Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Matemáticas e Informática

Trabajo Fin de Grado

Open Data - Movilidad

Autor: María Mencía Serrano Manzano
Tutor(a): Luis Mengual Galán

Madrid, Junio 2024

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Matemáticas e Informática

Título: Open Data - Movilidad

Junio 2024

Autor: María Mencía Serrano Manzano
Tutor: Luis Mengual Galán
Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

Resumen

En la era digital estamos ante un crecimiento urbano continuo, la comunicación eficiente entre los distintos puntos de una ciudad se ha convertido en uno de los desafíos principales para las ciudades modernas. En un contexto donde la movilidad se ha vuelto una necesidad cotidiana para miles de ciudadanos, Madrid, como muchas otras metrópolis, se enfrenta a la compleja tarea de optimizar su sistema de transporte para satisfacer las demandas de su población. Cada día, un flujo constante de personas se ve desafiado por la necesidad de llegar a sus lugares de trabajo, colegios o universidades a tiempo, lo que implica enfrentarse al tráfico y a la puntualidad del transporte público. Es en este contexto surge Open Data-Movilidad como una respuesta a la necesidad apremiante de comprender y mejorar la movilidad urbana.

Este proyecto se enfoca en el tratamiento de la información pública disponible sobre el tráfico y movilidad de la Comunidad de Madrid y, dado que el volumen es considerable, desarrollar el ciclo de vida del dato completo utilizando técnicas de *big data*. En primer lugar, se recopilarán todos los datos de interés, a continuación se procederá a su limpieza y tratamiento, para llevar a cabo un análisis de ellos mediante técnicas de aprendizaje supervisado y no supervisado como regresión lineal o *k-means*. Esto permitirá identificar tendencias de tráfico y desarrollar modelos predictivos que ayuden a comprender el contexto territorial y a tomar decisiones informadas respecto al transporte público y la mejora de las vías y carreteras de la ciudad.

Abstract

In this digital era and in the middle of a continuous urban growth, the efficient communication between different places in a city, has become one of the main challenges for modern cities. In a context where mobility has turned to be a daily necessity for thousands of citizens, Madrid, as many other metropolis, encounters the complex task of optimising its transport system to meet its population demands. Every day, a constant flow of people is challenged by the need of making it on time to their workplaces, schools or universities, which means facing traffic and public transport's punctuality. It is in this context that "Open data - Mobility" comes to play as an answer to this urgent need of understanding and improving the urban mobility.

This project focuses on the processing of public information about traffic and mobility in the Community of Madrid and, given the considerable volume, developing a complete data life cycle using "big data" techniques. In first instance, all relevant data will be gathered; afterwards, it will be cleaned and processed to carry out its analysis with supervised and unsupervised learning techniques such as lineal regression or "k-means". This will allow us to identify traffic patterns and develop predictive models that would help comprehend the territorial context and make informed decisions regarding public transport and the improvement of the roads and motorways of the city.

Tabla de contenidos

Capítulo 1

Introducción

1.1. Contexto y motivación

Debido al progreso de las tecnologías de información, el volumen de datos que se produce diariamente ha incrementado considerablemente y, por tanto, las herramientas tradicionales no son capaces de procesarlos. Por ello, las organizaciones y empresas han tenido que desarrollar nuevas técnicas capaces de analizar y comprender más allá de lo que las herramientas tradicionales ofrecen sobre los datos, aumentando así su competitividad.

Este proyecto se enmarca en el ámbito de transporte, pues se utiliza la abundancia de datos públicos disponibles sobre el entorno urbano de Madrid. Estos datos abarcan una amplia gama de información relacionada con el transporte, carreteras y accidentes, ofreciendo una visión detallada y completa del ecosistema de movilidad de la ciudad.

El propósito fundamental de este proyecto es extraer conocimientos significativos a partir de esta información, utilizando nuevas técnicas de tratamiento de datos que permitan gestionar el gran volumen de información existente, contribuyendo así a la mejora continua del sistema de transporte y red de carreteras madrileñas. A través de técnicas como minería de datos, análisis estadístico y visualización de datos, se pretende identificar los patrones subyacentes de movilidad, así como las áreas de congestión, con el fin de proponer modelos predictivos basados en datos históricos que puedan anticipar y prevenir cuellos de botella en el futuro. Este enfoque no solo permite una mayor comprensión de la movilidad sino que también pretende facilitar la toma de decisiones informadas orientadas a mejorar la eficiencia del sistema en su conjunto.

Open Data-Movilidad se propone no solo recopilar y analizar datos relacionados con la movilidad urbana, sino también establecer un sólido ciclo de vida del dato que asegure la calidad y la utilidad de la información en todas sus etapas. Se siguen estrictas prácticas de gestión de datos desde la recolección inicial hasta el

análisis y la representación visual para asegurar la autenticidad y confiabilidad de la información.

La verdadera utilidad de la información radica en su capacidad para ser comprendida y utilizada por una amplia variedad de usuarios. El hacer el proyecto accesible hace que pueda servir como modelo para realizar análisis similares de cualquier zona geográfica a nivel regional o global.

1.2. Estructura del documento

- Capítulo I
  - Contexto y motivación: se expone la motivación del proyecto y los objetivos del mismo, recogiendo la idea general del trabajo a realizar.
  - Estructura del documento.
- Capítulo II
  - Estado del arte: introducción al *big data* y sus conceptos básicos.
  - Herramientas: se detallan las tecnologías que se utilizarán en el proyecto.
- Capítulo III
  - Diseño: se relata la estructura del proyecto y las fases de este.
  - Implementación: se describe detalladamente los pasos llevados a cabo en cada etapa.
- Capítulo IV
  - Resultados
  - Conclusiones
  - Líneas futuras
- Capítulo V
  - Análisis de impacto

Capítulo 2

Estado del arte

2.1. *Big data*

2.1.1. Introducción

La importancia de los datos en la sociedad actual radica en su capacidad para ayudarnos a entender nuestro entorno. En las últimas décadas, gracias al auge de Internet, los dispositivos móviles y sensores en todo tipo de ámbitos, se ha generado una inmensa cantidad de información sin precedentes. Esta ingente cantidad de datos, sumado a las tecnologías de vanguardia disponibles en la actualidad, han dado lugar alsurgimiento del concepto de *big data*.

Este término se refiere a los conjuntos de datos voluminosos y complejos que exceden la capacidad de procesamiento de los sistemas informáticos convencionales. Estos datos cuya procedencia es diversa, desde redes sociales, transacciones comerciales, dispositivos conectados a Internet, sensores en tiempo real etc. tienen un gran potencial, ya que ofrecen perspectivas valiosas que se pueden utilizar para mejorar la toma de decisiones, optimizar procesos e identificar tendencias y patrones entre otras cosas.

Por otro lado, el simple hecho de contar con grandes cantidades de datos no garantiza automáticamente su utilidad. La verdadera clave radica en la capacidad de analizar, interpretar y extraer información significativa de estos datos. Es aquí donde entran en juego técnicas avanzadas de análisis de datos, como la inteligencia artificial o *machine learning*. Estas herramientas permiten descubrir correlaciones, identificar anomalías, predecir comportamientos futuros y generar conocimiento útil a partir de los datos masivos disponibles.

En última instancia, el *big data* no solo representa un desafío tecnológico, sino también un cambio de paradigma en la forma en que comprendemos y abordamos la información [1].

2.1.2. Las 5V's

Los conjuntos de datos se enfrentan a diversos desafíos gracias a las características del *big data*. Estos se conocen como las 5 Vs: volumen, variedad, veracidad, velocidad y valor, que definen la problemática del *big data*.

Estas 5 características provocan que las empresas tengan problemas para extraer datos reales y de alta calidad de conjuntos de datos tan masivos, cambiantes y complicados. [2]

- Volumen: se define como la cantidad de datos que se generan y recopilan en cada instante en este mundo digitalizado. El gran volumen de datos plantea muchas dificultades como puede ser su almacenamiento en un lugar seguro y accesible, su distribución en distintos puntos sin perder la disponibilidad y coherencia a tiempo real y su procesamiento, pues se requiere una gran capacidad de cómputo. Las principales dificultades que se encuentran son el coste, la escalabilidad y el rendimiento. El incremento del volumen también es consecuencia del aumento de las fuentes de datos, cada día hay más personas conectadas y, de la calidad y precisión de estos datos (por ejemplo, la de los sensores) [1].

- Velocidad: además de gestionar grandes volúmenes de datos, las empresas necesitan obtener información rápidamente. Maximizar la velocidad con la que se crean, transmiten y procesan los datos puede ser un gran desafío. La información puede tratarse a tiempo real o con algo de demora, lo que es crucial en aplicaciones que requieren respuestas instantáneas como la detección de fraudes en operaciones financieras o la monitorización del tráfico.

- Variedad: este concepto hace refernecia a la diversidad de formatos, tipos y fuentes de información. Los datos no necesariamente están estructurados o semiestructurados, es decir, pueden no tener un esquema y estructura fijos pensados para ser almacenados en una base de datos tradicional; pueden ser objetos, documentos, imágenes, tuits o datos geoespaciales. Además, su origen es diverso, como las máquinas, las personas y los procesos organizativos. Hay numerosos factores que promueven la variedad, pero entre otros encontramos las tecnologías móviles, las redes sociales, las geotecnologías o los vídeos.

- Veracidad: la veracidad hace referencia a la calidad y el origen de los datos. Es crucial asegurar su coherencia, completitud, integridad y que estén libres de ambigüedades. Entre los diversos factores que impulsan la veracidad están el coste y la necesidad de trazabilidad. Dado el gran volumen, velocidad y variedad de los datos que se generan, hay que asegurarse de que la información que recibimos no sea falsa [1]. El concepto *"Garbage in, garbage out "* junto a *"Garbage in, gospel out"* que se entienden como aceptar ciegamente la información generada automáticamente proveniente de máquinas, ilustran que la entrada de datos sin sentido provoca la salida de información carente de este. En algunos casos se pueden "limpiar" los

datos de entrada, pero hay contextos como la economía, condiciones climáticas o decisiones de empresas que generan una incertidumbre que los sistemas big data han de asumir y tolerar [3].

- Valor: posiblemente la V más importante, el despliegue de tecnologías solo tiene sentido si los datos aportan algún valor o beneficio tangible y significativo. Este valor viene de reconocer patrones que mejoren la eficiencia operativa, impulsen la innovación o proporcionen ventajas competitivas [4].

Figura 2.1: 5V's

2.1.3. Ciclo de vida del dato

El ciclo de vida del dato es una sucesión de etapas por las que transcurren los datos a lo largo de toda su vida útil. Estas fases se definen en función de distintos criterios y el dato pasa de una a otra a medida que se completan distintas tareas o cumplen ciertos requisitos. Este periodo abarca desde la generación del dato hasta su reutilización o eliminación, y se considera un ciclo pues la información generada a partir de unos datos puede servir de base para un proyecto posterior, consiguiendo así que la última etapa del proceso retroalimente la primera [5].

Este ciclo proporciona una visión general de las etapas que intervienen en la generación, uso y reutilización del dato. Llevar a cabo correctamente cada etapa permite tratar el dato de una forma más eficiente, preservar su calidad y generar información de mayor valor. Además, en el ámbito empresarial permite llevar a cabo un uso más seguro, evitando pérdidas y eliminaciones, definiendo el trato, uso, almacenamiento y compartición de la información [6]. A continuación se describen las etapas de este ciclo.

- Generación o captura: en esta primera fase se produce la creación del dato en bruto, que se obtiene a través de distintas técnicas que pueden abarcar

desde la compra del dato hasta la creación automática de este gracias a dispositivos y sistemas automáticos.

- Almacenamiento: los datos ocupan un espacio y han de ser almacenados adecuadamente en repositoorios como las bases de datos. El correcto almacenamiento es clave para garantizar la accesibilidad y el control sobre los datos. En este proceso es importante diferenciar los datos estructurados de los no estructurados pues cada uno se almacenará de una forma distinta o en lugares y formatos distintos.
- Tratamiento: en esta etapa el dato es preparado, organizado y transformado para su uso. Se pueden utilizar técnicas de análisis de datos y aprendizaje automático para extraer el valor y utilidad de los datos adquiridos. Un buen tratamiento del dato garantiza la calidad de los resultados y proporciona una buena base para la toma decisiones bien fundamentadas.
- Uso: gracias a los resultados del análisis se toman decisiones estratégicas como la optimización de procesos o reformas en un sistema logístico.
- Eliminación: una vez el dato ha dejado de proporcionar información útil o ya no tiene un propósito significativo, se destruye. El volumen de datos en cualquier organismo crece considerablemente con el paso del tiempo y no es factible el almacenamiento de todos ellos.

Figura 2.2: Ciclo de vida del dato.

2.1.4. Aprendizaje automático

Tras comprender aspectos básicos del *big data* y las etapas por las que pasan los datos, el siguiente paso es entender qué herramientas se utilizan en la fase de tratamiento del dato para obtener el valor de este. El aprendizaje automático también conocido como *machine learning* es una subcategoría de la inteligencia artificial que tiene como objetivo desarrollar sistemas capaces de mejorar su rendimiento de forma automática gracias a la experiencia y los datos que consumen, sin ser explícitamente programados [7]. El aprendizaje automático es capaz de, a través de algoritmos que identifican patrones, generar modelos que realizan predicciones. Cuantos más datos y de mayor calidad, mejores resultados se obtienen. Existen dos categorías dentro del aprendizaje automático, supervisado y no supervisado.

Aprendizaje automático supervisado

En este tipo de aprendizaje el modelo es entrenado a partir de un conjunto de datos previamente etiquetado, es decir, los datos ya contienen las repuestas correctas. El total de los datos se divide en dos, conjunto de entrenamiento y conjunto de prueba. Con el primer conjunto se entrena el modelo, es decir, se le enseñan los resultados que debe generar asociando las características de los datos a las etiquetas correspondientes. Una vez terminada esta fase, se utiliza el conjunto de prueba para que el modelo haga predicciones sobre este. Como ya se tiene la respuesta a esa predicción se puede evaluar la precisión del modelo.

Dentro del aprendizaje supervisado existen dos tipos de tareas, clasificación y regresión.

- Clasificación: las tareas de clasificación tratan de predecir etiquetas discretas, por ejemplo, cuando se recibe una llamada, saber si es *spam* o no. Esta clasificación puede ser binaria (llamada) o multiclase, por ejemplo, identificar qué enfermedad tiene un paciente dados unos síntomas.
- Regresión: los problemas de regresión tratan de predecir un valor continuo, como el precio de una casa basado en sus características o el consumo de luz de un hogar [8].

Aprendizaje automático no supervisado

El aprendizaje no supervisado utiliza datos que no están etiquetados. El objetivo aquí es explorar la estructura de los datos para encontrar algún patrón o secuencia y realizar una agrupación de los datos. Las técnicas de aprendizaje no supervisado son útiles para la segmentación de clientes, la organización de grandes bibliotecas de documentos y la detección de patrones atípicos o anomalías. Hay tres tareas principales que se resuelven con aprendizaje no supervisado: agrupación en *clusters*, asociación y reducción de dimensionalidad [9].

- Agrupación en *clusters*: consiste en la agrupación de datos que no están etiquetados en función de sus similitudes o diferencias. Los algoritmos de agrupación en *clusters* también conocidos como *clustering*, se emplean para

dividir información sin clasificar en grupos representados por estructuras o patrones en la información. Hay varios tipos de algoritmos de agrupación: exclusivos, superpuestos, jerárquicos y probabilísticos.

- Asociación: una regla de asociación es un método basado en reglas que trata de encontrar relaciones entre las distintas variables dentro de un conjunto de datos. Un ejemplo típico es el problema de la cesta de la compra, que permite a las empresas entender las relaciones entre los distintos productos y así desarrollar estrategias de venta cruzada y marketing. Por ejemplo, los clientes que compraron una barbacoa también compraron utensilios de cocina. Existen distintos algoritmos que generan reglas de asociación, como a priori, Eclat y FP-Growth, el algoritmo a priori es el más utilizado.

- Ajuste de dimensionalidad: en principio, puede parecer lógico que cuantos más datos se obtendrán resultados más precisos, pero también pueden afectar al rendimiento del modelo y dificultar la visualización del conjunto de datos. La técnica de reducción de dimensionalidad se basa en la reducción del número de entradas de datos a un tamaño gestionable y preservar la integridad del conjunto. Cuando el número de dimensiones es muy elevado, se aplica este método en la fase de preprocesamiento. Existen varias técnicas para realizar una reducción de dimensionalidad como el análisis de componentes principales, la descomposición en valores singulares o codificadores automáticos.

2.2. Técnicas y herramientas

En esta sección se describen las técnicas y herramientas que se utilizarán en las distintas fases del proyecto. Para desarrollar el trabajo se ha elegido el lenguaje python en el entorno Visual Studio Code. A continuación se describen las funcionalidades utilizadas.

2.2.1. Python

Python es un lenguaje de programación de alto nivel ampliamente utilizado, pues es simple, eficiente, fácil de aprender y se puede ejecutar en muchos entornos diferentes. Es utilizado para desarrollo de aplicaciones web, *software*, ciencia de datos y *machine learning* [10]. Además, es un lenguaje que tiene una gran biblioteca con códigos reutilizables para casi cualquier tarea. Entre ellos, para este proyecto se han utilizado:

- Pandas: es una librería especializada en el manejo y tratamiento de estructuras de datos. El origen de los datos puede ser archivos en formato CSV, Excel o bases de datos SQL. El acceso a estos se realiza mediante índices para filas y columnas y permite la reordenación y combinación de los datos mediante 3 estructuras, series, dataFrames y panel. En este proyecto se han utilizado las dos primeras [11].

- Re: una expresión regular consiste una secuencia de caracteres que da lugar a un patrón de búsqueda en un texto. Esta librería permite trabajar con expresiones regulares y encontrar patrones en el texto [12].
- Request: esta librería permite realizar peticiones http a una página web. Proporciona dos métodos principales, GET, para consumir de una API o extraer información de una página; y POST, para enviar contenido de un formulario de forma automática [13]. En este proyecto se ha utilizado GET.
- Os: permite usar funcionalidades del sistema operativo como abrir o guardar un archivo, manipular rutas o crear archivos y directorios temporales [14]. En este proyecto se utiliza para guardar de forma ordenada toda la información obtenida y poder acceder a ella para tratarla.
- Glob: este módulo, con un patrón especificado previamente, permite encontrar los nombres de rutas que se asemejan a este. Esta funcionalidad se ha utilizado para acceder a los distintos archivos almacenados en distintos bloques según la fase del proyecto.
- Numpy: es una librería especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. A través de *arrays* se representan colecciones de datos de un mismo tipo en varias dimensiones y proporciona funciones eficientes para la manipulación de estos [15] .
- JSON: utilizado para leer archivos de tipo JSON.
- Pyproj: utilizada para realizar operaciones con coordenadas geográficas. Los datos con los que se trabaja contienen ubicaciones geográficas de eventos y esta librería permite tratar estos datos.
- Unicode: es una herramienta que codifica cada elemento del lenguaje con un código propio y único. Cada carácter tiene asignado un número entero de 0 a 0x10FFF [16]. esta herramienta es utilizada para analizar cadenas de texto y realizar modificaciones en estas, como puede ser cambiar una letra de mayúscula a minúscula o suprimir una tilde a través de la codificación única de cada caracter.
- SqlAlchemy: es una librería que ofrece la posibilidad de conectarse a una base de datos, comunicarse con ella e intercambiar información en ambos sentidos, tanto para introducir información como para leerla. Se utilizará para tratar los datos una vez almacenados.

2.2.2. QGIS

QGIS, cuyas siglas se corresponden con Sistema de Información Geográfica, es una herramienta que permite la creación, visualización, análisis y edición de información geoespacial. Entre sus múltiples funcionalidades se encuentran:

- Edición de datos: QGIS permite editar y crear nuevos datos geográficos, como agregar nuevos puntos, líneas o polígonos a un mapa, o modificar los existentes según sea necesario.

- Etiquetado y simbolización: permite etiquetar elementos en un mapa con texto descriptivo y personalizable, así como aplicar diferentes estilos de simbología para representar los datos de manera visualmente atractiva y comprensible.
- Geocodificación: QGIS incluye herramientas para convertir direcciones o descripciones de lugares en coordenadas geográficas (geocodificación), lo que facilita la ubicación de lugares específicos en un mapa.
- Análisis avanzado: además de las herramientas básicas de análisis, QGIS también ofrece capacidades más avanzadas, como análisis de redes, interpolación espacial, análisis de visibilidad y análisis de terreno, que permiten realizar análisis complejos de datos geo espaciales. En este proyecto ha sido utilizado para representar los datos obtenidos de una forma intuitiva y sencilla, creando mapas de calor y visualizaciones dinámicas que muestran la evolución de los datos a lo largo del tiempo.
- Integración de datos externos: permite la integración de datos, como imágenes satelitales, datos climáticos e información demográfica, entre otros, para enriquecer y contextualizar los mapas y análisis realizados en QGIS.

2.2.3. Scikit-Learn

Es una biblioteca ampliamente utilizada para el preprocesamiento, la reducción de dimensionalidad, clasificación, regresión, *clustering* y selección de modelos. Ofrece una gran variedad de algoritmos y utilidades que hacen que sea una herramienta básica para realizar análisis de datos y modelado estadístico. Se caracteriza por tener una amplia variedad de módulos y algoritmos de aprendizaje automático, la posibilidad de extraer datos de repositorios y conjuntos de prueba [17].

Entre las distintas funcionalidades que ofrece se han utilizado las siguientes:

- Simple Imputer: esta función permite sustituir valores nulos por otros según distintas estrategias como la media, la moda o la mediana. Cuando se obtienen conjuntos de datos, es común que estén incompletos y hay que tratar adecuadamente estas situaciones. Con esta herramienta, todos los campos que no ofrecen información (nulos) pueden reemplazarse por valores aceptables conocidos.
- Standard Scaler: es una herramienta que permite estandarizar las características de los datos mediante la eliminación de la media y la varianza unitaria. Este proceso es crucial, debido a que muchos algoritmos de aprendizaje automático asumen que todas las características están centradas en cero y tienen varianza en la misma escala.

  En este proyecto se ha utilizado en la preparación de los datos para los modelos SVC y regresión logística.
- GridSearchCV: se trata de una herramienta que se utiliza para encontrar los hiperparámetros más óptimos para un modelo. Los modelos predictivos

tienen datos de entrada que el programador ha de elegir y, esta elección no es aleatoria. Para seleccionar las características que consigan el mejor rendimiento del modelo se utilizan distintas técnicas entre las cuales se encuentra esta. Funciona de tal forma que cuando le proporcionas una lista de posibles parámetros, prueba todas las combinaciones posibles y te retorna la combinación con la que se obtienen mejores resultados [18].

- Regresión logística: es un algoritmo de aprendizaje automático supervisado utilizado para la predicción de eventos binarios, es decir, solo hay dos resultados posibles, si o no. La regresión logística analiza la relación entre una o más variables independientes y clasifica los datos en clases discretas. Es utilizado en modelos predictivos, donde el modelo estima la probabilidad matemática de si una instancia pertenece a una categoría específica o no [19].
- Árbol de decisión: es un algoritmo de aprendizaje supervisado sin parámetros utilizado para tareas de clasificación y de regresión. Consta de un nodo raíz, ramas y nodos donde cada nodo representa un resultado posible dentro del conjunto de datos. Esta estructura proporciona una forma fácil de afrontar la toma de decisiones, permitiendo la mejor comprensión de por qué se tomó una decisión. Se trata de una estrategia de divide y vencerás [20].
- Bosque aleatorio: es un método que elabora múltiples árboles de decisión y los combina para obtener una predicción precisa y robusta. Es un modelo que tiene alta precisión y la capacidad de manejar conjuntos de datos grandes con numerosas variables. Destaca por su utilidad en la estimación de la importancia de ciertas características y evitar el sobre ajuste.
- XGBoost: utiliza una serie de árboles de decisión construidos de forma secuencial, donde cada árbol nuevo intenta corregir los errores cometidos por los árboles anteriores. Es un modelo con alta precisión, rendimiento y velocidad, capaz de trabajar con datos tanto numéricos como categóricos.
- SVC (Clasificación de vectores de soporte): este modelo de aprendizaje automático supervisado separa los puntos de datos mediante un hiperplano con la mayor cantidad de margen posible. Este margen es la distancia entre los dos puntos pertenecientes a clases distintas más cercanas. Es un algoritmo que ofrece alta precisión y gran velocidad [21].
- K-means: es una herramienta de aprendizaje no supervisado que agrupa un conjunto de datos en K grupos (*clusters*). Cada grupo tiene un representante llamado centroide, que es la media aritmética de los elementos que pertenecen al *cluster* y, este algoritmo, actúa de manera iterativa de tal forma que cada elemento esté más cerca de su centroide que de los centroides del resto de grupos [22].
- DBSCAN: es un algoritmo de *clustering* que consiste en que para cada observación se mira el número de puntos que se encuentran a una distancia máxima épsilon de ella, denominados vecinos. Si un punto tiene al menos un mínimo número de vecinos, se denota como observación central y, los

vecinos de este punto pertenecen al mismo *cluster*. Una observación que no es central y que no sea vecina de ninguna central es una anomalía o valor atípico [23].

- *Clustering* jerárquico: es un algoritmo de aprendizaje no supervisado que agrupa los datos en función de la distancia entre ellos. Para su explicación se utilizará un dendrograma. En este los datos de partida son A, B, C, D, E, F, G, H y se muestra como se van relacionando los elementos dos a dos.

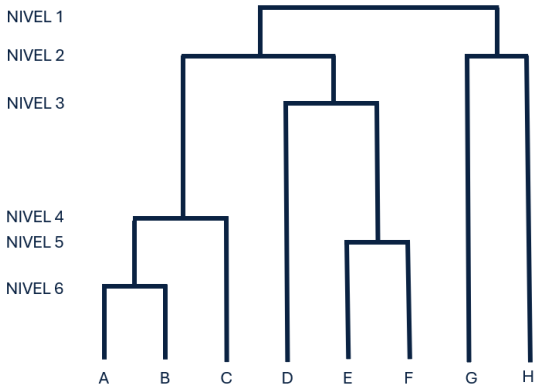

Figura 2.3: *Clustering* jerárquico.

Este muestra en qué orden se han ido realizando las agrupaciones hasta obtener un solo *cluster*, a continuación se elige un nivel y se podan las hojas de nivel igual o superior a este, los nodos hoja resultantes son los *clusters*.

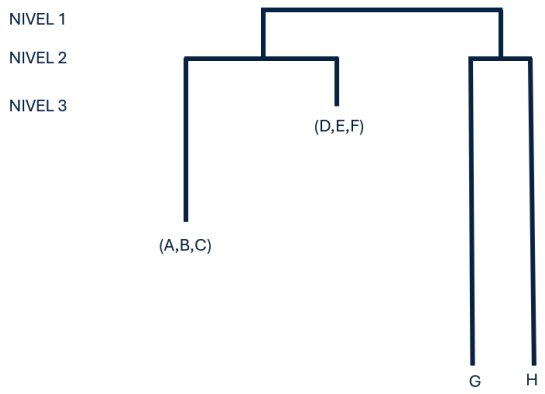

Figura 2.4: Poda en el nivel 3.

2.2.4. PostgreSQL

Es un motor de bases de datos relacional de código abierto. Destaca por su confiabilidad y robustez. Sigue un modelo relacional y es altamente escalable, es decir, permite manejar grandes volúmenes de datos. Este ha sido el motor elegido para almacenar la información del proyecto.

2.2.5. Docker

Docker es un software que permite crear y probar aplicaciones rápidamente. Empaqueta unidades de software en estructuras llamadas contenedores dentro de las cuales se encuentran todas las herramientas necesarias para la ejecución de aplicaciones. En este proyecto ha sido utilizado para levantar la base de datos PostgreSQL donde se almacenan los datos utilizados.

2.2.6. *Web scraping*

Este término hace referencia al proceso necesario para extraer contenidos e información de sitios web mediante software de manera automatizada. Por ejemplo, los comparadores de precios de vuelos y hoteles utilizan *web scraping* para rastrear varias páginas web en busca de la información demandada por el usuario. Hoy en día es posible rastrear todo tipo de datos en la web, desde *feeds* en RSS hasta información gubernamental, pero esta información no siempre se alcanza fácilmente. Dependiendo de la web, el rastreo se puede hacer usando APIs u otras herramientas [24].

Con el *web scraping* se extrae el contenido HTML de las webs para filtrar la información requerida y almacenarla. Para ello, en este proyecto se obtiene el HTML mediante el uso de la librería request mencionada anteriormente y se filtra esta información con las herramientas de la librería re.

Capítulo 3

Desarrollo

3.1. Metodología

Para llevar a cabo el proyecto se han establecido las siguientes fases o etapas. Extracción de datos, limpieza, preprocesamiento, almacenamiento, análisis y visualización. A continuación se describe el objetivo de cada fase.

- Extracción de datos: en esta primera fase del trabajo el objetivo es encontrar la información necesaria para llevar a cabo el proyecto. En esta etapa se identificarán las fuentes de datos disponibles como pueden ser datos en la nube, conjuntos de datos públicos o APIs externas con acceso a datos en tiempo real. Antes de descargar un conjunto de datos se ha de comprobar la integridad, consistencia y precisión de estos así como su actualidad. A continuación se ha de evaluar la relevancia de los datos para el objetivo del proyecto, su capacidad para proporcionar información valiosa y la relación que tienen entre ellos.

  Una vez identificadas las fuentes de datos a utilizar, se ha de iniciar el proceso de descarga mediante métodos como el *web scrapping*, la conexión a APIs o la descarga directa de datos disponibles en repositorios en línea. Por último, se crearán ejecutables automatizados para realizar la descarga, configurar conexiones seguras con las posibles bases de datos o implementar procesos de extracción.

- Limpieza: el objetivo de esta fase es garantizar la integridad, consistencia y calidad del conjunto de datos. Se han de identificar, corregir y eliminar posibles inconsistencias, errores y datos irrelevantes o duplicados. Este proceso contribuirá a mejorar la precisión de los resultados, optimizará el rendimiento de las soluciones analíticas y proporcionará una buena base para la toma de decisiones informadas.

- Preprocesamiento: esta fase es crucial para el análisis pues dota a los distintos conjuntos de datos de la estructura adecuada para poder ser cruzados correctamente. Esto puede incluir la selección de características relevantes, la normalización de datos, la codificación de variables categóricas,

la reducción de dimensionalidad y la división de datos en conjuntos de entrenamiento y prueba.

- Almacenamiento: en esta etapa, se determina el método más apropiado para almacenar los datos de manera eficiente y segura, teniendo en cuenta factores como el volumen de datos, la frecuencia de acceso y la necesidad de mantener la integridad y la confidencialidad de los mismos.

  Las opciones de almacenamiento pueden incluir bases de datos relacionales, sistemas de gestión de bases de datos NoSQL, sistemas de archivos distribuidos o almacenamiento en la nube. Se deben considerar aspectos como la escalabilidad, la disponibilidad y el rendimiento del sistema de almacenamiento seleccionado.

- Análisis: en esta etapa, se aplican técnicas y algoritmos analíticos para extraer información significativa de los datos preprocesados. Esto incluye tareas como la identificación de patrones, tendencias y relaciones en los datos, así como la construcción de modelos predictivos o descriptivos.

- Visualización: la visualización de datos es una herramienta fundamental a lo largo de todo el proyecto. Ayuda a entender los datos con los que se trabaja y su distribución. También es fundamental para el proceso de análisis, ya que permite comunicar de manera efectiva los hallazgos y resultados. Se utilizan diversas técnicas y herramientas de visualización para representar gráficamente los datos y facilitar su interpretación. Esto puede incluir gráficos de barras, diagramas de dispersión, mapas de calor y gráficos de líneas, entre otros. Una visualización efectiva puede ayudar a identificar patrones, tendencias y anomalías en los datos, así como a comunicar de manera clara y concisa los resultados del análisis.

  Aunque se haya representado como última etapa, la visualización es una herramienta que se utilizará a lo largo de todo el proyecto.

Figura 3.1: Fases del proyecto.

3.2. Extracción de datos

El proceso de extracción de datos es crucial pues sienta las bases para el resto del proyecto. La obtención de datos relevantes y confiables es el punto de partida

para llevar a cabo el análisis.

Se ha buscado en numerosas páginas de datos abiertos sobre el transporte público de Madrid. En ellas se encuentra una amplia gama de datos y, tras hacer una búsqueda exhaustiva, se ha decidido utilizar información del año 2022 pues es el año del cual se ha encontrado más información relevante completa. Del año 2023 hay información faltante de los últimos meses como noviembre o diciembre, y para poder hacer un estudio completo anual se han tomado los datos del año 2022. A continuación se describe la información encontrada y el procedimiento utilizado para extraerla.

- Datos Madrid: gracias a la página datos.madrid.es se ha encontrado información sobre accidentes de tráfico, accidentes de bicicleta y aforos de tráfico.

  Para obtener esta información se ha creado el archivo ”descar_csvs.py”. Este fichero descarga los siguientes archivos de la web en formato CSV: Aforos de tráfico en la ciudad de Madrid permanentes - Portal de datos abiertos del Ayuntamiento de Madrid, Accidentes de tráfico con implicación de bicicletas - Portal de datos abiertos del Ayuntamiento de Madrid y Accidentes de tráfico de la Ciudad de Madrid - Portal de datos abiertos del Ayuntamiento de Madrid.

  Este ejecutable trata dos escenarios, las webs que tienen la información mensual y las webs que contienen la información anual. En ambos casos el primer paso es el siguiente: se utiliza *web scraping*, se hace una llamada a la página web; después se parsea el HTML obtenido y así se consiguen los enlaces de descarga de los archivos. A continuación, se realiza un filtrado de los *href* para obtener los datos o bien mensuales o anuales del 2022. Estos se utilizan para hacer la petición de cada uno y así descargar todos los archivos.

  Se obtienen dos archivos anuales: “AccidentesBicicletas_2022.csv” y “AccidentesTrafico_2022.csv”. Ambos contienen información sobre los accidentes registrados en 2022 en Madrid.

  El primero consta de 877 entradas y recoge los accidentes con implicación de bicicletas y el segundo tiene 47.052 y data los accidentes de tráfico con implicación de vehículos, sin contener los accidentes entre bicicleta-bicicleta y bicicleta-peatón. Por ello nos interesa tener ambos. A continuación se muestra en la Tabla 1 los campos de cada registro de los ficheros y un ejemplo en la Figura 2.

CAMPO	DESCRIPCIÓN
num_expediente	Identificador
fecha	Fecha del accidente
hora	Hora del accidente
localización	Calle del accidente
numero	Número de la calle
distrito	Distrito del accidente
cod_distrito	Numeración del distrito
tipo_accidente	Alcance, atropello...
estado_meteorológico	Tiempo
tipo_vehiculo	Tipo vehículo
tipo_persona	Conductor, pasajero
rango_edad	Edad
sexo	Hombre o Mujer
lesividad	Lesividad
cod_lesividad	Numeración de la lesividad
coordenada_x_utm	Latitud
coordenada_y_utm	Longitud
positiva_alcohol	Alcohol
positiva_droga	Droga
Unassigned:19	Columna vacía
Unassigned:20	Columna vacía

Cuadro 3.1: Campos y descripción

2022S000034; 02/01/2022; 0:05:00; CALL. MARIA TERESA SAENZ DE HEREDIA, 6; 6; 15; CIUDAD LINEAL; Caída; Despejado; Bicicleta EPAC (pedaleo asistido); Conductor; De 25 a 29 años; Hombre; 7; Asistencia sanitaria solo en el lugar del accidente; 444.462.918; 4.474.808.752; S; NULL

Figura 3.2: Ejemplo del contenido de la tabla.

En cuanto al aforo, en la web se encuentran los ficheros mensuales y, por tanto, se obtienen 12 archivos que se unirán en un único CSV anual llamado "AforosTrafico_2022.csv" eliminando los mensuales. Este archivo contiene la información de la cantidad de vehículos que pasan por determinados lugares a lo largo del día y hora a hora los 365 días del año 2022. Consta de 292.430 entradas antes de realizar la limpieza de datos. A continuación se muestra en la Tabla 2 los campos de cada registro del fichero y un ejemplo en la Figura 3.

CAMPO	**DESCRIPCIÓN**
FDIA	Día
FEST	Calle donde se mide el aforo
FSEN	Sentido de la calle, AM/PM
HOR1	Hora 1 y 13
HOR2	Hora 2 y 14
HOR3	Hora 3 y 15
HOR4	Hora 4 y 16
HOR5	Hora 5 y 17
HOR6	Hora 6 y 18
HOR7	Hora 7 y 19
HOR8	Hora 8 y 20
HOR9	Hora 9 y 21
HOR10	Hora 10 y 22
HOR11	Hora 11 y 23
HOR12	Hora 12 y 24
Unnamed: 15	Columna vacía

Cuadro 3.2: Campos y descripción.

02/12/22; ES32; 2-; 271; 145; 119; 97; 70; 72; 125; 158; 285; 317; 341; 390;

Figura 3.3: Ejemplo del contenido de la tabla.

Al descargar los datos del aforo en formato CSV la ubicación geográfica de los sensores (FEST) de tráfico no aparece, solo había un id para distinguir cada estación, por tanto, se necesita un archivo que identifique cada sensor con su ubicación. Los archivos de aforo mensuales en formato Excel contienen 3 páginas y en una de ellas se encuentra la información deseada, por tanto, se escogió uno, noviembre de 2022 y con el ejecutable "descarga_excel.py": y pandas se iteró por las páginas del archivo hasta dar con "Ubicación estaciones". Se convirtió esa hoja a un *dataframe* para luego transformarlo a un CSV llamado "EstacionesTrafico_2022.csv". Este tiene 120 entradas, 2 por cada una de las 60 estaciones registradas, dependiendo de la orientación del tráfico (norte, sur, este, oeste). A continuación se muestra en la Tabla 3 los campos de cada registro del fichero y un ejemplo en la Figura 4.

CAMPO	DESCRIPCIÓN
Estación	Id de la estación
Nombre	Calle donde se mide el aforo
Latitud	Latitud
Longitud	Longitud
Sentido	Orientación codificada
Orientación	N-S S-N E-O O-E

Cuadro 3.3: Campos y descripción

```
1,Paseo de la Castellana,"40,4319272588958",3,68910874956933",1.0,S-N
```

Figura 3.4: Ejemplo del contenido de la tabla

- Opendatasoft: Gracias a la página opendatasoft.com se ha encontrado el contorno de la zona centro de Madrid en formato JSON y el contorno de la Comunidad de Madrid en formato GEOJSON. Esta información será útil para poder visualizar los datos y entender mejor su ubicación en el mapa. Para obtener estos datos se ha creado el archivo "descarga_json.py". Al igual que en el caso anterior, se hace una petición a la página web gracias a su href y la respuesta se guarda en "ContornoMadrid.json" y "ContronoC-Madrid.geojson" respectivamente.

3.3. Limpieza

Esta etapa consiste en encontrar y corregir errores o inconsistencias mediante la eliminación de errores o duplicados, correcciones ortográficas o tratamiento de inconsistencias. En primer lugar, en todos los archivos se ha tomado la decisión de eliminar tildes, cambiar la letra 'ñ' por la 'n' y suprimir las diéresis para evitar problemas a la hora de almacenar los datos. También se ha establecido como separador de cada CSV el punto y coma (;) para tener una estructura unificada entre todos los archivos.

A continuación, se describen los cambios realizados específicos de cada archivo.

- AforosTrafico: en primer lugar, en el archivo aparecían numerosas filas cuyo único contenido era un punto y coma (;) y, por tanto, se eliminaron, quedando así 86.140 filas. Por otro lado, el CSV tiene dos sentidos por cada calle y las calles de un solo sentido tenían una fila de ceros en el sentido inexistente, por tanto, se han eliminado las filas nulas de aquellas calles de un solo sentido, quedando así 79.498 entradas. Además, se ha eliminado la columna 'Unnamed:15' que se encontraba vacía.
- Accidentes: se ha realizado la limpieza de los archivos "AccidentesBicicletas_2022.csv" y "AccidentesTrafico_2022.csv" simultáneamente con el objetivo de unirlos en un solo CSV. En accidentes de tráfico hay dos columnas que reciben el nombre de 'Unassigned:19' y 'Unassigned:20', como no ofrecen ninguna información han sido eliminadas, teniendo así ambos archivos

el mismo número de columnas.

Las columnas 'coordenada_x_utm' y 'coordenada_y_utm' de "AccidentesBicicletas_2022.csv" aparecían en un formato extraño, con los números separados de 3 en 3 mediante puntos, como si fuera un número entero. Por otro lado, los mismos campos en "AccidentesTráfico_2022" aparecían como un número decimal con 3 decimales separados por una coma (,). En ambos archivos se ha unificado el formato de estos campos a WGS (World Geodetic System) similar al de los sistemas GPS. Además, han sido renombrados a 'Latitud' y 'Longitud'.

Por último, se han unificado los ficheros mediante el 'num_expediente' de cada accidente y eliminado las filas duplicadas, quedando 47.046 (solo 7 casos repetidos).

- DatosEstaciones: en este archivo los campos latitud y longitud tenían como indicador decimal una coma (,) y se ha sustituido por un punto para que esté en el mismo formato que en el resto de archivos.

3.4. Preprocesamiento

Para llevar a cabo la fase de análisis, toda la información ha de estar bien preparada para poder ser tratada por los modelos. Para ello se han de realizar ciertas modificaciones en "EstacionesTrafico_2022", "AforosTrafico_2022" y "AccidentesTraffico_2022".

En primer lugar, se ha modificado la estructura del dataset "AforosTrafico_2022". En lugar de tener un campo por cada hora del día, se ha creado uno solo llamado 'hora' y otro 'aforo' con el aforo correspondiente a la hora del día. Además, se han renombrado los campos 'FDIA', 'FEST', 'FSEN'. En la Tabla 4 se muestra como ha quedado el CSV y un ejemplo de este en la Figura 5.

CAMPO	**DESCRIPCIÓN**
fecha	Fecha
estacion	Id de la estación
sentido	Sentido de la calle
hora	Hora del día
aforo	Número de vehículos

Cuadro 3.4: Campos y descripción

01/01/2022; 1; 1; 19; 1266.0

Figura 3.5: Ejemplo del cambio realizado

A continuación, en este mismo archivo, 'estacion' se corresponde con el 'id' de "EstacionesTrafico_2022.csv" por lo tanto, para poder unir mediante ese ID, ambos datos han de ser del mismo tipo, 'estacion' es de tipo *String* y se ha convertido

a tipo *int*. A continuación se muestra en la Figura 6 un ejemplo del tipo de dato antes del preprocesamiento y después de este.

Antes del preprocesado, estacion: ES01, después, estacion: 1

Figura 3.6: Ejemplo del cambio realizado

Por otro lado, también se han realizado modificaciones en "EstacionesTrafico_2022.csv". Se han renombrado todos los campos poniendo en minúscula los nombres para unificar el formato con el del resto de archivos.

"AccidentesTrafico_2022.csv" ha sido el archivo en el cual se han realizado más cambios.

En el campo 'estado_metereológico' había 5.285 entradas nulas, en 'tipo_vehiculo' 199, en 'lesividad' y 'cod_lesividad' 22. Ante esta falta de datos se tomó la decisión de imputarlos y para elegir el método de imputación se observaron las distribuciones de las 4 categorías en histogramas. Las 4 eran asimétricas y con una categoría claramente superior al resto, por lo tanto, se imputó con la moda. En el caso de 'estado_meteorológico' la moda era "Despejado", en 'tipo_vehiculo' era "turismo" y en 'lesividad' y 'cod_lesividad' era " Sin asistencia sanitaria" con su código correspondiente que es el 14.

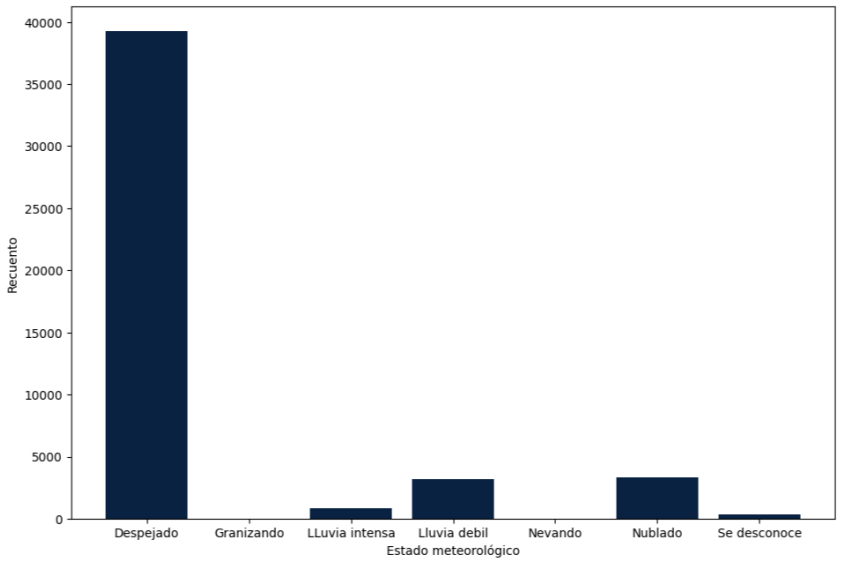

Figura 3.7: Histograma estado meteorológico

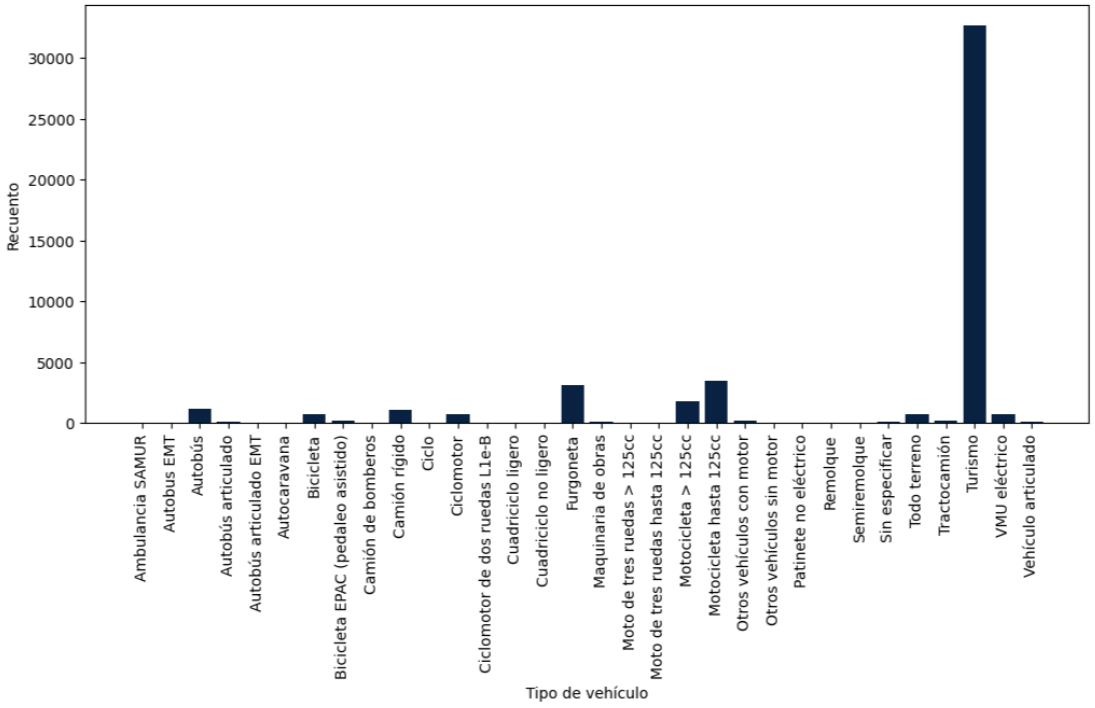

Figura 3.8: Histograma tipo de vehículo

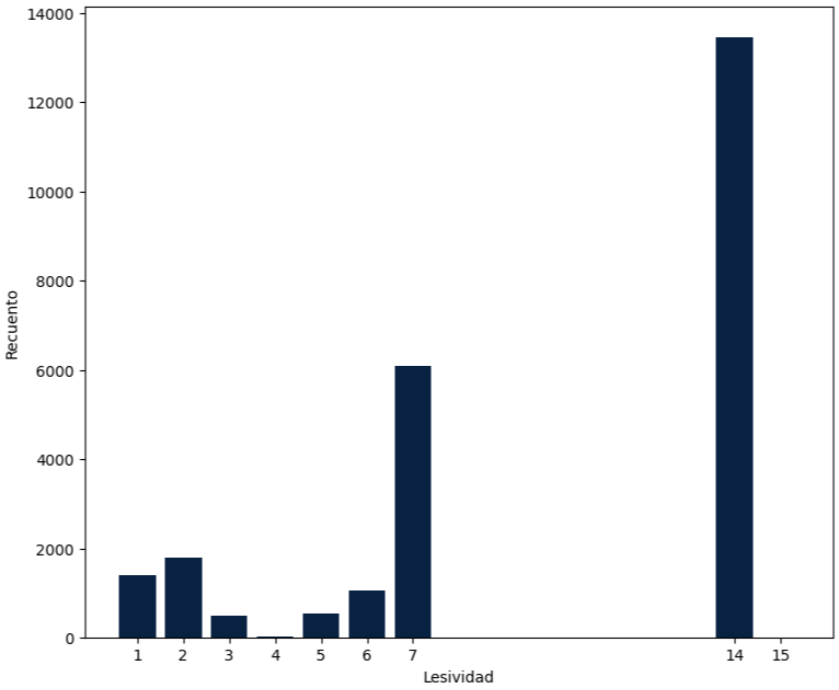

Figura 3.9: Histograma lesividad

De cara al análisis, a efectos prácticos, en el campo de lesividad el interés es saber si en un accidente se requiere la asistencia de medios sanitarios y, por tanto, su despliegue. Es por esto que se han creado dos clases, asignando un 0 a aquellos accidentes que no han requerido asistencia y un 1 a aquellos que si.

Por otro lado, en el campo 'rango_edad' no había entradas nulas, pero sí había 5.179 campos donde ponía "desconocido". Este ejemplo muestra la importancia de hacer un análisis exhaustivo preliminar, pues aunque a priori esa información parecía estar completa, no lo estaba. Se ha hecho el mismo procedimiento que en el caso anterior, se ha visto la distribución de la edad y al ver que era uniforme se han imputado los valores desconocidos con la media, que se encuentra en el rango 'De 40 a 44 años'.

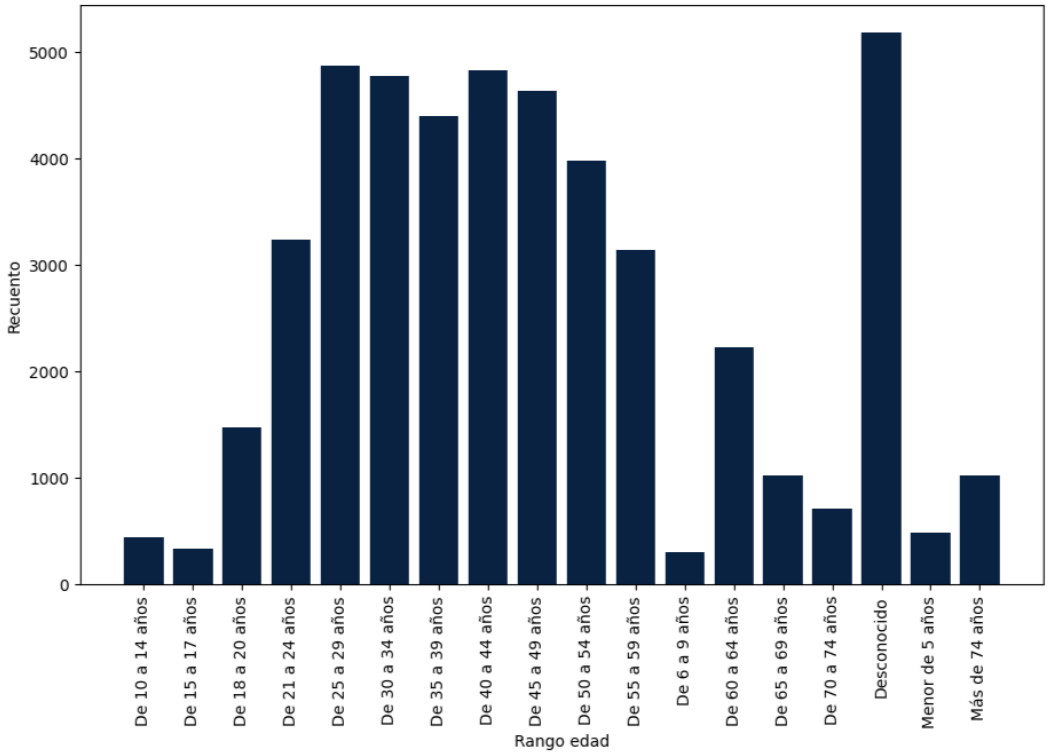

Figura 3.10: Histograma rango de edad

A continuación, el campo 'hora' proporciona información sobre la hora y el minuto del accidente; puesto que en los aforos tenemos la información por hora, se modificó este campo conservando solo esta, eliminando así los minutos.

Respecto al tipo de vehículo implicado en el accidente, se van a conservar todos. Existen 31 tipos en el CSV, para guardar esta información se ha decidido agrupar en categorías y crear una columna con cada una de ellas de tipo entero, que indica en número de vehículos de ese tipo implicado en el accidente. A

continuación se muestra la norma aplicada para realizar la agrupación.

- Turismo: turismo, VMU eléctrico, todoterreno.
- Bicicleta: bicicleta EPAC, bicicleta, patinete no eléctrico, otros vehículos sin motor.
- Moto: motocicleta hasta 125cc, motocicleta >125cc, ciclomotor, cuadriciclo ligero, cuadriciclo no ligero, ciclo, moto de tres ruedas >125cc, ciclomotor de dos ruedas L1e-B, moto de tres ruedas hasta 125cc.
- Furgoneta: furgoneta, autocaravana.
- Camión: camión rígido, vehículo articulado, tractocamión, semirremolque, remolque.
- Otros: camión de bomberos, otros vehículos con motor, ambulancia SAMUR, maquinaria de obras.

Los datos agrupados en la categoría "otros" podrían ser considerados datos atípicos y podrían ser eliminados, pero dado que el dataset no contiene un número de entradas considerablemente alto y esta categoría tiene 753 apariciones, se ha preferido preservar esta información.

En cuanto al tipo de accidente, también aparecen las distintas entradas con la siguiente frecuencia: 'Alcance' 4.089, 'Atropello a persona' 1403, 'Caída' 1.992, 'Choque contra obstáculo fijo' 3.073, 'Colisión frontal' 494, 'Colisión frontolateral' 4.690, 'Colisión lateral' 2.916, 'Colisión múltiple' 795, 'Otro' 753, "Despeñamiento" 1, 'Solo salida de la via' 131, 'Vuelco' 115 y 'Atropello a animal' 82 veces.

La entrada con el despeñamiento se ha eliminado debido a que solo aparece una vez y puede ser considerado dato atípico. Por otro lado, los 3 campos con menos frecuencia, es decir, 'solo salida de la vía', 'vuelco' y 'atropello animal', se han agrupado en la categoría 'Otros'.

Una vez terminadas las modificaciones campo por campo, vemos que en el dataset inicialmente aparecía una entrada por cada persona implicada en un accidente. El interés era tener una entrada por cada accidente y para ello se ha agrupado la información de las personas afectadas (sexo, rango_edad, lesividad, positiva_alcohol). La información general del accidente, es decir, la recogida en los campos 'fecha', 'hora', 'localización', 'numero', 'distrito', 'cod_distrito', 'estado_metereológico', 'tipo_accidente', 'latitud' y 'longitud' han permanecido intactos. Se ha creado un campo 'personas' en el que se recoge el número de estas implicadas en el accidente para no perder esta información. Por otro lado, se han tomado las siguientes decisiones respecto a los campos individuales de cada persona.

- 'positiva_droga' y 'positiva_alcohol': se han agrupado positivos en alcohol y drogas pues el número de campos nulos en 'positiva_droga' era considerable (solo 140 campos no nulos). En el caso de que uno de los involucrados en el accidente diera positivo en drogas o alcohol, aparece un 1 en este campo.

- 'lesividad' y 'codigo_lesividad': en estos campos se recoge si alguno de los afectados necesitó asistencia sanitaria o no en el accidente, por tanto, se guarda la información del mayor afectado.
- 'rango_edad': en este campo se ha conservado el rango de la persona con mayor edad, entendiendo que cuanto mayor sea una persona más gravemente puede afectarle un accidente.

Por último, se ha creado un campo 'geometry' con la combinación de las coordenadas latitud y longitud, que facilitará la visualización de los accidentes. A continuación se muestran los campos finales del archivo y un ejemplo de este.

CAMPO	**DESCRIPCIÓN**
num_expediente	Identificador
fecha	Fecha del accidente
localización	Calle del accidente
numero	Número de la calle
distrito	Distrito del accidente
cod_distrito	Numeración del distrito
tipo_accidente	Tipo de accidente
estado_meteorológico	Tiempo
rango_edad	Edad del mayor afectado
personas	Numero de personas implicadas
cod_lesividad	Asistencia sanitaria del mayor afectado
latitud	Latitud
longitud	Longitud
positiva_droga	(Longitud, Latitud)
turismo	Cantidad implicada en el accidente
autobús	Cantidad implicada en el accidente
bicicleta	Cantidad implicada en el accidente
camión	Cantidad implicada en el accidente
furgoneta	Cantidad implicada en el accidente
motocicleta	Cantidad implicada en el accidente
otros	Cantidad implicada en el accidente
geometry	Alcohol y drogas
hora	Hora del accidente

Cuadro 3.5: Campos y descripción

2022S000001; 01/01/2022; AVDA. ALBUFERA, 19; 19; 13; PUENTE DE VALLECAS; 0; 0; 49.0; 2; -3.667437045740308; 40.39741992179293; 0; 0; 0; 0; 0; 0; 0; 0; 2; 1; POINT (-3.667437045740308 40.39741992179293)

Figura 3.11: Ejemplo del contenido de la tabla

3.5. Almacenamiento

El almacenamiento es fundamental para gestionar grandes volúmenes de datos de manera eficiente y escalable. Tras explorar las distintas opciones disponibles, se optó por levantar una base de datos PostgreSQL en un *docker*. Esta opción ofrece varias ventajas como evitar problemas de incompatibilidades entre el sistema operativo y aplicaciones. Además, encapsula la base de datos y sus dependencias en un contenedor, lo que garantiza un entorno de desarrollo consistente.

La creación y levantamiento es un proceso rápido y sencillo que se realiza a través de dos comandos de *docker* y la conexión se realiza gracias a *sqlalchemy* y su opción create_engine. Gracias al método *tosql* se pueden subir los datos en escasos segundos. Además, los contenedores de *docker* ofrecen aislamiento de recursos, es decir, la base de datos se ejecuta en un entorno aislado y seguro, garantizadno que los procesos en ejecución dentro del contenedor no afecten a otros servicios y aplicaciones del sistema.

3.6. Análisis

En esta sección primero se va a realizar una exploración de los datos no solo para comprender el contenido de estos, sino para empezar a ver las principales tendencias de los conjuntos e identificar información relevante para su estudio. Una vez identificada, se aplicarán técnicas de aprendizaje supervisado y no supervisado para obtener información relevante de los datos. Comenzamos con el CSV de accidentes.

En primer lugar, gracias a la matriz de correlación nos podemos hacer una idea inicial de los datos que se tienen y la relación entre ellos. Esta, muestra información que puede parecer obvia a priori, como por ejemplo:

- El número de personas implicadas en un accidente tiene alta relación con accidentes en los que hay turismos involucrados, autobuses, a continuación furgonetas y con el vehículo que menos es con motos.
- El rango de edad está altamente relacionado con el tipo de vehículo 'turismo', es decir, el vehículo elegido por las personas mayores en primer lugar es el turismo y luego el autobús, y con el que tiene menos relación es con la moto.
- El número de positivos en droga tiene gran relación con la hora del día, siendo más alta cuanto más temprana es la hora, es decir, de madrugada.
- El vehículo más afectado por el estado meteorológico es la moto. Se entiende que es el vehículo en el que tanto conductor como pasajero son más afectados.
- El vehículo que tiene mayor lesividad es la motocicleta y después la bicicleta. Aunque la bicicleta sea más frágil que la motocicleta, esa alta relación tiene que ver con un mayor número de accidentes con motocicletas y tam-

bién con las altas velocidades que alcanzan estas en comparación con las bicicletas, lo que hace que el accidente pueda ser más peligroso a pesar de que la persona es menos vulnerable.

- El vehículo que tiene menor relación con la lesividad es el turismo, luego el camión y la furgoneta, es decir, son los más seguros.

Otras observaciones no tan obvias son las siguientes:

- Los casos de positivo en drogas tienen mayor relación con los accidentes con turismo implicados y con el tipo de vehículo con el que tienen menor relación es con la moto.
- Los casos en el que el vehículo es 'otro' tienen muy baja relación con todos los campos. Entendemos que esto se debe a que son casos muy aislados como accidentes con camiones de bomberos o tractores.
- El tipo de accidente tiene muy baja correlación con los tipos de vehículo.
- Los accidentes con motocicletas tienen muy poca relación con los accidentes en los que hay turismos implicados.

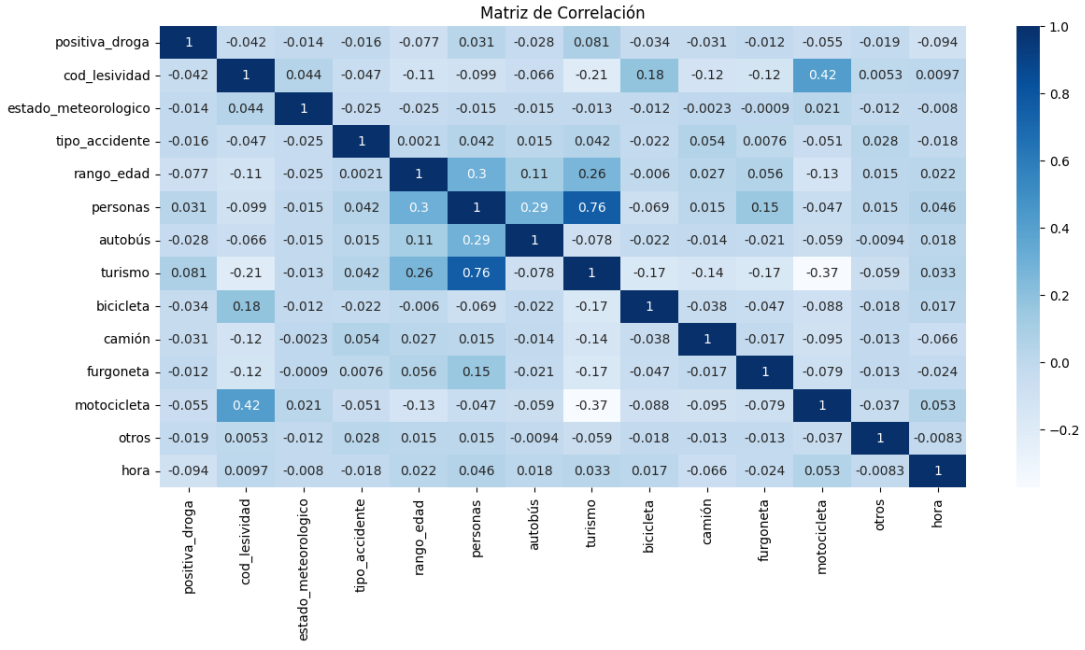

Figura 3.12: Matriz de correlación de accidentes.

A partir de aquí se entiende que se podrán hacer predicciones sobre variables como lesividad o el tipo de accidente. Por otro lado, la información geoespacial como el distrito será susceptible a técnicas de aprendizaje no supervisado para encontrar ubicaciones con las mismas características de accidentes.

Respecto al CSV de aforos cuya única información es el número de vehículos, la ubicación y la hora, se podrían hacer predicciones espacio temporales, pero es-

tas pueden ser complicadas y, en muchos casos, no son precisas debido a varios factores. En primer lugar, los datos temporales presentan patrones complejos y no lineales y, por tanto, modelarlos suele resultar inexacto. Además, son altamente sensibles a los datos de entrada y pequeñas variaciones en la información pueden afectar significativamente a los resultados. Por otro lado, a pesar de que se dispone de un volumen no muy reducido de datos, las predicciones temporales pueden verse afectadas por cambios en el entorno como eventos inesperados o tendencias emergentes (una obra, un festivo). Es por todo esto que a este archivo solo se le aplicarán técnicas de aprendizaje no supervisado, para intentar de esta forma obtener patrones y encontrar estaciones de medición de aforo con las mismas características.

3.6.1. Aprendizaje supervisado

En este apartado el objetivo es elaborar modelos predictivos sobre las variables vistas anteriormente.

Dados los tipos de datos que tenemos, nos encontramos ante problemas de clasificación como la edad, que viene dada en rangos 'De 40 a 44 años' o la lesividad (asistencia necesaria o no). En este apartado se buscan los modelos predictivos que mejor ajusten estas variables.

Para evaluar los modelos se utilizarán las métricas de *accuracy* (exactitud), *precision* (precisión), *recall* (recuerdo/memoria) y F1-score. Todas estas métricas se obtienen gracias a la matriz de confusión, que es una matriz que representa la calidad de un modelo. Tiene la siguiente estructura:

negativo	**VERDADERO NEGATIVO**	**FALSO POSITIVO**
positivo	**FALSO NEGATIVO**	**VERDADERO POSITIVO**
	negativo	positivo

Figura 3.13: Matriz de confusión

Las filas representan lo que el modelo debería predecir y las columnas lo que ha predicho. De esta forma, en la diagonal principal se encuentran los casos correctamente clasificados y en la anti-diagonal los erróneos; siendo un falso positivo un caso negativo que el modelo clasifica como positivo y un falso negativo, un caso positivo que el modelo clasifica como negativo [25]. A partir de estos datos se pueden calcular las métricas mencionadas anteriormente de la siguiente forma:

- *Accuracy* : es la métrica más usada y se define como la cantidad de veces que se acertó una afirmación sobre el total de los datos. Es decir, toma los

elementos de la diagonal principal en relación con el total.

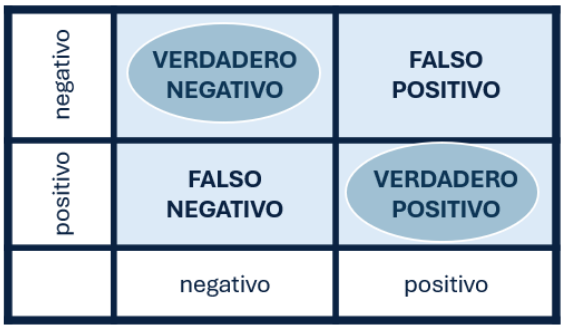

Figura 3.14: Matriz de confusión para *accuracy*

- *Precision*: esta métrica mide la cantidad de verdaderos positivos frente al total de positivos predichos, es decir, toma los verdaderos positivos sobre el total de la segunda columna.

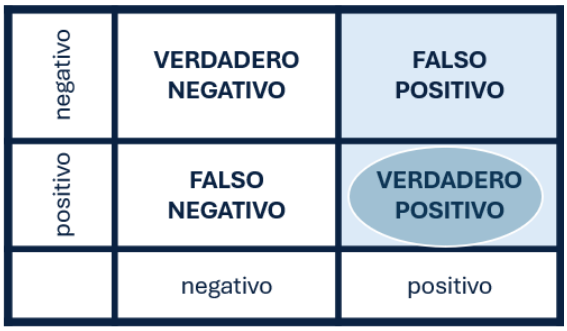

Figura 3.15: Matriz de confusión para precisión

- *Recall* : esta métrica compara la cantidad de verdaderos positivos sobre lo que realmente era positivo, es decir, toma los verdaderos positivos sobre el total de la segunda fila.

Figura 3.16: Matriz de confusión para *recall*

- *F1-score*: es el doble del producto de la precisión por el *recall* entre la suma de estos. Esta métrica es un indicador de alta precisión y alta sensibilidad.

$$F1 = 2 \times \frac{(\text{Precision} \times \textit{recall}\ )}{(\text{Precision} + \textit{recall}\ )}$$

Una vez conocemos las métricas para evaluar los modelos, empezamos a entrenarlos.

1. Positivo en droga: el caso positivo tiene asociado como etiqueta un 1 y el negativo un 0. Para abordar la predicción de esta situación se han probado 5 modelos: Regresión Lineal, Bosque aleatorio, Árboles de decisión, XGBoost y SVM. Antes de comenzar a aplicar los modelos hay que entender que este campo se encuentra considerablemente desbalanceado, esto quiere decir que las clases positiva y negativa en drogas no están representadas equitativamente, en este caso 3.709 negativos y 333 positivos. Este desequilibrio genera un sesgo significativo en los resultados de los modelos, lo que hace necesario un ajuste preciso para minimizar su impacto.

   En los primeros entrenamientos de los modelos se obtuvieron resultados no muy precisos y, tras realizar un ajuste de los hiperparámetros con técnicas como Grid Search, se obtuvieron los siguientes resultados.

Modelo	**Valor**	**Precision**	***Recall***	**F1-score**	***Accuracy***
Regresión	0	0.94	0.87	0.9	0.82
	1	0.20	0.39	0.27	0.82
Bosque aleatorio	0	0.92	1.00	0.96	0.92
	1	0.52	0.03	0.09	0.92
Árbol de decisión	0	0.94	0.88	0.92	0.85
	1	0.24	0.39	0.30	0.85
XGBoost	0	0.97	0.79	0.87	0.78
	1	0.23	0.72	0.35	0.78
SVM	0	0.95	0.64	0.71	0.64
	1	0.13	0.62	0.22	0.64

Cuadro 3.6: Resultados de los modelos aplicados

En primer lugar, se puede apreciar que la precisión de la predicción del valor negativo es muy alta en todos los modelos en comparación con la predicción del valor positivo, esto se debe al desbalance de los datos, sin embargo, este hecho puede resultar engañoso. Se observa que el modelo con mayor *accuracy* es el de bosque aleatorio, pero eso no significa que sea el más adecuado para nuestro caso. Este modelo solo clasifica correctamente el 0.03 % de los casos positivos, por lo tanto, queda descartado como mejor opción.

Hemos de buscar un modelo con la mayor precisión posible en la predicción de positivos y con un buen equilibrio o *recall*. Este es el caso del XGBoost, que es el modelo que más casos positivos clasificó correctamente.

En escenarios como este, los falsos positivos y negativos tienen implicaciones prácticas significativas. Los falsos positivos pueden resultar en gastos adicionales asociados con pruebas y análisis de drogas, mientras que los falsos negativos pueden comprometer la justicia y la seguridad pública al no identificar correctamente a los responsables de los accidentes. Por lo tanto, es esencial seleccionar un modelo que minimice tanto los falsos positivos como los falsos negativos.

2. Lesividad: este problema se trata de igual forma que el anterior, consiste en ver si se requiere asistencia sanitaria en el accidente o no. En este caso el problema está algo más balanceado pues hay 2.755 casos en los que no se necesitó asistencia y 1.286 en los que si. Se han probado los mismos 5 modelos que en el apartado anterior y las mismas métricas para evaluar el rendimiento del modelo, obteniendo los siguientes resultados.

Modelo	**Valor**	**Precision**	***recall***	**F1-score**	***Accuracy***
Regresión	0	0.85	0.87	0.86	0.80
	1	0.70	0.66	0.68	0.80
Bosque aleatorio	0	0.86	0.90	0.88	0.83
	1	0.75	0.68	0.71	0.83
Árbol de decisión	0	0.85	0.90	0.87	0.82
	1	0.75	0.66	0.70	0.82
XGBoost	0	0.86	0.89	0.87	0.82
	1	0.74	0.69	0.71	0.82
SVM	0	0.84	0.87	0.86	0.80
	1	0.70	0.66	0.68	0.80

Cuadro 3.7: Resultados de los modelos aplicados

En este caso se puede observar que tanto los valores obtenidos en *accuracy* como en *recall* son muy parecidos en todos los modelos. Cabe destacar la importancia del *recall* pues este indica, de todos los accidentes en los que se necesita una ambulancia, en cuantos se ha predicho la necesidad de esta. Es mejor enviar una ambulancia y que no haga falta (alta precisión) a no mandarla cuando si es necesaria (alto *recall*).

En este caso se pueden elegir dos modelos, el bosque aleatorio y el XGBoost que han sido los modelos con mejor desempeño. XGBoost tiene el *recall* en el caso positivo ligeramente superior al bosque aleatorio, pero este último tiene mejor *recall* en el caso negativo.

Estos resultados se deben a que ambos manejan bien el desbalance de clases mediante el ajuste de pesos de clases, que se refleja en mayor precisión y mejores puntuaciones en *F1-score*. Ambos modelos pueden capturar interacciones complejas entre características que podrían no ser captadas con modelos lineales como regresión logística.

Respecto al resto de categorías como la edad o el tipo de vehículo que se corresponden con ejercicios multiclase, se ha intentado hacer predicciones, pero debido al desbalance de estas clases y las pocas ocurrencias del algunas de ellas, las predicciones modeladas no han obtenido una precisión muy elevada, rondando entre el 30 %-40 %. Aquellas clases con mayor número de apariciones tenían una precisión, *recall* y *F1-score* razonables, pero al ponerlas en conjunto con las clases menos comunes. el rendimiento del modelo disminuye considerablemente.

En el apartado de preprocesamiento recordemos que se había llevado a cabo una agrupación de distintas categorías en una sola, como era el caso del tipo de vehículo o el tipo de accidente. Una posibilidad, ante esta situación, sería hacer agrupaciones en categorías con más elementos para que, así, cada clase tenga mayor número de apariciones. Por otro lado, si se realiza esta agrupación, tenemos la desventaja de que se perdería mucha información pues, al ser categorías más amplias, perderíamos la distinción entre los elementos pertenecientes a estas, dando lugar a una pérdida considerable de información y una generalización que puede resultar en pérdida de utilidad de las predicciones, a pesar de un posible mejor rendimiento.

3.6.2. Aprendizaje no supervisado

En este apartado, el objetivo será aplicar técnicas de clustering para agrupar la información en conjuntos con características similares y así entender mejor como se distribuyen tanto el tráfico como los accidentes en el espacio y en el tiempo.

Aforos

En este CSV se intenta agrupar las ubicaciones para entender cuáles tienen características similares. Las métricas que se han escogido para realizar la agrupación son la media y la varianza. La media ofrece una medida representativa

del tráfico habitual en cada estación, el volumen de tráfico en ella. La varianza informa sobre la estabilidad del tráfico, una alta variabilidad refleja picos con alto y bajo volumen.

Se han aplicado 3 métodos, k-means, DBSCAN y *clustering* jerárquico.

- K-means: el primer paso para llevar a cabo este método es elegir el número de *clusters* en que se pretende dividir la muestra. Esta elección no es aleatoria y se ha hecho mediante el método del codo. Esta técnica consiste en aplicar el algoritmo k-means con k=1, k=2 y así sucesivamente y calcular en cada k la variación total dentro de los *clusters*. Esta variación es la suma de las distancias al cuadrado de cada punto al centro de su *cluster* (centroide). El objetivo es encontrar el punto k donde la suma de las distancias al cuadrado no disminuye significativamente conforme se aumenta el número de *clusters*. A continuación se muestra el codo aplicado al CSV.

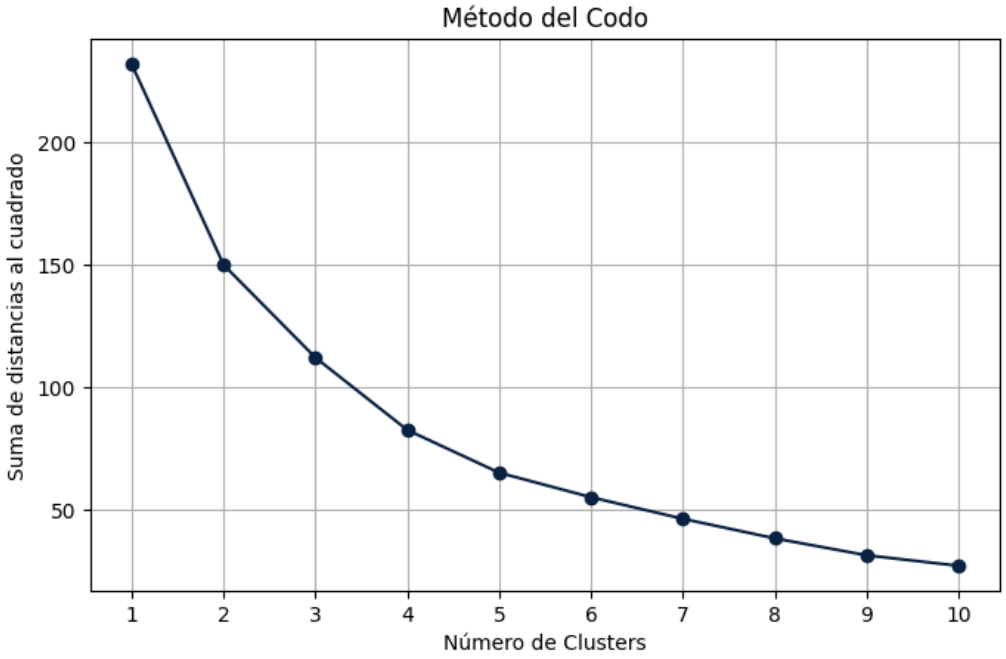

Figura 3.17: Método del codo.

Como se puede observar no hay un punto en la gráfica donde la disminución de la suma de los cuadrados se aplane significativamente, se podría elegir el 3, el 4 o el 5; por tanto, se ha probado con esos 3 valores. Para decidir cuál de los tres daba mejor resultado se ha utilizado el coeficiente de silueta promedio. Este toma valores entre -1 y 1 donde valores cercanos a -1 indican que los puntos se están asociando a *clusters* erróneos, valores cercanos a 0 que los *clusters* se están solapando y valores cercanos a 1 que los puntos están bien asignados a los correspondientes *clusters* y que estos están bien diferenciados. Tras escalar los datos y aplicar el modelo con los

3 valores de k, se han obtenido valores del coeficiente de silueta muy similares, siendo el mejor para k=5 con un valor de 0.321 y el más bajo para k=4 con un valor de 0.277. A continuación, en la Figura 3.18 se muestran en el mapa los 5 *clusters* obtenidos.

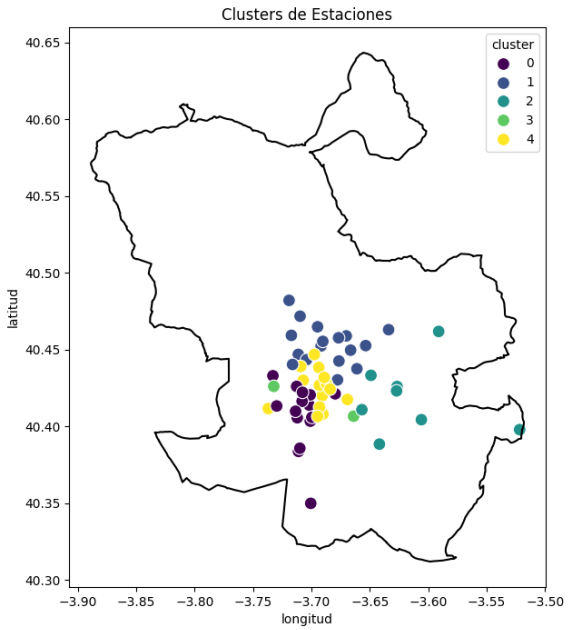

Figura 3.18: Visualización de los *clusters* obtenidos con k-means.

Para entender mejor las 5 agrupaciones y el valor medianamente bajo del coeficiente de silueta, vemos la varianza y media de los centroides de cada uno de los *clusters*, que recordemos que es el punto representante de cada uno de ellos.

Clúster	**Media**	**Varianza**
0	-0.477286	-0.537525
1	-0.281314	-0.216676
2	-0.454528	-0.329515
3	4.117412	4.714556
4	0.578688	0.407684

Cuadro 3.8: Medidas de los centroides de los clústeres.

En primer lugar, recordar que los valores son negativos y tan pequeños pues se ha aplicado la estandarización. Aunque a priori todos deberían estar entre 0 y 1, el *cluster* 3 indica que ha agrupado aquellos valores más dispersos y por ello tanto su media como su varianza tienen valores superiores a 1. Para representar los datos de forma más visual se crea la siguiente gráfica.

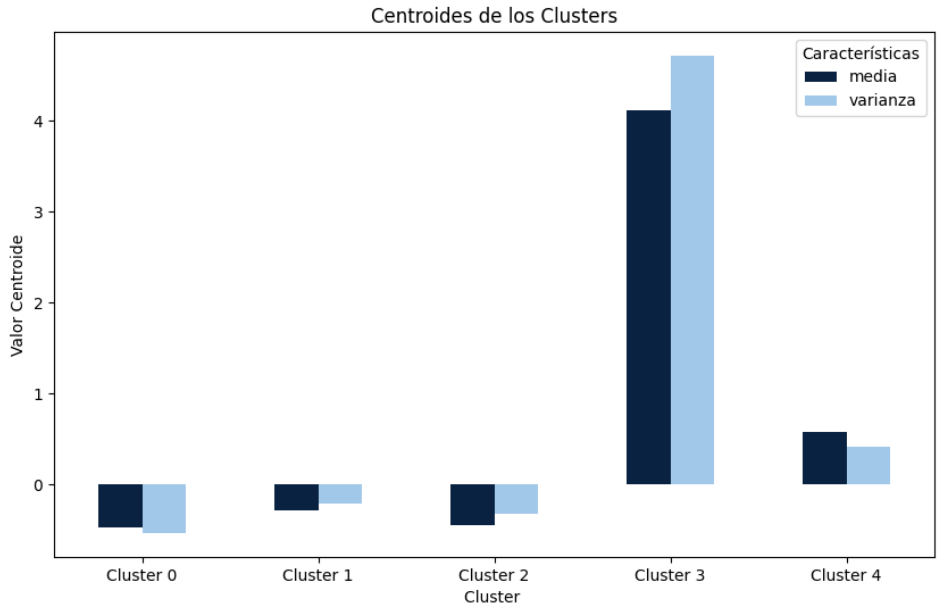

Figura 3.19: Centroides de los *clusters* para k=5.

Aquí se observa claramente que el *cluster* 0 recoge aquellas ubicaciones con bajo volumen de coches y menor varianza, es decir, zonas poco concurridas de forma constante. El *cluster* 1 recoge las zonas con menos volumen y variabilidad algo más moderada, dentro de que sigue siendo un valor bajo. El *cluster* 2 agrupa ubicaciones con muy poco volumen y poca variabilidad. El *cluster* 3 recoge claramente zonas con mucho aforo y mucha variabilidad, en este grupo se encuentran estaciones de medición como la M-30, lugares donde se explica claramente ese alto volumen que fluctúa considerablemente. Por último, el *cluster* 4 recoge ubicaciones con un volumen moderado y algo de variabilidad.

Al ver el centroide de cada *cluster* se explica el coeficiente de silueta pues se entiende que los *clusters* 0,1 y 2 corresponden a ubicaciones con características bastante similares. Si se aplica el algoritmo para k=3 se obtiene un coeficiente de silueta de 0.31 y 3 *cluster* con los siguientes centroides.

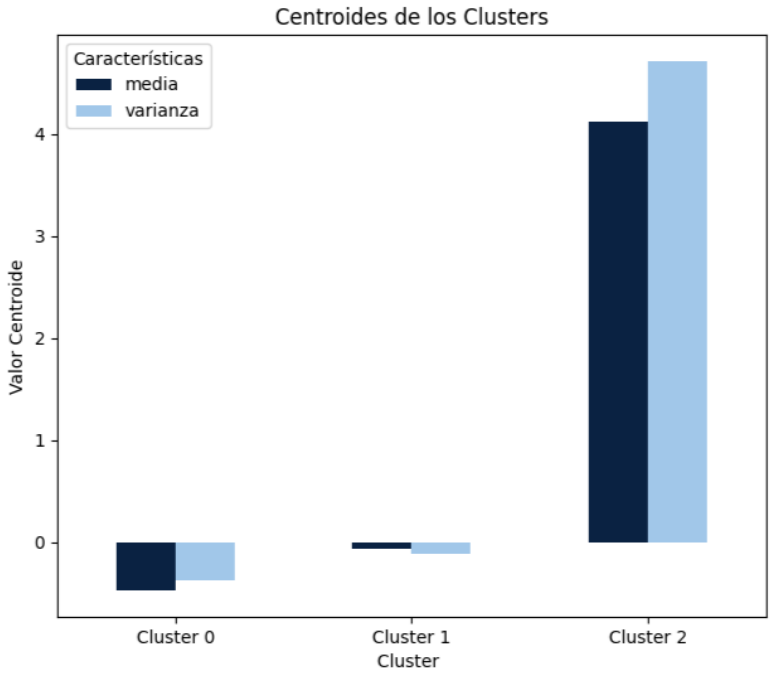

Figura 3.20: Centroides de los clústeres para k=3.

Aquí se observa que el *cluster* 2 es exactamente igual al *cluster* 3 anterior. Los *clusters* 0,1 y 4 anteriores han sido agrupados en el *cluster* 1 y el antiguo *cluster* 2 es ahora el *cluster* 0. Se entiende el bajo coeficiente de silueta pues ha unido estaciones que tienen un volumen de tráfico y varianza superior a 0 con estaciones cuyos valores están por debajo de 0.

- DBSCAN: para llevar a cabo el método de DBSCAN se han de escoger dos valores que utilizará el algoritmo, el mínimo número de elementos por *cluster* y la distancia máxima (eps) entre elementos del mismo *cluster*. El primer dato se puede elegir en función de las características del problema, para el segundo se ha utilizado la distancia al k vecino más cercano. Este método consiste en representar en una gráfica la distancia euclídea de cada punto al k vecino más cercano. Este valor de k se suele hacer que coincida con el mínimo de elementos del *cluster*, y los valores elegidos para calcular la distancia son de nuevo la media y la varianza. A continuación se muestra en la Fig. 3.21 la gráfica correspondiente a este método con 2 vecinos.

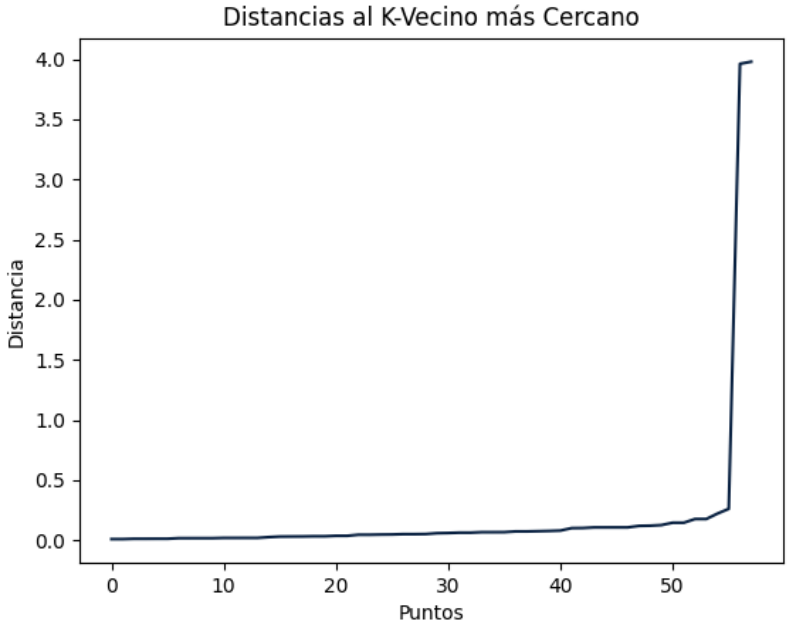

Figura 3.21: Método de 2 vecinos cercanos.

El punto que se elige como distancia máxima es aquel en el que se observa un pico. En este caso, a partir del valor 55 se ve que la distancia al segundo vecino es considerablemente grande y, por tanto, la distancia del dato que está en posición 55 a su segundo vecino más cercano es la elegida como eps, de esta forma nos aseguramos la densidad de los *clusters*. El hecho de que la distancia al k vecino más cercano se dispare a partir del dato número 55 sobre un total de 60 estaciones indica que los puntos están muy pegados. Esto puede ser un inconveniente para identificar grupos con diferencias significativas.

Al aplicar DBSCAN con estos parámetros se obtienen 4 *clusters* de 2 y 3 elementos y el resto los considera ruido (se observan en la figura como *cluster* -1). Además, el coeficiente de silueta es de -0.213.

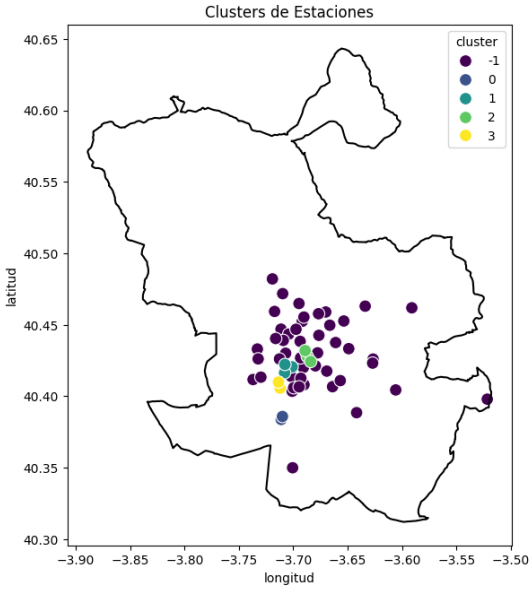

Figura 3.22: Visualización de los *cluster* obtenidos con DBSCAN.

Para entender estos resultados se ha realizado una gráfica de análisis de componentes principales (PCA). Esta técnica transforma las características de los datos combinándolas de manera que se maximice la varianza a lo largo de los ejes. A continuación se muestra la gráfica.

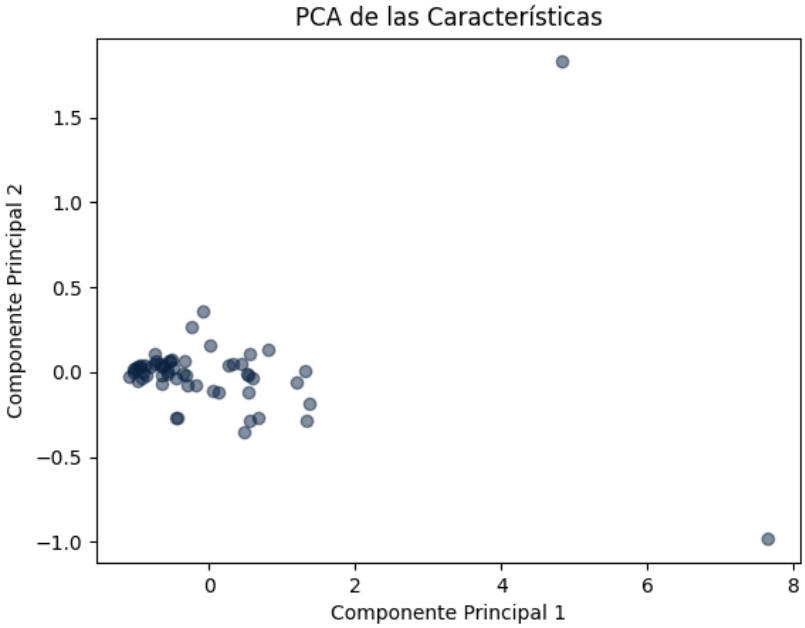

Figura 3.23: PCA

Como se observa, la mayoría de los datos se encuentran en torno al origen, pero hay dos de ellos que se encuentran considerablemente alejados. Esto indica que difieren significativamente en sus valores de media y varianza respecto a los otros datos. Por otro lado, el PCA 1 tiene una variabilidad de 0 hasta 8 (media) mientras que el PCA 2 de -1 a 1.5 (varianza) lo que muestra que la variable que tiene más impacto es la media.

El algoritmo de DBSCAN se basa en la densidad local y la conectividad. Casi la totalidad de los puntos se encuentran significativamente cerca, por tanto, cuando se elige una distancia (eps) superior a la proporcionada por el método de los vecinos, se agrupan todos los elementos en un solo *cluster* y los dos que están alejados se consideran ruido; cuando se elige un eps inferior o igual se crean *clusters* unipuntuales o con dos elementos, y el resto se considera ruido.

K-means es un método menos sensible a la distribución de los puntos en comparación con DBSCAN, dado que este se basa en la densidad de los puntos, esto justifica la baja calidad de los resultados.

- *Clustering* jerárquico: para aplicar este método se ha de elegir también el número de *clusters* que se desean. Esta elección se hace bajo el mismo criterio que en k-means, con el método del codo, y dado que en el anterior caso se escogieron 5 *clusters* se ha probado primeramente con este número. Se han obtenido las siguientes agrupaciones con un coeficiente de silueta de 0.37.

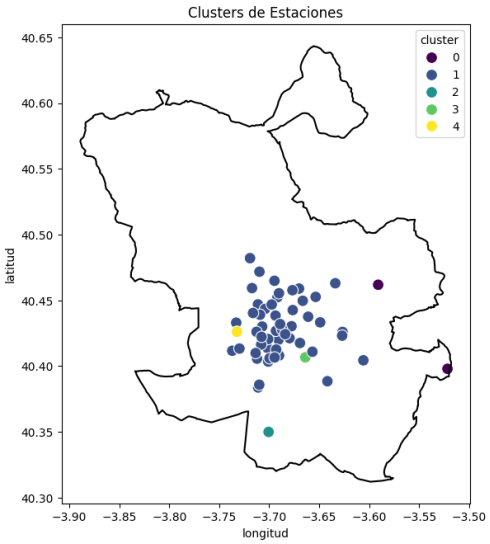

Figura 3.24: Visualización de los *clusters* obtenidos con método jerárquico para k=5.

A continuación, en la Fig. 3.25 se muestra el dendrograma, donde se puede observar el orden en el que el algoritmo ha ido haciendo agrupaciones hasta obtener un solo *cluster*. Para obtener esos 5 *clusters* hay que hacer una poda que se encuentra señalada con una línea horizontal roja entre los niveles 3 y 4, obteniendo así 5 líneas verticales que no se unen, cada una representa un cluster. Sombreados en cajas, se distinguen los elementos que forman cada uno de los *cluster*.

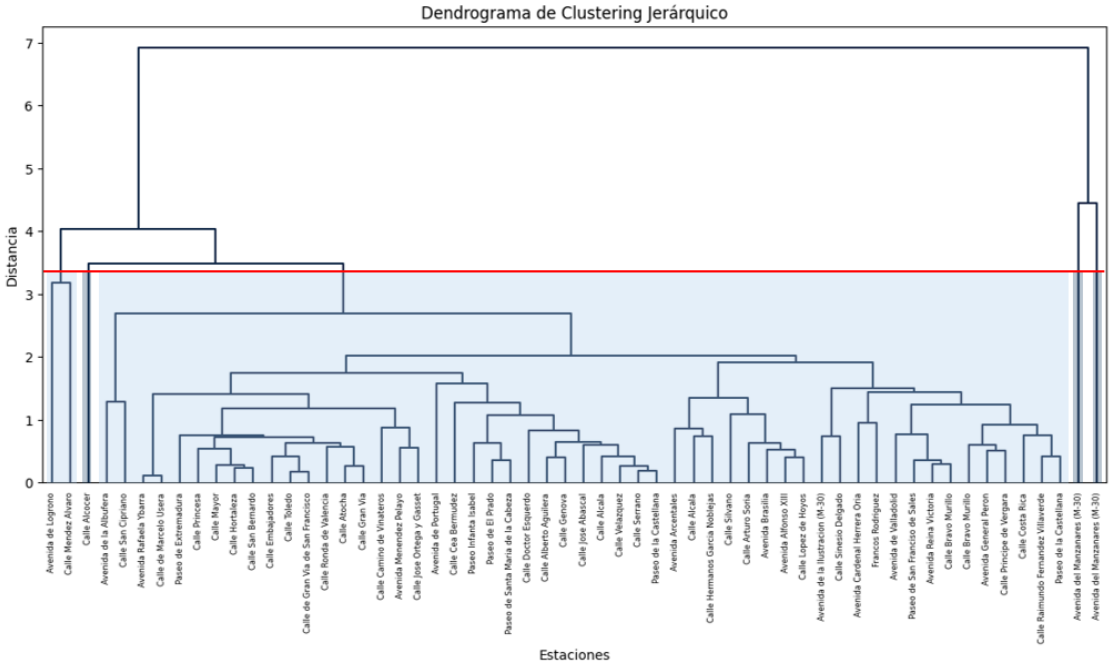

Figura 3.25: Visualización del dendrograma.

Por último, en este algoritmo no existen los centroides, pero para hacerse una idea de los elementos de cada *cluster* se ha calculado la media de las medias y varianzas de los elementos de cada conjunto, para así obtener un representante de cada uno de ellos.

Cluster	**Media**	**Varianza**
0	-0.730331	-0.825110
1	0.581823	-0.699909
2	-0.816976	-0.918567
3	1.837240	1.221793
4	0.291890	1.221793

Cuadro 3.9: Medidas de los representantes de cada *cluster*.

Igual que en el ejemplo anterior, lo representamos con una gráfica.

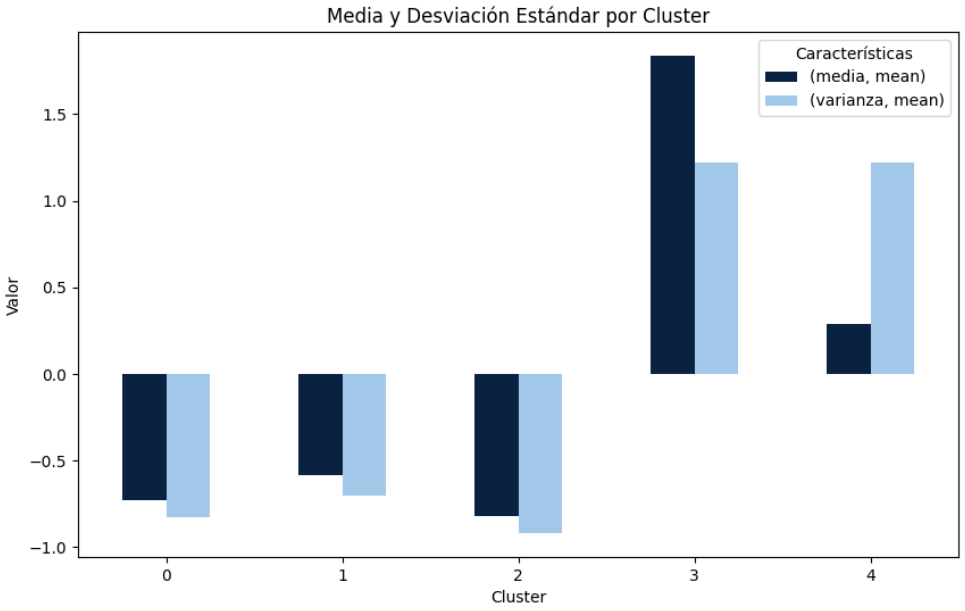

Figura 3.26: Representantes de los *cluster* para k=5.

Como podemos observar, los tres primeros *clusters* tienen medidas muy similares. Los 3 se caracterizan por tener bajo volumen y muy baja varianza. Por otro lado, el *cluster* 3 destaca por tener un alto volumen y varianza y el *cluster* 4 tiene un volumen moderado y una varianza elevada. Dados estos resultados y la similitud entre las tres primeras agrupaciones, se prueba el algoritmo con 3 y 4 *cluster* en lugar de 5, para intentar agrupar los 3 primeros.

Para el caso de k=4 se obtiene un coeficiente de silueta de 0.48 que refleja una gran mejoría y, a continuación, se muestran los representantes de esos 4 *clusters* en la Fig. 3.28 y los elementos de cada uno en el mapa en la Fig. 3.27.

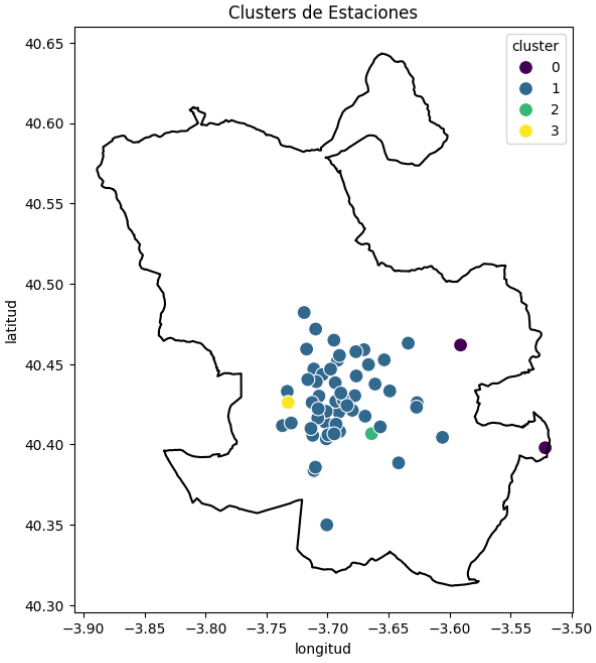

Figura 3.27: Visualización de los *clusters* obtenidos con método jerárquico para k=4.

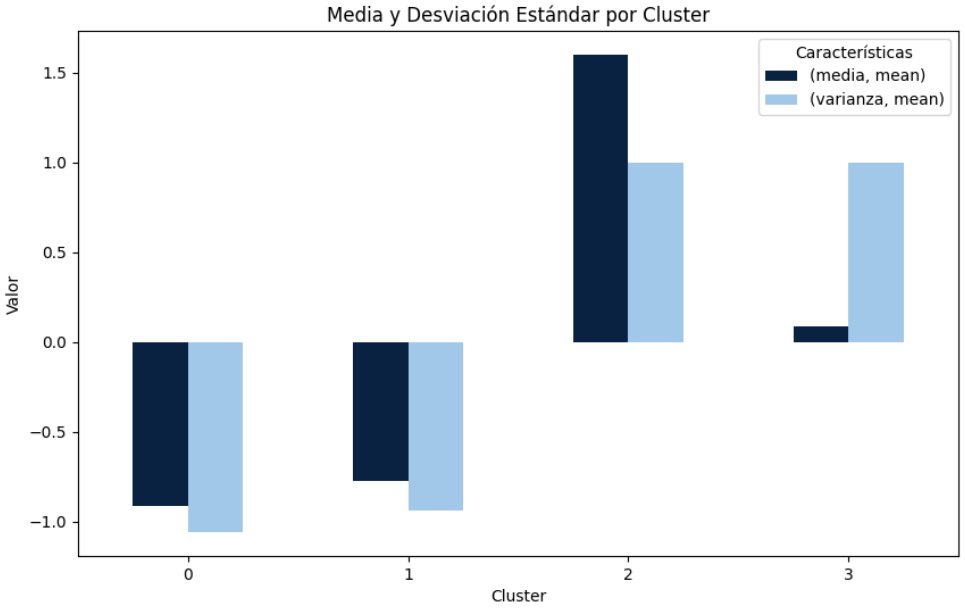

Figura 3.28: Representantes de los *clusters* para k=4

Como se puede observar, el elemento del anterior *cluster* 3 se ha añadido al *cluster* 2 y el resto han permanecido intactos.

Para el caso de k=3 se obtiene un coeficiente de silueta de 0.62, pero nos encontramos con 3 *clusters*, dos unipuntuales y otro con el resto del conjunto de datos.

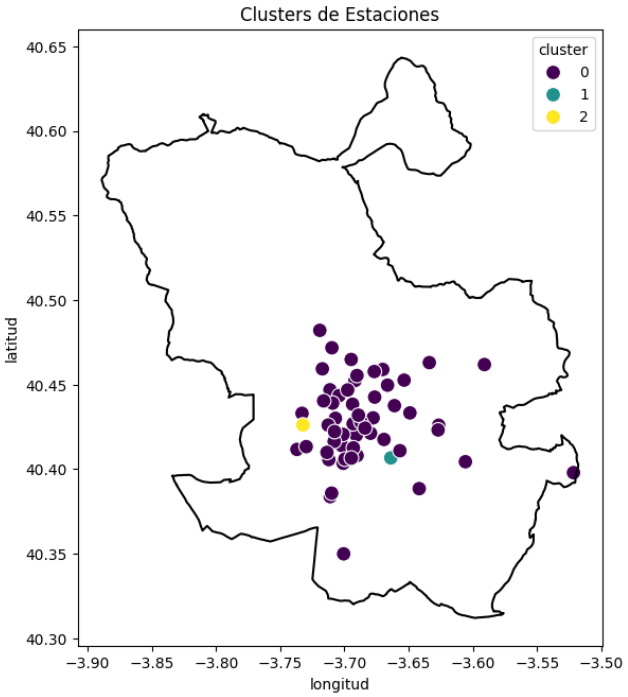

Figura 3.29: Visualización de los *clusters* obtenidos con método jerárquico para k=3.

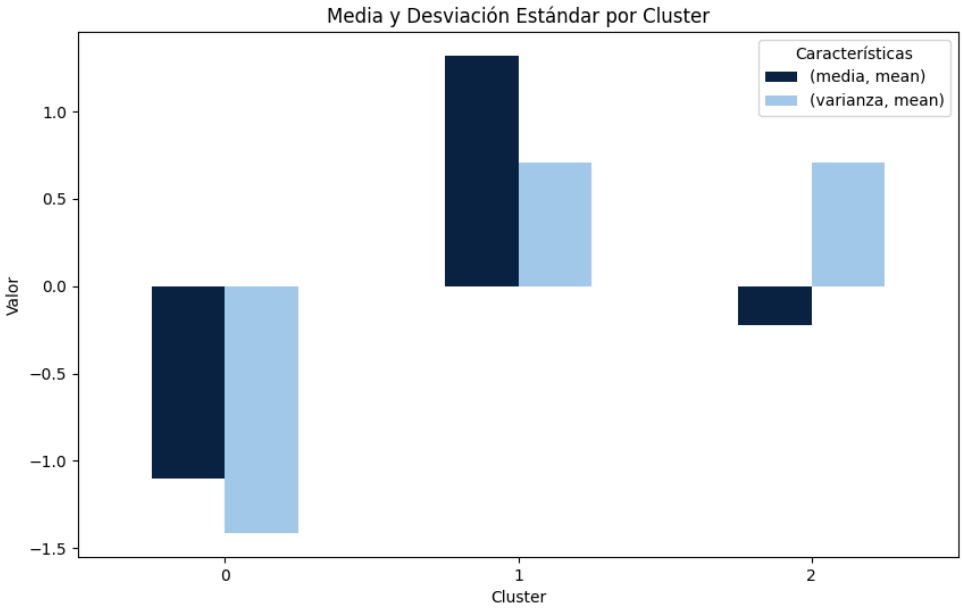

Figura 3.30: Representantes de los *clusters* para k=3

En este caso se observa que una mejoría del coeficiente de silueta no implica necesariamente mejores resultados en cuanto a su valor. Aunque para k=3 los *clusters* estén mejor definidos y separados, tener agrupaciones unipuntuales puede ser no tan útil en la práctica. Con un coeficiente bajo, si se entiende bien la información puede resultar de gran utilidad. Aquellas estaciones con mayor volumen y menor varianza son aquellas con un aforo alto constante y necesitarán alguna medida de transporte para desconges-tionarla (como nuevas carreteras) y aquellas con alta varianza y volumen en un determinado momento, medidas como el aumento de la frecuencia del transporte público en esas horas pico. Además, las zonas donde hay estaciones con bajo volumen y baja varianza podrían ser utilizados como modelo para la construcción de nuevas áreas en las que se busque estabilidad en el flujo de tráfico.

De este CSV también podemos visualizar las horas del día y los meses más concurridos. A continuación se muestran dos histogramas, uno con la media de aforo por hora (Fig. 3.31) del día y otro con la media de aforo por mes (Fig. 3.32).

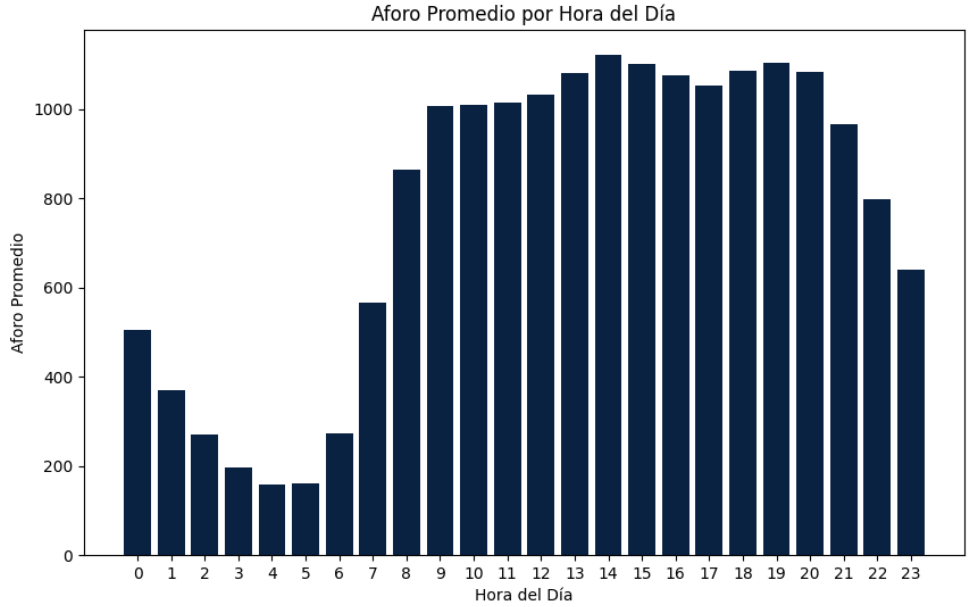

Figura 3.31: Media de aforo por hora

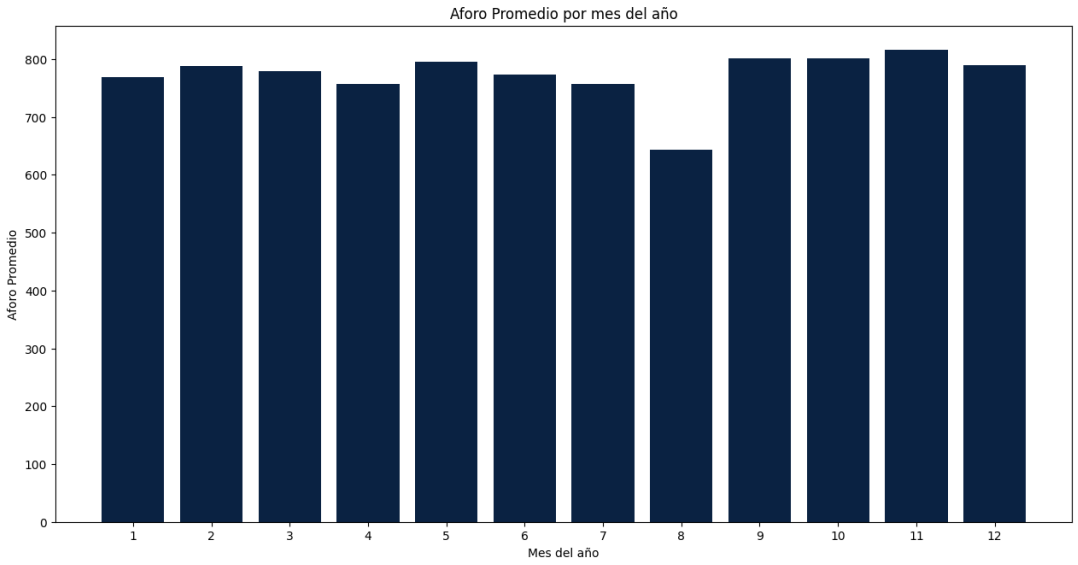

Figura 3.32: Media de aforo mensuales

Como se puede observar, las horas del día más tranquilas son las de la madrugada. A las 7 comienza una subida hasta alcanzar el pico a las 14, que se mantiene hasta las 20 donde empieza a bajar. Respecto a los meses, ninguno destaca notablemente por un aforo superior, sin embargo, en agosto se observa que disminuye considerablemente el tráfico pues es el mes en el que Madrid se vacía. Del mismo modo, pero no de forma tan marcada, sucede en julio y enero, meses también de carácter vacacional.

Accidentes

En este archivo, en primer lugar se va a agrupar los accidentes por distritos, para así entender cuáles son aquellos con mayor incidencia. Una vez hecha la agrupación, se va a aplicar el algoritmo k-means. En este caso, la medida que se va a utilizar como distancia para agrupar los distritos es el número de accidentes en cada uno de ellos. A continuación vamos a ver los distritos en función de la cantidad de accidentes que ocurren en cada uno (Fig. 3.33). Podemos observar que el distrito con más accidentes es el barrio de Salamanca y el que menos Vicálvaro. También se ha creado un mapa de calor para visualizar esta información en el territorio (Fig. 3.34).

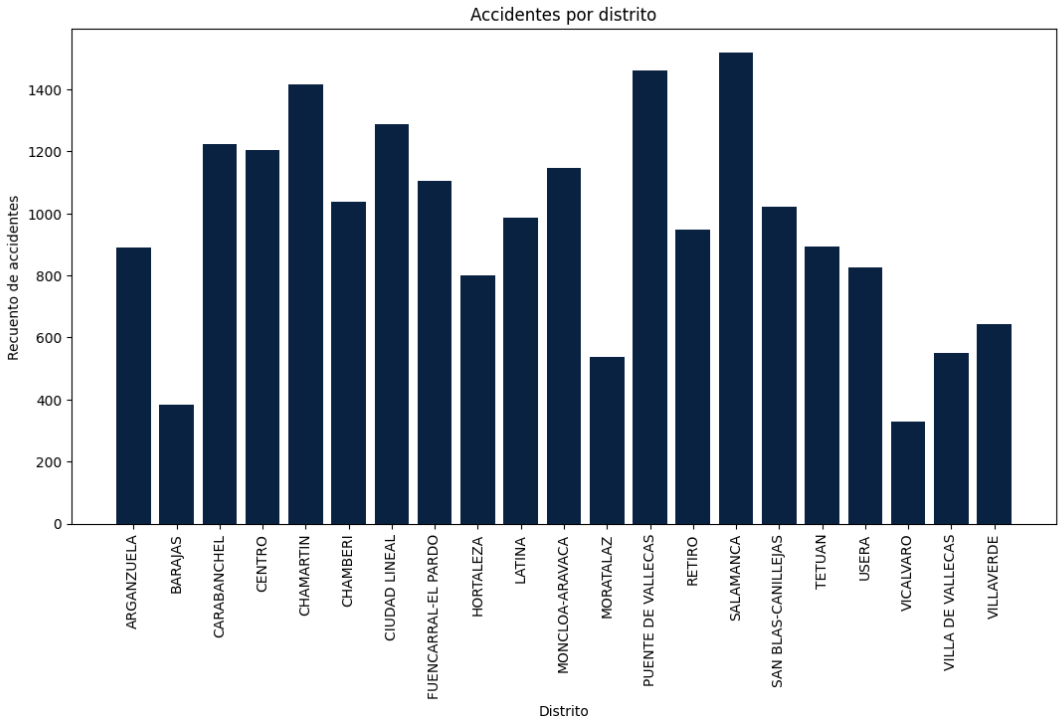

Figura 3.33: Recuento de accidentes por distrito

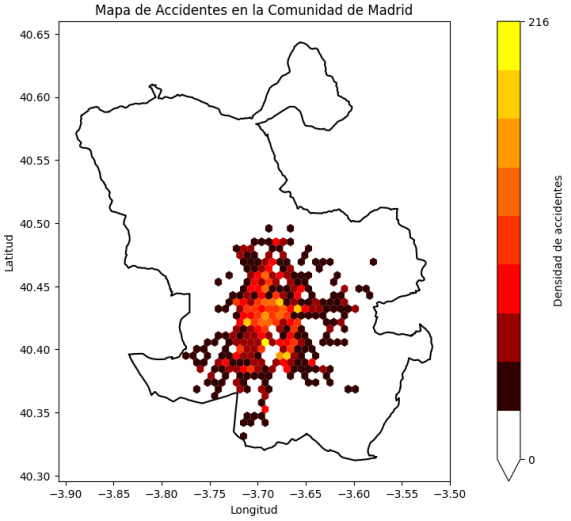

Figura 3.34: Mapa de calor de accidentes

Una vez se tiene una idea inicial sobre la distribución de los datos, se aplica el método del codo para determinar el número de *clusters* en los que hacer la división. En la gráfica del codo, se puede observar que los valores de k para los que la distancia a penas disminuye son k=3 y k=4, por tanto, se aplicará k-means para ambos valores.

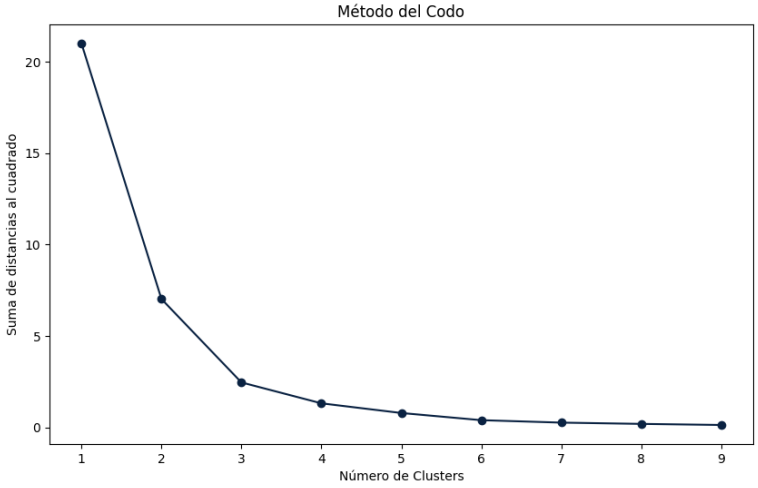

Figura 3.35: Método del codo

Para k=3 se obtiene un coeficiente de silueta de 0.568 y para k=4 toma un valor de 0.594. Ambos resultados son moderadamente altos, indicando una separación razonable y cohesión dentro de los *clusters*. Se muestran a continuación los *clusters* obtenidos para el mejor resultado, k=4.

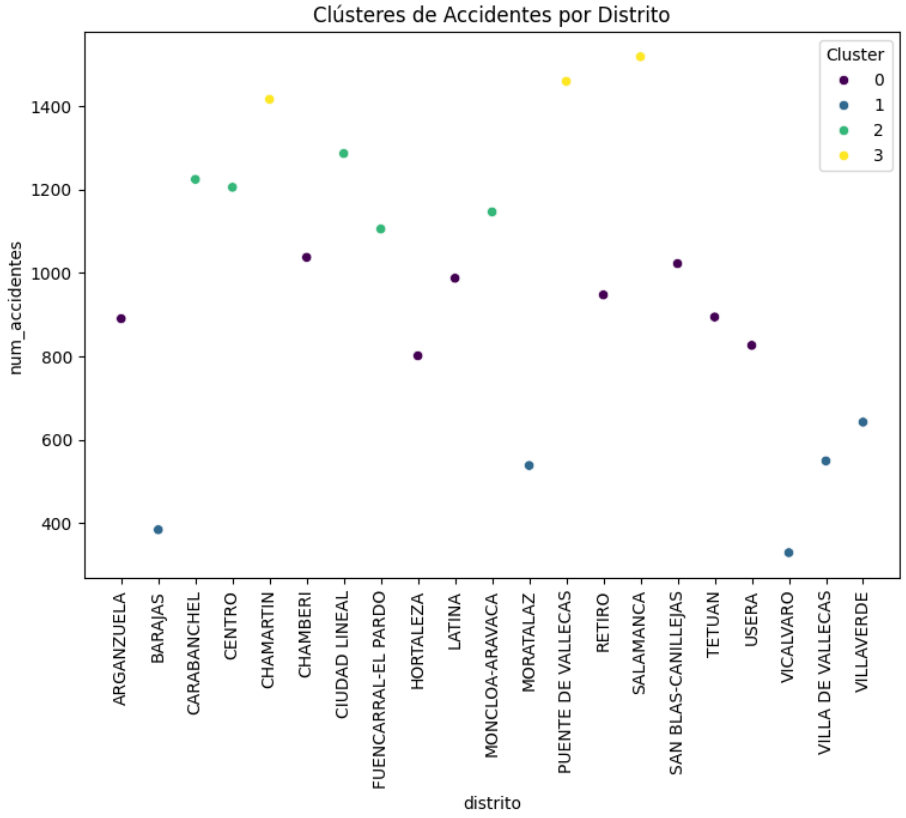

Figura 3.36: Visualización de los *clusters* obtenidos para k=4

Se han calculado los centroides de cada *cluster* para entender qué elementos forman parte de cada uno de ellos.

Cluster	**N.º accidentes**
0	926
1	489
2	1.193
3	1.464

Cuadro 3.10: Medidas de los centroides de los *cluster*s.

De esta forma vemos que el *cluster* 1 contiene las ubicaciones con menor número de accidentes, los *cluster*s 0 y 2 tienen un número de accidentes moderado y el *cluster* 3 es aquel con mayor incidencia.

A pesar de que los resultados obtenidos son razonablemente buenos, se va a

aplicar el *clustering* jerárquico para entender las relaciones y jerarquías entre las agrupaciones. A continuación se muestra el dendrograma correspondiente a este método (Fig. 3.37), con el cual se ha obtenido un coeficiente de silueta de 0.593.

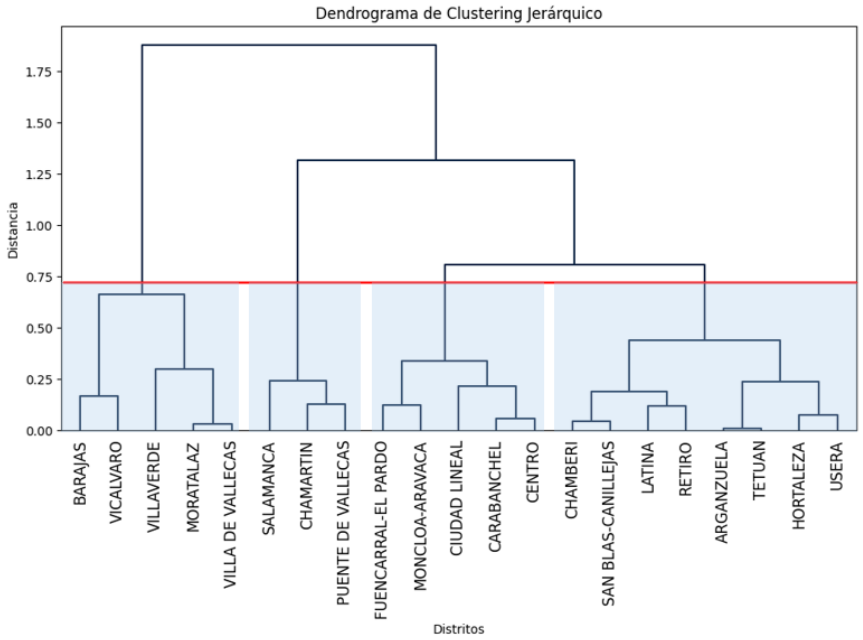

Figura 3.37: Visualización del dendrograma.

Con una línea horizontal roja queda señalado el nivel donde se ha realizado la poda, resultando así cuatro líneas verticales que representan cada *cluster*. Si se recorre el dendrograma de izquierda a derecha, en primer lugar nos aparece el *cluster* 1, a continuación el 3, seguidamente el 2 y por último el 0; viendo claramente los elementos de cada una de las agrupaciones.

Una vez identificadas las ubicaciones más problemáticas se va a realizar un análisis temporal, viendo qué meses y a qué horas hay más frecuencia de accidentes para así, poder contrastarlo con los resultados vistos en el CSV de aforos. Para ello se va a utilizar k-means y se va a hacer una agrupación de los accidentes en función del mes y la hora. De nuevo, el primer paso es realizar el diagrama del codo, en el que podemos observar que los mejores valores son k=3 y k=4.

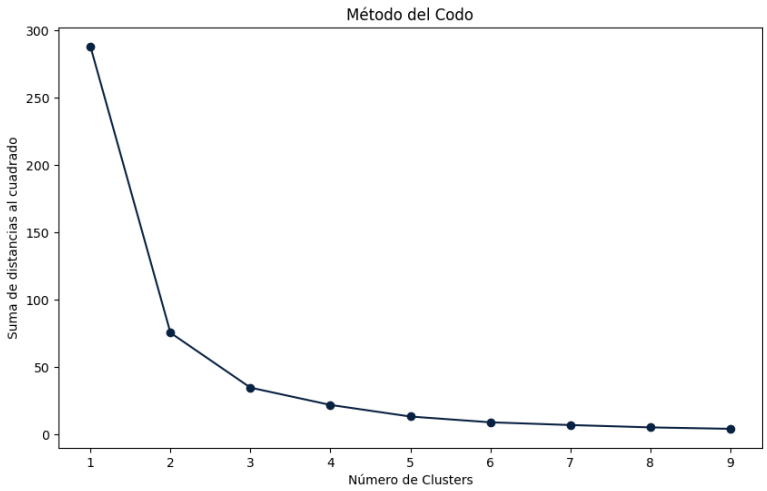

Figura 3.38: Método del codo.

Se ha aplicado el algoritmo para ambos valores y se han obtenido resultados muy similares con un coeficiente de silueta de 0.594 para k=3 y 0.59 para k=4. Por tanto, se muestran los *clusters* obtenidos para k=3.

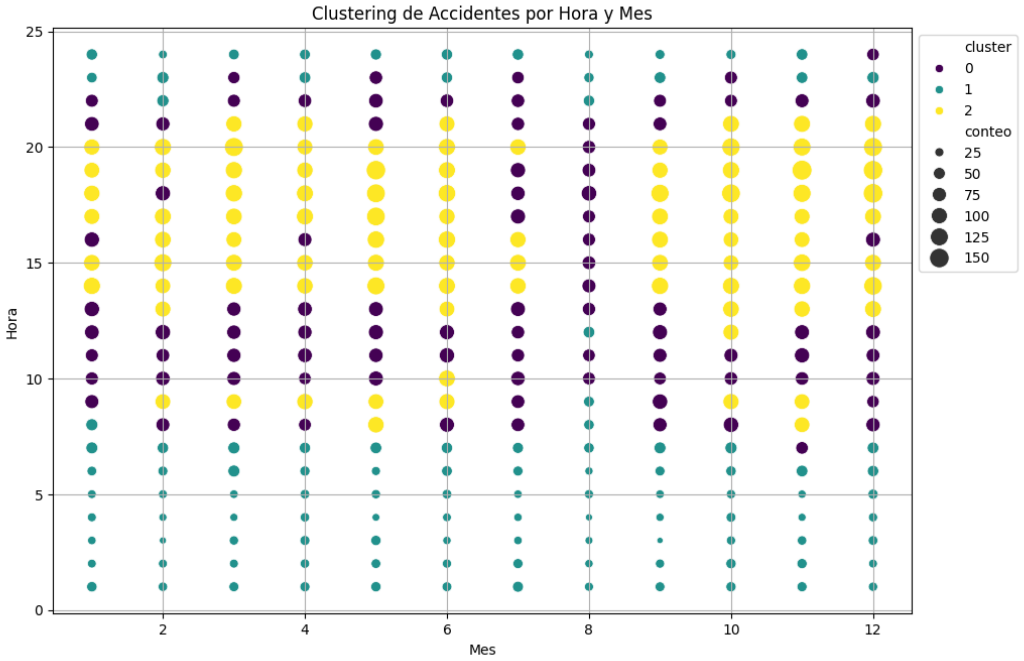

Figura 3.39: Visualización *clusters* obtenidos para k=3.

En la gráfica se ven los 3 *clusters* en distintos colores y distintos tamaños de círculos representando el número de accidentes, siendo 25 o menor el círculo más pequeño y 150 o superior el círculo más grande. Como se puede observar, el *cluster* 1 es aquel con menos accidentes, se reparte por todos los meses del año, destacando el mes de agosto, y se sitúan las horas del día desde las 23 hasta las 7 de la mañana aproximadamente. El *cluster* 0 es el siguiente con más accidentes, se reparte en todos los meses y se sitúa en las horas de la mañana desde las 7 hasta las 13 aproximadamente y las últimas horas de la tarde, de 21 a 23. El último cluster, el 2, se corresponde con las horas con mayor accidentes y se extiende en las horas centrales del día, de 14 a 20.

Cabe destacar que el mes de agosto no tiene ninguna hora perteneciente a este *cluster* lo que señala la baja cantidad de accidentes en este periodo, por otro lado, no hay ningún mes que destaque notablemente en el caso contrario, es decir, con un claro repunte de accidentes. También cabe señalar que los meses de enero y julio tienen algo menos de accidentes pues se observa menor presencia de puntos amarillos, esto se entiende dado que son meses típicamente vacacionales. Esta clasificación confirma la información ya obtenida gracias al CSV de aforos.

Capítulo 4

Resultados y conclusiones

4.1. Resultados

En este capítulo se recogen los mejores resultados obtenidos tras realizar toda la etapa de desarrollo. Recordemos que el objetivo final es encontrar patrones y tendencias en el tráfico y accidentes para poder mejorar el flujo del tráfico de la ciudad y disminuir la cantidad de accidentes o, en su defecto, minimizar sus consecuencias.

4.1.1. Aprendizaje supervisado

Gracias al aprendizaje supervisado hemos obtenido dos modelos predictivos clave:

- Modelo para detectar consumo de drogas o alcohol en accidentes: este modelo tiene una precisión del 78 %. Puede ser de gran utilidad para las autoridades al dirigirse al lugar del accidente, ya que les permite prever el estado de las personas involucradas y la probabilidad de que no estén en plenas facultades.
- Modelo para determinar la necesidad de asistencia sanitaria en accidentes: este modelo tiene una precisión del 83 %. Es fundamental para salvar vidas y optimizar recursos, especialmente en situaciones con múltiples accidentes o disponibilidad limitada de ambulancias. Permite evaluar rápidamente si se requiere asistencia sanitaria.

4.1.2. Aprendizaje no supervisado

En el aprendizaje no supervisado, el objetivo era obtener información geográfica y temporal sobre la afluencia de vehículos y la incidencia de accidentes. Se han obtenido los siguientes resultados:

- Los distintos puntos de medición de aforo se han clasificado según la media y varianza de vehículos que pasan por cada punto. Esto permite identificar patrones de tráfico en diferentes áreas de la ciudad. Se han creado 5 grupos

o *clusters*, el primero de ellos recoge aquellas ubicaciones con bajo volumen de coches y menor varianza, es decir, zonas poco concurridas de forma constante. El *cluster* 1 recoge las zonas con menos volumen y variabilidad algo más moderada, dentro de que sigue siendo un valor bajo. El *cluster* 2 agrupa ubicaciones con muy poco volumen y poca variabilidad. El *cluster* 3 recoge claramente zonas con mucho aforo y mucha variabilidad, en este grupo se encuentran estaciones de medición como la M-30, lugares donde se explica claramente ese alto volumen que fluctúa considerablemente. Por último, el *cluster* 4 recoge ubicaciones con un volumen moderado y algo de variabilidad. A continuación se muestran las ubicaciones en las que se encuentran los puntos de medición de cada *cluster*.

ID	**Dirección**	***Cluster***
1	Paseo de la Castellana	4
2	Calle Princesa	0
3	Calle Doctor Esquerdo	4
4	Paseo de San Francisco de Sales	1
5	Paseo de Santa María de la Cabeza	4
6	Calle Arturo Soria	1
7	Avenida de Portugal	4
8	Calle Gran Via	0
9	Calle Atocha	0
10	Avenida de Oporto	0
11	Avenida del Manzanares (M-30)	3
12	Calle Jose Abascal	4
13	Calle Génova	4
14	Calle Jose Ortega y Gasset	1
15	Avenida Reina Victoria	1
16	Calle Alberto Aguilera	4
17	Calle Cea Bermúdez	4
18	Avenida Menéndez Pelayo	0
19	Calle Bravo Murillo	1
20	Avenida del Manzanares (M-30)	3
21	Calle Príncipe de Vergara	1
22	Calle Ronda de Valencia	0
23	Paseo de El Prado	4
24	Calle de Gran Via de San Francisco	0
25	Calle Hortaleza	4
26	Calle San Bernardo	0
27	Calle Alcalá	4
28	Calle Méndez Álvaro	2
29	Paseo Infanta Isabel	4
30	Calle Embajadores	0
31	Francos Rodríguez	1
32	Calle Toledo	0
33	Calle Sinesio Delgado	1

ID	Dirección	*Cluster*
34	Calle Mayor	0
36	Paseo de la Castellana	1
37	Calle Costa Rica	1
38	Avenida Cardenal Herrera Oria	1
39	Avenida de la Ilustración (M-30)	1
40	Calle Raimundo Fernández Villaverde	4
41	Calle Bravo Murillo	1
42	Avenida General Perón	1
43	Paseo de Extremadura	0
44	Calle Serrano	4
45	Calle Velázquez	4
46	Avenida de la Albufera	2
47	Calle Alcalá	2
48	Calle Hermanos García Noblejas	2
49	Avenida de Valladolid	0
50	Calle López de Hoyos	1
51	Avenida Alfonso XIII	1
52	Avenida Brasilia	1
53	Calle de Marcelo Usera	0
54	Avenida Rafaela Ybarra	0
55	Calle Alcocer	0
56	Avenida Arcentales	2
57	Calle Silvano	1
58	Avenida de Logroño	2
59	Calle San Cipriano	2
60	Calle Camino de Vinateros	2

Cuadro 4.1: Elementos pertenecientes a cada *cluster*.

- Agrupación de distritos según la incidencia de accidentes: Los distritos se han agrupado en cuatro *clusters* en función de la incidencia de accidentes. El *cluster* 1 contiene las ubicaciones con menor número de accidentes, Los *clusters* 0 y 2 tienen un número de accidentes moderado y, por último, el *cluster* 3 es aquel con mayor incidencia. A continuación se muestran los distritos y el *cluster* al que pertenecen.

Distrito	*Cluster*
ARGANZUELA	3
BARAJAS	0
BARRIO DE SALAMANCA	1
CARABANCHEL	2
CENTRO	2
CHAMARTÍN	1
CHAMBERÍ	3

Distrito	***Cluster***
CIUDAD LINEAL	2
FUENCARRAL EL PARDO	2
HORTALEZA	3
LATINA	3
MONCLOA-ARAVACA	2
MORATALAZ	0
PUENTE DE VALLECAS	1
RETIRO	3
SAN BLAS-CANILLEJAS	3
TETUÁN	3
USERA	3
VICÁLVARO	0
VILLAVERDE	0

Cuadro 4.2: Elementos pertenecientes a cada *cluster*.

- Identificación de periodos con mayor incidencia de accidentes: los resultados han identificado los meses y horas con mayor riesgo. Entre ellos se encuentra que el mes de agosto es aquel con menor incidencia de accidentes y menor flujo de tráfico, a continuación, con unos niveles algo superiores al mes de agosto, pero aún bajos, se encuentran los meses de enero y julio. Por otro lado, el resto de meses del año se encuentran en un nivel superior, destacando ligeramente por encima de todos septiembre. Respecto a las horas del día, el rango de horas más tranquilas abarca desde las 11 de la noche hasta las 6 de la mañana, comenzando un gran incremento a las 7, que alcanza su pico de tráfico a las 15, manteniéndose hasta las 20, donde comienza a decrecer de nuevo.

4.2. Conclusiones

Los resultados obtenidos en este proyecto proporcionan herramientas valiosas para mejorar la movilidad urbana y la seguridad vial en Madrid. Algunas de las conclusiones clave incluyen:

- Utilidad práctica de los modelos predictivos: los modelos desarrollados para detectar consumo de sustancias y la necesidad de asistencia sanitaria pueden mejorar significativamente la respuesta de emergencia y la planificación de recursos.
- Mejora de la infraestructura vial y transporte público: los distritos pertenecientes a la agrupación con menor número de accidentes pueden ser tomados como modelo de cara a la creación de nuevos barrios, que puedan tomar la estructura de sus calles y los servicios de transporte público como referencia, imitando así también su bajo índice de accidentes. Por otro lado, los distritos con más accidentes pueden convertirse en el foco de las autoridades para llevar a cabo una mejora integral de la infraestructura

vial. En función de las características exactas de cada punto, se podría reforzar la señalización, mejorar el alumbrado público e incluso reconfigurar los cruces problemáticos para reducir el riesgo de accidentes. Asimismo, si la mejora de las infraestructuras viales no es factible, se podría intentar solventar la situación mejorando el transporte público de la zona, invitando así a los ciudadanos que transitan por ella a utilizarlo y, disminuir así el número de vehículos que circulan por la zona.

- Planificación basada en datos: la identificación de periodos de alto riesgo permite a las autoridades tomar medidas proactivas, como campañas de concienciación y mejoras en el sistema de transporte público en momentos críticos.

El uso de técnicas avanzadas de análisis de datos ha demostrado ser efectivo para manejar grandes volúmenes de datos y extraer información útil para la toma de decisiones informadas.

En resumen, este proyecto ha demostrado cómo el análisis de datos y el aprendizaje automático pueden contribuir de manera significativa a mejorar la movilidad y seguridad en una gran ciudad. Las herramientas y modelos desarrollados tienen el potencial de ser implementados y utilizados por las autoridades para hacer de Madrid una ciudad más segura y eficiente en términos de transporte.

4.3. Líneas futuras

Se encuentran varias líneas futuras que podrían tomarse a corto plazo. En primer lugar, llevar a cabo una expansión de la base de datos. Por un lado, sería interesante añadir información sobre días festivos, fines de semana y eventos especiales como manifestaciones y obras de larga duración. Esta información permitiría tener una mejor comprensión del tráfico y accidentes, pues en caso de eventos especiales como un puente, se entiende el aumento de accidentes y del flujo de vehículos. En caso de obra se podría dar la situación de una redirección del tráfico a otras zonas. De este modo, se podría tener una visión más exacta de la situación.

Por otro lado, respecto a los aforos, sería interesante realizar una propuesta para colocar más medidores en la Comunidad de Madrid con el fin de poder tener una visión más precisa del volumen de vehículos que se mueven por el territorio todos los días. Existe una gran afluencia de vehículos que se desplazan desde la periferia a través de carreteras como la A-1, A-6 o M-40, que influyen de manera determinante en el tránsito y, por tanto, su monitorización podría aportar información muy valiosa. También, podría aumentarse la frecuencia con la que se hace el recuento del número de vehículos, y tener registros cada media hora, en lugar de cada hora.

En cuanto al análisis, se podría realizar el *clustering* de este CSV con nuevas variables, que sean capaces de describir el estado del tráfico de manera más precisa y complementen a las que ya se tienen (media y varianza).

También, se podrían explorar algoritmos más complejos como redes neuronales

profundas (*deep learning*) para mejorar la precisión de las predicciones. Elaborar predicciones es una ardua tarea y más cuando se tienen datos desbalanceados, por lo tanto, sería interesante emplear más horas y nuevas técnicas en esta tarea con el fin de obtener modelos más precisos. Además, se podrían implementar técnicas de validación cruzada para evaluar la robustez de los modelos en diferentes subconjuntos de datos para garantizar su generalización.

En una visión más a largo plazo, se encuentran las siguientes utilidades del proyecto.

- Simulaciones de Tráfico: desarrollar y utilizar simulaciones que puedan modelar escenarios de tráfico complejos y sus interacciones. Se podrían llevar a cabo predicciones con series temporales para anticiparse a grandes flujos de tráfico.
- Planificación urbana: crear herramientas basadas en los modelos predictivos que resulten de utilidad a planificadores urbanos, para diseñar redes de carreteras y transporte público lo más óptimas posible.
- Herramientas educativas: gracias a la identificación de los rangos de edades y tipos de vehículos más afectados por los accidentes, se podrían llevar a cabo campañas de concienciación ciudadana orientadas a esas edades o, a los propietarios y futuros propietarios de esos tipos de vehículos.
- Análisis espacial: llevar a cabo un análisis espacial más riguroso que proporcione una buena segmentación de los distintos distritos madrileños e identifique las zonas más transitadas. El fin sería utilizar estos datos para elaborar una propuesta de mejora del transporte público personalizada a cada zona, contribuyendo así a una planificación urbana más eficiente y mayor seguridad para la ciudadanía.

  Por otro lado, aquellas zonas con redes de comunicación más eficientes podrían servir de referencia para la elaboración de estas propuestas de mejora.

En conclusión, este proyecto ofrece una base sólida para el análisis y la mejora del tráfico y la seguridad vial en Madrid. Las líneas futuras propuestas tienen el potencial de expandir y profundizar en el conocimiento obtenido, proporcionando herramientas prácticas y avanzadas para una planificación urbana más eficiente y segura. Con la implementación de estas mejoras, se espera que la movilidad en Madrid no solo sea más fluida y predecible, sino también más segura para todos sus ciudadanos

Capítulo 5

Análisis de impacto

La seguridad vial es una cuestión que afecta a todas las ciudades del mundo y un componente fundamental para lograr un desarrollo sostenible en ellas. Según los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, el ODS 3 que promueve la "salud y bienestar", y el ODS 11, que aboga por "ciudades y comunidades sostenibles", es fundamental promover el bienestar ciudadano, su seguridad y la sostenibilidad de las ciudades. El presente proyecto aborda los problemas de tráfico y accidentes en la Comunidad de Madrid, ofreciendo un análisis y propuestas de mejora, así como una metodología replicable para la continua actualización de este análisis.

Gracias a las técnicas utilizadas se han encontrado patrones temporales y tendencias que ayudan a entender las posibles causas de accidentes de tráfico, las condiciones típicas en las que estos se dan, las zonas con mayor incidencia y los rangos de edades más afectados por estos. A raíz de estas se proponen nuevas mejoras en el sistema de carreteras, ubicaciones que pueden ser utilizadas como referencia para replicar su estructura vial por su buen funcionamiento y eficiencia, y políticas de concienciación ciudadana, escogiendo de forma fundamentada la población objetivo de estas y el mensaje a transmitir.

Madrid es una ciudad con un crecimiento demográfico constante y, por tanto, sufre grandes aglomeraciones y movimiento intenso diario de vehículos. Para asegurar el desarrollo de la ciudad, es esencial tener carreteras seguras, libres de tráfico y embotellamientos, además de desarrollar y actualizar el sistema de transporte público en función de las necesidades demográficas, capaz de adaptarse al crecimiento de la ciudad. Todo ello contribuirá directamente en la calidad del aire, reduciendo la contaminación y mejorando la calidad de vida de sus habitantes.

Asimismo, es importante destacar que el centro de Madrid es típicamente conocido por los pronunciados desniveles del suelo, característica que limita el uso de la bicicleta en comparación con otras capitales europeas. Este hecho resalta la importancia de llevar a cabo un análisis exhaustivo de la movilidad urbana que favorezca las comunicaciones y disminuya los tiempos de desplazamiento.

Con este enfoque, el proyecto no solo busca mitigar los problemas de tráfico

y reducir la incidencia de accidentes, sino también contribuir a los esfuerzos globales para lograr las metas establecidas en los ODS 3 y 11, reforzando el compromiso con la salud, el bienestar y la sostenibilidad urbana.

Bibliografía

[1] MinnaLearn. (2022) Introducción al big data. [Online]. Available: https://courses.minnalearn.com/es/courses/digital-revolution/big-data-and-beyond/introduction-to-big-data/

[2] PowerData. (2022) Big data: ¿en qué consiste? su importancia, desafíos y gobernabilida. [Online]. Available: https://www.powerdata.es/big-data

[3] TechTarget. (2021) Garbage in, garbage out (gigo). [Online]. Available: https://www.techtarget.com/searchsoftwarequality/definition/garbage-in-garbage-out

[4] Teradata. (2021) What are the 5 v's of big data? [Online]. Available: https://www.teradata.com/glossary/what-are-the-5-v-s-of-big-data

[5] QuestionPro. (2021) Ciclo de vida de los datos: Qué es y qué etapas tiene. [Online]. Available: https://www.questionpro.com/blog/es/ciclo-de-vida-de-los-datos/

[6] Cyberlink. (2023) Ciclo de vida de los datos: qué es y cuales son sus fases. [Online]. Available: https://www.cyberclick.es/numerical-blog/ciclo-de-vida-de-los-datos-que-es-y-cuales-son-sus-fases

[7] OCI. (2024) ¿qué es el machine learning? [Online]. Available: https://www.oracle.com/es/artificial-intelligence/machine-learning/what-is-machine-learning/

[8] IBM. (2024) ¿qué es el aprendizaje supervisado? [Online]. Available: https://www.ibm.com/es-es/topics/supervised-learning

[9] ——. (2024) ¿qué es el aprendizaje no supervisado? [Online]. Available: https://www.ibm.com/es-es/topics/unsupervised-learning

[10] AES. (2023) ¿qué es python? [Online]. Available: https://aws.amazon.com/es/what-is/python/

[11] A. S. Alberca. (2022) La librería pandas. [Online]. Available: https://aprendeconalf.es/docencia/python/manual/pandas/

[12] E. R. Schmidt. (2019) re- expresiones regulares. [Online]. Available: https://rico-schmidt.name/pymotw-3/re/index.html

[13] J. J. L. Gómez. (2018-2023) Python requests. la librería para hacer peticiones http en python. [Online]. Available: https://j2logo.com/python/python-requests-peticiones-http/#requests-get

[14] P. S. Foundation. (2023) os — interfaces misceláneas del sistema operativo. [Online]. Available: https://docs.python.org/es/3.10/library/os.html

[15] A. S. Alberca. (2022) La librería numpy. [Online]. Available: https://aprendeconalf.es/docencia/python/manual/numpy/

[16] P. S. Foundation. (2024) Cómo (howto) unicode. [Online]. Available: https://docs.python.org/es/3/howto/unicode.html

[17] itop. (2023) Scikit-learn. [Online]. Available: https://www.itop.es/soluciones-tecnologicas/business-analytics-business-intelligence/scikit-learn.html

[18] A. Vidhya. (2024) Tune hyperparameters with gridsearchcv. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/

[19] S. Inc. (2006 - 2024) What is logistic regression? equation, assumptions, types, and best practices. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[20] IBM. (2006 - 2024) ?qué es un árbol de decisión? [Online]. Available: https://www.ibm.com/es-es/topics/decision-trees

[21] datacamps. (2024) Soporte para máquinas vectoriales con el tutorial scikit-learn. [Online]. Available: https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python

[22] L. Ramírez. (2023-04-01) Algoritmo k-means: ¿qué es y cómo funciona? [Online]. Available: https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/

[23] D. SCIENTIST. (2022-30-11) Machine learning clustering: el algoritmo dbscan. [Online]. Available: https://datascientest.com/es/machine-learning-clustering-dbscan

[24] Kinsta. (2024) ¿qué es el web scraping? cómo extraer legalmente el contenido de la web. [Online]. Available: https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/

[25] steemit. (2020) ¿explicación alternativa para accuracy, precision, recall y f1-score? [Online]. Available: https://steemit.com/spanish/@waster/explicacion-alternativa-para-accuracy-precision-recall-y-f1-score

Anexos

Apéndice A

Informe de originalidad Turn it in

Turnitin Informe de Originalidad

Procesado el: 03-jun.-2024 18:32 CEST
Identificador: 2394722304
Número de palabras: 16732
Entregado: 1

Open_Data___Movilidad.pdf Por MARIA MENCIA SERRANO MANZANO

Índice de similitud	Similitud según fuente	
14%	Internet Sources:	10%
	Publicaciones:	2%
	Trabajos del estudiante:	10%

1% match (trabajos de los estudiantes desde 04-feb.-2024)
Submitted to Universidad Politécnica de Madrid on 2024-02-04

< 1% match (trabajos de los estudiantes desde 27-may.-2024)
Submitted to Universidad Politécnica de Madrid on 2024-05-27

< 1% match (trabajos de los estudiantes desde 31-may.-2024)
Submitted to Universidad Politécnica de Madrid on 2024-05-31

< 1% match (trabajos de los estudiantes desde 28-may.-2024)
Submitted to Universidad Politécnica de Madrid on 2024-05-28

< 1% match (trabajos de los estudiantes desde 02-jun.-2024)
Submitted to Universidad Politécnica de Madrid on 2024-06-02

< 1% match (trabajos de los estudiantes desde 26-jul.-2017)
Submitted to Universidad Internacional de la Rioja on 2017-07-26

< 1% match (trabajos de los estudiantes desde 21-sept.-2022)
Submitted to Universidad Internacional de la Rioja on 2022-09-21

< 1% match (trabajos de los estudiantes desde 17-jul.-2023)
Submitted to Universidad Internacional de la Rioja on 2023-07-17

< 1% match (trabajos de los estudiantes desde 28-abr.-2024)
Submitted to Universidad Internacional de la Rioja on 2024-04-28

< 1% match (trabajos de los estudiantes desde 01-abr.-2024)
Submitted to Universidad Internacional de la Rioja on 2024-04-01

< 1% match (Internet desde 09-jul.-2023)
https://oa.upm.es/75053/1/TFG_DIEGO_LOPEZ_LOPEZ.pdf

< 1% match (Internet desde 07-jul.-2023)
https://oa.upm.es/74946/1/TFG_ADRIAN_SANCHEZ_RODERO.pdf

< 1% match (Internet desde 23-oct.-2019)
http://oa.upm.es/56152/1/TFG_AASHO_KUMAR.pdf

< 1% match (Internet desde 10-mar.-2023)
https://oa.upm.es/72896/1/TFG_DAVID_VINAS_MORALES.pdf

< 1% match (Internet desde 17-ago.-2022)
https://oa.upm.es/68034/1/TFG_YAEL_GARCIA_NOTARIO.pdf

< 1% match (Internet desde 07-jul.-2023)
https://oa.upm.es/74905/1/TFG_ADRIAN_ALONSO_LEDESMA.pdf

< 1% match (Internet desde 07-abr.-2021)
http://oa.upm.es/66257/1/TFG_DANIEL_JESUS_DE_LA_VEGA_MARTIN.pdf

< 1% match ()
Sava Leș, Nicolae. "Lost At Night: ciencia ciudadana sobre contaminación lumínica", E.T.S. de Ingenieros Informáticos (UPM), 2017

< 1% match ()
Biosca Valiente, Bárbara. "Optimización de los procesos de medida e interpretación de la tomografía geoeléctrica en la prospección superficial", E.T.S.I. Minas (UPM), 2012

< 1% match (trabajos de los estudiantes desde 02-oct.-2023)
Submitted to Corporación Universitaria Iberoamericana on 2023-10-02

< 1% match (trabajos de los estudiantes desde 16-abr.-2023)
Submitted to Corporación Universitaria Iberoamericana on 2023-04-16

< 1% match (trabajos de los estudiantes desde 21-jun.-2019)
Submitted to Mondragon Unibertsitatea on 2019-06-21

< 1% match (trabajos de los estudiantes desde 24-may.-2024)
Submitted to Mondragon Unibertsitatea on 2024-05-24

< 1% match (trabajos de los estudiantes desde 09-jun.-2023)
Submitted to Universidad Carlos III de Madrid - EUR on 2023-06-09

< 1% match (trabajos de los estudiantes desde 12-feb.-2023)
Submitted to Universidad Carlos III de Madrid - EUR on 2023-02-12

< 1% match (trabajos de los estudiantes desde 31-ago.-2023)
Submitted to Universidad Carlos III de Madrid - EUR on 2023-08-31

< 1% match (Internet desde 18-ene.-2024)
https://raw.githubusercontent.com/serwikk/TIA/main/AccidentesBicicletas_22-23.csv

< 1% match (trabajos de los estudiantes desde 30-may.-2024)
Submitted to Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-05-30

< 1% match (trabajos de los estudiantes desde 17-feb.-2024)
Submitted to Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-02-17

< 1% match (trabajos de los estudiantes desde 22-feb.-2024)
Submitted to Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-02-22

< 1% match (trabajos de los estudiantes desde 07-sept.-2021)
Submitted to Universidad Carlos III de Madrid on 2021-09-07

< 1% match (trabajos de los estudiantes desde 09-dic.-2023)
Submitted to Universidad Carlos III de Madrid on 2023-12-09

< 1% match (trabajos de los estudiantes desde 07-sept.-2022)
Submitted to Universidad Carlos III de Madrid on 2022-09-07

< 1% match (trabajos de los estudiantes desde 13-ene.-2024)
Submitted to Universidad de Manizales on 2024-01-13

< 1% match (trabajos de los estudiantes desde 08-may.-2023)
Submitted to Universidad de Manizales on 2023-05-08

< 1% match (trabajos de los estudiantes desde 23-feb.-2024)
Submitted to Universitat Politècnica de València on 2024-02-23

< 1% match (trabajos de los estudiantes desde 15-may.-2024)
Submitted to Universitat Politècnica de València on 2024-05-15

< 1% match (Internet desde 13-nov.-2020)
https://steemit.com/spanish/@waster/explicacion-alternativa-para-accuracy-precision-recall-y-f1-score

< 1% match (Internet desde 01-feb.-2023)
https://dokumen.tips/documents/trabajo-fin-de-grado-sistema-big-data-para-el-anlisis-de-figura-42-tiempos.html

< 1% match (Internet desde 29-abr.-2020)
https://issuu.com/ogryarts/docs/thesis

< 1% match (Internet desde 24-jul.-2016)
https://issuu.com/grupoduende/docs/edm147

< 1% match (Internet desde 28-sept.-2016)
https://issuu.com/uninortecolombia/docs/libro_ingenieri__as

< 1% match (Internet desde 19-may.-2016)
https://issuu.com/ambitoscomunicacion/docs/revista-comunicacion-ambitos-19?mode=window

< 1% match (Internet desde 06-mar.-2023)
http://robolabo.etsit.upm.es/publications/TFG/TFG_LauraJara.pdf

< 1% match (Internet desde 06-nov.-2020)
https://tiovictorcardenasvalderrama.blogspot.com/2017/10/big-data-en-que-consiste-su-importancia.html

< 1% match (trabajos de los estudiantes desde 17-dic.-2023)
Submitted to ITESM: Instituto Tecnologico y de Estudios Superiores de Monterrey on 2023-12-17

< 1% match (trabajos de los estudiantes desde 15-dic.-2023)
Submitted to ITESM: Instituto Tecnologico y de Estudios Superiores de Monterrey on 2023-12-15

< 1% match (Internet desde 23-feb.-2023)
https://www.researchgate.net/figure/The-six-phases-of-the-traditional-CRISP-DM-Model-Shearer-2000-p14_fig3_319937079

< 1% match (Internet desde 08-jun.-2023)
https://www.researchgate.net/publication/326583645_Artificial_Intelligence_and_the_Public_Sector-Applications_and_Challenges

< 1% match (Internet desde 08-feb.-2023)
https://www.researchgate.net/publication/220529437_A_fuzzy_extension_of_the_silhouette_width_criterion_for_cluster_analysis

< 1% match (Internet desde 10-oct.-2021)
https://zaguan.unizar.es/record/106897/files/TAZ-TFG-2021-2157.pdf

< 1% match (trabajos de los estudiantes desde 28-ago.-2023)
Submitted to Centro Europeo de Postgrado - CEUPE on 2023-08-28

< 1% match (trabajos de los estudiantes desde 20-ene.-2024)
Submitted to Universidad TecMilenio on 2024-01-20

< 1% match (trabajos de los estudiantes desde 25-ene.-2024)
Submitted to Universidad TecMilenio on 2024-01-25

< 1% match (Internet desde 23-abr.-2024)
https://repositorio.unitec.edu/xmlui/bitstream/handle/123456789/13022/Carlos_Jeff_Milton_Hernandez.docx?isAllowed=y&sequence=2

< 1% match (trabajos de los estudiantes desde 06-jun.-2023)
Submitted to Infile on 2023-06-06

< 1% match (trabajos de los estudiantes desde 16-jul.-2018)
Submitted to Infile on 2018-07-16

< 1% match (trabajos de los estudiantes desde 16-may.-2024)
Submitted to Universidad Francisco de Vitoria on 2024-05-16

< 1% match (Internet desde 28-feb.-2024)
https://digibug.ugr.es/bitstream/handle/10481/89442/88425.pdf?isAllowed=y&sequence=4

< 1% match (Internet desde 19-jul.-2023)
https://digibug.ugr.es/bitstream/handle/10481/82216/95215.pdf?isAllowed=y&sequence=4

< 1% match (trabajos de los estudiantes desde 24-jun.-2022)
Submitted to Universidad de Málaga - Tii on 2022-06-24

< 1% match (trabajos de los estudiantes desde 12-mar.-2022)
Submitted to University of Bolton on 2022-03-12

< 1% match (Internet desde 12-mar.-2024)
http://repositorio.ucsg.edu.ec/bitstream/3317/22588/1/T-UCSG-PRE-CEAE-CNI-108.pdf

< 1% match (trabajos de los estudiantes desde 04-sept.-2022)
Submitted to Aston University on 2022-09-04

< 1% match (Internet desde 14-abr.-2024)

https://www.coursehero.com/file/226849095/Electiva1-tarea-2-2022-0226pdf/

< 1% match (Internet desde 06-oct.-2022)
https://www.coursehero.com/file/81795264/QUIZ-1-BUENOdocx/

< 1% match (Internet desde 11-ene.-2024)
https://1library.co/document/qo3kk1kq-historias-del-laberinto.html

< 1% match (trabajos de los estudiantes desde 28-abr.-2024)
Submitted to Fundación Universitaria del Area Andina on 2024-04-28

< 1% match (trabajos de los estudiantes desde 26-abr.-2023)
Submitted to Universidad del Istmo de Panamá on 2023-04-26

< 1% match (Internet desde 29-sept.-2023)
https://acerosarequipa.com/bo/es/noticias/525/aceros-arequipa-y-bbva-firman-alianza-sostenible-por-us40-millones-para-impulsar-la-economia-circular-en-el-pais

< 1% match (trabajos de los estudiantes desde 22-jun.-2023)
Submitted to Consorcio CIXUG on 2023-06-22

< 1% match (Internet desde 17-mar.-2024)
https://repositorio.usm.cl/bitstream/handle/11673/56501/3560901064936UTFSM.pdf?isAllowed=y&sequence=1

< 1% match ()
Camisassa, María Eugenia. "Estrellas enanas blancas: Procesos físicos y aplicaciones", 2019

< 1% match (Abid Al-Akioui, Andres Monzon, Candela Martin. "Mobility patterns in healthcare centres. Case Study: La Paz University Hospital (Madrid, Spain)", Transportation Research Procedia, 2023)
Abid Al-Akioui, Andres Monzon, Candela Martin. "Mobility patterns in healthcare centres. Case Study: La Paz University Hospital (Madrid, Spain)", Transportation Research Procedia, 2023

< 1% match (trabajos de los estudiantes desde 24-may.-2024)
Submitted to De Montfort University on 2024-05-24

< 1% match (Internet desde 07-may.-2023)
https://polodelconocimiento.com/ojs/index.php/es/article/download/5457/13439

< 1% match ()
Red de Universidades con Carreras en Informática (RedUNCI). "CACIC 2016 | XXII Congreso Argentino de Ciencias de la Computación : Libro de Actas", Nueva Editorial Universitaria, 2016

< 1% match (trabajos de los estudiantes desde 27-ene.-2023)
Submitted to College of Estate Management on 2023-01-27

< 1% match (trabajos de los estudiantes desde 20-may.-2024)
Submitted to Universidad Catolica San Antonio de Murcia on 2024-05-20

< 1% match (trabajos de los estudiantes desde 04-dic.-2017)
Submitted to Universidad de Deusto on 2017-12-04

< 1% match (Internet desde 05-nov.-2022)
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default&vgnextoid=fabbf3e1de124610VgnVCM2000001f4a900aRCRD

< 1% match (Internet desde 05-may.-2023)
https://rstudio-pubs-static.s3.amazonaws.com/757826_1f7105d392be4d9997fafbd5ab4a4282.html

< 1% match (Internet desde 12-ene.-2023)
https://rstudio-pubs-static.s3.amazonaws.com/845322_30b72a4d18bb409bbafa1c974c9865de.html

< 1% match (trabajos de los estudiantes desde 29-ene.-2024)
Submitted to UWC Dilijan on 2024-01-29

< 1% match (trabajos de los estudiantes desde 09-nov.-2023)
Submitted to Universidad Santo Tomas on 2023-11-09

< 1% match ()
http://di002.edv.uniovi.es/~luciano/iblog/B1727748937/C97892319/E379996959/Media/ingcon_clasificacion.pdf

< 1% match (trabajos de los estudiantes desde 13-mar.-2023)
Submitted to Universidad Nacional de Educación a Distancia on 2023-03-13

< 1% match (Internet desde 24-may.-2024)
https://ph01.tci-thaijo.org/index.php/saujournalst/article/view/254394

< 1% match (Internet desde 16-may.-2024)
https://view.genially.com/663981cebcbaac0014fe9f29/presentation-fase-5-definir-la-importancia-de-la-ia-en-la-formacion-de-la-imagen-y

< 1% match (trabajos de los estudiantes desde 22-feb.-2024)
Submitted to Pontificia Universidad Javeriana Cali on 2024-02-22

< 1% match (Internet desde 20-nov.-2022)
https://riuma.uma.es/xmlui/bitstream/handle/10630/25049/Ruiz%20Su%c3%a1rez%2c%20Yeray%20Memoria.pdf?isAllowed=y&sequence=1

< 1% match (trabajos de los estudiantes desde 17-sept.-2021)
Submitted to Universidad de Alcalá on 2021-09-17

< 1% match (Internet desde 26-oct.-2023)
https://dspace.ups.edu.ec/bitstream/123456789/26427/4/UPS-CT010986.pdf

< 1% match (Internet desde 16-ene.-2023)
https://idus.us.es/bitstream/handle/11441/127072/TFG-3797%20GALLARDO%20G%C3%93MEZ%2C%20LUIS.pdf?isAllowed=y&sequence=1

< 1% match (Internet desde 30-ene.-2024)
https://repositorio.puce.edu.ec/server/api/core/bitstreams/9fc84446-0433-4c8e-a113-dc6fbc38d008/content

< 1% match (Internet desde 12-dic.-2023)
https://repository.unab.edu.co/handle/20.500.12749/23041?show=full

< 1% match (Internet desde 19-nov.-2022)
https://uvadoc.uva.es/bitstream/handle/10324/57251/TFG-G5795.pdf?isAllowed=y&sequence=1

< 1% match (Internet desde 06-mar.-2024)

https://www.scoop.it/topic/web-2-0-education/p/4028081745/2014/09/15/specialized-magazines-for-apps

< 1% match (Internet desde 19-dic.-2022)
https://www.slideshare.net/WilfredoFigueroaWjfi/07-investigacion-sobre-localizacion-y-distribucion-de-plantas

< 1% match (Internet desde 29-sept.-2021)
https://www.spreaker.com/user/urosarioradio/sistemas-de-informacion-georeferenciadas

< 1% match (Joyce Grabher Meier, Luciane Patrícia Andreani Cabral, Camila Zanesco, Clóris Regina Blanski Grden et al. "Factors associated with the frequency of medical consultations by older adults: a national study", Revista da Escola de Enfermagem da USP, 2020)
Joyce Grabher Meier, Luciane Patrícia Andreani Cabral, Camila Zanesco, Clóris Regina Blanski Grden et al. "Factors associated with the frequency of medical consultations by older adults: a national study", Revista da Escola de Enfermagem da USP, 2020

< 1% match (Internet desde 22-jun.-2020)
https://archive.org/stream/EnciclopediaSalvatDeCienciaYTecnicaVol05141986/Enciclopedia%20Salvat%20De%20Ciencia%20Y%20Tecnica%20vol%2005_14%201986_djvu

< 1% match (Internet desde 21-nov.-2020)
https://doku.pub/documents/memoriastecnicaspdf-4qz3637wx90k

< 1% match (Internet desde 26-nov.-2022)
https://olacefs.com/en/the-ccc-invites-you-to-the-7th-edition-of-the-international-seminar-on-data-analysis-in-public-administration/

< 1% match (Internet desde 21-sept.-2023)
https://openarchive.icomos.org/id/eprint/2955/1/K649-Monuments_and_Sites-v13-2005.pdf

< 1% match (Internet desde 18-may.-2010)
http://vodafone.es/fundacion/fundacion.vodafone.es/VSharedClient/FundacionVodafone/PDF/ELCEREBRO.PDF

< 1% match (Internet desde 22-nov.-2002)
http://www.diariomedico.com/enlared/not170200.html

< 1% match ()
(2-17-03) http://80.81.104.134/2003-01-20/palma/palma0.htm

< 1% match (Alberto García García. "Arquitectura de interoperabilidad para mejorar la gestión y coordinación de múltiples UXV y la toma de decisiones", Universitat Politecnica de Valencia, 2024)
Alberto García García. "Arquitectura de interoperabilidad para mejorar la gestión y coordinación de múltiples UXV y la toma de decisiones", Universitat Politecnica de Valencia, 2024

< 1% match (Silvia Marzal Romeu. "Concepción e integración de arquitecturas y protocolos de comunicación dentro de sistemas de supervisión y control de microrredes inteligentes", Universitat Politecnica de Valencia, 2019)
Silvia Marzal Romeu. "Concepción e integración de arquitecturas y protocolos de comunicación dentro de sistemas de supervisión y control de microrredes inteligentes", Universitat Politecnica de Valencia, 2019

< 1% match (Internet desde 01-jun.-2024)
https://doczz.es/doc/18045/prefacio---gaia-%E2%80%93-group-of-artificial-intelligence-applic...

< 1% match (Internet desde 05-oct.-2020)
https://es.qaz.wiki/wiki/Machine_learning

< 1% match (Internet desde 14-jun.-2008)
http://estudiantes.medicinatv.com/noticias/Default.asp?codigo=355143

< 1% match (Internet desde 07-may.-2021)
https://qdoc.tips/marketing-relacional-clientes-pdf-free.html

< 1% match (Internet desde 02-jun.-2022)
http://repositorio.mopt.go.cr:8080/xmlui/bitstream/handle/123456789/4702/Sector%20Tibas-Sto%20Domingo%20Informe%20Final.pdf?isAllowed=y&sequence=1

< 1% match (Internet desde 29-jul.-2022)
https://sct.uab.cat/estadistica/en/node/3380

< 1% match (Internet desde 12-dic.-2022)
https://tesis.ipn.mx/bitstream/handle/123456789/27154/TESIS.pdf?isAllowed=y&sequence=1

< 1% match ()
http://www.autoprofesional.com/noticias/2000/2000nov/30112000.html

< 1% match (Internet desde 01-ago.-2007)
http://www.ifemamotor.ifema.es/modules/news/index.php?storytopic=0&start=170

< 1% match ()
http://www.revistacompudata.com/imagesU/pdf/1060.pdf

< 1% match (Internet desde 07-dic.-2023)
https://www.tripadvisor.com.au/Attraction_Review-g297317-d7130174-Reviews-Iglesias_de_la_Chiquitania-Santa_Cruz_Santa_Cruz_Department.html

< 1% match (Internet desde 05-feb.-2021)
https://energynet.fronius.com/es/latin-america/tecnologia-de-carga-de-baterias/nuestra-experiencia/intralogistica/independencia-de-la-alimentacion-principal

< 1% match (Internet desde 16-dic.-2020)
https://es.wikihow.com/calcular-el-rango-estad%C3%ADstico

< 1% match ()
Pardo-Rodríguez, Jhindy Hasleyde, Sánchez-Suárez, María Alejandra. "Implementación de un prototipo funcional de aprendizaje de máquina para identificar correos electrónicos de Spear Phishing", 'World Scientific Pub Co Pte Lt', 2021

< 1% match (Internet desde 12-mar.-2021)
https://lookformedical.com/es/search/salud-mental

< 1% match (Internet desde 03-ene.-2007)
http://microsites.aprendemas.com/Elisava/p2.asp

< 1% match (Internet desde 06-may.-2016)
https://repositorio.uam.es/bitstream/handle/10486/14091/66095_campos%20soto%20pedro%20g..pdf

< 1% match ()
Red de Universidades con Carreras en Informática, Finochietto, Jorge. "CACIC 2013 : XIX Congreso Argentino de Ciencias de la Computación. Libro de actas", Fundación de Altos Estudios en Ciencias Exactas, 2013

< 1% match (Internet desde 16-sept.-2020)
https://www.humanlevel.com/articulos/google-adwords/nuevos-informes-graficos-en-google-adwords.html

< 1% match (Internet desde 06-oct.-2013)
http://www.mokhafaf.com/fsearch/GIGO/

< 1% match (Internet desde 27-nov.-2020)
https://www.toodledo.com/tasks/public.php?f=0&h=0&id=td555bf29ed6849&s=2

< 1% match ()
Soria Méndez, Bryan Andrés. "Desarrollo de un agente para la detección de Spam en el servicio de correo electrónico Zimbra aplicando técnica de machine learning de clasificación de texto para un GAD municipal.", La Libertad: Universidad Estatal Península de Santa Elena, 2023, 2023

Universidad Politécnica de Madrid Escuela Técnica Superior de Ingenieros Informáticos Grado en Matemáticas e Informática Trabajo Fin de Grado Open Data - Movilidad Autor: María Mencía Serrano Manzano Tutor(a): Luis Mengual Galán Madrid, Junio 2024 Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa. Trabajo Fin de Grado Grado en Matemáticas e Informática Título: Open Data - Movilidad Junio 2024 Autor: Tutor: María Mencía Serrano Manzano Luis Mengual Galán Departamento de Lenguajes y Sistemas Informáticos e Inge- niería de Software Escuela Técnica Superior de Ingenieros Informáticos Universidad Politécnica de Madrid Resumen En la era digital estamos ante un crecimiento urbano continuo, la comunicación eficiente entre los distintos puntos de una ciudad se ha convertido en uno de los desafíos principales para las ciudades modernas. En un contexto donde la mo- vilidad se ha vuelto una necesidad cotidiana para miles de ciudadanos, Madrid, como muchas otras metrópolis, se enfrenta a la compleja tarea de optimizar su sistema de transporte para satisfacer las demandas de su población. Cada día, un flujo constante de personas se ve desafiado por la necesidad de llegar a sus lugares de trabajo, colegios o universidades a tiempo, lo que implica enfrentarse al tráfico y a la puntualidad del transporte público. Es en este contexto surge Open Data-Movilidad como una respuesta a la necesidad apremiante de com- prender y mejorar la movilidad urbana. Este proyecto se enfoca en el tratamiento de la información pública disponible sobre el tráfico y movilidad de la Comunidad de Madrid y, dado que el volumen es considerable, desarrollar el ciclo de vida del dato completo utilizando técnicas de big data. En primer lugar, se recopilarán todos los datos de interés, a conti- nuación se procederá a su limpieza y tratamiento, para llevar a cabo un análisis de ellos mediante técnicas de aprendizaje supervisado y no supervisado como regresión lineal o k-means. Esto permitirá identificar tendencias de tráfico y de- sarrollar modelos predictivos que ayuden a comprender el contexto territorial y a tomar decisiones informadas respecto al transporte público y la mejora de las vías y carreteras de la ciudad. i Abstract In this digital era and in the middle of a continuous urban growth, the efficient communication between different places in a city, has become one of the main challenges for modern cities. In a context where mobility has turned to be a daily necessity for thousands of citizens, Madrid, as many other metropolis, encoun- ters the complex task of optimising its transport system to meet its population demands. Every day, a constant flow of people is challenged by the need of mak- ing it on time to their workplaces, schools or universities, which means facing traffic and public transport's punctuality. It is in this context that "Open data - Mobility" comes to play as an answer to this urgent need of understanding and improving the urban mobility. This project focuses on the processing of public information about traffic and mobility in the Community of Madrid and, given the considerable volume, de- veloping a complete data life cycle using "big data" techniques. In first instance, all relevant data will be gathered; afterwards, it will be cleaned and processed to carry out its analysis with supervised and unsupervised learning techniques such as lineal regression or "k-means". This will allow us to identify traffic patterns and develop predictive models that would help comprehend the terri- torial context and make informed decisions regarding public transport and the improvement of the roads and motorways of the city. iii Tabla de contenidos 1. Introducción 1.1. Contexto y motivación . . . . . 1.2. Estructura del documento . . . . . 2. Estado del arte 2.1. Big data . . . . . 2.1.1. Introducción . . . . . 2.1.2. Las 5V's . . . . . 2.1.3. Ciclo de vida del dato . . . . . 2.1.4. Aprendizaje automático . . . . . 2.2. Técnicas y herramientas . . . . . 2.2.1. Python . . . . . 2.2.2. QGIS . . . . . 2.2.3. Scikit-Learn . . . . . 2.2.4. PostgreSQL . . . . . 2.2.5. Docker . . . . . 2.2.6. Web scraping . . . . . 3. Desarrollo 3.1. Metodología . . . . . 3.2. Extracción de datos . . . . . 3.3. Limpieza . . . . . 3.4. Preprocesamiento . . . . . 3.5. Almacenamiento . . . . . 3.6. Análisis . . . . . 3.6.1. Aprendizaje supervisado . . . . . 3.6.2. Aprendizaje no supervisado . . . . . 4. Resultados y conclusiones 4.1. Resultados . . . . . 4.1.1. Aprendizaje supervisado . . . . . 4.1.2. Aprendizaje no supervisado . . . . . 4.2. Conclusiones . . . . . 4.3. Líneas futuras . . . . . 5. Análisis de impacto v 1 1 2 3 3 3 4 5 7 8 8 9 10 13 13 13 15 15 16 20 21 28 28 30 34 55 55 55 55 58 59 61 TABLA DE CONTENIDOS Bibliografía Anexos A. Informe de originalidad Turn it in 63 67 67 vi Capítulo 1 Introducción 1.1. Contexto y motivación Debido al progreso de las tecnologías de información, el volumen de datos que se produce diariamente ha incrementado considerablemente y, por tanto, las herramientas tradicionales no son capaces de procesarlos. Por ello, las orga- nizaciones y empresas han tenido que desarrollar nuevas técnicas capaces de analizar y comprender más allá de lo que las herramientas tradicionales ofrecen sobre los datos, aumentando así su competitividad. Este proyecto se enmarca en el ámbito de transporte, pues se utiliza la abun- dancia de datos públicos disponibles sobre el entorno urbano de Madrid. Estos datos abarcan una amplia gama de información relacionada con el transporte, carreteras y accidentes, ofreciendo una visión detallada y completa del ecosiste- ma de movilidad de la ciudad. El propósito fundamental de este proyecto es extraer conocimientos significati- vos a partir de esta información, utilizando nuevas técnicas de tratamiento de datos que permitan gestionar el gran volumen de información existente, contri- buyendo así a la mejora continua del sistema de transporte y red de carreteras madrileñas. A través de técnicas como minería de datos, análisis estadístico y visualización de datos, se pretende identificar los patrones subyacentes de mo- vilidad, así como las áreas de congestión, con el fin de proponer modelos pre- dictivos basados en datos históricos que puedan anticipar y prevenir cuellos de botella en el futuro. Este enfoque no solo permite una mayor comprensión de la movilidad sino que también pretende facilitar la toma de decisiones informadas orientadas a mejorar la eficiencia del sistema en su conjunto. Open Data-Movilidad se propone no solo recopilar y analizar datos relacionados con la movilidad urbana, sino también establecer un sólido ciclo de vida del dato que asegure la calidad y la utilidad de la información en todas sus etapas. Se si- guen estrictas prácticas de gestión de datos desde la recolección inicial hasta el Capítulo 1. Introducción análisis y la representación visual para asegurar la autenticidad y confiabilidad de la información. La verdadera utilidad de la información radica en su capacidad para ser com- prendida y utilizada por una amplia variedad de usuarios. El hacer el proyecto accesible hace que pueda servir como modelo para realizar análisis similares de cualquier zona geográfica a nivel regional o global. 1.2. Estructura del documento Capítulo I • Contexto y motivación: se expone la motivación del proyecto y los ob- jetivos del mismo, recogiendo la idea general del trabajo a realizar. • Estructura del documento. Capítulo II • Estado del arte: introducción al big data y sus conceptos básicos. • Herramientas: se detallan las tecnologías que se utilizarán en el pro- yecto. Capítulo III • Diseño: se relata la estructura del proyecto y las fases de este. • Implementación: se describe detalladamente los pasos llevados a cabo en cada etapa. Capítulo IV • Resultados • Conclusiones • Líneas futuras Capítulo V • Análisis de impacto Capítulo 2 Estado del arte 2.1. Big data 2.1.1. IntroducciónLa importancia de los datos en la sociedad actual radica en su capacidad para ayudarnos a entender nuestro entorno. En las últimas décadas, gracias al auge de Internet, los dispositivos móviles y sensores en todo tipo de ámbitos, se ha generado una inmensa cantidad de información sin precedentes. Esta ingente cantidad de datos, sumado a las tecnologías de vanguardia disponibles en la actualidad, han dado lugar alsurgimiento del concepto de big data. Este término se refiere a los conjuntos de datos voluminosos y complejos que exceden la capacidad de procesamiento de los sistemas informáticos convencio- nales. Estos datos cuya procedencia es diversa, desde redes sociales, transac- ciones comerciales, dispositivos conectados a Internet, sensores en tiempo real etc. tienen un gran potencial, ya que ofrecen perspectivas valiosas que se pue- den utilizar para mejorar la toma de decisiones, optimizar procesos e identificar tendencias y patrones entre otras cosas. Por otro lado, el simple hecho de contar con grandes cantidades de datos no garantiza automáticamente su utilidad. La verdadera clave radica en la capaci- dad de analizar, interpretar y extraer información significativa de estos datos. Es aquí donde entran en juego técnicas avanzadas de análisis de datos, como la inteligencia artificial o machine learning. Estas herramientas permiten descubrir correlaciones, identificar anomalías, predecir comportamientos futuros y gene- rar conocimiento útil a partir de los datos masivos disponibles. En última instancia, el big data no solo representa un desafío tecnológico, sino también un cambio de paradigma en la forma en que comprendemos y aborda- mos la información [1]. Capítulo 2. Estado del arte 2.1.2. Las 5V's Los conjuntos de datos se enfrentan a diversos desafíos gracias a las caracterís- ticas del big data. Estos se conocen como las 5 Vs: volumen, variedad, veracidad, velocidad y valor, que definen la problemática del big data. Estas 5 características provocan que las empresas tengan problemas para ex- traer datos reales y de alta calidad de conjuntos de datos tan masivos, cambian- tes y complicados. [2] Volumen: se define como la cantidad de datos que se generan y recopilan en cada instante en este mundo digitalizado. El gran volumen de datos plantea muchas dificultades como puede ser su almacenamiento en un lugar seguro y accesible, su distribución en distintos puntos sin perder la disponibilidad y coherencia a tiempo real y su procesamiento, pues se requiere una gran capacidad de cómputo. Las principales dificultades que se encuentran son el coste, la escalabilidad y el rendimiento. El incremento del volumen también es consecuencia del aumento de las fuentes de datos, cada día hay más personas conectadas y, de la calidad y precisión de estos datos (por ejemplo, la de los sensores) [1]. Velocidad: además de gestionar grandes volúmenes de datos, las empresas necesitan obtener información rápidamente. Maximizar la velocidad con la que se crean, transmiten y procesan los datos puede ser un gran desafío. La información puede tratarse a tiempo real o con algo de demora, lo que es crucial en aplicaciones que requieren respuestas instantáneas como la detección de fraudes en operaciones financieras o la monitorización del tráfico. Variedad: este concepto hace refernecia a la diversidad de formatos, tipos y fuentes de información. Los datos no necesariamente están estructurados o semiestructurados, es decir, pueden no tener un esquema y estructu- ra fijos pensados para ser almacenados en una base de datos tradicional; pueden ser objetos, documentos, imágenes, tuits o datos geoespaciales. Además, su origen es diverso, como las máquinas, las personas y los pro- cesos organizativos. Hay numerosos factores que promueven la variedad, pero entre otros encontramos las tecnologías móviles, las redes sociales, las geotecnologías o los vídeos. Veracidad: la veracidad hace referencia a la calidad y el origen de los da- tos. Es crucial asegurar su coherencia, completitud, integridad y que estén libres de ambigüedades. Entre los diversos factores que impulsan la vera- cidad están el coste y la necesidad de trazabilidad. Dado el gran volumen, velocidad y variedad de los datos que se generan, hay que asegurarse de que la información que recibimos no sea falsa [1]. El concepto " Garbage in, garbage out " junto a "Garbage in, gospel out" que se entienden como aceptar ciegamente la información generada automáticamente proveniente de máquinas, ilustran que la entrada de datos sin sentido provoca la salida de información carente de este. En algunos casos se pueden "limpiar" los 2.1. Big data datos de entrada, pero hay contextos como la economía, condiciones cli- máticas o decisiones de empresas que generan una incertidumbre que los sistemas big data han de asumir y tolerar [3]. Valor: posiblemente la V más importante, el despliegue de tecnologías solo tiene sentido si los datos aportan algún valor o beneficio tangible y sig- nificativo. Este valor viene de reconocer patrones que mejoren la eficiencia operativa, impulsen la innovación o proporcionen ventajas competitivas [4]. Figura 2.1: 5V's 2.1.3. Ciclo de vida del dato El ciclo de vida del dato es una sucesión de etapas por las que transcurren los datos a lo largo de toda su vida útil. Estas fases se definen en función de distin- tos criterios y el dato pasa de una a otra a medida que se completan distintas tareas o cumplen ciertos requisitos. Este periodo abarca desde la generación del dato hasta su

reutilización o eliminación, y se considera un ciclo pues la información generada a partir de unos datos puede servir de base para un pro- yecto posterior, consiguiendo así que la última etapa del proceso retroalimente la primera [5]. Este ciclo proporciona una visión general de las etapas que intervienen en la generación, uso y reutilización del dato. Llevar a cabo correctamente cada etapa permite tratar el dato de una forma más eficiente, preservar su calidad y generar información de mayor valor. Además, en el ámbito empresarial permite llevar a cabo un uso más seguro, evitando pérdidas y eliminaciones, definiendo el trato, uso, almacenamiento y compartición de la información [6]. A continuación se describen las etapas de este ciclo. Generación o captura: en esta primera fase se produce la creación del dato en bruto, que se obtiene a través de distintas técnicas que pueden abarcar desde la compra del dato hasta la creación automática de este gracias a dispositivos y sistemas automáticos. Almacenamiento: los datos ocupan un espacio y han de ser almacenados adecuadamente en repositoorios como las bases de datos. El correcto alma- cenamiento es clave para garantizar la accesibilidad y el control sobre los datos. En este proceso es importante diferenciar los datos estructurados de los no estructurados pues cada uno se almacenará de una forma distinta o en lugares y formatos distintos. Tratamiento: en esta etapa el dato es preparado, organizado y transformado para su uso. Se pueden utilizar técnicas de análisis de datos y aprendizaje automático para extraer el valor y utilidad de los datos adquiridos. Un buen tratamiento del dato garantiza la calidad de los resultados y proporciona una buena base para la toma decisiones bien fundamentadas. Uso: gracias a los resultados del análisis se toman decisiones estratégicas como la optimización de procesos o reformas en un sistema logístico. Eliminación: una vez el dato ha dejado de proporcionar información útil o ya no tiene un propósito significativo, se destruye. El volumen de datos en cualquier organismo crece considerablemente con el paso del tiempo y no es factible el almacenamiento de todos ellos. Figura 2.2: Ciclo de vida del dato. 2.1. Big data 2.1.4. Aprendizaje automático Tras comprender aspectos básicos del big data y las etapas por las que pasan los datos, el siguiente paso es entender qué herramientas se utilizan en la fase de tratamiento del dato para obtener el valor de este. El aprendizaje automático también conocido como machine learning es una subcategoría de la inteligen- cia artificial que tiene como objetivo desarrollar sistemas capaces de mejorar su rendimiento de forma automática gracias a la experiencia y los datos que consumen, sin ser explícitamente programados [7]. El aprendizaje automático es capaz de, a través de algoritmos que identifican patrones, generar modelos que realizan predicciones. Cuantos más datos y de mayor calidad, mejores re- sultados se obtienen. Existen dos categorías dentro del aprendizaje automático, supervisado y no supervisado. Aprendizaje automático supervisado En este tipo de aprendizaje el modelo es entrenado a partir de un conjunto de datos previamente etiquetado, es decir, los datos ya contienen las repuestas correctas. El total de los datos se divide en dos, conjunto de entrenamiento y conjunto de prueba. Con el primer conjunto se entrena el modelo, es decir, se le enseñan los resultados que debe generar asociando las características de los datos a las etiquetas correspondientes. Una vez terminada esta fase, se utiliza el conjunto de prueba para que el modelo haga predicciones sobre este. Como ya se tiene la respuesta a esa predicción se puede evaluar la precisión del modelo. Dentro del aprendizaje supervisado existen dos tipos de tareas, clasificación y regresión. Clasificación: las tareas de clasificación tratan de predecir etiquetas dis- cretas, por ejemplo, cuando se recibe una llamada, saber si es spam o no. Esta clasificación puede ser binaria (llamada) o multiclase, por ejemplo, identificar qué enfermedad tiene un paciente dados unos síntomas. Regresión: los problemas de regresión tratan de predecir un valor continuo, como el precio de una casa basado en sus características o el consumo de luz de un hogar [8]. Aprendizaje automático no supervisado El aprendizaje no supervisado utiliza datos que no están etiquetados. El obje- tivo aquí es explorar la estructura de los datos para encontrar algún patrón o secuencia y realizar una agrupación de los datos. Las técnicas de aprendizaje no supervisado son útiles para la segmentación de clientes, la organización de gran- des bibliotecas de documentos y la detección de patrones atípicos o anomalías. Hay tres tareas principales que se resuelven con aprendizaje no supervisado: agrupación en clusters, asociación y reducción de dimensionalidad [9]. Agrupación en clusters: consiste en la agrupación de datos que no están etiquetados en función de sus similitudes o diferencias. Los algoritmos de agrupación en clusters también conocidos como clustering, se emplean para dividir información sin clasificar en grupos representados por estructuras o patrones en la información. Hay varios tipos de algoritmos de agrupación: exclusivos, superpuestos, jerárquicos y probabilísticos. Asociación: una regla de asociación es un método basado en reglas que trata de encontrar relaciones entre las distintas variables dentro de un conjunto de datos. Un ejemplo típico es el problema de la cesta de la com- pra, que permite a las empresas entender las relaciones entre los distin- tos productos y así desarrollar estrategias de venta cruzada y marketing. Por ejemplo, los clientes que compraron una barbacoa también compra- ron utensilios de cocina. Existen distintos algoritmos que generan reglas de asociación, como a priori, Eclat y FP-Growth, el algoritmo a priori es el más utilizado. Ajuste de dimensionalidad: en principio, puede parecer lógico que cuan- tos más datos se obtendrán resultados más precisos, pero también pueden afectar al rendimiento del modelo y dificultar la visualización del conjunto de datos. La técnica de reducción de dimensionalidad se basa en la reduc- ción del número de entradas de datos a un tamaño gestionable y preservar la integridad del conjunto. Cuando el número de dimensiones es muy ele- vado, se aplica este método en la fase de preprocesamiento. Existen varias técnicas para realizar una reducción de dimensionalidad como el análisis de componentes principales, la descomposición en valores singulares o co- dificadores automáticos. 2.2. Técnicas y herramientas En esta sección se describen las técnicas y herramientas que se utilizarán en las distintas fases del proyecto. Para desarrollar el trabajo se ha elegido el len- guaje python en el entorno Visual Studio Code. A continuación se describen las funcionalidades utilizadas. 2.2.1. Python Python es un lenguaje de programación de alto nivel ampliamente utilizado, pues es simple, eficiente, fácil de aprender y se puede ejecutar en muchos entornos diferentes. Es utilizado para desarrollo de aplicaciones web, software, ciencia de datos y machine learning [10]. Además, es un lenguaje que tiene una gran biblioteca con códigos reutilizables para casi cualquier tarea. Entre ellos, para este proyecto se han utilizado: Pandas: es una librería especializada en el manejo y tratamiento de estruc- turas de datos. El origen de los datos puede ser archivos en formato CSV, Excel o bases de datos SQL. El acceso a estos se realiza mediante índices para filas y columnas y permite la reordenación y combinación de los datos mediante 3 estructuras, series, dataFrames y panel. En este proyecto se han utilizado las dos primeras [11]. 2.2. Técnicas y herramientas Re: una expresión regular consiste una secuencia de caracteres que da lugar a un patrón de búsqueda en un texto. Esta librería permite trabajar con expresiones regulares y encontrar patrones en el texto [12]. Request: esta librería permite realizar peticiones http a una página web. Proporciona dos métodos principales, GET, para consumir de una API o extraer información de una página; y POST, para enviar contenido de un formulario de forma automática [13]. En este proyecto se ha utilizado GET. Os: permite usar funcionalidades del sistema operativo como abrir o guar- dar un archivo, manipular rutas o crear archivos y directorios temporales [14]. En este proyecto se utiliza para guardar de forma ordenada toda la información obtenida y poder acceder a ella para tratarla. Glob: este módulo, con un patrón especificado previamente, permite encon- trar los nombres de rutas que se asemejan a este. Esta funcionalidad se ha utilizado para acceder a los distintos archivos almacenados en distintos bloques según la fase del proyecto. Numpy: es una librería especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. A través de arrays se representan colecciones de datos de un mismo tipo en varias dimensiones y proporciona funciones eficientes para la manipulación de estos [15] . JSON: utilizado para leer archivos de tipo JSON. Pyproj: utilizada para realizar operaciones con coordenadas geográficas. Los datos con los que se trabaja contienen ubicaciones geográficas de even- tos y esta librería permite tratar estos datos. Unicode: es una herramienta que codifica cada elemento del lenguaje con un código propio y único. Cada carácter tiene asignado un número entero de 0 a 0x10FFF [16]. esta herramienta es utilizada para analizar cadenas de texto y realizar modificaciones en estas, como puede ser cambiar una letra de mayúscula a minúscula o suprimir una tilde a través de la codificación única de cada caracter. SqlAlchemy: es una librería que ofrece la posibilidad de conectarse a una base de datos, comunicarse con ella e intercambiar información en ambos sentidos, tanto para introducir información como para leerla. Se utilizará para tratar los datos una vez almacenados. 2.2.2. QGIS QGIS, cuyas siglas se corresponden con Sistema de Información Geográfica, es una herramienta que permite la creación, visualización, análisis y edición de información geoespacial. Entre sus múltiples funcionalidades se encuentran: Edición de datos: QGIS permite editar y crear nuevos datos geográficos, como agregar nuevos puntos, líneas o polígonos a un mapa, o modificar los existentes según sea necesario. Etiquetado y simbolización: permite etiquetar elementos en un mapa con texto descriptivo y personalizable, así como aplicar diferentes estilos de simbología para representar los datos de manera visualmente atractiva y comprensible. Geocodificación: QGIS incluye herramientas para convertir direcciones o descripciones de lugares en coordenadas geográficas (geocodificación), lo que facilita la ubicación de lugares específicos en un mapa. Análisis avanzado: además de las herramientas básicas de análisis, QGIS también ofrece capacidades más avanzadas, como análisis de redes, inter- polación espacial, análisis de visibilidad y análisis de terreno, que permiten realizar análisis complejos de datos geo espaciales. En este proyecto ha si- do utilizado para representar los datos obtenidos de una forma intuitiva y sencilla, creando mapas de calor y visualizaciones dinámicas que muestran la evolución de los datos a lo largo del tiempo. Integración de datos externos: permite la integración de datos, como imá- genes satelitales, datos climáticos e información demográfica, entre otros, para enriquecer y contextualizar los mapas y análisis realizados en QGIS. 2.2.3. Scikit-Learn Es una biblioteca ampliamente utilizada para el preprocesamiento, la reducción de dimensionalidad, clasificación, regresión, clustering y selección de modelos. Ofrece una gran variedad de algoritmos y utilidades que hacen que sea una herramienta básica para realizar análisis de datos y modelado estadístico. Se caracteriza por tener una amplia variedad de módulos y algoritmos de aprendi- zaje automático, la posibilidad de extraer datos de repositorios y conjuntos de prueba [17]. Entre las distintas funcionalidades que ofrece se han utilizado las siguientes: Simple Imputer: esta función permite sustituir valores nulos por otros se- gún distintas estrategias como la media, la moda o la mediana. Cuando se obtienen conjuntos de datos, es común que estén incompletos y hay que tratar adecuadamente estas situaciones. Con esta herramienta, todos los campos que no ofrecen información (nulos) pueden reemplazarse por valores aceptables conocidos. Standard Scaler: es una herramienta que permite estandarizar las caracte- rísticas de los datos mediante la eliminación de la media y la varianza uni- taria. Este proceso es crucial, debido a que muchos algoritmos de apren- dizaje automático asumen que todas las características están centradas en cero y tienen varianza en la misma escala. En este proyecto se ha utilizado en la preparación de los datos para los modelos SVC y regresión logística. GridSearchCV: se trata de una herramienta que se utiliza para encontrar los hiperparámetros más óptimos para un modelo. Los modelos predictivos 2.2. Técnicas y herramientas tienen datos de entrada que el programador ha de elegir y, esta elección no es aleatoria. Para seleccionar las características que consigan el mejor rendimiento del modelo se utilizan distintas técnicas entre las cuales se encuentra esta. Funciona de tal forma que cuando le proporcionas una lista de posibles parámetros, prueba todas las combinaciones posibles y te retorna la combinación con la que se obtienen mejores resultados [18]. Regresión logística: es un algoritmo de aprendizaje automático supervisado utilizado para la predicción de eventos binarios, es decir, solo hay dos resul- tados posibles, si o no. La regresión logística analiza la relación entre una o más variables independientes y clasifica los datos en clases discretas. Es utilizado en modelos predictivos, donde el modelo estima la probabilidad matemática de si una instancia pertenece a una categoría específica o no [19]. Árbol de decisión: es un algoritmo de aprendizaje supervisado sin paráme- tros utilizado para tareas de clasificación y de regresión. Consta de un nodo raíz, ramas y nodos donde cada nodo representa un resultado posible den- tro del conjunto de datos. Esta estructura proporciona una forma fácil de afrontar la toma de decisiones, permitiendo la mejor comprensión de por qué se tomó una decisión. Se trata de una estrategia de divide y vencerás [20]. Bosque aleatorio: es un método que elabora múltiples árboles de decisión y los combina para obtener una predicción precisa y robusta. Es un mo- delo que tiene alta precisión y la capacidad de manejar conjuntos de datos grandes con numerosas variables. Destaca por su utilidad en la estimación de la importancia de ciertas características y evitar el sobre ajuste. XGBoost: utiliza una serie de árboles de decisión construidos de forma secuencial, donde cada árbol nuevo intenta corregir los errores cometidos por los árboles anteriores. Es un modelo con alta precisión, rendimiento y velocidad, capaz de trabajar con datos tanto numéricos como categóricos. SVC (Clasificación de vectores de soporte): este modelo de aprendizaje au- tomático supervisado separa los puntos de datos mediante un hiperplano con la mayor cantidad de margen posible. Este margen es la distancia en- tre los dos puntos pertenecientes a clases distintas más cercanas. Es un algoritmo que ofrece alta precisión y gran velocidad [21]. K-means: es una herramienta de aprendizaje no supervisado que agrupa un conjunto de datos en K grupos (clusters). Cada grupo tiene un repre- sentante llamado centroide, que es la media aritmética de los elementos que pertenecen al cluster y, este algoritmo, actúa de manera iterativa de tal forma que cada elemento esté más cerca de su centroide que de los centroides del resto de grupos [22]. DBSCAN: es un algoritmo de clustering que consiste en que para cada ob- servación se mira el número de puntos que se encuentran a una distancia máxima épsilon de ella, denominados vecinos. Si un punto tiene al menos un mínimo número de vecinos, se

denota como observación central y, los vecinos de este punto pertenecen al mismo cluster. Una observación que no es central y que no sea vecina de ninguna central es una anomalía o valor atípico [23]. Clustering jerárquico: es un algoritmo de aprendizaje no supervisado que agrupa los datos en función de la distancia entre ellos. Para su explicación se utilizará un dendrograma. En este los datos de partida son A, B, C, D, E, F, G, H y se muestra como se van relacionando los elementos dos a dos. Figura 2.3: Clustering jerárquico. Este muestra en qué orden se han ido realizando las agrupaciones hasta obtener un solo cluster, a continuación se elige un nivel y se podan las hojas de nivel igual o superior a este, los nodos hoja resultantes son los clusters. Figura 2.4: Poda en el nivel 3. 2.2. Técnicas y herramientas 2.2.4. PostgreSQL Es un motor de bases de datos relacional de código abierto. Destaca por su confiabilidad y robustez. Sigue un modelo relacional y es altamente escalable, es decir, permite manejar grandes volúmenes de datos. Este ha sido el motor elegido para almacenar la información del proyecto. 2.2.5. Docker Docker es un software que permite crear y probar aplicaciones rápidamente. Empaqueta unidades de software en estructuras llamadas contenedores dentro de las cuales se encuentran todas las herramientas necesarias para la ejecución de aplicaciones. En este proyecto ha sido utilizado para levantar la base de datos PostgreSQL donde se almacenan los datos utilizados. 2.2.6. Web scraping Este término hace referencia al proceso necesario para extraer contenidos e in- formación de sitios web mediante software de manera automatizada. Por ejem- plo, los comparadores de precios de vuelos y hoteles utilizan web scraping para rastrear varias páginas web en busca de la información demandada por el usua- rio. Hoy en día es posible rastrear todo tipo de datos en la web, desde feeds en RSS hasta información gubernamental, pero esta información no siempre se alcanza fácilmente. Dependiendo de la web, el rastreo se puede hacer usando APIs u otras herramientas [24]. Con el web scraping se extrae el contenido HTML de las webs para filtrar la información requerida y almacenarla. Para ello, en este proyecto se obtiene el HTML mediante el uso de la librería request mencionada anteriormente y se filtra esta información con las herramientas de la librería re. Capítulo 3 Desarrollo 3.1. Metodología Para llevar a cabo el proyecto se han establecido las siguientes fases o etapas. Extracción de datos, limpieza, preprocesamiento, almacenamiento, análisis y visualización. A continuación se describe el objetivo de cada fase. Extracción de datos: en esta primera fase del trabajo el objetivo es encon- trar la información necesaria para llevar a cabo el proyecto. En esta etapa se identificarán las fuentes de datos disponibles como pueden ser datos en la nube, conjuntos de datos públicos o APIs externas con acceso a datos en tiempo real. Antes de descargar un conjunto de datos se ha de comprobar la integridad, consistencia y precisión de estos así como su actualidad. A continuación se ha de evaluar la relevancia de los datos para el objetivo del proyecto, su capacidad para proporcionar información valiosa y la relación que tienen entre ellos. Una vez identificadas las fuentes de datos a utilizar, se ha de iniciar el proceso de descarga mediante métodos como el web scrapping, la conexión a APIs o la descarga directa de datos disponibles en repositorios en línea. Por último, se crearán ejecutables automatizados para realizar la descarga, configurar conexiones seguras con las posibles bases de datos o implemen- tar procesos de extracción. Limpieza: el objetivo de esta fase es garantizar la integridad, consistencia y calidad del conjunto de datos. Se han de identificar, corregir y eliminar posibles inconsistencias, errores y datos irrelevantes o duplicados. Este proceso contribuirá a mejorar la precisión de los resultados, optimizará el rendimiento de las soluciones analíticas y proporcionará una buena base para la toma de decisiones informadas. Preprocesamiento: esta fase es crucial para el análisis pues dota a los dis- tintos conjuntos de datos de la estructura adecuada para poder ser cruza- dos correctamente. Esto puede incluir la selección de características rele- vantes, la normalización de datos, la codificación de variables categóricas, Capítulo 3. Desarrollo la reducción de dimensionalidad y la división de datos en conjuntos de entrenamiento y prueba. Almacenamiento: en esta etapa, se determina el método más apropiado para almacenar los datos de manera eficiente y segura, teniendo en cuenta factores como el volumen de datos, la frecuencia de acceso y la necesidad de mantener la integridad y la confidencialidad de los mismos. Las opciones de almacenamiento pueden incluir bases de datos relaciona- les, sistemas de gestión de bases de datos NoSQL, sistemas de archivos distribuidos o almacenamiento en la nube. Se deben considerar aspectos como la escalabilidad, la disponibilidad y el rendimiento del sistema de almacenamiento seleccionado. Análisis: en esta etapa, se aplican técnicas y algoritmos analíticos para extraer información significativa de los datos preprocesados. Esto puede incluir la identificación de patrones, tendencias y relaciones en los datos, así como la construcción de modelos predictivos o descriptivos. Visualización: la visualización de datos es una herramienta fundamental a lo largo de todo el proyecto. Ayuda a entender los datos con los que se tra- baja y su distribución. También es fundamental para el proceso de análisis, ya que permite comunicar de manera efectiva los hallazgos y resultados. Se utilizan diversas técnicas y herramientas de visualización para represen- tar gráficamente los datos y facilitar su interpretación. Esto puede incluir gráficos de barras, diagramas de dispersión, mapas de calor y gráficos de líneas, entre otros. Una visualización efectiva puede ayudar a identificar patrones, tendencias y anomalías en los datos, así como a comunicar de manera clara y concisa los resultados del análisis. Aunque se haya representado como última etapa, la visualización es una herramienta que se utilizará a lo largo de todo el proyecto. Figura 3.1: Fases del proyecto. 3.2. Extracción de datos El proceso de extracción de datos es crucial pues sienta las bases para el resto del proyecto. La obtención de datos relevantes y confiables es el punto de partida 3.2. Extracción de datos para llevar a cabo el análisis. Se ha buscado en numerosas páginas de datos abiertos sobre el transporte pú- blico de Madrid. En ellas se encuentra una amplia gama de datos y, tras hacer una búsqueda exhaustiva, se ha decidido utilizar información del año 2022 pues es el año del cual se ha encontrado más información relevante completa. Del año 2023 hay información faltante de los últimos meses como noviembre o diciem- bre, y para poder hacer un estudio completo anual se han tomado los datos del año 2022. A continuación se describe la información encontrada y el procedi- miento utilizado para extraerla. Datos Madrid: gracias a la página datos.madrid.es se ha encontrado in- formación sobre accidentes de tráfico, accidentes de bicicleta y aforos de tráfico. Para obtener esta información se ha creado el archivo ”descar_csvs.py”. Es- te fichero descarga los siguientes archivos de la web en formato CSV: Aforos de tráfico en la ciudad de Madrid permanentes - Portal de datos abiertos del Ayuntamiento de Madrid, Accidentes de tráfico con implicación de bicicle- tas - Portal de datos abiertos del Ayuntamiento de Madrid y Accidentes de tráfico de la Ciudad de Madrid - Portal de datos abiertos del Ayuntamiento de Madrid. Este ejecutable trata dos escenarios, las webs que tienen la información mensual y las webs que contienen la información anual. En ambos casos el primer paso es el siguiente: se utiliza web scraping, se hace una llamada a la página web; después se parsea el HTML obtenido y así se consiguen los enlaces de descarga de los archivos. A continuación, se realiza un filtrado de los href para obtener los datos o bien mensuales o anuales del 2022. Estos se utilizan para hacer la petición de cada uno y así descargar todos los archivos. Se obtienen dos archivos anuales: “AccidentesBicicletas_2022.csv” y “Acci- dentesTrafico_2022.csv”. Ambos contienen información sobre los acciden- tes registrados en 2022 en Madrid. El primero consta de 877 entradas y recoge los accidentes con implica- ción de bicicletas y el segundo tiene 47.052 y data los accidentes de tráfico con implicación de vehículos, sin contener los accidentes entre bicicleta- bicicleta y bicicleta-peatón. Por ello nos interesa tener ambos. A continua- ción se muestra en la Tabla 1 los campos de cada registro de los ficheros y un ejemplo en la Figura 2. CAMPO DESCRIPCIÓN num_expediente Identificador fecha Fecha del accidente hora Hora del accidente localización Calle del accidente numero Número de la calle distrito Distrito del accidente cod_distrito Numeración del distrito tipo_accidente Alcance, atropello... estado_meteorológico Tiempo tipo_vehiculo Tipo vehículo tipo_persona Conductor, pasajero rango_edad Edad sexo Hombre o Mujer lesividad Lesividad cod_lesividad Numeración de la lesividad coordenada_x_utm Latitud coordenada_y_utm Longitud positiva_alcohol Alcohol positiva_droga Droga Unassigned:19 Columna vacía Unassigned:20 Columna vacía Cuadro 3.1: Campos y descripción 2022S000034; 02/01/2022; 0:05:00; CALL. MARIA TERESA SAENZ DE HE- REDIA, 6; 6; 15; CIUDAD LINEAL; Caída; Despejado; Bicicleta EPAC (pedaleo asistido); Conductor; De 25 a 29 años; Hombre; 7; Asistencia sanitaria solo en el lugar del accidente; 444.462.918; 4.474.808.752; S; NULL Figura 3.2: Ejemplo del contenido de la tabla. 3.2. Extracción de datos En cuanto al aforo, en la web se encuentran los ficheros mensuales y, por tanto, se obtienen 12 archivos que se unirán en un único CSV anual llamado “AforosTrafico_2022.csv” eliminando los mensuales. Este archivo contiene la información de la cantidad de vehículos que pasan por deter- minados lugares a lo largo del día y hora a hora los 365 días del año 2022. Consta de 292.430 entradas antes de realizar la limpieza de datos. A con- tinuación se muestra en la Tabla 2 los campos de cada registro del fichero y un ejemplo en la Figura 3. CAMPO DESCRIPCIÓN FDIA Día FEST Calle donde se mide el aforo FSEN Sentido de la calle, AM/PM HOR1 Hora 1 y 13 HOR2 Hora 2 y 14 HOR3 Hora 3 y 15 HOR4 Hora 4 y 16 HOR5 Hora 5 y 17 HOR6 Hora 6 y 18 HOR7 Hora 7 y 19 HOR8 Hora 8 y 20 HOR9 Hora 9 y 21 HOR10 Hora 10 y 22 HOR11 Hora 11 y 23 HOR12 Hora 12 y 24 Unnamed: 15 Columna vacía Cuadro 3.2: Campos y descripción. 02/12/22; ES32; 2-; 271; 145; 119; 97; 70; 72; 125; 158; 285; 317; 341; 390; Figura 3.3: Ejemplo del contenido de la tabla. Al descargar los datos del aforo en formato CSV la ubicación geográfica de los sensores (FEST) de tráfico no aparece, solo había un id para distinguir cada estación, por tanto, se necesita un archivo que identifique cada sensor con su ubicación. Los archivos de aforo mensuales en formato Excel con- tienen 3 páginas y en una de ellas se encuentra la información deseada, por tanto, se escogió uno, noviembre de 2022 y con el ejecutable ”descar- ga_excel.py”: y pandas se iteró por las páginas del archivo hasta dar con “Ubicación estaciones”. Se convirtió esa hoja a un dataframe para luego transformarlo a un CSV llamado “EstacionesTrafico_2022.csv”. Este tiene 120 entradas, 2 por cada una de las 60 estaciones registradas, dependien- do de la orientación del tráfico (norte, sur, este, oeste). A continuación se muestra en la Tabla 3 los campos de cada registro del fichero y un ejemplo en la Figura 4. CAMPO DESCRIPCIÓN Estación Id de la estación Nombre Calle donde se mide el aforo Latitud Latitud Longitud Longitud Sentido Orientación codificada Orientación N-S S-N E-O O-E Cuadro 3.3: Campos y descripción 1,Paseo de la Castellana,"40,4319272588958",3,68910874956933",1.0,S-N Figura 3.4: Ejemplo del contenido de la tabla Opendatasoft: Gracias a la página opendatasoft.com se ha encontrado el contorno de la zona centro de Madrid en formato JSON y el contorno de la Comunidad de Madrid en formato GEOJSON. Esta información será útil para poder visualizar los datos y entender mejor su ubicación en el ma- pa. Para obtener estos datos se ha creado el archivo ”descarga_json.py”. Al igual que en el caso anterior, se hace una petición a la página web gracias a su href y la respuesta se guarda en ”ContornoMadrid.json” y ”ContronoC- Madrid.geojson” respectivamente. 3.3. Limpieza Esta etapa consiste en encontrar y corregir errores o inconsistencias mediante la eliminación de errores o duplicados, correcciones ortográficas o tratamiento de inconsistencias. En primer lugar, en todos los archivos se ha tomado la decisión de eliminar tildes, cambiar la letra ‘ñ’ por la ‘n’ y suprimir las diéresis para evitar problemas a la hora de almacenar los datos. También se ha establecido como separador de cada CSV el punto y coma (;) para tener una estructura unificada entre todos los archivos. A continuación, se describen los cambios realizados específicos de cada archivo. AforosTrafico: en primer lugar, en el archivo aparecían numerosas filas cuyo único contenido era un punto y coma (;) y, por tanto, se eliminaron, quedando así 86.140 filas. Por otro lado, el CSV tiene dos sentidos por cada calle y las calles de un solo sentido tenían una fila de ceros en el sentido inexistente, por tanto, se han eliminado las filas nulas de aquellas calles de un solo sentido, quedando así 79.498 entradas. Además, se ha eliminado la columna ‘Unnamed:15’ que se encontraba vacía. Accidentes: se ha realizado la limpieza de los archivos ”AccidentesBicicle- tas_2022.csv” y ”AccidentesTrafico_2022.csv” simultáneamente con el ob- jetivo de unirlos en un solo CSV. En accidentes de tráfico hay dos columnas que reciben el nombre de ‘Unassigned:19’ y ‘Unassigned:20’, como no ofre- cen ninguna información han sido eliminadas, teniendo así ambos archivos 3.4. Preprocesamiento el mismo número de columnas. Las columnas ’coordenada_x_utm’ y ’coordenada_y_utm’ de ”AccidentesBi- cicletas_2022.csv” aparecían en un formato extraño, con los números se- parados de 3 en 3 mediante puntos, como si fuera un número entero. Por otro lado, los mismos campos en ”AccidentesTráfico_2022” aparecían como un número decimal con 3 decimales separados por una coma (,). En ambos archivos se ha unificado el formato de estos campos a WGS (World Geodetic System) similar al de los sistemas GPS. Además, han sido renombrados a ’Latitud’ y ’Longitud’. Por último, se han unificado los ficheros mediante el ’num_expediente’ de cada accidente y eliminado las filas duplicadas, quedando 47.046 (solo 7 casos repetidos). DatosEstaciones: en este archivo los campos latitud y longitud tenían como indicador decimal una coma (,) y se ha sustituido por un punto para que esté en el mismo formato que en el resto de archivos. 3.4. Preprocesamiento Para llevar a cabo la fase de análisis, toda la información ha de estar bien prepa- rada para poder ser tratada por los modelos. Para ello se han de realizar ciertas modificaciones en ”EstacionesTrafico_2022”, ”AforosTrafico_2022” y ”Acciden- tesTraffico_2022”. En primer lugar, se ha modificado la estructura del dataset ”AforosTrafico_2022”. En lugar de tener un campo por cada hora del día, se ha creado uno solo llama- do ’hora’ y otro ’aforo’ con el aforo correspondiente a la hora del día. Además, se han renombrado los campos ’FDIA’, ’FEST’, ’FSEN’. En la Tabla 4 se muestra como ha quedado el CSV y un ejemplo de este en la Figura 5. CAMPO DESCRIPCIÓN fecha Fecha estacion Id de la estación sentido Sentido de la calle hora Hora del día aforo Número de vehículos Cuadro 3.4: Campos y descripción 01/01/2022; 1; 1; 19; 1266.0 Figura 3.5: Ejemplo del cambio realizado A continuación, en este mismo archivo, ’estacion’ se corresponde con el ’id’ de ”EstacionesTrafico_2022.csv” por lo tanto, para poder unir mediante ese ID, am- bos datos

han de ser del mismo tipo, 'estacion' es de tipo String y se ha convertido a tipo int. A continuación se muestra en la Figura 6 un ejemplo del tipo de dato antes del preprocesamiento y después de este. Antes del preprocesado, estacion: ES01, después, estacion: 1 Figura 3.6: Ejemplo del cambio realizado Por otro lado, también se han realizado modificaciones en "EstacionesTrafi- co_2022.csv". Se han renombrado todos los campos poniendo en minúscula los nombres para unificar el formato con el del resto de archivos. "AccidentesTrafico_2022.csv" ha sido el archivo en el cual se han realizado más cambios. En el campo 'estado_metereológico' había 5.285 entradas nulas, en 'tipo_vehiculo' 199, en 'lesividad' y 'cod_lesividad' 22. Ante esta falta de datos se tomó la de- cisión de imputarlos y para elegir el método de imputación se observaron las distribuciones de las 4 categorías en histogramas. Las 4 eran asimétricas y con una categoría claramente superior al resto, por lo tanto, se imputó con la moda. En el caso de 'estado_meteorológico' la moda era "Despejado", en 'tipo_vehiculo' era "turismo" y en 'lesividad' y 'cod_lesividad' era " Sin asistencia sanitaria" con su código correspondiente que es el 14. Figura 3.7: Histograma estado meteorológico Figura 3.8: Histograma tipo de vehículo Figura 3.9: Histograma lesividad De cara al análisis, a efectos prácticos, en el campo de lesividad el interés es saber si en un accidente se requiere la asistencia de medios sanitarios y, por tanto, su despliegue. Es por esto que se han creado dos clases, asignando un 0 a aquellos accidentes que no han requerido asistencia y un 1 a aquellos que si. Por otro lado, en el campo 'rango_edad' no había entradas nulas, pero sí había 5.179 campos donde ponía "desconocido". Este ejemplo muestra la importancia de hacer un análisis exhaustivo preliminar, pues aunque a priori esa informa- ción parecía estar completa, no lo estaba. Se ha hecho el mismo procedimiento que en el caso anterior, se ha visto la distribución de la edad y al ver que era uni- forme se han imputado los valores desconocidos con la media, que se encuentra en el rango 'De 40 a 44 años'. Figura 3.10: Histograma rango de edad A continuación, el campo 'hora' proporciona información sobre la hora y el mi- nuto del accidente; puesto que en los aforos tenemos la información por hora, se modificó este campo conservando solo esta, eliminando así los minutos. Respecto al tipo de vehículo implicado en el accidente, se van a conservar to- dos. Existen 31 tipos en el CSV, para guardar esta información se ha decidido agrupar en categorías y crear una columna con cada una de ellas de tipo ente- ro, que indica en número de vehículos de ese tipo implicado en el accidente. A continuación se muestra la norma aplicada para realizar la agrupación. Turismo: turismo, VMU eléctrico, todoterreno. Bicicleta: bicicleta EPAC, bicicleta, patinete no eléctrico, otros vehículos sin motor. Moto: motocicleta hasta 125cc, motocicleta >125cc, ciclomotor, cuadriciclo ligero, cuadriciclo no ligero, ciclo, moto de tres ruedas >125cc, ciclomotor de dos ruedas L1e-B, moto de tres ruedas hasta 125cc. Furgoneta: furgoneta, autocaravana. Camión: camión rígido, vehículo articulado, tractocamión, semirremolque, remolque. Otros: camión de bomberos, otros vehículos con motor, ambulancia SA- MUR, maquinaria de obras. Los datos agrupados en la categoría "otros" podrían ser considerados datos atí- picos y podrían ser eliminados, pero dado que el dataset no contiene un número de entradas considerablemente alto y esta categoría tiene 753 apariciones, se ha preferido preservar esta información. En cuanto al tipo de accidente, también aparecen las distintas entradas con la siguiente frecuencia: 'Alcance' 4.089, 'Atropello a persona' 1403, 'Caída' 1.992, 'Choque contra obstáculo fijo' 3.073, 'Colisión frontal' 494, 'Colisión fronto- lateral' 4.690, 'Colisión lateral' 2.916, 'Colisión múltiple' 795, 'Otro' 753, "Des- peñamiento" 1, 'Solo salida de la via' 131, 'Vuelco' 115 y 'Atropello a animal' 82 veces. La entrada con el despeñamiento se ha eliminado debido a que solo aparece una vez y puede ser considerado dato atípico. Por otro lado, los 3 campos con menos frecuencia, es decir, 'solo salida de la vía', 'vuelco' y 'atropello animal', se han agrupado en la categoría 'Otros'. Una vez terminadas las modificaciones campo por campo, vemos que en el da- taset inicialmente aparecía una entrada por cada persona implicada en un ac- cidente. El interés era tener una entrada por cada accidente y para ello se ha agrupado la información de las personas afectadas (sexo, rango_edad, lesividad, positiva_alcohol). La información general del accidente, es decir, la recogida en los campos ' fecha', 'hora', 'localización', 'numero', 'distrito', 'cod_distrito', 'esta- do_metereológico', 'tipo_accidente', 'latitud' y 'longitud' han permanecido intac- tos. Se ha creado un campo 'personas' en el que se recoge el número de estas implicadas en el accidente para no perder esta información. Por otro lado, se han tomado las siguientes decisiones respecto a los campos individuales de ca- da persona. 'positiva_droga' y 'positiva_alcohol': se han agrupado positivos en alcohol y drogas pues el número de campos nulos en 'positiva_droga' era considera- ble (solo 140 campos no nulos). En el caso de que uno de los involucrados en el accidente diera positivo en drogas o alcohol, aparece un 1 en este campo. 'lesividad' y 'codigo_lesividad': en estos campos se recoge si alguno de los afectados necesitó asistencia sanitaria o no en el accidente, por tanto, se guarda la información del mayor afectado. 'rango_edad': en este campo se ha conservado el rango de la persona con mayor edad, entendiendo que cuanto mayor sea una persona más grave- mente puede afectarle un accidente. Por último, se ha creado un campo 'geometry' con la combinación de las coor- denadas latitud y longitud, que facilitará la visualización de los accidentes. A continuación se muestran los campos finales del archivo y un ejemplo de este. CAMPO DESCRIPCIÓN num_expediente Identificador fecha Fecha del accidente localización Calle del accidente numero Número de la calle distrito Distrito del accidente cod_distrito Numeración del distrito tipo_accidente Tipo de accidente estado_meteorológico Tiempo rango_edad Edad del mayor afectado personas Numero de personas implicadas cod_lesividad Asistencia sanitaria del mayor afectado latitud Latitud longitud Longitud positiva_droga (Longitud, Latitud) turismo Cantidad implicada en el accidente autobús Cantidad implicada en el accidente bicicleta Cantidad implicada en el accidente camión Cantidad implicada en el accidente furgoneta Cantidad implicada en el accidente motocicleta Cantidad implicada en el accidente otros Cantidad implicada en el accidente geometry Alcohol y drogas hora Hora del accidente Cuadro 3.5: Campos y descripción 2022S000001; 01/01/2022; AVDA. ALBUFERA, 19; 19; 13; PUENTE DE VA- LLECAS; 0; 0; 49.0; 2; -3.667437045740308; 40.39741992179293; 0; 0; 0; 0; 0; 0; 0; 0; 2; 1; POINT (-3.667437045740308 40.39741992179293) Figura 3.11: Ejemplo del contenido de la tabla 3.5. Almacenamiento El almacenamiento es fundamental para gestionar grandes volúmenes de datos de manera eficiente y escalable. Tras explorar las distintas opciones disponi- bles, se optó por levantar una base de datos PostgreSQL en un docker. Esta opción ofrece varias ventajas como evitar problemas de incompatibilidades en- tre el sistema operativo y aplicaciones. Además, encapsula la base de datos y sus dependencias en un contenedor, lo que garantiza un entorno de desarrollo consistente. La creación y levantamiento es un proceso rápido y sencillo que se realiza a través de dos comandos de docker y la conexión se realiza gracias a sqlalchemy y su opción create_engine. Gracias al método tosql se pueden subir los datos en escasos segundos. Además, los contenedores de docker ofrecen aislamiento de recursos, es decir, la base de datos se ejecuta en un entorno aislado y seguro, garantizadno que los procesos en ejecución dentro del contenedor no afecten a otros servicios y aplicaciones del sistema. 3.6. Análisis En esta sección primero se va a realizar una exploración de los datos no solo para comprender el contenido de estos, sino para empezar a ver las principales tendencias de los conjuntos e identificar información relevante para su estudio. Una vez identificada, se aplicarán técnicas de aprendizaje supervisado y no su- pervisado para obtener información relevante de los datos. Comenzamos con el CSV de accidentes. En primer lugar, gracias a la matriz de correlación nos podemos hacer una idea inicial de los datos que se tienen y la relación entre ellos. Esta, muestra infor- mación que puede parecer obvia a priori, como por ejemplo: El número de personas implicadas en un accidente tiene alta relación con accidentes en los que hay turismos involucrados, autobuses, a continua- ción furgonetas y con el vehículo que menos es con motos. El rango de edad está altamente relacionado con el tipo de vehículo 'turis- mo', es decir, el vehículo elegido por las personas mayores en primer lugar es el turismo y luego el autobús, y con el que tiene menos relación es con la moto. El número de positivos en droga tiene gran relación con la hora del día, siendo más alta cuanto más temprana es la hora, es decir, de madrugada. El vehículo más afectado por el estado meteorológico es la moto. Se entien- de que es el vehículo en el que tanto conductor como pasajero son más afectados. El vehículo que tiene mayor lesividad es la motocicleta y después la bicicle- ta. Aunque la bicicleta sea más frágil que la motocicleta, esa alta relación tiene que ver con un mayor número de accidentes con motocicletas y tam- 3.6. Análisis bién con las altas velocidades que alcanzan estas en comparación con las bicicletas, lo que hace que el accidente pueda ser más peligroso a pesar de que la persona es menos vulnerable. El vehículo que tiene menor relación con la lesividad es el turismo, luego el camión y la furgoneta, es decir, son los más seguros. Otras observaciones no tan obvias son las siguientes: Los casos de positivo en drogas tienen mayor relación con los accidentes con turismo implicados y con el tipo de vehículo con el que tienen menor relación es con la moto. Los casos en el que el vehículo es 'otro' tienen muy baja relación con todos los campos. Entendemos que esto se debe a que son casos muy aislados como accidentes con camiones de bomberos o tractores. El tipo de accidente tiene muy baja correlación con los tipos de vehículo. Los accidentes con motocicletas tienen muy poca relación con los acciden- tes en los que hay turismos implicados. Figura 3.12: Matriz de correlación de accidentes. A partir de aquí se entiende que se podrán hacer predicciones sobre variables como lesividad o el tipo de accidente. Por otro lado, la información geoespacial como el distrito será susceptible a técnicas de aprendizaje no supervisado para encontrar ubicaciones con las mismas características de accidentes. Respecto al CSV de aforos cuya única información es el número de vehículos, la ubicación y la hora, se podrían hacer predicciones espacio temporales, pero es- tas pueden ser complicadas y, en muchos casos, no son precisas debido a varios factores. En primer lugar, los datos temporales presentan patrones complejos y no lineales y, por tanto, modelarlos suele resultar inexacto. Además, son alta- mente sensibles a los datos de entrada y pequeñas variaciones en la información pueden afectar significativamente a los resultados. Por otro lado, a pesar de que se dispone de un volumen no muy reducido de datos, las predicciones tempora- les pueden verse afectadas por cambios en el entorno como eventos inesperados o tendencias emergentes (una obra, un festivo). Es por todo esto que a este ar- chivo solo se le aplicarán técnicas de aprendizaje no supervisado, para intentar de esta forma obtener patrones y encontrar estaciones de medición de aforo con las mismas características. 3.6.1. Aprendizaje supervisado En este apartado el objetivo es elaborar modelos predictivos sobre las variables vistas anteriormente. Dados los tipos de datos que tenemos, nos encontramos ante problemas de cla- sificación como la edad, que viene dada en rangos 'De 40 a 44 años' o la lesividad (asistencia necesaria o no). En este apartado se buscan los modelos predictivos que mejor ajusten estas variables. Para evaluar los modelos se utilizarán las métricas de accuracy (exactitud), pre- cision (precisión), recall (recuerdo/memoria) y F1-score. Todas estas métricas se obtienen gracias a la matriz de confusión, que es una matriz que representa la calidad de un modelo. Tiene la siguiente estructura: Figura 3.13: Matriz de confusión Las filas representan lo que el modelo debería predecir y las columnas lo que ha predicho. De esta forma, en la diagonal principal se encuentran los casos correc- tamente clasificados y en la anti-diagonal los erróneos; siendo un falso positivo un caso negativo que el modelo clasifica como positivo y un falso negativo, un caso positivo que el modelo clasifica como negativo [25]. A partir de estos da- tos se pueden calcular las métricas mencionadas anteriormente de la siguiente forma: Accuracy : es la métrica más usada y se define como la cantidad de veces que se acertó una afirmación sobre el total de los datos. Es decir, toma los elementos de la diagonal principal en relación con el total. Figura 3.14: Matriz de confusión para accuracy Precision: esta métrica mide la cantidad de verdaderos positivos frente al total de positivos predichos, es decir, toma los verdaderos positivos sobre el total de la segunda columna. Figura 3.15: Matriz de confusión para precisión Recall : esta métrica compara la cantidad de verdaderos positivos sobre lo que realmente era positivo, es decir, toma los verdaderos positivos sobre el total de la segunda fila. Figura 3.16: Matriz de confusión para recall F1-score: es el doble del producto de la precisión por el recall entre la suma de estos. Esta métrica es un indicador de alta precisión y alta sensibilidad. F 1 = 2 × (Precision + recall) (Precision × recall) Una vez conocemos las métricas para evaluar los modelos, empezamos a entre- narlos. 1. Positivo en droga: el caso positivo tiene asociado como etiqueta un 1 y el ne- gativo un 0. Para abordar la predicción de esta situación se han probado 5 modelos: Regresión Lineal, Bosque aleatorio, Árboles de decisión, XGBoost y SVM. Antes de comenzar a aplicar los modelos hay que entender que este campo se encuentra considerablemente desbalanceado, esto quiere decir que las clases positiva y negativa en drogas no están representadas equita- tivamente, en este caso 3.709 negativos y 333 positivos. Este desequilibrio genera un sesgo significativo en los resultados de los modelos, lo que hace necesario un ajuste preciso para minimizar su impacto. En los primeros entrenamientos de los modelos se obtuvieron resultados no muy precisos y, tras realizar un ajuste de los hiperparámetros con técnicas como Grid Search, se obtuvieron los siguientes resultados. Modelo Valor Precision Recall F1-score Accuracy Regresión 0 0.94 0.87 0.9 0.82 1 0.20 0.39 0.27 0.82 Bosque aleatorio 0 0.92 1.00 0.96 0.92 1 0.52 0.03 0.09 0.92 Árbol de decisión 0 0.94 0.88 0.92 0.85 1 0.24 0.39 0.30 0.85 XGBoost 0 0.97 0.79 0.87 0.78 1 0.23 0.72 0.35 0.78 SVM 0 0.95 0.64 0.71 0.64 1 0.13 0.62 0.22 0.64 Cuadro 3.6: Resultados de los modelos aplicados En primer lugar, se puede apreciar que la precisión de la predicción del valor negativo es muy alta en todos los modelos en comparación con la predicción del valor positivo, esto se debe al desbalance de los datos, sin embargo, este hecho puede resultar engañoso. Se observa que el modelo con mayor accuracy es el de bosque aleatorio, pero eso no significa que sea el más adecuado para nuestro caso. Este modelo solo clasifica correc- tamente el 0.03 % de los casos positivos, por lo tanto, queda descartado como mejor opción. Hemos de buscar un modelo con la mayor precisión posible en la predicción de positivos y con un buen equilibrio o recall. Este es el caso del XGBoost, que es el

modelo que más casos positivos clasificó correctamente. En escenarios como este, los falsos positivos y negativos tienen implicacio- nes prácticas significativas. Los falsos positivos pueden resultar en gastos adicionales asociados con pruebas y análisis de drogas, mientras que los falsos negativos pueden comprometer la justicia y la seguridad pública al no identificar correctamente a los responsables de los accidentes. Por lo tanto, es esencial seleccionar un modelo que minimice tanto los falsos po- sitivos como los falsos negativos. 2. Lesividad: este problema se trata de igual forma que el anterior, consiste en ver si se requiere asistencia sanitaria en el accidente o no. En este caso el problema está algo más balanceado pues hay 2.755 casos en los que no se necesitó asistencia y 1.286 en los que si. Se han probado los mismos 5 modelos que en el apartado anterior y las mismas métricas para evaluar el rendimiento del modelo, obteniendo los siguientes resultados. Modelo Valor Precision recall F1-score Accuracy Regresión 0 0.85 0.87 0.86 0.80 1 0.70 0.66 0.68 0.80 Bosque aleatorio 0 0.86 0.90 0.88 0.83 1 0.75 0.68 0.71 0.83 Árbol de decisión 0 0.85 0.90 0.87 0.82 1 0.75 0.66 0.70 0.82 XGBoost 0 0.86 0.89 0.87 0.82 1 0.74 0.69 0.71 0.82 SVM 0 0.84 0.87 0.86 0.80 1 0.70 0.66 0.68 0.80 Cuadro 3.7: Resultados de los modelos aplicados En este caso se puede observar que tanto los valores obtenidos en accuracy como en recall son muy parecidos en todos los modelos. Cabe destacar la importancia del recall pues este indica, de todos los accidentes en los que se necesita una ambulancia, en cuantos se ha predicho la necesidad de esta. Es mejor enviar una ambulancia y que no haga falta (alta precisión) a no mandarla cuando si es necesaria (alto recall). En este caso se pueden elegir dos modelos, el bosque aleatorio y el XGBoost que han sido los modelos con mejor desempeño. XGBoost tiene el recall en el caso positivo ligeramente superior al bosque aleatorio, pero este último tiene mejor recall en el caso negativo. Estos resultados se deben a que ambos manejan bien el desbalance de clases mediante el ajuste de pesos de clases, que se refleja en mayor preci- sión y mejores puntuaciones en F1-score. Ambos modelos pueden capturar interacciones complejas entre características que podrían no ser captadas con modelos lineales como regresión logística. Respecto al resto de categorías como la edad o el tipo de vehículo que se co- rresponden con ejercicios multiclase, se ha intentado hacer predicciones, pero debido al desbalance de estas clases y las pocas ocurrencias del algunas de ellas, las predicciones modeladas no han obtenido una precisión muy elevada, rondando entre el 30 %-40 %. Aquellas clases con mayor número de apariciones tenían una precisión, recall y F1-score razonables, pero al ponerlas en conjunto con las clases menos comunes. el rendimiento del modelo disminuye considera- blemente. En el apartado de preprocesamiento recordemos que se había llevado a cabo una agrupación de distintas categorías en una sola, como era el caso del tipo de vehículo o el tipo de accidente. Una posibilidad, ante esta situación, sería hacer agrupaciones en categorías con más elementos para que, así, cada clase tenga mayor número de apariciones. Por otro lado, si se realiza esta agrupa- ción, tenemos la desventaja de que se perdería mucha información pues, al ser categorías más amplias, perderíamos la distinción entre los elementos pertene- cientes a estas, dando lugar a una pérdida considerable de información y una generalización que puede resultar en pérdida de utilidad de las predicciones, a pesar de un posible mejor rendimiento. 3.6.2. Aprendizaje no supervisado En este apartado, el objetivo será aplicar técnicas de clustering para agrupar la información en conjuntos con características similares y así entender mejor como se distribuyen tanto el tráfico como los accidentes en el espacio y en el tiempo. Aforos En este CSV se intenta agrupar las ubicaciones para entender cuáles tienen ca- racterísticas similares. Las métricas que se han escogido para realizar la agru- pación son la media y la varianza. La media ofrece una medida representativa del tráfico habitual en cada estación, el volumen de tráfico en ella. La varianza informa sobre la estabilidad del tráfico, una alta variabilidad refleja picos con alto y bajo volumen. Se han aplicado 3 métodos, k-means, DBSCAN y clustering jerárquico. K-means: el primer paso para llevar a cabo este método es elegir el núme- ro de clusters en que se pretende dividir la muestra. Esta elección no es aleatoria y se ha hecho mediante el método del codo. Esta técnica consiste en aplicar el algoritmo k-means con k=1, k=2 y así sucesivamente y calcu- lar en cada k la variación total dentro de los clusters. Esta variación es la suma de las distancias al cuadrado de cada punto al centro de su cluster (centroide). El objetivo es encontrar el punto k donde la suma de las dis- tancias al cuadrado no disminuye significativamente conforme se aumenta el número de clusters. A continuación se muestra el codo aplicado al CSV. Figura 3.17: Método del codo. Como se puede observar no hay un punto en la gráfica donde la disminu- ción de la suma de los cuadrados se aplane significativamente, se podría elegir el 3, el 4 o el 5; por tanto, se ha probado con esos 3 valores. Para decidir cuál de los tres daba mejor resultado se ha utilizado el coeficiente de silueta promedio. Este toma valores entre -1 y 1 donde valores cercanos a -1 indican que los puntos se están asociando a clusters erróneos, valores cercanos a 0 que los clusters se están solapando y valores cercanos a 1 que los puntos están bien asignados a los correspondientes clusters y que estos están bien diferenciados. Tras escalar los datos y aplicar el modelo con los 3 valores de k, se han obtenido valores del coeficiente de silueta muy simi- lares, siendo el mejor para k=5 con un valor de 0.321 y el más bajo para k=4 con un valor de 0.277. A continuación, en la Figura 3.18 se muestran en el mapa los 5 clusters obtenidos. Figura 3.18: Visualización de los clusters obtenidos con k-means. Para entender mejor las 5 agrupaciones y el valor medianamente bajo del coeficiente de silueta, vemos la varianza y media de los centroides de cada uno de los clusters, que recordemos que es el punto representante de cada uno de ellos. Clúster Media Varianza 0 -0.477286 -0.537525 1 -0.281314 -0.216676 2 -0.454528 -0.329515 3 4.117412 4.714556 4 0.578688 0.407684 Cuadro 3.8: Medidas de los centroides de los clústeres. En primer lugar, recordar que los valores son negativos y tan pequeños pues se ha aplicado la estandarización. Aunque a priori todos deberían estar entre 0 y 1, el cluster 3 indica que ha agrupado aquellos valores más dispersos y por ello tanto su media como su varianza tienen valores superiores a 1. Para representar los datos de forma más visual se crea la siguiente gráfica. Figura 3.19: Centroides de los clusters para k=5. Aquí se observa claramente que el cluster 0 recoge aquellas ubicaciones con bajo volumen de coches y menor varianza, es decir, zonas poco concurri- das de forma constante. El cluster 1 recoge las zonas con menos volumen y variabilidad algo más moderada, dentro de que sigue siendo un valor bajo. El cluster 2 agrupa ubicaciones con muy poco volumen y poca variabilidad. El cluster 3 recoge claramente zonas con mucho aforo y mucha variabili- dad, en este grupo se encuentran estaciones de medición como la M-30, lugares donde se explica claramente ese alto volumen que fluctúa consi- derablemente. Por último, el cluster 4 recoge ubicaciones con un volumen moderado y algo de variabilidad. Al ver el centroide de cada cluster se explica el coeficiente de silueta pues se entiende que los clusters 0,1 y 2 corresponden a ubicaciones con carac- terísticas bastante similares. Si se aplica el algoritmo para k=3 se obtiene un coeficiente de silueta de 0.31 y 3 cluster con los siguientes centroides. Figura 3.20: Centroides de los clústeres para k=3. Aquí se observa que el cluster 2 es exactamente igual al cluster 3 ante- rior. Los clusters 0,1 y 4 anteriores han sido agrupados en el cluster 1 y el antiguo cluster 2 es ahora el cluster 0. Se entiende el bajo coeficiente de si- lueta pues ha unido estaciones que tienen un volumen de tráfico y varianza superior a 0 con estaciones cuyos valores están por debajo de 0. DBSCAN: para llevar a cabo el método de DBSCAN se han de escoger dos valores que utilizará el algoritmo, el mínimo número de elementos por clus- ter y la distancia máxima (eps) entre elementos del mismo cluster. El primer dato se puede elegir en función de las características del problema, para el segundo se ha utilizado la distancia al k vecino más cercano. Este método consiste en representar en una gráfica la distancia euclídea de cada punto al k vecino más cercano. Este valor de k se suele hacer que coincida con el mínimo de elementos del cluster, y los valores elegidos para calcular la distancia son de nuevo la media y la varianza. A continuación se muestra en la Fig. 3.21 la gráfica correspondiente a este método con 2 vecinos. Figura 3.21: Método de 2 vecinos cercanos. El punto que se elige como distancia máxima es aquel en el que se observa un pico. En este caso, a partir del valor 55 se ve que la distancia al segundo vecino es considerablemente grande y, por tanto, la distancia del dato que está en posición 55 a su segundo vecino más cercano es la elegida como eps, de esta forma nos aseguramos la densidad de los clusters. El hecho de que la distancia al k vecino más cercano se dispare a partir del dato número 55 sobre un total de 60 estaciones indica que los puntos están muy pegados. Esto puede ser un inconveniente para identificar grupos con diferencias significativas. Al aplicar DBSCAN con estos parámetros se obtienen 4 clusters de 2 y 3 elementos y el resto los considera ruido (se observan en la figura como cluster -1). Además, el coeficiente de silueta es de -0.213. Figura 3.22: Visualización de los cluster obtenidos con DBSCAN. Para entender estos resultados se ha realizado una gráfica de análisis de componentes principales (PCA). Esta técnica transforma las características de los datos combinándolas de manera que se maximice la varianza a lo largo de los ejes. A continuación se muestra la gráfica. Figura 3.23: PCA Como se observa, la mayoría de los datos se encuentran en torno al origen, pero hay dos de ellos que se encuentran considerablemente alejados. Esto indica que difieren significativamente en sus valores de media y varianza respecto a los otros datos. Por otro lado, el PCA 1 tiene una variabilidad de 0 hasta 8 (media) mientras que el PCA 2 de -1 a 1.5 (varianza) lo que muestra que la variable que tiene más impacto es la media. El algoritmo de DBSCAN se basa en la densidad local y la conectividad. Casi la totalidad de los puntos se encuentran significativamente cerca, por tanto, cuando se elige una distancia (eps) superior a la proporcionada por el método de los vecinos, se agrupan todos los elementos en un solo cluster y los dos que están alejados se consideran ruido; cuando se elige un eps inferior o igual se crean clusters unipuntuales o con dos elementos, y el resto se considera ruido. K-means es un método menos sensible a la distribución de los puntos en comparación con DBSCAN, dado que este se basa en la densidad de los puntos, esto justifica la baja calidad de los resultados. Clustering jerárquico: para aplicar este método se ha de elegir también el número de clusters que se desean. Esta elección se hace bajo el mismo criterio que en k-means, con el método del codo, y dado que en el anterior caso se escogieron 5 clusters se ha probado primeramente con este número. Se han obtenido las siguientes agrupaciones con un coeficiente de silueta de 0.37. Figura 3.24: Visualización de los clusters obtenidos con método jerárquico para k=5. A continuación, en la Fig. 3.25 se muestra el dendrograma, donde se puede observar el orden en el que el algoritmo ha ido haciendo agrupaciones hasta obtener un solo cluster. Para obtener esos 5 clusters hay que hacer una poda que se encuentra señalada con una línea horizontal roja entre los niveles 3 y 4, obteniendo así 5 líneas verticales que no se unen, cada una representa un cluster. Sombreados en cajas, se distinguen los elementos que forman cada uno de los cluster. Figura 3.25: Visualización del dendrograma. Por último, en este algoritmo no existen los centroides, pero para hacerse una idea de los elementos de cada cluster se ha calculado la media de las medias y varianzas de los elementos de cada conjunto, para así obtener un representante de cada uno de ellos. Cluster Media Varianza 0 -0.730331 -0.825110 1 0.581823 -0.699909 2 -0.816976 -0.918567 3 1.837240 1.221793 4 0.291890 1.221793 Cuadro 3.9: Medidas de los representantes de cada cluster. Igual que en el ejemplo anterior, lo representamos con una gráfica. Figura 3.26: Representantes de los cluster para k=5. Como podemos observar, los tres primeros clusters tienen medidas muy si- milares. Los 3 se caracterizan por tener bajo volumen y muy baja varianza. Por otro lado, el cluster 3 destaca por tener un alto volumen y varianza y el cluster 4 tiene un volumen moderado y una varianza elevada. Dados estos resultados y la similitud entre las tres primeras agrupaciones, se prueba el algoritmo con 3 y 4 cluster en lugar de 5, para intentar agrupar los 3 primeros. Para el caso de k=4 se obtiene un coeficiente de silueta de 0.48 que refleja una gran mejoría y, a continuación, se muestran los representantes de esos 4 clusters en la Fig. 3.28 y los elementos de cada uno en el mapa en la Fig. 3.27. Figura 3.27: Visualización de los clusters obtenidos con método jerárquico para k=4. Figura 3.28: Representantes de los clusters para k=4 Como se puede observar, el elemento del anterior cluster 3 se ha añadido al cluster 2 y el resto han permanecido intactos. Para el caso de k=3 se obtiene un coeficiente de silueta de 0.62, pero nos encontramos con 3 clusters, dos unipuntuales y otro con el resto del con- junto de datos. Figura 3.29: Visualización de los clusters obtenidos con método jerárquico para k=3. Figura 3.30: Representantes de los clusters para k=3 En este caso se observa que una mejoría del coeficiente de silueta no impli- ca necesariamente mejores resultados en cuanto a su valor. Aunque para k=3 los clusters estén mejor definidos y separados, tener agrupaciones uni- puntuales puede ser no tan útil en la práctica. Con un coeficiente bajo, si se entiende bien la información puede resultar de gran utilidad. Aquellas estaciones con mayor volumen y menor varianza son aquellas con un aforo alto constante y necesitarán alguna medida de transporte para desconges- tionarla (como nuevas carreteras) y aquellas con alta varianza y volumen en un determinado momento, medidas como el aumento de la frecuencia del transporte público en esas horas pico. Además, las zonas donde hay estaciones con bajo volumen y baja varianza podrían ser utilizados como modelo para la construcción de nuevas áreas en las que se busque estabi- lidad en el flujo de tráfico. De este CSV también podemos visualizar las horas del día y los meses más concurridos. A continuación se muestran dos histogramas, uno con la media de aforo por hora (Fig. 3.31) del día y otro con la media de aforo por mes (Fig. 3.32). Figura 3.31: Media de aforo por hora Figura 3.32: Media de aforo mensuales Como se puede observar, las horas del día más tranquilas son las de la ma- drugada. A las 7 comienza una subida hasta alcanzar el pico a las 14, que se mantiene hasta las 20 donde empieza a bajar. Respecto a los meses, ninguno destaca notablemente por un aforo superior, sin embargo, en agosto se observa que disminuye considerablemente el tráfico pues es el mes en el que Madrid se vacía. Del mismo modo, pero no de forma tan marcada, sucede en julio y enero, meses también de carácter vacacional. Accidentes En este archivo, en primer lugar se va a agrupar los accidentes por distritos, para así entender cuáles son aquellos con mayor incidencia. Una vez hecha la agrupación, se va a aplicar el algoritmo k-

means. En este caso, la medida que se va a utilizar como distancia para agrupar los distritos es el número de acci- dentes en cada uno de ellos. A continuación vamos a ver los distritos en función de la cantidad de accidentes que ocurren en cada uno (Fig. 3.33). Podemos ob- servar que el distrito con más accidentes es el barrio de Salamanca y el que menos Vicálvaro. También se ha creado un mapa de calor para visualizar esta información en el territorio (Fig. 3.34). Figura 3.33: Recuento de accidentes por distrito Figura 3.34: Mapa de calor de accidentes Una vez se tiene una idea inicial sobre la distribución de los datos, se aplica el método del codo para determinar el número de clusters en los que hacer la división. En la gráfica del codo, se puede observar que los valores de k para los que la distancia a penas disminuye son k=3 y k=4, por tanto, se aplicará k-means para ambos valores. Figura 3.35: Método del codo Para k=3 se obtiene un coeficiente de silueta de 0.568 y para k=4 toma un valor de 0.594. Ambos resultados son moderadamente altos, indicando una separa- ción razonable y cohesión dentro de los clusters. Se muestran a continuación los clusters obtenidos para el mejor resultado, k=4. Figura 3.36: Visualización de los clusters obtenidos para k=4 Se han calculado los centroides de cada cluster para entender qué elementos forman parte de cada uno de ellos. Cluster N.º accidentes 0 926 1 489 2 1.193 3 1.464 Cuadro 3.10: Medidas de los centroides de los cluster s. De esta forma vemos que el cluster 1 contiene las ubicaciones con menor número de accidentes, los clusters 0 y 2 tienen un número de accidentes moderado y el cluster 3 es aquel con mayor incidencia. A pesar de que los resultados obtenidos son razonablemente buenos, se va a aplicar el clustering jerárquico para entender las relaciones y jerarquías entre las agrupaciones. A continuación se muestra el dendrograma correspondiente a este método (Fig. 3.37), con el cual se ha obtenido un coeficiente de silueta de 0.593. Figura 3.37: Visualización del dendrograma. Con una línea horizontal roja queda señalado el nivel donde se ha realizado la poda, resultando así cuatro líneas verticales que representan cada cluster. Si se recorre el dendrograma de izquierda a derecha, en primer lugar nos aparece el cluster 1, a continuación el 3, seguidamente el 2 y por último el 0; viendo claramente los elementos de cada una de las agrupaciones. Una vez identificadas las ubicaciones más problemáticas se va a realizar un aná- lisis temporal, viendo qué meses y a qué horas hay más frecuencia de accidentes para así, poder contrastarlo con los resultados vistos en el CSV de aforos. Para ello se va a utilizar k-means y se va a hacer una agrupación de los accidentes en función del mes y la hora. De nuevo, el primer paso es realizar el diagrama del codo, en el que podemos observar que los mejores valores son k=3 y k=4. Figura 3.38: Método del codo. Se ha aplicado el algoritmo para ambos valores y se han obtenido resultados muy similares con un coeficiente de silueta de 0.594 para k=3 y 0.59 para k=4. Por tanto, se muestran los clusters obtenidos para k=3. Figura 3.39: Visualización clusters obtenidos para k=3. En la gráfica se ven los 3 clusters en distintos colores y distintos tamaños de círculos representando el número de accidentes, siendo 25 o menor el círculo más pequeño y 150 o superior el círculo más grande. Como se puede observar, el cluster 1 es aquel con menos accidentes, se reparte por todos los meses del año, destacando el mes de agosto, y se sitúan las horas del día desde las 23 hasta las 7 de la mañana aproximadamente. El cluster 0 es el siguiente con más accidentes, se reparte en todos los meses y se sitúa en las horas de la mañana desde las 7 hasta las 13 aproximadamente y las últimas horas de la tarde, de 21 a 23. El último cluster, el 2, se corresponde con las horas con mayor accidentes y se extiende en las horas centrales del día, de 14 a 20. Cabe destacar que el mes de agosto no tiene ninguna hora perteneciente a este cluster lo que señala la baja cantidad de accidentes en este periodo, por otro lado, no hay ningún mes que destaque notablemente en el caso contrario, es de- cir, con un claro repunte de accidentes. También cabe señalar que los meses de enero y julio tienen algo menos de accidentes pues se observa menor presencia de puntos amarillos, esto se entiende dado que son meses típicamente vacacio- nales. Esta clasificación confirma la información ya obtenida gracias al CSV de aforos. Capítulo 4 Resultados y conclusiones 4.1. Resultados En este capítulo se recogen los mejores resultados obtenidos tras realizar toda la etapa de desarrollo. Recordemos que el objetivo final es encontrar patrones y tendencias en el tráfico y accidentes para poder mejorar el flujo del tráfico de la ciudad y disminuir la cantidad de accidentes o, en su defecto, minimizar sus consecuencias. 4.1.1. Aprendizaje supervisado Gracias al aprendizaje supervisado hemos obtenido dos modelos predictivos cla- ve: Modelo para detectar consumo de drogas o alcohol en accidentes: este mo- delo tiene una precisión del 78 %. Puede ser de gran utilidad para las au- toridades al dirigirse al lugar del accidente, ya que les permite prever el estado de las personas involucradas y la probabilidad de que no estén en plenas facultades. Modelo para determinar la necesidad de asistencia sanitaria en accidentes: este modelo tiene una precisión del 83 %. Es fundamental para salvar vidas y optimizar recursos, especialmente en situaciones con múltiples acciden- tes o disponibilidad limitada de ambulancias. Permite evaluar rápidamente si se requiere asistencia sanitaria. 4.1.2. Aprendizaje no supervisado En el aprendizaje no supervisado, el objetivo era obtener información geográfica y temporal sobre la afluencia de vehículos y la incidencia de accidentes. Se han obtenido los siguientes resultados: Los distintos puntos de medición de aforo se han clasificado según la media y varianza de vehículos que pasan por cada punto. Esto permite identificar patrones de tráfico en diferentes áreas de la ciudad. Se han creado 5 grupos Capítulo 4. Resultados y conclusiones o clusters, el primero de ellos recoge aquellas ubicaciones con bajo volumen de coches y menor varianza, es decir, zonas poco concurridas de forma constante. El cluster 1 recoge las zonas con menos volumen y variabilidad algo más moderada, dentro de que sigue siendo un valor bajo. El cluster 2 agrupa ubicaciones con muy poco volumen y poca variabilidad. El cluster 3 recoge claramente zonas con mucho aforo y mucha variabilidad, en este grupo se encuentran estaciones de medición como la M-30, lugares donde se explica claramente ese alto volumen que fluctúa considerablemente. Por último, el cluster 4 recoge ubicaciones con un volumen moderado y algo de variabilidad. A continuación se muestran las ubicaciones en las que se encuentran los puntos de medición de cada cluster. ID Dirección Cluster 1 Paseo de la Castellana 4 2 Calle Princesa 0 3 Calle Doctor Esquerdo 4 4 Paseo de San Francisco de Sales 1 5 Paseo de Santa María de la Cabeza 4 6 Calle Arturo Soria 1 7 Avenida de Portugal 4 8 Calle Gran Via 0 9 Calle Atocha 0 10 Avenida de Oporto 0 11 Avenida del Manzanares (M-30) 3 12 Calle Jose Abascal 4 13 Calle Génova 4 14 Calle Jose Ortega y Gasset 1 15 Avenida Reina Victoria 1 16 Calle Alberto Aguilera 4 17 Calle Cea Bermúdez 4 18 Avenida Menéndez Pelayo 0 19 Calle Bravo Murillo 1 20 Avenida del Manzanares (M-30) 3 21 Calle Príncipe de Vergara 1 22 Calle Ronda de Valencia 0 23 Paseo de El Prado 4 24 Calle de Gran Via de San Francisco 0 25 Calle Hortaleza 4 26 Calle San Bernardo 0 27 Calle Alcalá 4 28 Calle Méndez Álvaro 2 29 Paseo Infanta Isabel 4 30 Calle Embajadores 0 31 Francos Rodríguez 1 32 Calle Toledo 0 33 Calle Sinesio Delgado 1 4.1. Resultados ID Dirección Cluster 34 Calle Mayor 0 36 Paseo de la Castellana 1 37 Calle Costa Rica 1 38 Avenida Cardenal Herrera Oria 1 39 Avenida de la Ilustración (M-30) 1 40 Calle Raimundo Fernández Villaverde 4 41 Calle Bravo Murillo 1 42 Avenida General Perón 1 43 Paseo de Extremadura 0 44 Calle Serrano 4 45 Calle Velázquez 4 46 Avenida de la Albufera 2 47 Calle Alcalá 2 48 Calle Hermanos García Noblejas 2 49 Avenida de Valladolid 0 50 Calle López de Hoyos 1 51 Avenida Alfonso XIII 1 52 Avenida Brasilia 1 53 Calle de Marcelo Usera 0 54 Avenida Rafaela Ybarra 0 55 Calle Alcocer 0 56 Avenida Arcentales 2 57 Calle Silvano 1 58 Avenida de Logroño 2 59 Calle San Cipriano 2 60 Calle Camino de Vinateros 2 Cuadro 4.1: Elementos pertenecientes a cada cluster. Agrupación de distritos según la incidencia de accidentes: Los distritos se han agrupado en cuatro clusters en función de la incidencia de accidentes. El cluster 1 contiene las ubicaciones con menor número de accidentes, Los clusters 0 y 2 tienen un número de accidentes moderado y, por último, el cluster 3 es aquel con mayor incidencia. A continuación se muestran los distritos y el cluster al que pertenecen. Distrito Cluster ARGANZUELA 3 BARAJAS 0 BARRIO DE SALAMANCA 1 CARABANCHEL 2 CENTRO 2 CHAMAR TÍN 1 CHAMBERÍ 3 Capítulo 4. Resultados y conclusiones Distrito Cluster CIUDAD LINEAL 2 FUENCARRAL EL PARDO 2 HOR TALEZA 3 LATINA 3 MONCLOA-ARAVACA 2 MORATALAZ 0 PUENTE DE VALLECAS 1 RETIRO 3 SAN BLAS-CANILLEJAS 3 TETUÁN 3 USERA 3 VICÁLVARO 0 VILLAVERDE 0 Cuadro 4.2: Elementos pertenecientes a cada cluster. Identificación de periodos con mayor incidencia de accidentes: los resulta- dos han identificado los meses y horas con mayor riesgo. Entre ellos se en- cuentra que el mes de agosto es aquel con menor incidencia de accidentes y menor flujo de tráfico, a continuación, con unos niveles algo superiores al mes de agosto, pero aún bajos, se encuentran los meses de enero y julio. Por otro lado, el resto de meses del año se encuentran en un nivel superior, destacando ligeramente por encima de todos septiembre. Respecto a las horas del día, el rango de horas más tranquilas abarca desde las 11 de la noche hasta las 6 de la mañana, comenzando un gran incremento a las 7, que alcanza su pico de tráfico a las 15, manteniéndose hasta las 20, donde comienza a decrecer de nuevo. 4.2. Conclusiones Los resultados obtenidos en este proyecto proporcionan herramientas valiosas para mejorar la movilidad urbana y la seguridad vial en Madrid. Algunas de las conclusiones clave incluyen: Utilidad práctica de los modelos predictivos: los modelos desarrollados pa- ra detectar consumo de sustancias y la necesidad de asistencia sanitaria pueden mejorar significativamente la respuesta de emergencia y la planifi- cación de recursos. Mejora de la infraestructura vial y transporte público: los distritos perte- necientes a la agrupación con menor número de accidentes pueden ser tomados como modelo de cara a la creación de nuevos barrios, que puedan tomar la estructura de sus calles y los servicios de transporte público co- mo referencia, imitando así también su bajo índice de accidentes. Por otro lado, los distritos con más accidentes pueden convertirse en el foco de las autoridades para llevar a cabo una mejora integral de la infraestructura 4.3. Líneas futuras vial. En función de las características exactas de cada punto, se podría re- forzar la señalización, mejorar el alumbrado público e incluso reconfigurar los cruces problemáticos para reducir el riesgo de accidentes. Asimismo, si la mejora de las infraestructuras viales no es factible, se podría intentar solventar la situación mejorando el transporte público de la zona, invitando así a los ciudadanos que transitan por ella a utilizarlo y, disminuir así el número de vehículos que circulan por la zona. Planificación basada en datos: la identificación de periodos de alto riesgo permite a las autoridades tomar medidas proactivas, como campañas de concienciación y mejoras en el sistema de transporte público en momentos críticos. El uso de técnicas avanzadas de análisis de datos ha demostrado ser efectivo para manejar grandes volúmenes de datos y extraer información útil para la toma de decisiones informadas. En resumen, este proyecto ha demostrado cómo el análisis de datos y el apren- dizaje automático pueden contribuir de manera significativa a mejorar la movili- dad y seguridad en una gran ciudad. Las herramientas y modelos desarrollados tienen el potencial de ser implementados y utilizados por las autoridades para hacer de Madrid una ciudad más segura y eficiente en términos de transporte. 4.3. Líneas futuras Se encuentran varias líneas futuras que podrían tomarse a corto plazo. En pri- mer lugar, llevar a cabo una expansión de la base de datos. Por un lado, sería interesante añadir información sobre días festivos, fines de semana y eventos especiales como manifestaciones y obras de larga duración. Esta información permitiría tener una mejor comprensión del tráfico y accidentes, pues en caso de eventos especiales como un puente, se entiende el aumento de accidentes y del flujo de vehículos. En caso de obra se podría dar la situación de una redi- rección del tráfico a otras zonas. De este modo, se podría tener una visión más exacta de la situación. Por otro lado, respecto a los aforos, sería interesante realizar una propuesta para colocar más medidores en la Comunidad de Madrid con el fin de poder tener una visión más precisa del volumen de vehículos que se mueven por el territorio todos los días. Existe una gran afluencia de vehículos que se desplazan desde la periferia a través de carreteras como la A-1, A-6 o M-40, que influyen de manera determinante en el tránsito y, por tanto, su monitorización podría aportar información muy valiosa. También, podría aumentarse la frecuencia con la que se hace el recuento del número de vehículos, y tener registros cada media hora, en lugar de cada hora. En cuanto al análisis, se podría realizar el clustering de este CSV con nuevas variables, que sean capaces de describir el estado del tráfico de manera más precisa y complementen a las que ya se tienen (media y varianza). También, se podrían explorar algoritmos más complejos como redes neuronales Capítulo 4. Resultados y conclusiones profundas (deep learning) para mejorar la precisión de las predicciones. Elaborar predicciones es una ardua tarea y más cuando se tienen datos desbalanceados, por lo tanto, sería interesante emplear más horas y nuevas técnicas en esta tarea con el fin de obtener modelos más precisos. Además, se podrían imple- mentar técnicas de validación cruzada para evaluar la robustez de los modelos en diferentes subconjuntos de datos para garantizar su generalización. En una visión más a largo plazo, se encuentran las siguientes utilidades del proyecto. Simulaciones de Tráfico: desarrollar y utilizar simulaciones que puedan modelar escenarios de tráfico complejos y sus interacciones. Se podrían llevar a cabo predicciones con series temporales para anticiparse a grandes flujos de tráfico. Planificación urbana: crear herramientas basadas en los modelos predicti- vos que resulten de utilidad a planificadores urbanos, para diseñar redes de carreteras y transporte público lo más óptimas posible. Herramientas educativas: gracias a la identificación de los rangos de eda- des y tipos de vehículos más afectados por los accidentes, se podrían llevar a cabo campañas de concienciación ciudadana orientadas a esas edades o, a los propietarios y futuros propietarios de esos tipos de vehículos. Análisis espacial: llevar a cabo un análisis espacial más riguroso que pro- porcione una buena segmentación de los distintos distritos madrileños e identifique las zonas más transitadas. El fin sería utilizar estos datos para elaborar una propuesta de mejora del transporte público personalizada a cada zona, contribuyendo así a una planificación urbana más eficiente y mayor seguridad para la ciudadanía. Por otro lado, aquellas zonas con redes de comunicación más eficientes podrían servir de referencia para la elaboración de estas propuestas de mejora. En conclusión, este proyecto ofrece una base sólida para el análisis y la mejora del tráfico y la seguridad vial en Madrid. Las líneas futuras propuestas tienen el potencial de expandir y profundizar en el conocimiento obtenido,

proporcio- nando herramientas prácticas y avanzadas para una planificación urbana más eficiente y segura. Con la implementación de estas mejoras, se espera que la mo- vilidad en Madrid no solo sea más fluida y predecible, sino también más segura para todos sus ciudadanos Capítulo 5 Análisis de impacto La seguridad vial es una cuestión que afecta a todas las ciudades del mundo y un componente fundamental para lograr un desarrollo sostenible en ellas. Según los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, el ODS 3 que promueve la "salud y bienestar", y el ODS 11, que aboga por "ciudades y comunidades sostenibles", es fundamental promover el bienestar ciudadano, su seguridad y la sostenibilidad de las ciudades. El presente proyecto aborda los problemas de tráfico y accidentes en la Comunidad de Madrid, ofreciendo un análisis y propuestas de mejora, así como una metodología replicable para la continua actualización de este análisis. Gracias a las técnicas utilizadas se han encontrado patrones temporales y ten- dencias que ayudan a entender las posibles causas de accidentes de tráfico, las condiciones típicas en las que estos se dan, las zonas con mayor incidencia y los rangos de edades más afectados por estos. A raíz de estas se proponen nuevas mejoras en el sistema de carreteras, ubicaciones que pueden ser utilizadas como referencia para replicar su estructura vial por su buen funcionamiento y eficien- cia, y políticas de concienciación ciudadana, escogiendo de forma fundamentada la población objetivo de estas y el mensaje a transmitir. Madrid es una ciudad con un crecimiento demográfico constante y, por tanto, sufre grandes aglomeraciones y movimiento intenso diario de vehículos. Para asegurar el desarrollo de la ciudad, es esencial tener carreteras seguras, libres de tráfico y embotellamientos, además de desarrollar y actualizar el sistema de transporte público en función de las necesidades demográficas, capaz de adap- tarse al crecimiento de la ciudad. Todo ello contribuirá directamente en la cali- dad del aire, reduciendo la contaminación y mejorando la calidad de vida de sus habitantes. Asimismo, es importante destacar que el centro de Madrid es típicamente cono- cido por los pronunciados desniveles del suelo, característica que limita el uso de la bicicleta en comparación con otras capitales europeas. Este hecho resalta la importancia de llevar a cabo un análisis exhaustivo de la movilidad urbana que favorezca las comunicaciones y disminuya los tiempos de desplazamiento. Con este enfoque, el proyecto no solo busca mitigar los problemas de tráfico Capítulo 5. Análisis de impacto y reducir la incidencia de accidentes, sino también contribuir a los esfuerzos globales para lograr las metas establecidas en los ODS 3 y 11, reforzando el compromiso con la salud, el bienestar y la sostenibilidad urbana. Bibliografía [1] MinnaLearn. (2022) Introducción al big data. [Online]. Avai- lable: https://courses.minnalearn.com/es/courses/digital-revolution/ big-data-and-beyond/introduction-to-big-data/ [2] PowerData. (2022) Big data: ¿en qué consiste? su importancia, desafíos y gobernabilida. [Online]. Available: https://www.powerdata.es/big-data [3] TechTarget. (2021) Garbage in, garbage out (gigo). [Online]. Available: https://www.techtarget.com/searchsoftwarequality/definition/ garbage-in-garbage-out [4] Teradata. (2021) What are the 5 v's of big data? [Online]. Available: https://www.teradata.com/glossary/what-are-the-5-v-s-of-big-data [5] QuestionPro. (2021) Ciclo de vida de los datos: Qué es y qué etapas tiene. [Online]. Available: https://www.questionpro.com/blog/es/ ciclo-de-vida-de-los-datos/ [6] Cyberlink. (2023) Ciclo de vida de los datos: qué es y cuales son sus fases. [Online]. Available: https://www.cyberclick.es/numerical-blog/ ciclo-de-vida-de-los-datos-que-es-y-cuales-son-sus-fases [7] OCI. (2024) ¿qué es el machine learning? [Online]. Availa- ble: https://www.oracle.com/es/artificial-intelligence/machine-learning/ what-is-machine-learning/ [8] IBM. (2024) ¿qué es el aprendizaje supervisado? [Online]. Available: https://www.ibm.com/es-es/topics/supervised-learning [9] ——. (2024) ¿qué es el aprendizaje no supervisado? [Online]. Available: https://www.ibm.com/es-es/topics/unsupervised-learning [10] AES. (2023) ¿qué es python? [Online]. Available: https://aws.amazon. com/es/what-is/python/ [11] A. S. Alberca. (2022) La librería pandas. [Online]. Available: https: //aprendeconalf.es/docencia/python/manual/pandas/ [12] E. R. Schmidt. (2019) re- expresiones regulares. [Online]. Available: https://rico-schmidt.name/pymotw-3/re/index.html BIBLIOGRAFÍA [13] J. J. L. Gómez. (2018-2023) Python requests. la librería para hacer peticiones http en python. [Online]. Available: https://j2logo.com/python/ python-requests-peticiones-http/#requests-get [14] P. S. Foundation. (2023) os — interfaces misceláneas del sistema operativo. [Online]. Available: https://docs.python.org/es/3.10/library/os.html [15] A. S. Alberca. (2022) La librería numpy. [Online]. Available: https: //aprendeconalf.es/docencia/python/manual/numpy/ [16] P. S. Foundation. (2024) Cómo (howto) unicode. [Online]. Available: https://docs.python.org/es/3/howto/unicode.html [17] itop. (2023) Scikit-learn. [Online]. Available: https://www.itop. es/soluciones-tecnologicas/business-analytics-business-intelligence/ scikit-learn.html [18] A. Vidhya. (2024) Tune hyperparameters with gridsearchcv. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/ tune-hyperparameters-with-gridsearchcv/ [19] S. Inc. (2006 - 2024) What is logistic regression? equation, assumptions, types, and best practices. [Online]. Available: https://www.spiceworks. com/tech/artificial-intelligence/articles/what-is-logistic-regression/ [20] IBM. (2006 - 2024) ?qué es un árbol de decisión? [Online]. Available: https://www.ibm.com/es-es/topics/decision-trees [21] datacamps. (2024) Soporte para máquinas vectoriales con el tutorial scikit-learn. [Online]. Available: https://www.datacamp.com/tutorial/ svm-classification-scikit-learn-python [22] L. Ramírez. (2023-04-01) Algoritmo k-means: ¿qué es y cómo funciona? [Online]. Available: https://www.iebschool.com/blog/ algoritmo-k-means-que-es-y-como-funciona-big-data/ [23] D. SCIENTIST. (2022-30-11) Machine learning clustering: el al- goritmo dbscan. [Online]. Available: https://datascientest.com/es/ machine-learning-clustering-dbscan [24] Kinsta. (2024) ¿qué es el web scraping? cómo extraer legalmente el contenido de la web. [Online]. Available: https://kinsta.com/es/ base-de-conocimiento/que-es-web-scraping/ [25] steemit. (2020) ¿explicación alternativa para accuracy, precision, recall y f1-score? [Online]. Available: https://steemit.com/spanish/@waster/ explicacion-alternativa-para-accuracy-precision-recall-y-f1-score Anexos Apéndice A Informe de originalidad Turn it in Capítulo 2. Estado del arte Capítulo 2. Estado del arte Capítulo 2. Estado del arte Capítulo 2. Estado del arteHASH(0x5565f6d7a400).4. Preprocesamiento Capítulo 3. Desarrollo Capítulo 3. Desarrollo 3.6. Análisis Capítulo 3. Desarrollo 3.6. Análisis Capítulo 3. Desarrollo HASH(0x5565f6d7ab20).6. Análisis Capítulo 3. Desarrollo 3.6. Análisis Capítulo 3. Desarrollo 1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 67

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Mon Jun 03 18:42:00 CEST 2024
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)