en España 2006-2015](https://www.miteco.gob.es/es/biodiversidad/temas/incendios-forestales/incendios-decenio-2006-2015_tcm30-521617.pdf).

<https://www.miteco.gob.es/es/biodiversidad/temas/incendios-forestales/incendios-decenio-2006-2015_tcm30-521617.pdf>

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* Por el motivo de interés que presentar determinar si existe relación entre los diferentes incendios que se han producido en nuestro país, y si es posible realizar un análisis que concluya la predicción de nuevos incendios o determinar ciertas zonas como zonas rojas, ya que, presentan mayor probabilidad de presenciar un incendio en un determinado espacio de tiempo.
* Porque el formato en el que se encuentra los datos es en un archivo PDF, por lo que, supone una mayor dificultad y por lo tanto un reto, extraer la información de valor, para más tarde proceder a un proceso de transformación y carga. Es una fuente de datos no estructurada, por lo que alberga un incremento de dificultad realizar los procesos de análisis.
* Por el cumplimiento de los objetivos del proyecto europeo Eulogh. Al realizar un análisis de este tipo de fuente de datos, se está aportando información de valor al proyecto, ya que, los incendios que se producen en España es uno de los ámbitos que se trata, más concretamente en el campo de clima, medio ambiente y biodiversidad, tal como podemos observar en la Figura 6.

A continuación, se muestra una tabla extraída del documento PDF que sirve como ejemplo para mostrar cómo está definida la información de valor, por tanto, la información que se necesita realizar una extracción. En la tabla se muestra un resumen de siniestros y superficies afectadas entre los años 2006 al 2015.

Tabla

Descripción generada automáticamente

Figura . Ejemplo de tabla extraída de la fuente de datos “Incendios producidos en España entre el 2006-2015”

4.2.2. Concesión de Nacionalidad Española entre el 2010 y el 2019.

Esta fuente de datos está suministrada por el gobierno de España, más concretamente esta suministrada por el ministerio español de inclusión, seguridad social y migraciones. Consiste en un archivo en formato xlsx, por lo tanto, es una fuente de datos estructura, donde los datos presentan una gran calidad. En el archivo se muestra información en función de la nacionalidad previa del ciudadano y en función del sexo (masculino o femenino). La fuente de datos abierta pertenece a la categoría de demografía y la podemos encontrar en el siguiente enlace *[17]*:

[Concesiones de Nacionalidad.](https://extranjeros.inclusion.gob.es/es/Estadisticas/operaciones/concesiones/index.html)

<https://extranjeros.inclusion.gob.es/es/Estadisticas/operaciones/concesiones/index.html>

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* Por el motivo de interés cultural. Es interesante saber de qué países hay más inmigración y como está relacionada y si es posible encontrar algún patrón o alguna conclusión de valor al realizar en análisis.
* Al ser el formato de la fuente de datos un xlsx es más simple realizar el proceso de extracción para posteriormente realizar proceso de transformación y carga. Es interesante abarcar todo tipo de formatos a la hora de realizar la extracción, de los más simples a los más complejos.

A continuación, se muestra un ejemplo de cómo se presenta la información en el archivo xlsx, corresponde a la primera página del archivo. Como se puede observar en la Figura 8, el histórico de concesiones de nacionalidad se presenta en función del país del que sea previamente residente el ciudadano, donde cada país esta divido por continente. También se presenta la información en función del sexo. Al estar la información presentada en formato tabla, al realizar el proceso de extracción utilizando Python es un proceso sencillo.

Tabla

Descripción generada automáticamente

Figura . Ejemplo de tabla extraída de la fuente de datos “Concesión de Nacionalidad entre el 2010 y el 2019”.

4.2.3. Capacidad Asistencial durante la Covid-19.

Esta fuente de datos está suministrada por el gobierno de España, más concretamente por el ministerio de sanidad. Consiste en un archivo CSV, en el que se muestra un histórico de datos entre el 1 de agosto de 2020 hasta la actualidad mostrando la capacidad asistencial frente al Covid-19 en función de provincias españolas. Es una fuente de datos estructurada, donde se caracteriza por la abundancia de datos (137.851 filas) y la calidad que presentan. La fuente de datos abierta pertenece a la categoría de salud y la podemos encontrar en el siguiente enlace *[18]*:

[Capacidad asistencial.](https://www.sanidad.gob.es/areas/alertasEmergenciasSanitarias/alertasActuales/nCov/situacionActual.htm)

<https://www>.sanidad.gob.es/adri/alertasEmergenciasSanitarias/alertasActuales/nCov/situacionActual.html

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* La crisis del Covid-19 tuvo un gran impacto que abarco todos niveles de la sociedad, por lo tanto, es interesante observar cómo va evolucionando la enfermedad vista desde una perspectiva de capacidad asistencial de nuestros hospitales. Además, es importante analizar el histórico para ver relaciones ocultas o descubrir nuevos conocimientos que nos ayuden a entender que realmente paso durante esta crisis.
* Al ser el formato de la fuente de datos un CSV también es más simple el proceso extracción. Es importante poder abarcar cuantos más tipos de archivos mejor.
* El tema del Covid-19 está estrechamente relacionado con los objetivos que quiere abarcar el proyecto europeo de Eulogh, por lo que, este análisis y las soluciones obtenidas implementan un gran valor al proyecto.

A continuación, se muestra un ejemplo de cómo se presenta la información en el archivo xlsx, la información se muestra en función de la provincia española y del día. Se muestra las camas ocupadas en la UCI, las camas ocupadas que tienen relación con el Covid-19 y los ingresos y altas de cada día. Esta información no se muestra como tabla, sino como histórico, por lo que su extracción es simple por ser un archivo xlsx, pero su transformación tiene un nivel más de complejidad.

Tabla

Descripción generada automáticamente

Figura . Ejemplo del histórico extraído de la fuente de datos “Capacidad Asistencial durante el Covid-19”.

4.2.4. Catálogo de Parques Municipales en la Ciudad de Madrid.

Esta fuente de datos está suministrada por el ayuntamiento de la Madrid, más concretamente por el ministerio de medio ambiente y movilidad. Consiste en un archivo PDF, en el que cada página de valor se muestra información correspondiente a un parque de la comunidad de Madrid. Se trata de una fuente de datos no estructurada donde la información que se expone se expone sin ningún tipo de estructura. Se trata de una fuente de datos abundante ya que nos encontramos con un PDF de 206, en donde cada parte tiene más de 60 propiedades. La fuente de datos abierta pertenece a la categoría de medio ambiente y la podemos encontrar en el siguiente enlace *[19]*:

[Catálogo de Parques Municipales.](https://www.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ContenidosTemporales/Cat%C3%A1logoParques/CATALOGO%20DE%20PARQUES%20MUNICIPALES%20MADRID_BAJA%20RESOLUCI%C3%93N%2003.08.2021.pdf)

<https://www>.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ContenidosTemporales/Cat%C3%A1logoParques/CATALOGO DE PARQUES MUNICIPALES MADRID\_BAJA RESOLUCI%C3%93N 03.08.2021.pdf

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* El aumento de la población en las grandes ciudades y como realizar estas ciudades los más “verdes” y sostenibles posibles en un reto de gran prioridad en la actualidad. La distribución de los diferentes parques a lo largo de la ciudad y como están constituidos es fundamental para abordar este tema. Es importante realizar un análisis que pueda aportar más conocimientos para una posterior toma de decisiones enfundada u optimización de los recursos utilizados en las ciudades.
* Al ser el formato de la fuente de datos un PDF mostrando la información en forma de texto sin la utilización en forma de tabla, hace que el método de extracción sea lo más complejo posible. Esta es una razón para elección de la base de datos porque se busca mostrar la capacidad de extracción de datos de todo tipo de fuentes de datos y como se encuentren estructuradas
* El tema de los parques y como están constituidos y sus propiedades abarca uno de los principales temas del proyecto internacional Eulogh, por lo tanto, es una razón de peso para la selección de la fuente de datos.

A continuación, se muestra un ejemplo de cómo está distribuida la información en el archivo PDF. En la *Figura 11* se puede observar la página 178 del PDF, donde muestra las propiedades del parque municipal del “Jardines del Buen Retiro”. Como se puede observar la información muestrada, se expone sin utilización de ninguna estructura, por lo que, el proceso de extracción es muy complejo, pero a través de la Python y sus multitudes de bibliotecas se puede realizar el proceso de extracción.

Una captura de pantalla de una computadora

Descripción generada automáticamente con confianza media

Figura . Ejemplo de cómo se expone la información en la fuente de datos “Catalogo de Parques Municipales de Madrid”.

4.2.5. Siniestralidad en la Carreteras en la Ciudad de Madrid.

Esta fuente de datos abierta esta suministrada por el gobierno de España, más concretamente por el ministerio de asuntos económicos y transformación digital. Consiste en archivo xlsx donde se muestra un histórico de siniestros en las carreteras de la ciudad de Madrid en el año 2022, donde cada siniestro se corresponde con una fila del archivo. Se trata de una fuente de datos estructurada con una cantidad de datos importante, en concreto existen 47.054 filas que es igual a 47.054 siniestros en el año de 2022. La fuente de datos abierta pertenece a la categoría de transporte y la podemos encontrar en el siguiente enlace *[20]*:

[Accidentes de Tráfico de la Ciudad de Madrid](https://datos.gob.es/en/catalogo/l01280796-accidentes-de-trafico-de-la-ciudad-de-madrid1)

<https://datos>.gob.es/en/catalogo/l01280796-accidentes-de-trafico-de-la-ciudad-de-madrid1

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* El interés que causa saber el por qué y donde se producen más accidentes de tráfico en la ciudad de Madrid y en un posterior análisis, encontrar relaciones ocultas o conclusiones e incluso predicciones en la que su valor obtenido puedan tener una implicación real en la vida de los ciudadanos.
* Al ser el formato de la fuente de datos un xlsx es más simple realizar el proceso de extracción para posteriormente realizar proceso de transformación y carga.

A continuación, se muestra un ejemplo de cómo está constituida la fuente de datos en el archivo xlsx. Cada siniestro de se corresponde a una fila de fuente de datos, cada siniestro tiene una serie de propiedades donde destacan la ubicación, el tipo de vehículo, el rango de edad, el sexo y si ha dado positivo en alcohol o drogas, tal y como se puede observar en la *figura 12*.

Texto

Descripción generada automáticamente con confianza media

Figura . Ejemplo de la composición de la fuente de datos “Siniestralidad en las Carreteras en la Ciudad de Madrid”.

4.2.6. Población por Provincias de España 1996-2021.

Esta fuente de datos esta suministrada por el instituto nacional de estadística (INE) *[21]*. Consiste en un archivo xlsx, en donde cada fila se corresponde a una provincia de España, recoge la población de cada provincia desde el año 1996 hasta el 2021. También la fuente de datos está dividida en función de sexo (masculino o femenino). Es una fuente de datos estructurada por lo que, proceso de extracción no tiene mucha complejidad. La fuente de datos la podemos encontrar en el siguiente enlace *[22]*:

[Población Por Provincias de España](https://www.ine.es/jaxiT3/Tabla.htm?t=2852).

<https://www.ine.es/jaxiT3/Tabla.htm?t=2852>

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* En la actualidad existe un problema muy serio en España con la despoblación de ciertas zonas rurales y como toda esta se está moviendo a las ciudades importantes, por ello es interesante realizar un análisis sobre los datos y observar cómo van evolucionando cada una de las provincias. También a través del análisis se pueden obtener conclusiones nuevas e incluso realizar una predicción de cómo será la evolución en los siguientes años.
* Al ser el formato de la fuente de datos un xlsx es más simple realizar el proceso de extracción para posteriormente realizar proceso de transformación y carga.
* Esta fuente de datos pertenece a la categoría de demografía. No solo España es el que está sufriendo este problema de despoblación de las zonas rurales, sino, que es un problema común en algunas ciudades europeas. Por lo tanto, esta fuente de datos y su posterior análisis puede contribuir al proyecto europeo de Eulogh

A continuación, se muestra un ejemplo de cómo está constituida la fuente de datos en el archivo xlsx. Las filas de la fuente de datos se corresponden a cada una de las provincias españolas y las columnas a los años, tal y como se puede observar en la *Figura 13*.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Figura . Ejemplo de la fuente de datos “Población por Provincias de España”.

4.2.7. Población por Países en la unión europea 2001-2022.

Esta fuente de datos esta suministrada por la oficina estadística oficial de la unión europea (adridt) *[23]*. Consiste en un archivo xlsx en donde se muestran la población en cada país perteneciente a la unión europea desde el año 2021 hasta el 2022. Se trata de una fuente de datos estructurada, donde el proceso de extracción a través de Python tiene una complejidad sencilla. La fuente de datos la podemos encontrar en el siguiente enlace *[24]*:

[Población por Países en Europa.](https://ec.europa.eu/eurostat/databrowser/view/DEMO_GIND__custom_2733962/settings_1/table?lang=en&bookmarkId=7084ed24-6b91-4cf3-b90d-d47565593505)

<https://ec>.europa.eu/adridt/databrowser/view/DEMO\_GIND\_\_custom\_2733962/settings\_1/table?lang=en&bookmarkId=7084ed24-6b91-4cf3-b90d-d47565593505

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* Es interesante observar cómo van creciendo los diferentes países de la unión europea en términos de población.
* Al ser el formato de la fuente de datos un xlsx es más simple realizar el proceso de extracción para posteriormente realizar proceso de transformación y carga.
* El análisis y el estudio de esta fuente de datos, comparte objetivos con el proyecto europeo Eulogh, por lo tanto, este estudio sirve de implementación a este proyecto más grande.

A continuación, se muestra un ejemplo de cómo está constituida la fuente de datos en el archivo xlsx, tal y como se puede observar en la *Figura 14*.

Tabla

Descripción generada automáticamente

Figura . Ejemplo de la fuente de datos “Población por Países de la unión Europa”.

4.2.7. Catálogo del Bosque Urbano de la Ciudad de Madrid.

Esta fuente de datos está suministrada por el ayuntamiento de la ciudad de Madrid. Consiste en un archivo PDF, donde expone la situación de los bosques de la ciudad de Madrid, explicando sus propiedades, como pueden ser explicar las diferentes especies de árboles que hay, el beneficio de los árboles para la ciudad, el almacenamiento de carbono.

Esta fuente de datos destaca por tener una estructura de datos semiestructurada, ya que, los datos se presentan con tablas, gráficos y demás imágenes, pero no es estructurada porque presenta la información usando diferentes herramientas. Al final del documento se expone una tabla de gran dimensión sobre el listado de especies (página 64), que contiene gran valor para el documento. La fuente de datos la podemos encontrar en el siguiente enlace *[25]*:

[Bosque Urbano de Madrid.](https://www.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ValorBosqueUrbanoMadrid/Valor%20Bosque%20Urbano%20de%20Madrid.pdf)

<https://www>.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ValorBosqueUrbanoMadrid/Valor Bosque Urbano de Madrid.pdf

Se ha seleccionado esta fuente de base de datos abierta por las siguientes razones:

* Tal y como se exponía en el punto 4.2.4 con el catálogo de parques municipales de Madrid, es fundamental tener en las ciudades cierta vegetación que ayude a la contaminación de la ciudad y de calidad de vida a la población. En este documento además de eso se expone cómo funcionan y que recursos necesitan cada una de las especies de los bosques, por lo tanto, al realizar un análisis se pueden logran conclusiones y relaciones que fundamenten el tipo de especia a utilizar en cada momento para ayudar al problema de contaminación u optimizar los recursos de la población.
* Al ser el formato de la fuente de datos un archivo PDF donde tiene una estructura semiestructura, dependiendo de que gráfico, tabla o imagen se quiera extraer la información va a tener un proceso u otro, lo que significa que van a tener diferentes grados de dificultad, por lo que, se abarca más posibilidades.
* Este tema está estrechamente relacionado con el tema de Smart City, por lo que, es una razón más de peso para seleccionar esta fuente de datos. Como se ha dicho anteriormente realizar un correcto análisis de datos que aparecen en el documento puede suponer una optimización de los recursos y aumento de la calidad de vida de la población.

A continuación, se muestra un ejemplo de cómo está constituida la fuente de datos en el archivo PDF. Dependiendo de la página en donde se quiera extraer la información se usa un proceso diferente porque, puede que la información ente expuesta en gráfico, en tabla o en imagen, incluso en una misma página puede estar mostrada utilizando varios métodos, tal y como se puede observar en la *Figura 15*. La *Figura 15* se corresponde a la página 51 del archivo PDF

Imagen que contiene Gráfico

Descripción generada automáticamente

Figura . Ejemplo de la exposición de la información de la fuente de datos “Valor del Bosque Urbano de Madrid”, de la página 51 del PDF.

4.2.8. Calificación de las diferentes fuentes de datos abiertas.

Como se ha puesto en los puntos anteriores, cada fuente de datos abierta seleccionada tiene una serie de propiedades por las que han sido seleccionadas para el proyecto. Donde las propiedades más características son las siguientes:

* Dependiendo del tipo de tema que abarquen (Smart Cities, Eulogh o los dos).
* Dependiendo del tipo de estructura que tenga la fuente de datos (estructural, no estructural o semi estructural).
* Dependiendo del tipo proceso de extracción que se aplica a la fuente de datos utilizando el lenguaje de programación Python (grado de complejidad).
* Dependiendo del tipo de archivo de donde estén almacenados las diferentes fuentes de datos abiertas (PDF, xlsx, CSV).
* Dependiendo de que organismo gubernamental haya puesto la fuente de datos a disposición de todos los ciudadanos (gobierno de España, INE, Eurostat, ayuntamiento de Madrid).

Con esta serie de propiedades se puede calificar cada una de las fuentes de datos seleccionadas, por ello, se constituye la siguiente tabla:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nombre** | **Temática** | **Tipo de Estructura** | **Complejidad de Extracción** | **Tipo de Archivo** | **Suministrador** |
| Incendios producidos en España entre el 2006 y el 2015. | Eulogh. | Semi Estructurada | Media | PDF | Gobierno de España |
| Concesión de Nacionalidad Española entre el 2010 y el 2019 | Eulogh. | Estructurada | Baja | xlsx | Gobierno de España |
| Capacidad Asistencial durante la Covid-19 | Eulogh. | Estructurada | Baja | CSV | Gobierno de España |
| Catálogo de Parques Municipales en la Comunidad de Madrid | Smart Cities y Eulogh | No estructurada | Alta | PDF | Ayuntamiento de Madrid |
| Siniestralidad en la Carreteras en la Ciudad de Madrid | Smart Cities | Estructurada | Baja | xlsx | Gobierno de España |
| Población por Provincias de España 1996-2021 | Smart Cities y Eulogh | Estructurada | Baja | xlsx | INE. |
| Población por Países en la unión europea 2001-2022 | Eulogh | Estructurada | Baja | xlsx | Eurostat |
| Catálogo del Bosque Urbano de la Ciudad de Madrid | Smart Cities y Eulogh | Semi Estructurada | Media | PDF | Ayuntamiento de Madrid |

Tabla . Calificación de las diferentes fuentes de datos seleccionadas.

4.3. Proceso ETL en las Fuentes de Datos.

Una vez seleccionada las fuentes de datos abiertas, se seleccionan en este caso 6 fuentes de datos de las anteriores a las que aplicar este proceso de extracción, transformación y carga. Se seleccionan seis fuentes de datos, ya que, muchas de ellas presentan muchas similitudes a la hora de realizar el proceso, y no aportan información añadida de valor al proyecto.

Las fuentes que se seleccionan para aplicar el proceso y el porqué de su selección son las siguientes:

1. Catálogo de Parques Municipales en la Comunidad de Madrid. Se selecciona esta fuente de datos debido a la complejidad y el reto que supone aplicar el proceso ETL en una fuente de datos no estructura, aportando conocimientos de valor al proyecto, abarcando todo tipo de estructuras de fuentes de datos.
2. Capacidad Asistencial durante la Covid-19. Se selecciona para ver, aunque tenga menos complejidad, como se produce el proceso ETL en una fuente de datos estructurada, además siendo el tipo de archivo un CSV.
3. Incendios producidos en España entre el 2006 y el 2015. Se selecciona esta fuente de datos para realizar el proceso ETL en una fuente de datos semi estructura y ya así, abarcar todas las estructuras posibles de fuentes de datos.
4. Concesión de Nacionalidad Española entre el 2010 y el 2019. Se selecciona esta fuente de datos para abarcar también más tipos de archivos, en este caso un archivo xlsx.
5. Catálogo del Bosque Urbano de la Ciudad de Madrid. Se selecciona esta fuente de datos también debido a su estructura semi estructurada y porque engloba el concepto de “Smart Cities” y el proyecto europeo “Eulogh”.
6. Población por Provincias de España 1996-2021. Se selecciona esta fuente de datos por el interés y la problemática actual sobre la pérdida de población en la zona rurales.

4.3.1. Proceso ETL en Catálogo de Parques Municipales en la ciudad de Madrid.

Como se ha podido observar en el punto 4.2.4 esta fuente de datos abiertas no tiene ningún tipo de estructura que defina la clasificación de la información, en cada página del PDF la información se muestra de una forma diferente y en una posición diferente por ello el proceso de ETL tiene una mayor complejidad. EL proceso ETL se ha llevado a cabo a través de Python y la utilización de las siguientes librerías:

* Fitz *[26]*. Esta librería tiene un grandísimo poder. A través de esta librería se puede manipular archivos PDF dando la posibilidad de realizar las siguientes acciones:
  + Leer texto de un archivo PDF, se puede leer todo el texto del archivo o indicar cual quieres concretamente. Para ello es necesario indicar las coordenadas de donde se quiere extraer el texto e indicar la página de donde se quiere extraer.
  + Extraer imágenes del PDF. Funciona igual que el punto anterior, indicando las coordenadas de la imagen y la pagina en cuestión.
  + Convertir archivos PDF a otros formatos.
  + Introducir imágenes o texto al archivo PDF.
  + Firmar digitalmente un archivo.
  + Encriptar el archivo con una contraseña.
* XlsxWriter *[27]*. Esta librería sirve para generar archivos de Excel en formato xlsx. Permite crear el archivo, escribir en él, crear gráficos y tablas y guardar una hoja de cálculo.

Lo primero que se hace en el programa desarrollado es abrir el archivo PDF utilizando la biblioteca *Fitz*, para poder usar todas sus ventajas. Después a partir de la página 7, que es donde empieza mostrarte la información de valor, y teniendo en cuenta que la pagina sea impar, ya que en todas las páginas impares es donde se muestra la información mientras que en las pares solo aparecen imágenes de los diferentes parques, se obtiene todo el texto de esa página.

El texto de cada página se obtiene a partir de las coordenadas de en la que se encuentra la información relevante, se sabe que coordenadas son ya que la librería *Fitz* lo indica. Utilizando las coordenadas no se extrae todo el texto de la página, sino la información que se considera relevante. Con todo el texto relevante de la página guardado en una lista se quitan los puntos y asteriscos que son innecesarios. Esta funcionalidad se implementa con la función *get\_text*.

Texto

Descripción generada automáticamente

Código . Función get\_text del proceso de extracción de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”.

Ya extraído el texto de cada página se procede a la limpieza, se borra toda la información que no es relevante y aun así se ha extraído y se elimina multitud de elementos que se han almacenado sin que fueran parte del texto. Después se quitan multitud de caracteres que estropean los datos para poder trabajar con ellos, por ejemplo, son *km*, *ud*, *m^2*, … Cuando se han limpiado la información de la página se repite el proceso extracción y limpieza con todas las páginas que albergan información para después guardar toda información limpia de todas las páginas en una lista.

Texto

Descripción generada automáticamente

Código . Función clean del proceso de transformación de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”.

Una vez ya realizado los procesos de extracción y transformación, se procede al proceso de carga de los datos en el data warehouse. Para ello se utiliza la función *export\_xlsx*, en la que se definen las cabeceras de las columnas del archivo xlsx en el que se va a guardar la información, a continuación, se crea el archivo xlsx y se introduce las cabeceras y posterior mente todo el texto limpio de todas las páginas del PDF. Para tratar con el archivo xlsx se usa la librería *XlsxWriter*.

Texto

Descripción generada automáticamente

Código . Función export\_xlsx del proceso de carga de la fuente de datos “Catalogo de Parques Municipales de la Ciudad de Madrid”.

Ya realizado el proceso ETL en la fuente de datos, parte del data warehouse se visualiza de la siguiente manera, donde una fila del documento se corresponde con una de las páginas como la que se puede visualizar en la *Figura 11*.

Tabla

Descripción generada automáticamente

Figura . Parte del data warehouse una vez realizado el proceso ETL.

4.3.2. Proceso ETL en Capacidad Asistencial durante la Covid-19.

Consiste en una fuente de datos estructura, en donde alberga muchísimos datos. Se realiza el un proceso ETL para obtener la información que se ajuste a nuestros objetivos y asi reducir el tamaño de los datos que se están tratando. En el archivo se tiene en cuenta todas las comunidades autónomas de España, pero solo se va a tener en cuenta la comunidad de Madrid. Tambien solo se va a considerar la fecha, las camas totales que estén disponibles y cuantas camas están ocupadas por enfermos de la Covid-19.

Aunque se trate de una fuente de datos estructura el proceso ETL, no va a ser sencillo, ya que hay multitud de datos que se van a desechar y la información de valor está escondida entre tanto dato.

Para la realización del proceso ETL ha sido necesario la implementación de las siguientes librerías:

* Pandas *[4]*. Es una librería famosa para el análisis de datos. Permite estructurar datos, manipularlos, analizarlos y visualizarlos usando gráficos y tablas. Se caracteriza por ser una herramienta muy flexible.
* NumPy *[6]*. Es una librería que se utiliza para el procesamiento numérico y científico de datos. Tiene multitud de acciones de gran relevancia, las mas famosas son las siguientes:
  + La principal característica de NumPy es el “*ndarray*”, es una matriz multidimensional homogénea y de tamaño fijo, que, a diferencia de las listas de Python, los arrays de NumPy son mejores para almacenar grandes conjuntos de datos y se pueden realizar operaciones vectorizadas, lo que provoca un mejor rendimiento.
  + Tiene una amplia gama de funciones matemáticas que se pueden aplicar a los arrays, están optimizadas para trabajar sobre un gran conjunto de datos.
* Datetime *[28]*. Proporciona una serie de clases y funciones para trabajar con todo tipo de formatos de fechas y horas. Facilita la manipulación, formateo y calculo de fechas y horas.

El programa desarrollado en Python esta divido en diferentes fases, donde cada fase se corresponde con los procesos de extracción, transformación y carga.

Lo primero que se realiza, es la importación del archivo xlsx a través de la librería *Pandas*. Después se procede con la extracción de la información de valor.

Primero se extrae todos los datos correspondientes con la comunidad de Madrid y en función de un rango de fechas, para extraer la información según el rango de fechas se utiliza la librería *Datetime*. El rango de fechas es del 1 de agosto del 2020 hasta el 1 de enero del 2021. Se escoge este rango debido a que fue el punto más álgido de la pandemia, por lo tanto, el punto donde mas datos interesantes existen. Este proceso se hace a través de la función *getByCCAAByData*, recorre todos los datos importados y almacena en una nueva lista los datos que estén en el rango de fechas y que cumplan la condición de ser de la comunidad de Madrid.

Después utilizando la función *getData* se extrae la información correspondiente a la fecha, al total de camas y las camas ocupadas por paciente de Covid-19, generando tres listas de igual tamaño.

Texto

Descripción generada automáticamente

Código . Proceso de extracción de la fuente de datos “Capacidad Asistencial durante la Covid-19”.

Ya con toda la información relevante extraída se procede a carga en un nuevo archivo xlsx, se utiliza la función *export\_xlsx* para llevar esta función a cabo. Primero en la función se pasa las fechas a un formato que se pueda introducir en el archivo Excel, después se unifican las tres listas (fechas, camas totales, camas ocupadas) formando una matriz con toda la información. Para ello se utiliza la librería *Numpy*. A continuación, se procede a cagar la matriz con las cabeceras en el archivo xlsx.

Texto

Descripción generada automáticamente

Código . Proceso de carga de la fuente de datos “Capacidad Asistencial durante la Covid-19”.

Se muestra en la Figura – un ejemplo de cómo ha quedado almacenada parte de la información tras el proceso ETL en el data warehouse.

Tabla

Descripción generada automáticamente

Figura . Resultado del Proceso ETL de la fuente de datos “Capacidad Asistencial durante la Covid-19”.

4.3.3. Proceso ETL en Incendios producidos en España entre el 2006 y el 2015.

Se trata de una fuente de datos semi estructurada, donde en el archivo PDF se puede encontrar información expuesta en forma de texto, tablas, gráficos o incluso mapas. Al tener una estructura semi estructurada, el nivel de complejidad de extracción va a ser intermedio. Para aplicar el proceso ETL se han utilizado las siguientes librerías de Python:

* Camelot *[29]*. Esta librería se utiliza para extraer tablas de archivos PDF, tiene gran poder ya que puede extraer tablas incluso si no están correctamente estructuradas o si tienen múltiples categorías en una misma tabla. Otra ventaja al usar esta librería que una vez extraída las tablas las almacena en una lista, por lo que, el proceso de extracción y carga se ve facilitado.
* Pandas *[4]*. Como se ha explicado en el punto anterior, es una biblioteca de gran poder que ayuda a los procesos de tratamiento de datos implementando multitud de opciones y funciones.

Lo primero que se realiza en el programa desarrollado en Python es leer el archivo PDF utilizando la librería *Camelot*. Esta librería permite que lea el archivo e indicándole la página que se quiera leer extraiga las tablas que se encuentren en esa página y almacenando las tablas en una lista de Python. Para el proyecto se esta utilizando la *página 26*, donde podemos encontrar una tabla que indica la evolución de los grandes incendios entre el 1968 hasta el 2015.

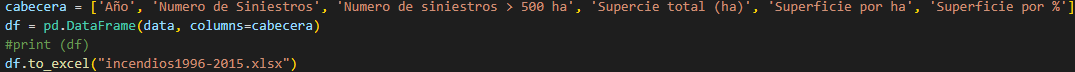
Una vez extraída toda la tabla se procede a la transformación, se procede a quitar determinados caracteres entre los números que entorpecen su utilización. Por ejemplo, se quita el símbolo “%” de todos números o se quitan los puntos que en documento se utilizan como separador de miles. Para poder realizar alguna de estas transformaciones es necesario pasar la tabla extraída al formato data frame.

Texto

Descripción generada automáticamente

Código . Proceso de extracción y transformación de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”.

Ya completados los procesos de extracción y transformación de los datos se procede a cargar los datos resultantes en un archivo xlsx, que actúa como data warehouse. Tambien es necesario incluir la cabecera de los datos, indicando que significa cada columna. Todo este proceso se realiza utilizando la librería *Pandas*. El código es el siguiente:



Código . Proceso de Carga de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”.

Se expone una comparación entre parte de los datos almacenados en la fuente de datos y parte de los datos almacenados en el archivo xlsx, que son frutos del resultado del proceso ETL llevado a cabo.

Tabla

Descripción generada automáticamente

Figura . Tabla extraída de la fuente de datos “Incendios producidos en España entre el 2006 y el 2015”, página 26.

Tabla

Descripción generada automáticamente

Figura . Parte del data warehouse una vez realizado el proceso ETL.

4.3.4. Proceso ETL en Concesión de Nacionalidad Española entre el 2010 y el 2019.

Se trata de una fuente de datos totalmente estructurada, por lo que el proceso ETL, es relativamente sencillo. El archivo xslx consiste en una tabla en función de los años y de los diferentes países. Entonces el proceso ETL consiste en extraer todos los datos del archivo, eliminar todos los datos menos las concesiones de nacionalidad a los hombres (se eliminan la categoría de ambos sexos y la de mujeres) y se carga estos datos en un archivo xlsx.

Solo se tienen en cuenta las concesiones de nacionalidad a los hombres, simplemente por escoger una categoría, no tiene relevancia alguna.

Para la realización del proceso ETL en esta fuente de datos se ha utilizado la siguiente librería de Python:

* Pandas *[4]*. Como se indicaba en el punto anterior, esta biblioteca es muy flexible. En el punto anterior se utilizaba solamente para cargar los datos en el data warehouse, mientras que en esta fuente de datos se utiliza durante todo el proceso ETL.

Lo primero que se realiza en el programa desarrollado en Python es leer el archivo xlsx utilizando la librería *Pandas,* el documento al tener más de una página es necesario indicar a que página se está accediendo.

Cuando ya se ha extraído toda la información, se genera una nueva lista indicando a partir de las filas y columnas que información de todo el archivo va a ser parte de la nueva lista.

A continuación, se introducen las cabeceras que va a tener el archivo a la lista y se crea el archivo xlsx, con el nombre “*nacionalidad-Hombres.xlsx*” y se introduce toda la lista en el archivo. El cogido más relevante del programa es el siguiente:

Texto

Descripción generada automáticamente

Código . Proceso de ETL de la fuente de datos “Concesiones de Nacionalidad Española entre el 2010 y el 2019”.

Ya realizado el proceso ETL, parte del data warehouse se visualiza de la siguiente manera, donde se puede hacer una comparación con la *Figura 8*, que aunque no sea la misma categoría, porque el de la *Figura 8* corresponde a la categoría de ambos sexos, tiene el mismo tipo de estructura.

Tabla

Descripción generada automáticamente con confianza baja

Figura . Parte del data warehouse una vez realizado el proceso ETL de la fuente de datos “Concesiones de Nacionalidad Española entre el 2010 y el 2019”.

4.3.5. Proceso ETL en Catálogo del Bosque Urbano de la Ciudad de Madrid.

Esta fuente de datos se caracteriza por tener una estructura semi estructurada, donde en el documento PDF presenta la información de diferentes formas, tal y como se podía observar en la *Figura 14*.

De este documento se extrae la información que aparece desde la *página 64* hasta la *página 75*. En estas páginas se muestra una tabla con el listado de especies a lo largo del bosque urbano de la ciudad de Madrid, esta información esta presentada en forma de tabla, pero, aunque tenga esta estructura presenta dificultades a la hora de aplicar el proceso de extracción.

Se decide extraer esta información del PDF y no otra debido a la cantidad de datos que se presentar, es exactamente una tabla constituida por 446 filas y 19 columnas, donde cada fila es una especie en concreto.

Para la realización del proceso ETL en esta fuente de datos se ha utilizado las siguientes librerías de Python:

* Fitz *[26]*. Como se ha explicado anteriormente esta librería tiene un grandísimo poder y una gran flexibilidad ya que se pueden realizar multitud de acciones. Aunque ya se haya utilizado esta librería en otro documento ETL, la funcionalidad que tiene en proceso ETL difiere totalmente con el que se describió. En esta fuente de datos la librería extrae todo el texto de la página en cuestión, da igual que este en formato de tabla o no. Por lo que, el proceso de transformación es mucho más arduo. Por ejemplo, como se puede observar en la *Figura 18*, además de extraer la tabla extrae el título, el subtítulo, el numero de la página y demás elementos que carecen de valor.
* XlsxWriter *[27]*. Esta librería se utiliza para crear el archivo xlsx que actúa como data warehouse e insertar toda la información en forma de tabla.

El programa desarrollado en Python esta divido en diferentes funciones siguiendo los procesos de extracción, transformación y carga. El programa está divido en una función principal, en una función que implementa el proceso de extracción (*get\_text*), en una función que implementa el proceso de transformación (*clean*) y en una función que implementa el proceso de carga (*export\_xlsx*).

Lo primero que se realiza es importar todo el archivo PDF a través de la biblioteca *Fitz*, a continuación, se realiza un bucle para que cada pagina entre la 64 y 75 realice el proceso de extracción (*get\_text*) y de transformación (*clean*) en cada iteración. Cuando se realiza estos dos procesos el resultado se almacena en una lista. Una vez terminado el bucle y realizado estos dos procesos en todas las páginas indicadas se procede al proceso de carga, toda la información almacenada se introduce al archivo xlsx a través de la función *export\_xlsx*.

El proceso de extracción se extraen todas las palabras del archivo, sin tener en cuenta algunos caracteres, los puntos, los doble puntos, asteriscos, … El texto se va extrayendo en función de filas, cuando es una nueva fila, se indica en el texto gracias a librería *Fitz* con un 0.

Captura de pantalla de computadora

Descripción generada automáticamente

Código . Proceso de extracción de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”.

Una vez obtenido todo el texto de la página en cuestión se procede a transformarlo, con la función *clean*. Primero se elimina del texto extraído toda la información que no es relevante con la tabla el titulo o el numero de pagina del documento. Tambien se eliminan los caracteres que desvirtúan la información. Algunos de estos caracteres son los puntos, comas, el tanto por ciento, … Cada carácter de texto del texto se extrae por separado y se extrae con formato char, por lo que hay también hay que concatenar los que son necesarios.

Texto

Descripción generada automáticamente

Código . Proceso de transformación de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”.

Ya con todas las páginas extraídas y limpiadas una a una se procede a guardar la información en el data warehouse. Este proceso es similar a los procesos de carga de los puntos anteriores. Se proceso la carga a través de la librería *XlsxWriter*, lo primero que se realiza es la introducción de la cabecera de los datos en la tabla, después se introduce cada dato de la tabla, en cada casilla del archivo xlsx.

Texto

Descripción generada automáticamente

Código . Proceso de carga de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”.

Una vez completo todo el proceso ETL, la visualización de la información en el data warehouse corresponde con la *Figura 19*. También se ofrece una comparativa de como esta dispuesta la tabla en el documento PDF y como queda la información almacenada tras el proceso ETL.

Tabla

Descripción generada automáticamente

Figura . Parte de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”.

Imagen que contiene Texto

Descripción generada automáticamente

Figura . Parte de la visualización del data warehouse de la fuente de datos “Catalogo de Bosque Urbano de la Ciudad de Madrid”.

4.3.6. Proceso ETL en Población por Provincias de España 1996-2021.

Se trata de una fuente de datos estructurada, por lo que, el proceso ETL es mas simple. La fuente de datos está constituida en función de las provincias españolas y los años, también esta divida en función de sexo (ambos sexos, masculino o femenino). El proceso ETL se realiza para obtener solo la categoría de ambos sexos para todas las provincias, se elige esta categoría por elegir alguna categoría, no hay ninguna razón de peso detrás de la decisión.

Para la realización del proceso ETL en esta fuente de datos se ha utilizado las siguientes librerías de Python:

* Pandas *[4]*. Esta librería ya se ha explicado en los puntos anteriores. La implementación desarrollada por esta librería es la misma que realizado en el punto *4.3.4* con la fuente de datos de *Concesiones de Nacionalidad Española*.

Para desarrollar el programa en Python, todo se implementa a través de la biblioteca *Pandas*. Se importa todo el archivo, a continuación, extrae de toda la información que se ha importado el rango de filas y columnas que abarcan solo a la categoría de ambos sexos.

Después se produce la limpieza de los datos a través de la función *clean.* La primera columna del archivo, donde se tiene el nombre de cada provincia tiene un numero asignado y nombre, tal y como se puede observar en la *Figura 12*, se limpia el archivo para que solo se tenga en cuenta el nombre.

A continuación, se introducen las cabeceras que van a tener en el data warehouse y a través de la misma librería Pandas, se crea el archivo xlsx y se introduce todos los datos en el archivo.

Texto

Descripción generada automáticamente

Código . Proceso ETL de la fuente de datos “Población por Provincias de España 1996-2021”.

Una vez aplicado todo el proceso ETL, parte de la visualización del resultado es la siguiente:

Tabla

Descripción generada automáticamente

Figura . Parte de la visualización del data warehouse de la fuente de datos “Población por Provincias de España 1996-2021”.

4.4. Comparativa entre las librerías ETL.

Para realización de los procesos ETL aplicados en las diferentes fuentes de datos abiertas se ha utilizado una serie de librerías como ya se han indicado, donde una librería puede realizar el proceso de extracción o transformación o de carga o todos los procesos.

Al realizar cada proceso se ha realizado una investigación tanto teórica como experimental de que librería se adecuaba o funcionaba de la forma mas optima teniendo en cuenta la fuente de datos abierta.

Dependiendo de cómo se presentan los datos en las diferentes fuentes de datos o cuales son los objetivos para analizar de cada una de ellas, es conveniente utilizar unas u otras. Por todo esto, se realiza una comparativa entre las diferentes librerías. Tanto las que se han llegado a implementar como las librerías que se han investigado pero que no se han utilizado, pero pueden ser útiles para otros objetivos.

La comparativa entre las diferentes librerías para la realización de los procesos ETL se visualiza en forma de tabla. La tabla está compuesta por las siguientes categorías:

* Tipo de Proceso ETL. Se indica que procesos ETL (extracción, transformación y carga) abarca la librería.
* Complejidad de Uso. Indica la complejidad de la librería a la hora del aprendizaje y de la implementación.
* Rendimiento. Cuantifica el rendimiento de la librería en términos de velocidad y su eficiencia en la ejecución. Califica la calidad de los resultados una vez aplicada la librería.
* Tipo de Estructura. Indica que tipo de estructura se presenta los datos para garantizar la mayor calidad a la hora de realizar los procesos ETL (tablas, imágenes, gráficos, grafos, texto) Cuando una información se encuentra presentada en forma de tabla, se puede indicar que tiene una tabular.
* Documentación/Soporte. Indica si la librería cuanta con una documentación de calidad y si la librería esta siendo activamente desarrollada u actualizada.

La comparativa entre las librerías en Python utilizadas para los diferentes procesos ETL es la siguiente:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Librería** | **Tipo de Proceso ETL** | **Complejidad de Uso** | **Rendimiento** | **Tipo de Estructura (Extracción)** | **Documentación/Soporte** |
| Fitz | Extracción y transformación de datos de archivos PDF. | Difícil | Medio | Texto | Baja |
| XlsxWriter | Carga en archivos Excel (xlsx). | Fácil | Alto | - | Buena |
| Pandas | Extracción, transformación y carga de datos tabulares. | Fácil | Alto | Tabular | Excelente |
| NumPy | Transformación en arreglos y matrices. | Medio | Alto | - | Buena |
| Camelot | Extracción de tablas de archivos PDF. | Fácil | Alto | Tabular | Media |
| Tabula-Py | Extracción de tablas de archivos PDF | Medio | Media | Tabular | Media. |
| PySpark SQL | Transformación de datos con SQL. | Media | Alto | - | Buena |
| PyPDF2 | Extracción y transformación en archivos PDF. | Medio | Baja | Texto | Buena |

Tabla . Comparación entre las diferentes librerías para los procesos ETL.

Cabe destacar que la mayoría de los procesos de transformación realizados, se han llevado a cabo a partir de las propias funciones que proporciona el lenguaje de programación Python.

La gran mayoría de la información de las fuentes de datos abiertas se presenta en formato tabular, en formato de tabla, por ello, es de importancia utilizar las librerías correctas para cada tipo de tabla.

Como se puede observar en la *tabla 2*, hay múltiples librerías para realizar el proceso de extracción cuando la información esta presentada en forma tabular. Pero no todas las tablas están constituidas de la misma forma, ni presentadas en el mismo archivo. Por ello, dependiendo de la tabla se utiliza una librería u otra. A continuación, en la *tabla 3*, se presenta que librerías son más adecuadas para que tipo de tablas.

|  |  |  |
| --- | --- | --- |
| **Librería** | **Estructura de la Tabla** | **Archivo** |
| Pandas | Tablas bien estructuradas. | CSV, Xlsx. |
| Camelot | Tablas bien estructuradas y simples con varios subtítulos. | PDF |
| Tabula-Py | Tablas más complejas con celdas fusionadas o diseños complicados. | PDF |

Tabla . Comparación entre las librerías para la extracción de tablas.

4.5. Modelos Estadísticos.

4.5.1. Análisis Descriptivo.

Texto

Descripción generada automáticamente

Capítulo 7

1. Bibliografía.

[1] European University Alliance for Global Health. “*Eulogh"*. [en línea]. Available: <https://www.eugloh.eu/>

[2] Enhancing Skills for Smart City Tech. “*Smacite”*. [en línea]. Available: <https://smacite.eu/>

[3] The official portal for European Data. “*Open Data Maturity*”. [en línea]. Available:

<https://data.europa.eu/en/publications/open-data-maturity/2022>

[4] Powerful Data Structures for Data Analysis. “*Pandas*”. [en línea]. Available:

<https://pandas.pydata.org/>

[5] Beautiful Soup Documentation 4.12.2. “*Beautiful Soup*”. [en línea]. Available:

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

[6] The Fundamental package for Scientific Computing with Python. “*NumPy*”. [en línea]. Available: <https://numpy.org/>

[7] Popularity of Programming Languages Index. “*PYPL*”. [en línea]. Available: <https://pypl.github.io/PYPL.html>

[8] Python Programming Language. “*Python*”. [en línea]. Available: <https://www.python.org/>

[9] Programming environment Visual Studio Code. [en línea]. “*Visual Studio Code*”. Available:

<https://code.visualstudio.com/>

[10] Microsoft Software Excel. “*Excel*”. [en línea]. Available: [https://www.microsoft.com/es-es/microsoft- 365/excel](https://www.microsoft.com/es-es/microsoft-%20%20365/excel)

[11] Microsoft Software PowerBi. “*PowerBi*”. [en línea]. Available: <https://powerbi.microsoft.com/es-es>

[12] Microsoft Software Word. “*Word*”. [en línea]. Available: [https://www.microsoft.com/es-es/microsoft-365/word?](https://www.microsoft.com/es-es/microsoft-365/word?ms.officeurl=word&rtc=1&activetab=tabs%3afaqheaderregion3)

[13] Visualization Software LuciChart. “*LuciChart*” [en línea]. Available: <https://lucid.app/documents#/documents?folder_id=recent>

[14] “*GitHub*” [en línea]. Available: <https://github.com/>

[15] Enlace a mi repositorio de GitHub

[16] Ministerio para la Transición Ecológica y el Reto Demográfico. “*Incendios forestales en España. Decenio 2006-2015*” [en línea]. Available: <https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/temas/incendios-forestales/incendios-decenio-2006-2015_tcm30-521617.pdf>

[17] Ministerio de Inclusión, Seguridad Social y Migraciones (España). “*Estadísticas de concesiones de autorizaciones de residencia y trabajo*”. [en línea]. Available: <https://inclusion.seg-social.es/web/migraciones/homees/Estadisticas/operaciones/concesiones/index.html>

[18] Ministerio de Sanidad (España). “*Situación actual del COVID-19*”. [en línea]. Available: [https://www.sanidad.gob.es/areas/alertasEmergenciasSanitarias/alertasActuales/nCov/situacionActual/index.html](https://www.sanidad.gob.es/areas/alertasEmergenciasSanitarias/alertasActuales/nCov/situacionActual.html)

[19] Ayuntamiento de Madrid. “Catálogo de parques municipales de Madrid”. [en línea]. Available: <https://www.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ContenidosTemporales/Cat%C3%A1logoParques/CATALOGO%20DE%20PARQUES%20MUNICIPALES%20MADRID_BAJA%20RESOLUCI%C3%93N%2003.08.2021.pdf>

[20] Datos del Gobierno de España. “*Accidentes de tráfico de la ciudad de Madrid*”. [en línea]. Available:

<https://datos.gob.es/en/catalogo/l01280796-accidentes-de-trafico-de-la-ciudad-de-madrid1>

[21] “*Instituto Nacional de Estadística (España)*”. [en línea]. Available: <https://www.ine.es/>

[22] INE (España) Tabla de Datos Estadísticos. “*Población en España por Provincia y Sexo*”. [en línea]. Available: <https://www.ine.es/jaxiT3/Tabla.htm?t=2852>

[23] “*Eurostat*”. [en línea]. Available: <https://ec.europa.eu/eurostat>

[24] Eurostat. “*Demographic balance and crude rates at national level*”. [en línea]. Available: <https://ec.europa.eu/eurostat/databrowser/view/DEMO_GIND__custom_2733962/settings_1/table?lang=en&bookmarkId=7084ed24-6b91-4cf3-b90d-d47565593505>

[25] Ayuntamiento de Madrid Zonas Verdes. “*Valor Bosque Urbano de Madrid*”. [en línea]. Available: <https://www.madrid.es/UnidadesDescentralizadas/ZonasVerdes/TodoSobre/ValorBosqueUrbanoMadrid/Valor%20Bosque%20Urbano%20de%20Madrid.pdf>

[26] PyPI Fitz Documentation 1.18.9. “*Fitz Documentation*”. [en línea]. Available: <https://pypi.org/project/fitz/>

[27] PyPI XlsxWriter Documentation 3.1.2. “*XlsxWriter Documentation*”. [en línea]. Available: <https://pypi.org/project/XlsxWriter/>

[28] Python 3.9.7 Sotfware Foundation Datetime. “*Datetime*”. [en línea]. Available: <https://docs.python.org/3/library/datetime.html>

[29] PyPI Camelot Documentation 0.11.0. “*Camelot* *Documentation*”. [en línea]. Available: <https://pypi.org/project/camelot-py/>

[30] Características de Python. *“¿Python el lenguaje del futuro?”.* [en línea]. Available: <https://www.paradigmadigital.com/dev/es-python-el-lenguaje-del-futuro/>

[31] TYS Magazine. “*Las 10 ciudades más inteligentes de Europa*”. [en línea]. Available: <https://tysmagazine.com/las-10-ciudades-mas-inteligentes-de-europa/>

[32] Trabajo Académicos de la Universidad de Alcalá. “*e\_Bu@h*”. [en línea]. Available: <https://ebuah.uah.es/dspace/handle/10017/17681>

[33] Wikipedia. Cross Industry Standard Process for Data Mining. “*CRISP-DM*”. [en línea]. Available: <https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining>

El documento debe igualmente incluir los siguientes apartados obligatorios:

* Índice.
* Resumen en español del trabajo en un máximo de cien (100) palabras.
* Resumen en inglés del trabajo en un máximo de cien (100) palabras.
* Palabras clave con un máximo de cinco.
* Introducción en la que se indique el planteamiento del trabajo y los objetivos a conseguir.
* Descripción del trabajo desarrollado, estructurado como proceda según su tipo, y con los contenidos que se consideren publicables si hay un acuerdo de confidencialidad.
* Conclusiones.
* Bibliografía, que incluirá el conjunto de referencias. Se recomienda el estilo de citación del IEEE.

Se recomienda que la estructura se complete de acuerdo con el siguiente esquema:

* Índice.
* Resumen en español del trabajo en un máximo de cien (100) palabras.
* Resumen en inglés del trabajo en un máximo de cien (100) palabras.
* Resumen extendido del trabajo en un máximo de 4 páginas.
* Glosario de acrónimos y abreviaturas.
* Introducción en la que se indique el planteamiento del trabajo y los objetivos a conseguir.
* Descripción del trabajo desarrollado estructurado con el siguiente esquema:
  + Base teórica en la que se expongan los conceptos teóricos utilizados para la realización del trabajo, así como los cálculos realizados.
  + Descripción experimental, cuando sea necesario, descripción del diseño, resultados, etc.
* Conclusiones y, en su caso, trabajo futuro.
* Bibliografía, que incluirá el conjunto de referencias. Se recomienda el estilo de citación del IEEE.
* Anexos/apéndices:
  + Planos y diagramas, entendiendo por tales los generales, diagramas de bloques, esquemas de detalle y planos, ajustados a la normativa existente sobre el análisis y el diseño de sistemas hardware y software. En caso de restricciones de confidencialidad se excluirá este apartado.
  + Pliego de condiciones (en su caso), que incluya las condiciones generales (normativas), condiciones de materiales y de equipos (características técnicas) y condiciones de ejecución. En caso de restricciones de confidencialidad se excluirá este apartado.
  + Presupuesto, que incluya: ejecución material (materiales y mano de obra), gastos generales y beneficio industrial, honorarios de dirección y redacción (tarifas del Colegio, en su caso), coste de ejecución por contrata y presupuesto total. En caso de confidencialidad se excluirá este apartado.
  + Manual de usuario y/o de instalación y/o de mantenimiento (en su caso), en todos aquellos equipos o programas generados en el trabajo y que vayan a utilizarse posteriormente.