Texto

Descripción generada automáticamente

Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos/ Visual Analytics and Big Data

Biker: un innovador sistema de análisis de datos sobre el ciclismo en el País Vasco

|  |  |
| --- | --- |
| Trabajo fin de estudio presentado por: | Iker Sebastián Pérez |
| Tipo de trabajo: | Desarrollo Software |
| Director/a: | Henry Eduardo Baquero Vega |
| Fecha: |  |

Resumen

La continua evolución en las tecnologías de la información ha posicionado los datos como epicentro de numerosos ámbitos de la sociedad. Desde su recogida y limpieza hasta la aplicación de sus más novedosas técnicas de Inteligencia Artificial (IA), el dato continúa aportando innumerables ventajas a aquellos quienes lo usan en mayor o menor medida. No obstante, ninguna de estas fases del ciclo de vida de los datos ha tenido tanto impacto en los últimos años como la visualización de datos.

Es la visualización de los datos la que ha aumentado su uso exponencialmente, entre otros factores, debido a la aparición de diversas herramientas dedicadas a esta finalidad. Estas herramientas destacan por su alta usabilidad y su intuitiva interfaz, lo que las hace accesibles y fáciles de integrar en los sistemas ya existentes.

En este proyecto, se lleva a la práctica el ciclo de vida completo de los datos. Dicho proceso se inicia con la recogida de datos a través de interfaces de programación de aplicaciones (APIs). Con los datos recogidos, se lleva a cabo la etapa de limpieza y procesamiento para posteriormente almacenar dichos datos. Seguidamente, se aplican técnicas de IA como regresiones o redes neuronales, las cuales son necesarias para la obtención de información relevante para el usuario final. Finalmente, se culmina el ciclo con la visualización de la información obtenida, permitiendo al usuario convertir la información presentada en conocimiento.

La razón principal para la realización de este proyecto es el poder aplicar en un entorno real el ciclo de vida de un conjunto de datos al completo, desde su recogida hasta su visualización. Con ello, se podrá comprender el potencial de cada fase, así como el de las herramientas utilizadas, desde un punto de vista totalmente práctico.

**Palabras clave:** Ciclo de vida del dato, recogida de datos, gestión de datos, análisis de datos, visualización de datos.

Abstract

The continuous evolution in information technologies has positioned data as the core of many areas of society. From its collection and cleaning to the application of the most innovative Artificial Intelligence (AI) techniques, data continues to provide great advantages to those who use it. However, none of these phases of data life cycle has generated as much impact in recent years as data visualization has.

It is visualization of data what has increased its use exponentially, among other factors, due to the appearance of various tools dedicated to this purpose. These tools stand out for their high usability and intuitive interface, which makes them accessible and easy to integrate into existing systems that deal with data.

In this project, the full data lifecycle is implemented. This process begins with the collection of data through application programming interfaces (APIs). Once data is collected, the cleaning and processing stage go on, to subsequently store the data. Next, AI techniques such as regressions or neural networks are applied, which are necessary to get high-relevance information for the end user. Finally, the process ends up with the visualization of the information obtained, allowing the user to turn the information presented into knowledge.

The main reason for going through this project is to be able to apply the life cycle of a complete dataset in a real environment, from its collection to its visualization. With this, it will be possible to understand the potential of each phase, as well as that of the tools used, from a totally practical point of view.

**Keywords**: Data life cycle. Data collection. Data management. Data analysis. Data visualization.

Índice de contenidos

[1. Introducción 1](#_Toc173519173)

[1.1. Motivación 2](#_Toc173519174)

[1.2. Planteamiento del trabajo 3](#_Toc173519175)

[1.3. Estructura del trabajo 4](#_Toc173519176)

[2. Contexto y estado del arte 7](#_Toc173519177)

[2.1. Contexto del problema 7](#_Toc173519178)

[2.2. Estado del arte de herramientas similares existentes 8](#_Toc173519179)

[2.2.1. MeteoRuta 8](#_Toc173519180)

[2.2.2. Comobity 9](#_Toc173519181)

[2.2.3. TuTiempo.net 10](#_Toc173519182)

[2.2.4. Flare 11](#_Toc173519183)

[2.3. Estado del arte de las tecnologías a utilizar 12](#_Toc173519184)

[2.3.1. Recogida de datos 12](#_Toc173519185)

[2.3.2. Limpieza y procesamiento de datos 14](#_Toc173519186)

[2.3.3. Almacenamiento de datos 15](#_Toc173519187)

[2.3.4. Análisis de datos 16](#_Toc173519188)

[2.3.5. Visualización de datos 19](#_Toc173519189)

[2.4. Conclusiones 23](#_Toc173519190)

[3. Objetivos concretos y alcance 25](#_Toc173519191)

[3.1. Objetivo general 25](#_Toc173519192)

[3.2. Objetivos específicos 25](#_Toc173519193)

[3.3. Alcance 26](#_Toc173519194)

[4. Metodología de trabajo y arquitectura 28](#_Toc173519195)

[4.1. Metodología de trabajo 28](#_Toc173519196)

[4.2. Arquitectura 29](#_Toc173519197)

[5. Marco normativo 31](#_Toc173519198)

[6. Desarrollo específico de la contribución 32](#_Toc173519199)

[6.1. Recolección de los datos 33](#_Toc173519200)

[6.2. Limpieza y procesamiento de datos 36](#_Toc173519201)

[6.2.1. Datos ligados al tráfico 37](#_Toc173519202)

[6.2.2. Datos ligados a la meteorología 39](#_Toc173519203)

[6.3. Almacenamiento de datos 43](#_Toc173519204)

[6.4. Análisis de datos 44](#_Toc173519205)

[6.4.1. Análisis exploratorio de datos (EDA) 44](#_Toc173519206)

[6.4.2. Análisis de los subconjuntos de datos 45](#_Toc173519207)

[6.4.3. Exportar resultados del análisis 51](#_Toc173519208)

[6.5. Visualización de datos 52](#_Toc173519209)

[7. Código fuente y datos analizados 53](#_Toc173519210)

[7.1. Repositorio 53](#_Toc173519211)

[7.2. Código fuente 53](#_Toc173519212)

[7.3. Datos Analizados 53](#_Toc173519213)

[8. Plan de trabajo y gestión del proyecto 54](#_Toc173519214)

[8.1. Planificación del proyecto 54](#_Toc173519215)

[8.1.1. Cronograma 54](#_Toc173519216)

[8.1.2. Análisis DAFO y CAME 54](#_Toc173519217)

[8.2. Problemas encontrados y desvíos 54](#_Toc173519218)

[8.3. Limitaciones del proyecto 54](#_Toc173519219)

[9. Resultados 55](#_Toc173519220)

[9.1. Indicadores del proyecto 55](#_Toc173519221)

[9.2. Resultados de los algoritmos 55](#_Toc173519222)

[9.3. Indicadores del proyecto 55](#_Toc173519223)

[10. Conclusiones 56](#_Toc173519224)

[11. Trabajo futuro 57](#_Toc173519225)

[Referencias bibliográficas 58](#_Toc173519226)

[Glosario 65](#_Toc173519227)

[Lista de siglas y acrónimos 67](#_Toc173519228)

[Anexo A. Privacidad y protección de datos 68](#_Toc173519229)

Índice de figuras

[**Figura 1.** *Representación visual de las fases del ciclo de vida del dato.* 4](#_Toc173573057)

[**Figura 2.** *Captura de la web MeteoRuta.* 8](#_Toc173573058)

[**Figura 3.** *Captura de la aplicación Comobity.* 10](#_Toc173573059)

[**Figura 4.** *Captura de la web TuTiempo.net.* 11](#_Toc173573060)

[**Figura 5.** *Ejemplo de visualización con librerías de Python.* 19](#_Toc173573061)

[**Figura 6.** *Ejemplo de visualización con D3.* 20](#_Toc173573062)

[**Figura 7.** *Ejemplo de visualización con Power BI.* 21](#_Toc173573063)

[**Figura 8.** *Arquitectura de las fases del ciclo de vida del dato.* 29](#_Toc173573064)

[**Figura 9.** *Colecciones de la base de datos “TFM” en MongoDB* 43](#_Toc173573065)

[**Figura 10.** *Ejemplo de una matriz de correlación* 45](#_Toc173573066)

[**Figura 11.** *Clustering de los flujos de tráfico* 46](#_Toc173573067)

[**Figura 12.** *Resultados de las predicciones de la calidad del aire* 48](#_Toc173573068)

[**Figura 13.** *Clustering de las incidencias* 49](#_Toc173573069)

[**Figura 14.** *Resultados de las predicciones de los datos meteorológicos* 51](#_Toc173573070)

[**Figura 15.** *Resultados de las predicciones en la base de datos* 52](#_Toc173573071)

[**Figura 16.** *Resultados del clustering en la base de datos* 52](#_Toc173573072)

Índice de tablas

[Tabla 1. *Tabla comparativa de las herramientas analizadas.* 12](#_Toc169627483)

[Tabla 2. *Tabla comparativa de herramientas de visualización.* 21](#_Toc169627484)

# Introducción

La evolución tecnológica de los últimos años ha propulsado la investigación de muchos ámbitos que afectan de forma directa al ser humano. Estos avances tecnológicos han tenido un considerable impacto en la sociedad, cambiado hábitos tanto a nivel individual, con el aumento de dispositivos Internet of Things (IoT), como en el ámbito industrial, donde empresas multinacionales se han visto obligadas a digitalizar.

En todos estos casos, el dato ha ganado relevancia, convirtiéndose en la unidad esencial para el ser humano en su día a día. Por ello, y sumado a que la transición a la era digital no ha hecho más que comenzar, se ha querido enfatizar la relevancia del dato, posibilitando así un proyecto donde se trabaje cada una de sus fases del ciclo de vida en mayor detalle.

En concreto, se presenta el proyecto de “Biker”, una aplicación para todo aquel aficionado al ciclismo en el País Vasco, donde se ofrece al usuario final la posibilidad de conocer cierta información de un alto interés a la hora de practicar dicho deporte. En la aplicación, se tratan conjuntos de datos que recogen las incidencias existentes en las carreteras del País Vasco, así como el número de vehículos que transitan las mismas. También se utilizan datos del clima para analizar ambos conjuntos de manera unificada y encontrar relaciones e información que aporte un valor añadido a los ciclistas.

Dicho tratamiento de los datos comienza con la extracción de éstos desde las APIs públicas de Open Data Euskadi y Euskalmet. Desde estas entidades, se facilita la recogida de datos públicamente, siendo los datos relacionados con el tráfico y el clima los almacenados para este proyecto. Entre todos estos datos, se seleccionan los que interesen especialmente para llevar a cabo el análisis. Posteriormente, se formatean los datos seleccionados y se guardan en distintas colecciones de la base de datos de MongoDB en local. Esta base de datos realiza la función de almacén de datos (Data Warehouse).

Seguidamente, se explotan los datos con tal de extraer las relaciones que se consideren más relevantes. De esta forma, se obtienen predicciones de índices tanto meteorológicos como de tráfico, para estimar los potenciales valores de los días siguientes. Estas predicciones son posibles gracias a la aplicación de varias técnicas de IA, entre las que destacan la regresión lineal, Random Forest y la red neuronal.

Finalmente, se visualiza la información extraída de todo el análisis, utilizando herramientas de Business Intelligence (BI). Como línea futura, se podría valorar la creación de una aplicación móvil, de cara a ganar en accesibilidad y manejabilidad en el uso de esta herramienta.

En definitiva, mediante este proyecto se propone realizar el desarrollo del ciclo de vida completo de un conjunto de datos referente al tráfico y el clima en el País Vasco. Dicho desarrollo, se enfoca al sector ciclista, cuyos usuarios pueden obtener información novedosa y útil, gracias a las diferentes técnicas de IA utilizadas. Asimismo, resaltar que una de las finalidades principales del proyecto consiste en facilitar la información recogida a través de las visualizaciones a los usuarios finales, siendo necesario la realización de todo el proceso previo, y permitiendo interpretar a su vez los diferentes índices e indicadores que resumen el estado del clima y tráfico en el País Vasco.

## Motivación

Las principales razones para realizar este proyecto se explican a continuación:

Durante estos últimos años, el auge en la explotación de los datos ha incrementado de forma exponencial, siendo un aspecto esencial en la transformación digital que se está llevando a cabo en cada vez más empresas. Esto se debe a los tantos beneficios que una buena gestión de los datos puede aportar, siendo clave en la toma de decisiones de cualquier ámbito. Por ello, mediante este proyecto se ha decidido profundizar en la totalidad de las fases que conforman el ciclo de vida de los datos, tomando un conjunto de datos y aplicando su ciclo de vida al completo, desde la recogida y almacenamiento hasta su análisis y visualización.

Asimismo, el tema del que trata el conjunto de datos y el enfoque que se le proporciona tiene cierta afección e interés personal, ya que la aplicación toma como base el ciclismo, un deporte muy común en la región del País Vasco. Al ser nativo de esta zona geográfica y haber podido vivir de primera mano todo el impacto que dicho deporte genera en la población vasca, supone un aliciente más para el desarrollo de esta idea, pudiendo tal vez en un futuro generar una aplicación con un alcance mayor.

Por otra parte, la decisión de realizar este proyecto también engloba la linealidad de trabajar sobre la seguridad, no en cuanto a sistemas informáticos se refiere en este caso, si no en la seguridad física de los ciclistas. La seguridad es un tema que siempre ha sido relevante y que se ha ido adaptando a los continuos cambios que ha sufrido cualquier sector. Sin embargo, en la gran mayoría de circunstancias ha sido un aspecto infravalorado, siendo un variable más de cualquier ecuación cuando en realidad debería ser el cimiento principal sobre el cual se debería construir. En cuanto a la aplicación se refiere, se permite la consulta de tanto el tráfico como la climatología, lo que deduce el nivel de seguridad que se estima para la práctica del ciclismo bajo dichas condiciones.

Algo similar ocurre con la actividad física, ya que mediante este proyecto se quiere propulsar e incentivar sobre todo a las jóvenes generaciones a la práctica del deporte, en este caso el ciclismo. Mediante herramientas de este calibre, se facilita la planificación de las sesiones que se quieran realizar y permitir un acercamiento mayor hacia el mismo.

En líneas generales, una de las ideas principales que recogen los puntos previamente comentados hace énfasis en la relevancia de aplicar prácticamente un proceso que tiene un uso muy notorio en la sociedad actual, como es el ciclo de vida de los datos. Dicho proceso aporta un gran valor añadido en un amplio abanico de escenarios, el cual se prevé que siga ampliándose en los próximos años. Asimismo, se destaca la importancia de fomentar la práctica del deporte, así como de hacerlo en un entorno seguro.

## Planteamiento del trabajo

Ante la carencia de una aplicación dedicada al ciclismo que integre tanto información meteorológica como propiedades del tráfico, se ha identificado una considerable necesidad en la sociedad, la cual se intenta cumplimentar a través del presente proyecto.

Consecuentemente emerge “Biker”, un proyecto que busca facilitar la planificación en la práctica del ciclismo mediante la fácil obtención de información sobre el clima y el tráfico en el País Vasco. Se presenta una solución que cubre una necesidad existente en la actualidad, la cual es solventada hoy en día por aplicaciones de una índole más genérica o mediante varias aplicaciones que se complementan entre sí. Además, es destacable el hecho de que la información a compartir se pretende mostrar mayoritariamente a través de representaciones gráficas, consiguiendo una mayor sencillez en su interpretación.

Por consiguiente, a la hora de querer practicar ciclismo, el usuario es capaz de ahorrar tiempo en la planificación, permitiendo por lo tanto su práctica durante más tiempo. Asimismo, se obtiene la información necesaria para realizar este deporte siendo conocedor de las condiciones y pudiendo evaluar el riesgo de su práctica.

## Estructura del trabajo

Mediante este proyecto se trabaja cada una de las fases del ciclo de vida de los datos. Se comienza con la investigación y el análisis de varias herramientas existentes, las cuales tienen cierto grado de similitud respecto al proyecto presentado. En esta misma sección se valoran todas las posibles tecnologías a utilizar para su desarrollo, justificando la elección de cada una de las utilizadas.

En la sección posterior, se define el objetivo principal y los objetivos secundarios del proyecto, englobando desde la exploración y recolección de datos hasta su visualización. Seguidamente, se presenta la metodología utilizada para la gestión del proyecto, con sus respectivas herramientas de gestión.

El grosso del proyecto se divide mediante subapartados, uno por cada fase del ciclo de vida del conjunto de datos: captura, limpieza y procesamiento, almacenamiento, análisis y visualización.

**Figura 1.** *Representación visual de las fases del ciclo de vida del dato.*

Diagrama

Descripción generada automáticamente

Fuente: Javier Esmerado, Medium.

Respecto a la imagen previa, la única diferencia existente es que las fases de almacenamiento y tratamiento están intercambiadas. Como se menciona en más de una ocasión a lo largo del documento. Este hecho se debe a la conformación de un Data Warehouse en vez de un Data Lake, almacenando únicamente los datos que son de interés en lugar de la totalidad de los mismos, habiendo sufrido una transformación previa a su almacenamiento.

Entrando en detalle, en la primera fase se comienza explicando la recogida de datos desde unas APIs públicas de Open Data Euskadi y Euskalmet, mediante las cuales se recogen datos meteorológicos y relacionados con el tráfico.

En la segunda fase se presenta el procesamiento y la limpieza de los datos recogidos. De todos los datos devueltos por las APIs, se seleccionan los que mayor relevancia tengan para la aplicación. Una vez seleccionados y previo a la inserción en la base de datos de MongoDB en local, se realiza la limpieza y el tratamiento de estos, con el fin de almacenar únicamente datos con valor y que por consiguiente generen un Data Warehouse.

En la fase de almacenamiento, se muestra el guardado de los datos seleccionados y su estructuración en distintas colecciones, lo que al ser una base de datos no relacional simula las diferentes tablas de lo que sería una base de datos estructurada (SQL).

Para la cuarta fase, fase de análisis, se comienza realizando un análisis exploratorio de datos (EDA), con el fin de entender en mayor profundidad los valores que dichos datos pueden tomar. Tras realizar el análisis, se plantean diversas técnicas de IA para poder obtener predicciones de los diferentes indicadores meteorológicos y referentes al tráfico que podrían existir durante los próximos días. Para ello, se utilizan varios tipos de regresiones y redes neuronales. Se tienen en cuenta todos los históricos registrados en la base de datos, así como diferentes atributos que ponderan según su correlación y permiten realizar el entrenamiento óptimo para el modelo en cuestión.

Una vez obtenidas las predicciones, se inicia la fase de visualización. El desarrollo de esta fase se lleva a cabo utilizando herramientas de BI como Power BI. De esta forma, se pueden visualizar los datos que mayor valor aporten al proyecto. Asimismo, los datos son ubicados en varios dashboards, de acuerdo con el área al que pertenezca cada dato: tráfico, incidencias en carretera, temperatura, precipitaciones…

Tras el desarrollo del proyecto, se detallan aspectos de una índole más genérica como son el plan de trabajo y gestión del proyecto. En este apartado, se desglosa la planificación realizada para poder cumplir con el proyecto además de presentar las diferentes limitaciones y problemáticas que han podido existir durante el transcurso del mismo.

Finalmente, se exponen los resultados y conclusiones obtenidas, permitiendo conocer el resultado final del proyecto. Se concluye con posibles líneas futuras de trabajo que tratar sobre el proyecto.

# Contexto y estado del arte

## Contexto del problema

Cada año, decenas de personas pierden la vida o sufren graves lesiones practicando ciclismo en las carreteras del País Vasco. Estas tragedias no solo representan pérdidas humanas, sino que evidencian la preocupante situación existente en cuanto a la seguridad sobre la bicicleta. Como varias noticias avalan, el número de accidentes de ciclistas en el País Vasco ha aumentado considerablemente en los últimos años (Press, 2023). Esto lo convierte en una situación insostenible, donde los datos registran que casi la mitad de los fallecidos en accidentes de carreteras vascas son ciclistas, motoristas o viandantes («45 personas murieron en las carreteras en Euskadi en 2023, casi la mitad motoristas, viandantes o ciclistas», 2024). Con estos datos, queda claro que se deben tomar varias medidas de concienciación y seguridad, para posibilitar la práctica de este deporte de forma segura, lo que en la actualidad no es un hecho.

El caso más reciente ocurrió apenas un año vista, con el fallecimiento de un joven ciclista de 19 años cerca de la capital vasca al ser arroyado por un vehículo (Eurosport, 2023). Aunque no todos los incidentes que ocurren se deben a esta causa, la inmensa mayoría de los accidentes se pueden agrupar en que han sido causados por el tráfico (Michael T. Gibson et al., 2020). No obstante, hay un segundo grupo que recoge gran cantidad de accidentes ciclistas también como son las circunstancias meteorológicas (Cesvimap, 2021).

Ante esta situación, una mayor prevención es innegociable, así como una mejor planificación de las rutas a realizar o del momento en el que llevarlas a cabo. Para ayudar en el objetivo de disminuir y quizás algún día erradicar estos accidentes, se propone este proyecto. Este, se puede entender como una herramienta adicional para aquellos ciclistas que quieran tomar un mayor control sobre las variables altamente complicadas de controlar que afectan durante el transcurso de esta actividad.

## Estado del arte de herramientas similares existentes

Al realizar un estudio de variantes similares al proyecto presentado, se han encontrado varias herramientas ya existentes que comparten algunas de las funcionalidades que se buscan obtener. No obstante, cabe mencionar que ninguna de las alternativas encontradas ofrece la visión completa que se pretende obtener con “Biker”. A continuación, se presentan las alternativas con mayor similitud y relevancia a la propuesta realizada.

### MeteoRuta

Trata de una colaboración entre la Dirección General de Tráfico (DGT) y la Agencia Estatal de Meteorología (AEMET). Dicha colaboración surgió a inicios del año 2014 y puso a disposición de los ciudadanos españoles una web en la que poder visualizar el estado de las carreteras en tiempo real (Agencia Estatal de Meteorología, s. f.). Esta web, salió a la luz como consecuencia del fuerte temporal ocurrido durante esas fechas («Mapa interactivo de la DGT y AEMET, esencial para planificar tu ruta cuando hace mal tiempo», 2023). Actualmente, esta web sigue en activo.

Esta herramienta se puede utilizar de diferentes formas, pudiendo analizar una ruta concreta o bien ir explorando zonas del mapa manualmente, conociendo información más específica a medida que se recorren las diferentes vías.

Una captura de pantalla de una computadora

Descripción generada automáticamente**Figura 2.** *Captura de la web MeteoRuta.*

Fuente: Elaboración propia.

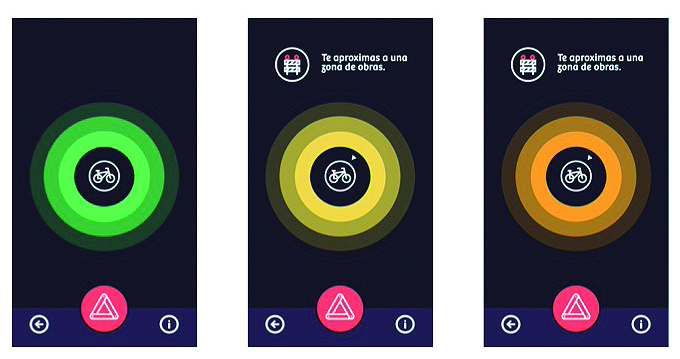
En lo que a la comparativa con “Biker” respecta, se puede considerar como la alternativa con mayor similitud encontrada, dado que engloba tanto un análisis meteorológico como datos de tráfico en una misma aplicación. Sin embargo, existen diferencias considerables como que el enfoque de MeteoRuta no está principalmente ligado con las bicicletas, que se engloba un territorio mucho mayor o que no se realiza un análisis tan detallado del tráfico como pretende realizar “Biker”. Además, otro aspecto a considerar es que MeteoRuta únicamente trabaja sobre una visualización de un mapa donde poder interactuar, mientras con “Biker” se compone de varios grafismos que aporten información variada.

### Comobity

Otra aplicación que puede tener relación en lo que al ámbito de movilidad se refiere, es la aplicación de Comobity («Comobity, la App “segura” para ciclistas», 2015). Comobity es una aplicación tanto para peatones, ciclistas y vehículos donde priorizan a las bicicletas y los peatones por delante de los vehículos de motor.

La finalidad principal de esta aplicación va dirigida hacia la búsqueda de una mayor seguridad para cualquiera de los individuos que transiten en estos tres medios de transporte. Para ello, el principio a seguir se basa en que cualquier individuo debe llevar la aplicación de Comobity abierta, pudiendo conformar así una red de usuarios con sus respectivas ubicaciones. De esta forma, ante potenciales incidentes como acercamientos imprudentes o zonas conflictivas, Comobity puede avisar al usuario de dicho riesgo. De igual forma, en caso de detectar alguna alteración en la ruta, como puede ser una zona de obras, se notificará al usuario a medida que se aproxime a la zona conflictiva.

En comparación con “Biker”, el análisis que Comobity realiza sobre el tráfico tiene cierta relación, a pesar de que es una herramienta pensada para ser utilizada en tiempo real. Por el contrario, la propuesta que se realiza en este documento tiene como finalidad ayudar en la planificación y que el usuario sea conocedor de las condiciones existentes previo al inicio de la práctica del deporte, por lo que no requiere esa característica de actualización de datos inminente. Se ha de mencionar también, que Comobity no hace ninguna referencia al análisis meteorológico, lo que en “Biker” toma un importante peso.

**Figura 3.** *Captura de la aplicación Comobity.*

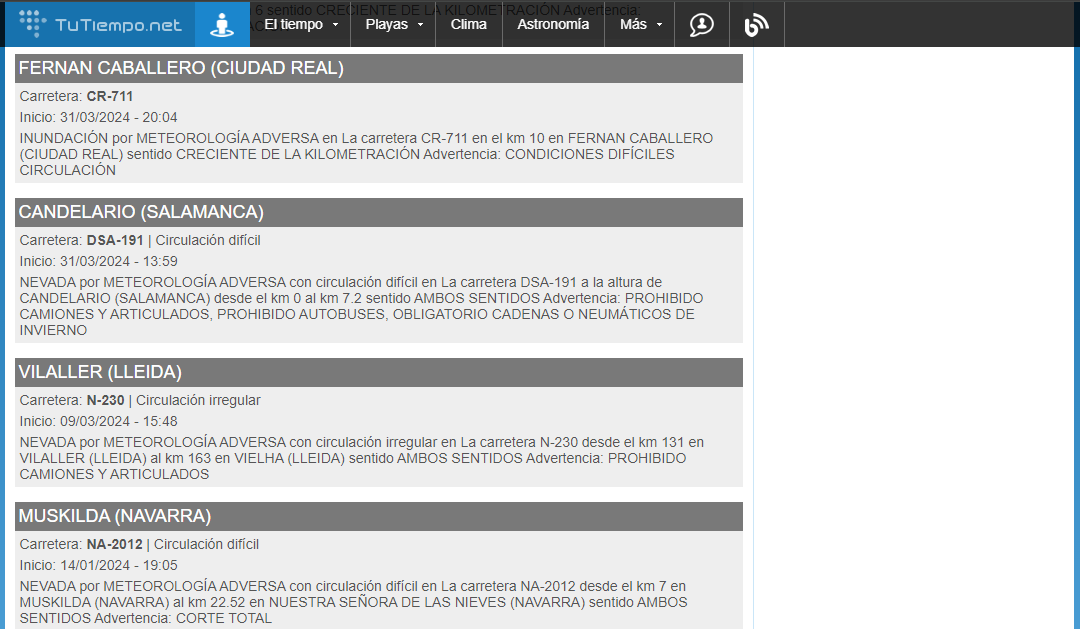
Fuente: Elaboración propia.

### TuTiempo.net

La web de TuTiempo.net (*Información meteorológica tráfico - Incidencias en carretera*, s. f.) es una herramienta que permite conocer tanto la información meteorológica como los incidentes ocurridos en la carretera en tiempo real. Por ello, es considerado una alternativa muy competitiva de no ser porque no dispone de una representación gráfica de dichos datos. El registro de incidencias en las carreteras, por ejemplo, se presenta mediante entradas de texto, lo que dificulta considerablemente el poder detectar la localización de una incidencia en una cierta zona geográfica de un rápido vistazo. Es por ello que la gran diferencia entre ambas herramientas viene en la representación utilizada para mostrar los datos.

Por otro lado, como ha ocurrido con otras herramientas previamente analizadas, mencionar que en “Biker” se espera obtener un análisis sobre el tráfico con mayor detalle, pudiendo ofrecer información clave para el ámbito del ciclismo como es la densidad del tráfico.

**Figura 4.** *Captura de la web TuTiempo.net.*



Fuente: Elaboración propia.

### Flare

Por último, se presenta la aplicación de Flare («Lanzan una app para detectar automáticamente los accident...», 2022), desarrollada por la Federación Internacional del Automóvil (FIA). Esta aplicación notifica en tiempo real los accidentes sufridos por los ciclistas, permitiendo así una asistencia sanitaria en menor tiempo y compartiendo automáticamente la geolocalización del dispositivo.

En comparación con “Biker”, uno de los objetivos finales que ambas soluciones comparten es que pretenden mejorar la seguridad de los ciclistas. No obstante, más allá del ámbito que tratan, no tienen similitud en ninguna de las funcionalidades que ofrecen, ya que una de ellas está destinada al rescate en caso de sufrir un accidente mientras que la otra pretende mejorar la planificación previa a la práctica del deporte.

A continuación, se puede apreciar una tabla comparativa entre las herramientas presentadas a lo largo de la sección.

Tabla 1. *Tabla comparativa de las herramientas analizadas.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Herramienta | MeteoRuta | Comobity | TuTiempo.net | Flare |
| Finalidad principal | Informar | Proteger | Informar | Socorrer |
| Año de lanzamiento | 2014 | 2015 | 2005 | 2022 |
| Alcance | España | España | España | - |
| Información meteorológica | Sí | No | Sí | No |
| Información del tráfico | Sí | Sí | Sí | No |
| Visualización gráfica | Sí | Sí | No | Sí |
| Aplicación móvil | App de AEMET | Sí | Sí | Sí |
| Activa | Sí | Sí | Sí | Sí |

Fuente: Elaboración propia.

## Estado del arte de las tecnologías a utilizar

Una vez realizado el análisis sobre las herramientas existentes, se investiga y justifica el uso de cada tecnología utilizadas en este proyecto. En dicho análisis, se realiza una comparativa entre las principales opciones a utilizar en cada fase del ciclo de vida del dato y se detalla la elección de la herramienta a utilizar.

### Recogida de datos

La recogida de los datos es una fase crítica en el ciclo de vida de los datos. Esto se debe a que, si no existen unos datos adecuados para su análisis, difícilmente se obtendrán unos resultados útiles. Por ello, se presentan diferentes opciones para realizar la recogida de datos, con las ventajas e inconvenientes de cada posibilidad.

#### APIs

Permite recoger los datos más recientes de forma automática sin tener la necesidad de reemplazar el origen de los datos (Castro, 2022). Esto es posible porque se trabaja sobre unos datos dinámicos, permitiendo conformar un flujo de datos actualizado a través de una recogida de datos continua.

Aditivamente, es destacable el hecho de que recogiendo los datos mediante APIs, se obtiene una mayor libertad para elegir los datos con los que se desea trabajar, seleccionando únicamente los campos de interés. De esta forma se puede permutar el orden en la que se lleve a cabo la fase de almacenamiento y la fase de la limpieza y procesamiento, lo que permite conformar un Data Warehouse en lugar de un Data Lake.

Un punto a tener en cuenta es que existen infinidad de APIs, teniendo cada una sus propias parametrizaciones y llegando a requerir claves para su uso en algunos casos.

#### Repositorios

Son conjuntos de datos estáticos sobre los cuales trabajar (Tehreem Naeem, 2020). Generalmente suelen encontrarse en archivos de texto con diferentes formatos, entre los que pueden ser comunes los archivos Comma Separated Value (CSV), Extensible Markup Language (XML) o JavaScript Object Notation (JSON).

#### Web Scraping

Técnica utilizada para obtener datos de las páginas web de forma indirecta (Datademia, 2021). El funcionamiento de esta técnica consiste en recoger todo el contenido de la web en la que se hallan los datos, y posteriormente distinguir la localización de los mismos dentro de toda la estructura HyperText Markup Language (HTML) de dicha web.

#### Formularios y encuestas

Los formularios y las encuestas son técnicas mediante las cuales se obtiene información de una muestra de la población que interesa conocer (Ortega, 2019). Esta herramienta posibilita su uso tanto de forma tradicional como online. Con la digitalización cada vez más presente, la versión online aumenta su índice de uso, ya no solo por la facilidad de recoger los datos de manera digital si no por la mayor publicidad y participación que se consigue obtener.

De cara a analizar los datos que se recojan, destacar la importancia del tipo de pregunta que se realice en la encuesta o formulario en cuestión. Si se presenta una pregunta con respuesta abierta, a la hora de tratar los datos puede existir pérdida de información. Por el contrario, si se proponen preguntas con respuesta cerrada, los datos son fácilmente analizables sin tener ninguna perdida en la información recogida. Por ello, este último es el principal tipo de preguntas que se recomienda realizar en formularios y encuestas.

### Limpieza y procesamiento de datos

Tanto la fase de limpieza y procesamiento de datos, así como la de almacenamiento de datos se encuentran estrechamente ligadas. Esto se debe a que, en el ciclo de vida de un conjunto de datos, en muchas de ocasiones se alterna el orden de estas fases, dependiendo de si se aplica un ETL (Extract Transform Load) o ELT (Extract Load Trasnform). En el primer caso, se conforma un Data Warehouse, almacenando únicamente la información relevante mientras que en la segunda opción se crea un Data Lake.

En el caso de Biker, se plantean las dos opciones para valorar sus puntos positivos y negativos, teniendo siempre en mente el objetivo general de completar el ciclo de vida de los datos previamente recogidos. Seguidamente, se exponen sus diferencias y se valoran ambas alternativas.

#### Data Lake

Consiste en una estructura donde se permiten almacenar grandes cantidades de datos en su formato de origen, sin ningún procesamiento previo y para almacenar durante una gran cantidad de tiempo (*¿Qué son los lagos de datos?*, s. f.). Estos datos pueden ser estructurados, semiestructurados o no estructurados, admitiendo todo tipo de formatos dentro de dicho repositorio (*Soluciones de data lake | IBM*, s. f.).

Esta es justamente una de las principales funcionalidades de los Data Lake, ya que no es fácil encontrar configuraciones que permitan el almacenar distintos formatos de datos en una misma ubicación, y, además, durante un intervalo tan amplio en el tiempo (Navarro, 2022).

#### Data Warehouse

Consiste en una estructura donde se permiten almacenar grandes cantidades de datos tras ser analizados, transformados y modificados (*¿Qué son los almacenes de datos?*, s. f.). Por ello, únicamente se pueden almacenar datos estructurados o semiestructurados.

La finalidad principal de estos repositorios es el almacenar información para su posterior explotación por parte de la entidad propietaria, generalmente empresas. Consecuentemente, tienen una gran velocidad en el acceso a los datos, siendo dicho acceso fácilmente realizable (Daniel, 2022).

### Almacenamiento de datos

Existen diferentes posibilidades para almacenar los datos, siendo las bases de datos la opción más habitual. No obstante, existen varias alternativas dentro de esta solución, las cuales son comentadas a continuación.

#### Base de datos relacionales

Las bases de datos relacionales son un tipo de bases de datos que toma como base las relaciones entre los distintos grupos de datos que se conforman. Estos grupos se conocen como tablas, las cuales se relacionan entre sí y conforman un modelo relacional, el pilar de la base de datos (*¿Qué es una base de datos relacional?*, s. f.).

El principal beneficio de este modelo es que el fundamento sobre el cual se construye es muy básico, pudiendo complicarse a medida que se escale la dimensión de los datos. Asimismo, la alta usabilidad existente con este tipo de bases de datos y sus propiedades ACID la hacen ser una opción perfecta para almacenar conjuntos de datos estructurados (*¿Qué es una base de datos relacional?*, s. f.).

Algunas de las bases de datos relacionales más comunes son:

* MySQL (*MySQL*, s. f.).
* PostgreSQL (PostgreSQL Global Development Group, 2024).
* Microsot SQL Server (*SQL Server 2022 | Microsoft*, s. f.).
* Oracle Database (*Base de datos rentable y de alto rendimiento*, s. f.).

#### Base de datos no relacionales

Las bases de datos no relacionales son un tipo de bases de datos que están diseñadas para ofrecer una alta flexibilidad y escalabilidad en los modelos de datos (*Bases de datos no relacionales | Bases de datos de gráficos | AWS*, s. f.).

A diferencia de las bases de datos relacionales, los datos que se guardan en ellas no tienen una estructura predefinida. Consecuentemente, existen varios tipos de bases de datos no relacionales, teniendo cada una de ellas sus virtudes y limitaciones. Entre las principales, se pueden encontrar bases de datos de documentos, de clave-valor y de grafos (*¿Qué es una base de datos NoSQL?*, 2024).

Las principales bases de datos no relacionales son:

* MongoDB (*MongoDB*, s. f.)
* Cassandra (*Apache Cassandra | Apache Cassandra Documentation*, s. f.)
* Redis (*Redis - La plataforma de datos en tiempo real*, s. f.)
* Neo4j (*Neo4j Graph Database & Analytics – The Leader in Graph Databases*, s. f.).

### Análisis de datos

La fase de análisis de datos se compone por los dos tipos principales de algoritmos. Tanto los algoritmos de aprendizaje supervisado como los algoritmos de aprendizaje no supervisado ofrecen diferentes posibilidades de cara a obtener información de utilidad para el usuario final.

Por ello, con el fin de determinar la utilidad de estos algoritmos en el proyecto de Biker, es necesario valorar en detalle varias de sus propiedades. La finalidad, las características del conjunto de datos que requieren y los parámetros de configuración son algunos de los muchos aspectos a comparar entre ellos. Por ello, se presentan ambos conjuntos.

#### Algoritmos de aprendizaje supervisado

Son algoritmos que tratan con datos etiquetados. Estos algoritmos utilizan las etiquetas de los datos para entrenarse con el fin de realizar una correcta clasificación o predicción (*¿Qué es el aprendizaje supervisado?*, 2024).

Para ello, se requiere dividir el conjunto de datos en dos grupos. El primer grupo es el de los datos de entrenamiento. Este conjunto es aplicado al algoritmo como base, definiendo las entradas y sus correspondientes salidas. De esta forma, el modelo obtiene un pequeño conocimiento de cómo actuar. El segundo grupo lo conforman los datos de prueba, unos datos que son totalmente independientes de los datos de entrenamiento y que permiten comprobar el funcionamiento real del algoritmo una vez entrenado.

Para evaluar los resultados obtenidos, se utilizan diferentes métricas como la precisión o la pérdida. En caso de que la precisión obtenida no sea la esperada, existen metodologías para intentar mejorarla. La validación cruzada (Cross-Validation) precisamente, consiste en realizar varias iteraciones de entrenamiento al algoritmo, cambiando los datos que componen los conjuntos de entrenamiento y prueba. De esta forma se entrena con un mayor número de datos que ayuda a obtener un mejor entrenamiento (Na8, 2023).

Entre los algoritmos de aprendizaje supervisado se encuentran los siguientes:

* Regresión lineal: el objetivo principal es poder realizar la predicción de valores continuos. Se predice el valor de la variable dependiente (output) a través de la relación existente con las variables independientes (inputs), generando una recta entre 2 puntos que interprete de la mejor forma posible los datos de entrenamiento. De esta forma, al evaluar el modelo con los datos de validación, se obtiene un rendimiento similar.
* Regresión logística: el objetivo principal es poder realizar la predicción de valores categóricos, ocurriendo con cierta frecuencia clasificaciones binarias o multiclase. A diferencia de las regresiones lineales, en este tipo se predice la probabilidad de ocurrencia, obteniendo consecuentemente valores entre 0 y 1. Otra diferencia respecto a la regresión lineal es que, en este caso, la recta que se conforma puede tener curvas.
* Máquinas de soporte vectorial (SVM): el objetivo principal es poder realizar clasificaciones y predicciones de los datos, diferenciando la mayor cantidad de clases posibles. Para ello se define un hiperplano, el cual separa las clases independientemente de las dimensiones que lo conformen.
* Árboles de decisión: el objetivo principal es poder solucionar problemas de clasificación y predicción de los datos. No obstante, son herramientas muy versátiles y que permiten una amplia usabilidad. El uso de estos algoritmos se basa en la división continua del conjunto de datos, hasta alcanzar el criterio de parada definido.
* Redes neuronales: herramienta con una amplia versatilidad que simula el funcionamiento de un cerebro humano, permitiendo un amplio abanico de opciones entre las que destacan principalmente la clasificación y la regresión.

La base de esta herramienta se fundamenta en las neuronas artificiales y las capas que la conforman, así como en las conexiones ponderadas y funciones de activación.

* Naive Bayes: algoritmo sencillo, pero realmente eficaz, con el objetivo principal de poder solucionar problemas de clasificación de variables categóricas. Este algoritmo se basa en el teorema de Bayes (López, 2018).

#### Algoritmos de aprendizaje no supervisado

Son algoritmos que buscan agrupar los datos según propiedades o patrones que el propio algoritmo detecte, no teniendo ninguna etiqueta, a diferencia de los algoritmos de aprendizaje supervisado (*¿Qué es el aprendizaje no supervisado?*, 2023).

Dentro de los algoritmos de aprendizaje no supervisado, las agrupaciones de mayor relevancia son:

* Clustering exclusivo: K-Means.

El objetivo principal del algoritmo K-Means es poder dividir el conjunto de datos en K grupos, intentando minimizar la suma de los cuadrados de las distancias dentro de cada clúster. Para ello se definen los centroides y se itera el cálculo de la distancia existente desde cada dato hasta los centroides. Dicho dato es asignado al clúster con el centroide más cercano.

* Clustering jerárquico: agrupamiento jerárquico aglomerativo.

El objetivo principal del agrupamiento jerárquico aglomerativo es agrupar conjuntos de datos de forma jerárquica. El escenario inicial se sitúa con varios clústeres, los cuales se fusionan hasta converger en un único clúster. El proceso de fusionado de los clústeres está ligado a la distancia existente entre los mismos, agrupando de los clústeres existentes los que mayor similitud tengan.

* Clustering probabilista: Expectation-Maximization (EM).

El objetivo principal del EM se basa en estimar los parámetros incompletos de los modelos estadísticos. Dicho algoritmo se compone de dos fases (expectación y maximización) las cuales se llevan a cabo de forma iterativa. La primera busca obtener una primera aproximación de los valores, siendo dichos valores optimizados por la segunda. Este proceso se realiza durante varias iteraciones hasta que el valor deje de ser cambiante y se estabilice.

* Clustering difuso: Fuzzy C-Means.

El objetivo principal del algoritmo Fuzzy C-Means es poder dividir el conjunto de datos en varios grupos, al igual que realiza el algoritmo K-Means. La principal diferencia respecto a este algoritmo se encuentra en que K-Means únicamente permite asignar un dato a un clúster, mientras que Fuzzy C-Means permite que un dato pertenezca a varios clústeres. Este algoritmo sigue un proceso prácticamente idéntico al algoritmo K-means, siendo la diferencia más notoria que la asignación de un dato a un clúster se realiza en base a un grado de pertenencia. De esta forma, un único dato puede pertenecer a varios clústeres distintos.

### Visualización de datos

En la fase de visualización de los datos existen infinidad de opciones sobre como mostrar los datos obtenidos. De igual forma, existen millones de herramientas con las cuales conformar estas visualizaciones. Se valoran seguidamente algunas de ellas.

#### Visualizaciones con Python

Python es un lenguaje de programación ampliamente conocido y utilizado, teniendo un gran crecimiento en los últimos años. Esto se debe a su amplia funcionalidad, así como su facilidad tanto para el aprendizaje como en su uso. A más, existen abundancia de librerías a importar que permiten realizar infinidad de tareas como es el caso de la visualización de datos (Zapata, 2024).

**Gráfico, Gráfico de barras

Descripción generada automáticamenteFigura 5.** *Ejemplo de visualización con librerías de Python.*

Fuente: Elaboración propia.

De las librerías más relevantes dentro de este lenguaje de programación destacan:

* Matplotlib
* Seaborn
* Plotly

#### Visualizaciones con Data-Driven Documents (D3)

D3 es una biblioteca de JavaScript (JS) que permite tratar datos integrándolos en visualizaciones interactivas. Dicha herramienta se compone de infinidad de soluciones pudiendo hasta generar gráficos que deriven de un subconjunto de datos. Además, es una solución que destaca por su flexibilidad, así como por su fácil integración en las páginas web (*¿Qué es D3?*, s. f.).

**Figura 6.** *Ejemplo de visualización con D3.*

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Fuente: Elaboración propia.

#### Visualizaciones con herramientas de BI

Por último, se explora la alternativa de utilizar herramientas de BI, las cuales además de permitir la visualización de los gráficos permite el análisis previo y posterior. Una de las principales ventajas de estas aplicaciones es la alta personalización que ofrece al usuario, difícilmente existiendo dos dashboards iguales (Analytics, 2023). Conociendo su utilidad, se presentan las opciones más relevantes en lo que a herramientas BI se refiere:

* Power BI: herramienta novedosa que ofrece un fácil uso y permite graficar los datos que han podido ser previamente tratados. Ideal para la generación de informes y dashboards de una forma sencilla pero eficaz (*Power BI*, s. f.).
* Tableau: herramienta globalmente conocida y líder absoluta hasta la aparición de Power BI (*Software de análisis e inteligencia de negocios | Tableau*, s. f.). Se requiere de una mayor formación en la herramienta, a pesar de que es intuitiva. En cuanto a funcionalidades, muy similar a Power BI, permitiendo también el tratamiento de datos.
* QlikView: herramienta con cierta similitud a las dos previamente explicadas, pero con un tratamiento de los datos bastante más limitado. Se basa en una exploración más asociativa a pesar de tener una interfaz menos intuitiva que Tableau o Power BI (*QlikView – Analítica y cuadros de mando interactivos y eficaces | Qlik*, s. f.).
* SAP BI: herramienta del Enterprise Resource Planning (ERP) que le añade la funcionalidad especifica de la inteligencia de negocio (*SAP*, s. f.). Las principales diferencias son la integración en el entorno de desarrollo de programas de sistemas de análisis (SAP), el mayor requerimiento de conocimiento para su uso y que en cuanto a coste, ya que requiere una mayor inversión que otras alternativas.

Gráfico, Aplicación

Descripción generada automáticamente**Figura 7.** *Ejemplo de visualización con Power BI.*

Fuente: Elaboración propia.

#### Comparativa de herramientas de visualización

Tabla 2. *Tabla comparativa de herramientas de visualización.*

|  |  |  |
| --- | --- | --- |
| Herramienta | Ventajas | Desventajas |
| Visualizaciones con Python | * Gran personalización a través de diferentes librerías. * Amplia documentación. * Opensource. | * Curva de aprendizaje pronunciada. * Conocimientos de Python. |
| Visualizaciones con D3 | * Altamente personalizable. * Visualizaciones interactivas. * Amplia documentación. * Opensource. | * Curva de aprendizaje muy pronunciada. * Conocimientos de JavaScript. |
| Power BI | * Fácil uso. * Amplia conectividad. * Destinado específicamente para el análisis de datos. * Visualizaciones interactivas. * Amplia documentación. * Versión gratuita existente. | * Menor personalización. * Algunas visualizaciones requieren conexión a internet. |
| Tableau | * Visualizaciones interactivas. * Destinado específicamente para el análisis de datos. * Amplia documentación. * Versión gratuita muy limitada. | * Coste elevado mediante licencias. * Lentitud con grandes conjuntos de datos. * Necesaria cierta adaptación al entorno. |
| QlikView | * Destinado específicamente para el análisis de datos. * Gran personalización. | * Curva de aprendizaje pronunciada. * Coste elevado. |
| SAP BI | * Integración con SAP. * Soporte y actualizaciones. * Destinado específicamente para el análisis de datos. | * Coste elevado. * Requiere gran capacidad de recursos. * Destinado a grandes empresas. |

Fuente: Elaboración propia.

## Conclusiones

Una vez realizado el análisis, se puede afirmar que ninguna de las herramientas presentadas coincide con todos los pilares que conforman dicho proyecto. No obstante, estas opciones pueden llegar a tener un uso similar, a pesar de que su función principal no sea esta.

Los pilares mencionados sobre los que se basa “Biker” se pueden agrupar en un análisis del tráfico y un análisis meteorológico. El primero de estos mide la afluencia de vehículos en las carreteras, incidentes existentes en las mismas y otras propiedades de las vías que puedan afectar al practicante durante el desarrollo de la actividad. En este análisis, se espera abarcar un mayor nivel de detalle que las herramientas competidoras, llegando al punto de facilitar datos muy específicos como la densidad del tráfico esperada.

Por su parte, el análisis meteorológico incluye una serie de mediciones de diferentes índices meteorológicos, como puede ser el viento, la lluvia y la temperatura, permitiendo la planificación y valoración de si es adecuada la práctica de este deporte en ese instante. En este caso, sí que existen diversas herramientas que se pueden aproximar más al nivel de detalle de este ámbito, ya que la fortaleza principal de “Biker” consiste en el análisis conjunto realizado entre ambos conjuntos de datos.

De todas las opciones valoradas en las diversas fases del ciclo de vida del dato, las elecciones tomadas son anunciadas a continuación, con su respectiva justificación.

En la fase inicial del proceso, se ha decidido utilizar la recolección de datos a través de APIs, ya que se cree que puede ser una solución que aporte grandes beneficios en su uso posterior. En este proyecto se han utilizado APIs más sencillas de integrar y otras más complejas, que han requerido de claves públicas y privadas. Para estas últimas, ha sido necesario realizar una investigación más exhaustiva, dado la pobre documentación que existía al respecto.

Como se puede intuir al ser la segunda etapa del proceso, en la fase de limpieza y procesamiento de los datos, se utiliza un Data Warehouse. Mediante esta estructura, se entiende que existe un procesando y una recogida selectiva de los datos obtenidos desde las APIs, siendo únicamente aquellos datos considerados de utilidad los que se almacenan en la siguiente fase. Dicho procesamiento y selección de los datos es realizado con Python.

Esta elección se toma mayoritariamente por la limitación tecnológica existente en los equipos con los que se desarrolla este proyecto. Al realizar primero una limpieza, se consiguen mover menos datos a la base de datos, siendo aquellos almacenados indispensables para el análisis posterior. Consecuentemente existe un considerable ahorro de tiempo y de capacidad, permitiendo obtener un resultado más eficiente con los medios de los que se disponen.

Para la fase de almacenamiento, se ha seleccionado una base de datos no relacional como es MongoDB dado que en ciertas ocasiones los datos recogidos dentro de una misma colección pueden diferir en formato. Aun así, se podría haber llegado a utilizar una base de datos relacional, ya que la mayoría de los datos recogidos son procedentes de sensores y se consideran datos estructurados o semiestructurados. No obstante, dentro de la amplia variedad de bases de datos existentes, se ha decidido que MongoDB es la mejor alternativa, tanto por la mayor velocidad en el acceso a los registros, así como por la experiencia previa existente con la herramienta.

Con el análisis realizado, se ha podido conocer que MongoDB dispone de un alto rendimiento gracias a los índices que utiliza y el almacenamiento en memoria del que dispone. Sumado a su flexibilidad, escalabilidad y soporte, lo convierten en una herramienta con un alto potencial, superando en varios puntos a otras herramientas similares, lo que la convierte en la elección para este proyecto.

En la fase de análisis de los datos, se ha decidido utilizar mayoritariamente algoritmos de aprendizaje supervisado. El uso de estos algoritmos ha sido argumentado por la finalidad principal de predecir valores, siendo en la mayoría de los casos índices relacionados tanto con la meteorología como con el tráfico. No obstante, también se ha utilizado el clustering para ciertos conjuntos de datos.

Por último, en la etapa de visualización de datos se ha decidido utilizar Power BI ya que es una herramienta que está en auge y sobre la que cada vez existe una mayor demanda. Al ser una herramienta que dispone de una amplia capa de conectividad con diversas herramientas y al disponer de una versión gratuita, ha sido fácil encontrar soporte sobre la misma. Adicionalmente, era una herramienta que se quería aprender a nivel personal por lo que se ha encontrado el momento idóneo a través de este proyecto.

# Objetivos concretos y alcance

El siguiente apartado presenta los objetivos que se pretenden logran en este TFM, siendo diferenciado el objetivo principal de los objetivos secundarios.

## Objetivo general

El objetivo principal de este proyecto es diseñar, desarrollar y evaluar una aplicación para ciclistas en el País Vasco que gestione el ciclo de vida completo de datos meteorológicos y referentes al tráfico. Se busca mediante la información facilitada, que los ciclistas sean capaces de tomar decisiones fundamentadas, pudiendo conocer aspectos tan desconocidos como el clima o estado de las carreteras, de forma previa a iniciar la ruta.

Para ello se requiere del ciclo de vida al completo, el cual debe incluir desde la recogida de los datos, hasta su visualización, pasando por las fases intermedias de limpieza y procesamiento, almacenamiento y análisis.

## Objetivos específicos

Además de cumplir el objetivo principal, el proyecto pretende completar los denominados como objetivos secundarios, los cuales se mencionan a continuación:

* Obtener datos de 2 orígenes de datos diferentes.
* Filtrar los datos útiles antes de ser almacenados, conformando un ETL.
* Almacenar los datos de una forma semiestructurada en la base de datos.
* Obtener un conjunto de datos mayor que 80.000 instancias.
* Comparar 2 algoritmos de aprendizaje supervisado y justificar la elección tomada.
* Utilizar 1 algoritmo de aprendizaje no supervisado.
* Utilizar técnicas de validación cruzada de modelos en los algoritmos utilizados.
* Definir los hiperparámetros de forma óptima para evitar el sobreajuste (overfitting).
* Desarrollar y evaluar la predicción de los principales índices meteorológicos a través de algoritmos de aprendizaje supervisado con un R2-Score superior a 0.8.
* Desarrollar y evaluar la predicción de los principales índices de la calidad del aire a través de algoritmos de aprendizaje supervisado con un R2-Score superior a 0.5.
* Obtener visualizaciones interactivas a través de una herramienta de BI.
* Utilizar 5 tipos de visualizaciones en los informes.

La visión con la que se establecen estos objetivos es la de obtener un proceso completo y detallado sobre la gestión del ciclo de vida de un conjunto de datos. En este caso, el ámbito a tratar es el ciclismo, pudiendo en un futuro transversalizarse a cualquier otro campo del que se dispongan datos.

Asimismo, se pretende corroborar el cumplimiento general de este proyecto a través del cumplimiento individual de cada uno de los objetivos secundarios, siendo pequeñas muestras del cómputo global del proyecto.

## Alcance

Este proyecto tiene como alcance el desarrollo de una aplicación para ciclistas que integre el ciclo de vida completo de un conjunto de datos, desde la recogida de datos hasta su visualización. Sobre este conjunto de datos se pretende llevar a cabo tanto un análisis meteorológico, así como un análisis sobre las carreteras del País Vasco.

En este proyecto se incluye:

1. El diseño del proceso que engloba el ciclo de vida completo del conjunto de datos utilizado.
2. La recogida de los datos desde las APIs de Open Data Euskadi y Euskalmet. Adicionalmente se llevan a cabo una serie de pruebas a través de herramientas externas como Postman para asegurar el correcto funcionamiento de las APIs, al igual que se integra el Bearer Token en los casos que se requiere.
3. El tratamiento y limpieza de datos previo a la inserción en la base de datos. Se manipulan los datos con el fin de prepararlos y convertirlos en datos semiestructurados.
4. El almacenamiento de los datos en una base de datos. Únicamente serán almacenados aquellos datos que se consideren relevantes, los cuales han sido tratados y estructurados en la fase anterior.
5. Un análisis de datos sobre las técnicas de IA a utilizar, valorando las diferentes propiedades de cada algoritmo y realizando una comparativa entre ellos. El desarrollo de todos los modelos es realizado para su evaluación, tomando posteriormente la decisión sobre que algoritmo se utiliza.
6. La visualización de los datos a través de diferentes informes que permiten al usuario conocer las circunstancias meteorológicas, así como las referentes al tráfico.
7. La documentación del proyecto, incluyendo los resultados de este y la aportación realizada.

En contraposición a lo previamente citado, este proyecto no incluye los siguientes puntos:

1. A pesar de la existir la posibilidad de recoger los datos en tiempo real con un desarrollo más profundo, se ha decidido acotar este aspecto debido al límite temporal existente. Ante esta circunstancia, para la fase práctica se toma un conjunto de datos estático, siendo dichos datos recogidos en fechas recientes a la publicación de este proyecto.
2. De forma similar al punto previamente comentado, tampoco se realiza el almacenamiento de los datos recogidos a través de un clúster. La principal limitación se presenta por el tiempo disponible para la investigación e implementación de esta funcionalidad en el proyecto.

# Metodología de trabajo y arquitectura

## Metodología de trabajo

La metodología del trabajo principal en este proyecto sigue los principios de la metodología ágil Scrum, siendo adaptada para ser utilizada con un único desarrollador. Asimismo, cada uno de los sprints que conforma el proyecto son organizados inicialmente en base al Big Data Lifecyle Managment (Esmerado, 2024, p. 5). Este modelo resume el uso completo de los datos en proyectos Big Data en 5 fases principales, a pesar de que en ciertas ocasiones aparezcan fases adicionales. No obstante, al ser una metodología ágil pueden existir modificaciones en la organización de los sprints durante el transcurso del proyecto, requiriendo dividir alguna fase en varios sprints.

Para gestionar esta metodología se utilizan varias herramientas entre las cuales destaca Github para la gestión de versiones de código. Github es un portal web que funciona como repositorio distribuido utilizando la tecnología Git (Fernández, 2019). Actualmente, esta plataforma es mundialmente conocida y existen millones de usuarios en ella. El uso que se le atribuye a esta herramienta básicamente tiene como objetivo el almacenamiento de la última versión del software, existiendo un registro de las modificaciones realizadas y un control de sus versiones (Casero, 2023). Al ser un proyecto personal, no se utilizan todas las funcionalidades que esta tecnología puede aportar, ya que no es necesario. No obstante, en otras situaciones en la que se involucren más participantes, podría ser interesante utilizar funcionalidades adicionales de esta tecnología como la creación de ramas, destinada principalmente para el desarrollo colaborativo.

Otra herramienta que se utiliza, principalmente para tener una visión más general del proyecto, es GanttProject. Mediante la versión gratuita, se puede realizar la planificación general del proyecto, definiendo los sprints mediante un cronograma con sus respectivos plazos y recursos. GanttProject permite tener un enfoque más genérico del proyecto, permitiendo su gestión de una forma más organizada y consecuentemente eficiente (*GanttProject*, 2024).

## Arquitectura

En el esquema siguiente se puede apreciar la arquitectura que define las fases de la solución presentada. Estas fases engloban el ciclo de vida completo de un conjunto de datos, pudiendo apreciarse adicionalmente la tecnología utilizada para llevar a cabo dicha fase. En siguientes secciones se desglosa en mayor detalle.

**Figura 8.** *Arquitectura de las fases del ciclo de vida del dato.*

Diagrama

Descripción generada automáticamente con confianza media

Fuente: Elaboración propia.

La primera fase comienza con la obtención de los datos desde APIs públicas, tanto de Open Data Euskadi como de Euskalmet. Esta recogida de datos se ha realizado mediante librerías de Python y los datos devueltos desde las APIs han sido almacenados en variables de dicho lenguaje de programación. Los datos recogidos están relacionados con la meteorología y el tráfico del País Vasco.

La segunda fase consiste en recoger esas variables que contienen la totalidad de los datos y aplicarles una limpieza y tratamiento. Posteriormente, se realiza la selección únicamente de los datos que se crean útiles para el proyecto. Esta fase también es realizada con el lenguaje de programación Python.

La fase de almacenamiento consiste únicamente en almacenar los datos obtenidos tras la limpieza y el procesamiento realizado. Mediante librerías de Python se crea el enlace con MongoDB y posteriormente se añaden los datos tratados a la base de datos.

En la cuarta fase se comienza realizando un EDA a los datos hábiles con Python. Seguidamente se aplican técnicas de IA gracias a librerías más específicas de este mismo lenguaje. Estas técnicas son utilizadas para obtener una información que aporte un mayor valor añadido al análisis realizado, permitiendo obtener información valiosa para la posterior visualización.

En la última fase, la fase de visualización, se recogen los datos tras su análisis y se plasman en un dashboard mediante la herramienta de Power BI. Gracias a los distintos tipos de visualizaciones dentro de este dashboard, se aporta información variada y relevante al usuario final, la cual es de gran utilidad para llevar a cabo la planificación del deporte.

# Marco normativo

Como se ha mencionado a lo largo del escrito, la mayoría de los datos recogidos desde las diferentes APIs son de un carácter público, lo que no exime su uso en este proyecto. No obstante, el resto de datos se ha corroborado que se rigen bajo la licencia Creative Commons, lo que posibilita su uso con fines académicos (*Información legal*, 2010). Es necesario aclarar que este conjunto de datos no es utilizado para ninguna otra finalidad más allá de este proyecto educativo, lo que se posibilita tanto por las licencias que poseen como por las normativas que se describen a continuación.

Otro punto de alta relevancia en el uso de este conjunto de datos es que en la totalidad del proyecto se utilizan datos anónimos por lo que no se requiere de la aplicación de la normativa de datos. No obstante, se cumplen las dos principales normas en lo que a protección de datos se refiere. La primera de estas es la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD) (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, 2018).

El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos es considerada como la segunda norma a respetar para cumplir con la legislatura en torno a la protección de datos (RGPD) (BOE.es - DOUE-L-2016-80807 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)., 2018).

Por ello, como se puede conocer en la página oficial de Open Data Euskadi, los datos reutilizados de la propia web son aptos en su totalidad para fines como el empleado en este proyecto, permitiendo su recogida, tratamiento, almacenamiento, análisis y visualización (*Normativa*, 2017).

# Desarrollo específico de la contribución

El ciclo de vida de un conjunto de datos hace referencia a una secuencia de etapas por las que pasan los datos a lo largo de su vida útil. Este proceso engloba desde la generación o recogida de los datos hasta la visualización y publicación de estos. Además, son varias las razones que hacen de este proceso una solución altamente relevante, destacando sobre todo en el ámbito empresarial.

La primera razón, que a su vez es la más destacable, se resume en que conforma un procedimiento estandarizado para la gestión eficaz de la información. De esta forma, se obtiene el máximo rendimiento a la explotación de los datos, pudiendo transversalizar el proceso a cualquiera de las situaciones posibles, independientemente del ámbito al que pertenezcan los datos. Además, se garantiza la regulación, así como la seguridad de los datos con los que se trata.

Otra ventaja ciertamente relevante es la amplia disponibilidad de los datos. Mediante este método, se posibilita la obtención de los datos de una forma mucho más veloz, conociendo la distribución en la que se almacenan los datos, así como la información que reportan. Haciendo mención al punto anterior, se enfatiza la integridad estructural conformada a través de este proceso. Con una estructura claramente definida y las medidas de seguridad pertinentes, se brinda la posibilidad de tomar decisiones de una forma mucho más eficiente.

Por todas las ventajas mencionadas, así como las facilidades a más que puede aportar este proceso, el ciclo de vida de un conjunto de datos es una herramienta con una utilidad inmensa, siendo mayor aun si cabe el potencial del que dispone.

En esta sección se explica el desarrollo realizado en este proyecto, analizando cada fase en detalle y mostrando el valor aportado.

## Recolección de los datos

Tras identificar un boceto con la idea del proyecto, el primer aspecto a tratar ha sido el encontrar un origen del cual obtener los datos para llevar a cabo dicha propuesta. Gracias a otros proyectos previos ya realizados, se conocía el repositorio de datos público de Open Data Euskadi. Además, como consecuencia de la investigación realizada para el estado del arte, se han podido valorar distintas alternativas tanto en el formato de los datos, como en el origen de los mismos.

Finalmente, se ha decidido recoger los datos mediante APIs no solo de Open Data Euskadi, sino que también de Euskalmet, la agencia de meteorología más conocida en el País Vasco. Ambas organizaciones poseen una alta confiabilidad en los datos que facilitan, ya que son entidades oficiales descendentes del Gobierno Vasco. Consecuentemente, están recogidas en la página oficial del Gobierno de España, bajo la supervisión del Ministerio para la transformación digital y de la función pública (*Open Data Euskadi | datos.gob.es*, s. f.).

Mediante estas APIs se han recogido datos referentes al tráfico y a la meteorología del País Vasco, permitiendo conformar la base del proyecto Biker. No obstante, la obtención de estos datos no ha sido únicamente mediante la ejecución de un código Python, sino que ha seguido una metodología con sus respectivas fases.

La fase inicial de este procedimiento se ha fundamentado en investigar los diferentes parámetros y usos que permite cada API. Tras conocer estas funcionalidades en cada una de las circunstancias, se ha realizado una serie de pruebas con la aplicación de Postman. Adicionalmente, se ha añadido el Bearer Token para las APIs que lo requerían, que concretamente han sido aquellas facilitadas por Euskalmet. Una vez asegurado el correcto funcionamiento de las APIs, se han incluido dichas consultas en el código Python con el fin de recoger los datos y continuar con el ciclo de vida del conjunto de datos.

Es destacable el hecho de que, durante el proceso de recogida de los datos desde las APIs, se han incluido mecanismos de tolerancia a fallos, permitiendo reintentar la solicitud en varias ocasiones más como medida de prevención a los posibles inconvenientes que pudiesen llegar a ocurrir en el trascurso del proceso.

Para tener una visión más detallada, a continuación, se presentan las APIs que han recogido la totalidad de los datos utilizados en este proyecto.

* Número de páginas en los que se agrupan los puntos de medición:

<https://api.euskadi.eus/traffic/v1.0/meters>

Se utiliza esta API de Open Data Euskadi para conocer el número de puntos de medición existentes en todo el País Vasco, y consecuentemente en cuantas páginas se agrupan. El valor obtenido es utilizado en posteriores APIs.

* Listado de los puntos de medición:

[https://api.euskadi.eus/traffic/v1.0/meters?\_page={config.contador\_pags}](https://api.euskadi.eus/traffic/v1.0/meters?_page=%7bconfig.contador_pags%7d)

Se utiliza esta API de Open Data Euskadi para recoger un listado con los identificadores únicos de todos los puntos de medición existentes en el País Vasco. Estos valores retroalimentan otras APIs que aportan los valores que realmente son utilizados. La variable “*config.contador\_pags*” recorre todas las páginas en las que existen datos para recoger sus respectivos identificadores únicos.

* Número de páginas en los que se agrupan los datos del tráfico por cada punto de medición:

[https://api.euskadi.eus/traffic/v1.0/flows/byYear/{year\_month[0]}/byMeter/{meter}](https://api.euskadi.eus/traffic/v1.0/flows/byYear/%7byear_month%5b0%5d%7d/byMeter/%7bmeter%7d)

Se utiliza esta API de Open Data Euskadi para conocer el número de páginas que agrupan los datos de un punto de medición concreto, referente a los flujos de tráfico. Se obtiene el recuento de los datos recogidos en un año. El valor obtenido es utilizado en la API posterior. Mediante la primera posición del parámetro *“year\_month”* se devuelve el año que interesa conocer y mediante *“meter”*, el punto de medición. Este último itera a través del array que recoge la totalidad de estos.

* Valores de los flujos de circulación:

[https://api.euskadi.eus/traffic/v1.0/flows/byYear/{year\_month[0]}/byMeter/{meter}1?\_page={config.contador\_pags}](https://api.euskadi.eus/traffic/v1.0/flows/byYear/%7byear_month%5b0%5d%7d/byMeter/%7bmeter%7d1?_page=%7bconfig.contador_pags%7d)

Esta API de Open Data Euskadi obtiene la totalidad de los datos de cada punto de medición en un año concreto. En la variable “*year\_month*” se almacenan parejas de valores que hacen referencia a la fecha, siendo el primer valor el año y el segundo el mes. Únicamente se utiliza el primero de estos valores.

Para el otro parámetro, “*meter*”, se utilizan todos los valores recogidos en la API que previamente ha obtenido el listado de los puntos de medición, aplicando cada uno de forma iterativa. De esta forma, se recorren todas las fechas establecidas entre la fecha inicial y fecha final para cada punto de medición. Y de igual forma, este proceso se lleva a cabo para todos los identificadores únicos de los puntos de medición, lo que conforma un amplio catálogo de datos sobre los que trabajar. Hay que tener en cuenta que también se itera el parámetro “*config.contador\_pags*”.

* Valores de la calidad del aire:

[https://api.euskadi.eus/air-quality/measurements/daily/counties/{prov}/from/{inicio}/to/{fin}](https://api.euskadi.eus/air-quality/measurements/daily/counties/%7bprov%7d/from/%7binicio%7d/to/%7bfin%7d)

Mediante esta API de Open Data Euskadi se obtienen los valores referentes a la calidad del aire en diferentes estaciones meteorológicas ubicadas por el País Vasco. Esta API la componen 3 diferentes parámetros. El primero de ellos hace referencia a la provincia en la que se quieren recoger datos. Como tal, cada una de las 3 provincias que componen el País Vasco tiene asociado un código numérico que referencia a dicha zona geográfica y es utilizado en esta ocasión. El segundo y tercer parámetro limitan el uso de la API temporalmente, ya que se define la fecha de inicio y fin de la recogida de dichos datos.

* Listado de las estaciones de Euskalmet:

<https://api.euskadi.eus/euskalmet/stations>

Se utiliza esta API de Euskalmet para recoger un listado con datos de las estaciones existentes en el País Vasco. Los principales datos que se recogen son tanto los identificadores únicos de las estaciones, así como la fecha de su snapshot, lo que puede asemejarse a una copia de seguridad.

* Datos de las incidencias en las carreteras del País Vasco:

[https://api.euskadi.eus/traffic//v1.0/incidences/byDate/{year}/{month}/{day}](https://api.euskadi.eus/traffic//v1.0/incidences/byDate/%7byear%7d/%7bmonth%7d/%7bday%7d)

Se utiliza esta API de Open Data Euskadi para recoger los datos relacionados con las incidencias en las carreteras del País Vasco. Entre estos datos, se puede encontrar la carretera en la que ha ocurrido el incidente, el tipo de incidente y el intervalo de su duración, a través de la fecha de inicio y de fin. Los parámetros que se utilizan únicamente hacen referencia a la fecha concreta en la que interesa conocer las incidencias ocurridas.

* Valores de las estaciones de Euskalmet:

[https://api.euskadi.eus/euskalmet/stations/{est}/{snap}](https://api.euskadi.eus/euskalmet/stations/%7best%7d/%7bsnap%7d)

Mediante esta API de Euskalmet se obtienen los datos de las estaciones en la fecha de una de sus copias de seguridad realizadas, permitiendo conocer los valores que recogen los diferentes sensores que componen la estación.

El procedimiento que se sigue consiste en alimentar la propia API desde un array que se compone a su vez por varios objetos. Estos objetos más concretamente son parejas de datos, donde se almacena el identificador único de la estación como primer dato y la fecha del snapshot más reciente de esa estación en segunda posición. Estos valores se facilitan a la API mediante los parámetros *“est”* y *“snap”* respectivamente.

* Valores de los sensores ubicados en las estaciones de Euskalmet:

[https://api.euskadi.eus/euskalmet/readings/summarized/byDay/forStation/{estacion}/{sensor}/measures/{tipo\_medida}/{medida}/at/{year}/{month}/{day}](https://api.euskadi.eus/euskalmet/readings/summarized/byDay/forStation/%7bestacion%7d/%7bsensor%7d/measures/%7btipo_medida%7d/%7bmedida%7d/at/%7byear%7d/%7bmonth%7d/%7bday%7d)

Mediante esta última API se obtienen los datos que facilitan los sensores ubicados en las diferentes estaciones de Euskalmet. De todas estas mediciones, se han definido como relevantes algunas métricas en concreto que son las recogidas causa de la iteración de los identificadores únicos de las estaciones, que son parametrizados en la variable *“estación”.* Además, se facilita el sensor del cual recoger los datos mediante la variable *“sensor”.* Sobre la medición que se busca obtener, se conoce tanto el tipo como la medición exacta, gracias a los parámetros *“tipo\_medida”* y *“medida”.* Por último, hacer mención de que cada consulta se realiza en una fecha específica, la cual se conoce a través de las variables *“year”*, *“month”* y *“day”.*

## Limpieza y procesamiento de datos

Una vez recogida la totalidad de los datos desde las diferentes APIs, se les ha aplicado la limpieza y el procesamiento correspondiente, mediante el lenguaje de programación de Python. Para comprender este proceso con una mayor claridad, esta sección se divide en base a los diferentes datos recogidos, explicando los procesos que involucren a cada uno de ellos respectivamente.

Sin embargo, existe un factor altamente relevante que explicar antes de comenzar a describir los procesos a los que se han expuesto los datos. Este punto hace referencia a la definición de las fechas que conforman el periodo en el que se recogen datos, estableciendo una fecha inicial y una fecha final. Este periodo tiene varias utilidades dentro de la aplicación, entre las que destacan la alimentación de diferentes APIs y el procesado de los datos previamente obtenidos.

Dada las limitaciones tanto capacitivas como temporales, el conjunto de datos que se utiliza en Biker es del primer semestre de 2024, recogiendo desde el mes de enero hasta el mes de junio, ambos incluidos. Esta decisión viene tomada en primera instancia por el límite de memoria existente en los dispositivos hardware utilizados, lo que al ser contrastado con el número tan elevado de carreteras existentes y su recogida de al menos un registro diario, hace inviable el aumento temporal.

Si a esto, se le suma la inmensa cantidad de indicadores meteorológicos recogidos, se obtiene un conjunto de datos considerablemente amplio, y por consiguiente, pesado. Esto hace inviable la recogida de datos en un periodo de tiempo mayor.

Ligado a este punto, aparece la limitación temporal, ya que, con el largo tiempo de ejecución destinado a la recogida y tratamiento de datos, no es posible realizar una recogida a mayor escala. No obstante, como línea futura puede valorarse el aumento de los datos históricos, permitiendo así a posteriores técnicas de IA un aprendizaje basado en ciclos temporales como las estaciones. Para ello, se precisarían mayores recursos de computación.

Tras este pequeño inciso, se procede a presentar la limpieza y procesamiento de los diferentes subconjuntos de datos.

### Datos ligados al tráfico

Este subapartado recoge todos los datos que tienen relación directa con el tráfico. Principalmente, el conjunto de datos se puede dividir en 2 agrupaciones como son los datos de las incidencias y los datos de los flujos de las carreteras.

El primero de estos subgrupos recoge ciertas propiedades de los datos devueltos por la API consultada, realizando un filtrado que distingue entre los datos interesantes o de utilidad y los datos que no aportan valor al proyecto. A continuación, se pueden conocer las propiedades de los datos seleccionadas:

* *“\_id”*: identificador numérico único de la incidencia registrada.
* *“sourceId”*: identificador numérico único del origen desde el que se ha detectado la incidencia.
* *“incidenceType”*: tipo de la incidencia detectada.
* *“province”*: provincia en la que se ha detectado la incidencia.
* *“cause”*: causa de la incidencia detectada.
* *“citytown”*: ciudad en la que se ha detectado la incidencia.
* *“startDate”*: fecha de comienzo de la incidencia.
* *“road”*: carretera que ha sufrido la incidencia.

Seguidamente, se formatean los datos seleccionados y se conforman diccionarios de datos, los cuales siguen una estructura definida. De esta forma se facilita el análisis posterior, almacenando estas estructuras de datos en un array.

En lo que a los datos sobre el tráfico se refiere, existe un tratamiento de estos mucho más exhaustivo. La principal diferencia viene dada en la existencia de un mayor volumen de datos en comparación con los registros de las incidencias, conformando un proceso más laborioso.

El primer paso consiste en obtener el dato del número de páginas en las que se recogen los diferentes identificadores únicos de los puntos de medición. Este dato se obtiene mediante una API y su única utilidad es alimentar otra API posterior, para conocer la totalidad de estos identificadores únicos.

Seguidamente, se realizan solicitudes a la misma API de forma iterativa, compartiendo como parámetro el número de la página a la que se accedía, incrementado este valor hasta alcanzar el número de páginas máximo en el que se recogen datos. De esta forma se consiguen almacenar todos los identificadores únicos de los puntos de medición.

Con los identificadores previamente recogidos, es posible realizar las solicitudes pertinentes a la última API que se involucra en este conjunto de datos. Asimismo, las parejas de datos de año y mes confeccionadas previamente también son requeridas para realizar solicitudes a esta API. A continuación, se presentan los datos devueltos por la API tras ser formateados:

* *“\_id”*: identificador único que recoge el identificador único del punto de medición, la fecha de la medición y la hora de esta.
* *“meterId”*: identificador numérico único del punto de medición.
* *“source”*: identificador numérico único del origen desde el que se ha realizado la medición.
* *“fecha”:* fecha de la medición.
* *“año”*: primero de los valores de la variable “*year\_month*”, que almacena el año y es utilizada en la API.
* *“mes”*: el segundo valor de la pareja de datos de la variable “*year\_month*”, que almacena el mes.
* *“vel\_media”*: velocidad media de los vehículos que han recorrido la carretera en esa medición. En caso de no tener dato, se asigna un 0 automáticamente.
* *“vehiculos”*: número de vehículos que recorren la carretera durante la medición.

Tras obtener todos los datos de los diversos flujos de tráfico, se detecta que existen varios registros en un mismo día y que se sitúan en una misma carretera. Para simplificar el análisis a realizar posteriormente, se decide agrupar todas las casuísticas que cumplan las condiciones previamente mencionadas. Para ello, se realiza el sumatorio de todos los vehículos recogidos durante cada día en esa vía, así como el cálculo de la media ponderada de la velocidad media alcanzada por los dichos vehículos.

De esta forma se obtienen unos datos más organizados, obteniendo un único diccionario por carretera y día. Estos diccionarios se almacenan en un array.

### Datos ligados a la meteorología

Por otra parte, esta sección involucra todos aquellos datos del ámbito meteorológico, que completan el conjunto de datos utilizados en el proyecto Biker. En rasgos generales, los datos recogidos en este grupo se dividen en datos sobre la calidad del aire, las estaciones meteorológicas y todos los datos recogidos de los diferentes sensores que componen las estaciones.

Respecto a los datos que describen la calidad del aire, son los que menor procesamiento y limpieza necesitan de todo el subconjunto definido. Esto ocurre porque únicamente se accede a una API para la obtención de los mismos, sin requerir ningún otro dato o parámetro adicional. Sin embargo, es necesario realizar una solicitud a la API por cada una de las provincias sobre las que se quieren obtener datos, repitiendo el proceso en 3 ocasiones. Seguidamente, se muestran los datos de interés recogidos mediante la API:

* *“\_id”*: identificador numérico único de la medición de la calidad del aire en una estación.
* *“fecha”*: fecha de la medición de la calidad del aire.
* *“id\_estacion”:* identificador numérico único de la estación de la que se obtiene la medición de la calidad del aire.
* *“nombre”*: nombre de la estación donde se obtiene la medición de la calidad del aire.
* *“provincia”*: identificador numérico que hace referencia a la provincia. Se actualiza según el bucle que se esté recorriendo, haciendo referencia a la provincia pertinente en cada caso.
* *“measure[‘name’]”*: nombre de la métrica que se recoge de la API en la medición.
* *“unidad\_medicion”*: unidad en la que se mide la métrica recogida.

Todos los datos recogidos, independiente de la provincia de origen, se almacenan en un mismo array. Esto permite agrupar todos los datos organizados en diccionarios estructurados, facilitando el análisis que se realizará en las siguientes fases.

En el subconjunto de datos que se engloban las estaciones meteorológicas, se requiere de un tratamiento ligeramente más exhaustivo. En gran medida, esto ocurre por la gestión de los snapshots.

Este proceso comienza accediendo a la API que devuelve los identificadores únicos de las estaciones y sus snapshots. Cada pareja de estación-snapshot se almacena en un array, que a su vez, se almacena en un segundo array, donde se almacenarán la totalidad de los datos. Como cada estación tiene varios snapshots, existen parejas de datos en los que la estación coincide, diferenciándose únicamente en el snapshot. Estas casuísticas no aportan información, por lo que se realiza una comprobación posterior con la finalidad de almacenar únicamente la pareja de datos que recoja el snapshot más reciente de cada estación.

Para esta comprobación, se conforma un nuevo array con los valores de las estaciones y su último snapshot. El procedimiento a seguir consiste en recorrer el array conformado por todos los datos y verificar la estación de cada pareja. Si se detecta que la estación ha sido previamente revisada, se contrastan los snapshots, almacenando únicamente la pareja que tenga una fecha más reciente. Si por el contrario dicho identificador de la estación no ha sido detectado hasta el momento, es añadido directamente al array.

Con el array ya conformado, se procede a iterar cada uno de esos valores en la última API de las estaciones meteorológicas. Se realiza una solicitud a esta API para cada pareja de datos que consecuentemente recoge una estación diferente. Mediante ella, se obtiene un listado de los nombres de los sensores que se hallan en la propia estación meteorológica. Este listado es añadido al diccionario que se devuelve por cada petición a la API, siendo posteriormente almacenado. Para conocer en mayor detalle la estructura del diccionario tipo, se presentan los parámetros a continuación:

* *“\_id”*: identificador único de la estación, compuesto por valores alfanuméricos.
* *“tipo\_estacion”*: tipo de la estación meteorológica.
* *“nombre”:* nombre de la estación en español, que generalmente hace referencia a la localización de esta.
* *“municipio”*: nombre del municipio en la que se sitúa la estación, en español.
* *“provincia”*: nombre de la provincia en la que se sitúa la estación, en español.
* *“snapshot”*: identificador numérico del snapshot, que recoge la fecha del mismo.
* *“sensores”*: listado anidado de los sensores que conforman la estación.

Al igual que con las casuísticas anteriores, estos diccionarios son almacenados en un array que se utilizará en posteriores etapas del ciclo de vida del conjunto de datos.

Tratando ya el conjunto de datos más complejo de esta sección, que engloba todos esos datos recogidos desde los múltiples sensores que contienen las estaciones meteorológicas, se procede a explicar su limpieza y tratamiento.

El procedimiento que se sigue consiste en iterar cada diccionario existente en el array que recoge los datos sobre las estaciones. De cada diccionario, se obtiene el listado de los sensores que componen cada estación, y consecuentemente, de cada sensor se obtienen a su vez todas las mediciones posibles, dentro de las métricas meteorológicas que sean interesantes de analizar. A continuación, se definen dichas métricas:

* Humedad: humedad relativa del aire. La unidad de medición es el porcentaje de agua existente en el aire.
* Temperatura: temperatura del aire. La unidad de medición es el grado Celsius.
* Precipitaciones: cuantía de las precipitaciones. La unidad de medición es el milímetro por metro cuadrado.
* Velocidad media del viento: velocidad promedio del viento. La unidad de medición es metro por segundos.
* Dirección del viento: dirección que toma el viento. La unidad de medición son los grados de ángulo.

Para ello, se realiza una solicitud a la API por cada sensor de cada estación, adjuntando a su vez el tipo de la métrica a obtener, así como la métrica en cuestión. Además, se añade la fecha como parámetro, con un desglose de día, mes y año.

Mediante la respuesta de la API, se recogen los valores considerados como más relevantes de cara al análisis posterior, conformando un diccionario como en los casos anteriores. Posteriormente, se muestran los datos recogidos en lo que sería un diccionario tipo:

* *“\_id”*: identificador único que recoge el año, mes, día junto al nombre de la medida, nombre de la estación y nombre del sensor. Todos estos valores son concatenados en un único nombre, permitiendo diferenciarse con cualquier otro registro.
* *“año”*: año en la que se realizó la medición.
* *“mes”:* mes en la que se realizó la medición.
* *“dia”:* día en la que se realizó la medición.
* *“estacion”:* estación que realizó la medición.
* *“sensor”:* sensor de la estación mencionada que realizó la medición.
* *“tipo\_medida”:* tipo de la métrica que se mide.
* *“medida”:* métrica concreta que se mide.
* *“max”:* de todos los valores obtenidos durante la medición, el valor más elevado.
* *“max\_acumulado”*: el valor máximo acumulado se obtiene en determinadas métricas como el valor máximo registrado durante la medición. En las mediciones que no se recoge esta métrica, se devuelve el valor max.
* *“min”*: de todos los valores obtenidos durante la medición, el valor ínfimo.
* *“media”*: de todos los valores obtenidos durante la medición, el valor promedio.
* *“total”*: de todos los valores obtenidos durante la medición, el valor total.

## Almacenamiento de datos

Dando por finalizada la fase de limpieza y procesamiento de datos, se obtiene el conjunto de datos que finalmente será utilizado en el proyecto. Seguidamente, en la fase de almacenamiento, estos datos son almacenados en una base de datos, la cual es seleccionada dependiendo la necesidad de cada proyecto. En la situación de Biker, MongoDB fue la opción más acorde para las necesidades del proyecto, permitiendo tener una mayor flexibilidad en las estructuras de los documentos de datos. Además, la experiencia previa con esta herramienta fue influyente cuanto menos en la decisión de utilizar esta opción.

En lo que a la distribución del almacenamiento de los datos se refiere, se conforman varias colecciones dentro de la base de datos creada bajo el nombre de *“TFM”*.

La primera de ellas se conoce bajo el nombre de *“calidad\_aire”* y se recogen todos los datos de la calidad del aire en las diferentes provincias del País Vasco. La segunda colección tiene el nombre de *“estaciones”*, y se almacenan los datos que tienen relación con las estaciones meteorológicas. La tercera colección se conoce como *“incidencias”*, siendo el almacén de los datos que reflejan las incidencias ocurridas en las carreteras. La cuarta colección toma el nombre de *“flows”* y se recogen todos los datos de los flujos de tráfico. Y por último, existe la colección *“meteo”*, donde se recogen los datos de los sensores de las estaciones meteorológicas. Las métricas almacenadas son las siguientes: humedad, temperatura, precipitaciones, velocidad media y dirección del viento.

**Figura 9.** *Colecciones de la base de datos “TFM” en MongoDB*

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

Fuente: Elaboración propia.

Además, se ha de mencionar que para cada una de las colecciones se debe definir la inserción de los datos mediante la librería pymongo de Python, con el fin de establecer una conexión entre la base de datos y el código ejecutado.

## Análisis de datos

Superando el ecuador del ciclo de vida del conjunto de datos, comienza la etapa de análisis de datos. Esta etapa se centra en la explotación de los datos que recogen la calidad del aire así como todos aquellos datos que tratan sobre las diferentes métricas de las estaciones meteorológicas y flujos de tráfico. En todas estas ocasiones, y en base a los históricos almacenados, se predicen valores futuros de todos estos datos gracias a diferentes técnicas de IA como la regresión lineal, Random Forest o redes neuronales. A continuación, se define el proceso de análisis realizado de forma completa.

### Análisis exploratorio de datos (EDA)

Inicialmente se realiza un EDA que aplica a todos los conjuntos de datos donde se pueden conocer diferentes características de cada uno.

En un primer instante, se muestran los 5 primeros registros de cada conjunto de datos, permitiendo tener una primera visión general de los datos. Seguidamente se conoce su estructura en mayor detalle, mostrando tanto el número de columnas como de filas que definen cada conjunto de datos. Se indaga en mayor profundidad en las columnas, mostrando el nombre y el tipo de dato que las componen, además de mostrar el número de registros no nulos existentes.

Se visualizan varios estadísticos, como el número total de valores, cuantos de ellos son únicos, el que más veces se ha repetido y el número de repeticiones. Asimismo, se realiza un desglose por cada una de las columnas que componen el conjunto de datos mostrando las veces que dicho valor ha aparecido. Al final de cada columna, aparece el nombre de la misma, el tipo de dato que se almacena y el número de registros que la compone.

Por último, se genera una matriz de correlación donde se pueden apreciar las relaciones existentes entre las columnas numéricas en cada conjunto de datos.

**Figura 10.** *Ejemplo de una matriz de correlación*

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Fuente: Elaboración propia.

### Análisis de los subconjuntos de datos

Tras un EDA genérico aplicado a todos los conjuntos de datos, se realiza un análisis más detallado por cada uno de los conjuntos de datos existentes. En ellos se puede encontrar desde técnicas de aprendizaje supervisado como la regresión hasta técnicas de aprendizaje no supervisado como el clustering.

#### Flujos de tráfico

En el conjunto de datos que se tratan los flujos de tráfico, únicamente se utiliza una técnica de aprendizaje no supervisado como es el clustering. El clustering tiene la finalidad de agrupar los datos existentes en un número determinado de grupos, sin la existencia de ninguna etiqueta por cada uno de los grupos conformados.

Por ello, las agrupaciones se realizan en base a la similitud de los elementos. En este caso concreto, se realizan 2 agrupaciones, en base a la afluencia de vehículos y a la velocidad de circulación. De esta forma, se puede clasificar la saturación de vehículos existente en la carretera, así como la velocidad media de circulación en las mismas, permitiendo al usuario final conocer estos datos.

Ambas agrupaciones se han dividido en 5 subclases que varían en el caso de la afluencia de vehículos desde una saturación mínima de las carreteras hasta una afluencia más densa del tráfico. En el caso de la velocidad media ocurre algo similar, existiendo un abanico que abarca desde una velocidad de circulación considerablemente reducida hasta una velocidad elevada.

A continuación, se puede apreciar la distribución de ambos clústeres.

**Figura 11.** *Clustering de los flujos de tráfico*

Fuente: Elaboración propia.

#### Calidad del aire

En este conjunto de datos se opta por el uso de algoritmos de aprendizaje supervisado para obtener la predicción de algunos indicadores. Para ello, se definen 3 algoritmos, regresión lineal, Random Forest y una red neuronal, con el fin de realizar una comparativa entre sus resultados y concluir que solución obtiene mejores resultados.

El algoritmo de regresión lineal se inicializa a través de la función *“LinearRegression()”* de la librería *“sklearn”*, al igual que el algoritmo Random Forest con la función *“RandomForestRegressor()”*.

En el caso de la red neuronal es necesario un proceso ligeramente más complejo. En primera instancia es necesario definir el tipo de algoritmo, que en este caso es *“Sequential()”*. Seguidamente se añaden tantas capas de neuronas como se quiera. En este caso la red neuronal es consta de 3 capas, 2 de ellas son conformadas por 512 neuronas con activación *“relu”* además de la capa de salida que tiene activación *“linear”*, dado que está construida para predecir valores numéricos. Entre las capas se define un *“Dropout”* lo que ocasiona un descarte controlado de resultados para evitar el sobreajuste. Por último, se compila la red neuronal con el optimizador *“adam”* y se definen tanto la perdida como las métricas que evalúan la red neuronal. Se obtiene un resumen final mediante el atributo *“summary”.*

Con la red neuronal definida, se ha de aplicar el algoritmo de *“KerasRegressor”*, definiendo a su vez los ciclos (epochs) a ejecutar, así como la cantidad de datos que se utilizarán en cada ciclo (batch\_size).

Una vez configurados los 3 modelos, se define la validación cruzada a través de KFold. KFold es una técnica utilizada para evaluar el rendimiento de un modelo con la validación cruzada, importada también a través de la librería *“sklearn”.* En este caso se ha definido una validación cruzada en 5 subgrupos, para todos los modelos.

Con los modelos definidos y el KFold listo para aplicarse, únicamente son necesarias unas mínimas transformaciones en el conjunto de datos para poder realizar las predicciones. La primera adaptación consiste en eliminar una serie de columnas que imposibilita la predicción. De igual forma se deben eliminar las instancias de datos inconcluyentes, detectados por el EDA previamente realizado.

Seguidamente se agrupa el conjunto de datos por fecha y provincia, y se rellenan los valores nulos con ceros, dando a entender la ausencia de datos en esas casuísticas.

Con el conjunto ya preparado, se empiezan a realizar las predicciones para cada uno de los indicadores siguientes: Tolueno, S02, PM10, PM2,5, 03 8h y NOX.

El primer paso para cada predicción consiste en dividir la variable a predecir del resto, agrupándolas en las variables X (variables independientes) e Y (variable objetivo). Tras la división, se comienza con la predicción gracias a la función *“cross\_val\_predict”.*  Se obtienen los resultados de la predicción y se calculan las métricas utilizadas para su evaluación, en este caso el Mean Squared Error (MSE), Mean Absolute Error (MAE) y R2 Score.

**Figura 12.** *Resultados de las predicciones de la calidad del aire*

Texto

Descripción generada automáticamente

Fuente: Elaboración propia.

#### Incidencias

El conjunto de datos que engloba las incidencias utiliza una técnica de aprendizaje no supervisado como es el clustering, al igual que se hace con los datos de los flujos de tráfico. No obstante, existen algunas diferencias en ambos procedimientos. La más notoria es que se requiere de una adaptación de los tipos de fecha existentes ya que algunos registros contienen un formato con segundos mientras que otros los omiten.

Para ello se unifica el formato, añadiendo los segundos a aquellos registros faltantes y se convierte dicha columna a tipo fecha. Posteriormente se conforma un nuevo conjunto de datos agrupando por la fecha, provincia y tipo de incidencia.

Con esta agrupación realizada, se obtienen 3 clústeres en base al número de incidentes registrados, pudiendo considerar alta, media o baja la cantidad de incidentes ocurridos.

Se muestra la asignación en la siguiente ilustración.

**Figura 13.** *Clustering de las incidencias*

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Fuente: Elaboración propia.

Una vez realizada la asignación de a que clúster pertenece cada registro, se ha de hacer una unión para poder corresponder la información con el conjunto de datos previo a la agrupación. De esta forma se genera una clave común en el conjunto de datos sin agrupar y en el agrupado para obtener la totalidad de la información y poder realizar las visualizaciones posteriores.

#### Datos meteorológicos

Y para el último conjunto de datos, el cual engloba todos los datos correspondientes a los sensores meteorológicos, se sigue una estructura relativamente parecida a la aplicada en el conjunto de datos de la calidad del aire.

En primera instancia, existe una subdivisión en el conjunto de datos, causa de las diferentes mediciones que le aplica a cada uno de ellos. Por ello, se genera un conjunto de datos individual para cada una de las siguientes mediciones: humedad, temperatura, precipitaciones, velocidad media y dirección del viento.

Con los conjuntos conformados, se definen los mismos 3 algoritmos utilizados en el conjunto de datos de la calidad del aire, la regresión lineal, Random Forest y una red neuronal. Se ha de mencionar la existencia de pequeñas diferencias en su configuración, como la diferente dimensión de entrada en la red neuronal. Adicionalmente, se aplica el KFold, posibilitando la aplicación de la validación cruzada.

Tras definir los modelos y el KFold, da comienzo la iteración de las predicciones de los distintos conjuntos de datos. Primero se diferencian las variables independientes y objetivo, seguidamente se obtiene la predicción de cada conjunto y con cada modelo y por último se obtienen las métricas. Una vez obtenidos todos los resultados, se valora en líneas generales cual es el modelo que mejor rendimiento ha tenido.

**Figura 14.** *Resultados de las predicciones de los datos meteorológicos*

Imagen que contiene texto, celular, teléfono, sostener

Descripción generada automáticamente

Fuente: Elaboración propia.

### Exportar resultados del análisis

Con el análisis finalizado, se guardan los resultados en nuevos diccionarios, a través de los cuales son insertados en nuevas colecciones de la base de datos. De ahí, se exportan en formato CSV, con la finalidad posterior de importar esos datos en Power BI para su visualización. No obstante, existen 2 variantes, los resultados de las predicciones y las del clustering.

Los resultados obtenidos de las predicciones conforman un diccionario donde se almacenan las propias predicciones y las métricas que referencian. A continuación, se presenta un ejemplo.

**Figura 15.** *Resultados de las predicciones en la base de datos*

EJEMPLO RESULTADO PREDICCION MONGO DB

Fuente: Elaboración propia.

En los casos que se aplica clustering, se recoge el propio conjunto de datos, pero con la información adicional obtenida. Por ello, únicamente se utiliza este conjunto de datos en la visualización, a diferencia de las predicciones, que utilizan tanto el conjunto de datos original como los resultados del mismo.

**Figura 16.** *Resultados del clustering en la base de datos*

EJEMPLO RESULTADO clustering MONGO DB

Fuente: Elaboración propia.

## Visualización de datos

# Código fuente y datos analizados

## Repositorio

Durante el transcurso del proyecto, se ha ido publicando en el repositorio Github del proyecto las novedades y actualizaciones del proyecto con el fin de tener una copia de seguridad en la nube. De igual forma, si hubiera sido un proyecto en equipo, mediante este repositorio se podría haber compaginado el desarrollo del mismo de forma paralela, gracias a las herramientas que el mismo facilita.

Al ser un repositorio privado, en caso de requerir acceso a dicho repositorio contactar con [2001ikersebastian@gmail.com](mailto:2001ikersebastian@gmail.com).

## Código fuente

Es recomendable que el estudiante incluya en su memoria la URL del repositorio donde tiene alojado el código fuente desarrollado durante el TFE. El estudiante debe ser el único autor del código y único propietario del repositorio. En el repositorio no debe haber commit de ningún otro usuario del repositorio

## Datos Analizados

De igual forma, los datos que hayan utilizado para el análisis, siempre que así se considere oportuno, también deberían están alojamos en el mismo repositorio.

Si el TFE está asociado a una actividad o proyecto de Empresa, se debe justificar en la memoria que, por temas de confidencialidad, no se deja disponible ni el código fuente ni los datos utilizados.

# Plan de trabajo y gestión del proyecto

En un proyecto es esencial el contenido que se recoge en el mismo, así como su planificación. En esta sección se realizará un enfoque mayormente centrado en la gestión del trabajo, describiendo la estructura y planificación que conforma al proyecto.

## Planificación del proyecto

Este proyecto se ha decidido planificar siguiendo una metodología ágil, en este caso Scrum. Mediante esta metodología, se han ido desarrollando las diferentes fases que componen el ciclo de vida de un conjunto de datos de forma iterativa. Tras cada iteración, se ha revisado el correcto funcionamiento de todo el desarrollo previo, así como la correcta integración p

### Cronograma

### Análisis DAFO y CAME

## Problemas encontrados y desvíos

## Limitaciones del proyecto

Una vez concluido el trabajo, deberás hacer una **valoración crítica sobre el mismo y exponer las limitaciones que has encontrado** y que han marcado la realización de tu trabajo. Aquí se deberán hacer las consideraciones pertinentes sobre qué problemas o carencias se ha encontrado el autor para el desarrollo del trabajo (necesidad de valorar otras variables, ampliar la muestra, utilizar otros instrumentos, etc.); estas serán las limitaciones del trabajo

# Resultados

## Indicadores del proyecto

## Resultados de los algoritmos

## Indicadores del proyecto

# Conclusiones

Este último apartado es habitual en todos los tipos de trabajos y presenta el resumen final de tu trabajo y debe servir para informar del alcance y relevancia de tu aportación.

Suele estructurarse empezando con un resumen del problema tratado, de cómo se ha abordado y de por qué la solución sería válida.

Es recomendable que incluya también un resumen de las contribuciones del trabajo, en el que relaciones las contribuciones y los resultados obtenidos con los objetivos que habías planteado para el trabajo, discutiendo hasta qué punto has conseguido resolver los objetivos planteados. Las conclusiones ofrecidas deberán ser consecuencia del trabajo realizado y, por lo tanto, deberán marcar el grado de consecución de los objetivos propuestos (cada objetivo del trabajo se enlazará con una conclusión).

# Trabajo futuro

Finalmente, se suele dedicar un último apartado a hablar de líneas de trabajo futuro que podrían aportar valor añadido al trabajo realizado. La sección debería señalar las perspectivas de futuro que abre el trabajo desarrollado para el campo de estudio definido. En el fondo, debes justificar de qué modo puede emplearse la aportación que has desarrollado y en qué campos.

Referencias bibliográficas

*Una vez que el trabajo está terminado, hay que revisar el apartado “Referencias bibliográficas”. Si has usado un sistema automático (un gestor bibliográfico tipo Endnote, Refworks o Mendeley), inserta la bibliografía en la opción adecuada (APA).*

*Si lo has ido haciendo manualmente, repasa que todo es correcto: aparecen todas las referencias citadas en el texto, los autores están ordenados alfabéticamente por apellidos, las cursivas son correctas, los artículos tienen números de páginas, no faltan años ni ciudades de edición, se cumple en todas las referencias la normativa APA, etc.*

*Se recomienda evitar citas que hagan referencia a Wikipedia y que no todas las referencias sean solo enlaces de internet, es decir, que se vea alguna variabilidad entre libros, congresos, artículos y enlaces puntuales de internet.*

*Ejemplos:*

*Swanson, E., Barnes, M., Fall, A. M., & Roberts, G. (2017). Predictors of Reading Comprehension Among Struggling Readers Who Exhibit Differing Levels of Inattention and Hyperactivity. Reading & Writing Quarterly, 34(2), 132-146. doi:10.1080/10573569.2017.1359712*

45 personas murieron en las carreteras en Euskadi en 2023, casi la mitad motoristas, viandantes o ciclistas. (2024, enero 4). *elDiario.es*. https://www.eldiario.es/euskadi/45-personas-murieron-carreteras-euskadi-2023-mitad-motoristas-viandantes-ciclistas\_1\_10813252.html

Agencia Estatal de Meteorología. (s. f.). *MeteoRuta—Agencia Estatal de Meteorología—AEMET. Gobierno de España*. Recuperado 15 de abril de 2024, de https://www.aemet.es/es/eltiempo/prediccion/meteoruta

Analytics, B. (2023, septiembre 14). *Business intelligence y visualización de datos: Beneficios y herramientas imprescindibles - Bimex Analytics*. Business intelligence y visualización de datos: beneficios y herramientas imprescindibles - Bimex Analytics. https://bimexanalytics.com/blog/la-visualizacion-de-datos-es-un-elemento-fundamental-del-business-intelligence-bi-se-trata-de-representar-graficamente-datos-complejos-para-convertirlos-en-informacion-comprensible-y-accesible/, https://bimexanalytics.com/blog/la-visualizacion-de-datos-es-un-elemento-fundamental-del-business-intelligence-bi-se-trata-de-representar-graficamente-datos-complejos-para-convertirlos-en-informacion-comprensible-y-accesible/

*Apache Cassandra | Apache Cassandra Documentation*. (s. f.). [Apache Cassandra]. Recuperado 19 de mayo de 2024, de https://cassandra.apache.org/\_/index.html

*Base de datos rentable y de alto rendimiento*. (s. f.). [Oracle]. Recuperado 19 de mayo de 2024, de https://www.oracle.com/es/database/

*Bases de datos no relacionales | Bases de datos de gráficos | AWS*. (s. f.). Amazon Web Services, Inc. Recuperado 19 de mayo de 2024, de https://aws.amazon.com/es/nosql/

BOE.es - DOUE-L-2016-80807 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)., Reglamento (UE) 2016/679 § 1 (2018). https://www.boe.es/buscar/doc.php?id=DOUE-L-2016-80807

Casero, A. (2023, septiembre 14). Control de versiones con Git: Conceptos básicos. *Control de versiones con Git: conceptos básicos*. https://keepcoding.io/blog/control-de-versiones-con-git/

Castro, D. (2022, febrero 21). ¿API? Todo lo que debes saber. *Medium*. https://davidcasr.medium.com/api-todo-lo-que-debes-saber-d8b011d30aff

Cesvimap. (2021, septiembre 4). Influencia de la climatología en los accidentes de tráfico. *Revista CESVIMAP*. https://www.revistacesvimap.com/influencia-de-la-climatologia-en-los-accidentes-de-trafico/

Comobity, la App «segura» para ciclistas. (2015, noviembre 13). *13 de noviembre de 2015*. https://revista.dgt.es/es/reportajes/2015/11NOVIEMBRE/1113Comobity-la-App-segura-para-ciclistas.shtml

Daniel. (2022, enero 10). Data Warehouse: ¿qué es y cómo utilizarlo? *Formación en ciencia de datos | DataScientest.com*. https://datascientest.com/es/data-warehouse-que-es-y-como-utilizarlo

Datademia. (2021, septiembre 6). ¿Qué es Web Scraping? *Datademia*. https://datademia.es/blog/que-es-web-scraping

Esmerado, J. (2024, enero 20). Las 5 etapas del ciclo de vida de los datos. *Medium*. https://medium.com/@esmeradovela/las-5-etapas-del-ciclo-de-vida-de-los-datos-f4ddc3c1ed3f

Eurosport. (2023, noviembre 24). Muere un ciclista vasco de 19 años al ser atropellado mientras entrenaba—Eurosport. *Eurosport Espana*. https://www.eurosport.es/ciclismo/muere-un-ciclista-vasco-de-19-anos-horas-despues-de-ser-arrollado-por-un-coche-mientras-entrenaba\_sto9894016/story.shtml

Fernández, Y. (2019, octubre 30). Qué es Github y qué es lo que le ofrece a los desarrolladores. *Xataka*. https://www.xataka.com/basics/que-github-que-que-le-ofrece-a-desarrolladores

*GanttProject: Características, ventajas y desventajas*. (2024, abril 4). Gantt Chart GanttPRO Blog. https://blog.ganttpro.com/es/ganttproject-caracteristicas-ventajas-y-desventajas/

*Información legal*. (2010, marzo 17). [Open Data Euskadi]. https://opendata.euskadi.eus/general/-/informacion-legal-opendata/

*Información meteorológica tráfico—Incidencias en carretera*. (s. f.). www.tutiempo.net. Recuperado 15 de abril de 2024, de https://www.tutiempo.net/trafico.html

Lanzan una app para detectar automáticamente los accident... (2022, noviembre 26). *Ciclismo a Fondo*. https://www.ciclismoafondo.es/noticias/lanzan-app-detectar-automaticamente-accidentes-ciclistas\_268495\_102.html

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, Pub. L. No. Ley Orgánica 3/2018, BOE-A-2018-16673 119788 (2018). https://www.boe.es/eli/es/lo/2018/12/05/3

López, J. F. (2018, febrero 21). *Teorema de Bayes—Qué es, fórmula y ejemplos*. Economipedia. https://economipedia.com/definiciones/teorema-de-bayes.html

Mapa interactivo de la DGT y AEMET, esencial para planificar tu ruta cuando hace mal tiempo. (2023, febrero 24). *La Vanguardia*. https://www.lavanguardia.com/motor/consejos/20230224/8780046/mapa-carretera-interactivo-aemet-dgt-planificar-mejor-ruta-nieve-lluvia.html

Michael T. Gibson, Todd Curtin, & Amit Jhalli. (2020, junio 30). Los accidentes de bicicleta más comunes | Michael T. Gibson P.A., Auto Justice Attorney. *Michael T. Gibson P.A. Auto Justice Attorney*. https://autojusticeattorney.com/es/common-bicycle-accidents/

*MongoDB: The Developer Data Platform*. (s. f.). MongoDB. Recuperado 19 de mayo de 2024, de https://www.mongodb.com

*MySQL*. (s. f.). [MySQL]. Recuperado 19 de mayo de 2024, de https://www.mysql.com/

Na8. (2023, marzo 3). *Sets de Entrenamiento, Test y Validación | Aprende Machine Learning*. Aprende Machine Learning en Español. https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/

Navarro, S. (2022, abril 25). ¿Qué son Data Lakes? | KeepCoding Bootcamps. *¿Qué son Data Lakes?* https://keepcoding.io/blog/data-lakes/

*Neo4j Graph Database & Analytics – The Leader in Graph Databases*. (s. f.). [Neo4j]. Graph Database & Analytics. Recuperado 19 de mayo de 2024, de https://neo4j.com/

*Normativa*. (2017, junio 14). [Normativa]. Normativa. https://opendata.euskadi.eus/sobre-open-data/-/normativa-open-data/

*Open Data Euskadi | datos.gob.es*. (s. f.). Recuperado 7 de junio de 2024, de https://datos.gob.es/es/iniciativas/open-data-euskadi

Ortega, C. (2019, agosto 21). Métodos de recolección de datos más efectivos. *QuestionPro*. https://www.questionpro.com/blog/es/metodos-de-recoleccion-de-datos/

PostgreSQL Global Development Group. (2024, mayo 19). *PostgreSQL*. PostgreSQL. https://www.postgresql.org/

*Power BI: Visualización de datos | Microsoft Power Platform*. (s. f.). [Power BI]. Recuperado 20 de mayo de 2024, de https://www.microsoft.com/es-es/power-platform/products/power-bi

Press, E. (2023, enero 7). Los accidentes con ciclistas en Euskadi crecen un 20% en los últimos cinco años y los heridos graves y fallecidos un 9%. *7 de enero de 2023*. https://www.europapress.es/euskadi/noticia-accidentes-ciclistas-euskadi-crecen-20-ultimos-cinco-anos-heridos-graves-fallecidos-20230107120751.html

*QlikView – Analítica y cuadros de mando interactivos y eficaces | Qlik*. (s. f.). QlikView – Analítica y cuadros de mando interactivos y eficaces | Qlik. Recuperado 20 de mayo de 2024, de https://www.qlik.com/es-es/products/qlikview

*¿Qué es D3? | D3 por observable*. (s. f.). ¿Qué es D3? Recuperado 20 de mayo de 2024, de https://d3js.org/what-is-d3

*¿Qué es el aprendizaje no supervisado? | IBM*. (2023, mayo 4). [IBM]. https://www.ibm.com/es-es/topics/unsupervised-learning

*¿Qué es el aprendizaje supervisado? | IBM*. (2024, mayo 10). [IBM]. https://www.ibm.com/es-es/topics/supervised-learning

*¿Qué es una base de datos NoSQL? | IBM*. (2024, abril 14). [IBM]. https://www.ibm.com/es-es/topics/nosql-databases

*¿Qué es una base de datos relacional?* (s. f.). Oracle. Recuperado 19 de mayo de 2024, de https://www.oracle.com/es/database/what-is-a-relational-database/

*¿Qué son los almacenes de datos?* (s. f.). Google Cloud. Recuperado 19 de mayo de 2024, de https://cloud.google.com/learn/what-is-a-data-warehouse?hl=es

*¿Qué son los lagos de datos?* (s. f.). Google Cloud. Recuperado 19 de mayo de 2024, de https://cloud.google.com/learn/what-is-a-data-lake?hl=es

*Redis—La plataforma de datos en tiempo real*. (s. f.). [Redis]. Redis. Recuperado 19 de mayo de 2024, de https://redis.io/es/

*SAP BusinessObjects | Plataforma y suite de Business Intelligence (BI)*. (s. f.). SAP. Recuperado 20 de mayo de 2024, de https://www.sap.com/spain/products/technology-platform/bi-platform.html

*Software de análisis e inteligencia de negocios | Tableau*. (s. f.). [Tableau]. Recuperado 20 de mayo de 2024, de https://www.tableau.com/es-es

*Soluciones de data lake | IBM*. (s. f.). [IBM]. Recuperado 19 de mayo de 2024, de https://www.ibm.com/es-es/data-lake

*SQL Server 2022 | Microsoft*. (s. f.). [Microsoft SQL Server]. Recuperado 19 de mayo de 2024, de https://www.microsoft.com/es-es/sql-server/sql-server-2022

Tehreem Naeem. (2020, noviembre 7). ¿Qué es el repositorio de datos? (Definición, tipos y beneficios). *Astera*. https://www.astera.com/es/type/blog/data-repository/

Zapata, J. R. (2024, marzo 14). Visualización de Datos con Python. *Jose Ricardo Zapata*. https://joserzapata.github.io/courses/python-ciencia-datos/visualizacion/

Glosario

Por orden alfabético:

* Adam:
* Array:
* Bearer Token:
* Creative Commons:
* Cross Validation:
* Dashborad:
* Data Lake:
* Data Warehouse:
* Euskalmet:
* Fuzzy C-Means:
* Hardware:
* Input:
* K-Means:
* LinearRegression:
* Matplotlib:
* MongoDB:
* Naive Bayes:
* Open Data Euskadi:
* Output:
* Overfitting:
* Plotly:
* Postman:
* Power BI:
* Pymongo
* Relu:
* Seaborn:
* Snapshot:
* Summary:

Lista de siglas y acrónimos

Por orden alfabético:

* ACID:
* AEMET:
* API:
* BI:
* CSV:
* D3:
* DGT:
* EDA:
* EM:
* ERP:
* FIA:
* HTML:
* IA:
* IOT:
* JS:
* JSON:
* LOPDGDD:
* MAE:
* MSE
* RGPD:
* SAP:
* SQL:
* SVM:
* TFM:
* XML:

1. Privacidad y protección de datos

El presente anexo establece las directrices a seguir por el alumno en la elaboración de su memoria, cuando requiera cumplir con la normativa de privacidad y protección de datos personales. (**ver instruciiones**)