Facultad de Ciencias Económicas y Empresariales, ICADE

Implementación y Análisis de Modelos PHYD (*Pay How You Drive*) Utilizando Telemática y Machine Learning para la Reducción de Accidentes de Tráfico en España

Autor: María Dolores Roca Morlán

Director: María Coronado Vaca

Madrid | Junio 2024

Implementación y Análisis de Modelos PHYD (*Pay How You Drive*) Utilizando Telemática y Machine Learning para la Reducción de Accidentes de Tráfico en España

Autor: Roca Morlán, María Dolores
Director: Coronado Vacas, María
Entidad Colaboradora: ICADE – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este Trabajo de Fin de Grado investiga la implementación y el impacto de los modelos PHYD (*Pay How You Drive*) en los accidentes de tráfico en España. El objetivo principal es analizar los datos de accidentes de tráfico para comprender las variables que influyen en las fatalidades y lesiones, y proponer medidas preventivas para reducir las tasas de accidentes.

El estudio adopta un enfoque deductivo, combinando métodos cuantitativos y cualitativos. Se recopilaron datos cuantitativos sobre accidentes de tráfico, información de vehículos y lesiones de la DGT (Dirección General de Tráfico) para el período de 2017 a 2022. Estos datos fueron limpiados y preparados para el análisis utilizando técnicas estadísticas y modelos de aprendizaje automático.

Los conjuntos de datos incluyen registros detallados de accidentes de tráfico, registros de vehículos e información demográfica sobre los conductores. La limpieza de datos implicó eliminar duplicados, corregir inconsistencias y asegurar que los datos numéricos estuvieran correctamente formateados. Luego, los datos limpios se visualizaron usando Power BI y Python para identificar patrones y tendencias en las tasas de accidentes.

El análisis se centró en la siniestralidad en España y las características de los vehículos involucrados en accidentes. Se utilizaron mapas de calor y otras herramientas visuales para resaltar las distribuciones geográficas y demográficas de los accidentes, revelando áreas críticas y poblaciones en mayor riesgo.

Se desarrollaron varios modelos de aprendizaje automático para predecir el número de fallecidos y heridos en accidentes de tráfico. Los modelos incluían Random Forest, ANN (*Artificial Neural Network*) y Regresión Lasso. Estos modelos se evaluaron utilizando métricas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2). Los modelos identificaron factores significativos que contribuyen que haya heridos y fallecidos en los accidentes de tráfico.

Esta investigación destaca el potencial del aprendizaje automático para mejorar la seguridad vial y reducir fatalidades en los accidentes de tráfico. Al aprovechar datos detallados de accidentes y análisis predictivos, las partes interesadas pueden desarrollar intervenciones específicas para reducir las fatalidades de tráfico y mejorar la seguridad vial en general.

Palabras clave: Aprendizaje Automático, Accidentes de Tráfico, PHYD (*Pay How You Drive*), DGT (Dirección General de Tráfico), Seguridad Vial, Modelos de Seguros, España, Random Forest, ANN (*Artificial Neural Network*), Regresión Lasso.

Implementation and Analysis of PHYD (Pay How You Drive) Models Using Telematics and Machine Learning for Reducing Traffic Accidents in Spain

Author: Roca Morlán, María Dolores
Supervisor: Coronado Vacas, María
Collaborating Entity: ICADE– Universidad Pontificia Comillas

ABSTRACT

This Final Degree Project investigates the implementation and impact of PHYD (Pay How You Drive) models on traffic accidents in Spain. The main objective is to analyse traffic accident data to understand the variables influencing fatalities and injuries and to propose preventive measures to reduce accident rates.

The study adopts a deductive approach, combining quantitative and qualitative methods. Quantitative data on traffic accidents, vehicle information, and injuries were collected from the DGT (*Dirección General de Tráfico*) for the period from 2017 to 2022. These datasets were cleaned and prepared for analysis using statistical techniques and machine learning models.

The datasets include detailed records of traffic accidents, vehicle registrations, and demographic information about drivers. Data cleaning involved removing duplicates, correcting inconsistencies, and ensuring numerical data were properly formatted. The data was then visualized using Power BI and Python to identify patterns and trends.

The analysis focused on the sinistrality in Spain and the characteristics of vehicles involved in accidents. Heat maps and other visual tools were used to highlight geographic and demographic distributions of accidents, revealing critical areas and populations at higher risk.

Several machine learning models were developed to predict the number of fatalities and injuries in traffic accidents. The models included Random Forest, ANN (Artificial Neural Networks), and Lasso Regression. These models were evaluated using metrics such as Mean Squared Error (MSE) and the coefficient of determination (R^2). The models identified significant factors contributing to injuries and fatalities in traffic accidents.

This research highlights the potential of machine learning to improve road safety and reduce fatalities in traffic accidents. By leveraging detailed accident data and predictive analytics, stakeholders can develop targeted interventions to reduce traffic fatalities and improve overall road safety.

Keywords: Machine Learning, Traffic Accidents, PHYD (Pay How You Drive), DGT (*Dirección General de Tráfico*), Road Safety, Insurance Models, Spain, Random Forest, ANN (Artificial Neural Network), Lasso Regression

Índice de la memoria

Índice de figuras

Índice de Ecuaciones

Índice de tablas

Glosario de acrónimos

ACM	Análisis de Correspondencias Múltiples
ANN	*Artificial Neural Network*/Red Neuronal Artificial
BD	Big Data
CAN	*Controller Area Network*/Red de Área de Controladores
ConvNets	*Convolutional Neural Networks*/Redes Neuronales Convolucionales
DTR	*Decision Tree Regression*/Regresión de Árbol de Decisión
GAM	*Generalized Additive Models*/Modelos Generalizados Aditivos
GAMLSS	*Generalized Additive Models for Location, Scale, and Shape*/Modelos Generalizados Aditivos Ubicación, Forma y Escala
IoV	*Internet of Vehicles*/Internet de Vehículos
MHYD	*Manage How You Drive*/Gestiona Cómo Conduces
MLP	*Multilayer Perceptron*/Perceptrón Multicapa
PAYD	*Pay As You Drive*/Paga Como Conduces
PCA	*Principal Component Analysis*/Análisis de Componentes Principales
PHYD	*Pay How You Drive*/Paga Cómo Conduces
SHR2	*Strategic Highway Research Program*/Programa Estratégico de Investigación de Carreteras
SVR	*Support Vector Regression*/Regresión Vectorial de Soporte
UBI	*Usage Based Insurance* / Seguro Basado en el Uso

Capítulo 1. INTRODUCCIÓN

1.1 OBJETIVOS

El objetivo principal de este Trabajo de Fin de Grado es explorar, analizar y evaluar en profundidad la situación española en relación con los accidentes de tráfico. Los objetivos principales incluyen:

- **Analizar las causas que provocan los accidentes de tráfico** en España realizando un análisis exhaustivo de las variables que causan que haya fallecidos y heridos en los accidentes.
- **Proponer recomendaciones y medidas preventivas** basadas en los hallazgos con el objetivo de reducir la incidencia de accidentes relacionados con los factores identificados. Mejorando así el comportamiento de los conductores

1.2 JUSTIFICACIÓN DEL TEMA

La adopción de PHYD (*Pay How You Drive*) representa una oportunidad significativa para mejorar la precisión en la evaluación de riesgos en la industria de seguros vehiculares. Al permitir una evaluación más personalizada del comportamiento del conductor se pueden fomentar conductas de conducción más seguras. Esta investigación busca aportar una comprensión profunda de los beneficios y desafíos asociados con la implementación de PHYD, contribuyendo así al desarrollo de estrategias más efectivas y éticas en la gestión de riesgos automovilísticos.

Además, la implementación de PHYD tiene el potencial de contribuir a la reducción de accidentes de tráfico. Al fomentar conductas de conducción seguras, se puede esperar una disminución en la frecuencia y gravedad de los accidentes, lo que a su vez reduce los costos asociados con los siniestros y las reclamaciones de seguros. Esto no solo beneficia a las

aseguradoras, sino también a la sociedad en general, al mejorar la seguridad vial y reducir las lesiones y muertes en la carretera.

1.2.1 MOTIVACIÓN DEL PROYECTO

En la era digital actual, la convergencia de la tecnología y la industria automotriz ha dado lugar a innovaciones revolucionarias en el ámbito de los seguros de automóvil. Una de estas innovaciones notables es el concepto de "*Pay How You Drive*" (PHYD), también conocido como identificación del conductor. Este enfoque representa un cambio paradigmático en la forma en que se evalúan y gestionan los riesgos asociados a la conducción, al introducir la monitorización y evaluación del comportamiento del conductor como un factor clave en la determinación de las primas de seguro.

Durante 2023 los accidentes en la carretera han sumado 1145 personas que han perdido la vida, según datos de la DGT (Dirección General de Tráfico, 2023). Estos accidentes pueden ocurrir aparentemente de manera inevitable, pero hay varios factores que contribuyen a su frecuencia y gravedad como las condiciones meteorológicas, la velocidad excesiva, las distracciones o las condiciones de la carretera.

Según la OMS (Organización Mundial de la Salud) los accidentes de tráfico son una de las principales causas de muerte en el mundo. Cada año los accidentes causados por vehículos se cobran la vida de 1,19 millones de personas. Las personas más afectadas por estos accidentes son los jóvenes y niños de 5 a 19 años. (World Health Organizationç, 2023)

Este no es simplemente un tema de cifras y estadísticas; es una cuestión que puede cambiar vidas en un instante y que, con frecuencia, podría prevenirse. Detrás de cada accidente hay personas que han sufrido y familias que han sido afectadas y que han perdido a seres queridos. En la figura 1 vemos la variación de víctimas mortales durante el 2023.

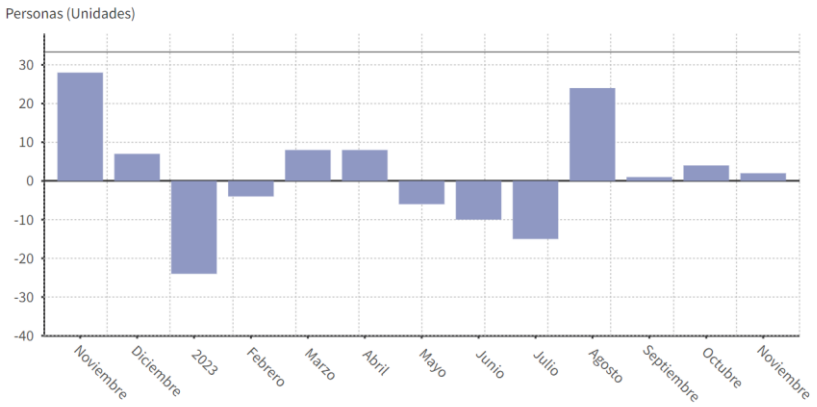

Figura 1: Variación de víctimas mortales durante el 2023. (Europa Press, 2023.)

Con este trabajo se pretende contribuir a que disminuya el número de accidentes de tráfico y así contribuir a que los viajes por carretera sean más seguros. De esta manera se podrían evitar tragedias.

Hay que tener en cuenta que la seguridad vial es responsabilidad de uno mismo, no de los peatones o de las autoridades. Por mi edad, 23 años, pertenezco a unos de los grupos de población entre los más afectados por este tipo de accidentes. Creo que es necesario realizar un estudio en profundidad sobre estos factores que provocan accidentes y poder realizar una aplicación que nos muestre cómo conducimos y que estas acciones tengan consecuencias.

1.3 METODOLOGÍA

La investigación adoptará un enfoque deductivo, combinando métodos cuantitativos y cualitativos. Este enfoque permitirá una comprensión integral de la implementación y el impacto de los modelos PHYD en la industria de seguros vehiculares.

- Métodos Cuantitativos: Se recopilarán y analizarán datos relevantes sobre accidentes de tráfico en España, incluyendo información sobre vehículos y heridos. Este análisis incluirá la limpieza y preparación de los datos, y la aplicación de técnicas de análisis

estadístico y aprendizaje automático para identificar patrones y tendencias que puedan informar la evaluación de riesgos.

- Métodos Cualitativos: Se realizará una revisión bibliográfica para comprender el estado actual del conocimiento sobre PHYD, incluyendo sus beneficios, desafíos y aplicaciones en diversos contextos. Además, se llevarán a cabo estudios de caso para examinar la implementación de PHYD en diferentes países y contextos. Estos estudios de caso proporcionarán información valiosa sobre las estrategias utilizadas, los resultados obtenidos y las lecciones aprendidas, que pueden ser aplicadas al contexto español.

1.4 ESTRUCTURA

El trabajo se estructurará en varios capítulos, comenzando con una introducción que abarcará la motivación, los objetivos, la justificación del tema y la metodología. Seguirá un estado de la cuestión que revisará la literatura existente y los conceptos clave. A continuación, se describirán los datos utilizados y el proceso de limpieza y preparación de los datos. También se realizará un análisis exploratorio de estos. Posteriormente, se detallarán los modelos desarrollados y su evaluación. Finalmente, se presentarán los resultados obtenidos y se concluirá con las implicaciones prácticas, limitaciones y recomendaciones para futuras investigaciones. También se incluirán apéndices con documentación adicional relevante.

Capítulo 2. ESTADO DE LA CUESTIÓN

Este TFG (Trabajo de Fin de Grado) explora el análisis de datos y el aprendizaje automático para identificar perfiles de comportamiento de conducción. La selección de textos para este estudio se ha realizado con el objetivo de abarcar un amplio espectro de metodologías y la personalización de programas de entrenamiento de conductores.

Los documentos seleccionados presentan avances en la identificación de comportamientos de conducción agresiva, la evaluación de riesgos de conductores mediante modelos predictivos, y la identificación personalizada del conductor. Estos estudios representan la actualidad de la investigación aplicada y ofrecen una base sólida para desarrollar un marco analítico que responda a los desafíos actuales en la gestión de la conducta del conductor.

Para llevar a cabo este análisis se va a empezar definiendo cada uno de los documentos seleccionados y su contextualización. Se va a explorar cómo se define el comportamiento de conducción de los distintos estudios y como estos se clasifican, para hallar similitudes y diferencias entre los mismos. Después se van a analizar las teorías y modelos explicativos y así encontrar problemas sin resolver y preguntas de investigación. Se van a estudiar también las metodologías empleadas en los estudios revisados y las variables empleadas.

2.1 DEFINICIÓN Y CONTEXTUALIZACIÓN DE LA DOCUMENTACIÓN

Para poner al lector en contexto en este apartado se va a ir buscando cómo los autores definen el fenómeno de estudio, así como la identificación de las teorías o marcos conceptuales que los autores utilizan para fundamentar su investigación. Se quiere conseguir aclarar las metodologías utilizadas y la interpretación de los resultados obtenidos para extraer los hallazgos clave de cada estudio. Todo el trabajo realizado se muestra de forma resumida en las tablas 1,2 y 3.

AUTORES	OBJETIVOS	FENÓMENO	METODOLOGÍA	RESULTADOS
Usami et al. (2017) ITALIA	Identificar perfiles de comportamiento para entender la relación entre estos perfiles y la implicación en accidentes de tráfico.	Comportamiento de conducción, específicamente enfocado en aspectos como la agresividad y la inseguridad.	Se usa ACM (Análisis de Correspondencia Múltiple) y análisis de clúster para la encuesta a conductores italianos identificando patrones	Se han identificado 7 perfiles de conductores y se encontró una relación significativa entre estos perfiles y su implicación en accidentes
Abdelrahman et al. (2018) EE.UU.	Desarrollar un enfoque robusto basado en datos para perfilar el comportamiento del conductor, prediciendo la probabilidad de riesgo.	Predecir el riesgo con comportamientos de conducción, utilizando modelos de aprendizaje automático para su predicción	SVR (*Support Vectorial Regression*) y DTR (*Decision Tree Regression*) sobre los datos para modelar la probabilidad de riesgo	Los modelos validan la capacidad de predecir la probabilidad de riesgo basada en comportamientos de conducción.
Martinelli y Mercaldo (2018) ITALIA	Investigar y categorizar la agresividad de conducción utilizando datos telemáticos para mejorar la seguridad vial.	Se basa en modelos conductuales y telemáticos para definir y medir la agresividad en la conducción	Utiliza análisis de clúster sobre datos del bus CAN (*Controller Area Network*) para identificar patrones de conducción agresiva	Identifica niveles de agresividad entre conductores demostrando la viabilidad de clasificar comportamientos
Martinelli y Mercaldo (2018) ITALIA	Identificar al conductor y el género de este mediante un conjunto de características extraídas durante la conducción	Para la identificación se emplea un enfoque de clasificación de series temporales y aprendizaje supervisado	Análisis de series temporales, seguida de un entrenamiento y clasificación utilizando redes MLP (*Multilayer perceptron*)	Los experimentos demuestran la capacidad de las características propuestas para identificar al conductor.
Martinelli et al. (2018) ITALIA	Sistema de autenticación de conductores en tiempo real usando aprendizaje automático para diferenciar entre el propietario o no.	Se basa en el comportamiento de conducción. Se aprendizaje automático y análisis de patrones	Extracción de características de datos de sensores del vehículo y análisis mediante un sistema de clasificación basado en reglas.	Alta precisión en la identificación del conductor, demostrando la viabilidad del sistema para la autenticación del conductor
Ezzini et al. (2018) ITALIA	Desarrollar un método para la identificación precisa de conductores usando patrones de conducción y datos del vehículo	Basado en teorías de reconocimiento de patrones y aprendizaje automático para interpretar los datos de conducción	Uso de técnicas de aprendizaje automático y análisis de datos de conducción y diferenciar entre conductores	Demostración de la capacidad de identificar conductores con alta precisión en los primeros minutos de conducción
Huang y Meng (2019) CHINA	Utilizar datos telemáticos de conducción para mejorar la precisión en la tarificación de seguros automóviles	Se apoya en teorías de riesgo y modelos estadísticos, como la regresión logística y técnicas de aprendizaje automático	Implementa regresión logística y algoritmos de aprendizaje automático para predecir la probabilidad de siniestros	Se muestra una mejora en la precisión en la tarificación de seguros, destacando el valor de los datos sobre los factores de riesgo.

Tabla 1: Comparación de estudios sobre comportamientos de conducción. Fuente: Elaboración propia

AUTORES	OBJETIVOS	FENÓMENO	METODOLOGÍA	RESULTADOS
Subramanian et al. (2019) MUNDIAL	Analizar el comportamiento de conducción utilizando BD (Big Data) para mejorar la precisión en la tarificación de seguros	Explora modelos como PAYD (*Pay As You Drive*) , PHYD (*Pay How You Drive*) y MHYD (*Manage How You Drive*)	Utiliza tecnologías de BD y aprendizaje automático para analizar patrones de conducción agresiva y otros comportamientos	Taridas de seguros personalizadas basadas en el análisis del comportamiento, considerando factores emocionales
Gao y Wüthrich (2019) MUNDIAL	Clasificar viajes de conducción individuales usando datos GPS a través de ConvNets (*Convolutional Networks*)	Se centra en patrones de velocidad, aceleración y cambios de sentido. Usa aprendizaje profundo para clasificar comportamientos	ConvNets para procesar y clasificar datos telemáticos, demostrando la aplicación de técnicas de aprendizaje profundo	Logra una precisión del 75% en predicciones, mostrando la efectividad de las ConvNets en la clasificación precisa del comportamiento
Gao y Wüthrich (2019) MUNDIAL	Analizar el riesgo de conducción asociado a diferentes velocidades, utilizando datos telemáticos	Teorías de análisis de riego y modelos K-medoids y PCA (*Principal Component Analysis*)	K-medoids y PCA sobre datos telemáticos para identificar covariables significativas que impactan la frecuencia de reclamaciones	Demuestra que los datos de conducción pueden mejorar la evaluación de riesgo y clasificación de asegurados
Winlaw et al. (2019) MUNDIAL	Identificar comportamientos de conducción que aumentan el riesgo de accidentes utilizando datos telemáticos	Teorías de análisis de riesgo y comportamiento del conductor, utilizando modelos explicativos	Estudio de casos y controles, analizando datos de conductores que han estado involucrados en accidentes	Ciertos comportamientos, especialmente la velocidad excesiva, están asociadaos a un aumento de riesgo de accidente
Boucher y Rurcotte (2020)	Investigar la relación entre la distancia conducida y la probabilidad de accidente, usando análisis longitudinal	Aplica GAM (*Generalized Additive Model*) y GAMLSS (*Generalized Additive Model* for Location, Shape and Scale)	Análisis de datos telemáticos y técnicas estadísticas para modelar la relación entre la distancia recorrida y la ocurrencia de accidentes	Correlación entre mayor distancia recorrida y aumento en la probabilidad de accidentes
Xun et al. (2020) CHINA	Busca desarrollar un esquema de autenticación de conductores para crear una “huella digital” del conductor	Teorías de identificación biométrica y comportamental, aplicando modelos de aprendizaje automático	Técnicas de aprendizaje automático para procesar y analizar datos de comportamiento, con el fin de identificar características únicas	La efectividad del esquema propuesto para identificar conductores con alta precisión
Martinelli et al. (2020) ITALIA	Desarrollar y validar un modelo de aprendizaje automático que sea capaz de identificar conductores basándose en patrones	Variabilidad en los patrones de conducción. Incluye aspectos como aceleración, frenada, estilo de toma de curvas y otros comportamientos	Modelos como redes neuronales, máquinas de vectores de soporte o árboles de decisión.	Precisión, sensibilidad y especificidad de aprendizaje automático para identificar conductores

Tabla 2: Comparación de estudios sobre comportamientos de conducción. Fuente: Elaboración propia

AUTORES	OBJETIVOS	FENÓMENO	METODOLOGÍA	RESULTADOS
Ebrahim et al. (2020)	Perfilar el comportamiento de los conductores, buscando perfiles de riesgo de los conductores basándose en eventos específicos	Compara diferentes modelos de aprendizaje automático supervisado para ofrecer mejores predicciones de riesgo	Utiliza datos de SHRP2 (*Stratgic Highway Research Program*) para identificar 13 predictores de riesgo.	Los modelos que manejan complejidad y no linealidad de datos son particularmente efectivos en predecir el riesgo
Gao et al. (2021)	Mejorar los modelos de regresión de Poisson utilizados para predecir la frecuencia de reclamación de seguros	Relación entre el comportamiento de conducción y la frecuencia de reclamaciones de seguros	Métodos de boosting para integrar y analizar datos. Combina modelos predictivos débiles para formar un modelo fuerte y preciso.	Mejora en la precisión de predicción de la frecuencia de reclamaciones al incorporar datos telemáticos en los modelos de regresión de Poisson.
Abdennour at al. (2021) TÚNEZ	Encontrar un equilibrio óptimo entre la precisión, la latencia y el tiempo de entrenamiento en la identificación del conductor	51 tipos diferentes de señales recogidas durante un total de 23 horas de conducción realizadas por 10 participantes en un estudio llevado a cabo por Ocslab.	Modelos de aprendizaje automático clásico	Precisión y eficiencia del modelo en comparación con otros modelos existentes
Niño de Zepeda et al. (2021) EE.UU.	Perfilar estilos de conducción, capturando patrones. Comprender los comportamientos cambiantes y la respuesta al entorno	Un agrupamiento dinámico que asigna a un conductor una secuencia de clusters, así se identifican los comportamientos	Datos de Los Ángeles, explorando y analizando los estilos de conducción. Algoritmos de agrupamiento estático y dinámico.	Efectividad del agrupamiento dinámico para capturar patrones de conducción en constante evolución.
Zhao et al. (2021) CHINA	Si la inclusión de datos de comportamiento offline puede mejorar la precisión de la clasificación del riesgo de accidentes de tráfico	Clasificación del riesgo de accidentes en función de IoV (*Internet Of Vehicles*) y datos de comportamiento offline	Se construyen modelos utilizando datos online y offline y se usan técnicas para identificar variables de alto poder que mejoran la clasificación	Incluir datos offline mejora la precisión de clasificación de riesgo del conductor. Las variables de estos datos son significativas.
Martinelli et al. (2021) ITALIA	Investigar si un conjunto de características es útil para discriminar entre diferentes conductores, estilos de conducción y tipos de carretera	Evaluación en tres etapas: comparación de estadísticas descriptivas de las poblaciones, estilos de conducción y tipos de carreteras.	Se diseñan varias redes neuronales y se encuentra que la arquitectura óptima tiene 25 capas ocultas para la identificación	Logra altos niveles de precisión y baja pérdida para todas las tareas de detección. Las redes neuronales superan a la clasificación tradicional.

Tabla 3: Comparación de estudios sobre comportamientos de conducción. Fuente: Elaboración propia

2.2 HECHOS Y CONSENSOS

2.2.1 HECHOS

Los hechos identificados en los estudios se basan en datos y resultados experimentales, como la efectividad de las redes neuronales en la clasificación de comportamientos de conducción. Martinelli y Mercaldo (2018) demuestra cómo las redes neuronales, específicamente las arquitecturas profundas, pueden identificar con precisión distintos patrones de conducción basados en datos telemáticos.

Otro hecho importante es la eficacia de los algoritmos de aprendizaje automático para diferenciar entre el propietario del vehículo y posibles impostores basándose en características recogidas del bus CAN del vehículo, identificado en el estudio de Martinelli et al. (2020).

2.2.2 CONSENSOS

Hay consensos entre los autores, como la utilidad de los datos telemáticos para mejorar los modelos de riesgo. Los estudios sobre PHYD y la identificación de comportamientos de conducción mediante redes neuronales u otras técnicas de aprendizaje automático demuestran cómo los datos telemáticos, que incluyen información detallada sobre el estilo de conducción o la velocidad pueden ser utilizados para evaluar el riesgo.

Gao y Wüthrich (2019) demuestran cómo las redes neuronales pueden identificar con precisión diferentes estilos de conducción y tipos de carretera. Abdelrahman et al. (2018) destacan en su estudio cómo los modelos de aprendizaje automático pueden mejorar la precisión en la predicción de riesgos utilizando datos telemáticos. Por último, también hay que destacar el estudio de Huang y Meng (2019) donde demuestran que la integración de datos telemáticos mejora significativamente la capacidad predictiva de los modelos de seguros tradicionales

Otra aplicación propuesta es el uso de estos modelos en sistemas de vehículos autónomos para mejorar la capacidad del vehículo de adaptarse al estilo de conducción del conductor humano en entornos de conducción mixtos.

Martinelly y Mercaldo (2018) investigan cómo los algoritmos de aprendizaje automático pueden identificar patrones únicos de comportamiento de conducción, lo cual es fundamental para mejorar la interacción entre vehículos autónomos y humanos

Además, Niño de Zepeda et al. (2021) introducen un enfoque de agrupamiento para identificar estilos de conducción dinámicos, demostrando cómo los datos telemáticos no solo pueden mejorar la precisión de los modelos de riesgo, sino también ofrecer perspectivas sobre la evolución del comportamiento de conducción en el tiempo.

Se ilustra el consenso entre los autores sobre la importancia de integrar datos telemáticos en los modelos de riesgo. Al personalizar las primas de seguro basadas en el comportamiento real del conductor y mejorar la capacidad de los vehículos autónomos para adaptarse a los estilos de conducción humana, estas tecnologías pueden contribuir significativamente a la seguridad vial y a la eficiencia del tráfico.

2.3 PREGUNTAS Y PROBLEMAS SIN RESOLVER

Los estudios han respondido a preguntas sobre la capacidad de los datos telemáticos y las características del conductor para predecir el riesgo de conducción. Sin embargo, persisten problemas sin resolver relacionados con la privacidad de los datos y la aceptación por parte de los usuarios de la monitorización continua.

2.3.1 PRIVACIDAD DE LOS DATOS

La preocupación por la privacidad de los datos surge del hecho de que los datos telemáticos recopilan información detallada sobre los patrones de conducción de los individuos. Esto incluye, pero no se limita a, la ubicación en tiempo real, la velocidad, las aceleraciones, las frenadas y otros comportamientos de conducción. Mientras que esta información es

invaluable para la personalización de los seguros y la mejora de la seguridad vial, también plantea preguntas sobre cómo se recopila, almacena y utiliza esta información, y quién tiene acceso a ella. La protección de la privacidad del conductor es fundamental.

2.3.2 ACEPTACIÓN DE LOS USUARIOS

La aceptación de los usuarios de la monitorización continua es otro desafío importante. Aunque muchos conductores pueden estar dispuestos a compartir sus datos a cambio de primas de seguro más bajas o incentivos similares, otros pueden sentirse incómodos con la idea de ser monitoreados constantemente.

La percepción de la monitorización como una intrusión en la privacidad personal puede limitar la disposición de los usuarios a participar en programas de seguros basados en el uso. Además, la aceptación varía significativamente entre diferentes culturas y jurisdicciones, lo que complica la implementación global de estas tecnologías.

2.4 DISCIPLINAS INVOLUCRADAS

La naturaleza de los estudios sobre PHYD refleja la convergencia de varias disciplinas que aportan enfoques únicos y complementarios al análisis del comportamiento de conducción y la evaluación del riesgo. Esta fusión de conocimientos es esencial para desarrollar modelos precisos y aplicaciones prácticas que benefician tanto a conductores como a compañías de seguros.

2.4.1 INFORMÁTICA

La informática juega un papel crucial en el desarrollo de algoritmos y modelos de aprendizaje automático que procesan y analizan grandes volúmenes de datos telemáticos. En el estudio realizado por Marinelli et al. (2021) se investiga si las características basadas en datos del acelerómetro pueden discriminar entre diferentes conductores. Los resultados muestran que las redes neuronales pueden identificar estos patrones de conducción,

destacando la importancia de los algoritmos de aprendizaje automático en la mejora de la evaluación del riesgo de conducción.

2.4.2 ACTUARIAL

La disciplina actuarial se centra en la evaluación del riesgo y la modelización de la incertidumbre, lo que es esencial para la tarificación de los seguros. La integración de datos telemáticos en modelos actuariales permite una evaluación más precisa del riesgo basada en el comportamiento de conducción real, lo que lleva a una tarificación más justa y personalizada. Por ejemplo, Huang y Meng (2019) demuestran cómo los datos telemáticos pueden ser utilizados para mejorar los modelos de clasificación y tarificación de seguros, permitiendo una evaluación más detallada del riesgo asociado con el comportamiento de conducción real.

2.4.3 CONTRIBUCIONES MULTIDISCIPLINARES

La colaboración entre estas disciplinas permite abordar el fenómeno del PHYD desde múltiples ángulos, combinando la capacidad técnica para analizar datos con una comprensión profunda del comportamiento humano y los sistemas de tráfico. Esto no solo mejora la precisión de los modelos de riesgo, sino que también facilita el desarrollo de soluciones que son prácticas, éticamente responsables y aceptables para los conductores.

2.5 METODOLOGÍAS DE ESTUDIO

Los estudios sobre PHYD y la identificación del comportamiento de conducción aprovechan las metodologías de aprendizaje automático y análisis estadístico para extraer perspectivas de los datos telemáticos. Estas metodologías se aplican de manera diversa dependiendo del objetivo específico del estudio, ya sea para clasificar estilos de conducción, predecir riesgos de accidentes, o personalizar tarifas de seguros.

2.5.1 APRENDIZAJE AUTOMÁTICO

Las técnicas de aprendizaje automático, como las redes neuronales y los algoritmos de clasificación, se utilizan para analizar patrones complejos en los datos telemáticos. Por ejemplo, en el estudio de Martinelli et al. (2018) se demostró cómo las redes neuronales pueden identificar con precisión patrones de conducción basados en datos del acelerómetro. Estas técnicas permiten modelar comportamientos de conducción complejos y dinámicos que no se podrían analizar eficazmente con métodos estadísticos tradicionales.

2.5.2 ANÁLISIS ESTADÍSTICO

El análisis estadístico se emplea para evaluar la significancia de las variables y sus relaciones. Esto incluye pruebas de hipótesis para determinar la importancia de ciertas características de conducción en la predicción del riesgo. Por ejemplo, Abdennour, et al. (2021) utilizaron técnicas estadísticas para validar la efectividad de las características seleccionadas en la diferenciación entre conductores.

2.6 VARIABLES CLAVE

La selección de variables adecuadas es fundamental en el análisis del comportamiento de conducción para desarrollar modelos predictivos precisos.

- Variables de Comportamiento de Conducción: Incluyen la velocidad, la aceleración, la distancia de frenado, y el tiempo de reacción. Estas variables son indicadores directos del estilo de conducción y pueden ser utilizados para evaluar el riesgo asociado con comportamientos específicos. Por ejemplo, la aceleración o frenado brusco pueden indicar un estilo de conducción agresivo.
- Factores Contextuales: El tipo de carretera (urbana, rural, autopista), las condiciones meteorológicas (lluvia, nieve, niebla), y la hora del día (día, noche) son ejemplos de factores contextuales que influyen en el comportamiento de conducción. La adaptación del conductor a estas condiciones puede proporcionar características

adicionales sobre su capacidad para gestionar diferentes situaciones de conducción y su predisposición al riesgo.

Hay estudios, como el de Abdennour et al. (2021) o el de Niño de Zepeda et al. (2021) que ilustran cómo la combinación de variables de comportamiento y factores contextuales puede mejorar significativamente la precisión de los modelos predictivos.

2.7 DIFERENCIAS Y SIMILITUDES

2.7.1 SIMILITUDES

Una similitud destacada es el reconocimiento de la importancia de los datos telemáticos. Los datos telemáticos, que incluyen información sobre la velocidad del vehículo, la aceleración, la ubicación GPS, y otros parámetros de conducción, se han reconocido ampliamente por su valor en la mejora de los modelos de riesgo para las compañías de seguros y en la promoción de la seguridad vial. Estos datos permiten una evaluación más precisa y personalizada del comportamiento de conducción, lo que lleva a mejores predicciones sobre el riesgo de accidentes y, por lo tanto, a una tarificación más justa y personalizada de los seguros de automóviles.

2.7.2 DIFERENCIAS EN TÉCNICAS DE ANÁLISIS E INTERPRETACIÓN

- Técnicas de Análisis de Datos: La principal diferencia entre los estudios radica en las técnicas específicas utilizadas para analizar los datos telemáticos. Algunos estudios emplean métodos estadísticos tradicionales como el realizado por Huan y Meng (2019) o el realizado por Zhao et al. (2021), mientras que otros aprovechan algoritmos de aprendizaje automático y aprendizaje profundo para identificar patrones complejos en los datos como el de Martinelli et al. (2021).
- Interpretación e Integración en Modelos de Riesgo: Otra área donde pueden surgir diferencias es en cómo los investigadores interpretan los datos telemáticos y los integran en modelos de riesgo. Algunos estudios como el de Martinelli et al. (2021) se enfocan en la correlación directa entre comportamientos específicos de

conducción y la probabilidad de accidentes, mientras que otros buscan entender el contexto más amplio de esos comportamientos, incluyendo factores externos como las condiciones del tráfico y el clima como el estudio de Niño de Zepeda et al. (2021).

2.8 GRANDES DEBATES Y PREGUNTAS

El debate sobre la mejor manera de integrar los datos telemáticos en los modelos de seguros y la preocupación por equilibrar la precisión del modelo con la privacidad del usuario reflejan dos de las cuestiones más críticas en el desarrollo de tecnologías de seguros basadas en el uso y PHYD. Este debate se centra en varios aspectos clave:

2.8.1 INTEGRACIÓN DE DATOS TELEMÁTICOS EN MODELOS DE SEGUROS

1. Precisión vs. Generalización: Existe un debate sobre cómo desarrollar modelos que sean a la vez precisos en la predicción del riesgo, pero sin limitar la aplicabilidad del modelo a contextos específicos.
2. Datos Relevantes vs. Redundantes: Otro aspecto del debate se centra en determinar qué datos telemáticos son verdaderamente relevantes para la evaluación del riesgo. Mientras más datos se recopilen, más complejo puede volverse el modelo, aumentando la dificultad de interpretar los resultados y potencialmente introduciendo ruido en lugar de señales claras.

2.8.2 OPTIMIZACIÓN DE MODELOS PARA DIFERENTES CONTEXTOS

- Adaptabilidad Cultural: Los modelos de seguros basados en telemática deben ser sensibles a las normas culturales y las expectativas de privacidad, que varían significativamente entre regiones. Lo que es aceptable en términos de monitoreo y recolección de datos en una cultura puede no serlo en otra.
- Variabilidad Geográfica: Las diferencias en las infraestructuras de transporte, las normas de tráfico y los patrones de conducción entre regiones requieren que los modelos sean adaptados o ajustados para reflejar estas variabilidades. Esto plantea

preguntas sobre cómo desarrollar modelos que sean tanto específicos del contexto como escalables a nivel global.

2.9 *LAGUNAS EN LA INVESTIGACIÓN*

Una de las principales lagunas es comprender cómo los modelos PHYD, que ajustan las primas de seguro basadas en el comportamiento de conducción real, afectan a dicho comportamiento a lo largo del tiempo. Existen preguntas sin respuesta sobre si los incentivos proporcionados por los modelos PHYD conducen a mejoras sostenidas en la seguridad vial o si los efectos positivos disminuyen después de cierto tiempo.

Otra área poco explorada es cómo los modelos PHYD se adaptan a los rápidos avances en la tecnología de los vehículos, incluyendo la creciente adopción de vehículos autónomos y asistidos. Los sistemas de asistencia al conductor y otras tecnologías emergentes pueden cambiar fundamentalmente los patrones de riesgo asociados con la conducción, lo que requiere una evolución constante de los modelos PHYD para reflejar estas nuevas realidades.

Además, otra gran laguna es la falta de datos españoles en los estudios, ya que la mayoría de los datos provienen de otros países, lo que limita la aplicabilidad y relevancia de los hallazgos para el contexto español.

2.10 *CONCLUSIÓN*

A lo largo de este capítulo, he analizado en detalle el concepto y las implicaciones de los modelos de seguro PHYD, basándome en múltiples estudios sobre datos telemáticos, aprendizaje automático y su integración en la evaluación del riesgo de conducción. He identificado tres áreas principales que presentan lagunas significativas en la comprensión actual de estos modelos.

En respuesta a estas lagunas y a la predominancia de datos internacionales en la literatura existente, he decidido enfocar mi investigación en el contexto español, analizando datos específicos de accidentes y vehículos de España.

Esta decisión se basa en la convicción de que los datos locales proporcionarán una perspectiva más relevante y precisa para comprender los efectos y desafíos de los modelos PHYD entre los conductores españoles. Mediante el uso de esta información específica, aspiro a realizar una investigación profunda que arroje luz sobre cómo los accidentes, los vehículos y otros factores influyen en la seguridad vial en España, contribuyendo así a la adaptación y mejora continua de los modelos PHYD en nuestro entorno específico.

Con mi estudio pretendo abordar dos de las tres lagunas identificadas en la investigación actual. Analizaré datos españoles sobre accidentes de tráfico para evaluar cómo los conductores en España reaccionan a las primas de seguro ajustadas según su comportamiento a lo largo de los años. Esto permitirá entender mejor el impacto sostenido de los modelos PHYD en la seguridad vial dentro del contexto español. No obstante, la cuestión de cómo los modelos PHYD se adaptan a los rápidos avances en la tecnología de los vehículos, queda fuera del alcance de mi investigación y deberá ser abordada en futuros estudios.

Capítulo 3. DESCRIPCIÓN DE LOS DATOS

Este capítulo describe la base de datos utilizada, las características y las técnicas empleadas para su limpieza y preparación. El objetivo es asegurar la comprensión profunda de los datos. Se centra en el análisis empírico para desarrollar un modelo predictivo que pueda estimar la probabilidad de accidentes vehiculares basándose en variables como sexo, comunidad autónoma, edad y tipo de vehículo.

3.1.1 ORIGEN DE LOS DATOS

Los datasets utilizados en este estudio provienen de la Dirección General de Tráfico (DGT), lo que proporciona una base sólida y oficial para el análisis de la siniestralidad vial y los vehículos. Cada conjunto de datos corresponde a un año específico, cubriendo un periodo continuo desde 2017 hasta 2022, lo que permite realizar un análisis exhausto de la evolución de la siniestralidad y las características relativas a los accidentes en España.

El uso de esta información oficial asegura un alto grado de fiabilidad y consistencia en los datos. Estas bases de datos no solo reflejan las cifras absolutas, sino que también permiten profundizar en dinámicas y en relaciones causales detrás de las estadísticas.

3.1.2 DATASETS USADOS

Se cuenta con dos series de datasets:

1. **Datos sobre los vehículos**: Incluyen información detallada sobre la cantidad y características de los vehículos registrados, desglosados por tipo y antigüedad, así como datos demográficos de los conductores. Estos datos están estructurados por municipio, lo que permite un análisis geográfico detallado.

2. **Datos de Siniestralidad**: Contienen registros de accidentes, especificando el tipo de vehículo involucrado y la gravedad de las consecuencias para los ocupantes y peatones. Al igual que el otro dataset, estos datos están organizados por municipio y se proporcionan coordenadas geográficas para posibilitar estudios de localización específica.

Dado que los conjuntos de datos contienen información complementaria, es beneficioso unirlos. Para fusionarlos se ha utilizado identificadores comunes como el código INE, provincia y comunidad autónoma y año. El resultado es un dataset de 50.000 datos de entrada.

3.1.3 INFORMACIÓN DEL DATASET

Las variables de interés en los datasets sobre los vehículos incluyen, entre otras, la cantidad total de vehículos por tipo y su antigüedad media, así como la distribución de la población y del censo de conductores por género. En los datasets de siniestralidad, se destacan las variables relativas al número de accidentes con víctimas, la cantidad de fallecidos y heridos por tipo de vehículo, y la localización geográfica de estos eventos. Estas variables serán fundamentales para entender la relación entre la composición del parque automotor y la incidencia de los accidentes.

La correcta caracterización y entendimiento de estas bases de datos son esenciales para el desarrollo de un modelo predictivo robusto que busca estimar la probabilidad de accidentes basándose en factores demográficos y técnicos relevantes.

3.2 LIMPIEZA Y PREPARACIÓN DE LOS DATOS

El proceso de limpieza y preparación de los datos es una fase crítica en cualquier análisis estadístico y modelado predictivo. Este proceso asegura que los datos sean precisos y estén en un formato adecuado para realizar el análisis deseado. Dada la naturaleza de los datos recolectados de la DGT y su importancia en el estudio, se adoptaron pasos metódicos para garantizar su integridad.

3.2.1 LIMPIEZA DE DATOS

Se realizó una primera limpieza con la herramienta Excel, ya que los datos se descargaron en ese formato y eran más manipulables. El primer paso fue la creación de las coordenadas. Es decir, los dataset mencionados mostraban la información respecto a cada pueblo o ciudad de España. Mediante una función de Excel se ha creado otras dos columnas referentes a la latitud y longitud. Con esta información adicional se ha podido crear después mapas de calor.

Posteriormente se ha proseguido con la eliminación de registros duplicados, lo que asegura la unicidad de cada evento. También se realizó una verificación de coherencia, donde se comprobó que las categorías y rangos de las variables estuviesen dentro de los parámetros esperados. Por ejemplo, se revisaron las edades de los conductores y la antigüedad de los vehículos para identificar y rectificar valores atípicos o errores de entrada.

Además, se identificaron y corrigieron errores en las entradas de texto, tales como variaciones en la nomenclatura de los municipios o errores tipográficos, estandarizando los nombres. Para asegurar que todos los datos eran numéricos se ha comprobado mediante filtros que estuvieran convertidos en datos tipo float.

3.2.2 TRANSFORMACIÓN DE LOS DATOS

La transformación de los datos geográficos se realizó utilizando el lenguaje de programación Python y la biblioteca `pandas`, una herramienta muy útil para el análisis y manipulación de datos. Los pasos específicos en este proceso incluyeron:

1. **Incorporación del Año**: Se añadió una nueva columna al conjunto de datos para indicar el año correspondiente al dataset. Esto facilita el análisis longitudinal posterior y permite una fácil agregación de datos a lo largo del tiempo.
2. **Estandarización de Coordenadas Geográficas**: Se corrigió el formato de las coordenadas geográficas para asegurar la consistencia.

3. **Limpieza de Valores Nulos**: Se eliminaron las filas donde la latitud o longitud eran valores nulos. Además, en el caso de las columnas relacionadas con los accidentes, también se eliminaron los registros sin información sobre accidentes con víctimas para enfocarse en datos completos y precisos.
4. **Filtrado Geográfico**: Se establecieron y aplicaron límites geográficos para incluir solo aquellos registros dentro del territorio español, asegurando que el análisis sea pertinente al contexto nacional. Esto se logró mediante el establecimiento de límites mínimos y máximos para la latitud y longitud que abarcan la geografía de España.

Este proceso de transformación es crítico para la confiabilidad de los análisis geográficos y asegura que los datos estén en un formato adecuado para ser utilizados en herramientas avanzadas de análisis espacial y visualización.

Capítulo 4. ANÁLISIS EXPLORATORIO

El análisis exploratorio es una fase crítica en la que se examinan los datos preliminares para descubrir patrones, detectar anomalías, probar hipótesis y verificar supuestos con la ayuda de resúmenes estadísticos y representaciones gráficas. Vamos a dividir este epígrafe en dos secciones, una para los datos sobre la siniestralidad en España y otro sobre los vehículos. Separando los dos tipos de datos que se tienen se podrán analizar de forma más exacta los datos.

La base de este estudio es una recopilación de datos proporcionada por la Dirección General de Tráfico (DGT) de España, que incluye detallados registros de accidentes vehiculares. Este capítulo visualiza los datos para una comprensión más completa. El objetivo es asegurar la comprensión profunda de los datos.

4.1 SINIESTRALIDAD EN ESPAÑA

En esta sección se presentan los resultados del análisis exploratorio de los datos de siniestralidad en España, utilizando una serie de visualizaciones desarrolladas en Power BI y en Python.

4.1.1 VISUALIZACIÓN EN POWER BI

Se ha creado un panel (Figura 2) que proporciona una vista general de los accidentes ocurridos en el año seleccionado, en este caso 2021 y en las comunidades autónomas seleccionadas, en este caso Madrid, Comunidad Valenciana y Castilla La Mancha. Este panel permite una rápida comprensión de la gravedad y el alcance de la siniestralidad en el año en cuestión. Se puede observar e interactuar con el panel en el siguiente link:

Link al Cuadro de Mando (Siniestralidad)

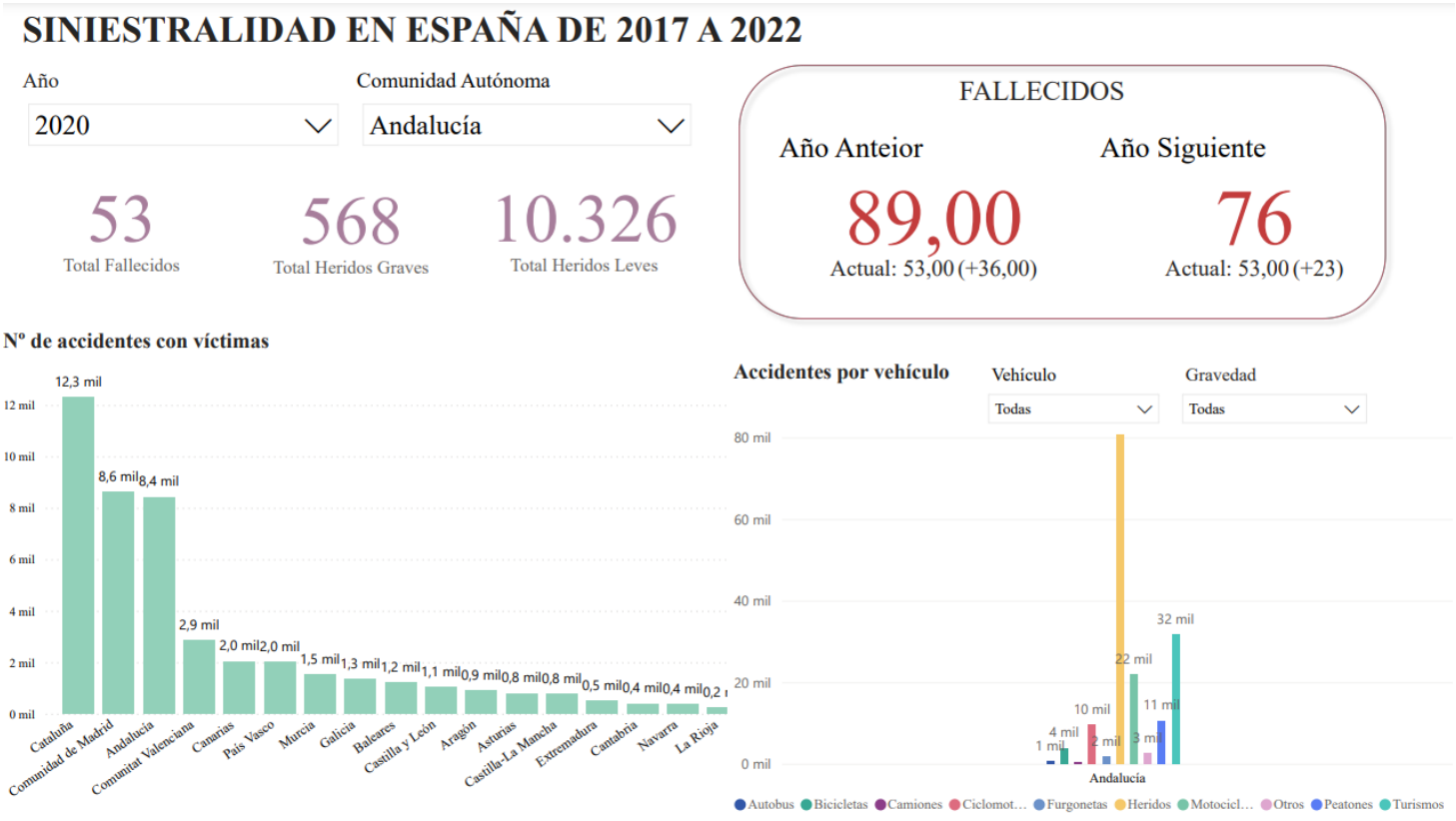

Figura 2: Panel principal de siniestralidad en España entre los años 2017 y 2022. Fuente: Elaboración propia a partir de datos de la DGT (DGT,2017-2022)

- Comparativa Anual de Fallecidos: Se muestra (Figura 3) una comparativa entre el número de fallecidos del año actual, el anterior y el año siguiente. Esta visualización destaca los cambios en la tasa de mortalidad, proporcionando una perspectiva sobre la efectividad de las medidas implementadas o cambios significativos en el comportamiento de conducción. Se observa que en 2021 respecto al 2020 ha disminuido en 29 fallecidos menos y respecto al 2022 hay 46 fallecidos más.

Figura 3: Comparativa anual de fallecidos para el año 2020 en España Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2021)

- Accidentes por Comunidad Autónoma: Utilizando un gráfico de barras, se comparan las comunidades autónomas en términos de número de accidentes con víctimas. Esta representación gráfica permite identificar rápidamente las áreas con mayor incidencia de siniestros y sugiere regiones donde se deben concentrar esfuerzos en prevención. En la figura 4 se muestra lo respectivo al año 2021.

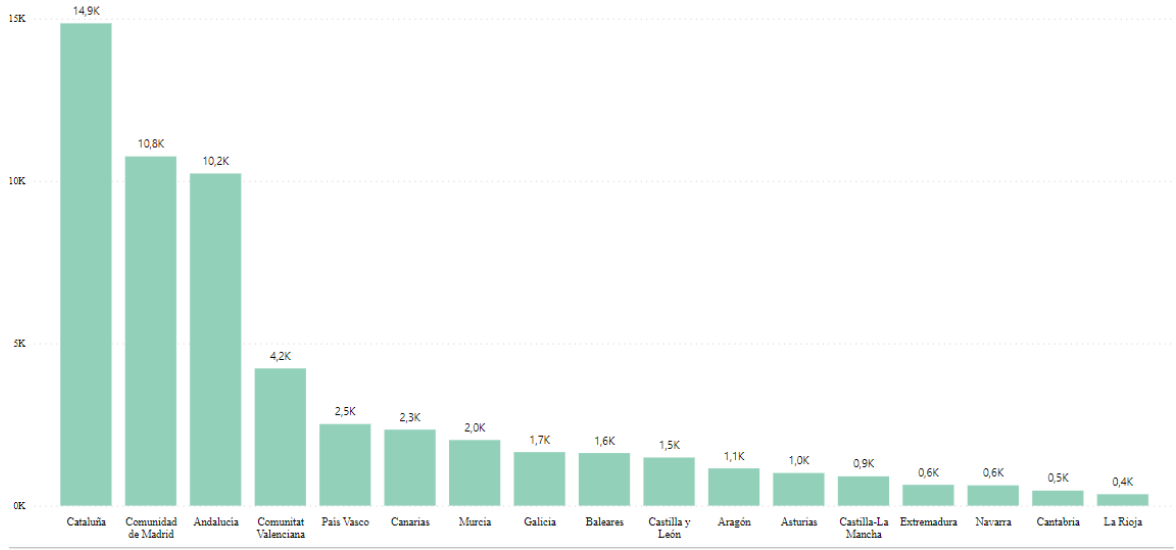

Figura 4: Comparativa en número de víctimas por comunidad autónoma para el año 2021 en España.
Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2021)

- Accidentes por Tipo de Vehículo: En un gráfico de dispersión, se comparan las comunidades autónomas en función del tipo de vehículo involucrado en los accidentes. Se puede filtrar por vehículo y gravedad del accidente. Esta visualización es esencial para entender cómo el tipo de vehículo puede influir en la siniestralidad. En la figura 5 vemos los datos para 2021 en la Comunidad de Madrid, Castilla la Mancha y País Vasco y sin tener ningún filtro. En la figura 6 se muestra lo mismo, pero filtrado para fallecidos en turismos.

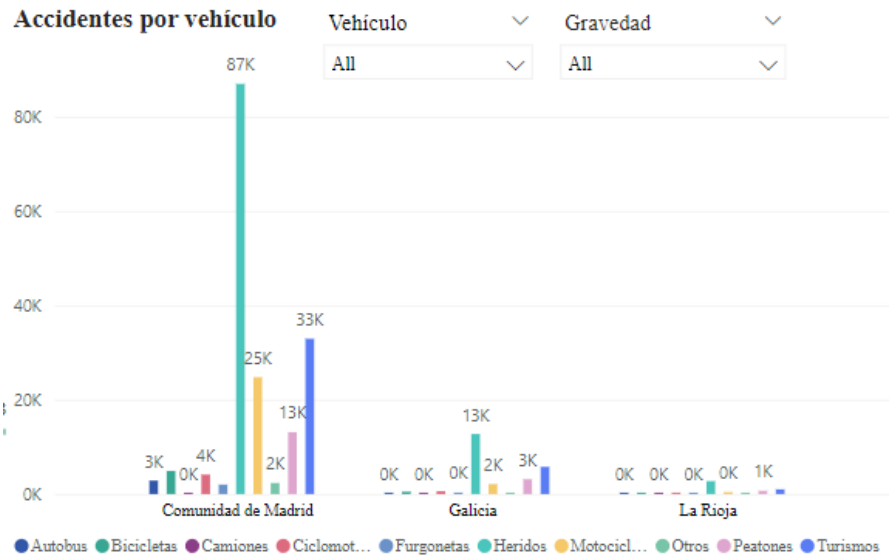

Figura 5: Accidentes por tipo de vehículo y gravedad en 2021 en España. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2021)

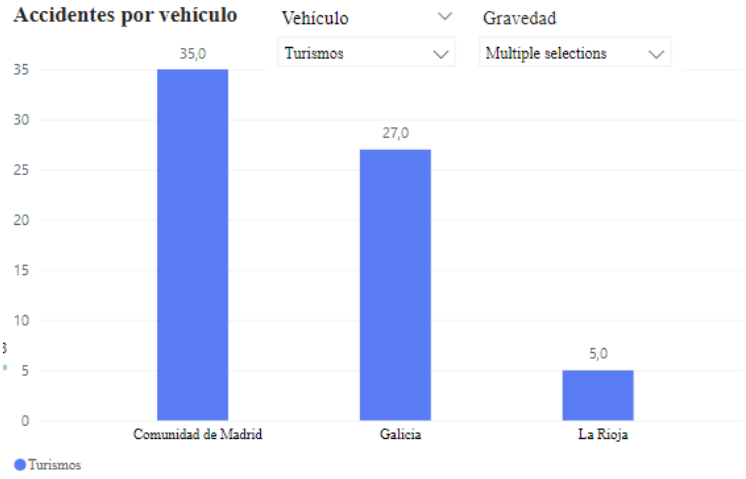

Figura 6: Fallecidos en 2021 por accidentes en turismos. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2021)

Para poder realizar una visualización con filtro independiente ha sido necesaria la creación de otra tabla con los mismos parámetros, pero utilizando la función de Power BI de dinamización de columnas. Con esto se ha conseguido que los datos estén separados por

vehículo y por tipo de accidente. Consiguiendo tres columnas como las mostradas en la figura 7.

1.2 Nº de accidentes	Vehículo	Gravedad
0	Bicicletas	Fallecidos
0	Bicicletas	Heridos Graves
0	Bicicletas	Heridos Leves
0	Ciclomotores	Fallecidos
0	Ciclomotores	Heridos Graves
0	Ciclomotores	Heridos Leves
0	Motocicletas	Fallecidos

Figura 7: Columnas resultantes al dinamizar las columnas de la tabla original. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2017-2022)

- Totales: Por último, para mostrar la siniestralidad en las carreteras de España durante un año específico. La interfaz permite la selección de una o varias Comunidades Autónomas, proporcionando así una visión adaptada y detallada del impacto de los accidentes de tráfico en las regiones seleccionadas y en año seleccionado. Se muestra en el periodo y comunidades seleccionadas el **total de fallecidos, total de heridos graves y total heridos leves en accidentes de tráfico**. En la figura 8 se muestra un ejemplo para el año 2020.

Figura 8: Totales de afectados en accidentes en el 2021. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2021)

Para realizar este panel se han necesitado una serie de funciones de DASH. En primer lugar, se ha sumado para cada vehículo el total de fallecidos, heridos graves y heridos leves. Con esta información se ha podido relacionar el número de accidentes para cada vehículo. Lo vemos en las ecuaciones 1 a la 7.

```
Bicicletas =
SUM(DGT[Bicicletas Fallecidos]) +
SUM(DGT[Bicicletas Heridos Graves]) +
SUM(DGT[Bicicletas Heridos Leves])
```

Ecuación 1: Total de accidentes para bicicletas

```
Camiones =
SUM(DGT[Camiones Fallecidos]) +
SUM(DGT[Camiones Heridos Graves]) +
SUM(DGT[Camiones Heridos Leves])
```

Ecuación 2: Total de accidentes para camiones

```
Ciclomotores =
SUM(DGT[Ciclomotores Fallecidos]) +
SUM(DGT[Ciclomotores Heridos Graves]) +
SUM(DGT[Ciclomotores Heridos Leves])
```

Ecuación 3: Total de accidentes para ciclomotores

```
Furgonetas =
SUM(DGT[Furgonetas Fallecidos]) +
SUM(DGT[Furgonetas Heridos Graves]) +
SUM(DGT[Furgonetas Heridos Leves])
```

Ecuación 4: Total de accidentes para furgonetas

```
Otros =
SUM(DGT[Otros Fallecidos]) +
SUM(DGT[Otros Fallecidos]) +
SUM(DGT[Otros Heridos Leves])
```

Ecuación 5: Total de accidentes para otros (otro tipo de vehículo)

```
Peatones =
SUM(DGT[Peatones Fallecidos]) +
SUM(DGT[Peatones Heridos Graves]) +
SUM(DGT[Peatones Heridos Leves])
```

Ecuación 6: Total de accidentes para peatones

```
Turismos =
SUM(DGT[Turismos Fallecidos]) +
SUM(DGT[Turismos Heridos Graves]) +
SUM(DGT[Turismos Heridos Leves])
```

Ecuación 7: Total de accidentes para turismos

Para sumar todos los fallecidos que hay se ha utilizado la ecuación 8. Con esta función se han sumado todos los fallecidos para todos los vehículos. Se ha hecho lo mismo para los heridos leves y graves, como se ve en las ecuaciones 9 y 10. Con esta información se muestran cifras totales de fallecidos o heridos en un año determinado.

```
Total Fallecidos =
SUM(DGT[Bicicletas Fallecidos]) +
SUM(DGT[Ciclomotores Fallecidos]) +
SUM(DGT[Motocicletas Fallecidos]) +
SUM(DGT[Turismos Fallecidos]) +
SUM(DGT[Furgonetas Fallecidos]) +
SUM(DGT[Camiones Fallecidos]) +
SUM(DGT[Autobus Fallecidos]) +
SUM(DGT[Otros Fallecidos]) +
SUM(DGT[Peatones Fallecidos])
```

Ecuación 8: Total de fallecidos para el año seleccionado

```
Total Heridos Graves =
SUM(DGT[Bicicletas Heridos Graves]) +
SUM(DGT[Ciclomotores Heridos Graves]) +
SUM(DGT[Motocicletas Heridos Graves]) +
SUM(DGT[Turismos Heridos Graves]) +
SUM(DGT[Furgonetas Heridos Graves]) +
SUM(DGT[Camiones Heridos Graves]) +
SUM(DGT[Autobus Heridos Graves]) +
SUM(DGT[Otros Heridos Graves]) +
SUM(DGT[Peatones Heridos Graves])
```

Ecuación 9: Total de heridos graves para el año seleccionado

```
Total Heridos Leves =
SUM(DGT[Bicicletas Heridos Leves]) +
SUM(DGT[Ciclomotores Heridos Leves]) +
SUM(DGT[Motocicletas Heridos Leves]) +
SUM(DGT[Turismos Heridos Leves]) +
SUM(DGT[Furgonetas Heridos Leves]) +
SUM(DGT[Camiones Heridos Leves]) +
SUM(DGT[Autobus Heridos Leves]) +
SUM(DGT[Otros Heridos Leves]) +
SUM(DGT[Peatones Heridos Leves])
```

Ecuación 10: Total de heridos leves para el año seleccionado

Se ha necesitado una serie de cálculos para poder representar los fallecidos en el año actual, en el año posterior y en el año anterior. En la ecuación 11 vemos la suma total de fallecidos, que va a coincidir con la ecuación 8. Se ha hecho para tener un control más organizado de las ecuaciones

```
Fallecidos Actual = SUM(DGT[Fallecidos])
```

Ecuación 11: Total de fallecidos para el año seleccionado

En la ecuación 12 vemos la función utilizada para calcular los fallecidos del año anterior. Para ello se ha utilizado una función de dash que utiliza el año seleccionado y le resta 1. Para este año calcula el total de fallecidos

```
Fallecidos Año Anterior =
CALCULATE(
    [Fallecidos Actual],
    DATEADD(DGT[Año], -1, YEAR)
)
```

Ecuación 12: Fallecidos del año anterior al seleccionado

En la ecuación 13 se ha hecho lo mismo que en la ecuación 12, pero sumando un año.

```
Fallecidos Año Siguiente =
CALCULATE(
    [Fallecidos Actual],
    DATEADD(DGT[Año], 1, YEAR)
)
```

Ecuación 13: Fallecidos del año siguiente al seleccionado

Cada una de estas visualizaciones aporta un valor único al análisis, permitiendo un entendimiento integral de la siniestralidad desde varias perspectivas. Los patrones observados establecen una base para las investigaciones más detalladas que seguirán y ayudan a formular hipótesis sobre las causas y factores que contribuyen a los accidentes en España.

4.1.2 VISUALIZACIONES EN PYTHON

En pyhton se han realizado mapas de calor para profundizar en la distribución geográfica de los accidentes. Estos mapas utilizan las coordenadas de latitud y longitud para mostrar concentraciones de accidentes, proporcionando una comprensión visual del panorama de la siniestralidad a lo largo de la geografía española.

En la figura 9 se muestra el mapa de calor relacionado con el número de accidentes. Se ha utilizado las columnas latitud y longitud para representar las coordenadas y la columna de número de accidentes con víctimas. Hay un filtro para poder visualizar el año y los círculos son más grandes o pequeños según la cantidad de accidentes. Si ampliamos y pinchamos en los círculos vemos el número de accidentes con víctimas en esa región o pueblo.

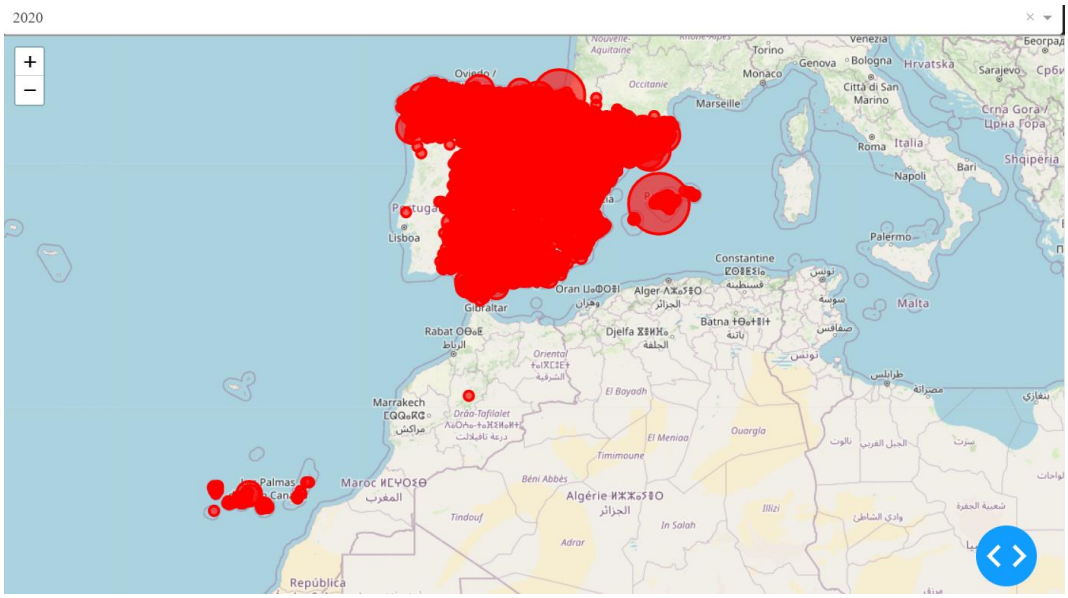

Figura 9: Mapa de calor que muestra el número de accidentes con víctimas en España. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2017-2022)

En la figura 10 vemos la distribución total en España para el tipo de severidad en España. El color amarillo quiere decir heridos leves, el naranja son los heridos graves y el rojo el número de personas fallecidas.

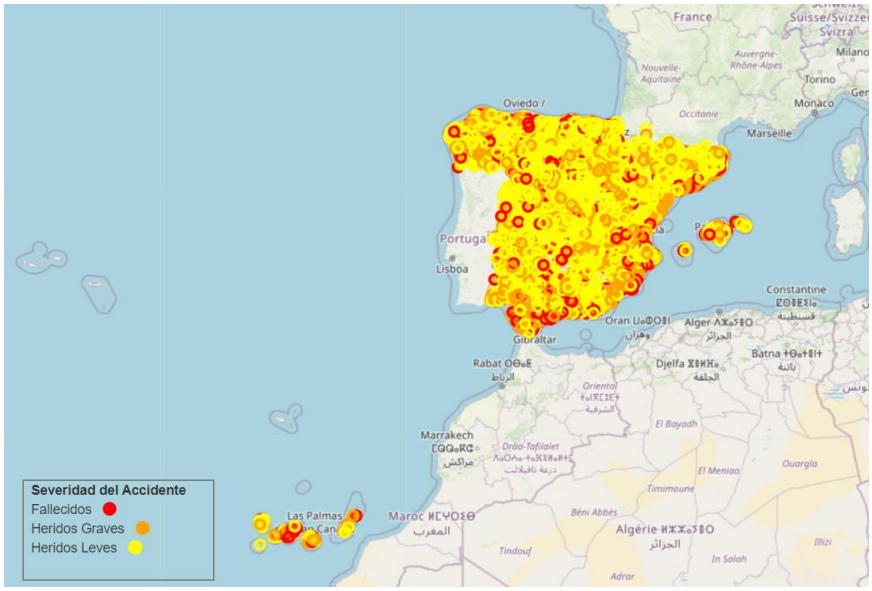

Figura 10: Severidad de los accidentes desde 2017 a 2022. Fuente: Elaboración propia a partir de datos de la DGT (DGT2017-2022)

4.2 DATOS SOBRE LOS VEHÍCULOS

En este apartado, se presenta un análisis exhaustivo de los datos municipales en España durante el periodo comprendido entre 2017 y 2022. La exploración de estos conjuntos de datos se llevó a cabo mediante las herramientas Power BI y Python.

4.2.1 VISUALIZACIONES EN POWER BI

Este panel (Figura 11) de datos ofrece una visión comprensiva de diferentes métricas relacionadas con el parque automovilístico y la demografía de conductores en España, abarcando el periodo de 2017 a 2022. A través de un conjunto de visualizaciones, se representan estadísticas clave sobre la evolución del censo de población, la distribución de vehículos por comunidades autónomas, la proporción de conductores por género y la vida media de los vehículos. Para ver e interactuar con el panel en el siguiente link:

Link al Cuadro de Mando (Datos Municiapales)

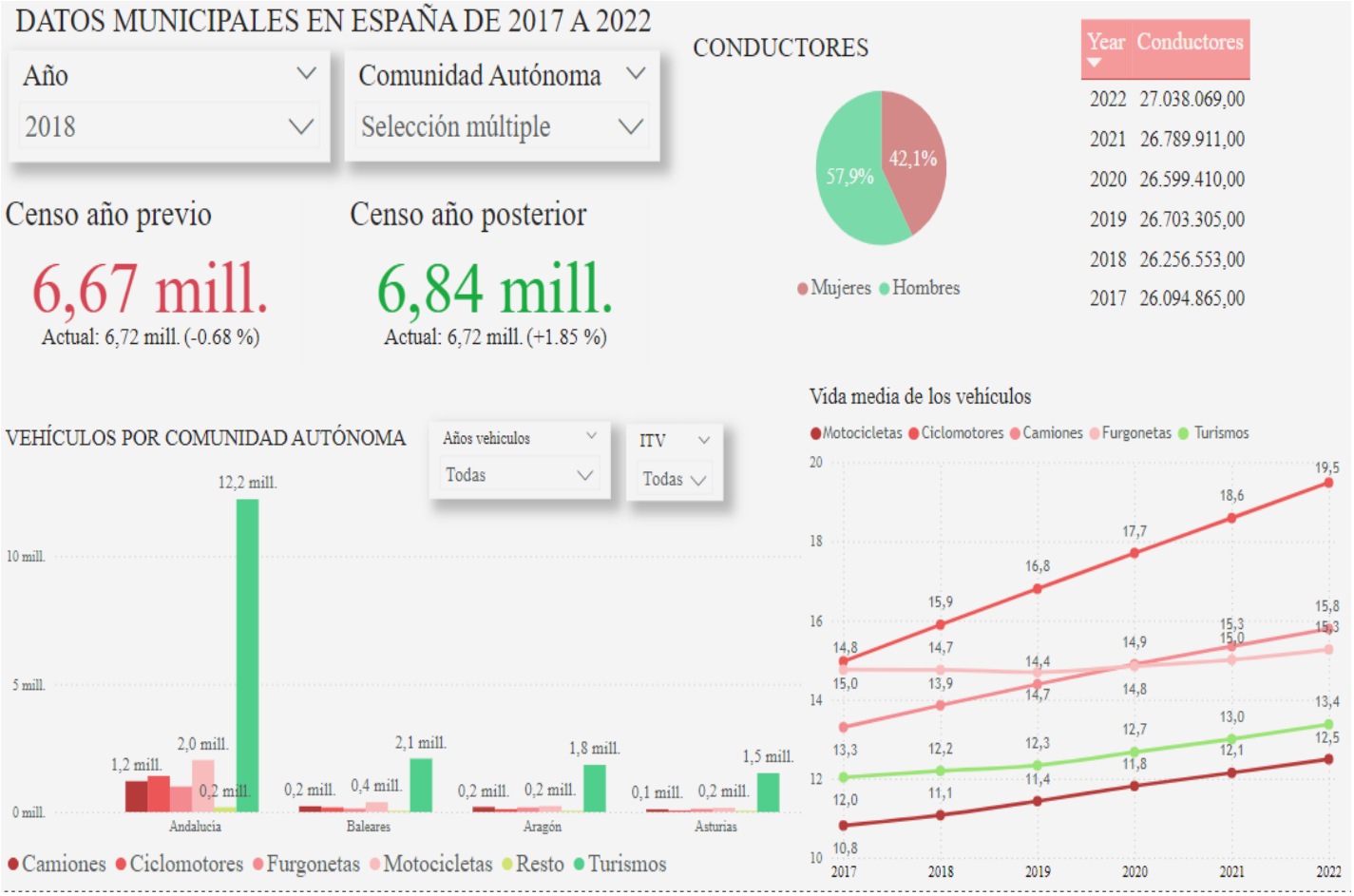

Figura 11: Panel sobre los datos municipales en España entre 2017 y 2022. Fuente: Elaboración propia a partir de datos de la DGT (DGT. 2017-2022)

- Evolución del censo poblacional: Esta visualización compara el censo del año previo y posterior. Según el año y comunidades autónomas que se seleccionen en los filtros muestra los datos correspondientes a ese periodo. Para la figura 12 se ha capturado el censo relativo al año 2018 y de las comunidades Andalucía, Baleares, Aragón y Asturias. Resaltando un leve descenso en el año anterior y un incremento en el año posterior. Se muestra tanto el número absoluto del censo como el porcentaje de cambio, proporcionando un contexto claro sobre el crecimiento o decrecimiento demográfico en ese periodo.

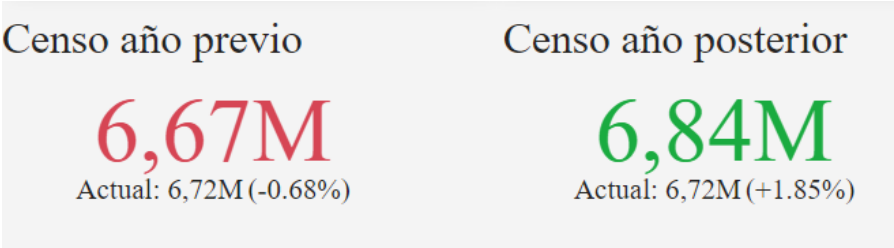

Figura 12: Evolución del censo poblacional en España en el año 2018. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2018)

- Vehículos por comunidad autónoma: Esta visualización de barras muestra la distribución de diferentes tipos de vehículos en varias comunidades autónomas. La categorización de los vehículos incluye **camiones, ciclomotores, furgonetas, motocicletas,** entre otros. El gráfico destaca la predominancia de determinados tipos de vehículos en cada región, lo cual puede ser indicativo de las características geográficas, económicas o de infraestructura de cada comunidad. El filtro general determina el año y las comunidades autónomas. Hay un filtro específico para esta visualización en el que se puede seleccionar los años del vehículo: más de 25 años, más de 15 años, más de 4 años, más de 8 años o menos de 4 años. Hay otro filtro específico que se selecciona si se quieren ver los vehículos con ITV o sin. En la figura 13 se muestra lo relativo al año 2018 y a las comunidades autónomas de Andalucía, Baleares, Aragón y Asturias. **No se ha seleccionado nada en el filtro específico**.

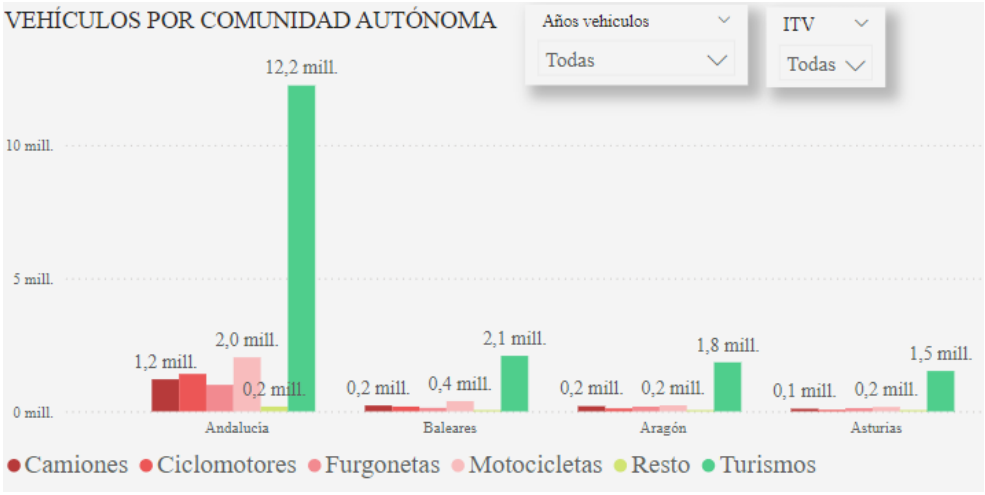

Figura 13: Distribución de vehículos en el año 2018 para Andalucía, Baleares, Aragón y Asturias. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2018)

- Proporción de hombres y mujeres: El gráfico circular representa la proporción de conductores hombres y mujeres. Este dato es fundamental para entender la distribución por género en la conducción. En la figura 14 se observa lo relativo al año 2018.

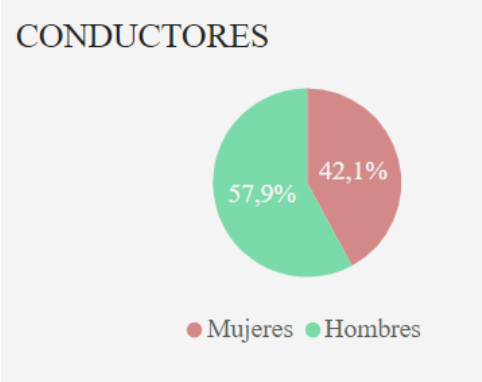

Figura 14: Proporción de conductores por sexo en el año 2018. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2018)

- Conductores al año: La tabla de la figura 15 muestra el número total de conductores registrados cada año desde 2017 hasta 2022. Esta información es crucial para evaluar el crecimiento o la contracción del censo.

Año	Conductores
2017	26.094.865,00
2018	26.256.553,00
2019	26.703.305,00
2020	26.599.410,00
2021	26.789.911,00
2022	27.038.069,00

Figura 15: Número de conductores desde 2017 hasta 2022. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2017-2022)

- Vida media de los vehículos: La figura 16 muestra la tendencia en la vida media de diferentes tipos de vehículos a lo largo del tiempo. Los incrementos o disminuciones en la durabilidad promedio de los vehículos pueden señalar cambios en los hábitos de consumo, la calidad de los vehículos y las políticas de renovación.

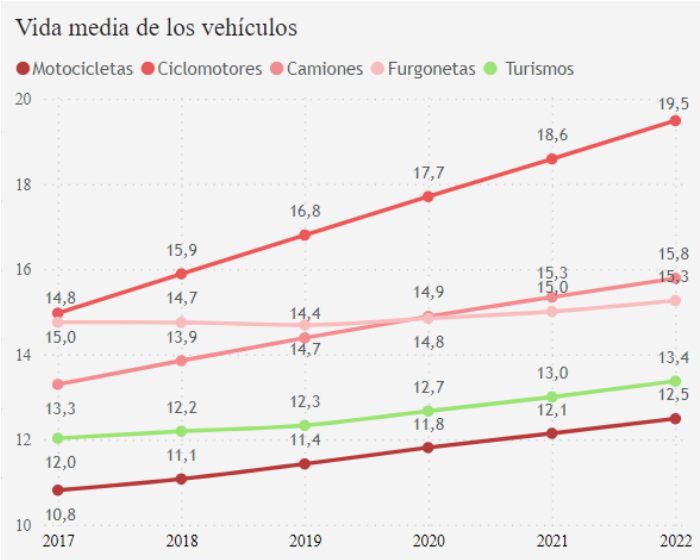

Figura 16: Vida media de los vehículos en España entre 2017 y 2022. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2017-2022)

Para realizar este panel (Figura 16) se han necesitado una serie de funciones de DASH. Para la ecuación 14 se ha sumado el valor de la columna censo conductores cuando la combinación de la columna código INE y Año es único. Hay que hacer eso debido a que anteriormente se ha dinamizado columnas y se repetían los valores.

```
Censo Actual =
SUMX(
    DISTINCT(
        SUMMARIZE('vehículos', 'vehículos'[Año ], 'vehículos'[Codigo INE],
"CensoUnico", MAX('vehículos'[Censo Conductores]))
    ),
    [CensoUnico]
)
```

Ecuación 14: Suma de la columna del censo de conductores

En las ecuaciones 15 y 16 se ha realizado lo mismo que anteriormente. Se ha calculado el censo en el año anterior y posterior.

```
Censo año anterior =
CALCULATE(
    [Censo Actual],
    DATEADD('vehículos'[Año ], -1, YEAR)
)
```

Ecuación 15: Suma de la columna censo conductores para el año anterior al seleccionado

```
Censo año siguiente =
CALCULATE(
    [Censo Actual],
    DATEADD('vehículos'[Año ], 1, YEAR)
)
```

Ecuación 16: Suma de la columna censo conductores para el año siguiente al seleccionado

4.2.2 VISUALIZACIONES EN PYTHON

En pyhton se han realizado mapas de calor para profundizar en la distribución geográfica de los accidentes. Estos mapas utilizan las coordenadas de latitud y longitud para mostrar la distribución de conductores, proporcionando una comprensión visual del panorama vehicular a lo largo de la geografía española.

En la figura 17 muestra un mapa de la península ibérica, marcando en rojo donde hay más conductores hombres y en azul donde hay más conductores mujeres. Este patrón denso de puntos rojos sugiere áreas donde la población masculina es significativamente mayor, y la ausencia de puntos que representen a la población femenina podría reflejar disparidades demográficas importantes o áreas con concentraciones específicas de población masculina. Para esta visualización hay un filtro de año.

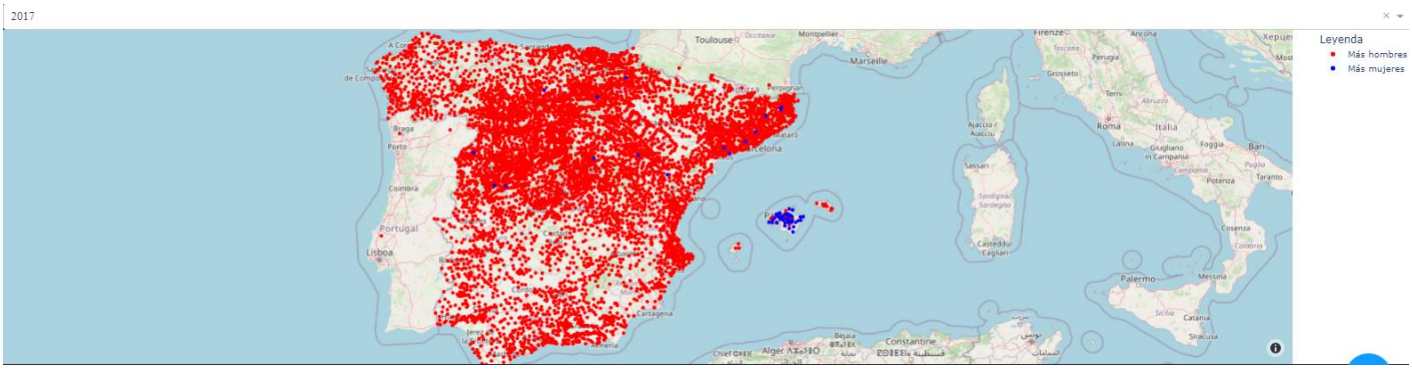

Figura 17: Mapa de calor que muestra la distribución de conductores en España para el año 2017. Fuente: Elaboración propia a partir de datos de la DGT (DGT, 2017)

4.3 CONCLUSIÓN

A través de visualizaciones avanzadas en Power BI y Python, se ha logrado una comprensión detallada de las dinámicas de accidentes y distribución demográfica de conductores. Las diferencias anuales en mortalidad y la comparativa por comunidades y tipo de vehículos han proporcionado datos sobre la efectividad de políticas de seguridad vial y han ayudado a identificar focos críticos para la prevención de accidentes. Por otro lado, la representación de los datos vehiculares ha revelado patrones en la evolución del parque automotor. Este análisis ha establecido una base sólida para poder continuar con el análisis.

Capítulo 5. MODELOS DESARROLLADOS

Este capítulo se dedica al desarrollo y evaluación de modelos predictivos basados en los datos descritos en el capítulo 3 para estimar el número de fallecidos en accidentes de tráfico. La finalidad es identificar factores significativos que contribuyan a los accidentes mortales y entender cómo estos pueden ser utilizados para informar políticas de seguridad vial y estrategias de tarificación en seguros de automóviles, enmarcado dentro del concepto de PHYD.

5.1 APRENDIZAJE AUTOMÁTICO

5.1.1 DEFINICIÓN

El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender a partir de datos e información sin ser explícitamente programadas. La principal característica del aprendizaje automático es su capacidad de mejorar el rendimiento en la resolución de tareas específicas mediante la experiencia.

5.1.2 UTILIDAD DEL APRENDIZAJE AUTOMÁTICO

La utilización de datos de la DGT en modelos de aprendizaje automático proporciona una oportunidad única para identificar y entender las variables más influyentes en los fallecimientos o heridos en accidentes de tráfico. Estos modelos pueden analizar grandes volúmenes de datos históricos sobre accidentes

Al aplicar técnicas de aprendizaje automático a estos datos, es posible identificar variables predictoras y así determinar qué características están más frecuentemente asociadas con los accidentes. Esto no solo ayuda a entender mejor las causas comunes de estos trágicos eventos, sino también a priorizar intervenciones.

5.1.3 PREDICCIÓN DE LAS VARIABLES

En la figura 18 se observa el boxplot que representa la distribución de tres categorías diferentes de resultados en accidentes de tráfico**: fallecidos, heridos hospitalizados y heridos no hospitalizados**. Estas son las tres variables que se han escogido para los modelos predictivos, ya que reflejan diferentes niveles de gravedad de los accidentes.

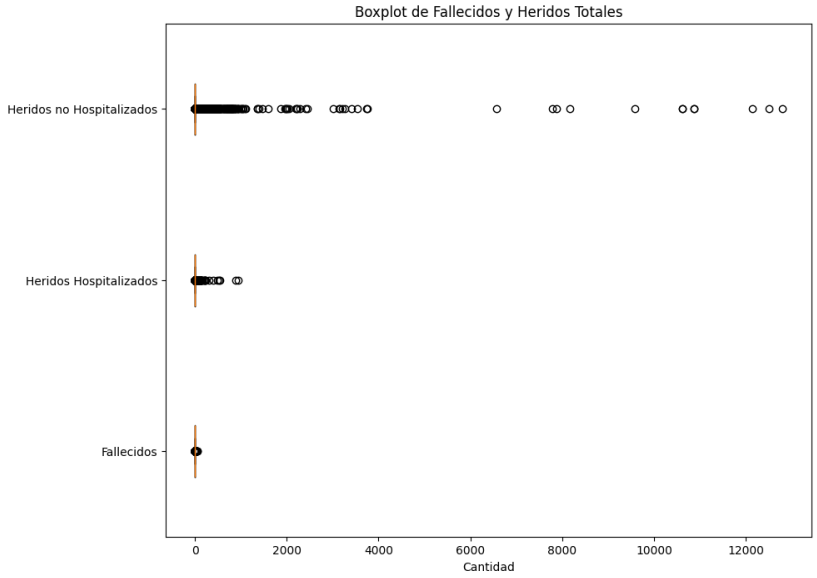

Figura 18: Outliers detectados en las tres variables a predecir: Fallecidos, Heridos Hospitalizados y Heridos no Hospitalizados. Fuente: Elaboración propia

En la figura 18, tenemos tres categorías: "Fallecidos", "Heridos Hospitalizados" y "Heridos no Hospitalizados". Cada caja indica la mediana (la línea central) y los cuartiles superior e inferior, mostrando dónde se concentra la mitad central de los datos en cada categoría. Las líneas verticales (bigotes) se extienden hasta los valores máximos y mínimos que no se consideran outliers (puntos fuera de los bigotes), mostrando el rango completo de los datos.

Se observa que hay una mayor cantidad de "Heridos no Hospitalizados" en comparación con las otras dos categorías, lo que indica que los accidentes que resultan en heridas sin necesidad de hospitalización son más comunes que aquellos que resultan en heridas graves (hospitalizaciones) o en fallecimientos.

5.2 MODELOS DE APRENDIZAJE AUTOMÁTICO

5.2.1 MODELO DE RANDOM FOREST

5.2.1.1 Descripción del Modelo

El modelo Random Forest es un modelo predictivo que forma parte de los métodos de ensemble learning. Se compone de una multitud de árboles de decisión, cada uno construido a partir de un subconjunto aleatorio de los datos de entrenamiento y de las características. Estos árboles operan como un conjunto donde cada uno vota por su predicción y la predicción final es determinada por la mayoría de los votos (en clasificación) o por el promedio de las predicciones (en regresión).

5.2.1.2 Justificación de su Uso

Se ha seleccionado Random Forest para este análisis ya que es capaz de manejar un gran número de entradas sin la necesidad de seleccionar manualmente las características. Esto es particularmente útil en datos de tráfico, donde pueden existir interacciones complejas entre las variables. Además, Random Forest puede manejar datos no lineales y proporciona una buena estimación de la importancia de las características.

5.2.1.3 Desarrollo del modelo

1. Preparación de los Datos: Se ha definido una lista de columnas que no contribuyen a la predicción de las variables y que podrían incluir datos redundantes o puede llevar a la fuga de información. Se eliminan estas columnas del dataset y las columnas restantes se definen como *features.*
2. Variable Objetivo: La columna a predecir se establece como la variable objetivo (*target*), que es la que el modelo intentará predecir.
3. Conversión de Categorías: Las variables categóricas se convierten en variables dummy para que el modelo pueda procesarlas adecuadamente. Una variable dummy es una variable binaria que se convierte en 1 si la variable está presente y en 0 si no.

4. División de Datos: Se dividen los datos en un conjunto de entrenamiento y otro de prueba. Usando el 80% para el entrenamiento y el 20% para la prueba.
5. Construcción del Modelo: Se crea un modelo de Random Forest (*RandomForestRegressor*) con 100 árboles, que es conocido por su capacidad para manejar una gran cantidad de características y por ser menos propenso al sobreajuste.
6. Entrenamiento del Modelo: Se entrena el modelo con el conjunto de datos de entrenamiento ("*X_train" e "y_train"*).
7. Evaluación del Modelo: Se hacen predicciones en el conjunto de prueba y se calculan métricas como el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2) para evaluar el rendimiento del modelo.
8. Visualización: Finalmente, se visualizan las 10 características que más influyen en la predicción de las variables

5.2.2 MODELO DE RED NEURONAL ARTIFICIAL (ANN)

4.4.2.1 Descripción del Modelo

La Red Neuronal Artificial es un modelo predictivo que imita la forma en que los humanos toman decisiones al utilizar unidades de procesamiento interconectadas llamadas neuronas. Está compuesta por capas de neuronas que procesan la información en secuencia, desde la entrada hasta la salida. Las capas intermedias, conocidas como capas ocultas, permiten al modelo aprender patrones complejos en los datos. Cada neurona en una capa está conectada con varias neuronas en la capa siguiente, permitiendo que el modelo realice cálculos complejos y ajuste sus pesos internos basados en el error de la salida predicha.

5.2.2.1 Justificación de su Uso

Se ha seleccionado una ANN para este análisis debido a su capacidad para modelar relaciones no lineales complejas y su eficacia en grandes volúmenes de datos, características comunes en los datos de tráfico. Las ANN son particularmente útiles en la detección de patrones intrincados que pueden ser difíciles de capturar con modelos lineales o incluso con métodos de ensemble como Random Forest.

5.2.2.2 Desarrollo del Modelo

Los primeros pasos son iguales que en el modelo de Random Forest. Una vez que hemos preparado los datos, seleccionando la variable objetivo, conversión de las variables categóricas y dividido los datos se ha construido el modelo de la red neuronal con varias capas ocultas y funciones de activación *'relu'*, junto con capas de dropout para reducir el riesgo de sobreajuste. El modelo se ha entrenado utilizando los datos de entrenamiento con un total de 20 épocas y un tamaño de lote de 10, ajustando los pesos de la red para minimizar el error cuadrático medio. Después del entrenamiento, el modelo se evalúa utilizando el conjunto de datos de prueba. Se calculan el MSE y el R^2 para comparar el rendimiento con el resto de las modelos. También vamos a visualizar las métricas de entrenamiento como la pérdida y el MSE durante cada época para monitorear la progresión del aprendizaje y ajustar si es necesario.

5.2.3 MODELO DE REGRESIÓN LASSO

5.2.3.1 Descripción del modelo

La regresión Lasso (*Least Absolute Shrinkage and Selection Operator*) es un modelo lineal que se utiliza tanto para la selección de características como para la regularización con el objetivo de mejorar la precisión predictiva y la interpretabilidad del modelo resultante. Lasso introduce una penalización L1.

La penalización L1 se observa en la ecuación 17 donde al error (MSE) se le suma el parámetro de regularización α multiplicado por la suma de los de los coeficientes del modelo (w_i). Con esto se consigue una reducción del sobreajuste ya que al agregar la penalización el modelo se vuelve menos propenso a ajustar el ruido en los datos de entrenamiento, mejorando su generalización a datos no vistos.

$$Función\ de\ coste = MSE + \alpha \sum_{i=1}^{n} |w_i|$$

Ecuación 17: Función de coste del modelo Lasso

5.2.3.2 Justificación de su Uso

Se ha seleccionado el modelo de regresión LASSO debido a su capacidad para manejar datasets con un gran número de características, especialmente cuando algunas de ellas no son relevantes. Esta característica es particularmente útil en el análisis de nuestro dataset, donde puede haber muchas variables que no contribuyen significativamente a la predicción. La capacidad de LASSO para realizar selección automática de características ayuda a identificar las variables más importantes y a mejorar la interpretabilidad del modelo.

5.2.3.3 Desarrollo del Modelo

Los primeros pasos son iguales que en el modelo de Random Forest y en ANN. Una vez que hemos preparado los datos, seleccionando la variable objetivo, conversión de las variables categóricas y dividido los datos se ha construido el modelo LASSO. Se ajusta el parámetro de regularización (α) para encontrar el equilibrio adecuado entre el ajuste del modelo y la complejidad. El modelo se ha entrenado con el conjunto de datos de entrenamiento, ajustando los coeficientes para minimizar el error cuadrático medio con la penalización L1. Se han hecho predicciones en el conjunto de prueba y se calculan métricas como el MSE y el R) para evaluar el rendimiento del modelo. Finalmente, se visualizan los coeficientes de las características en un gráfico de barras, destacando las características más importantes.

5.3 EVALUACIÓN DE LOS MODELOS

Al analizar un conjunto de datos con 50.000 entradas, he obtenido dos métricas importantes que miden el rendimiento de mi modelo predictivo: el MSE y el R^2

El MSE mide la calidad del estimador cuantificando la diferencia entre los valores predichos y los valores observados. Un MSE bajo indica un modelo más preciso. El R^2 indica la proporción de la varianza para la variable dependiente que es predecible a partir de las variables independientes, con valores más cercanos a 1 que sugieren un mejor ajuste del modelo.

El MSE es una medida que calcula el promedio de los cuadrados de los errores, es decir, la diferencia cuadrada entre los valores observados y los valores predichos por el modelo. Matemáticamente, se define como vemos en la ecuación 18.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

Ecuación 18: Expresión matemática de MSE

donde "***y***" son los valores reales, "$\bar{y}_i$" son los valores predichos, y "***n***" es el número total de observaciones.

El R^2 es una medida que indica qué tan bien las predicciones de un modelo se ajustan a los datos reales. Se define como la proporción de la varianza total de la variable explicada por el modelo. Matemáticamente se expresa como vemos en la ecuación 19.

$$R^2 = 1 - \frac{Suma\ de\ cuadrados\ de\ los\ residuos\ (SSR)}{Suma\ total\ de\ cuadrados\ (SST)}$$

Ecuación 19: Expresión matemática del coeficiente de determinación

Donde SSR es la suma de los cuadrados de las diferencias entre los valores observados y los predichos, y SST es la suma de los cuadrados de las diferencias entre los valores observados y su media. Un valor de R^2 de 1 indica un ajuste perfecto, y un valor de 0 indica que el modelo no explica nada de la variabilidad de los datos de respuesta.

5.3.1 RANDOM FOREST

5.3.1.1 Predecir Fallecidos

En la tabla 4 observamos las métricas de evaluación para la variable fallecidos. El MSE es de 0,0093, un valor bajo indica que el modelo tiene un mejor rendimiento en términos de precisión de las predicciones.

Vemos que el Random Forest ha demostrado un R^2 de 0.89, lo que indica que el modelo puede explicar aproximadamente el 89% de la variabilidad en la cantidad de fallecidos en accidentes de tráfico. Esto es un indicador sólido de un buen rendimiento, especialmente en conjuntos de datos con muchas variables predictoras y relaciones no lineales.

MSE	0.09303617979882649
R^2	0.8973393156769732

Tabla 4: Métricas de evaluación para Random Forest para la variable Fallecidos. Fuente: Elaboración propia

La importancia de las características mostradas en la figura 19 enseña qué variables tienen más peso a la hora de predecir fallecidos, Un coeficiente positivo indica que, manteniendo todas las demás variables constantes, un aumento en esa característica está asociado con un aumento en el número de fallecidos.

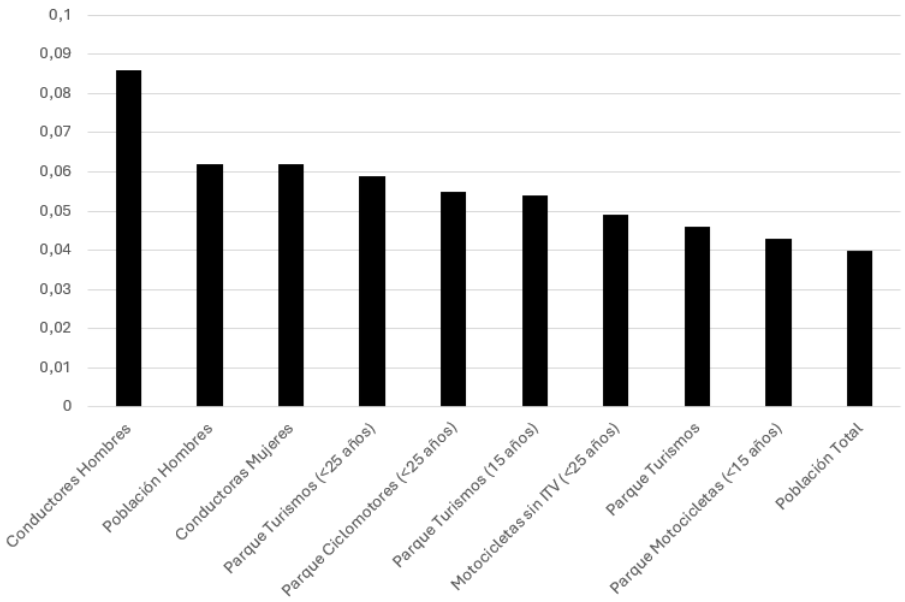

Figura 19: 10 características clave para predecir Fallecidos en el modelo Random Forest. Fuente: Elaboración propia

5.3.1.2 Predecir Heridos Hospitalizados

En la tabla 5 vemos que el MSE es de 51 y el R^2 es de 0,78, lo que significa que aproximadamente el 78% de la variabilidad en el número de heridos hospitalizados es explicado por las variables independientes utilizadas en el modelo.

Este es un resultado significativo, demostrando que el modelo es bastante eficaz en capturar las relaciones entre las características y la cantidad de heridos hospitalizados.

MSE	51.00899462489523
R^2	0.7857939739491225

Tabla 5: Métricas de evaluación para Random Forest para la variable Heridos Hospitalizados. Fuente: Elaboración propia

El análisis de la importancia de las características revela cuáles son los predictores más influyentes en la estimación de heridos hospitalizados. Lo vemos en la figura 20. Por ejemplo, los conductores hombres es el predictor más significativo, seguido por la población de hombres y de mujeres.

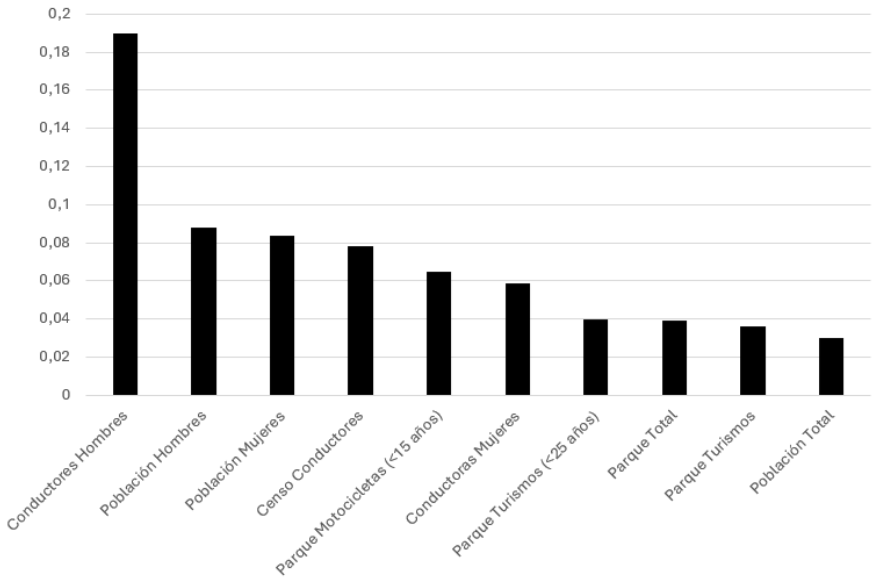

Figura 20: 10 características clave para predecir Heridos Hospitalizados en el modelo Random Forest. Fuente: Elaboración propia

5.3.1.3 Predecir heridos no hospitalizados

En la tabla 6 vemos el MSE y el R². El MSE es 1719. Este valor es tan alto debido a varías circunstancias:

- El dataset tiene 50.000 filas y el total de Heridos no Hospitalizados es de 400.000. El MSE es sensible a valores extremos porque eleva al cuadrado las diferencias entre los valores reales y los predichos.
- El MSE es una función cuadrática, por lo que penaliza más fuertemente los errores grandes. Aunque el modelo solo tenga algunos errores grandes y muchos pequeños, el MSE seguiría siendo alto.
- Algunos registros cuentan con decenas o cientos de heridos, entonces este valor no es tan alto, está dentro del rango apropiado.

Un R² de 0,976 significa que el modelo explica el 97,6% de la variabilidad en los datos de Heridos no Hospitalizados, lo cual es extremadamente alto y sugiere que el modelo tiene un ajuste muy bueno a los datos históricos.

MSE	1719.1421008067898
R^2	0.9761978142730545

Tabla 6: Métricas de evaluación para Random Forest para la variable Heridos no Hospitalizados. Fuente: Elaboración propia

Además, identificamos y visualizamos las características más importantes que contribuyen a las predicciones de heridos no hospitalizados en la figura 21. Destacando que si mantenemos todas las variables constantes un aumento de parque motocicletas en general va a estar asociado con un aumento en el número de fallecidos.

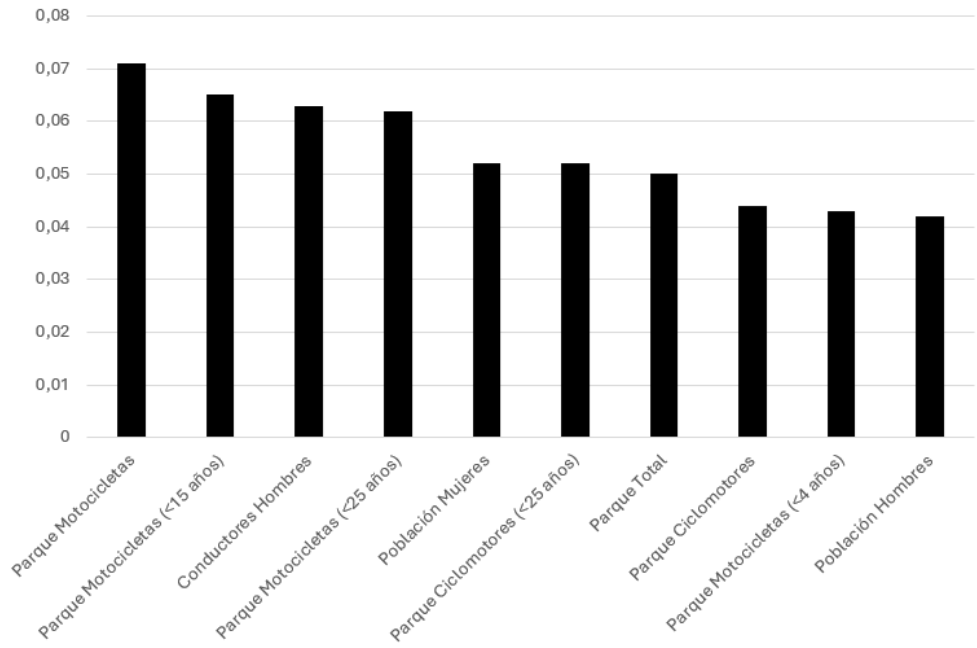

Figura 21: 10 características clave para predecir Heridos no Hospitalizados en el modelo Random Forest. Fuente: Elaboración propia

5.3.2 ANN

Se ha desarrollado un modelo ANN con una capa de entrada adaptada al número de características del conjunto de datos, seguida de dos capas ocultas con activación ReLU y Dropout para evitar el sobreajuste, y una capa de salida para regresión. El modelo se ha entrenado durante 20 épocas con un batch size de 10, utilizando un 20% del conjunto de entrenamiento para validación.

5.3.2.1 Predecir fallecidos

Los resultados que he obtenido al aplicar ANN a mi conjunto de datos se muestran en la tabla 7. Los valores sugieren que el modelo tiene un buen rendimiento en términos de minimizar los errores de predicción. Un R^2 de 0,8898 significa que aproximadamente el 88,98% de la varianza en la variable dependiente puede ser explicado por las variables independientes incluidas en el modelo.

MSE	0,09983004043282089
R^2	0,8898426366066452

Tabla 7: Métricas de evaluación para ANN para la variable Fallecidos. Fuente: Elaboración propia

La imagen 22 muestra el MSE de entrenamiento a lo largo de 20 épocas. El MSE de entrenamiento disminuye de manera consistente y suave, lo que es un buen indicador de que el modelo está aprendiendo de manera correcta de los datos.

El MSE de validación tiene picos significativos indicando que hay lotes de validación que son particularmente difíciles de predecir para el modelo. La tendencia general del MSE de validación, a pesar de los picos, es más o menos constante. Esto sugiere que el modelo no está mejorando su capacidad de generalización después de cierto punto, probablemente alrededor de la época 10. En resumen, aunque hay picos altos, el modelo generaliza correctamente.

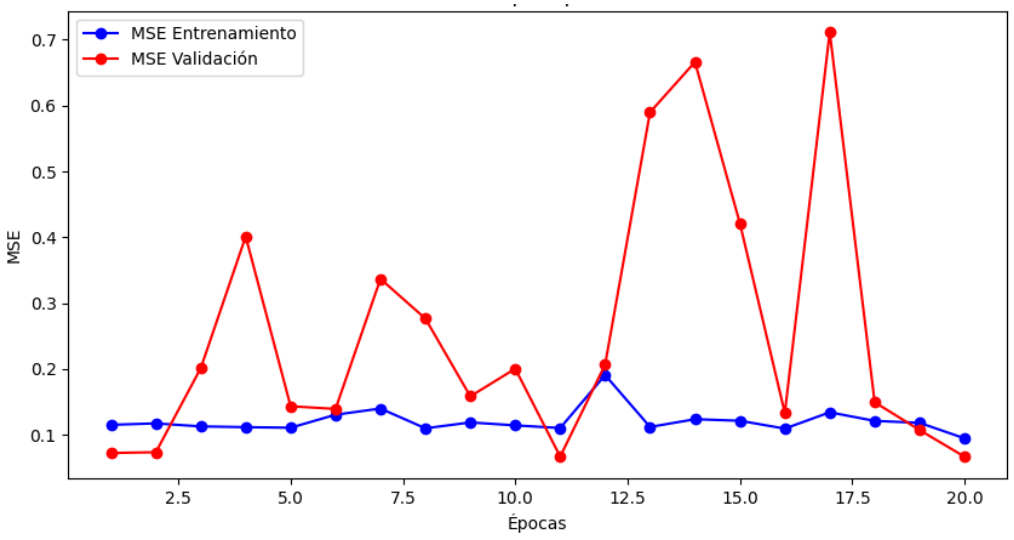

Figura 22: MSE por época en el modelo ANN para la variable Fallecidos. Fuente: Elaboración propia

5.3.2.2 Predecir Heridos Hospitalizados

En el caso de aplicar el modelo ANN para predecir los heridos hospitalizados obtenemos los resultados de la tabla 8.

MSE	74,55384326090025
R^2	0,6869202655497303

Tabla 8: Métricas de evaluación para ANN para la variable Heridos Hospitalizados. Fuente: Elaboración propia

En la gráfica 23 observamos el MSE de entrenamiento que comienza con un valor alto y va disminuyendo gradualmente, estabilizándose alrededor de la época 5. Va teniendo fluctuaciones, pero son menores que el conjunto de validación. El MSE de validación va teniendo mucha variabilidad a lo largo de las épocas con picos significativos en la época 5, 12 y 18.

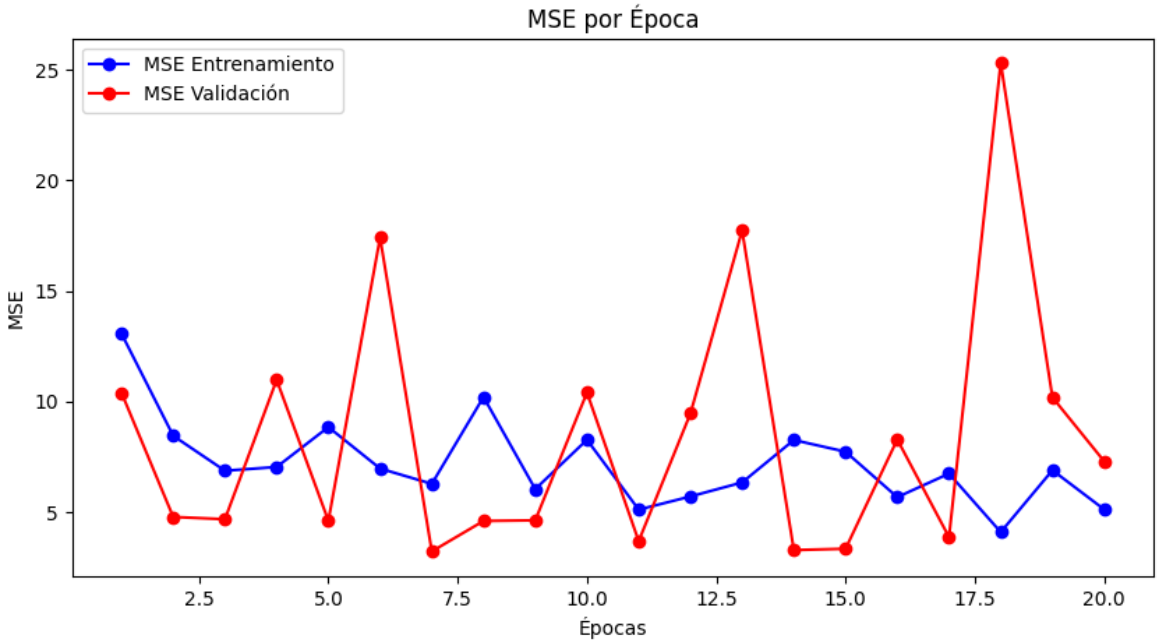

Figura 23: MSE por época en el modelo ANN para la variable Heridos Hospitalizados. Fuente: Elaboración propia

5.3.2.3 Predecir Heridos no Hospitalizados

En el caso de aplicar el modelo ANN para predecir los heridos no hospitalizados obtenemos los resultados de la tabla 9.

MSE	3057,94914123227
R^2	0,9576615142116492

Tabla 9: Métricas de evaluación para ANN para la variable Heridos no Hospitalizados. Fuente: Elaboración propia

En la figura 25 vemos el MSE de entrenamiento que empieza en un valor muy alto, pero se reduce drásticamente en las primeras épocas. Aunque hay ciertas fluctuaciones el MSE de entrenamiento es bastante estable. El MSE de validación también presenta fluctuaciones y comienza con un valor muy alto, aunque después se estabilizada. Destacan los picos 2 y 12. En este caso, aunque el modelo muestra una mejora general en el MSE, la alta variabilidad del MSE de validación sugiere la necesidad de un ajuste más fino de los hiperparámetros.

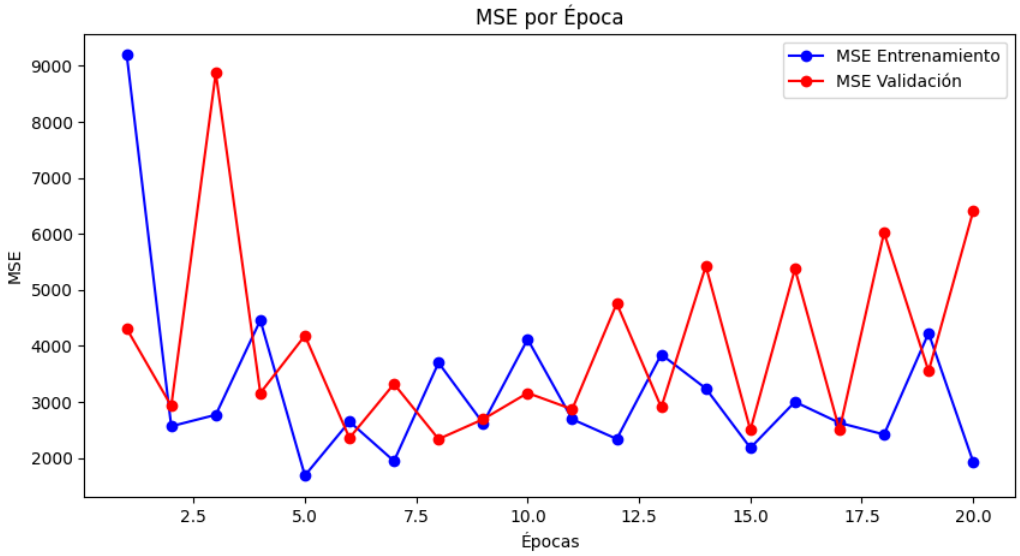

Figura 24: MSE por época en el modelo ANN para la variable Heridos no Hospitalizados. Fuente: Elaboración propia

5.3.3 LASSO

5.3.3.1 Predecir Fallecidos

En la tabla 10 observamos las métricas. Un MSE bajo indica que el modelo tiene un mejor rendimiento en términos de precisión de las predicciones. El valor de R^2 es bastante alto, se explica el 83% de la variabilidad de los datos.

MSE	0.14740066162816132
R^2	0.8373508797853195

Tabla 10: Métricas del modelo Lasso para la variable Fallecidos. Fuente: Elaboración propia

La importancia de las características mostradas en la figura 25 enseña qué variables tienen más peso a la hora de predecir fallecidos. Un mayor de población de mujeres podría estar relacionado con más accidentes ya que tiene la importancia más alta.

El hecho de que varias características, como Población Total, tenga coeficientes iguales a cero indica que la regularización Lasso ha determinado que estas características no contribuyen significativamente a la predicción y las ha eliminado del modelo.

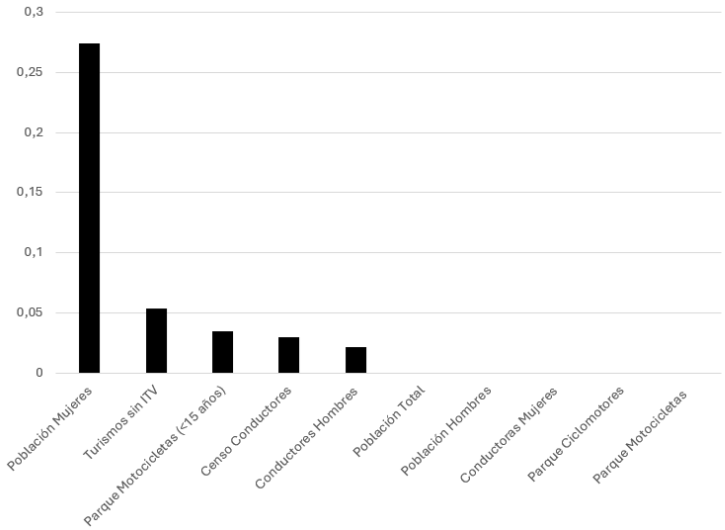

Figura 25: Características más relevantes para predecir los Fallecidos con el modelo Lasso. Fuente: Elaboración propia

5.3.3.2 Predecir Heridos Hospitalizados

En la tabla 11 observamos que el MSE es de 56,89. En este caso el MSE es más alto que anteriormente debido a que entre las 50.000 filas del dataset hay en total 25,144 heridos hospitalizados. La variabilidad sigue siendo alta como se observa en la tabla 8.

MSE	56.89454946532474
R^2	0.7610783071781172

Tabla 11: Métricas del modelo Lasso para la variable Heridos Hospitalizados. Fuente: Elaboración propia

La importancia de las características mostradas en la figura 26 enseña qué variables tienen más peso a la hora de predecir los heridos hospitalizados. En este caso la variable que más influye es la población de mujeres y las conductoras mujeres. En menor medida también influyen la antigüedad media del parque (<25 años) y la antigüedad de furgonetas.

También hay variables a 0 que el modelo ha decidido eliminar para facilitar la predicción. En el caso de una relación negativa quiere decir que si aumentan los ciclomotores (<4 años) entonces el número de fallecidos disminuye

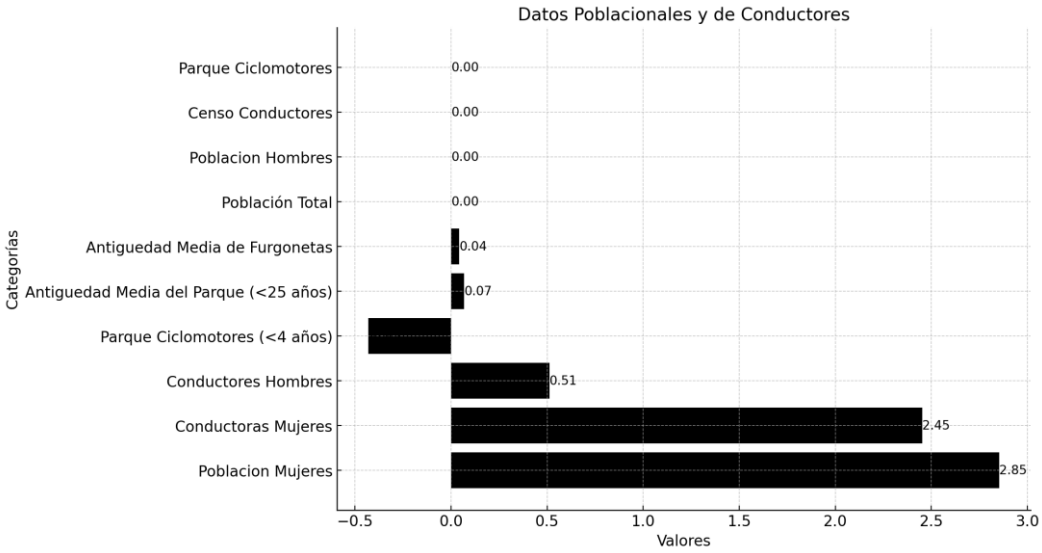

Figura 26: Características más relevantes para predecir los Heridos Hospitalizados con el modelo Lasso. Fuente: Elaboración propia

5.3.3.3 Predecir Heridos no Hospitalizados

Al haber 419.111 heridos no hospitalizado al sumar todas las entradas del dataset, el MSE va a ser más alto. Observamos en la tabla 12 que el valor es superior a los calculados con anterioridad. La variabilidad sigue siendo alta

MSE	4887.042421021981
R^2	0.9323370119863643

Tabla 12 Métricas del modelo Lasso para la variable Heridos no Hospitalizados. Fuente: Elaboración propia

La importancia de las características mostradas en la figura 27 enseña qué variables tienen más peso a la hora de predecir los heridos hospitalizados. En este caso la variable que más influye es el parque de motocicletas, Hay varias variables que si aumentan van a reducirse el número de heridos no hospitalizados como las motocicletas sin ITV.

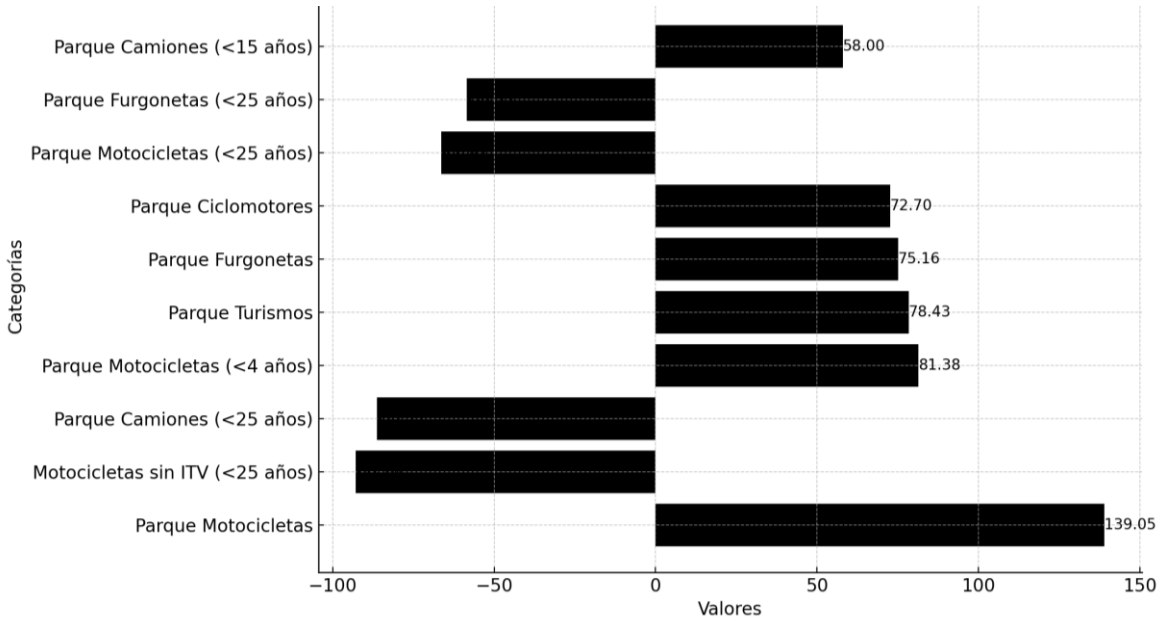

Figura 27: Características más relevantes para predecir los Heridos no Hospitalizados con el modelo Lasso. Fuente: Elaboración propia

Capítulo 6. ANÁLISIS DE RESULTADOS

En este capítulo se van a analizar y comparar de los distintos modelos de machine learning utilizados para predecir el número de fallecidos y heridos en accidentes de tráfico. Los modelos utilizados en el análisis son Random Forest, ANN y Regresión Lasso.

Se va a destacar cual es el modelo más efectivo y el que mejor ha realizado las predicciones. Se van a analizar con detalle los resultados obtenidos y la importancia de las variables utilizadas.

6.1 ANÁLISIS DE RESULTADOS

Los resultados obtenidos demuestran que el modelo Random Forest es el más efectivo para predecir el número de fallecidos y heridos en accidentes de tráfico. Esto se refleja en sus altos valores de R^2 y bajos valores de MSE en comparación con los otros modelos.

6.1.1 FALLECIDOS

Como podemos observar en las figuras 28 y 29 el modelo de Random Forest tiene el menor MSE y junto al modelo ANN el mayor R^2, lo que indica una alta precisión en las predicciones y una excelente capacidad de ajuste a los datos observados.

La variable conductores hombres es la variable más influyente para predecir fallecidos. Esto puede deberse a que los hombres, especialmente los jóvenes, tienden a tener comportamientos de conducción más arriesgados. Estudios han demostrado que los conductores hombres son más propensos a exceder los límites de velocidad, conducir bajo la influencia del alcohol y participar en comportamientos peligrosos, lo que aumenta la probabilidad de accidentes fatales (Saz, 2023).

La variable de población hombres también es bastante relevante a la hora de predecir fallecidos. En áreas con una mayor proporción de hombres, es probable que haya un mayor número de conductores hombres, lo que contribuye a un mayor riesgo de accidentes graves.

Por último, la variable de conductoras mujeres que, aunque menos influyente que los hombres, las conductoras mujeres también juegan un papel en la predicción de fallecidos. Las mujeres tienden a tener comportamientos de conducción más cautelosos, pero su influencia en el modelo sugiere que, en ciertas circunstancias, su participación en accidentes fatales no puede ser ignorada.

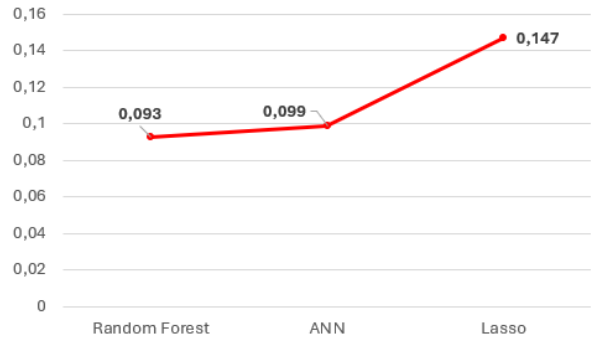

Figura 28: Comparación de la métrica MSE entre modelos. Fuente: Elaboración propia

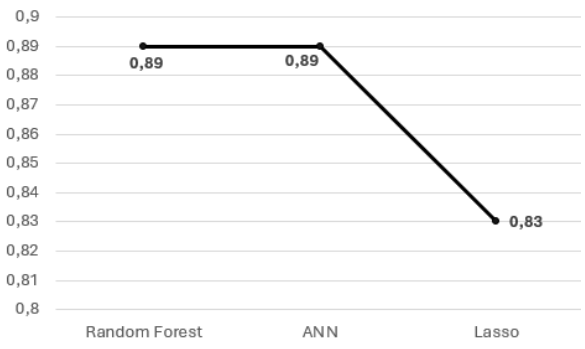

Figura 29: Comparación de la métrica R^2 entre modelos. Fuente: Elaboración propia

6.1.2 HERIDOS HOSPITALIZADOS

En las figuras 30 y 31 se aprecia las métricas calculadas para cada modelo a la hora de predecir heridos hospitalizados. En este caso Random Forest también destaca con el menor MSE y con el mayor R^2. Los valores son más altos que con la predicción de fallecidos debido a que en 50.000 entradas hay un total de 25.000 heridos hospitalizados, mientras que solo hay 2.500 fallecidos. Por eso, el valor de MSE es más alto.

En este caso, las medidas de los otros dos modelos son bastante más altas en MSE y bajas en R^2 que las de Random Forest. Sobre todo, las métricas del modelo ANN.

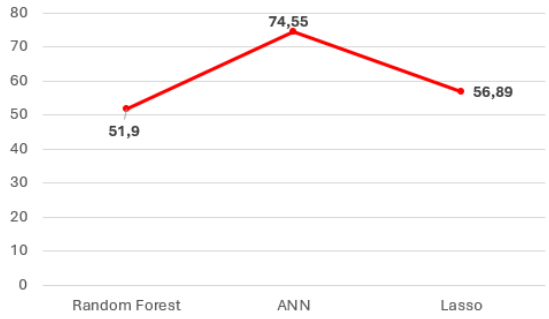

Figura 30: Comparación MSE entre modelos

Fuente: Elaboración propia

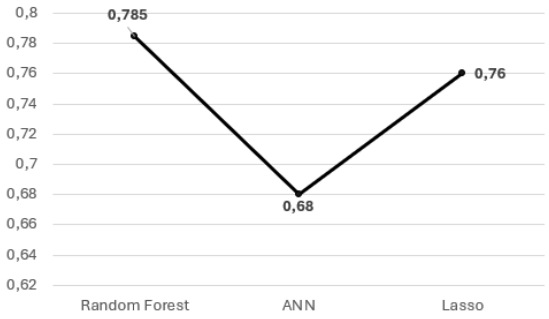

Figura 31: Comparación R^2 entre modelos

Fuente: Elaboración propia

Los conductores hombres provocan más heridos hospitalizados en accidentes de tráfico ya que los hombres tienden a hacer una conducción más temeraria. Esta variable está relacionada con la población de hombres, que también es muy importante a la hora de predecir heridos hospitalizados. La población de mujeres también afecta a la predicción, sugiriendo que las características demográficas tienen un gran impacto en las predicciones de heridos hospitalizados.

6.1.3 HERIDOS NO HOSPITALIZADOS

En las figuras 32 y 32 vemos que Random Forest demostró ser extremadamente efectivo en esta categoría, indicando una gran capacidad de ajuste y predicción. Al haber un total de 42.000 heridos no hospitalizados en 50.000 entradas el MSE es más alto.

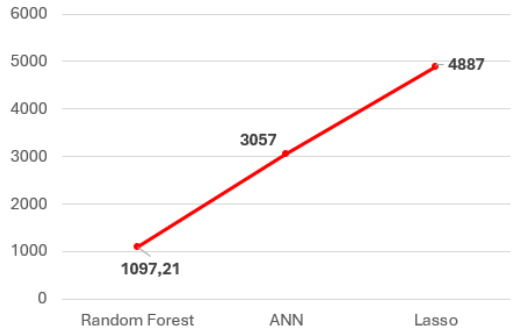

Figura 32: Comparación MSE entre modelos

Fuente: Elaboración propia

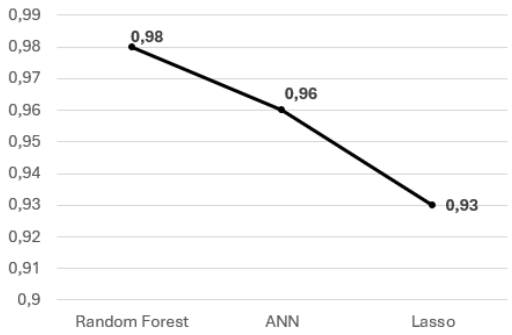

Figura 33: Comparación R^2 entre modelos

Fuente: Elaboración propia

La principal variable que afecta a heridos leves es el parque de motocicletas, y sobre todo las que tienen más de 15 años. Si se aumentan el número de estas en España puede llevar a un aumento de heridos en accidentes de tráfico. Los conductores hombres también afectan a el aumento de heridos en accidentes. Hay que destacar también la importancia del tipo y edad del vehículo en la gravedad de los accidentes.

Capítulo 7. CONCLUSIONES

Se ha demostrado que el mejor modelo para predecir fallecidos y heridos en accidentes de tráfico es el modelo de Random Forest, por ello se han utilizado las variables predichas por este modelo para identificar las variables principales en accidentes de tráfico en España. Sus altos valores de R^2 y bajos MSE demuestran su capacidad para manejar datos complejos y captar relaciones no lineales entre las variables. Las variables demográficas y características del parque vehicular resultaron ser las más influyentes, sugiriendo que cualquier estrategia de mejora de seguridad vial debe centrarse en estos factores.

Por otro lado, las Redes Neuronales Artificiales mostraron un rendimiento competitivo, especialmente en la predicción de heridos no hospitalizados, pero no superaron a Random Forest en ninguna categoría.

La Regresión Lasso, aunque útil para identificar la importancia de las características, no alcanzó la precisión de los otros dos modelos. Sin embargo, proporcionó información valiosa sobre las características más influyentes.

7.1 RECOMENDACIONES

Gracias al modelo Random Forest se han identificado las variables que más afectan a que se produzcan heridos y fallecidos en accidentes de tráfico. Una vez identificadas las variables se van a explicar una serie de recomendaciones que afectan a estas variables y así poder disminuir los accidentes de tráficos.

- Campañas específicas para hombres: Desarrollar campañas de sensibilización dirigidas a hombres jóvenes, destacando los riesgos de los comportamientos de conducción peligrosos y promoviendo hábitos de conducción responsables. Utilizar canales de comunicación efectivos para llegar a este público, como redes sociales, eventos deportivos o colaboraciones con figuras influyentes.

- Educación vial en los colegios: Implementar programas de educación vial específicos para niños y adolescentes, inculcando desde pequeños valores como la responsabilidad, el respeto a las normas y la conducción segura.
- Control de velocidad: Intensificar los controles de velocidad en zonas de riesgo, especialmente aquellas con alta presencia de conductores jóvenes. Implementar sistemas de control automatizado de la velocidad y sanciones más severas para los infractores.
- Control del alcohol y las drogas: Reforzar los controles de alcoholemia y drogas al volante, especialmente en horarios nocturnos y fines de semana.
- Aplicaciones móviles de seguridad vial: Promover el uso de aplicaciones móviles que den información sobre el estado de las carreteras, incidencias de tráfico o zonas de riesgo, ayudando a los conductores a planificar sus viajes de manera más segura.
- Cursos de conducción segura para mujeres: Ofrecer cursos de conducción segura específicos para mujeres, diseñados para abordar sus necesidades e inquietudes particulares. Estos cursos pueden incluir técnicas de conducción en situaciones de riesgo, manejo del estrés al volante o estrategias para afrontar situaciones de acoso o violencia vial.
- Revisión técnica obligatoria: Implementar una revisión técnica obligatoria más exhaustiva para motocicletas con más de 15 años, incluyendo aspectos como el estado de los frenos, neumáticos, suspensiones, sistema eléctrico y componentes de seguridad.
- Bonificaciones fiscales: Ofrecer bonificaciones fiscales o subvenciones a los propietarios de motocicletas antiguas que opten por renovarlas por modelos más nuevos y seguros.
- Programas de desguace: Implementar programas de desguace que premien la entrega de motocicletas antiguas para su baja definitiva, incentivando así la renovación del parque vehicular.

- Cursos de conducción segura para motoristas: Desarrollar cursos de conducción segura específicos para motoristas, con especial énfasis en la conducción de motocicletas antiguas, sus particularidades y los riesgos asociados.

El uso de técnicas de machine learning como Random Forest no solo proporciona predicciones precisas, sino que también ofrece una visión detallada de los factores más relevantes en la ocurrencia y gravedad de accidentes de tráfico, permitiendo a los responsables de políticas dirigir sus esfuerzos de manera más efectiva. Este estudio demuestra la validez y utilidad de estas técnicas para mejorar la seguridad vial y reducir el número de accidentes y sus consecuencias.

DECLARACIÓN USO DE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL GENERATIVA EN TRABAJOS DE FIN DE GRADO

Por la presente, yo, María Dolores Roca Morlán, estudiante de GITT + BA de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Implementación y Análisis de Modelos PHYD (*Pay How You Drive*) Utilizando Telemática y Machine Learning para la Reducción de Accidentes de Tráfico en España" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
4. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 19/06/2024

Firma: _____

Capítulo 8. BIBLIOGRAFÍA

1. Abdelrahman, A., Hassanein, H. S. y Abu-Ali, N. (2018). Data-driven robust scoring approach for driver profiling applications. *Proceedings of the IEEE International Conference on Communications* (ICC 2018), 1-8. https://doi.org/10.1109/ICC.2018.8422589
2. Abdelrahman, A. E., Hassanein, H. S. y Abu-Ali, N. (2020). Robust data-driven framework for driver behavior profiling using supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2020.3035700
3. Abdennour, N., Ouni, T. y Ben Amor, N. (2021). Driver identification using only the CAN-Bus vehicle data through an RCN deep learning approach. *Robotics and Autonomous Systems*. https://doi.org/10.1016/j.robot.2020.103707
4. Bernardi, M. L., Cimitile, M., Martinelli, F. y Mercaldo, F. (2018). Driver identification: A time series classification approach. *Proceedings of the International Joint Conference on Neural Networks* (IJCNN 2018**)**, 1-7. https://doi.org/10.1109/IJCNN.2018.8489087
5. Gao, G., Wüthrich, M. V. y Yang, H. (2019). Evaluation of driving risk at different speeds using clustering techniques**.** *Insurance: Mathematics and Economics*, 88, 108-119. https://doi.org/10.1016/j.insmatheco.2019.06.004
6. Huang, Y. y Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, 121, 113156. https://doi.org/10.1016/j.dss.2019.113156
7. Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A. y Santone, A. (2018). Cluster analysis for driver aggressiveness identification. *Proceedings of the 4th International Conference on Information Systems Security and Privacy* (ICISSP 2018)**,** 562-569. https://doi.org/10.5220/0006755205620569
8. Martinelli, F., Mercaldo, F. y Santone, A. (2020). Machine learning for driver detection through CAN bus. *Proceedings of the 2020 IEEE International Conference on Communications* (ICC 2020)**.** https://doi.org/10.1109/ICC.2020.9148856
9. Niño de Zepeda, M. V., Meng, F., Su, J., Zeng, X.-J. y Wang, Q. (2021). Dynamic clustering analysis for driving styles identification. *Engineering Applications of Artificial Intelligence*, 97, 104096. https://doi.org/10.1016/j.engappai.2020.104096

10. Usami, D. S., Persia, L., Picardi, M., Saporito, M. R. y Corazziari, I. (2017). Identifying driving behaviour profiles by using multiple correspondence analysis and cluster analysis. *Proceedings of the AIIT International Congress on Transport Infrastructure and Systems* (TIS 2017)**,** 108-119. https://doi.org/10.1201/9781315281896-108
11. Wang, Y., Zhao, T., Tahmasbi, F., Cheng, J., Chen, Y. y Yu, J. (2020). Driver identification leveraging single-turn behaviors via mobile devices. *Proceedings of the 2020 IEEE International Conference on Communications* (ICC 2020)**.** https://doi.org/10.1109/ICC.2020.9148856
12. Winlaw, M., Steiner, S. H., MacKay, R. J. y Hilal, A. R. (2019). Using telematics data to find risky driver behaviour. *Accident Analysis and Prevention*, 131, 131-136. https://doi.org/10.1016/j.aap.2019.06.003
13. Xun, Y., Liu, J., Kato, N., Fang, Y. y Zhang, Y. (2020). Automobile driver fingerprinting: A new machine learning based authentication scheme. *IEEE Transactions on Industrial Informatics***.** https://doi.org/10.1109/TII.2019.2946626
14. Zhao, X., Lu, T. y Dai, Y. (2021). Individual driver crash risk classification based on IoV data and offline consumer behavior data. *Mobile Information Systems*, 2021, 6784026. https://doi.org/10.1155/2021/6784026
15. Zhu, R. y Wüthrich, M. V. (2020). Clustering driving styles via image processing. *Annals of Actuarial Science***,** 1-15. https://doi.org/10.1017/S1748499520000317

Otras fuentes:

16. Dirección General de Tráfico. (2017). Mujeres conductoras en España. *Revista DGT*. https://revista.dgt.es/es/noticias/nacional/2017/05MAYO/0530mujeres-conductoras-en-Espana.shtml
17. Dirección General de Tráfico. (2017*). Datos municipales de siniestralidad 2017*. https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2017/
18. Dirección General de Tráfico. (2018). *Datos municipales de siniestralidad 2018*. https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2018/

19. Dirección General de Tráfico. (2019). *Datos municipales de siniestralidad 2019*. https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2019/
20. Dirección General de Tráfico. (2020*). Datos municipales de siniestralidad 2020.* https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2020/
21. Dirección General de Tráfico. (2021). *Datos municipales de siniestralidad 2021*. https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2021/
22. Dirección General de Tráfico. (2022*). Datos municipales de siniestralidad 2022.* https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Datos-municipales-siniestralidad-2022/
23. EPDATA. (s.f.). *Variación de víctimas mortales en accidentes de tráfico*. https://www.epdata.es/variacion-victimas-mortales-accidentes-trafico/4d59c35c-010c-407a-a1f4-b54573579a24
24. Saz, A. P. (2023, 2 septiembre). ¿Es cierto que los conductores jóvenes suelen tener más accidentes en España? *Motor y Movilidad*. https://www.20minutos.es/motor/movilidad/jovenes-tener-mas-accidentes-espana-5165039/

ANEXO I: INFORMACIÓN SOBRE EL CÓDIGO

Se incluye el archivo README.md del repositorio de Github para clarificar la organización y estructura del código. AUTOR: María Dolores Roca Morlán

AUTOR: María Dolores Roca Morlán
GRADO: 5ºGITT + BA

Descripción del proyecto

Este proyecto se centra en predecir las variables que más afectan en los accidentes de tráfico. Se han obtenido datos desde 2017 a 2022 sobre las distintas variables que afectan en los accidentes de tráfico, estos datos se han obtenido de la web oficial de la DGT. Para realizar esta predicción se han utilizado 3 modelos: Random Forest, ANN y Regresión Lasso. También se ha realizado un exhaustivo análisis de los datos y la limpieza de estos.

Estructura del Proyecto

Carpeta DatosMunicipales: Principales cifras de siniestralidad vial a nivel municipal. Datos consolidados. Son datasets que abarcan desde el año 2017 a 2022. Se realiza un análisis exploratorio de estos datos y una limpieza para el posterior tratamiento de los datos. Variables de interés

- Fallecidos: personas fallecidas en los 30 días posteriores a la ocurrencia del accidente.
- Heridos hospitalizados: personas que requirieron ingreso hospitalario de al menos 24 horasa consecuencia de un accidente de tráfico
- Heridos no hospitalizados: personas heridas en un accidente de tráfico que no hayan precisado hospitalización superior a veinticuatro horas

Carpeta Datos Siniestralidad Información sobre la composición del parque de vehículos y su antigüedad, y del censo de conductores. Datos consolidados. Son datasets que abarcan desde el año 2017 a 2022. Se realiza un análisis exploratorio de estos datos y una limpieza para el posterior tratamiento de los datos. Variables de interés

- Municipios Información sobre el municipio: Código municipio, nombre del municipio, provincia del municipio y CCAA del municipio.
- Población a 1 de enero del año de referencia Padrón Municipal.
- Censo de conductores Titulares de algún permiso o licencia de conducción en vigor.

- Número de automóviles registrados. Se excluyen vehículos en situación de baja definitiva o temporal.
- Número de los vehículos con menos de 25 años en los que no consta ITV en vigor. N
- Número de vehículos en función de la antigüedad del parque con menos de 25 años para cada clase de vehículos.
- Antigüedad media Promedio de la antigüedad del parque con menos de 25 años para cada clase de vehículos.

Carpeta ModelosRealizados En esta carpeta se encuentra el archivo en el que se aplican los 3 modelos a las 3 variables a predecir con las medidas realizadas para poder compararlos entre sí

LINK AL REPOSITORIO