Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN

ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

RIESGOS MEDIOAMBIENTALES APLICADOS AL SECTOR INMOBILIARIO

NOMBRE:

RAFAEL J. MAYORAL GONZÁLEZ

CURSO 2021-2022

Universidad Europea

TÍTULO: RIESGOS MEDIOAMBIENTALES APLICADOS AL SECTOR INMOBILIARIO

AUTOR: RAFAEL J. MAYORAL GONZÁLEZ

TITULACIÓN: MÁSTER EN BIG DATA ANALYTICS

DIRECTOR DEL PROYECTO: NICOLAS COCA LÓPEZ

FECHA: OCTUBRE de 2022

RESUMEN

En España, en los tiempos que vivimos actualmente, aspectos relacionados con el medioambiente como el cambio climático, el calentamiento global y la falta de recursos tan importantes como el agua van adquiriendo cada vez mayor importancia y hacen que aumente la preocupación por parte de las personas. Es muy común ver diariamente en los periódicos como va creciendo la población afectada por estos fenómenos.

Dada esta situación, el interés por desarrollar herramientas que nos permitan lidiar con estos fenómenos medioambientales está creciendo inevitablemente y son cada vez más las compañías y organizaciones que apuestan por aplicar nuevas tecnologías a este campo, en concreto las técnicas de Big Data y aprendizaje automático, que son las que se utilizarán en este proyecto.

Gracias a que la cantidad de datos que se generan diariamente en el mundo no para de incrementar, resulta indispensable disponer de estas tecnologías en cualquier campo para poder tratar dicha cantidad de datos, entenderlos y extraer conclusiones. El uso de herramientas de aprendizaje automático en el campo de los riesgos medioambientales no es extraño de ver, ya que hoy en día se utiliza bastante para otros fines relacionados, como por ejemplo, en predicción de incendios, predicción meteorológica e incluso en predicción del movimiento de asteroides.

Por otro lado, el sector inmobiliario es un sector que no se quiere quedar atrás en cuanto a tecnología se refiere y es uno de los más interesados en estos temas, puesto que los riesgos medioambientales afectan directamente a la toma de decisiones de compradores, inversores y todas aquellas personas interesadas en adquirir o vender un inmueble.

Es por eso por lo que, en este proyecto, en colaboración con una empresa del sector inmobiliario, se lleva a cabo el desarrollo de herramientas software de Big Data y aprendizaje automático relacionadas con los riesgos medioambientales. El desarrollo consta de dos partes, la primera en la que se lleva a cabo un proceso de extracción, transformación y carga (ETL) de las cinco fuentes de datos seleccionadas sobre diferentes riesgos medioambientales y la segunda parte en la que se aplican técnicas de aprendizaje automático sobre esos datos.

Las técnicas de aprendizaje automático que se utilizan en este proyecto son técnicas de aprendizaje no supervisado, en concreto de agrupamiento, con el propósito de obtener patrones de agrupación en los datos que se usarán para obtener un índice global de riesgo medioambiental. Posteriormente se utilizará dicho índice para clasificar el riesgo medioambiental de cada uno de los municipios del territorio español y para crear una visualización de ello en un mapa.

Además, gracias al tratamiento de datos que se realiza en este proyecto, se obtienen otro tipo de herramientas que son también de mucha utilidad, como un endpoint en el que dadas una latitud y una longitud nos muestra los riesgos medioambientales que afectan a esa localización o un mapa de riesgos medioambientales en el que podemos visualizar los datos espaciales de cada uno de los riesgos por capas.

ABSTRACT

In Spain, in the times we live in today, aspects related to the environment such as climate change, global warming and the lack of resources as important as water are becoming increasingly important and step-up people´s concern. It is very common to see daily in the newspapers how is growing the affectation of these phenomena to a big part of the population.

Given this situation, the interest in developing tools that allow us to deal with these environmental phenomena is inevitably growing. This makes that more and more companies are betting on applying new technologies to this field, specifically Big Data and machine learning techniques that will be used in this project.

Thanks to the fact that the amount of data generated daily in the world is increasing each time more, it is essential to apply these technologies in any field, to be able to process this data, understand it and draw conclusions. The use of machine learning tools in the field of environmental risks is not uncommon, as it is widely used today for other related purposes, such as fire forecasting, weather forecasting, and even asteroids movement prediction.

On the other hand, the real estate area is an area that does not want to be left behind in terms of technology and is one of the most interested in these issues, because environmental risks directly affect the decision-making of buyers, investors and all those interested in buying or selling a property.

That is why, in this project, in collaboration with a real estate company, the development of Big Data and machine learning software tools related to the environmental risks is carried out. The development of the project consists of two parts, the first one consists in extraction, transformation, and loading (ETL) process of the five selected data sources of different environmental risks and the second part in which machine learning techniques are applied automatically to that data.

The machine learning techniques used in this project are unsupervised learning techniques, specifically clustering, with the purpose of obtaining clustering patterns in the data that will be used to obtain a global risk index. Later, this index will be used to classify the environmental risk of each of the townships existing in Spain and to create a visualization of it on a map.

In addition, thanks to the data processing carried out in this project, other types of tools are obtained that are also very useful, such as an endpoint in which, given a latitude and a longitude, it shows us the environmental risks that affect that location or a map of environmental risks in which we can visualize the spatial data of each of the risks by layers.

Universidad Europea

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. INTRODUCCIÓN

Con el incesante crecimiento de los precios en el sector inmobiliario, la demanda y la escasez de inmuebles, resulta cada vez más difícil encontrar buenas oportunidades para hacerse con una vivienda o invertir en este sector.

A este problema, además, se incorpora el aumento de la preocupación del consumidor por el crecimiento de episodios de desastres medioambientales. Es sabido por todos, sobre todo por su recurrente aparición cada día en las noticias, que el cambio climático y el calentamiento global es un problema del presente y que se está haciendo notar cada vez más.

Sin ir más lejos, este mismo verano el intenso calor que hemos vivido en España ha hecho que la superficie quemada provocada por incendios supere a la de las últimas tres décadas, liderando España la lista de los países europeos con más hectáreas quemadas en incendios.
Y no sólo incendios, también han incrementado los casos de inundaciones, de sequías extremas y otros desastres medioambientales como el de La Palma de este mismo año, que sufrió innumerables daños procedentes de la erupción del volcán Cumbre Vieja.

Todo este tipo de desastres medioambientales trae consigo que los futuros compradores e interesados en el sector inmobiliario demanden cada vez más información antes de hacer cualquier tipo de movimiento y es por este motivo por el que cada vez va cogiendo más fuerza el término del Big Data, y por el cuál las empresas y los inversores cada vez invierten más capital en este tipo de tecnologías. Con ellas, se puede acceder a herramientas, a las cuales no todo el mundo tiene acceso, que proporcionan información que supone una clara ventaja en la toma de decisiones con respecto a los que no las utilicen.

Con esa ventaja encima de la mesa, es mucho más fácil encontrar las opciones que mejor se adapten a los criterios de búsqueda del interesado e incluso es posible que a largo plazo sean las opciones más rentables.

1.1 Planteamiento del problema

La motivación de este proyecto surge a raíz de la petición de los clientes de acceso a información menos trivial sobre los activos inmobiliarios que poseen o en los que quieren invertir.

La necesidad del inversor hoy en día es la de minimizar al máximo posible la aleatoriedad de cualquiera de las variables que puedan afectar a uno de sus activos y puedan provocarles perdidas, por eso es cada vez más recurrente hacer uso de herramientas de Big Data e Inteligencia Artificial para extraer toda la información posible del entorno y analizarla para extraer las conclusiones pertinentes antes de llevar a cabo cualquier operación.
En este caso particular, se pretende analizar y extraer información acerca de los riesgos medioambientales asociados a un activo inmobiliario, por lo que el problema consistirá principalmente en cumplir con dos hitos.

El primero de ellos, será llevar a cabo una exploración y análisis de las fuentes de datos que nos puedan ser de utilidad para lograr nuestro objetivo, extraerlas, procesarlas, transformarlas y almacenarlas, es decir, desarrollar un proceso de ETL para tener disponibles esas fuentes de datos.

El segundo hito, será una vez se tengan las fuentes de datos disponibles. Consistirá en llevar a cabo un procesamiento y un análisis de esas fuentes de datos para extraer la información necesaria para obtener patrones y conclusiones que ayuden a la toma de decisiones. Para lidiar con ello, se hará uso de distintas técnicas de machine learning, las cuales se evaluarán para seleccionar de ellas la más idónea para este problema.

1.2 Objetivos del proyecto

Con este proyecto se pretende lograr el desarrollo de un software inteligente que sea capaz de extraer automáticamente datos de las fuentes seleccionadas para transformarlos y darnos información sobre los riesgos medioambientales asociados al sector inmobiliario dentro del territorio de España.

Para alcanzar ese objetivo final, será necesario descomponer el problema en varios objetivos más pequeños:

1. Análisis y exploración de las fuentes de datos que sean de interés para el proyecto.

2. Desarrollo de software para la extracción, transformación y carga (ETL) automática de los datos.

3. Análisis y selección de las técnicas de machine learning oportunas para el proyecto.

4. Comparación de resultados entre las técnicas, representación de los resultados y desarrollo de software para su uso

1.3 Estructura del proyecto

Este proyecto se divide en siete capítulos. En esta sección se facilitará una breve descripción de los contenidos de cada capítulo.

- **Capítulo 1: Introducción**. Este capítulo será un capítulo introductorio donde se explicarán los objetivos y motivaciones que han dado lugar al desarrollo de este Trabajo de Fin de Máster.

- **Capítulo 2: Conceptos relacionados y Estado del Arte**. En este capítulo se detallarán todos los conceptos que sirven de pilar fundamental para el desarrollo de este proyecto. Será un capítulo en el que se expondrá el dominio del problema y toda su base teórica. Se presentarán como los aspectos más relevantes en este proyecto los conceptos de indicador clave de rendimiento y cuadro de mando integral.

- **Capítulo 3: Preparación de los datos**. En el capítulo 3 se explicarán todos los procesos llevados a cabo para lograr alojar todas las fuentes de datos de manera consistente y estructurada en un sistema de almacenamiento accesible y consultable de forma ágil.

- **Capítulo 4: Construcción del modelo**. En el capítulo 4 se tratarán de forma teórica y profundizando en los detalles matemáticos, todas las técnicas y algoritmos de machine-learning que se han utilizado para construir el modelo que trabajará con los datos obtenidos en el Capítulo 3.

- **Capítulo 5: Evaluación de los modelos y de las soluciones obtenidas**. El capítulo 5 se ha dedicado explicar cómo se han construido los modelos y a analizar el rendimiento de las soluciones obtenidas para compararlas y elegir la que más se adapte a las necesidades de este proyecto.
- **Capítulo 6: Herramientas obtenidas**. En este breve capítulo se explicará a modo de guía cómo funcionan las herramientas software consumibles que se han construido durante este proyecto.
- **Capítulo 7: Conclusiones y futuras líneas de trabajo**. Para finalizar, se introduce un capítulo donde se detallan las conclusiones obtenidas tras la realización de este proyecto y se comentan las posibles actualizaciones que podría recibir en el futuro para su mejora.

Capítulo 2. Conceptos relacionados y Estado del arte

2.1 Introducción.

En este capítulo se definirán todos los conceptos y tecnologías relevantes para el desarrollo de este proyecto. Para la consecución del Objetivo 2: Desarrollar un software ETL es necesario contar con un sistema de almacenamiento en la nube para almacenar ahí los datos en bruto y posteriormente realizar los procesamientos pertinentes hasta tener los datos con el formato deseado y almacenarlos, en este caso, en una base de datos no relacional como es MongoDB para su rápido acceso y consumo.

Puesto que se va a trabajar también con datos que contienen geometrías y coordenadas será necesario conocer los aspectos y características de los archivos GeoJSON, que será el formato de datos que utilicemos para trabajar con dicha información.

Por último, para mostrar los resultados que se han obtenido en este proyecto será necesario recurrir a una herramienta capaz de visualizar y trabajar con datos geométricos, como es CARTO.

2.2 Sistemas de almacenamiento en la nube: AWS.

Amazon Web Services (Amazon Web Services (AWS), 2022) (AWS para abreviar) es uno de los principales proveedores de computación en la nube en todo el mundo. La computación en la nube generalmente incluye los tres componentes infraestructura como servicio (IaaS), plataforma como servicio (PaaS) y software como servicio (SaaS). O en pocas palabras: en lugar de operar sus propios servidores o centros de datos, los clientes aquí obtienen la infraestructura y los servicios de TI según sea necesario y no tienen que preocuparse por mantener y actualizar el hardware y los centros de datos subyacentes.

AWS ofrece una gama muy amplia de funciones y productos, además de servicios de computación. Esto incluye soluciones de infraestructura, como recursos informáticos, almacenamiento y bases de datos, así como soluciones para las tendencias actuales, como aprendizaje automático, inteligencia artificial o Internet de las cosas.

De entre los servicios más populares de AWS, destacaremos Amazon Simple Storage Service (Amazon S3), que es un servicio de almacenamiento en la nube calculable, de alta velocidad y basado en la web que será protagonista en este proyecto puesto que haremos uso de él para almacenar todos nuestros datos. El servicio permite realizar copias de seguridad y archivar

datos y aplicaciones en línea en Amazon Web Services (AWS). Amazon S3 se creó con un conjunto mínimo de funciones y está destinado a facilitar a los desarrolladores el trabajo con la informática a escala web.

Ilustración 1. Logotipo AWS.

2.3 Python.

2.4 Bases de datos no relacionales: MongoDB.

MongoDB (MongoDB, 2022) es una base de datos NoSQL de código abierto. NoSQL significa que la base de datos no usa tablas relacionales como una base de datos SQL tradicional.

Hay varios tipos de bases de datos NoSQL, pero MongoDB almacena datos en objetos similares a los que se utilizan en JavaScript llamados documentos, cuyo contenido está compuesto por diccionarios clave-valor.

A diferencia de las tablas SQL, en MongoDB no existen limitaciones estructurales en los datos que se pueden almacenar. Los esquemas de datos no se aplican: se puede almacenar lo que quiera, donde quiera. Esto hace que MongoDB sea ideal para estructuras de datos orgánicas o desordenadas.

Ilustración 2. Logotipo MongoDB.

2.5 Concepto de ETL (Extraction, Transform y Load).

El término ETL (Kimball, 2004) es un término que se utiliza para referenciar una técnica para el tratamiento de la información, la cual se compone de varios procesos individuales que se usan para integrar datos de varias fuentes de datos en un almacén de datos común extrayéndolas y preparándolas. Este proceso es comúnmente muy utilizado para procesar grandes cantidades de datos en el entorno de Big Data y Business Intelligence.

La abreviatura ETL está compuesta por las iniciales de tres términos en inglés “Extract”, “Transform” y “Load” que se traducen como extraer, transformar y cargar. De ahora en adelante nos referiremos a este proceso como ETL.

El uso de este tipo de técnicas es muy ventajoso a la hora del procesamiento de grandes cantidades de datos porque si la información está distribuida en diferentes sistemas de información, es redundante o presenta diferentes estructuras, utilizar el proceso ETL nos permite que los datos heterogéneamente estructurados de diferentes fuentes se combinen y procesen dando el resultado de una mejor calidad en los datos y una consistencia asegurada, establecida en el almacén de datos.

Para las aplicaciones existentes en el entorno Big Data, es importante que los pasos del proceso ETL se ejecuten a alta velocidad y con baja latencia. Los principales fabricantes de sistemas de administración de bases de datos y aplicaciones de Big Data, como IBM, SAP, Oracle o Microsoft, ofrecen productos que respaldan el proceso ETL. También están disponibles numerosas herramientas ETL de código abierto. En el caso de este proyecto el software construido para el proceso ETL será expresamente desarrollado por el autor haciendo uso de las herramientas que se han mencionado anteriormente en este capítulo.

2.6 JSON y GeoJSON.

JSON son las siglas de JavaScript Object Notation (JavaScript Object Notation, 2022), que es un formato ligero de intercambio de datos muy fácilmente comprensible para las personas, además de ser generado (en términos computacionales) y manipulado fácilmente por las máquinas. Está basado en un subconjunto del Lenguaje de Programación JavaScript, Standard ECMA-262.

JSON está constituido por dos estructuras:

- Una colección de pares clave/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla hash, lista de claves o un arreglo asociativo.
- Una lista ordenada de valores. En la mayoría de los lenguajes, esto se implementa como vectores, listas o secuencias.

Este tipo de estructuras son universales puesto que virtualmente cualquier lenguaje de programación es capaz de soportarlas de una forma u otra. Es muy común que un formato de intercambio de datos que es independiente del lenguaje de programación se base en estas estructuras.

GeoJSON (GeoJSON, 2022) es un formato de intercambio de datos espaciales de estándar abierto para representar elementos geográficos simples y sus atributos no espaciales. GeoJSON está basado en JSON y es un formato para codificar diferentes estructuras de datos geográficos que es ampliamente utilizado en aplicaciones de cartografía en entornos web gracias que permite el intercambio de datos de forma rápida y fácilmente comprensible. Este formato se basa en el sistema de referencia de coordenadas geográficas del Sistema Geodésico Mundial de 1984 y utiliza grados decimales como unidad.

2.7 Visualización de datos geográficos: CARTO.

CARTO (CARTO, 2022) es la plataforma líder mundial para visualización de datos espaciales de manera inteligente. Una de las principales ventajas competitivas que presenta CARTO es que permite a las organizaciones hacer uso de datos espaciales para analizar las rutas de entrega más eficientes, mejorar sus propuestas de marketing, ubicaciones estratégicas para espacios físicos… entre otros muchos usos. Además sirve de mucha utilidad para científicos de datos, desarrolladores y analistas para optimizar sus procesos de negocio, predecir futuros resultados gracias a la ciencia de datos espaciales.

Ilustración 3. Logotipo de CARTO.

Capítulo 3. Preparación de los datos

3.1 Análisis y exploración de fuentes

Antes de comenzar con los aspectos técnicos de machine learning que se van a aplicar en este proyecto, será necesario esclarecer qué datos vamos a utilizar y la preparación que será necesaria para poder aplicar estas técnicas.

Para ello, existen dos objetivos previos al análisis y selección de técnicas de machine learning del capítulo posterior, que se describirán a continuación.

Con respecto al primer objetivo, el análisis y exploración de las fuentes de datos, se explorarán y analizarán fuentes de datos que cumplan con los siguientes requisitos:

- Que sean fuentes oficiales del gobierno de España.
- Que proporcionen datos sobre el territorio español, en concreto, sobre riesgos medioambientales.
- Que sean descargables.
- Que presenten licencia de datos abiertos.

Teniendo estos requisitos presentes se explorarán los datos contenidos en las siguientes páginas web:

- Ministerio de transportes, movilidad y agenda urbana (MITMA) (Ministerio de Transportes Movilidad y Agenda Urbana, 2022).
- Consejo Superior de Investigaciones Científicas (CSIC) (Consejo Superior de Investigaciones Científicas, 2022).
- Instituto Nacional de Estadística (INE) (Instituto Nacional de Estadística, 2022).
- Centro Nacional de Información Geográfica (CNIG) (Centro Nacional de Información Geográfica, 2022).
- Instituto Geográfico Nacional (IGN) (Instituto Geográfico Nacional, 2022).
- Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO) (Ministerio para la Transición Ecológica y el Reto Demográfico, 2022).
- Instituto Geológico y Minero de España (IGME) (Instituto Geológico y Minero de España, 2022).

El objetivo de la exploración consistirá en encontrar conjuntos de datos sobre alguno de los riesgos medioambientales en los que exista dentro de su diccionario de datos atributos principalmente sobre la localización geográfica a la que afecta (coordenadas, geometrías, localidad…) y una calificación sobre el riesgo que supone, la frecuencia con la que ocurre o que este campo pueda ser inferido.

Un ejemplo de un conjunto de datos que encajaría perfectamente con los requisitos anteriormente detallados serían los Mapas de peligrosidad por inundación costera localizados en la página web del Ministerio para la Transición Ecológica y el Reto Demográfico.

En este conjunto de datos, no se especifica expresamente una calificación sobre el riesgo pero estos vienen separados en diferentes archivos correspondientes cada uno a un escenario con

una probabilidad distinta de inundación (períodos de retorno de 100 años o 500 años) por lo que podemos añadir este campo posteriormente.

En la ilustración, se muestra un ejemplo de un registro correspondiente a este conjunto de datos, donde podemos ver que incluye información geográfica para poder localizar el territorio en el que se ha producido esta inundación, que es un requisito necesario para posteriormente construir el modelo.


```
{'type': 'Feature',
 'id': '0',
 'properties': {'id_poblaci': 'ES060_ARPS_0132_T500_POB_001',
  'cod_arpsi': 'ES060_ARPS_0132',
  'superficie': 0.02,
  'id_municip': '11004',
  'nom_munici': 'Algeciras',
  'n_hab_muni': 116917,
  'num_hab_zi': 83,
  'victim_ori': 0,
  'herid_ori': 0,
  'otras_cons': None},
 'geometry': {'type': 'Polygon',
  'coordinates': [[(-5.441415983088461, 36.16102383587657),
    (-5.441450362844116, 36.160963080472015),
    (-5.441758528447916, 36.16110088076272),
    (-5.441672430554897, 36.161248884655684),
    (-5.441641916908702, 36.161301337602325),
    (-5.441556595196858, 36.161448005146184),
...}
```


Ilustración 4. Ejemplo de registro contenido en los mapas de peligrosidad por inundación costera en formato JSON.

3.2 Extraction, Transform and Load (ETL).

Una vez se tengan claros los conjuntos de datos que se van a utilizar, se deberá proceder con el segundo objetivo de desarrollar el proceso de ETL de los datos. El proceso de ETL es el objetivo más costoso y que más tiempo requiere de todo el proyecto, ya que en la mayoría de proyectos suele representar el 80% del tiempo dedicado al desarrollo de ese software. En el caso del proyecto actual, ha supuesto alrededor de dos meses de trabajo.

Este proceso en este proyecto consistirá básicamente en tres fases (ilustración 2):

- **Extracción**. En esta fase los datos se descargan en bruto de donde han sido localizados y se almacenan en un contenedor. En este proyecto se utilizará como contenedor un clúster en Amazon Web Services, para tener los datos disponibles desde la nube.
  La extracción de datos se llevará a cabo utilizando Python como lenguaje de programación junto a librerías como Requests (Python Requests, 2022), que servirá para procesar las respuestas y las peticiones a los servidores donde se alojan los datos

y en algunos casos BeautifulSoup (Beautiful Soup Documentation, 2022) para leer esas respuestas y extraer los datos que nos interesan, además para poder automatizar la extracción de datos y posteriormente comprobar si hay actualizaciones de estos.
Los scripts desarrollados en Python para llevar a cabo esta fase en cada una de las fuentes recibirán el nombre de “download”.

- **Transformación**. El proceso de transformación tiene como principal objetivo que se establezca en los diferentes conjuntos de datos que se han seleccionado el formato deseado, que en este caso será que todos los conjuntos de datos se adapten a un formato común para poder relacionar unos con otros.
  Para lidiar con esta tarea, se recurrirá a librerías que nos permitan realizar fácilmente alteraciones sobre los datos como Pandas o similares.
  En este paso, añadiremos campos a nuestro dataset como por ejemplo los centroides de las geometrías de cada riesgo incluyendo también el territorio al que pertenecen (municipio, provincia, comunidad…), inferiremos el campo riesgo en aquellos conjuntos de datos que no lo incorporen y los bounding-box de cada geometría para posteriormente acelerar las búsquedas en la base de datos.
  Los scripts desarrollados en Python para llevar a cabo esta fase en cada una de las fuentes recibirán el nombre de “parse”.

- **Carga**. La fase de carga es la menos compleja de las tres, ya que básicamente se exportan los datos transformados a un sistema de salida, como puede ser una base de datos, donde se almacenan para ser consumidos por otros software o por otros usuarios. En este proyecto se utilizará una base de datos no relacional, como MongoDB.
  Los scripts desarrollados en Python para llevar a cabo esta fase en cada una de las fuentes recibirán el nombre de “clean”.

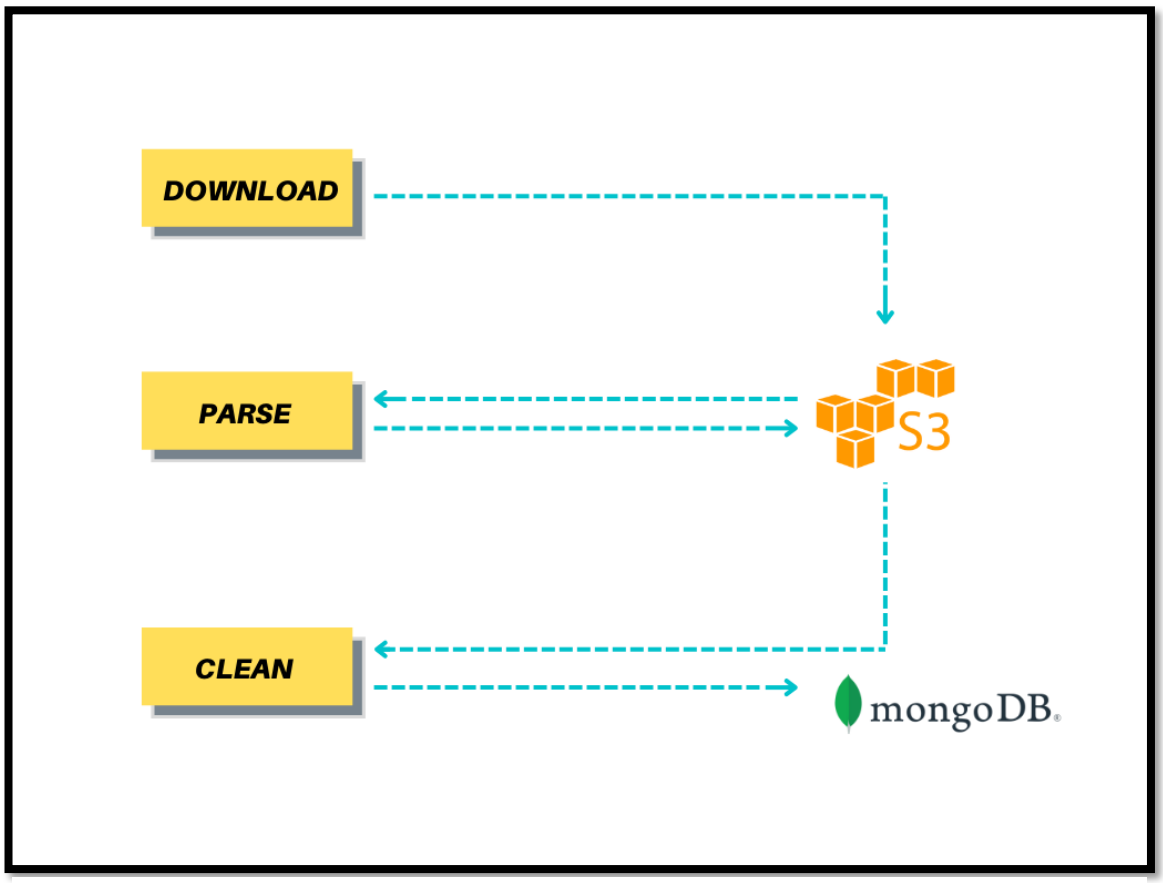

Ilustración 5. Diagrama de flujo ETL.

3.3 Selección de fuentes

Definitivamente, tras llevar a cabo los análisis e investigaciones pertinente se ha decidido incorporar al proyecto las fuentes correspondientes a los siguientes riesgos medioambientales:

- **Riesgo por desertificación**. Esta fuente contiene información relativa a los mapas de riesgo por desertificación proporcionados por el Programa de Acción Nacional contra la desertificación (PAND) que se publican en el Ministerio para la Transición Ecológica y el Reto Demográfico.

  En España, las zonas susceptibles a sufrir desertificación son aquellas en las que la proporción entre la precipitación anual y la evapotranspiración potencial se comprende entre los valores 0.5 y 0.65.

  El término de evapotranspiración potencial se refiere a la cantidad de agua devuelta en forma de vapor por una superficie completamente cubierta de vegetación sin que exista limitación por el suministro de agua para su óptimo crecimiento. Se utiliza para determinar la demanda de agua por las plantas y atmosfera de un lugar concreto.

Siguiendo esa definición, en España existen muchas zonas susceptibles a sufrir riesgo por desertificación, puesto que dos terceras partes del territorio geográfico español son áreas áridas o semiáridas.

La fuente se encuentra localizada en la página del Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO). Los datos vienen en formato Shapefile y su diccionario de datos se puede observar en la Tabla 1 adjunta en los anexos.

- **Riesgo por incendio forestal**. Una de las fuentes que se ha utilizado para incorporar este riesgo medioambiental ha sido la Frecuencia de Incendios por Término Municipal que es elaborada por el Centro de Coordinación de la Información Nacional de Incendios Forestales (CCINIF) que recoge datos de todos los incendios que se originan en España, información anual que proporcionan las Comunidades Autónomas.

  La Estadística General de Incendios Forestales (EGIF) recolecta los datos mediante el Parte de Incendio, un formulario con 130 campos sobre el incendio sucedido. El formulario es revisado por todas las organizaciones responsables y una vez completada la información, anualmente es añadido a la base de datos de la Estadística Forestal Española.

  En este conjunto de datos, que se publica tanto en formato .xls como .kml en la página del Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), su diccionario de datos se localiza en la Tabla 2 de los anexos.

  Puesto que al granularidad del resto de fuentes es a nivel de geometría y en este caso los datos vienen a nivel de municipio, se decide incrementar la granularidad mediante la incorporación de una fuente adicional para llevar a cabo un enriquecimiento de los datos.

  Dicha fuente son los Mapas Forestales de España (MFE) de máxima actualidad que incluye cartografía de todas las masas forestales que se encuentran en el territorio español a nivel de comunidad y en las cuales podemos encontrar las geometrías y polígonos que necesitamos para enriquecer nuestra fuente de Riesgos por Incendios Forestales, ya que podemos geolocalizar las masas forestales donde se producen dichos incendios y evaluar su riesgo.

  MFE también se encuentra disponible para su descarga en el Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO) en formato Shapefile y la información que incluye se puede visualizar en la Tabla 3 de los anexos.

- **Riesgo por sismicidad.** La fuente que se ha decidido incorporar para lidiar con este riesgo medioambiental ha sido la proporcionada por el Instituto Geológico y Minero de España, que recibe el nombre de ZESIS.

  ZESIS es una base de datos de zonas sismogénicas que comprenden las áreas de la península ibérica y territorios que tienen influencia sobre esta. El fin de esta base de

datos es proporcionar información para el cálculo y actualización del mapa de peligrosidad sísmica de España.

Este conjunto de datos es publicado en numerosos formatos, pero para este proyecto utilizaremos el formato Shapefile por su compatibilidad con la librería Fiona de Python. El diccionario de datos correspondiente a este fuente se encuentra en la Tabla 4 de los anexos.

- **Riesgo por inundación fluvial**. Los datos disponibles en esta fuente incluyen la cartografía de las áreas definidas como zonas inundables asociadas a un periodo de retorno determinado. Dichos datos son obtenidos mediante estudios llevados a cabo por las autoridades competentes en materia de aguas, ordenación del territorio y Protección Civil.

  Los periodos de retorno se determinan mediante un estudio hidrológico para definir los caudales y los niveles alcanzados por las láminas de agua con su extensión del área inundada asociada a esa frecuencia.

  Posteriormente, a través de estudios geomorfológicos-históricos se delimitan las zonas con su probabilidad de inundación en función de las evidencias históricas y geomorfológicas identificadas.

  Las zonas se clasifican en zonas no inundables, con alta probabilidad menor o igual a 10 años, frecuente con probabilidad menor a 50 años, media u ocasional con probabilidad menor o igual a 100 años y baja o excepcional con probabilidad igual o menor a 500 años.

  El diccionario de datos asociado a esta fuente de datos se localiza en la Tabla 5 de los anexos.

- **Riesgo por inundación costera.** Los datos disponibles en esta fuente incluyen la cartografía de las áreas definidas como zonas susceptibles de sufrir inundación costera asociadas a un periodo de retorno determinado. Dichos datos han sido elaborados por el Ministerio para la Transición Ecológica (MITECO) y se encuentran descargables desde su portal en distintos formatos, de entre los cuales se seleccionará el formato Shapefile.

  Existen tres posibles escenarios de inundación con diferentes periodos de retorno: Probabilidad media de inundación con periodo de retorno de 100 años y probabilidad baja de inundación con periodo de retorno de 500 años.

  La metodología que se ha llevado a cabo para realizar estos cálculos ha sido a través de distintas hipótesis de oleaje y nivel de mar con modelos matemáticos.

  El diccionario de datos asociado a esta fuente de datos se localiza en la Tabla 6 de los anexos.

3.3 Linaje de datos

El linaje de datos (DAMA Internacional, 2009) es una técnica que se utiliza para analizar el flujo de datos y mantener una trazabilidad sobre estos y todas las transformaciones que sufran. Desde el origen, se muestra la ruta y como llegan al almacenamiento objetivo. Gracias al linaje de datos podemos entender y mantener gobernado el ciclo de vida de los datos.

En el caso de este proyecto, se llevarán a cabo las siguientes tareas para cumplir con el linaje de datos de cada una de las fuentes:

1. **Identificación de los elementos técnicos**.
   a. **Origen**.
      i. Riesgo por desertificación: https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/descarga_pand.aspx
      ii. Riesgo por incendio forestal: https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/incendios-forestales.aspx https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/biodiversidad/mfe.aspx
      iii. Riesgo sísmico: http://info.igme.es/catalogo/resource.aspx?portal=1&catalog=3&ctt=1&lang=por&dlang=eng&llt=dropdown&master=infoigme&shdt=false&shfo=false&resource=51
      iv. Riesgo por inundación fluvial: https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/agua/riesgo-inundacion-fluvial-t500.aspx
      v. Riesgo por inundación costera: https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/costas-medio-marino/Mapas-peligrosidad-por-inundacion-costera.aspx
   b. **Transformaciones**. Para poder homogeneizar todas las fuentes se han llevado a cabo distintas transformaciones en cada una de ellas.
      i. **Riesgo por desertificación**:
         1. Los nombres de todos los campos se han traducido al inglés y en minúscula.
         2. El campo “DESER_CLA” se ha renombrado como “risk” y además se ha dividido en los campos “risk.string” y “risk.value”.
         3. Se han transformado las coordenadas del campo “GEOMETRY” de geométricas a geográficas.
         4. Se ha añadido el campo “area” calculado a partir de la geometría.
         5. Se ha añadido el campo “location” que es el centroide de la geometría.
         6. Se ha añadido el campo “bbox” con los “bounding-box” de las geometrías para acelerar las consultas.
         7. Se ha añadido el campo “territory” con los subcampos “community”, “province”, “township”, “district” y “section” calculados a partir del centroide de la geometría.
      ii. **Riesgo por incendio forestal**:

1. La fuente de datos principal es la de las Masas Forestales de España (MFE) que se ha enriquecido con la de CCINIF añadiendo los datos de los incendios forestales del municipio en el que se encuentra la geometría.
2. Para consolidar la fuente CCINIF se han aglutinado los datos de ambos intervalos de fechas en uno solo incluyendo el total de incendios.
3. Se ha añadido el campo “risk” calculado a partir del número total de incendios en el municipio segregado en distintos intervalos con los subcampos “risk.string” y “risk.value”.
4. Los nombres de todos los campos se han traducido al inglés y en minúscula.
5. Se han transformado las coordenadas del campo “GEOMETRY” de geométricas a geográficas.
6. Se ha añadido el campo “area” calculado a partir de la geometría.
7. Se ha añadido el campo “location” que es el centroide de la geometría.
8. Se ha añadido el campo “bbox” con los “bounding-box” de las geometrías para acelerar las consultas.
9. Se ha añadido el campo “territory” con los subcampos “community”, “province”, “township”, “district” y “section”

iii. **Riesgo sísmico**:

1. Los nombres de todos los campos se han traducido al inglés y en minúscula.
2. El campo “Peligrosiid” se ha renombrado como “risk” y además se ha dividido en los campos “risk.string” y “risk.value”.
3. Se han transformado las coordenadas del campo “GEOMETRY” de geométricas a geográficas.
4. Se ha añadido el campo “area” calculado a partir de la geometría.
5. Se ha añadido el campo “location” que es el centroide de la geometría.
6. Se ha añadido el campo “bbox” con los “bounding-box” de las geometrías para acelerar las consultas.
7. Se ha añadido el campo “territory” con los subcampos “community”, “province”, “township”, “district” y “section” calculados a partir del centroide de la geometría.

iv. **Riesgo por inundación costera y fluvial**:

1. Los nombres de todos los campos se han traducido al inglés y en minúscula.
2. El campo “risk” se ha añadido a partir del periodo de retorno en el que se agrupan los datos y se ha dividido en los campos “risk.string” y “risk.value”.
3. Se han transformado las coordenadas del campo “GEOMETRY” de geométricas a geográficas.
4. Se ha añadido el campo “area” calculado a partir de la geometría.
5. Se ha añadido el campo “location” que es el centroide de la geometría.
6. Se ha añadido el campo “bbox” con los “bounding-box” de las geometrías para acelerar las consultas.
7. Se ha añadido el campo “territory” con los subcampos “community”, “province”, “township”, “district” y “section” calculados a partir del centroide de la geometría.

c. **Destino**. El destino será la base de datos "environment-risks" creada en MongoDB en la que se creará una colección distinta para cada uno de los riesgos medioambientales.

2. **Generación de la trazabilidad y validación de los datos**. La trazabilidad se generará automáticamente en los scripts de Python correspondientes para cada una de las etapas de la ETL que realizarán de manera sistemática todas las transformaciones y movimientos necesarios para almacenar correctamente los datos en la base de datos MongoDB.
3. **Mantenimiento**. En los scripts de Python se incluirá código para que periódicamente se ejecute y se compruebe si alguna de las fuentes ha recibido actualizaciones, y en caso afirmativo se inicie el proceso de actualización.

3.4 Diccionario de datos

El diccionario de datos (DAMA Internacional, 2009) es un herramienta necesaria para contextualizar y ayudar a identificar correctamente los datos en un proyecto. En él se recoge información técnica a nivel de conjunto de datos y todos los campos asociados.

Lo que compone un diccionario de datos es básicamente un conjunto de metadatos formado por propiedades y descripciones de datos que nos ayudan dentro de un proyecto a reducir la documentación, ya que a medida que incluimos más fuentes de datos se vuelve difícil de gestionar. También nos permite localizar más rápido los datos y detalles relevantes de ellos a través de los metadatos.

En este apartado se mostrará el diccionario de datos final producto de todas las transformaciones y procesos realizados durante el procedimiento ETL. El formato de los atributos será JSON que es el que utiliza MongoDB para sus documentos. Los atributos con un "." se entienden como sub-atributos.

La base de datos recibe el nombre de "environmental-risks" y cada uno de los riesgos se almacena en una colección independiente.

- Colección "desertification".

Atributo	Descripción	Tipo de dato
_id	Identificador del documento	String
area	Área de la geometría en metros cuadrados	Float
perimeter	Perímetro de la geometría en kilómetros cuadrados	Float
territory.community	Código INE de la comunidad autónoma	String
territory.province	Código INE de la provincia	String
territory.township	Código INE del municipio	String
territory.district	Código INE del distrito	String
territory.section	Código INE de la sección	String

location	Centroide de la geometría	GeoJson
bbox	Bounding-box de la geometría	GeoJson
geometry	Geometría del riesgo	GeoJson
risk.string	Valor textual del riesgo	String
risk.value	Valor numérico del riesgo	Integer

Tabla 1. Diccionario de datos final (colección "desertification")

- Colección “forest_fires”.

Atributo	Descripción	Tipo de dato
_id	Identificador del documento	String
area	Área de la geometría en metros cuadrados	Float
surface	Área de la geometría en hectáreas	Float
territory.community	Código INE de la comunidad autónoma	String
territory.province	Código INE de la provincia	String
territory.township	Código INE del municipio	String
territory.district	Código INE del distrito	String
territory.section	Código INE de la sección	String
location	Centroide de la geometría	GeoJson
bbox	Bounding-box de la geometría	GeoJson
geometry	Geometría del riesgo	GeoJson
risk.string	Valor textual del riesgo	String
risk.value	Valor numérico del riesgo	Integer
structure_type	Indica los distintos usos del suelo que pueden aparecer	String
fcc.arb	Fracción de cabida cubierta arbórea. Valores entre 0 y 100% del conjunto de las especies del estrato arbóreo.	Integer
fcc.mat	Fracción de cabida cubierta matorral. Valora entre 0 y 100% del conjunto de las especies del estrato de matorrales.	Integer
fcc.herb	Fracción de cabida cubierta (5 y 100%) del conjunto de	Integer

	espacies de la formación herbácea.	
fcc.her_ds	Fracción de cabida cubierta (5 y 100%) del conjunto de espacies de la formación herbácea por distribución de especie.	Integer
fcc.total	Fracción de cabida cubierta (5 y 100%) de la suma de las fracciones de las demás FCC.	Integer
tree_form	Formación arbolada que representa la comunidad vegetal arbórea de orden superior con fisiología y biología homogénea.	String
distribution	Formas en las que puede aparecer agrupada la vegetación.	String
species.X.name	Nombre de la especie predominante X	String
species.X.surface_%	Porcentaje de ocupación del suelo de la especie X	Integer
species.X.status	Fase de desarrollo en la que se encuentra la especie X	String
herb_form	Formación arbustiva. Código para indicar si existen dos formaciones arbustivas claramente representadas.	String
avg_height	Altura media del matorral en decímetros.	Integer
attribute	Característica asignada al territorio.	String
fuel_model	Modelo de combustible mayoritario de la tesela.	String
strat_ifn	Estrato del IFN4.	Integer
forest_type	Tipología del bosque predominante.	String
floor.ifn	Clasificación del suelo según Inventario Forestal Nacional.	String
floor.mfe	Uso general mediante una pasarela entre el Tipo Estructural y la clasificación IFN.	String
bio_region	Pertenencia a una de las cuatro regiones biogeográficas existentes en España.	String
lulucf	Uso y cobertura del suelo para el cálculo de las emisiones y absorciones del sector LULUCF.	Class
township_forest_fires.conatus	Número de pequeños fuegos que no se han convertido en incendios en el municipio	Integer
township_forest_fires.fires	Número de incendios en el municipio	Integer

township_forest_fires.total_fires	Número total de conatos + incendios producidos en el municipio	Integer
township_forest_fires.burnt_surface.wooded	Superficie arbolada quemada en hectáreas	Integer
township_forest_fires.burnt_surface.no_wooded	Superficie no arbolada quemada en hectáreas	Integer
township_forest_fires.burnt_surface.total	Total de superficie quemada en hectáreas	Integer

Tabla 2. Diccionario de datos final (colección "forest_fires")

- Colección “seismic”.

Atributo	Descripción	Tipo de dato
_id	Identificador del documento	String
zone	Nombre de la zona sismogénica	String
cortex.type	Tipo de corteza predominante en la zona sismogénica	String
cortex.thickness	Grosor de la corteza en la zona sismogénica	String
cortex.thermal_flow	Flujo térmico en la zona sismogénica	String
cortex.comment	Comentarios asociados a la corteza	String
cortex.shortening_direction	Dirección de acortamiento de la corteza	String
cortex.effort_regime	Régimen de esfuerzo	String
cortex.effort_comment	Comentario sobre régimen de esfuerzo	String
surface.dominant_tectonic	Placa tectónica dominante	String
surface.QAFI	índice QAFI	String
surface.MNSE	índice MNSE	String
seismicity.sample.size	Tamaño de la muestra	String
seismicity.sample.distribution	Distribución normal de la muestra	String
seismicity.sample.comment	Comentario sobre la muestra	String
seismicity.major_earthquakes	Mayor seísmo registrado	Integer
seismicity.magnitude.max	Magnitud máxima registrada	String
seismicity.magnitude.avg	Media de las magnitudes registradas	Float
seismicity.magnitude.std	Desviación estándar de las magnitudes registradas	Float

seismicity.magnitude.criterio	Criterio de la magnitud máxima registrada	String
seismicity.ratio_Mw4	Ratio de seísmos con magnitud 4	Float
seismicity.gutenberg	Escala de Gutenberg	Float
seismicity.def_rupture_mechanism	Mecanismo de ruptura en la zona sismogénica	String
seismicity.comment	Comentario general	String
seismicity.params.prob.years_mw_4	Tiempo estimado en años para que se produzca un seismo de magnitud 4	Float
seismicity.params.prob.years_mw_5	Tiempo estimado en años para que se produzca un seismo de magnitud 5	Float
seismicity.params.prob.years_mw_6	Tiempo estimado en años para que se produzca un seismo de magnitud 6	Float
seismicity.params.prob.years_mw_max	Tiempo estimado en años para que se produzca un seismo de la magnitud máxima registrada	Float
area	Área de la geometría en metros cuadrados	Float
risk.value	Valor numérico del riesgo	Integer
risk.string	Valor textual del riesgo	String
territory.section	Código INE de la sección	String
territory.district	Código INE del distrito	String
territory.township	Código INE del municipio	String
territory.province	Código INE de la provincia	String
territory.community	Código INE de la comunidad autónoma	String
location	Centroide de la geometría	GeoJson
bbox	Bounding-box de la geometría	GeoJson
geometry	Geometría del riesgo	GeoJson

Tabla 3. Diccionario de datos final (colección "seismic")

- Colección “river_flood”.

Atributo	Descripción	Tipo de dato
_id	Identificador del documento	String
territory.community	Código INE de la comunidad autónoma	String
territory.province	Código INE de la provincia	String
territory.district	Código INE del distrito	String
territory.section	Código INE de la sección	String
territory.township	Código INE del municipio	String
year	Año de actualización	Integer
risk.value	Valor numérico del riesgo	Integer
risk.string	Valor textual del riesgo	String
surface	Superficie en hectáreas	Float
area	Área de la geometría en metros cuadrados	Float
location	Centroide de la geometría	GeoJson
bbox	Bounding-box de la geometría	GeoJson
geometry	Geometría del riesgo	GeoJson
estimated_affected_population.section	Población estimada afectada en esa sección	Integer
estimated_affected_population.township	Población estimada afectada en el municipio	Integer
demarcation.name	Nombre de la demarcación	String
demarcation._id	Código de identificación de la demarcación	String

Tabla 4. Diccionario de datos final (colección "river_flood").

- Colección “sea_flood”.

Atributo	Descripción	Tipo de dato
_id	Identificador del documento	String
territory.community	Código INE de la comunidad autónoma	String
territory.province	Código INE de la provincia	String
territory.district	Código INE del distrito	String
territory.section	Código INE de la sección	String

territory.township	Código INE del municipio	String
year	Año de actualización	Integer
risk.value	Valor numérico del riesgo	Integer
risk.string	Valor textual del riesgo	String
surface	Superficie en hectáreas	Float
area	Área de la geometría en metros cuadrados	Float
location	Centroide de la geometría	GeoJson
bbox	Bounding-box de la geometría	GeoJson
geometry	Geometría del riesgo	GeoJson
estimated_affected_population.victims	Victimas estimadas	Integer
estimated_affected_population.township	Población estimada afectada en el municipio	Integer
estimated_affected_population.injured	Heridos estimados	Integer
demarcation.name	Nombre de la demarcación	String
demarcation._id	Código de identificación de la demarcación	String

Tabla 5. Diccionario de datos final (colección "sea_flood).

Tal y como se observa, en todos los diccionarios de datos de las distintas fuentes de riesgos medioambientales existen atributos comunes a todas ellas como son “territory”, “risk”, “area”, “location”, “bbox” y “geometry” que son los atributos más importantes y en los que se ha hecho más hincapié porque son los que utilizaremos más adelante para los entrenamientos de los modelos.

Dentro de estos atributos, uno de los más importantes es el atributo “risk” que es el que nos indica la escala que se utiliza para medir ese riesgo medioambiental. Este campo ha sido inferido en algunas fuentes como en las de inundaciones. A continuación en la siguiente tabla se muestra la escala de valores para los diferentes riesgos medioambientales.

Riesgo Medioambiental	Descripción	Valores de evaluación
Desertificación	Indica el grado de riesgo de desertificación o su no aplicación, correspondiente a los mapas forestales del territorio español	- 6 – Muy Alto - 5 – Alto - 4 – Medio - 3 – Bajo - 2 – Láminas de agua - 1 – Urbano - 0 – Índice de aridez húmedo o subhúmedo
Incendio Forestal	Indica si existe una masa forestal a menos de 5km y su frecuencia de incendio por término municipal en caso afirmativo	- 0 - No existe masa forestal cercana - 1 - De 1 a 5 incendios - 2 - Entre 6 y 10 - 3 - Entre 11 y 25 - 4 - Entre 26 y 50 - 5 - Entre 51 y 100 - 6 - Entre 101 y 500 - 7 - Entre 501 y 1000 - 8 - Entre 1001 y 1511 - 9 - Más de 1511
Sismicidad	Indica el nivel de actividad de la zona sismogénica en la que se localiza	- 4 - Muy Alto - 3 - Alto - 2 - Medio - 1 - Bajo
Inundación costera	Indica la peligrosidad asociada a las zonas inundables correspondientes diferentes escenarios de probabilidad de inundación (períodos de retorno de 100 y de 500 años) calculado a partir de distintas hipótesis de oleaje y nivel de mar con modelos matemáticos.	- 0 - Zona No Inundable - Z.I. Con probabilidad Baja o Excepcional (T=500 años) - 2- Z.I. Con probabilidad Media u Ocasional (T=100 años)
Inundación fluvial	Indica la peligrosidad asociada a las zonas inundables correspondientes diferentes escenarios de probabilidad de inundación (períodos de retorno de 10, 100 y 500 años) calculado a partir de distintas hipótesis de inundación fluvial con modelos matemáticos.	- 0 - Zona No Inundable - 1 - Z.I. Con probabilidad Baja o Excepcional (T=500 años) - 2 - Z.I. Con probabilidad Media u Ocasional (T=100 años) - 3 - Z.I. Con Alta probabilidad (T=10 años)

Tabla 6. Escala de valores para los diferentes riesgos medioambientales (variable "risk").

Capítulo 4. Construcción del modelo.

Como se ha comentado anteriormente, teniendo en cuenta la naturaleza del problema se observa cómo encaja perfectamente con los algoritmos no supervisados, en concreto las técnicas de agrupamiento o clustering, puesto que se desconoce la variable objetivo y lo que se pretende conseguir es hallar agrupaciones de los riesgos medioambientales en las que encasillar cada uno de los municipios de España para relacionar a cada municipio un nivel de riesgo.

Los algoritmos de clustering se utilizan para organizar y comprender la colección de datos , generando grupos de ejemplos parecidos entre sí, clasificando los ejemplos en clases no predefinidas. Uno de los aspectos fundamentales dentro de los algoritmos de clustering es lograr que los ejemplos de cada grupo sean lo más parecidos entre sí y a su vez que los grupos sean lo más diferentes posibles.

En la siguiente ilustración se puede apreciar un ejemplo visual de los fundamentos en los que se basan los algoritmos de clustering.

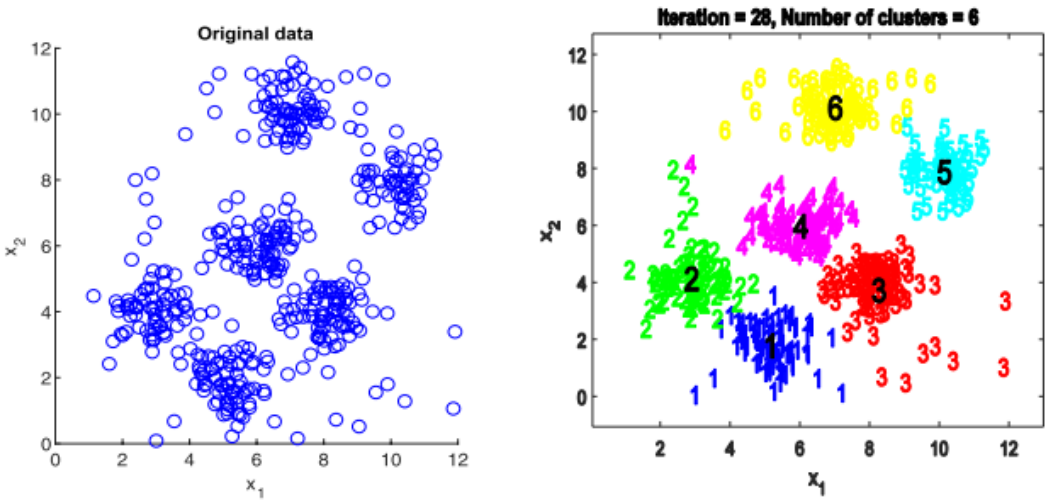

Ilustración 6. Ejemplo gráfico de un algoritmo de clustering. (IEEE – Unsupervised K-Means Clustering Algorithm)

4.1 Definición de la medida.

Es común, dentro de los algoritmos de clustering que utilice como medida la distancia euclídea o la similitud del coseno.

La distancia euclídea (Archana Singh, 2013) se define con la siguiente formula:

$$d_e(U,V) = \sqrt{(u_1 - v_1)^2 + \cdots + (u_n - v_n)^2}$$

Donde *U* y *V* son dos vectores representados en un espacio de n dimensiones y *u* y *v* son cada uno de los elementos de esos vectores. Los valores serán siempre positivos y mayores o iguales que cero (intervalo [0, 1].

Y la similitud del coseno se define como el coseno del ángulo que forman los dos vectores correspondientes a los puntos dada por la siguiente fórmula:

$$sim_c(U,V) = \cos(\widehat{U,V}) = \frac{U \,.\, V}{|U||V|} = \frac{u_1 v_1 + \cdots + u_n v_n}{\sqrt{u_1^2 + \cdots + u_n^2}\sqrt{v_1^2 + \cdots + v_n^2}}$$

Donde el valor obtenido se encuentra entre [-1 , 1] al tratarse de un coseno, siendo calculado como el producto escalar de los vectores *U* y *V*, dividido por el producto de sus normas.

También es importante tener en cuenta que nuestro algoritmo puede sufrir graves alteraciones si no se aplican correctamente estas medidas, puesto que el aplicar las medidas a campos con diferentes rangos de valores se producen salidas con valores muy dispares.

En este proyecto utilizaremos como medida la distancia euclídea. Se nos presentan campos que utilizaremos que se encuentran en diferentes escalas, como los riesgos o la latitud y longitud de los centroides de las geometrías. Para solucionar esto es necesario normalizar correctamente nuestro conjunto de datos.

4.2 Normalización de los datos.

Existen diferentes técnicas de normalización de datos, en este proyecto destacaremos la normalización Min-Max puesto que son conocidos los valores máximos y mínimos de cada variable.

La normalización Min-Max (S. Gopal Krishna Patro, 2015) consiste en que todos los valores numéricos de las diferentes variables se escalen a un rango específico denotado por $[Min_A, Max_A] \rightarrow [N_{Min_A}, N_{Max_A}]$.

La expresión matemática que se utiliza para la obtención de ese nuevo valor es:

$$v' = \frac{v - Min_a}{Max_a - Min_a}\left(N_{Max_A} - N_{Min_A}\right) + N_{Min_A}$$

Donde Max_A y Min_A son los máximos y mínimo de la variable en cuestión. El intervalo final obtenido se moverá en el rango $[-1, 1]$ puesto que es óptimo para algoritmos de aprendizaje basados en distancias.

4.3 Selección de algoritmos.

Dentro del campo de los algoritmos no supervisados, en concreto los de clustering, existen diferentes alternativas que serán sometidas a prueba en este proyecto, para evaluar cuál de ellas nos ofrece mejores resultados.

4.3.1. K-Means

Es el algoritmo de clustering por excelencia (KRISTINA P. SINAGA AND MIIN-SHEN YANG, 2020), el más simple y el más popular. El objetivo de este algoritmo es básicamente agrupar los datos que sean similares en grupos para descubrir patrones en ellos.

Las agrupaciones de datos son comúnmente llamados clusters y son las colecciones de datos similares que serán agregados a un grupo porque contienen ciertas similitudes entre ellos.

El número de clusters (k) será también el número de centroides que definamos en el conjunto de datos. Un centroide es un punto que representa real o imaginariamente el centro de un clúster.

Con cada iteración, este algoritmo irá alojando cada punto al clúster más cercano, manteniendo los clusters lo más pequeños posibles. El significado de “Means” en el nombre del algoritmo se refiere a que el centroide será la media de los puntos que componen ese clúster.

El funcionamiento del algoritmo y su implementación sería la siguiente:

1. Se seleccionan los k puntos iniciales como los centroides. Esta acción se puede dar aleatoriamente o eligiendo los primeros k.

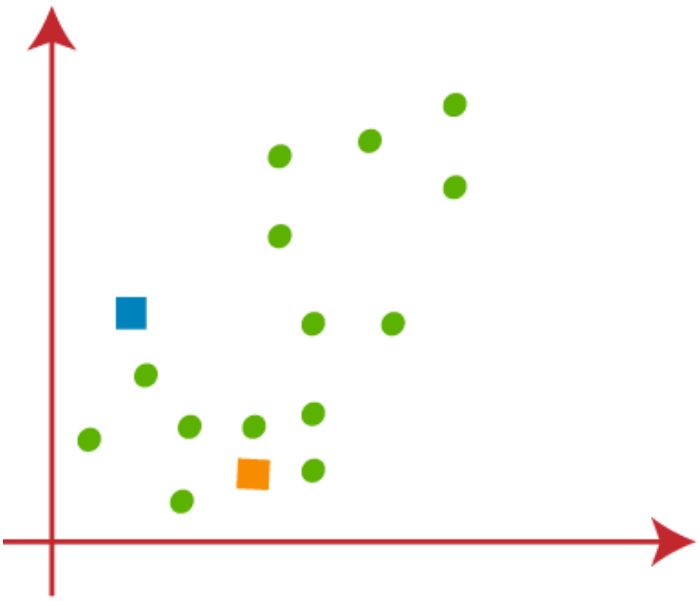

Ilustración 7. K-Means, paso 1

2. Se calcula la distancia euclídea, tal y como se ha especificado en el apartado anterior, de cada punto en el conjunto de datos a los k centroides definidos.

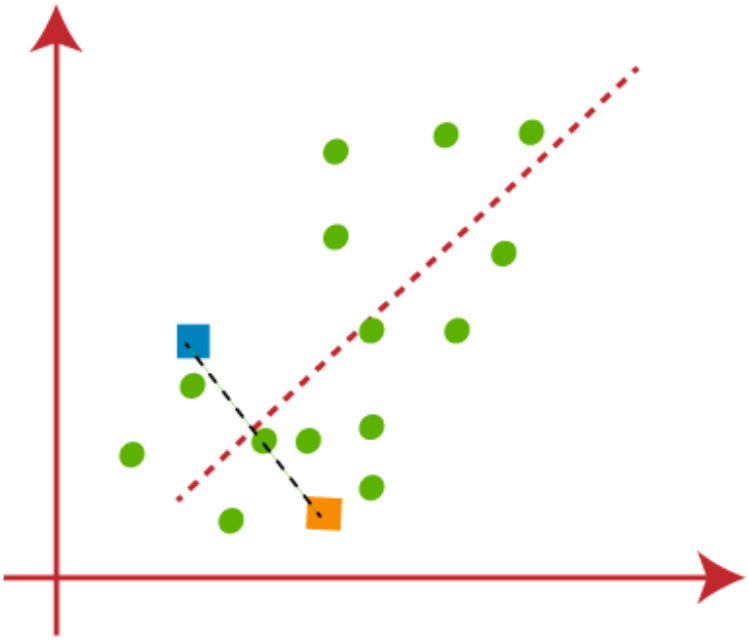

Ilustración 8. K-Means, paso 2

3. A cada punto se le asignará el identificador del clúster con el centroide más cercano al clúster usando la distancia previamente calculada. Matemáticamente correspondería a la siguiente expresión:

$$\arg\min_{c_i \in C} dist(c_i, x)^2$$

Siendo el centroide de cada clúster c_i, x el punto seleccionado y dist la distancia Euclidea.

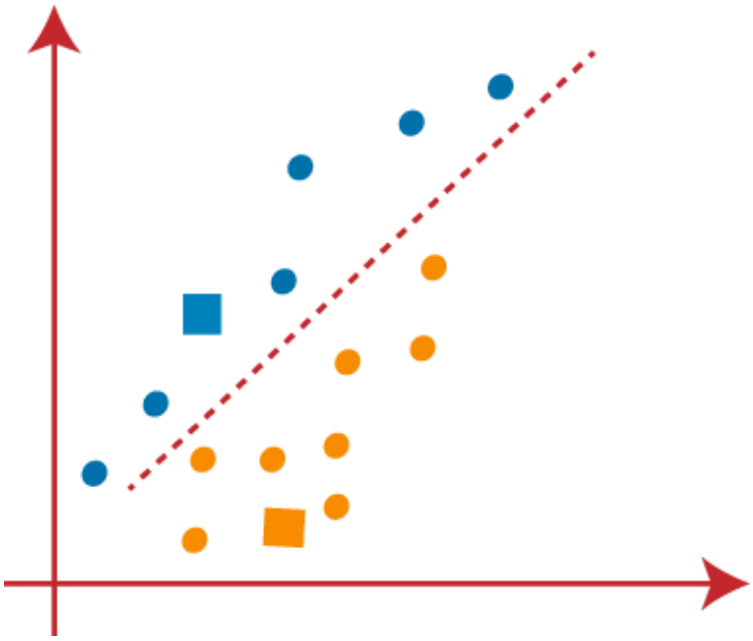

Ilustración 9. K-Means, paso 3

4. Definir el nuevo centroide usando la media de los puntos en cada clúster para definirlo. Matemáticamente correspondería a la siguiente expresión:

$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x_i$$

Donde S_i es el conjunto de todos los puntos asignados a ese cluster y x cada punto de ese cluster.

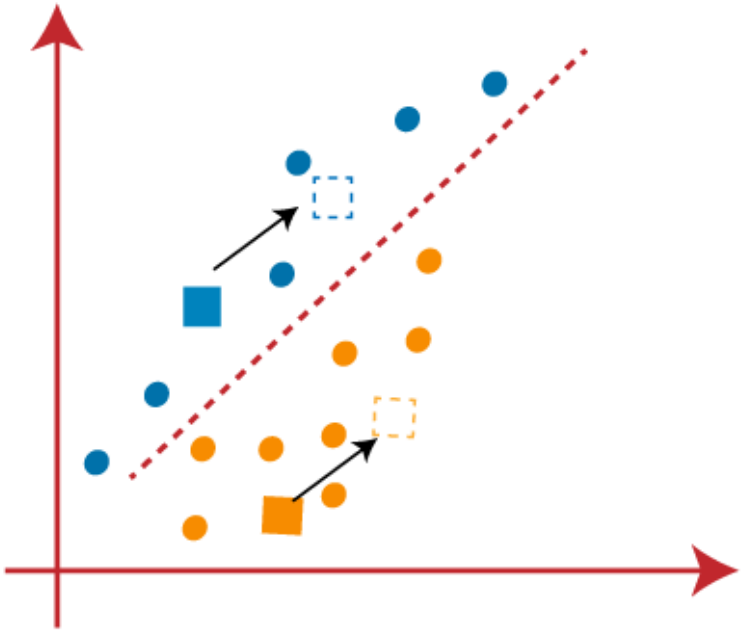

Ilustración 10. K-Means, paso 4

5. Repetir los pasos 2, 3 y 4 hasta que los centroides no cambien (lo que se conoce como loss function) o un número determinado de iteraciones

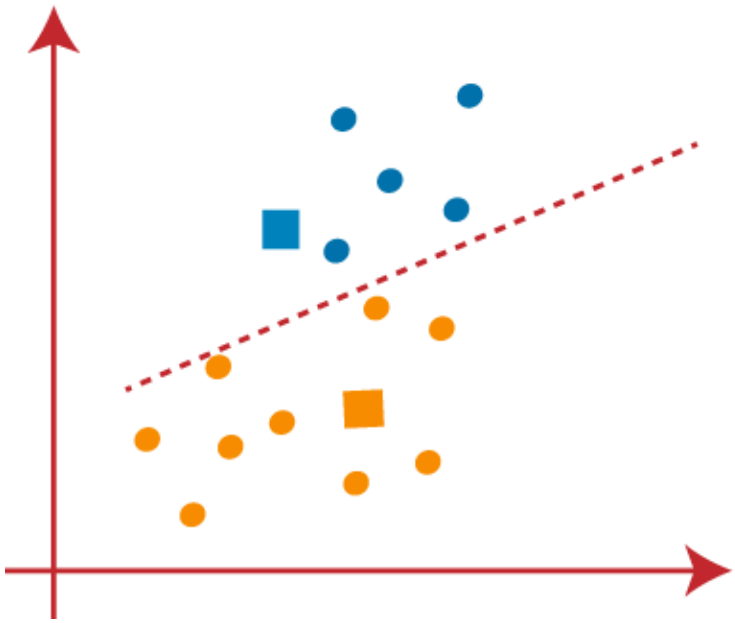

Ilustración 11. K-Means, resultado final

Como principales desventajas, se podrían destacar las siguientes:

- Cuando se forman clusters de tamaños muy distintos, no funciona correctamente.
- Su sensibilidad a outliers es muy alta y hace que sea muy inestable.
- La inicialización del algoritmo puede definir cómo van a ser los resultados.

4.3.2. Gaussian Mixture

El algoritmo de clustering Gaussian Mixture (Douglas Reynolds, 2009), es al igual que al algoritmo K-Means un algoritmo no supervisado, en el que la variable objetivo se desconoce, pero existen algunas diferencias entre ellos.

La principal diferencia con respecto al K-Means, es que este es un algoritmo de clustering cerrado, lo que quiere decir que cada punto se asocia a un clúster o no se asocia, no hay más posibilidades. El algoritmo Gaussian Mixture nos ofrece la posibilidad de asociar una probabilidad que nos dice cuan asociado está un punto a un clúster.

Este algoritmo se basa en una función compuesta por diferentes Gaussianas, cada una de las cuáles identificada como $k \in \{1, \dots, K\}$ siendo K el número de clusters en nuestro dataset.

Cada Gaussiana está compuesta por los siguientes parámetros:

- μ. La media que define su centroide.
- Σ. La covarianza que define su anchura.
- Π. La mezcla de probabilidades que definen como de grande o pequeña será la función Gaussiana.

Este algoritmo se entrena con el algoritmo Expectation-Maximization (EM) (Dempster, 1977) cuya función principal es la de encontrar los valores de los parámetros.

Funciona de la siguiente forma:

Siendo los parámetros de nuestro modelo los siguientes: $\theta = \{\mu, \Sigma, \Pi\}$ se siguen los siguientes pasos:

1. Inicialización de las estimaciones para θ.
2. Expectation Step. Se computan las responsabilidades para cada punto del conjunto de datos. Se utiliza la fórmula:

$$\gamma_i = \frac{\Pi \emptyset(x_i; \mu_2; \Sigma_2)}{(1 - \Pi)\emptyset(x_i; \mu_1; \Sigma_1) + \Pi \emptyset(x_i; \mu_2; \Sigma_2)}$$

   Donde x_i es la observación i, Π es la probabilidad del cluster 2, μ_k es la media del cluster k, Σ_k es la desviación estándar del cluster k, $\emptyset$ es la distribución normal de probabilidad (PDF) y γ_i es una estimación suavizada del valor Δ_i que es la variable 0/1 en la que se evalúa la pertenencia al clúster para la observación i.

3. Maximization Step. Optimiza los parámetros de la distribución utilizando la fórmula de máxima verosimilitud. Cada uno de los parámetros se actualiza conforme a las siguientes formulas:

$$\mu_2 = \frac{\sum \gamma_i \, x_i}{\sum \gamma_i}$$

$$\Sigma_2 = \frac{\sum \gamma_i (x_i - \mu_2)^2}{\sum \gamma_i}$$

$$\Pi = \frac{1}{n} \sum \gamma_i$$

4. Repetir los pasos 2 y 3 hasta que los parámetros converjan a un óptimo local.

4.3.3. Hierarchical Clustering

Es una técnica de clustering muy popular (Frank Nielsen, 2016) que se utiliza para generar una jerarquía de clusters anidados para dividir o unificar iterativamente el conjunto de datos.

Existen dos tipos:

- Aglomerativo. Utilizando esta técnica, inicialmente cada punto es considerado como un clúster individual y en cada iteración, los clusters similares se mezclan con otros clusters hasta que se forma un único clúster o K clusters.
  Funciona de la siguiente forma:
  1. Se calcula la proximidad de cada punto individualmente utilizando por ejemplo la distancia euclídea entre los centroides y se define un umbral de similitud. Se consideran todos los puntos como clusters individuales.
  2. Los clusters similares se unen formando uno.
  3. Se vuelve a calcular la proximidad de los clusters y se vuelven a unir los similares.
  4. Finalmente todos los clusters se unen formando un único clúster

En la siguiente ilustración se puede observar el funcionamiento.

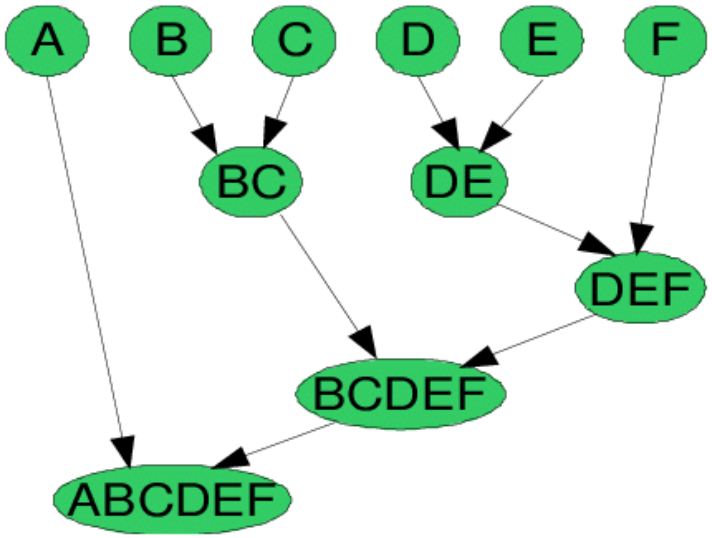

Ilustración 12. Clustering Jerárquico Aglomerativo.

- Divisivo. En pocas palabras, se puede definir los algoritmos de clustering jerárquicos divisivos funcionan de forma opuesta a como lo hacen los aglomerativos. Se parte con un único clúster y en cada iteración se van haciendo divisiones para dividir el clúster en n clusters.

4.4 Tecnologías utilizadas

Para la construcción del modelo se ha hecho uso de las siguientes tecnologías:

- **Lenguaje de programación Python.**

Python (Python, 2022) es un lenguaje de programación que, gracias a su clara sintaxis y fácil lectura, es fácil de aprender y muy versátil. Python está disponible gratuitamente para los sistemas operativos comunes. Se admiten los paradigmas de programación habituales, como la programación funcional u orientada a objetos.

Python tiene la reputación de ser un lenguaje de programación simple y limpio con una estructura clara. Su código es intuitivo de usar y fácil de leer al mismo tiempo. A pesar de su simplicidad, Python ofrece una buena escalabilidad y puede usarse para proyectos de software complejos. Debido a la sintaxis expresiva y minimalista, las aplicaciones se pueden implementar con solo unas pocas líneas de código y una baja susceptibilidad a errores de programación. Para garantizar la simplicidad y la claridad, Python hace uso de muy pocas palabras clave y utiliza la sangría como elemento estructural.

Ilustración 13. Logotipo Python.

- **Jupyter Notebook.**

Jupyter (Jupyter Notebook, 2022) es una iniciativa sin fines de lucro que tiene como objetivo desarrollar y proporcionar software de código abierto y estándares para el trabajo interactivo. Uno de los productos más famosos del proyecto es Jupyter Notebook. Es una aplicación cliente-servidor que funciona con Python como lenguaje de programación para crear y compartir hojas de trabajo interactivas. Otros productos incluyen JupyterLab, JupyterHub y Voilà.

Ilustración 14. Logotipo Jupyter Notebook.

- **Librería de Python Scikit-learn.**

Scikit-learn (Scikit-learn, 2022) es una biblioteca de software para aprendizaje automático. El software está disponible gratuitamente en GitHub bajo una licencia BSD de 3 cláusulas y está diseñado para el lenguaje de programación Python. Varios algoritmos, como agrupación en clústeres, regresión o clasificación están disponibles para el aprendizaje automático. Scikit-learn se utiliza junto con las bibliotecas científicas de Python “SciPy” y “NumPy”. La biblioteca se caracteriza por sus funciones robustas y bien documentadas. Scikit-learn está escrito principalmente en Python.

Ilustración 15. Logotipo Scikit-Learn.

Capítulo 5. Evaluación de modelos y soluciones obtenidas.

En este capítulo se procederá a detallar los resultados obtenidos que resultan de la generación de modelos que han surgido siguiendo los patrones detallados en el capítulo anterior.

Para cada uno de los algoritmos propuestos (K-Means, Gaussian Mixture y Hierarchical Clustering) se realizarán uno o varios entrenamientos con los datos obtenidos mediante el proceso de ETL descrito en el capítulo 3, los cuales requerirán de un preprocesamiento extra para ser adaptados a la entrada necesaria de los algoritmos, hasta alcanzar la solución deseada.

Para alcanzar la solución final, se analizarán los resultados obtenidos en cada entrenamiento con cada modelo y se descartarán aquellos que no cumplan con las expectativas necesarias.

5.1 Medición de resultados

Para evaluar los resultados obtenidos en cada uno de los modelos se utilizarán las siguientes métricas:

- **Tiempo de entrenamiento**. El tiempo de entrenamiento es uno de los parámetros más importantes dentro del paradigma del Big Data, puesto que a medida que la cantidad de datos aumenta, los tiempos de entrenamiento para los modelos de machine-learning también y en ocasiones resulta crucial para poder obtener una buena respuesta del modelo que se diseñe a tiempo.

- **Complejidad del algoritmo**. En muchas ocasiones la balanza se decanta para algoritmos menos complejos aunque con estos se obtengan peores resultados puesto que en algunos proyectos es necesario conocer y comprender como funciona el algoritmo a la perfección para que sea entendido por aquellos que no son expertos en la materia pero que tienen un papel importante dentro de ese proyecto.

- **Recursos consumidos**. Los recursos consumidos son importantes a la hora de determinar en qué máquina se va a entrenar el modelo y si es posible con las características que presente dicha máquina. Muchas veces se elige un algoritmo con el que se obtienen peores resultados pero que es más simple y consume menos recursos porque no se dispone de una máquina lo suficientemente potente como para lidiar con algoritmos más complejos.

- **Coeficiente de Silhouette**. Es una métrica que se utiliza comúnmente en los algoritmos de clustering para la evaluación de la calidad con la que el algoritmo genera los agrupamientos para identificar cuál es el número óptimo de agrupamientos.

  El intervalo de valores en el que trabaja el coeficiente de Silhouette es [1, -1] siendo 1 el mejor valor posible y -1 el peor. Los valores que se aproximan a 0 indican que los

clusters se están sobreponiendo unos a otros. Los valores negativos generalmente indican que una observación ha sido asignada al peor clúster y que existe un clúster más similar al asignado.

La fórmula es la siguiente:

$$s = \frac{b - a}{\max(a, b)}$$

Donde **a** es la distancia media a los elementos del clúster al que se asignaría y **b** la distancia media a los elementos del clúster más cercano.

5.2 Aplicación de coeficientes

Puesto que se asume que cada uno de los riesgos medioambientales no ocurre con la misma frecuencia y no son todos de la misma relevancia, para mejorar la predicción de los modelos se aplicará un coeficiente para cada riesgo medioambiental basado en la probabilidad de que este ocurra y el riesgo que supone.

Según el informe generado en 2020 por la Dirección General de Protección Civil y Emergencias en el que se muestra el número de fallecidos por desastres naturales en España en los intervalos comprendidos entre los años 2000 y 2019 (Dirección General de Protección Civil y Emergencias, 2022) se puede observar cómo los desastres medioambientales con mayor riesgo y que más relevancia tienen en España son las altas temperaturas, que está directamente relacionado con los incendios forestales y las sequías seguido de las inundaciones, que además son los desastres medioambientales que más daños producen y más afectan al sector inmobiliario en España.

Teniendo todo lo anterior en cuenta y además entendiendo que a mayor número de fallecidos mayor es la probabilidad de que ocurra ese desastre medioambiental, repartiremos los coeficientes como se muestran en la Tabla 7.

Riesgo Medioambiental	Coeficiente	Coeficiente Normalizado [0, 1]
Incendio forestal	118	0.173
Desertificación	173	0.253
Sismicidad	9	0.013
Inundación fluvial	209	0.306
Inundación costera	173	0.253

Tabla 7. Coeficientes asignados a los riesgos medioambientales.

Año	Inundaciones	Deslizamiento de terrenos	Incendios forestales	Aludes	Episodios de nieve y frio	Rayos	Vientos fuertes	Altas temperaturas	Terremotos	Temporales marítimos	Total
2000	14	0	6	4	2	4	24	"NA"	0	37	**91**
2001	9	1	1	2	4	4	13	"NA"	0	27	**61**
2002	13	1	6	4	0	2	10	"NA"	0	15	**51**
2003	9	2	11	4	0	1	7	60	0	5	**99**
2004	7	0	4	5	3	4	2	25	0	20	**70**
2005	8	0	19	1	3	1	7	9	0	"NA"	**48**
2006	9	5	8	0	0	1	8	23	0	"NA"	**54**
2007	11	2	1	0	0	1	2	9	0	2	**28**
2008	6	1	1	4	0	2	2	3	0	5	**24**
2009	6	2	11	3	1	1	11	6	0	2	**43**
2010	12	2	9	11	1	1	5	16	0	5	**62**
2011	9	3	12	2	1	1	1	6	9	2	**46**
2012	15	0	10	0	0	1	0	6	0	7	**39**
2013	5	2	1	4	0	1	6	4	0	9	**32**
2014	2	0	4	0	0	1	4	0	0	18	**29**
2015	17	0	3	1	0	1	2	33	0	2	**59**
2016	10	3	1	1	2	1	0	8	0	4	**30**
2017	3	0	6	0	0	1	3	20	0	1	**34**
2018	24	4	1	3	2	0	2	42	0	10	**88**
2019	20	4	3	1	0	0	1	21	0	2	**52**
Total	**209**	**32**	**118**	**50**	**19**	**29**	**110**	**291**	**9**	**173**	**1040**
%	20	3	11	5	2	3	11	28	1	17	100,00

Fuente: Dirección General de Protección Civil y Emergencias. 2020

Ilustración 16. Número de fallecidos por desastres naturales en España, 2000-2019.

5.3 Ranking de clusters

Una vez que se obtengan los clusters mediante el entrenamiento del modelo, se procederá a clasificarlos por orden de riesgo, siendo el clúster 0 el que presenta menor riesgo y el 9 el de mayor riesgo.

Para llevar a cabo esta clasificación se calcularán los estadísticos sobre las observaciones que se asignan a cada clúster calculando la media aritmética para cada tipo de riesgo y multiplicándola por el coeficiente mostrado en la Tabla 7.

La fórmula que se aplicará será la siguiente:

$$R_c = \sum_{i \in c}^{n} I + \mu_i$$

Siendo $R_c$la puntuación de riesgo obtenida por el cluster por la que se clasificará, I el coeficiente aplicado según la Tabla 7, μ_i la media aritmética de las observaciones asignadas al cluster c para el tipo de riesgo i y n los distintos riesgos que existen en ese clúster (n sería un valor comprendido entre [0, 5], siendo 0 si no existe ningún riesgo medioambiental en ese cluster y 5 si existen todos los que se tratan en este proyecto).

5.4 Primer entrenamiento.

5.4.1 Selección de atributos

Uno de los aspectos más relevantes en la construcción de un modelo de machine-learning es la correcta selección de los atributos que se van a utilizar para su entrenamiento, puesto que de ello dependerán aspectos tan importantes como el tiempo de entrenamiento y los resultados obtenidos. Además, es necesario para construir los clusters que los atributos existan en todas las fuentes de riesgos medioambientales para poder compararlas.

En este primer entrenamiento seleccionaremos los siguientes atributos (el significado de cada campo viene detallado en los diccionarios de datos del capítulo 3):

- "area".
- "territory.township".
- "territory.community".
- "territory.province".
- "location.coordinates".
- "risk.value"
- "risk.type". Nueva variable utilizada únicamente para este entrenamiento. Se utiliza un entero comprendido en el intervalo [0, 4] para indicar el tipo de riesgo medioambiental.

Puesto que hay riesgos que afectan a áreas más grandes y otros que afectar a áreas más pequeñas se decide incluir el campo "area" en el entrenamiento para tener ese aspecto en cuenta. Además, en este primer entrenamiento se tendrá muy en cuenta la localización del riesgo, por ello se incluyen los campos "territory" y "location". Por último, "risk.value" es el campo más importante, puesto que nos aporta información sobre la peligrosidad de ese riesgo.

5.4.2 Preprocesamiento de datos.

Es necesario, para poder adaptar los datos a la entrada de los diferentes algoritmos que vamos a utilizar realizar un procesamiento adicional de los datos. Para ello se llevarán a cabo las siguientes transformaciones sobre los datos:

- **De JSON a formato Array de Numpy**. Los algoritmos de Scikit-Learn trabajan con el formato Array de Numpy y los datos sobre los riesgos medioambientales se almacenan en formato JSON que es el que se utiliza en MongoDB.
- **Separación de "location.coordinates" en dos variables independientes**. Dicha variable se almacena en formato GeoJson que viene dado por una lista y para que Scikit-Learn pueda leerlo es necesario dividir los valores de latitud y longitud de las coordenadas en dos campos distintos.
- **Columnas a valores numéricos**. Todas las columnas de "territory" como se especifica en los diccionarios de datos vienen almacenados en formato String y no es compatible con los algoritmos de Scikit-Learn por lo que hay que transformar los códigos de los municipios, provincias y comunidades a enteros.

- **Normalización de "risk.value"**. Como se ha especificado en la Tabla 6, la variable "risk" para cada uno de los riesgos medioambientales viene acotada en rangos de valores distintos, por lo que hay que normalizarlos todos antes de realizar el entrenamiento.
- **Normalización del resto de variables**. Todas las variables se normalizarán en el intervalo [0, 1].

Por último pero no menos importante entrenar el modelo con todos los datos es costoso a la par que ineficiente, puesto que podemos caer en uno de los problemas más comunes dentro del mundo del machine-learning, el sobreajuste. Para evitar este problema es necesario llevar a cabo una validación cruzada y dividir el conjunto de datos en un conjunto para entrenamiento y otro para test. En este caso se reservará un 33% de los datos para test y el resto se utilizarán para entrenar el modelo.

Puesto que las transformaciones que se van a llevar a cabo no son excesivamente complejas, no se hará uso de librerías como Pandas puesto que es una librería muy pesada (computacionalmente hablando) y se estaría añadiendo otra dependencia más a un sitio externo por lo que se desarrollarán métodos en Python para llevar a cabo estas transformaciones.

El conjunto de datos de entrenamiento resultante que se utilizará tiene un tamaño de 845785 registros por 8 variables y el de test un tamaño de 416582 registros por 8 variables.

5.4.3 Selección de metaparámetros.

En un primer momento se llevaron a cabo pruebas iniciales para decidir el número correcto de clusters y se comenzó con una prueba estableciendo 5 clusters pero se observó que los datos se agrupaban demasiado y no era útil a la hora de obtener los clusters asociados a cada observación, porque dos observaciones distintas bastante diferentes acababan en el mismo clúster por lo que se tomó la decisión de aumentar a 10 el número de clusters para que observaciones con un mínimo de diferencia acabasen en clusters distintos y además porque la empresa tenía interés en que ese valor estuviera en el intervalo [0, 9] para clasificar el riesgo del 1 al 10. Esta decisión se mantiene para todo el proyecto.

5.4.4 Análisis de resultados.

Tras realizar el entrenamiento, se procede a analizar los resultados obtenidos utilizando como métricas las expuestas en el apartado 5.1.

Algoritmo	Tiempo de entrenamiento	Recursos consumidos	Coeficiente de Silhouette	Complejidad
K-Means	25.16	5.84 Gb	-0.6595549	**
Gaussian Mixture	84.97	7.06 Gb	-0.4034776	***
Hierarchical Clustering	-	2.60 Tb	-	*

Tabla 8. Resultados primer entrenamiento.

Tal y como se observa en la tabla 8 el consumo de memoria del algoritmo Hierarchical Clustering es muy elevado ($O(n^2)$), significa que crece exponencialmente comparado al tamaño del conjunto de datos. Si es cierto que la computación puede reducirse de $O(n^3)$ a $O(n^2)$ con un único enlace, pero desafortunadamente esto no se aplica a la memoria que consume el algoritmo. Una de las posibles soluciones sería la de reducir el conjunto de entrenamiento y aumentar el de test, pero se descarta esa solución puesto que se quiere entrenar el modelo con una cantidad suficiente de datos por lo que el uso de este algoritmo para construir el modelo queda descartado completamente.

Aunque Gaussian Mixture presenta un Coeficiente de Silhouette más cercano a 0 que el obtenido en K-Means, lo que quiere decir que la calidad de agrupamiento de los clusters es mejor, el tiempo de entrenamiento es superior y el consumo de recursos también es más elevado.

Que el tiempo de entrenamiento sea superior no es tan relevante como para descartar su elección pero el consumo de recursos sí que lo es porque se acerca mucho al límite de memoria disponible en la máquina en la que se va a llevar a cabo el entrenamiento y puede suponer problemas en el futuro, por lo que la opción seleccionada será K-Means.

Como último paso para finalizar este entrenamiento, se realiza el ranking de clusters (apartado 5.3) para clasificarlos conforme a su riesgo medioambiental y se descubre el problema mostrado en la ilustración 17.

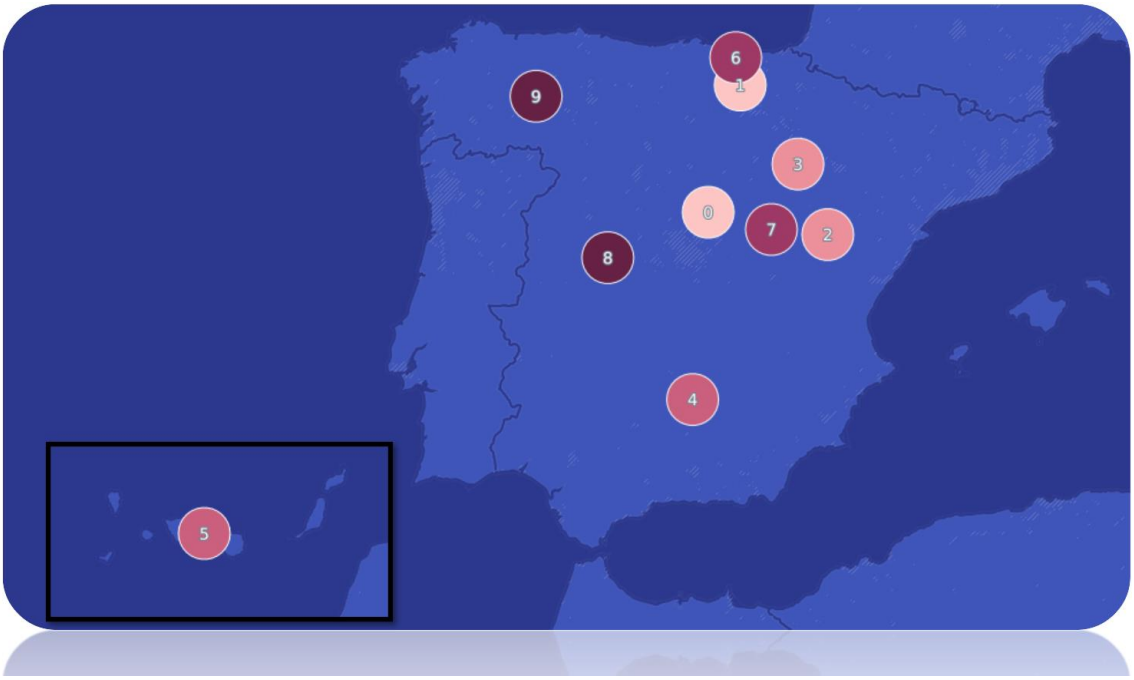

Ilustración 17. Centroides de los clusters obtenidos en el primer entrenamiento.

Puesto que en este primer entrenamiento se han utilizado distintos atributos relacionados con la geolocalización de los riesgos, es posible mostrar sus centroides en un mapa. El problema reside en que en el mapa se observa como el número asociado a cada centroides coincide perfectamente con la imagen que obtendríamos si representamos los riesgos de la fuente de

riesgos por incendio forestal, por lo que estaríamos ante un problema de sobreajuste a esta fuente, que viene dado porque esta fuente es de un tamaño muy superior al del resto y al incluir información sobre su geolocalización está teniendo un peso mucho mayor que el resto y no queremos que tenga tanta relevancia dentro del entrenamiento, por lo que se descartan estos resultados y se procede a realizar un segundo entrenamiento con otras variables.

5.5 Segundo entrenamiento.

5.5.1 Selección de variables.

En este segundo entrenamiento, no se seleccionarán los atributos relacionados con la geolocalización de los riesgos medioambientales para evitar el sobreajuste mencionado en el primer entrenamiento, por lo que se realizará únicamente con los siguientes atributos:

- "area".
- "risk.value"
- "risk.type"

5.5.2 Preprocesamiento de datos.

El preprocesamiento de datos en este entrenamiento será el mismo que en el anterior per únicamente para las variables que se han seleccionado.

- **De JSON a formato Array de Numpy**. Los algoritmos de Scikit-Learn trabajan con el formato Array de Numpy y los datos sobre los riesgos medioambientales se almacenan en formato JSON que es el que se utiliza en MongoDB.
- **Normalización de "risk.value"**. Como se ha especificado en la Tabla 6, la variable "risk" para cada uno de los riesgos medioambientales viene acotada en rangos de valores distintos, por lo que hay que normalizarlos todos antes de realizar el entrenamiento.
- **Normalización del resto de variables**. Todas las variables se normalizarán en el intervalo [0, 1].

El conjunto de datos de entrenamiento resultante que se utilizará tiene un tamaño de 845785 registros por 3 variables y el de test un tamaño de 416582 registros por 3 variables.

5.5.3 Análisis de resultados.

Como se ha especificado anteriormente, Hierarchical Clustering está muy lejos de poder ser viable y por lo tanto se descartará esa opción para este entrenamiento.

Algoritmo	Tiempo de entrenamiento	Recursos consumidos	Coeficiente de Silhouette	Complejidad
K-Means	13.11	5.84 Gb	-0.3132555	**
Gaussian Mixture	25.82	6.31 Gb	-0.1739882	***

Tabla 9. Resultados segundo entrenamiento.

En este segundo entrenamiento, como se puede observar gracias a la reducción del número de variables en el conjunto de datos se ha reducido tanto el tiempo de entrenamiento como los

recursos consumidos en ambos algoritmos y también ha mejorado en ambos el coeficiente de Silhouette.

Llegados a este punto, serían ambas opciones viables, ya que a diferencia del primer entrenamiento, Gaussian Mixture consume menos recursos y existe margen hasta alcanzar el límite de la máquina en la que se realiza el entrenamiento, además el tiempo de entrenamiento es similar al de K-Means en el primer entrenamiento. En este caso utilizaremos para desempatar la complejidad del algoritmo, por lo que finalmente utilizaremos K-Means como algoritmo para el entrenamiento del modelo.

5.6 Visualización de los resultados.

Una vez obtenido el modelo y generados los resultados del entrenamiento, se procede a llevar a cabo el ranking de clusters siguiendo los pasos especificados en el apartado 5.3 para obtener la puntuación que define el riesgo de cada clúster.

El principal objetivo para el que fue construido este modelo fue para utilizar los clusters obtenidos para calcular los municipios de España con mayor riesgo medioambiental, por lo tanto para obtener estos resultados será necesario precalcular un pequeño conjunto de datos con todos los municipios de España obtenido gracias la fuente de datos de incendios forestales proporcionada por CCINIF, que ya teníamos procesada, en la que vienen incluidas dichas geometrías.

Dicha fuente contiene cada municipio y su geometría con las coordenadas y las usaremos para calcular las geometrías que intersectan con las geometrías de los riesgos medioambientales. Una vez obtenido ese pequeño conjunto de datos, usaremos nuestro modelo entrenado para ver en que clúster asigna cada uno de los riesgos que han intersectado con el municipio y calcularemos la media de todas esas asignaciones para establecer un único clúster para el municipio.

Como último paso, haciendo uso de la herramienta CARTO podemos cargar todas las geometrías de los municipios que hemos utilizado y estableciendo un rango de colores para cada clúster podemos obtener el siguiente mapa:

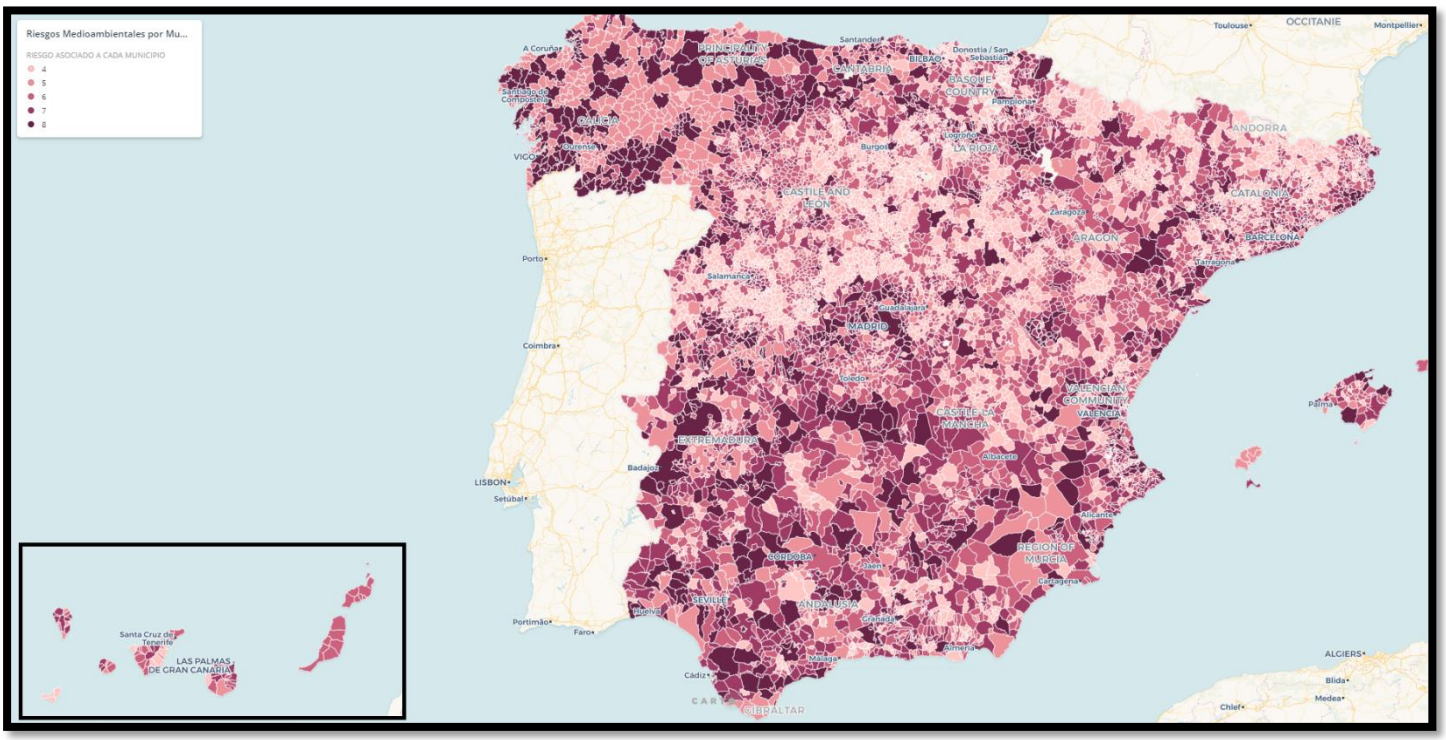

Ilustración 18. Municipios de España con mayores riesgos medioambientales.

En la ilustración 14 podemos visualizar el resultado y la principal aplicación del modelo de machine-learning que hemos creado, el **mapa con los municipios de España que mayores riesgos medioambientales presentan**. Dicho mapa se puede acceder mediante el siguiente enlace:

- https://mappingspain.carto.com/u/mayoralius/builder/f5203ac6-869a-4cab-aa4e-e9e5d91d9216/embed

Como se puede intuir en el mapa, se ha establecido una gama de colores rojos más oscuros para los municipios con mayor riesgo medioambiental y colores más claros para los municipios con menor riesgo medioambiental.

Entendiendo el mapa, se ve como tiene mucho sentido ya que las zonas con mayor riesgo son las zonas de Galicia y Extremadura que son zonas donde los incendios forestales son muy frecuentes y adquieren mucha importancia lo que aumenta mucho el riesgo en esas zonas, en la zona de Valencia y Murcia se ve como la desertificación está afectando de manera grave y en zonas del sur como Huelva y Cádiz las inundaciones han sido el riesgo medioambiental que más estragos ha causado.

Capítulo 6. Herramientas obtenidas.

Además de lo expuesto en el capítulo anterior donde se muestra el principal objetivo y desafío de este proyecto que era el de calcular el índice global de riesgo medioambiental y los municipios de España con mayor riesgo medioambiental, en este capítulo se detallarán otros usos y herramientas obtenidas gracias a la realización de este proyecto.

6.1 Visualizador de riesgos medioambientales.

Gracias a las fuentes de datos que hemos procesado y a las herramientas de tratamiento de datos espaciales de las que se ha hecho uso en este proyecto, podemos simplemente proyectar las geometrías sobre un mapa con CARTO y podemos visualizar individualmente cada uno de los riesgos medioambientales. He de mencionar que puesto que el plan de CARTO que se facilita al estudiante solo incluye un máximo de 100 MB de espacio de almacenamiento, no se puede visualizar en él la fuente de riesgos por incendio forestal puesto que su tamaño supera los 100 MB.

Los códigos de colores que se han utilizado en las ilustraciones que se muestran a continuación siguen el siguiente patrón:

- Colores oscuros: indican un nivel de riesgo medioambiental mayor.
- Colores claros: indican un nivel de riesgo medioambiental menor.

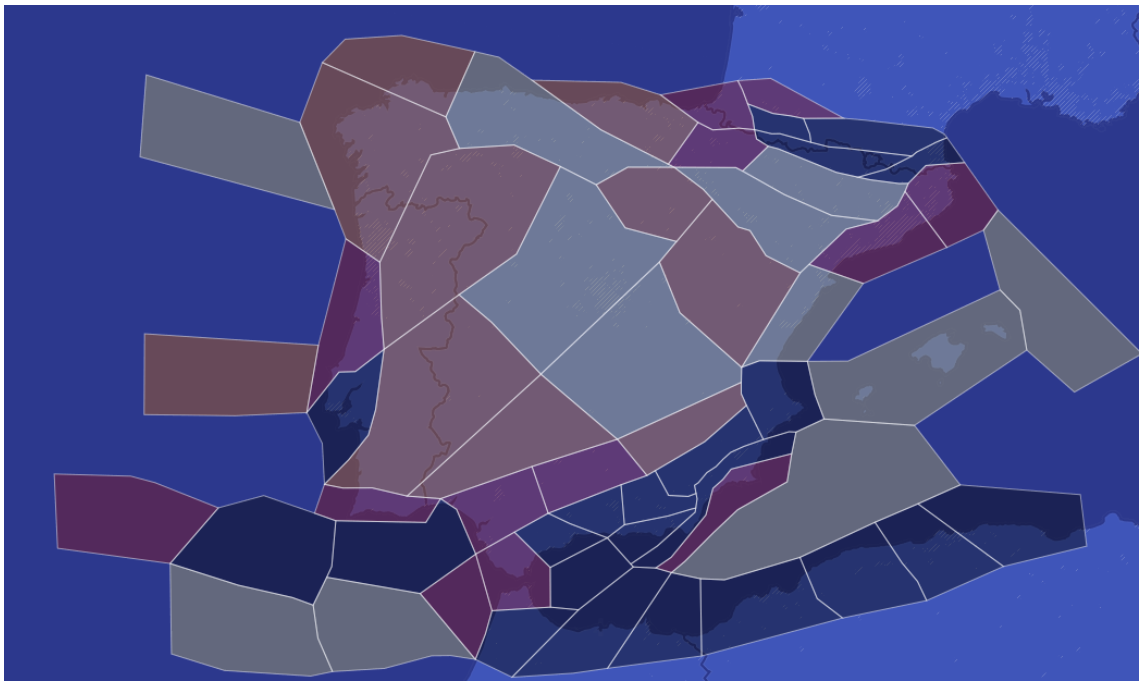

Ilustración 19. Visualización de datos espaciales con CARTO: Riesgo Sísmico.

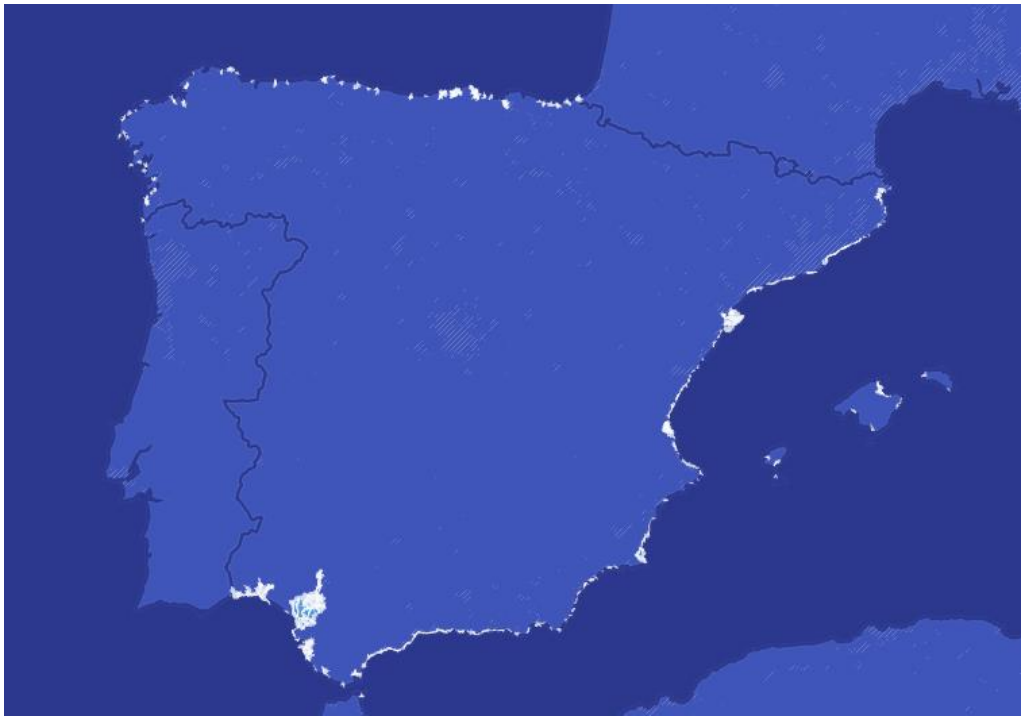

Ilustración 20. Visualización de datos espaciales con CARTO: Riesgo por inundación costera.

Ilustración 21. Visualización de datos espaciales con CARTO: Riesgo por inundación fluvial.

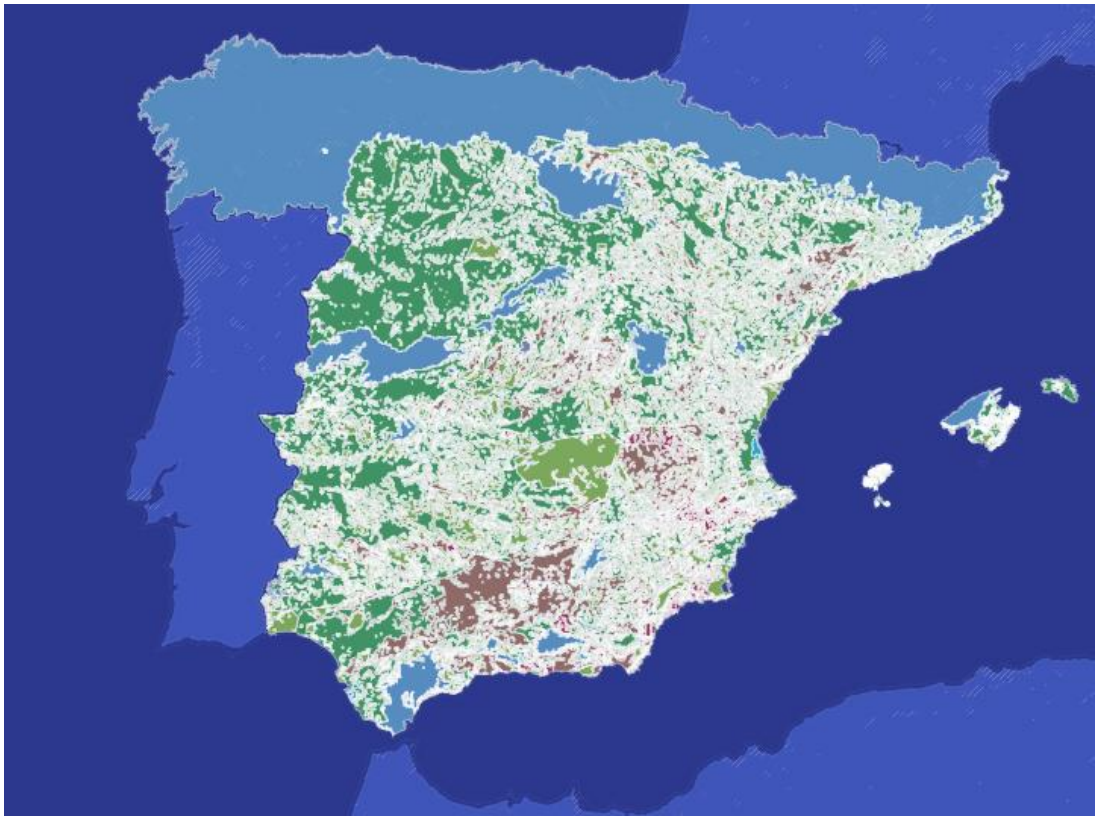

Ilustración 22. Visualización de datos espaciales con CARTO: Riesgo por desertificación.

A continuación se muestra en la siguiente tabla los códigos de colores para cada una de las imágenes:

Fuente	Riesgo	Color
Sísmico	Muy Alto	
Sísmico	Alto	
Sísmico	Medio	
Sísmico	Bajo	
Inundación Costera	Probabilidad 100 años	
Inundación Costera	Probabilidad 500 años	
Inundación Fluvial	Probabilidad 500 años	
Inundación Fluvial	Probabilidad 100 años	

Inundación Fluvial	Probabilidad 100 años	
Desertificación	Muy alto	
Desertificación	Alto	
Desertificación	Medio	
Desertificación	Bajo	
Desertificación	Zonas húmedas	
Desertificación	Láminas de agua	
Desertificación	Urbano	

Tabla 10. Código de colores para las visualizaciones de datos espaciales.

6.2 Endpoint de riesgos medioambientales.

Otra de las herramientas que se ha desarrollado durante este proyecto ha sido un endpoint accesible desde el navegador web en el que introduciendo las coordenadas de la localización deseada en formato latitud y longitud, nos devuelve la información asociada a los riesgos medioambientales que existen en esa localización.

Este software ha sido desarrollado como herramienta interna para la empresa con la que se ha realizado conjuntamente este proyecto y por lo tanto no es accesible por personas ajenas a la organización. Adjunto a este proyecto se incluye únicamente el código que se ha desarrollado para llevar a cabo la función principal que se utiliza en este endpoint.

En las siguientes ilustraciones se incluyen capturas de pantalla del funcionamiento del endpoint.

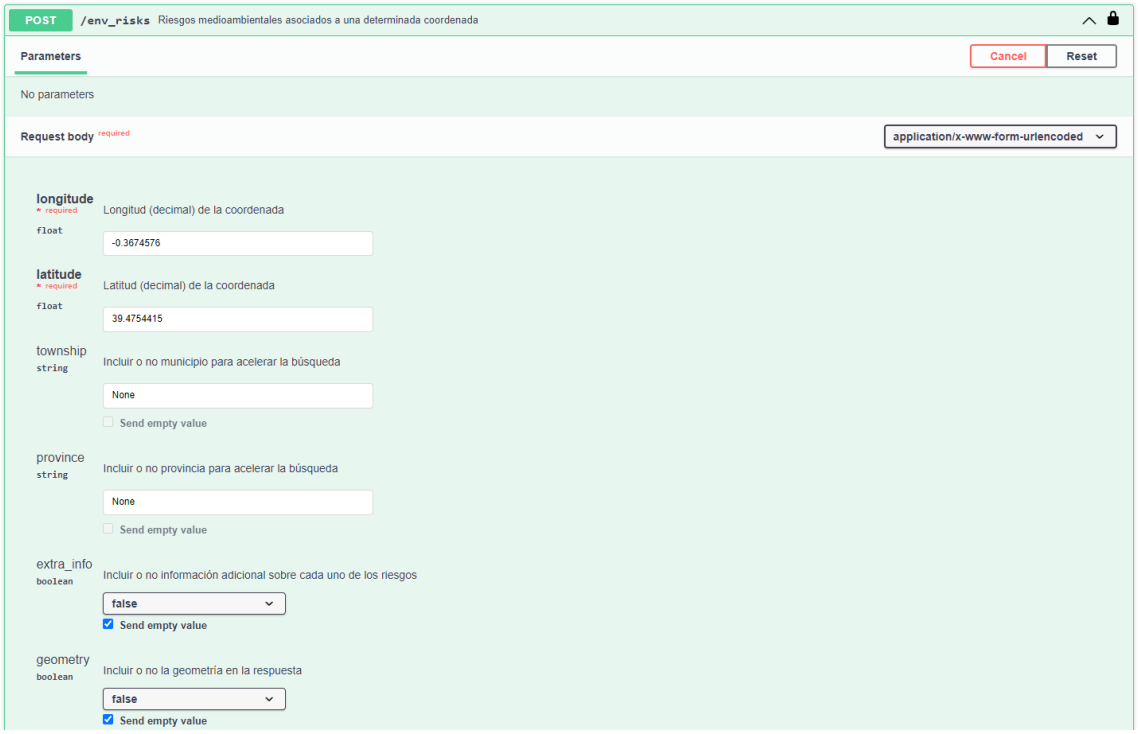

Tabla 11. Endpoint riesgos medioambientales (1).

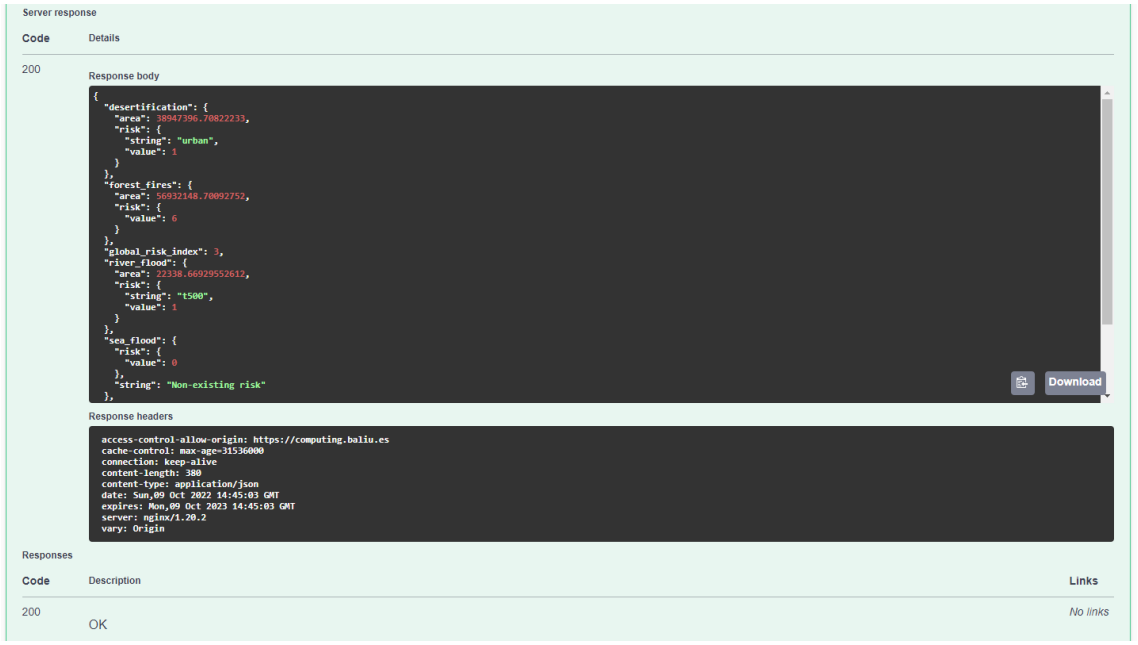

Tabla 12. Endpoint riesgos medioambientales (2).

Capítulo 7. Conclusiones y futuras líneas de trabajo.

Con el simple hecho de acceder a cualquier noticiario o periódico se puede entender la creciente importancia del tema escogido en este proyecto, que cada vez tiene mas preocupados a los inversores, a futuros compradores y en general, a personas interesadas en el sector inmobiliario.

Las posibilidades que este tipo de herramientas nos ofrecen son inmensas, por ello su uso en este tipo de problemas es de gran utilidad, además de ser un tema muy interesante para ser tratado como trabajo de fin de máster y donde se puede exprimir al máximo todos los conocimientos adquiridos durante el curso.

Como se ha visto en el capítulo 5, los resultados obtenidos han sido de gran utilidad ya que se han podido cumplir los objetivos que se pretendían alcanzar en este proyecto, calculando los municipios con más riesgo medioambiental en España. Además, como se expone en el capítulo 6, se han podido desarrollar otras herramientas que también nos permiten analizar el contexto medioambiental en el que se sitúa un inmueble, permitiéndonos conocer de forma rápida y sencilla todos los riesgos medioambientales que podrían afectar a esa localización.

Puesto que ya existe interés en esta herramienta tras mostrarla a ciertos clientes por parte de la empresa con la que se ha colaborado, se podría decir que ha sido un éxito. Todas estas herramientas utilizadas de forma conjunta suponen una clara ventaja para inversores, compradores, bancos… y todas aquellas personas interesadas en el invertir su dinero en el sector inmobiliario.

Como líneas futuras de trabajo, es muy probable que se añadan otras fuentes de riesgos medioambientales para aumentar el alcance de la herramienta y se desarrollen otros endpoint que permitan calcular los riesgos medioambientales de localizaciones de forma masiva para agilizar el proceso cuando se trabaje con volúmenes de datos considerables.

A nivel personal, este proyecto ha supuesto un reto puesto que enfrentarse al desarrollo de una ETL con tantas fuentes y tantos datos que procesar no es tarea fácil y requiere de mucho esfuerzo y tiempo. Además de que existía el miedo de que el proyecto terminase en fracaso puesto que al principio no se conocía exactamente cómo abordar este problema, sobretodo la parte de machine-learning.

Finalmente, gracias a la colaboración de compañeros de trabajo y del tutor, ha sido posible condensar todo el conocimiento aprendido durante el curso en este proyecto y gracias a ello he podido mejorar mis conocimientos y mis habilidades como profesional en este campo.

ANEXOS

1.1 Diccionarios de datos de las fuentes originales

Atributo	Descripción	Tipo de Dato
_id	Identificador de la geometría.	String
AREA	Área en hectáreas de la geometría.	Entero
PERIMETER	Perímetro de la geometría en kilómetros.	Entero
DESER_CLA	Índica el grado o riesgo de desertificación o su no aplicación.	Entero
geometry	Geometría representada mediante puntos con coordenadas.	GeoJson

Tabla 13. Diccionario de datos de la fuente PAND

Atributo	Descripción	Tipo de Dato
NOMBRE	Nombre del municipio.	String
CODIGOINE	Código INE del municipio en cuestión.	Integer
GEOMETRY	Geometría del municipio en cuestión.	GeoJson
NUT2	Nomenclatura de las unidades territoriales.	String
PRO_INE	Código INE de la provincia.	Integer
PRO_N_INE	Nombre de la provincia.	String
CONATOS	Número de pequeños fuegos que no se han convertido en incendios en el municipio.	Integer
INCENDIOS	Número de incendios en el municipio.	Integer
ARBOLADO	Superficie arbolada quemada en hectáreas.	Integer
NOARBOLADO	Superficie no arbolada quemada en hectáreas.	Integer
TOTAL	Total de superficie quemada en hectáreas.	Integer
TOTAL_INCENDIOS	Número total de conatos + incendios producidos en el municipio.	Integer
COM_N_INE	Código INE de la comunidad.	Integer

Tabla 14. Diccionario de datos de la fuente proporcionada por CCINIF.

Atributo	Descripción	Tipo de Dato
OBJECTID	Identificador del objeto.	String
POLIGON_OR	Identificador del polígono.	String

PROV_NOM	Nombre de la provincia en la que se localiza.	String
CCAA_NOM	Nombre de la comunidad autónoma en la que se localiza.	String
SUPERFICIE	Superficie en hectáreas.	Integer
TIPO_ESTRU	Indica los distintos usos del suelo que pueden .aparecer	String
FCCARB	Fracción de cabida cubierta arbórea. Valores entre 0 y 100% del conjunto de las especies del estrato arbóreo.	Integer
FORM_ARB_D	Formación arbolada que representa la comunidad vegetal arbórea de orden superior con fisiología y biología homogénea.	String
DISTRIBUCI	Formas en las que puede aparecer agrupada la vegetación.	String
ESPECIEX	Nombre de la especie predominante X.	String
OX	Porcentaje de ocupación del suelo de la especie X.	Integer
ESTADOX	Fase de desarrollo en la que se encuentra la especie X.	String
FCCMAT	Fracción de cabida cubierta matorral. Valora entre 0 y 100% del conjunto de las especies del estrato de matorrales.	Integer
FORMARBUST	Formación arbustiva. Código para indicar si existen dos formaciones arbustivas claramente representadas.	String
HMMAT	Altura media del matorral en decímetros.	Integer
FCCHER	Fracción de cabida cubierta (5 y 100%) del conjunto de espacies de la formación herbácea.	Integer
FCCHER_DS		
GEOMETRY	Geometría de la masa forestal en coordenadas geográficas.	GeoJson
ATRIBUTO	Característica asignada al territorio.	String
FCCTOT	Fracción de cabida cubierta (5 y 100%) de la suma de las fracciones de las demás FCC.	Integer
MODELOCOMB	Modelo de combustible mayoritario de la tesela.	String
ESTRATOIFN	Estrato del IFN4.	Integer
TIPOBOSQUE	Tipología del bosque predominante.	String
USOIFN	Clasificación del suelo según Inventario Forestal	String

	Nacional.	
USOMFE	Uso general mediante una pasarela entre el Tipo Estructural y la clasificación IFN.	String
REGBIO	Pertenencia a una de las cuatro regiones biogeográficas existentes en España.	String
LULUCF	Uso y cobertura del suelo para el cálculo de las emisiones y absorciones del sector LULUCF.	Class

Tabla 15. Diccionario de datos de la fuente MFE.

Atributo	Descripción	Tipo de dato
ID	Identificador de la zona sismogénica.	String
nombreZona	Nombre de la zona sismogénica.	String
Peligrosiid	Riesgo de la zona sismogénica.	String
tipoCortez	Tipo de corteza predominante en la zona sismogénica.	String
espesorCor	Grosor de la corteza en la zona sismogénica.	String
flujoTermi	Flujo térmico en la zona sismogénica.	String
comentCort	Comentarios asociados a la corteza.	String
dirAcortam	Dirección de acortamiento de la corteza.	String
regimenEsf	Régimen de esfuerzo.	String
comentEsf	Comentario sobre régimen de esfuerzo.	String
tectDom	Placa tectónica dominante.	String
fallasQAFI	índice QAFI.	String
fallasMNSE	índice MNSE.	String
tamanoMues	Tamaño de la muestra.	String
distribMue	Distribución normal de la muestra.	String
comentMues	Comentario sobre la muestra.	String
sismosSign	Mayor seísmo registrado.	Integer
magMaxRegi	Magnitud máxima registrada.	String

magMaxMedi	Media de las magnitudes registradas.	Float
errorMaxMa	Desviación estándar de las magnitudes registradas.	Float
critMaxMag	Criterio de la magnitud máxima registrada.	String
tasaMw_4	Ratio de seísmos con magnitud 4.	Float
parámetro_	Escala de Gutenberg.	Float
MecRotuPre	Mecanismo de ruptura en la zona sismogénica.	String
comentPara	Comentario general.	String
TiempoMw_4	Tiempo estimado en años para que se produzca un seísmo de magnitud 4.	Float
TiempoMw_5	Tiempo estimado en años para que se produzca un seísmo de magnitud 5.	Float
TiempoMw_6	Tiempo estimado en años para que se produzca un seísmo de magnitud 6.	Float
TiempoMax	Tiempo estimado en años para que se produzca un seismo de la magnitud máxima registrada.	Float
Geometry	Geometría de la zona sismogénica en coordenadas geométricas.	GeoJson

Tabla 16. Diccionario de datos ZESIS

Atributo	Descripción	Tipo de dato
ID_POBLACI	Identificador de la geometría.	String
COD_ARPSI	Identificador del área de riesgo potencial significativo de inundación.	String
SUPERFICIE	Superficie en hectáreas.	Float
ID_MUNICIP	Código INE del municipio.	Integer
NOM_MUNICI	Nombre del municipio.	String
N_HAB_MUNI	Número de habitantes en el municipio.	Integer
NUM_AFE_ZI	Población estimada afectada en esa sección.	Integer

NUM_AFE_MU	Población estimada afectada en el municipio.	Integer
FECHA	Año de actualización.	Integer
OTRAS_CON	Comentarios.	String
DEMARCACIO	Nombre de la demarcación.	String
ID_DEMAR	Código de identificación de la demarcación.	String
GEOMETRY	Geometría de la zona en coordenadas geométricas.	GeoJson

Tabla 17. Diccionario de datos inundación fluvial

Atributo	Descripción	Tipo de dato
ID_POBLACI	Identificador de la geometría.	String
COD_ARPSI	Identificador del área de riesgo potencial significativo de inundación.	String
SUPERFICIE	Superficie en hectáreas.	Float
ID_MUNICIP	Código INE del municipio.	Integer
NOM_MUNICI	Nombre del municipio.	String
N_HAB_MUNI	Número de habitantes en el municipio.	Integer
NUM_AFE_ZI	Población estimada afectada en esa sección.	Integer
VICTIM_ORI	Victimas estimadas.	Integer
HERID_ORI	Heridos estimados.	Integer
OTRAS_CONS	Comentarios adicionales.	String

Tabla 18. Diccionario de datos inundación costera.

BIBLIOGRAFÍA

Amazon Web Services (AWS). (2022). *aws.amazon.com*.

Archana Singh, A. Y. A. R. (2013). *K-Means with Three different Distance Metrics*.

Beautiful Soup Documentation. (2022). *beautiful-soup-4.readthedocs.io/en/latest/*.

CARTO. (2022). *carto.com*.

Centro Nacional de Información Geográfica. (2022). *centrodedescargas.cnig.es*.

Consejo Superior de Investigaciones Científicas. (2022). *csic.es*.

DAMA Internacional. (2009). *DAMA-Guía de fundamentos para la gestión de datos*.

Dempster, A. , L. N. , R. D. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society*.

Dirección General de Protección Civil y Emergencias. (2022). *Fallecidos por riesgos naturales en España en 2019*.

Douglas Reynolds. (2009). *Gaussian Mixture Models*.

Frank Nielsen. (2016). *Hierarchical Clustering*.

GeoJSON. (2022). *geojson.io*.

Instituto Geográfico Nacional. (2022). *ign.es*.

Instituto Geológico y Minero de España. (2022). *igme.es*.

Instituto Nacional de Estadística. (2022). *ine.es*.

JavaScript Object Notation. (2022). *json.org*.

Jupyter Notebook. (2022). *jupyter.org*.

Kimball, R. & C. J. (2004). *The Data Warehouse ETL Toolki*.

KRISTINA P. SINAGA AND MIIN-SHEN YANG. (2020). Unsupervised K-Means Clustering Algorithm. *Https://Ieeexplore.Ieee.Org/Stamp/Stamp.Jsp?Tp=&arnumber=9072123*.

Ministerio de Transportes Movilidad y Agenda Urbana. (2022). *mitma.gob.es*.

Ministerio para la Transición Ecológica y el Reto Demográfico. (2022). *miteco.gob.es*.

MongoDB. (2022). *mongodb.com*.

Python. (2022). *python.org*.

Python Requests. (2022). *requests.readthedocs.io/en/latest/*.

S. Gopal Krishna Patro, K. K. S. (2015). *Normalization: A Preprocessing Stage*.

Scikit-learn. (2022). *scikit-learn.org*.