COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales
ICADE

ANÁLISIS EXPLORATORIO Y MODELIZACIÓN DE LESIONES EN ACCIDENTES DE TRÁFICO EN EL CONDADO DE MONTGOMERY

Autor: Carmen Humanes Fernández
Director: Carlos Miguel Vallez Fernández

MADRID | Junio, 2026

RESUMEN

El presente trabajo analiza los patrones de siniestralidad vial en el condado de Montgomery (Maryland, Estados Unidos) y desarrolla un modelo capaz de predecir la gravedad de los accidentes, entendida como la presencia o ausencia de lesiones en las personas implicadas. Para ello se utiliza el conjunto de datos público Crash Reporting – Drivers Data, sobre el que se aplica una metodología basada en el proceso ETL (*Extract*, *Transform*, *Load*). Una parte sustancial del trabajo se dedica a la limpieza y transformación de unos datos reales, mayoritariamente categóricos y con numerosas inconsistencias, así como a la creación de nuevas variables mediante *feature engineering.*

A partir del conjunto de datos depurado se construyen y comparan distintos modelos de clasificación, abordando el fuerte desbalanceo de la variable objetivo mediante técnicas específicas como la ponderación de clases, el sobremuestreo (SMOTE) y el Balanced Random Forest, además del ajuste del umbral de decisión. Los resultados muestran una capacidad predictiva moderada, ya que la información disponible antes del accidente no captura por completo los factores que determinan la gravedad de las lesiones. Para interpretar el comportamiento del modelo se aplica la técnica SHAP, que revela que las variables más asociadas a las lesiones son las relativas a la dinámica del propio siniestro. Finalmente, los principales patrones se presentan en un panel de visualización. Más allá del rendimiento del modelo, el trabajo aporta una reflexión sobre las dificultades reales de predecir la siniestralidad a partir de datos abiertos y sobre las limitaciones que imponen los datos disponibles.

Palabras Clave: *analítica de datos, seguridad vial, accidentes de tráfico, modelos predictivos, desbalanceo de clases, interpretabilidad, SHAP.*

ABSTRACT

This project analyses road accident patterns in Montgomery County (Maryland, United States) and develops a model to predict accident severity, understood as the presence or absence of injuries among those involved. The study uses the public dataset Crash Reporting – Drivers Data, to which an ETL (Extract, Transform, Load) methodology is applied. A substantial part of the work is devoted to cleaning and transforming real-world data, mostly categorical and containing numerous inconsistencies, as well as to creating new variables through feature engineering.

Based on the cleaned dataset, several classification models are built and compared, addressing the strong imbalance of the target variable through specific techniques such as class weighting, oversampling (SMOTE) and the Balanced Random Forest, together with decision-threshold tuning. The results show a moderate predictive capacity, since the information available before the accident does not fully capture the factors that determine injury severity. To interpret the model's behaviour, the SHAP technique is applied, revealing that the variables most associated with injuries are those describing the dynamics of the accident itself. Finally, the main patterns are presented in a visualisation dashboard. Beyond the model's performance, this work offers a reflection on the real difficulties of predicting road accidents from open data and on the limitations imposed by the available data.

Keywords: *data analytics, road safety, traffic accidents, predictive models, class imbalance, interpretability, SHAP.*

ÍNDICE

1. Introducción 6

2. Estado del Arte 7

2.1 Aplicaciones del análisis de datos en movilidad y seguridad vial 7

2.2. Estudios previos sobre modelos predictivos de accidentes 11

3. Objetivos 16

4. Metodología (ETL) 16

4.1. Extract 18

4.2. Transform 20

4.3. Load 23

5. Desarrollo 24

5.1. Descripción del dataset y variables empleadas 24

5.2. Resultados del análisis exploratorio 26

5.3. Resultados del modelo predictivo 33

5.4. Interpretación y visualización en dashboard 46

6. Conclusiones 50

6.1. Principales hallazgos 50

6.2. Limitaciones del estudio 51

6.3. Futuras líneas de investigación 52

7. Bibliografía 56

8. Anexo 59

ÍNDICE DE FIGURAS

Figura 1: Hospitalizaciones por 100.000 personas en Nueva York comparadas con condados cercanos 8
Figura 2: Flujo metodológico del trabajo (proceso ETL y fases de análisis) 17
Figura 3: Distribución inicial de la variable Injury Severity 27
Figura 4: Proporción de accidentes con lesión según la hora del día 30
Figura 5: Proporción de accidentes con lesión según el tipo de vehículo 31
Figura 6: Circunstancias más frecuentes de los accidentes (nube de palabras de la variable 'Circumstance') 32
Figura 7: Proceso de construcción del modelo predictivo 34
Figura 8: Matriz de confusión del modelo final (Balanced Random Forest, umbral 0,5) 42
Figura 9: Importancia media de las variables según SHAP (modelo explicativo) 43
Figura 10: Efecto de las variables sobre la predicción de lesión (beeswarm) 44
Figura 11: Distribución espacial de accidentes de tráfico en el condado de Montgomery 47
Figura 12: Top 10 calles con más accidentes 48
Figura 13: Número de accidentes por hora del día 49
Figura 14: Tasa de lesionados por hora del día (Tableau) 49

ÍNDICE DE TABLAS

Tabla 1: Comparativa de estudios e iniciativas sobre análisis de datos en seguridad vial 10
Tabla 2: Principales estudios sobre modelos predictivos de accidentes (orden cronológico) 14
Tabla 3: Comparativa de los modelos probados (clase "con lesión") 38
Tabla 4: Resultados del modelo final (Balanced Random Forest) según el umbral de decisión 40

1. Introducción

En las últimas décadas, la movilidad urbana se ha vuelto uno de los principales desafíos que enfrentan las ciudades modernas, tanto por su impacto económico como por sus implicaciones en la seguridad y la sostenibilidad. Según la Organización Mundial de la Salud (OMS, 2023), los accidentes de tráfico provocan cada año aproximadamente 1,19 millones de muertes, y entre 20 y 50 millones de personas sufren lesiones no mortales, muchas de las cuales derivan en discapacidades físicas permanentes y en secuelas psicológicas de carácter prolongado. Además, estos accidentes generan pérdidas económicas equivalentes a alrededor del 3% del producto interior bruto (PIB) de la mayoría de los países, lo que convierte su análisis en un campo de creciente interés para administraciones públicas, entidades internacionales y la comunidad científica. En este contexto, la Asamblea General de las Naciones Unidas ha establecido como objetivo reducir a la mitad el número mundial de muertes y lesiones causadas por accidentes de tráfico para el año 2030 (A/RES/74/299), reforzando la necesidad de adoptar enfoques más eficaces en cuanto a seguridad vial. Bajo este marco, la analítica de datos se muestra como una herramienta fundamental para comprender patrones de siniestralidad urbana y así diseñar estrategias preventivas más eficaces.

El presente estudio se basa en el condado de Montgomery (Maryland, Estados Unidos), que proporciona un entorno ideal para el análisis de movilidad, al combinar áreas urbanas densas con zonas residenciales y carreteras estatales. Dicho análisis se apoya en un conjunto de datos oficial publicado por el gobierno del condado, bajo el nombre *Crash Reporting – Drivers Data.* El *dataset* reúne información detallada acerca de miles de accidentes de tráfico, incluyendo la fecha, hora, tipo de colisión, comportamiento del conductor y condiciones meteorológicas. La granularidad de estas variables permite estudiar con precisión los factores vinculados a la siniestralidad y su relación con el entorno urbano (Montgomery County of Maryland, 2025).

A partir de esta base de datos, el trabajo utiliza técnicas de *Data Analytics* para explorar y modelizar los patrones de accidentes en Montgomery. El proyecto se estructura siguiendo el proceso ETL (*Extract, Transform, Load*), que asegura la calidad de los datos antes de realizar un análisis exploratorio y la construcción de modelos predictivos. Esta clase de enfoque permite identificar relaciones no evidentes entre variables y anticipar así situaciones de riesgo, lo que ayuda a elaborar políticas de movilidad más seguras e

inteligentes. De este modo, el estudio se alinea con los objetivos de las *smart cities*, que buscan incorporar el análisis de datos en la planificación y gestión urbana.

En conclusión, el propósito principal de este TFG es analizar los patrones de siniestralidad urbana en Montgomery y desarrollar un modelo capaz de predecir la gravedad de los accidentes, entendida como la presencia o ausencia de lesiones en las personas implicadas. Más allá de la predicción, el trabajo presta especial atención a la interpretación del modelo, con el objetivo de identificar qué factores se asocian en mayor medida a la aparición de lesiones. La combinación de análisis exploratorio, modelización predictiva, interpretabilidad y visualización de resultados ofrece una visión integral del fenómeno. Asimismo, el estudio permite reflexionar sobre las dificultades reales de predecir un fenómeno tan complejo a partir de datos abiertos, así como sobre las limitaciones que imponen los datos disponibles. En última instancia, se pretende mostrar cómo la analítica de datos puede convertirse en una herramienta de apoyo para comprender la siniestralidad vial y contribuir a una movilidad urbana más segura y eficiente.

2. Estado del Arte

En este TFG se usarán los términos de “accidente” y “colisión” como sinónimos. Sin embargo, se usará “siniestralidad” para referirse al fenómeno agregado (frecuencia y gravedad) en un lugar y tiempo determinado.

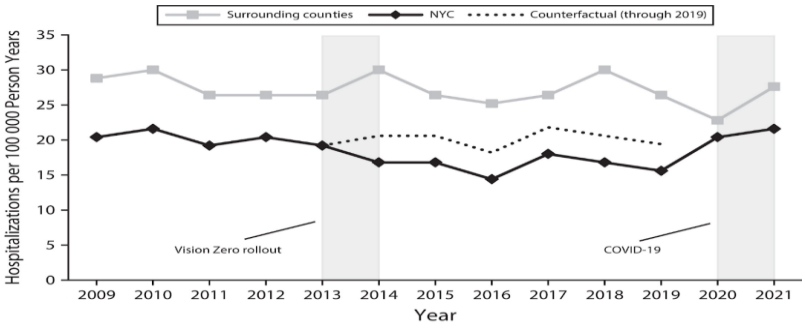
2.1 Aplicaciones del análisis de datos en movilidad y seguridad vial

El análisis masivo de datos se ha convertido en los últimos años en una herramienta esencial para comprender y gestionar la movilidad urbana y la seguridad vial. El creciente acceso a datos procedentes de dispositivos IoT, sensores y GPS está cambiando la forma en la que se tratan estos problemas, posibilitando modelos predictivos más precisos y estudios más integrales que superan las limitaciones de los análisis descriptivos tradicionales. Según una revisión reciente, el uso cada vez más extendido de sensores conectados ha dado lugar a numerosas investigaciones basadas en *big data* en casi todos los ámbitos de la seguridad vial. Esto abarca desde la monitorización de infraestructuras (condición del pavimento, detección de baches) hasta el seguimiento de la conducta del conductor (fatiga, distracciones) y el análisis en tiempo

real de los flujos de tráfico para reducir la congestión. En comparación con los métodos tradicionales centrados en análisis descriptivos, procesos manuales, y estadísticas simples, las aproximaciones modernas permiten una visión más integral del comportamiento del tráfico y de sus factores de riesgo asociados (Sohail et al., 2023).

En este ámbito, diversos ejemplos prácticos muestran el potencial de la analítica de datos. Dentro de la seguridad vial proactiva, que busca anticiparse al riesgo antes de que ocurra, las autoridades pueden detectar los llamados "puntos negros" o tramos donde se concentran los accidentes (TCA) a partir del análisis de datos históricos de accidentes georreferenciados, es decir, localizados geográficamente. Iniciativas como *Vision Zero*, que promueven la eliminación total de muertes y lesiones graves en el tráfico mediante estrategias preventivas basadas en datos, han usado datos de accidentes y modelos estadísticos para orientar intervenciones como el rediseño de intersecciones peligrosas o la adaptación de límites de velocidad. Un caso representativo es el de Nueva York, donde la implementación de *Vision Zero* se asoció con una reducción significativa de lesiones relacionadas con el tráfico entre 2014 y 2019, así como con ahorros de más de 90 millones de dólares en costes sanitarios, principalmente derivados de la reducción de hospitalizaciones y lesiones graves. Además, los resultados muestran un impacto positivo en grupos socioeconómicamente más vulnerables, lo que refuerza el valor de esta iniciativa no solo desde una perspectiva de seguridad vial, sino también de salud pública e igualdad social (*Vision Zero Network*, 2024).

Figura 1: Hospitalizaciones por 100.000 personas en Nueva York comparadas con condados cercanos



Fuente: Vision Zero Network (2024), Big cost savings & equity benefits highlighted in new Vision Zero study.

Estos enfoques en el ámbito de política pública se basan, en gran parte, en la incorporación de fuentes de datos novedosas y tecnologías emergentes, que permiten obtener información más detallada y en tiempo real sobre el comportamiento del tráfico y de los conductores. En este sentido, el análisis de datos de los vehículos (*telematics*) permite supervisar las flotas en tiempo real y de esta manera corregir comportamientos de riesgo (frenazos bruscos, excesos de velocidad) antes de que acaben en accidentes. La literatura existente evidencia que los datos procedentes de acelerómetros de *smartphones* pueden usarse para detectar irregularidades en la infraestructura vial, como baches o resaltos, con implicaciones directas en la seguridad (Salau et al., 2019, citado por Sohail et al., 2023). De forma similar, se han creado aplicaciones de visión artificial capaces de identificar a conductores distraídos o somnolientos a través de cámaras y algoritmos entrenados (Kaplan et al., 2015, citado por Sohail et al., 2023). Estas aplicaciones demuestran cómo la combinación de Big Data y algoritmos de inteligencia artificial tiene el potencial de abordar simultáneamente varias facetas del problema de la siniestralidad.

A nivel de gestión urbana, el análisis de datos de movilidad se utiliza para optimizar tanto la eficiencia del tráfico como la seguridad. Por una parte, las ciudades inteligentes están utilizando datos en tiempo real (procedentes de GPS, redes de telefonía móvil o plataformas como Google y Waze) para regular la temporización de los semáforos y gestionar el tráfico de forma dinámica, reduciendo así atascos y, con ello, la probabilidad de accidentes. Por otra parte, el uso de datos abiertos y colaborativos posibilita diagnósticos más exactos en cuanto a seguridad vial. Un ejemplo significativo es el de Bogotá, donde un estudio impulsado por el Banco Mundial combinó imágenes de las calles (obtenidas mediante Mapillary) con datos de accidentes y tráfico de Waze for Cities. A partir de estas dos fuentes, el modelo de machine learning estimaba el nivel de riesgo de cada tramo de carretera y lo comparaba con los accidentes reales. Los resultados fueron buenos, ya que el modelo identificaba correctamente los tramos de bajo riesgo en un 70% de los casos y los de alto riesgo en un 75%, lo que demuestra la utilidad de combinar distintos tipos de datos para detectar zonas peligrosas. En Nairobi (Kenia), el proyecto smarTTrans analizó más de 30.000 informes de accidentes recogidos a través de Twitter, donde los usuarios comparten información sobre atascos, accidentes e incidencias de tráfico. Con estos datos se creó una base de datos abierta que permitió estudiar dónde se localizaban los accidentes en la ciudad. Así se detectaron las vías y cruces donde más se concentraban, lo que ayudó a identificar las zonas de mayor riesgo

y a dirigir hacia ellas medidas concretas, como mejoras en la infraestructura o un mayor control policial. Estos ejemplos demuestran cómo las técnicas de analítica de datos, que van desde fuentes tradicionales hasta *Big Data* y *crowdsourcing*, se están utilizando con éxito en diferentes ciudades para detectar problemas relacionados con la seguridad vial y apoyar decisiones de política pública orientadas a una movilidad más segura y sostenible (World Bank, 2023).

La Tabla 1 recoge, a modo de resumen, una comparativa de los principales estudios e iniciativas mencionados:

Tabla 1: Comparativa de estudios e iniciativas sobre análisis de datos en seguridad vial

Iniciativa/ Estudio	Ubicación	Fuentes de datos	Enfoque/ Técnica	Principal Resultado
Vision Zero (Vision Zero Network, 2024)	Nueva York (EE. UU.)	Datos históricos de accidentes	Modelos estadísticos y estrategias preventivas basadas en datos	Reducción de lesiones de tráfico (2014-2019) y ahorro de más de 90 M$ en costes sanitarios, con impacto positivo en grupos vulnerables
Estudio del Banco Mundial (World Bank, 2023)	Bogotá (Colombia)	Imágenes de calles (Mapillary) y datos de tráfico y accidentes (Waze for Cities)	Machine learning combinando visión por ordenador y datos de siniestros	Identificación correcta de tramos de bajo riesgo (70%) y de alto riesgo (75%)
Proyecto smarTTrans (World Bank, 2023)	Nairobi (Kenia)	Más de 30.000 informes de accidentes recogidos en Twitter (crowdsourcing)	Creación de una base de datos abierta y análisis espacial	Detección de vías y cruces con mayor concentración de accidentes para dirigir intervenciones

Fuente: Elaboración propia

En general, los estudios que se han revisado muestran cómo la analítica de datos se ha vuelto una herramienta muy importante para entender y prevenir la siniestralidad vial. Sin embargo, gran parte de esta literatura se enfoca en grandes áreas metropolitanas o en el uso de fuentes de datos complejas, como IoT, imágenes o plataformas colaborativas. Dentro de este contexto, este TFG quiere aportar una aplicación práctica de estos métodos usando exclusivamente datos abiertos que provienen del condado de Montgomery. Así, se busca analizar hasta qué punto se pueden identificar factores de riesgo y construir modelos predictivos a partir de información accesible para administraciones públicas y organismos locales.

2.2. Estudios previos sobre modelos predictivos de accidentes

Si los trabajos anteriores muestran cómo la analítica de datos se puede usar para comprender y gestionar la movilidad urbana, también resulta interesante analizar cómo estas herramientas han evolucionado hacia la predicción de accidentes de tráfico. Durante décadas, el desarrollo de modelos predictivos de accidentes de tráfico ha sido un tema central en la investigación sobre seguridad vial. Históricamente, los primeros enfoques se basaban en modelos estadísticos tradicionales que estimaban la frecuencia de accidentes a partir de las características de la vía y el volumen de tráfico. Su finalidad era principalmente descriptiva y explicativa, más que predictiva. Por ejemplo, los modelos de regresión de Poisson y binomial negativa (GLM) se popularizaron para estimar el número anual de accidentes en un tramo o intersección concretos. De ellos, la binomial negativa resulta especialmente adecuada para los datos de accidentes, ya que estos suelen presentar sobredispersión, es decir, una variabilidad entre tramos mayor de lo que asume el modelo de Poisson (Greibe, 2003). Greibe (2003) desarrolló uno de los primeros modelos de predicción de accidentes para vías urbanas, relacionando variables de la infraestructura con la siniestralidad observada. De igual manera, se produjeron avances metodológicos como el método Empírico Bayesiano, que combina la información observada en un lugar concreto con los valores medios de referencia de sitios similares. De este modo, se reduce el efecto del azar y se mejora la estimación del riesgo en ubicaciones específicas (Huenupi et al., 2018). Estos enfoques tradicionales fueron las bases de manuales técnicos como el *Highway Safety Manual* (AASHTO, 2010), una guía de referencia muy utilizada por las administraciones de transporte que proporciona

funciones de predicción calibradas (ajustadas) para distintos tipos de vías y condiciones. Estas herramientas se usan habitualmente como apoyo para evaluar la seguridad y los beneficios de los proyectos viales (Elvik, 2007). Sin embargo, aunque resultan útiles para análisis generales, los modelos estadísticos lineales presentan limitaciones: parten de relaciones fijas entre las variables y no consiguen captar las interacciones más complejas que aparecen en los datos de accidentes.

Con el auge del aprendizaje automático, numerosos investigadores han explorado modelos de *machine learning* para predecir accidentes, con el objetivo de mejorar la precisión y la capacidad para identificar patrones no lineales. Entre las técnicas utilizadas se encuentran redes neuronales artificiales (ANN), árboles de decisión, bosques aleatorios o *Random* Forest y modelos bayesianos, entre otros. Çodur y Tortum (2015) llevaron a cabo un estudio emblemático en el que entrenaron una red neuronal artificial (ANN) con el objetivo de modelizar la frecuencia de accidentes en autopistas de Turquía, incorporando variables de tráfico y características geométricas de la vía. Los resultados mostraron un alto grado de ajuste entre los valores estimados y observados, con un coeficiente de determinación cercano a la unidad ($R^2 \approx 0{,}98$), lo que permitió un buen rendimiento en la identificación de segmentos de alto riesgo y evidenció que las ANN son capaces de captar relaciones no lineales en los datos sobre accidentalidad. Por otro lado, Hossain y Muromachi (2012) desarrollaron un marco basado en redes bayesianas para la predicción de accidentes en tiempo real en tramos básicos (tramos donde el flujo del tráfico es continuo, es decir, segmentos homogéneos) de autopistas urbanas en Japón. El modelo integró variables dinámicas como la velocidad, el flujo de tráfico y las condiciones meteorológicas, permitiendo estimar la probabilidad de ocurrencia de accidentes con varios minutos de antelación. Los resultados demostraron la viabilidad de este enfoque para anticipar situaciones de riesgo un tiempo antes de que ocurra y apoyar la gestión proactiva de la seguridad vial. Otra importante contribución es la de Basso et al. (2018), quienes utilizaron datos desagregados de tráfico (velocidad, volumen, densidad) para crear un modelo predictivo de colisiones en tiempo real en una vía expresa urbana (de alta capacidad), mejorando la detección de altos niveles de riesgo inmediatamente antes de un accidente. En general, estos métodos de machine learning han mostrado una tendencia a superar los modelos convencionales en capacidad predictiva, especialmente cuando se trata de clasificar la gravedad de accidentes o la probabilidad de colisión bajo determinadas condiciones, al capturar relaciones no lineales

y complejas entre múltiples factores de riesgo (Marcillo et al., 2022). Esta tendencia también se refleja en revisiones sistemáticas recientes, que analizan el uso cada vez mayor de algoritmos de clasificación y métodos de ensemble en la predicción de accidentes. Estas revisiones destacan que existe un compromiso entre la precisión y la interpretabilidad: los modelos más avanzados predicen los accidentes con mayor exactitud, pero resultan más difíciles de interpretar (Verma & Agarwal, 2025).

En los últimos años, la investigación se ha orientado hacia el aprendizaje profundo (*deep learning*), un tipo de técnica que emplea redes neuronales con múltiples capas capaces de aprender patrones complejos a partir de grandes volúmenes de datos. Combinado con datos masivos y de distinta naturaleza, este enfoque permite predecir los accidentes con mayor detalle, tanto en el espacio (dónde ocurren) como en el tiempo (cuándo ocurren). Ren et al. (2018) propusieron un modelo basado en *deep learning* para predecir el riesgo de accidentes a escala urbana, utilizando redes neuronales profundas capaces de incorporar información espacial y temporal de toda la ciudad. Este modelo permite analizar cómo el riesgo de siniestros varía tanto según la zona urbana como a lo largo del tiempo, lo que resulta útil para la planificación y gestión proactiva de la seguridad vial. Igualmente, Bao et al. (2019) desarrollaron un modelo de aprendizaje profundo espaciotemporal que combinaba varias fuentes de datos, como el tráfico en tiempo real, condiciones meteorológicas, características de las vías, etc., para anticipar colisiones a corto plazo en una ciudad. Estos estudios muestran que, al explotar grandes volúmenes de datos de distinta naturaleza, se logran mejoras significativas en la precisión de las predicciones, algo que no era posible con los métodos tradicionales. Sin embargo, también evidencian retos importantes: la calidad y la integridad de los datos se vuelve crítica, al igual que la necesidad de abordar problemas de desequilibrio (por ejemplo, los accidentes graves son eventos raros) y de alta dimensionalidad en las variables, ya que demasiadas variables pueden hacer el modelo más complejo y aumentar el riesgo de *overfitting*. Una revisión sistemática reciente de modelos predictivos basados en *machine learning* corrobora que la eficacia de estos métodos está fuertemente relacionada con la riqueza de los datos disponibles y con una adecuada integración de fuentes heterogéneas. Para investigaciones futuras, sugiere tener en cuenta factores aún poco explotados, como datos detallados del comportamiento del conductor o imágenes del entorno, y utilizar arquitecturas avanzadas de redes neuronales profundas (Marcillo et al., 2022)

Otro aspecto emergente en la literatura es la atención creciente a la interpretabilidad de los modelos predictivos de accidentes. Debido a que los responsables públicos necesitan entender qué factores explican el riesgo para diseñar intervenciones eficaces, algunos estudios recientes han empezado a explorar técnicas de inteligencia artificial explicable, que permiten interpretar las predicciones y vincularlas con variables concretas del tráfico y del entorno, ayudando a entender cómo cambia el riesgo. Por ejemplo, Parra et al. (2020) evaluaron el desempeño de modelos de *machine learning* explicables para la predicción de accidentes en California, combinando métodos como Random Forest y árboles de decisión con métodos de interpretación de variables. Los modelos alcanzaron un rendimiento predictivo superior al 70% en términos de F1-score, al tiempo que ayudaron a identificar los factores relacionados con la siniestralidad. Entre los resultados, se observó una fuerte relación entre las condiciones meteorológicas y la ocurrencia de accidentes, lo que demuestra el valor de estos enfoques no solo para predecir, sino también para entender y apoyar la toma de decisiones en seguridad vial. Esta forma de investigación apunta a un equilibrio entre precisión y transparencia, que permite que los modelos predictivos no solo anticipen dónde o cuándo ocurrirá un accidente, sino que también aporten *insights* accionables sobre por qué ocurriría (por ejemplo, si influye más la velocidad del vehículo, el diseño de la vía, la hora del día, etc.).

La Tabla 2 resume, en orden cronológico, los principales estudios sobre modelos predictivos de accidentes analizados en este apartado.

Tabla 2: Principales estudios sobre modelos predictivos de accidentes (orden cronológico)

Autor(es) y año	Técnica / Modelo	Ámbito y datos	Principal Resultado
Greibe (2003)	Modelos de regresión (GLM)	Vías urbanas	Uno de los primeros modelos de predicción de accidentes en entorno urbano, relacionando variables de infraestructura con la siniestralidad
Hossain y Muromachi (2012)	Redes bayesianas	Autopistas urbanas (Japón); variables dinámicas (velocidad, flujo, meteorología)	Predicción de accidentes en tiempo real, con capacidad de anticipar situaciones de riesgo unos minutos antes

Çodur y Tortum (2015	Red neuronal artificial (ANN)	Autopistas (Turquía); variables de tráfico y geometría de la vía	Alto ajuste ($R^2 \approx 0,98$) y buena identificación de segmentos de alto riesgo; confirma la capacidad de las ANN para captar relaciones no lineales
Basso et al. (2018)	Machine learning	Vía expresa urbana; datos desagregados de tráfico (velocidad, volumen, densidad)	Predicción de colisiones en tiempo real, mejorando la detección de riesgo justo antes del accidente
Ren et al. (2018)	Deep learning	Escala urbana; información espacial y temporal	Análisis de cómo varía el riesgo de accidentes según la zona y el momento, útil para la gestión proactiva
Bao et al. (2019)	Deep learning espaciotemporal	Ciudad; múltiples fuentes (tráfico, meteorología, características de las vías)	Anticipación de colisiones a corto plazo combinando datos heterogéneos
Parra et al. (2020)	ML explicable (Random Forest y árboles de decisión)	California; variables de tráfico y entorno	Rendimiento superior al 70% (F1-score) e identificación de factores de riesgo (p. ej., meteorología), combinando predicción e interpretabilidad

*Fuente**:** Elaboración propia*

En conclusión, el estado del arte sugiere que los modelos predictivos de accidentes han ido avanzando hacia enfoques cada vez más sofisticados, incorporando *big data, deep learning,* y fuentes de datos que antes estaban inexploradas, logrando así mejorar la predicción de accidentes en entornos urbanos complejos. En esta línea, el presente trabajo busca explorar hasta qué punto es posible aplicar estos métodos a un caso concreto como el del condado de Montgomery, valorando tanto sus posibilidades como sus limitaciones.

Asimismo, aunque la literatura reciente demuestra el potencial que tienen las técnicas avanzadas de *machine learning* y *deep learning* para mejorar la capacidad de predicción de los modelos de accidentes, todavía existe la necesidad de desarrollar

soluciones que combinen buen rendimiento predictivo con interpretación clara de factores de riesgo. Esto es especialmente importante para este TFG, cuyo objetivo no es solo predecir la ocurrencia de lesiones en accidentes, sino también identificar qué variables contribuyen más a este resultado. La metodología usada en este trabajo, que se desarrollará más adelante, busca precisamente mantener este equilibrio entre capacidad predictiva e interpretabilidad.

3. Objetivos

El objetivo principal de este trabajo es analizar los patrones de siniestralidad vial en el condado de Montgomery y desarrollar un modelo capaz de predecir la gravedad de los accidentes, entendida como la presencia o ausencia de lesiones en las personas implicadas, a partir de datos abiertos.

Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- Limpiar, transformar y preparar un conjunto de datos reales e imperfectos, dejándolo en condiciones adecuadas para su análisis.
- Explorar los datos para identificar los principales patrones y factores asociados a la siniestralidad en el condado.
- Construir, evaluar e interpretar distintos modelos predictivos, aplicando técnicas adecuadas para tratar el desbalanceo e identificando qué variables se asocian en mayor medida con la gravedad de los accidentes.
- Presentar de forma visual los principales patrones detectados mediante un panel de visualización.

4. Metodología (ETL)

Para el desarrollo de este trabajo se optó por una metodología basada en el proceso ETL (*Extract, Transform, Load*), muy habitual en proyectos de Business Analytics y análisis de datos. Este método permite estructurar el tratamiento de la información en varias fases diferenciadas, garantizando así que los datos empleados en el análisis sean consistentes, utilizables y adecuados para la construcción de modelos predictivos.

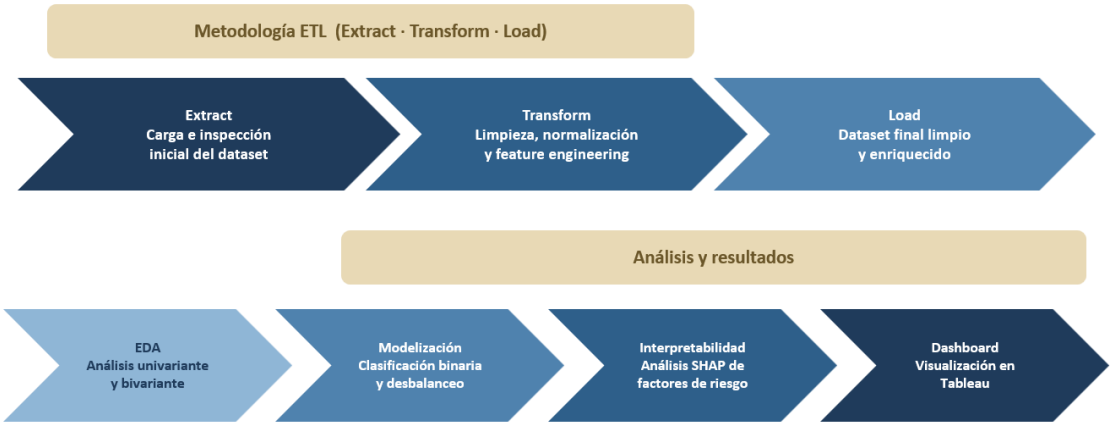
La primera fase, *Extract*, consiste en obtener y cargar de forma inicial los datos desde la fuente original. En este caso, se trabajó con un *dataset* público de accidentes de tráfico sacados del portal Data.gov. En esta etapa se realizó una inspección preliminar del conjunto de datos y se identificaron varios problemas relacionados con el formato y la calidad de la información.

La segunda fase, *Transform*, se refiere al proceso de limpieza y transformación de los datos. En esta etapa se trataron valores nulos, inconsistencias y categorías redundantes, además de llevar a cabo procesos de normalización y creación de nuevas variables a través de técnicas como el *feature engineering*. El objetivo de esta etapa era convertir los datos originales en un conjunto de información más estructurado y útil para la fase posterior: análisis exploratorio y predictivo.

Finalmente, la fase de *Load* consistió en la preparación del dataset final para utilizarlo en los modelos analíticos y herramientas de visualización. Una vez hechas las transformaciones pertinentes, los datos quedaron preparados para realizar el análisis exploratorio, la modelización predictiva y la visualización final.

La Figura 2 resume de forma esquemática el flujo de trabajo seguido en el trabajo, desde la obtención y el tratamiento de los datos mediante el proceso ETL hasta las fases posteriores de análisis exploratorio, modelización predictiva, interpretabilidad y visualización.

Figura 2: Flujo metodológico del trabajo (proceso ETL y fases de análisis)



Fuente: Elaboración propia con PowerPoint

Gracias a esta metodología, ha sido posible trabajar con un conjunto de datos reales y complejos, reproduciendo un flujo de trabajo similar al utilizado en proyectos profesionales de analítica de datos.

4.1. Extract

La primera fase del proceso ETL consistió en la extracción y revisión inicial del conjunto de datos utilizado para el análisis. Para esto, se usó el dataset público "Crash Reporting – Drivers Data", disponible en el portal oficial de datos abiertos del gobierno de Estados Unidos (Data.gov). Este conjunto de datos recoge información sobre los conductores involucrados en accidentes de tráfico que ocurren en carreteras locales y del condado de Montgomery, en el estado de Maryland (EE. UU.).

Según la descripción oficial del *dataset*, la información se recoge a través del sistema Automated Crash Reporting System (ACRS) de la Policía Estatal de Maryland, una plataforma donde distintos cuerpos policiales del condado registran digitalmente la información de los accidentes de tráfico. El conjunto de datos recoge información detallada sobre cada accidente y los conductores implicados, incluyendo variables relacionadas con el contexto del accidente, el estado de la vía, las condiciones meteorológicas, el comportamiento del conductor, las características del vehículo y la gravedad/severidad de este.

Sin embargo, la propia fuente advierte de que los datos son preliminares y que pueden presentar errores humanos, inconsistencias o registros que todavía no están verificados oficialmente. Esta observación fue particularmente importante en las primeras fases del trabajo, ya que los datos originales presentaban numerosos problemas de calidad y estructura que dificultaron su tratamiento de forma directa.

Inicialmente, se intentó cargar el archivo CSV directamente en Python con la librería pandas. Sin embargo, durante esta primera carga, empezaron a aparecer múltiples errores relacionados con la estructura interna del fichero. El *dataset* presentaba inconsistencias en el número de columnas, registros incompletos y líneas mal formateadas, lo que impedía la lectura correcta de los datos.

Además, en esta primera fase también surgieron problemas de compatibilidad entre diferentes versiones de librerías dentro del entorno de Anaconda, lo que impedía

importar algunas librerías básicas necesarias para el análisis. Tras revisar el origen de este problema, se concluyó que el entorno utilizado contenía dependencias antiguas que generaban conflictos entre versiones.

Para solucionar este problema, se decidió crear un entorno virtual independiente específico para el TFG utilizando Anaconda y Python 3.10. En este nuevo entorno se instalaron nuevamente todas las librerías necesarias para el análisis. Esto permitió trabajar en un entorno más estable y evitar conflictos de dependencias entre paquetes.

Una vez configurado correctamente el entorno de trabajo, se volvió a intentar cargar el *dataset*. Sin embargo, volvieron a aparecer los numerosos errores estructurales del fichero. En una primera revisión se eliminaron filas vacías y registros con un número insuficiente de columnas. Aun así, seguían existiendo inconsistencias relacionadas con registros desalineados y diferencias entre el número de columnas esperadas y las realmente encontradas por pandas.

Ante esta situación, se decidió inspeccionar manualmente el archivo CSV utilizando un editor de texto para identificar el origen de los errores. Durante esta revisión se detectó que muchos registros contenían dobles o incluso triples comillas consecutivas ("" o """), lo que generaba problemas de interpretación en la lectura del fichero. También se observó que numerosos registros comenzaban y terminaban con comillas innecesarias, además de existir varias variables mal escritas o con errores de formato.

Dado que la propia descripción oficial del *dataset* ya advertía de la existencia de posibles errores manuales y registros preliminares, se optó por aplicar una lectura más tolerante del fichero. Para ello, se utilizó el parámetro “on_bad_lines='skip'”, lo que permitía descartar automáticamente aquellas filas que seguían generando errores de lectura tras haber hecho una primera limpieza. Esta decisión permitió conservar la gran mayoría de los registros válidos sin comprometer la consistencia global de los datos.

Por último, tras realizar esta primera fase de limpieza estructural y corregir los principales errores de formato, el *dataset* pudo cargarse correctamente en Python. Con esto, el conjunto de datos quedó preparado para la siguiente fase del proceso ETL, que se centraría en la transformación, limpieza avanzada y creación de nuevas variables para el análisis exploratorio y predictivo posterior.

4.2. Transform

Una vez extraído y cargado el conjunto de datos, se procedió a la fase de transformación (*Transform*) del proceso ETL, una de las más importantes del trabajo debido a la naturaleza de los datos. El objetivo de esta etapa fue limpiar y adaptar la información original para hacerla adecuada tanto para el análisis exploratorio como para la posterior construcción de los modelos predictivos. Como ya se ha mencionado, al tratarse de un *dataset* real procedente de informes policiales, los datos presentaban numerosas inconsistencias, valores faltantes y categorías poco homogéneas, por lo que fue necesario llevar a cabo un proceso de limpieza exhaustivo.

En una primera revisión del *dataset* se analizaron los valores nulos presentes en cada variable. Se observó que algunas variables presentaban porcentajes extremadamente elevados de datos faltantes, superando en ciertos casos el 90% de valores NaN. Por ello, se decidió eliminar aquellas variables cuya información resultaba insuficiente o cuyo nivel de completitud no permitía obtener conclusiones fiables. No obstante, algunas variables con un alto porcentaje de valores nulos se mantuvieron inicialmente, al considerarse que podían aportar información relevante en fases posteriores del análisis, como por ejemplo la variable '*Circumstance'*.

Además de los valores faltantes, uno de los principales problemas detectados fue la falta de homogeneidad en las variables categóricas. Muchas categorías aparecían duplicadas debido a diferencias en mayúsculas, abreviaturas, errores ortográficos o ligeras variaciones en la escritura. Por este motivo, se realizó un proceso de normalización y agrupación de categorías, unificando etiquetas equivalentes y reduciendo la fragmentación de variables con demasiados valores únicos. Este paso permitió reducir ruido en los datos y mejorar posteriormente la interpretabilidad del análisis y de los modelos predictivos.

Durante esta fase también se eliminaron variables identificadoras que no aportaban valor al análisis. Variables como '*Report Number'*, '*Local Case Number*', '*Person ID*'o '*Vehicle ID'* actuaban solamente como identificadores únicos del accidente o del conductor, y presentaban una cardinalidad demasiado alta sin aportar información explicativa sobre la severidad del accidente.

Otro aspecto importante de la transformación fue la detección de posibles problemas de *data leakage*. Algunas variables contenían información demasiado ligada a la variable objetivo (salir ileso o no del accidente) y podían introducir sesgos en el modelo predictivo. Por ejemplo, la variable '*ACRS Report Type*' mostraba de forma indirecta la severidad del accidente, por lo que se decidió eliminarla también para evitar que el modelo basara sus predicciones en información que únicamente está disponible con posterioridad al accidente.

A continuación, se llevó a cabo una limpieza específica de las variables relevantes para el análisis. En variables relacionadas con el entorno del accidente, como '*Weather'*, '*Light*' o '*Surface Condition*', se agruparon categorías redundantes y se unificaron formatos. En otros casos, como el de Traffic Control, debido a la gran cantidad de categorías originales, se reorganizaron los valores en grupos más representativos, diferenciando en este caso entre semáforos, señales de stop, señales peatonales, controles policiales o ausencia de control vial.

También se transformaron varias variables relacionadas con el comportamiento del conductor. La variable '*Driver At Fault*', originalmente categórica, se codificó como una variable binaria para facilitar la interpretación. Por su parte, la variable '*Driver Distracted By'* también presentaba numerosas categorías muy específicas y poco homogéneas, por lo que se agruparon en categorías más amplias, como distracciones por uso de teléfono, distracciones mentales, factores externos o manipulación del vehículo.

Uno de los procesos de transformación más complejos estuvo relacionado con las variables de abuso de sustancias. Tanto '*Driver Substance Abuse*'como '*Non-Motorist Substance Abuse'* tenían valores muy desestructurados, múltiples etiquetas y combinaciones contradictorias. Para solucionar este problema se diseñaron funciones específicas que permitiesen identificar de forma correcta si existía consumo de drogas o alcohol, evitando falsos positivos derivados de la multietiqueta. Por último, estas variables se recodificaron en variables binarias que indicaban simplemente la presencia o ausencia de consumo de sustancias.

En relación con los peatones y usuarios vulnerables, se transformó la variable '*Related Non-Motorist'* en una variable binaria que mostraba si había o no algún peatón o ciclista implicado en el accidente. Inicialmente se planteó dividir esta variable en categorías más específicas, como peatón, ciclista u otros usuarios vulnerables. Sin

embargo, debido al elevado número de valores nulos y al desbalanceo entre las mismas categorías, se consideró más apropiado simplificar la variable para reducir ruido en el modelo.

También se llevaron a cabo transformaciones sobre las variables relacionadas con el vehículo. En la variable '*Vehicle Year'* se detectaron valores inconsistentes y claramente erróneos, como años futuros u outliers. Por ello, se estableció un rango de años válidos y se eliminaron aquellos registros que quedaban fuera del intervalo. Además, variables como '*Vehicle Make*' o '*Vehicle Model'* presentaban una cardinalidad muy elevada, por lo que se conservaron únicamente las marcas más frecuentes, con el fin de evitar problemas de sobreajuste.

Asimismo, variables relacionadas con el tipo de vehículo, el movimiento del mismo o el tipo de colisión se agruparon en categorías más generales y homogéneas. Esto permitió mantener la información relevante reduciendo al mismo tiempo la complejidad del *dataset*.

Una vez se completó la limpieza básica de variables, se llevó a cabo un proceso de *feature engineering* con el objetivo de generar nuevas variables útiles para el análisis. A partir de la variable '*Crash Date/Time'* se crearon variables temporales, como la hora del día, el día de la semana, o variables binarias que indicaban si el accidente se había producido en fin de semana o durante la noche. Estas nuevas variables permitieron analizar los patrones temporales de la accidentalidad.

También se crearon nuevas variables relacionadas con la localización y el vehículo. Por ejemplo, se identificaron calles con mayor número de accidentes y se creó una variable binaria que indicaba si el accidente ocurría en estas calles o no. Además, se creó la variable de '*Vehicle Age*', calculada a partir de la diferencia entre el año del accidente y el año de fabricación del vehículo.

Finalmente, se transformó la variable objetivo del análisis. Inicialmente, la variable '*Injury Severity'* contenía varias categorías relacionadas con distintos niveles de gravedad de las lesiones. Sin embargo, el fuerte desbalanceo existente entre accidentes graves y no graves hacía complicado el modelado. Por esto, se optó por construir una nueva variable binaria como variable objetivo, que diferenciara entre accidentes en los que el conductor resultaba ileso y accidentes con algún tipo de lesión. Este cambio

permitió simplificar el problema y adaptar mejor el análisis a un modelo de clasificación binaria.

En conclusión, esta fase de transformación permitió convertir un *dataset* original altamente inconsistente y desestructurado en un conjunto de datos limpio, coherente y preparado para el análisis exploratorio, la modelización predictiva y la visualización posterior de resultados.

4.3. Load

La última fase del proceso ETL, *Load*, consiste en dejar los datos ya tratados disponibles para su uso en las fases de análisis. A diferencia de un proyecto empresarial, en el que esta etapa suele implicar la carga de los datos en una base de datos o un *data warehouse*, en este trabajo la fase de *Load* se entiende como la preparación del conjunto de datos final, una vez completadas la limpieza, la transformación y la creación de nuevas variables mediante *feature engineering.*

En este punto, el conjunto de datos resultante, estructurado a nivel de cada conductor implicado en un accidente, ya se encuentra limpio y enriquecido con las variables necesarias, quedando listo para su tratamiento y uso en los modelos. Este *dataset* es el que alimenta las tres fases posteriores del trabajo: el análisis exploratorio, la construcción y evaluación de los modelos predictivos y la elaboración del panel de visualización.

En cuanto a las herramientas empleadas, se llevaron a cabo el análisis exploratorio y la modelización predictiva en Python, utilizando principalmente las librerías pandas, scikit-learn e imbalanced-learn. Para la parte de visualización, una vez preparado el conjunto de datos final, se exportó desde Python a un archivo Excel, que posteriormente se cargó en Tableau para elaborar el panel con los distintos gráficos y mapas. De este modo, la fase de *Load* actúa como puente entre el tratamiento de los datos y su análisis posterior, garantizando que toda la información utilizada parte de una base común, limpia y coherente.

5. Desarrollo

En este apartado se va a hablar de los principales resultados obtenidos a lo largo del TFG, desde la descripción y análisis inicial del conjunto de datos hasta la construcción e interpretación de modelos predictivos. El objetivo principal del desarrollo es identificar qué factores están más relacionados con la probabilidad de no salir ileso en un accidente de tráfico, es decir, sufrir algún tipo de lesión, y analizar si es posible predecir este tipo de situaciones a partir de información disponible en el *dataset*.

En primer lugar, se describe el conjunto de datos final utilizado en el análisis, así como las variables seleccionadas tras el proceso de limpieza y transformación llevados a cabo durante el proceso de ETL. Dada la naturaleza de los datos, gran parte de las variables son categóricas, algo habitual en datos procedentes de informes policiales y accidentes de tráfico reales. Por esto, una parte muy importante del trabajo ha consistido en tratar los datos no estructurados, corregir inconsistencias y transformar la información en variables más útiles y homogéneas para el análisis.

Después, se realiza un análisis exploratorio de los datos (*Exploratory Data Analysis* o EDA), dividido en análisis univariante y bivariante. Esta fase permite comprender mejor la distribución de las variables y detectar posibles patrones relacionados con los accidentes con lesión. A través de gráficos y visualizaciones se estudian factores relacionados con el entorno, el comportamiento del conductor, el tipo de vehículo o las características del accidente.

A continuación, se desarrollan distintos modelos predictivos con el objetivo de clasificar si un conductor sale ileso o no en un accidente. Para ello, se utilizan técnicas de preparación de datos, codificación de variables categóricas y tratamiento del desbalanceo existente en la variable objetivo. Finalmente, se interpretan los resultados obtenidos a través de herramientas de explicabilidad.

5.1. Descripción del dataset y variables empleadas

Después de hacer el proceso de extracción, limpieza y transformación de los datos, el conjunto final utilizado para el análisis quedó compuesto por 203.890 registros y 63 variables (muchas de ellas binarias). Cada registro representa un conductor implicado en un accidente de tráfico ocurrido en el condado de Montgomery (Maryland, EE. UU.), por

lo que un mismo accidente puede aparecer varias veces en el *dataset* si existen varios conductores involucrados.

Como se ha mencionado anteriormente, el conjunto de datos tiene principalmente variables categóricas relacionadas con las características del accidente, el entorno, el conductor y el vehículo. Este tipo de estructura es común en *datasets* reales de tráfico, ya que gran parte de la información se recoge a través de informes policiales y descripciones cualitativas del accidente. Por esto, una parte muy importante del TFG ha consistido en tratar datos no estructurados, corregir inconsistencias y transformar la información en variables más homogéneas y útiles para el análisis.

Aunque el *dataset* contiene una alta proporción de variables categóricas, esto no supone una limitación para el análisis. Al contrario, trabajar con este tipo de datos permite acercarse a situaciones reales de analítica de datos, donde gran parte del valor del proyecto reside principalmente en la limpieza, tratamiento y transformación de los datos antes de construir modelos predictivos. Además, mediante técnicas como *one-hot encoding* y *feature engineering,* es posible convertir estas variables en información cuantificable y útil para algoritmos de *machine learning*.

Las variables utilizadas pueden agruparse en distintos bloques según la información que aportan. En primer lugar, se incluyen las variables relacionadas con el entorno y las condiciones de la vía, como '*Weather'*, '*Surface Condition'*, '*Light'*, '*Traffic Control'* o '*Route Type'*. Estas variables permiten analizar cómo factores externos como el tiempo, la iluminación o el tipo de control vial pueden influir en la probabilidad de sufrir lesiones en un accidente.

En segundo lugar, se usaron variables relacionadas con el conductor y su comportamiento durante el accidente. Entre ellas destacan '*Driver At Fault'*, '*Driver Distracted By'*, '*Driver Substance Abuse'* o las variables derivadas de '*Circumstance'*. Estas últimas fueron transformadas en múltiples variables binarias que permiten identificar comportamientos específicos, como exceso de velocidad, conducción distraída, invasión de carril, pérdida de control del vehículo o saltarse señales de tráfico.

También se incluyeron variables relacionadas con las características del vehículo y la dinámica del accidente, como '*Vehicle Body Type'*, '*Vehicle Movement'*, '*Collision Type'*, '*Vehicle Going Dir'*, '*Vehicle Make'* o '*Speed Limit'*. Estas variables permiten

estudiar cómo el tipo de vehículo, el movimiento previo al impacto o la naturaleza de la colisión afectan a la severidad del accidente.

Asimismo, durante la fase de *feature engineering,* se crearon nuevas variables temporales a partir de la fecha y hora del accidente. Entre ellas se encuentran variables como la hora del día (*hour*), el día de la semana (*day_of_week*), el mes (*month*), una variable binaria de fin de semana (*is_weekend*) y otra que identifica accidentes ocurridos en horario nocturno (*is_night*). Estas variables permiten detectar posibles patrones temporales asociados a la accidentalidad.

Por otro lado, también se generaron variables relacionadas con la localización y el contexto del accidente. Se crearon variables binarias que identifican si el accidente ocurrió en una de las calles con mayor número de accidentes (*Road_Top* y *Top_Cross_Street*), así como variables relacionadas con la presencia de peatones u otro tipo de usuarios vulnerables implicados (*NonMotorist_Involved*).

La variable objetivo utilizada, sobre la que se hablará en más profundidad en el apartado 5.2, fue *Injured,* una variable creada a partir de la variable original *Injury Severity.* En esta nueva variable, el valor 0 representa accidentes en los que el conductor resultó ileso, mientras que el valor 1 agrupa a aquellos casos en los que existió algún tipo de lesión. Esta transformación permite simplificar el problema y adaptarlo a un modelo de clasificación binaria más interpretable.

Finalmente, cabe destacar que el *dataset* presenta un desbalanceo importante en la variable *Injured*, ya que aproximadamente un 92% de los registros corresponden a conductores ilesos frente a un 8% de accidentes con lesión. Esto se tuvo en cuenta durante la fase posterior de modelización predictiva, aplicando técnicas específicas para reducir el sesgo hacia la clase mayoritaria.

5.2. Resultados del análisis exploratorio

Una vez preparado el conjunto de datos final, se llevó a cabo un análisis exploratorio de los datos (*Exploratory Data Analysis* o EDA) con el objetivo de entender mejor la distribución de las variables y detectar posibles patrones relacionados con la probabilidad de sufrir lesiones en un accidente. Esta fase es importante especialmente

antes de construir modelos predictivos, ya que permite identificar tendencias, relaciones entre variables y posibles factores de riesgo.

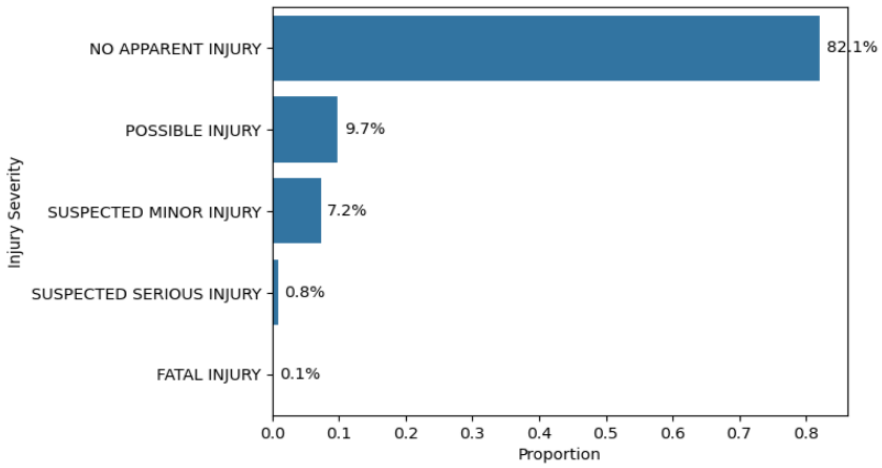
El análisis exploratorio se dividió en dos partes principales. En primer lugar, se llevó a cabo un análisis univariante centrado en estudiar de forma individual las variables más relevantes y la distribución general del *dataset*. Después, se realizó un análisis bivariante orientado a analizar la relación entre distintas variables explicativas y la variable objetivo (*Injured)*. Por último, se incorporaron algunas visualizaciones complementarias para facilitar la interpretación de determinados patrones en los datos.

Análisis univariante

El primer paso del análisis exploratorio fue estudiar la distribución de la variable objetivo (*Injured)*, que diferencia entre conductores ilesos y conductores que sufrieron algún tipo de lesión durante el accidente. Para ello, se usó la variable '*Injury Severity'*, que clasificaba los accidentes según el nivel de lesión sufrido por el conductor. La distribución original de esta variable era la siguiente:

- *NO APPARENT INJURY*: 165.678 registros
- *POSSIBLE INJURY*: 19.635 registros
- *SUSPECTED MINOR INJURY*: 14.604 registros
- *SUSPECTED SERIOUS INJURY*: 1.672 registros
- *FATAL INJURY*: 186 registros
- Valores nulos: 2.115 registros

Figura 3: Distribución inicial de la variable Injury Severity



Fuente: Elaboración propia con Python

Como puede observarse en la Figura 3, la gran mayoría de los registros corresponden a accidentes sin lesiones aparentes, mientras que las categorías relacionadas con accidentes graves o fatales representan únicamente una pequeña parte del *dataset*.

En un primer instante, se planteó construir la variable objetivo diferenciando únicamente los accidentes graves, agrupando las categorías *FATAL INJURY* y *SUSPECTED SERIOUS INJURY*. Sin embargo, esta opción generaba un problema importante de desbalanceo, ya que los accidentes graves representaban aproximadamente solo un 1% del total de registros del *dataset*.

Debido a este fuerte desbalanceo, se decidió reformular la variable objetivo y construir una clasificación algo más equilibrada entre conductores ilesos y conductores con lesión. El objetivo era tener una variable más adecuada para el desarrollo de modelos predictivos, evitando que el modelo quedase demasiado condicionado por una clase minoritaria extremadamente pequeña.

Por esto, se creó una nueva variable binaria denominada '*Injured'*. Para construirla, se agruparon las categorías *FATAL INJURY*, *SUSPECTED SERIOUS INJURY* y *SUSPECTED MINOR INJURY* dentro de la categoría "1", correspondiente a accidentes con lesión. Por otro lado, las categorías *NO APPARENT INJURY* y *POSSIBLE INJURY* se incluyeron en la categoría "0", que corresponde a conductores ilesos o sin lesiones aparentes.

Tras esta transformación, la nueva variable objetivo presentó una distribución considerablemente más equilibrada que la que se planteó inicialmente para accidentes graves. Aproximadamente un 92% de los registros correspondían a conductores ilesos, mientras que cerca del 8% eran accidentes con alguna lesión.

Aunque el *dataset* seguía presentando un desbalanceo importante entre clases, esta nueva distribución resultaba más adecuada para el análisis predictivo y permitía trabajar con una cantidad significativamente mayor de casos positivos dentro del modelo. Además, este tipo de distribución suele resultar habitual en *datasets* reales de accidentes de tráfico, donde la mayoría suelen corresponder a accidentes leves o sin consecuencias graves.

Por otra parte, también se analizaron de forma individual otras variables además de '*Injured'*. En términos generales, se observó que gran parte de los accidentes

ocurrieron en condiciones meteorológicas normales, con buena iluminación y en vías urbanas. Variables como '*Weather'*, '*Light*'o '*Surface Condition'* mostraban una fuerte concentración en unas pocas categorías dominantes, mientras que otras categorías aparecían únicamente en un número reducido de registros.

En cuanto a las características del accidente, algunos tipos de colisión y movimientos del vehículo aparecían con mucha mayor frecuencia que otros. Por ejemplo, las colisiones por alcance y los movimientos a velocidad constante representaban una parte importante de los accidentes registrados. También se observó que ciertos tipos de vehículos, como turismos y SUV, predominaban claramente dentro del *dataset*, algo coherente con el tipo de vehículos habituales en entornos urbanos.

Asimismo, el análisis univariante permitió detectar variables muy fragmentadas o con categorías poco representadas, lo que confirmó la necesidad de hacer agrupaciones y transformaciones durante la fase ETL para reducir ruido y mejorar la interpretabilidad del análisis.

Análisis Bivariante

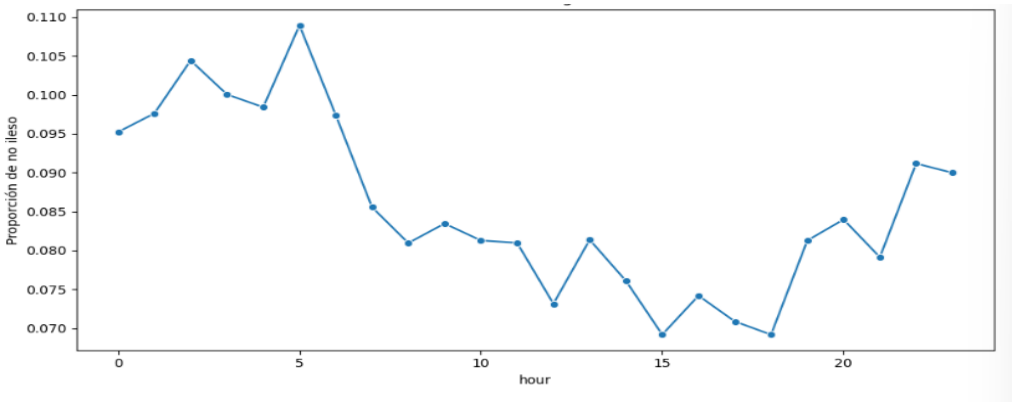
Una vez analizada la distribución individual de las variables, se estudió la relación entre distintas características del accidente y la probabilidad de que el conductor resultase lesionado.

Para facilitar la interpretación de los resultados, las variables se han agrupado en diferentes bloques temáticos relacionados con las variables temporales, condiciones del entorno y vía, el comportamiento del conductor y las características del vehículo y la dinámica.

En relación con las condiciones del entorno y la vía, se analizaron variables como '*Weather'*, '*Light'*, '*Surface Condition'* y '*Traffic Control'.* En términos generales, estas variables no presentaron diferencias significativas en la proporción de lesionados entre sus distintas categorías. Aunque algunas condiciones adversas, como la nieve, el hielo o la lluvia, aparecían con frecuencia en los accidentes, las tasas de lesión se mantenían bastante similares en la mayoría de los grupos. Esto sugiere que, aunque las condiciones meteorológicas pueden influir en la ocurrencia del accidente, no siempre parecen determinar directamente la gravedad de las lesiones.

También se analizaron varios patrones temporales a partir de las variables creadas durante la fase de *feature engineering.* Variables como la hora del día, el día de la semana o la conducción nocturna permitieron detectar ciertas tendencias relacionadas con la probabilidad de lesión. En particular, se observó que los accidentes ocurridos durante la noche presentaban una proporción algo superior de conductores lesionados respecto a los accidentes diurnos. De igual forma, como podemos observar en la Figura 4, algunos tramos horarios concentraban mayor frecuencia de accidentes con lesión, especialmente periodos asociados a menor visibilidad o mayor intensidad de tráfico. Por otro lado, no se observó gran diferencia entre las categorías en variables como el día de la semana o el mes del año.

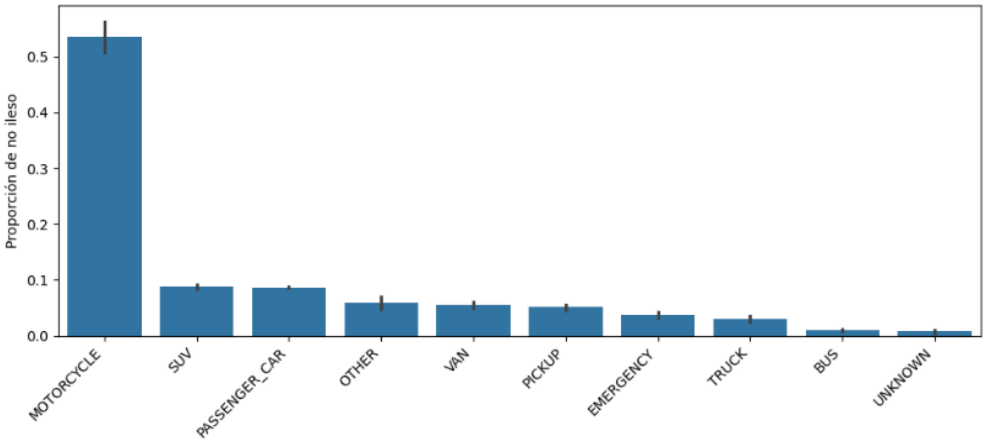
Figura 4: Proporción de accidentes con lesión según la hora del día



Fuente: Elaboración propia con Python

Por otro lado, sí se observaron patrones más relevantes en variables relacionadas con la dinámica del accidente y el tipo de vehículo implicado. El análisis mostró que determinados tipos de vehículos presentan una probabilidad considerablemente mayor de que el conductor resulte lesionado. Destaca especialmente el caso de la motocicleta, donde aproximadamente la mitad de los conductores implicados en accidentes sufrieron algún tipo de lesión. Este resultado resulta coherente con la menor protección estructural que ofrece este tipo de vehículo frente a impactos.

Figura 5: Proporción de accidentes con lesión según el tipo de vehículo



Fuente: Elaboración propia con Python

También se observaron diferencias importantes según el movimiento del vehículo antes del accidente. Situaciones relacionadas con pérdida de control, deslizamiento del vehículo (*skidding)* o determinadas maniobras, como giros a la izquierda, aparecían relacionadas con una mayor proporción de accidentes con lesión. Del mismo modo, algunos tipos de colisiones tenían patrones claramente diferenciados. Las colisiones frontales (*head-on collisions*) tenían una proporción mucho más elevada de conductores lesionados en comparación con colisiones laterales o accidentes por alcance.

En cuanto al comportamiento del conductor, variables relacionadas con distracciones, consumo de sustancias o determinadas circunstancias del accidente también mostraron asociaciones interesantes con la variable objetivo. En particular, situaciones relacionadas con conducción distraída, pérdida de control del vehículo o incumplimiento de señales de tráfico aparecían con más frecuencia en accidentes con lesión.

Además, las variables derivadas de '*Circumstance'* permitieron detectar ciertos patrones recurrentes relacionados con errores humanos y condiciones adversas de conducción. Aunque muchas de estas variables presentaban originalmente una estructura compleja y multietiqueta, su transformación en variables binarias permitió analizarlas de una forma mucho más interpretable.

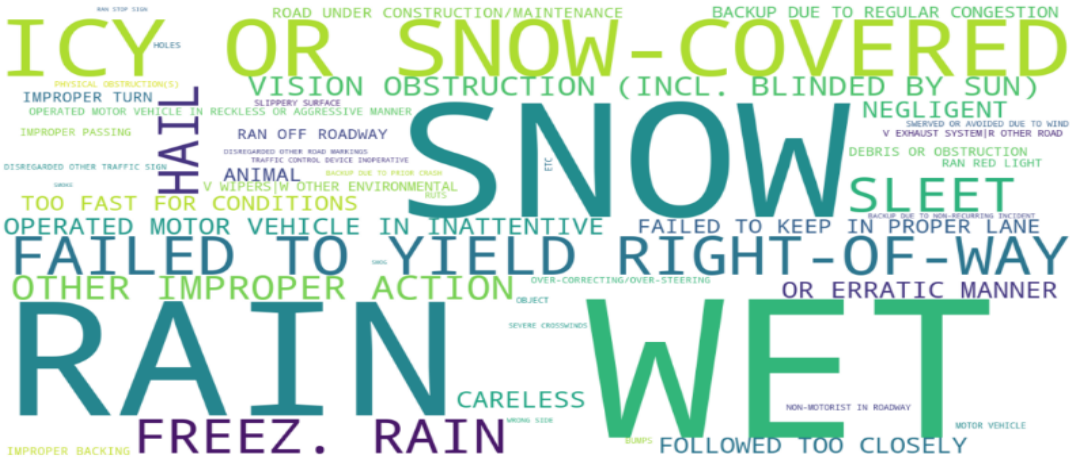
En conjunto, el análisis bivariante permitió identificar qué variables parecían tener una mayor relación con la probabilidad de lesión y ayudó a detectar patrones relevantes que posteriormente serían contrastados mediante modelos predictivos.

Patrones y visualizaciones destacadas

Con el objetivo de complementar el análisis exploratorio, se realizaron visualizaciones adicionales sobre algunas de las variables más complejas y menos estructuradas del *dataset*.

Una de las variables más interesantes fue '*Circumstance',* ya que combinaba información relacionada con el comportamiento del conductor, las condiciones de la vía, factores meteorológicos y otras circunstancias específicas del accidente. Debido a su alta heterogeneidad y presencia de múltiples etiquetas dentro de un mismo registro, además de transformarla en variables binarias durante el proceso de ETL, se decidió realizar una nube de palabras utilizando los términos más frecuentes presentes en los informes policiales.

Figura 6: Circunstancias más frecuentes de los accidentes (nube de palabras de la variable '*Circumstance*')



Fuente: Elaboración propia con Python

Como puede observarse en la Figura 6, destacan especialmente términos relacionados con condiciones meteorológicas adversas como "snow", "wet", "ice" o "freez rain", así como expresiones relacionadas con errores humanos y comportamiento

de riesgo, como “failed to yield right-of-way”, “inattentive”, “followed too closely” o “careless”.

La presencia de estos términos permite identificar patrones generales dentro del *dataset* y aporta una visión más cualitativa sobre el contexto en el que ocurren muchos accidentes. Aunque este tipo de visualización no permite establecer relaciones causales ni cuantificar directamente el impacto de cada circunstancia sobre la gravedad del accidente, sí es útil para complementar el análisis y detectar tendencias de forma visual e intuitiva.

En conjunto, el análisis exploratorio permitió identificar múltiples factores potencialmente relacionados con la probabilidad de lesión, así como comprender mejor la estructura y complejidad del *dataset* antes de empezar a desarrollar los modelos.

5.3. Resultados del modelo predictivo

Tras las fases de limpieza, transformación y análisis exploratorio, esta sección trata el objetivo central del trabajo: la construcción de un modelo capaz de predecir la gravedad de los accidentes registrados en el condado de Montgomery. Más concretamente, se plantea un problema de clasificación binaria que tiene el propósito de estimar si una persona implicada en un accidente resultará ilesa o, por el contrario, sufrirá algún tipo de lesión.

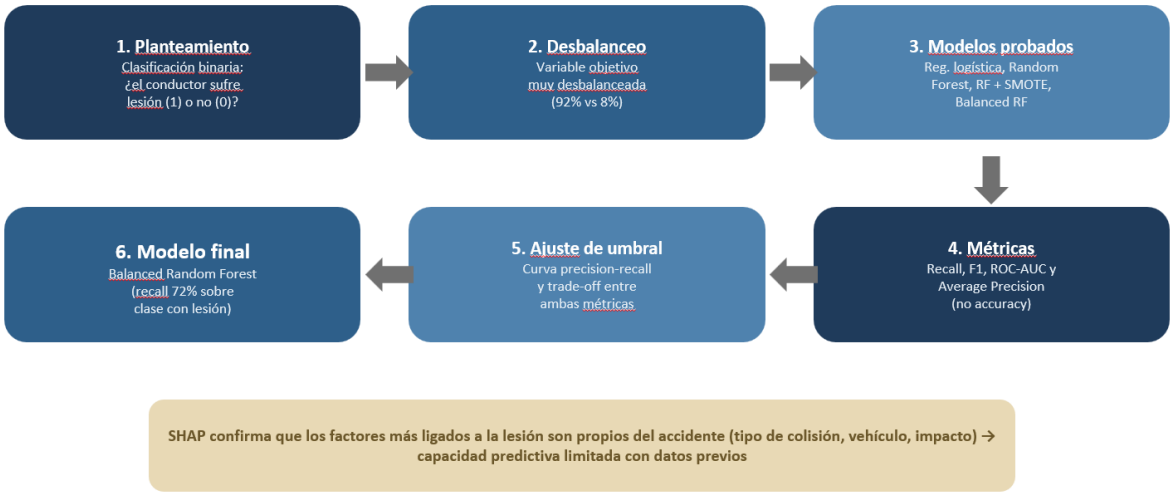
Como ya se comentaba en el estado del arte, en este tipo de problemas no solo interesa predecir, sino también entender qué factores se asocian a estas predicciones. Por esto, además del modelo predictivo principal, que utiliza únicamente información disponible antes o en el momento del accidente, también se desarrolla de forma complementaria un enfoque explicativo que permite estudiar qué variables influyen más en la gravedad del accidente. La idea es poder comparar ambas perspectivas y ver hasta qué punto las condiciones previas al accidente bastan para anticipar lesiones.

En los siguientes apartados se explica primero cómo se ha planteado el problema y cómo se ha definido la variable objetivo. Después se trata el desbalanceo de clases, que es uno de los principales retos del *dataset*, y se presentan los modelos que se han probado junto con las métricas utilizadas para evaluarlos. A continuación, se recogen las distintas técnicas que se han aplicado para intentar mejorar los resultados y se explica por qué

finalmente se ha optado por el Balanced Random Forest. Por último, se incluye un análisis de interpretabilidad con la técnica SHAP y una valoración crítica de la capacidad predictiva alcanzada, así como las principales limitaciones que presenta el modelo, que se retomarán con más detalle en el apartado 6.2.

La Figura 7 sintetiza las etapas seguidas en la construcción del modelo predictivo, desde el planteamiento del problema y el tratamiento del desbalanceo hasta la selección del modelo final y su posterior interpretación. Cada una de estas fases se desarrolla en detalle en los apartados siguientes

Figura 7: Proceso de construcción del modelo predictivo



Fuente: Elaboración propia con PowerPoint

Planteamiento del problema predictivo y variable objetivo

El problema se ha planteado como una tarea de clasificación binaria, en la que el modelo debe estimar si una persona implicada en un accidente sufrirá o no una lesión. La variable objetivo, llamada '*Injured*', se construyó a partir de la variable original '*Injury Severity*', que recoge la gravedad de las lesiones siguiendo la escala estándar KABCO, muy utilizada en los informes de tráfico de Estados Unidos. Esta escala distingue cinco niveles: lesión mortal (K), lesión grave (A), lesión leve (B), lesión posible (C) y sin lesión aparente (O) (Maryland Highway Safety Office, 2024; NHTSA, 2024).

A partir de estos niveles se definió una variable binaria que agrupa las lesiones confirmadas (leve, grave y mortal) en la clase positiva (1), mientras que las categorías de lesión posible y sin lesión aparente se incluyen en la clase negativa (0). Esta decisión se justifica por la naturaleza de la categoría "lesión posible" , que no refleja una lesión confirmada, sino un estado algo ambiguo que se asume cuando no hay evidencia física visible y que depende de la apreciación y opinión del agente o persona implicada. Por esto, agruparla con la categoría "sin lesión aparente" resulta más coherente, de forma que la variable objetivo acabe representando la presencia o ausencia de una lesión evidente.

Una vez definida la variable objetivo, se establecieron dos conjuntos de variables explicativas con finalidades diferentes. El primero, con carácter predictivo, incluye únicamente aquellas variables que se conocen antes o en el mismo momento del accidente, como el tipo de vía, las condiciones meteorológicas, iluminación, estado de la superficie, límite de velocidad, tipo de vehículo o diferentes variables temporales. De este conjunto se excluyeron las variables que solo pueden conocerse una vez se había investigado el accidente, como el tipo de colisión, el movimiento del vehículo, si el conductor era responsable o las circunstancias del accidente, ya que su inclusión podría suponer un *data leakage* y restaría validez al modelo como herramienta de predicción real.

El segundo conjunto, de carácter explicativo, sí incluye estas variables, con el objetivo de analizar en qué medida las circunstancias propias del accidente impactan en la gravedad de las lesiones, y poder también comparar ambos enfoques.

Es importante señalar que la variable objetivo presenta un marcado desbalanceo, ya que aproximadamente el 92% de los registros corresponden a la clase sin lesión evidente y solo el 8% aproximadamente a la clase con lesión. Esta característica, que suele ser habitual en los conjuntos de datos de siniestralidad, condiciona de forma importante el proceso de modelización y se aborda con detalle en el siguiente apartado.

<u>El problema del desbalanceo de clases</u>

Como ya se ha comentado a lo largo del trabajo, la variable objetivo presenta un fuerte desbalanceo: alrededor del 92% de los registros corresponden a accidentes sin lesión evidente (clase 0) y solo en torno al 8% a accidentes con lesión (clase 1). Este desequilibrio es un problema frecuente en los conjuntos de datos de siniestralidad, ya que

la mayoría de los accidentes no provocan lesiones confirmadas, pero condiciona de forma importante la construcción y la evaluación de los modelos.

El principal reto es que los algoritmos de clasificación tienden a favorecer a la clase mayoritaria, ya que así suelen conseguir un mayor número de aciertos globales. En la práctica, esto significa que un modelo puede conseguir una *accuracy* muy alta limitándose a predecir que casi ningún accidente acaba en lesión, sin que aporte un valor real. De hecho, un modelo que dijera siempre que no hay lesión acertaría aproximadamente el 92% de las veces, pero sería inútil, porque lo que realmente interesa es detectar los accidentes en los que sí hay lesión.

Por este motivo, el desbalanceo obliga a tomar dos decisiones importantes a lo largo del análisis. Por un lado, a utilizar métricas de evaluación que no estén distorsionadas por la clase mayoritaria y que reflejen el comportamiento del modelo sobre la clase minoritaria, que es la interesante. Por otro, a aplicar técnicas específicas para tratar este desequilibrio durante el entrenamiento. Ambos aspectos se desarrollan en los siguientes apartados.

Modelos probados y métricas de evaluación

Para tratar el problema se probaron varios modelos de clasificación, empezando por los más sencillos e interpretables y siguiendo con modelos que gestionan mejor el desbalanceo.

Inicialmente se entrenó una regresión logística, por tratarse de un modelo simple, rápido y muy interpretable, que sirve como referencia inicial (*baseline*) para tener una idea del rendimiento mínimo que se puede esperar. A continuación, se entrenó un Random Forest, un modelo más complejo que, a diferencia de la regresión logística, puede captar relaciones no lineales e interacciones entre variables, además de tratar bien los datos categóricos. La razón de probar ambos fue poder comparar un modelo lineal sencillo con otro más potente y comprobar si el aumento de complejidad se traducía realmente en una mejora de los resultados. Como el Random Forest ofreció un rendimiento ligeramente superior y se adapta mejor a la naturaleza de los datos, se decidió continuar el análisis con este modelo.

A partir de aquí, y siguiendo la estructura mencionada, el Random Forest se entrenó sobre los dos conjuntos de variables definidos: uno predictivo, con la información

disponible antes del accidente, y otro explicativo, que incluye también las variables propias del accidente. Esto permitió comparar los dos enfoques sobre un mismo algoritmo.

Ya desde estos primeros modelos se observó que el rendimiento predictivo era bastante limitado, especialmente en cuanto a la precisión sobre la clase con lesión. Como los resultados seguían estando muy condicionados por el desbalanceo, se decidió probar dos técnicas pensadas específicamente para este tipo de problema, tratando de comprobar si realmente conseguían mejorar el rendimiento: un Random Forest combinado con la técnica de sobremuestreo SMOTE, y un Balanced Random Forest. Como se verá, ninguna de las dos logró una mejora realmente significativa de los resultados, lo cual es en sí mismo un hallazgo importante sobre la dificultad del problema. Ambas técnicas se explican con más detalle en el siguiente apartado.

En lo que respecta a la evaluación de los modelos, la elección de las métricas es un aspecto importante en este TFG, ya que está directamente condicionada por el desbalanceo de la variable objetivo. La métrica más intuitiva, la *accuracy*, no es útil en este caso, ya que al haber un 92% de accidentes sin lesión, un modelo podría alcanzar valores muy altos limitándose a predecir la clase mayoritaria, sin aportar nada al objetivo del trabajo. Por este motivo, el análisis no se centra en la *accuracy*, sino en las métricas que reflejan el comportamiento del modelo sobre la clase minoritaria, que es la que realmente interesa: los accidentes con lesión.

En concreto, se ha prestado especial atención al *recall*, ya que mide la capacidad del modelo para identificar correctamente los accidentes con lesión, que es precisamente lo que se quiere detectar; un *recall* bajo significaría que se están dejando escapar casos relevantes. Sin embargo, fijarse solo en el *recall* tampoco es suficiente, porque un modelo puede detectar muchas lesiones a costa de generar muchas falsas alarmas, lo que se reflejaría en una precisión baja. Por ello se ha tenido en cuenta también el F1-score, que combina ambas en un único valor y permite valorar el equilibrio entre detectar lesiones y no equivocarse en exceso. Finalmente, se han incluido el ROC-AUC y, sobre todo, la Average Precision (AP), esta última especialmente recomendable en problemas desbalanceados por resumir el comportamiento del modelo a lo largo de los distintos umbrales de decisión. En conjunto, estas métricas ofrecen una visión mucho más realista del rendimiento del modelo que la simple *accuracy*.

A modo de resumen, la Tabla 3 recoge los resultados de los modelos sobre la clase de interés (accidentes con lesión), utilizando el umbral por defecto de 0,5.

Tabla 3: Comparativa de los modelos probados (clase "con lesión")

Modelo	Precision	Recall	F1-score	ROC-AUC	Avg. Precision
Regresión logística (baseline)	0,130	0,711	0,219	0,711	-
RF Predictivo (baseline)	0,115	0,681	0,197	0,669	0,168
RF Explicativo (referencia)	0,157	0,763	0,260	0,776	0,234
RF + SMOTE (predictivo)	0,593	0,013	0,026	0,634	0,132
Balanced RF (predictivo)	0,114	0,722	0,197	0,676	0,167

Fuente: Elaboración propia

Nota: resultados obtenidos con el umbral por defecto (0,5). La variable objetivo presenta un desbalanceo del 92% (sin lesión) frente al 8% (con lesión). En el caso de la regresión logística no se calculó la Average Precision.

Al observar la tabla se aprecian varias cosas. En primer lugar, los resultados son en general modestos: todos los modelos presentan una precisión baja y un *recall* relativamente alto, lo que indica que detectan buena parte de los accidentes con lesión, pero a costa de generar bastantes falsas alarmas. El modelo explicativo tiene mejores resultados que los predictivos en todas las métricas, lo cual era esperable, ya que incorpora variables que describen el propio accidente. En cuanto a las técnicas de balanceo, el Random Forest con SMOTE muestra un comportamiento incluso contraproducente, con una precisión alta pero un *recall* casi nulo, es decir, apenas predice la clase con lesión, por lo que en la práctica no resulta útil. El Balanced Random Forest,

en cambio, sí consigue una mejora, aunque ligera, respecto al baseline, ofreciendo el mejor equilibrio entre las métricas dentro del enfoque predictivo, con un *recall* del 72% y valores algo superiores de ROC-AUC y F1-score. Por este motivo se seleccionó como modelo final y se trabajó sobre él en los apartados siguientes, donde se explica con más detalle. El hecho de que ninguna de las técnicas probadas haya conseguido una mejora significativa constituye en sí mismo una conclusión relevante, que se analiza con más detalle en la síntesis final del apartado.

Técnicas aplicadas para mejorar el modelo

Dado que los primeros modelos ofrecieron un rendimiento limitado, se aplicaron varias técnicas con el objetivo de tratar de mejorarlo, centradas sobre todo en el problema del desbalanceo y en la revisión de las variables. A continuación, se describen las principales.

Tratamiento del desbalanceo: En primer lugar, en los modelos basados en Random Forest se utilizó el parámetro class_weight='balanced', que asigna un mayor peso a la clase minoritaria durante el entrenamiento, y así el modelo "penaliza" más los errores sobre los accidentes con lesión. Además, más tarde se probaron dos técnicas específicas para este problema: el sobremuestreo mediante SMOTE, que crea ejemplos sintéticos de la clase minoritaria para equilibrar las clases, y el Balanced Random Forest, que equilibra las clases de forma distinta: en lugar de inventar datos nuevos, utiliza para entrenar el mismo número de accidentes con lesión y sin lesión, descartando parte de los casos sin lesión para que el modelo no se incline siempre hacia la clase mayoritaria. Como ya se ha comentado, SMOTE no mejoró los resultados (de hecho, los empeoró, reduciendo el *recall* casi a cero), mientras que el Balanced Random Forest sí ofreció una ligera mejora, por lo que se mantuvo como modelo final.

Revisión de variables: También se llevó a cabo una revisión de las variables utilizadas. Por un lado, se analizaron las categorías poco frecuentes de algunas variables, valorando la posibilidad de agruparlas, aunque finalmente se decidió mantener la mayoría por tener sentido dentro del problema y porque ya se habían agrupado en el tratamiento y limpieza de datos (por ejemplo, la categoría motocicleta, asociada a un mayor riesgo de lesión). Por otro lado, y más importante, se detectó que algunas variables incluidas inicialmente en el modelo predictivo, como las circunstancias del accidente, la distracción del conductor o el consumo de sustancias, en realidad solo se conocen tras la investigación

del siniestro, por lo que constituían una fuga de información. Estas variables se eliminaron del modelo predictivo para garantizar su validez. Curiosamente, su eliminación apenas modificó los resultados, lo que confirma que no estaban aportando información predictiva relevante.

Ajuste del umbral de decisión: Por último, se probó a ajustar el umbral de decisión del modelo. Por defecto, los modelos de clasificación asignan la clase positiva (1) cuando la probabilidad estimada supera el 0,5. Sin embargo, en problemas desbalanceados este umbral no tiene por qué ser el más adecuado. Por ello, se calculó la curva *precision-recall* y se buscó el umbral que maximizaba el F1-score, con la idea de conseguir un mejor equilibrio entre detectar lesiones (*recall*) y no generar demasiadas falsas alarmas (precisión). El umbral óptimo resultó ser de 0,556. La Tabla 4 muestra la comparación de los resultados del modelo final con el umbral por defecto y con el umbral ajustado.

Tabla 4: Resultados del modelo final (Balanced Random Forest) según el umbral de decisión

Métrica (clase "con lesión")	Umbral 0,5 (por defecto)	Umbral 0,556 (ajustado)
Precision	0,114	0,154
Recall	0,717	0,343
F1-score	0,196	0,213

Fuente: Elaboración propia

Como se observa en la tabla, el ajuste del umbral mejora la precisión (de 0,114 a 0,154) y el F1-score (de 0,196 a 0,213), pero a costa de reducir el *recall* (de 0,717 a 0,343). Es decir, el modelo se vuelve más exigente al clasificar un accidente como "con lesión", lo que reduce las falsas alarmas, pero hace que se escapen más casos reales. Esta es una muestra clara del compromiso (trade-off) que existe entre *precision* y *recall*, ya que mejorar una implica casi siempre empeorar la otra. En cualquier caso, la mejora global que se obtiene con el ajuste del umbral es modesta, lo que vuelve a apuntar a que la limitación principal del modelo no está en su configuración, sino en la información disponible. Por este motivo, y dado que en el contexto de la seguridad vial interesa más

no dejar escapar accidentes con lesión, se ha optado por mantener el umbral por defecto en el modelo final, ya que prioriza el *recall* y, por tanto, la detección de los casos con lesión.

Modelo final: Balanced Random Forest

Tras comparar todos los modelos y técnicas previamente explicados, se seleccionó como modelo final el Balanced Random Forest entrenado con conjunto de variables predictivas y con el umbral de decisión por defecto. Esto se justifica porque, dentro del enfoque predictivo, fue el modelo que ofreció el mejor equilibrio entre las métricas de interés, gestionando el desbalanceo sin necesidad de generar datos sintéticos y consiguiendo un *recall* razonable sobre la clase con lesión.

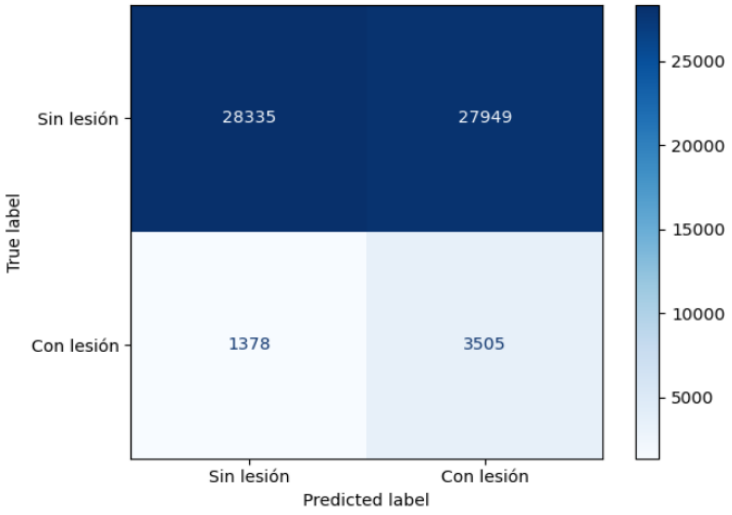
El modelo final se entrenó con el conjunto de variables ya limpio, una vez eliminadas las variables que suponían *data leakage*, quedando un total de 19 variables predictivas. Igual que en el resto de modelos, los datos se separaron en dos partes: una para entrenar el modelo (70%) y otra para probarlo (30%), cuidando que las dos tuvieran la misma proporción de accidentes con y sin lesión. Las variables categóricas se transformaron con *one-hot encoding* para que el modelo pudiera trabajar con ellas. En cuanto al umbral de decisión, se mantuvo el valor por defecto (0,5), ya que, como se ha visto, prioriza el *recall* y, por tanto, la detección de los accidentes con lesión, que es el objetivo principal en este contexto.

Con esta configuración, el modelo alcanza un *recall* del 72% sobre la clase con lesión, aunque con una precisión baja (en torno al 11%) y unos valores de ROC-AUC y Average Precision de 0,67 y 0,16, respectivamente. Se trata, por tanto, de un modelo con una capacidad predictiva moderada y con limitaciones claras a la hora de identificar los accidentes con lesión, lo cual es coherente con la dificultad del problema y se analiza con más detalle en la síntesis final.

Para ver el comportamiento del modelo de forma más visual y detallada, la Figura 8 muestra la matriz de confusión del modelo final. En ella se ve el principal problema del modelo: aunque detecta una buena parte de los accidentes con lesión (verdaderos positivos), clasifica erróneamente como "con lesión" un alto número de accidentes que en realidad no lo son (falsos positivos), lo que explica la baja precisión obtenida. Esta

visualización confirma que el modelo tiende a "sobreavisar", priorizando la detección de lesiones a costa de generar bastantes falsas alarmas.

Figura 8: Matriz de confusión del modelo final (Balanced Random Forest, umbral 0,5)



Fuente: Elaboración propia

Interpretabilidad del modelo mediante SHAP

Una vez evaluado el modelo predictivo, se incorporó un análisis de interpretabilidad con el objetivo de entender mejor en qué se basa el modelo para realizar sus predicciones. Para ello se utilizó el método SHAP (SHapley Additive exPlanations), una técnica que permite medir cuánto influye cada variable en las predicciones del modelo. Es decir, ayuda a saber qué variables tienen más peso a la hora de que el modelo decida si un accidente tendrá lesión o no.

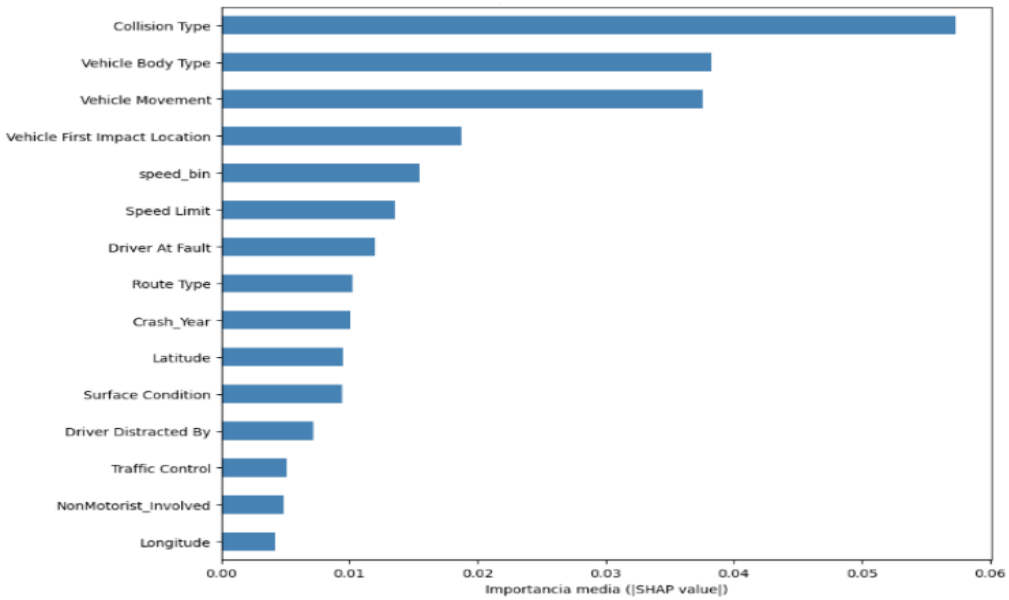
A diferencia del resto del análisis, este apartado se hizo sobre el modelo explicativo y no sobre el predictivo. La razón es que, para la interpretación, resulta mucho más útil este modelo: el modelo explicativo incluye las variables que describen el propio accidente (tipo de colisión, movimiento del vehículo, punto de impacto, etc.), que son precisamente las que más relación tienen con la gravedad de las lesiones. Por ello, analizar qué variables pesan más en este modelo permite extraer conclusiones más influyentes sobre los factores asociados a la siniestralidad. Es importante recordar, eso sí, que este enfoque no tiene utilidad predictiva real, ya que utiliza información que solo se conoce

tras el accidente, sino que se plantea como un análisis complementario de los factores de riesgo y para saber qué está usando el modelo.

Antes de calcular los valores SHAP, se entrenó una versión del modelo explicativo (Balanced Random Forest) excluyendo la variable '*Vehicle Damage Extent*'. Esta variable está tan relacionada con la variable objetivo que casi dice lo mismo (si el vehículo tiene daños graves, lo normal es que haya lesiones). Por eso acaparaba casi toda la importancia del modelo y no dejaba ver el peso del resto de variables. Al excluirla, se consigue identificar con mayor claridad los demás factores asociados a la gravedad. La variable '*Vehicle Damage Extent*' solo se quitó para este análisis SHAP; en la tabla comparativa de modelos se sigue utilizando el modelo explicativo completo, que la incluye.

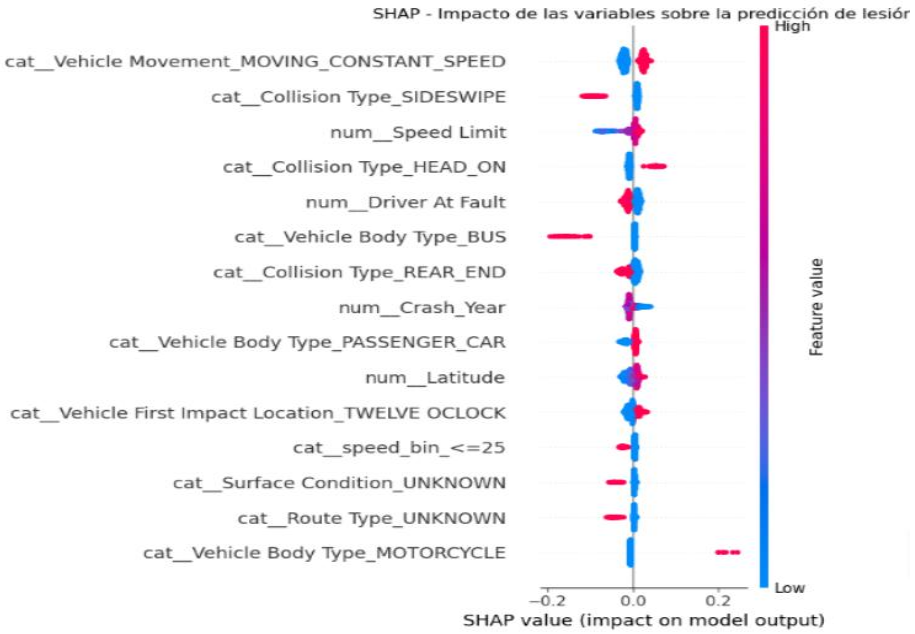
Los valores SHAP se calcularon sobre una muestra de 1.000 accidentes del conjunto de prueba. Como las variables categóricas se habían dividido en varias columnas con el *one-hot encoding*, se prepararon dos gráficos complementarios. El primero (Figura 9) es un gráfico de barras que muestra qué variables influyen más en el modelo, juntando de nuevo todas las categorías de cada variable original para que se entienda mejor. El segundo (Figura 10) es un gráfico de tipo beeswarm, que además de la importancia de cada variable permite ver en qué dirección influye en la predicción.

Figura 9: Top 15 variables por importancia según SHAP (modelo explicativo)



Fuente: Elaboración propia

Figura 10: Efecto de las variables sobre la predicción de lesión (beeswarm)



Fuente: Elaboración propia

En el gráfico de barras (Figura 9) se observa que las variables con mayor peso en las predicciones del modelo son el tipo de colisión (*Collision Type*), el tipo de vehículo (*Vehicle Body Type*) y el movimiento del vehículo (*Vehicle Movement*), seguidas del lugar de impacto (*Vehicle First Impact Location*) y de la velocidad (*speed_bin* y *Speed Limit*). En cambio, las variables relacionadas con el entorno (tipo de vía, condición de la superficie, control de tráfico) y las coordenadas geográficas presentan una contribución bastante menor. Esto refleja que el modelo se basa sobre todo en factores relacionados con la dinámica del propio accidente, más que en las condiciones previas o ambientales.

El gráfico beeswarm (Figura 10) permite analizar también la dirección de estos efectos, y los patrones que muestra resultan coherentes con lo que cabría esperar en seguridad vial. Por ejemplo, cuando el accidente es una colisión frontal (HEAD_ON), el modelo tiende a predecir que sí habrá lesión, mientras que en las colisiones laterales por roce (SIDESWIPE) se inclina hacia que no la habrá. Esto tiene sentido, ya que un choque frontal suele ser bastante más grave que un simple roce lateral. El tipo de vehículo

también influye bastante, ya que ir en motocicleta es lo que más empuja la predicción hacia que haya lesión, mientras que ir en autobús la reduce, lo cual tiene sentido porque sus pasajeros están más protegidos. Además, cuanto mayor es el límite de velocidad, más tiende el modelo a predecir que habrá lesión.

Aun así, estos resultados deben interpretarse con cuidado por varios motivos. En primer lugar, los valores de SHAP que se obtienen son bastante pequeños, algo que tiene que ver con el fuerte desbalanceo de la variable objetivo. Por eso, lo importante no es el número en sí, sino comparar unas variables con otras para ver cuáles pesan más, es decir, la importancia relativa. En segundo lugar, algunas categorías con bastante peso corresponden a valores desconocidos (UNKNOWN), lo que muestra limitaciones en la calidad del registro de los datos y obliga a tener cuidado al sacar conclusiones. Por último, conviene recordar que SHAP solo explica cómo funciona el modelo por dentro, es decir, qué variables usa para tomar sus decisiones. Esto no significa que esas variables sean la causa real de las lesiones, así que los resultados hay que interpretarlos como una ayuda para entender el modelo, no como una relación de causa y efecto.

En conjunto, este análisis refuerza una de las ideas centrales del TFG: la gravedad de un accidente se explica en gran medida por factores que solo se conocen cuando ya ha ocurrido el accidente (tipo de colisión, punto de impacto, dinámica del vehículo). Esto ayuda a entender por qué el modelo predictivo, que solo dispone de variables previas al accidente, tiene una capacidad predictiva limitada: los factores que más se relacionan con la gravedad son, en buena parte, propios del accidente y no se conocen de antemano.

Síntesis de la capacidad predictiva

En general, los resultados muestran que la capacidad predictiva del modelo es limitada. El mejor modelo predictivo (Balanced Random Forest) consigue detectar buena parte de los accidentes con lesión, pero a costa de una precisión baja y un elevado número de falsas alarmas, con unos valores de ROC-AUC y Average Precision que, aunque superiores al azar, son modestos.

Lo más relevante es que ninguna de las estrategias probadas a lo largo del análisis (cambiar de algoritmo, usar técnicas de balanceo como SMOTE o el Balanced Random Forest, revisar y eliminar variables o ajustar el umbral de decisión) consiguió una mejora

significativa de los resultados. Esto apunta a que la limitación principal no está en el modelo ni en la técnica empleada, sino en la propia información disponible: las variables que se conocen antes del accidente no capturan los factores que realmente determinan la gravedad de las lesiones. De hecho, como se ha visto en el análisis SHAP, los factores más asociados a la gravedad son propios del accidente (tipo de colisión, tipo de vehículo, punto de impacto) y no están disponibles en un escenario de predicción real.

Aun así, el hecho de no haber obtenido un modelo con un rendimiento alto no resta valor al trabajo. Al contrario, el proceso que se ha seguido ha permitido aprender a manejar un problema real de clasificación con datos imperfectos y desbalanceados, a elegir y justificar las métricas adecuadas, y a probar de forma crítica distintas técnicas, comprobando que no siempre las soluciones más sofisticadas mejoran los resultados. En este sentido, el principal aprendizaje del trabajo no es tanto el modelo final en sí, como comprender los retos reales que suponen predecir un fenómeno como la siniestralidad y las limitaciones que imponen los datos disponibles. Estas limitaciones, así como las posibles líneas de mejora, se desarrollan con más detalle en el apartado 6.2.

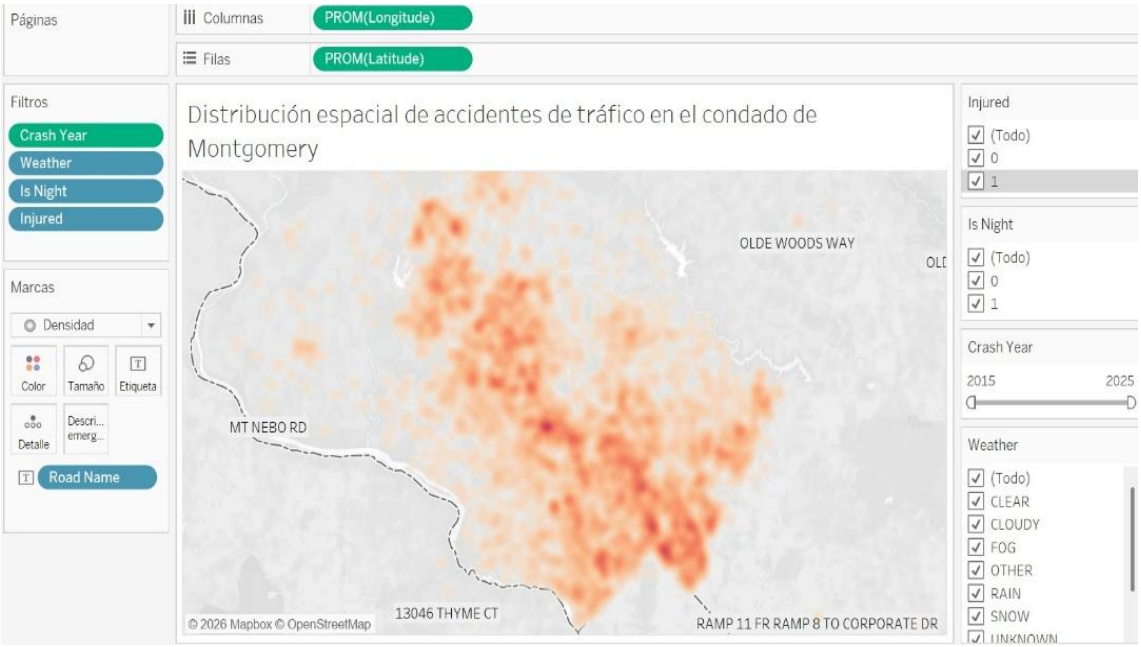
5.4. Interpretación y visualización en dashboard

Como complemento al análisis exploratorio y al modelo predictivo, se elaboró una serie de visualizaciones en Tableau para presentar de forma más visual e intuitiva algunos de los patrones de siniestralidad identificados a lo largo del trabajo. La elección de Tableau se debe a que permite crear gráficos interactivos y mapas de forma sencilla, facilitando la exploración de los datos y la comunicación de los resultados a un público no técnico, algo muy útil en el ámbito de la seguridad vial. Estas visualizaciones no pretenden aportar un análisis nuevo, sino reforzar visualmente algunas conclusiones obtenidas anteriormente.

En primer lugar, se elaboró un mapa de calor que muestra la distribución espacial de los accidentes en el condado de Montgomery (Figura 11), a partir de las coordenadas de latitud y longitud, donde la intensidad del color refleja el número de accidentes en cada zona. Este mapa incorpora además varios filtros interactivos por año, condiciones meteorológicas y franja horaria (día/noche), y, sobre todo, un filtro por la variable objetivo '*Injured'*, que es probablemente el más relevante, ya que permite distinguir entre los accidentes en los que hubo lesión y aquellos en los que no. De esta forma, no solo se

observa dónde se concentran los accidentes, sino también en qué zonas tienden a producirse con mayor frecuencia accidentes con lesión, lo que conecta directamente con el objetivo principal del trabajo y aporta una dimensión espacial que el resto del análisis no recogía.

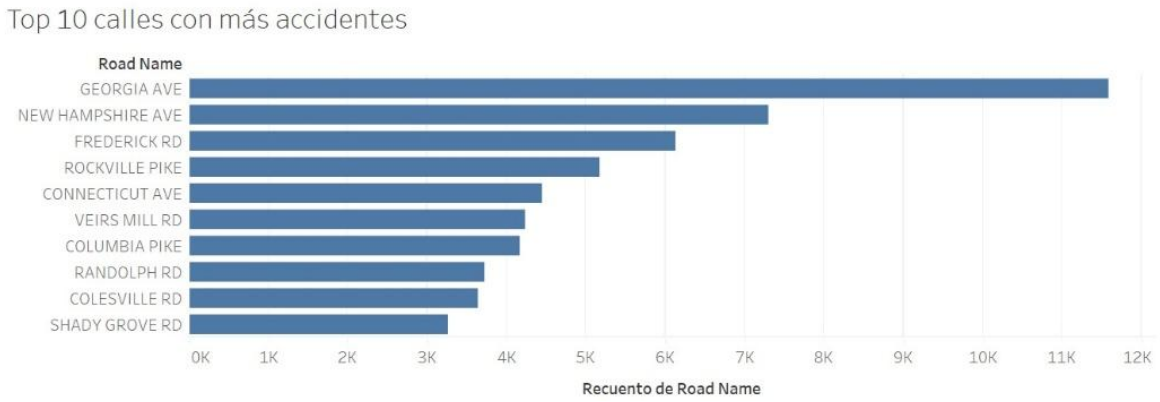
Figura 11: Distribución espacial de accidentes de tráfico en el condado de Montgomery



Fuente: Elaboración propia con Tableau

En segundo lugar, se incluyó un gráfico con las diez calles con mayor número de accidentes (Figura 12), elaborado a partir de la variable '*Road Name'*, que permite identificar rápidamente los puntos más conflictivos de la red viaria del condado. Su interés reside en que traslada el análisis a un nivel más concreto y operativo: mientras que el mapa anterior ofrece una visión general por zonas, este gráfico señala las vías concretas sobre las que podrían orientarse posibles medidas de seguridad vial.

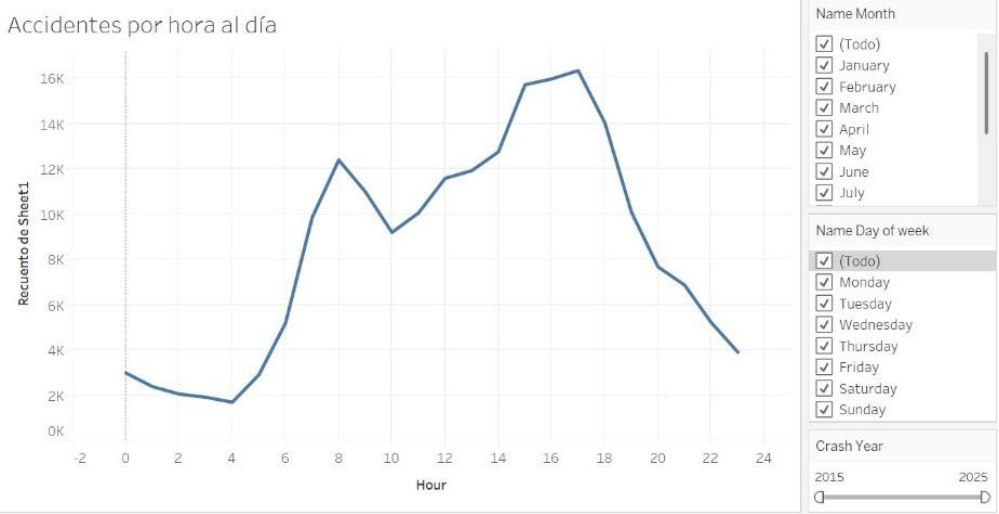
Figura 12: Top 10 calles con más accidentes



Fuente: Elaboración propia con Tableau

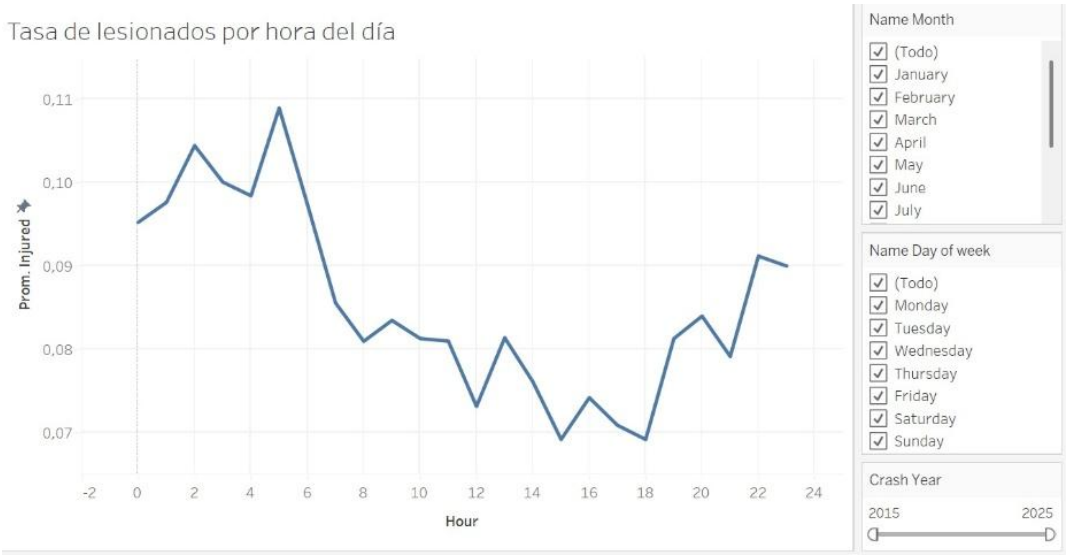
Por último, se elaboraron dos gráficos de líneas centrados en la dimensión temporal. El primero muestra el número de accidentes según la hora del día (Figura 13), lo que permite identificar las franjas horarias con mayor siniestralidad y comprobar que los accidentes no se reparten de manera uniforme a lo largo del día, sino que se concentran en determinados tramos, probablemente asociados a una mayor intensidad de tráfico. El segundo representa la tasa de lesionados según la hora del día (Figura 14); este último ya se había analizado en el apartado bivariante, pero se decidió rehacerlo en Tableau para presentarlo de forma más clara y coherente con el resto del panel y, sobre todo, para poder aprovechar los filtros interactivos que permite la herramienta, lo que ofrece una visualización mucho más dinámica del mismo patrón y permite explorarlo cruzándolo con otras variables.

Figura 13: Número de accidentes por hora del día



Fuente: Elaboración propia con Tableau

Figura 14: Tasa de lesionados por hora del día (Tableau)



Fuente: Elaboración propia con Tableau

En general, estas visualizaciones dan una visión complementaria y más accesible de los patrones principales de siniestralidad detectados en el análisis, tanto desde el punto de vista espacial como temporal.

6. Conclusiones

6.1. Principales hallazgos

Una vez finalizado el trabajo, se pueden extraer varias conclusiones en relación con los objetivos planteados al principio del TFG.

En cuanto al tratamiento de los datos, uno de los principales aprendizajes ha sido enfrentarse a un conjunto de datos real y bastante imperfecto. Gran parte del esfuerzo del trabajo se ha dedicado a la limpieza y transformación de la información, lo que ha permitido comprobar de primera mano que, en un proyecto de analítica de datos, esta fase es tan importante como la modelización en sí. El hecho de trabajar con un *dataset* mayoritariamente categórico y con numerosas inconsistencias ha supuesto un reto que ha aportado un aprendizaje valioso sobre el manejo de datos reales.

En el análisis exploratorio se identificaron algunos patrones de interés: la siniestralidad varía según la hora del día y se concentra en determinadas zonas y vías del condado, mientras que variables como el tipo de vehículo (especialmente las motocicletas) o el tipo de colisión muestran una relación clara con la probabilidad de sufrir lesiones.

Respecto al modelo predictivo, la principal conclusión es que su capacidad para anticipar la gravedad de un accidente a partir de la información disponible antes del mismo es limitada. Se probaron distintos algoritmos y varias técnicas para tratar el desbalanceo, así como el ajuste del umbral de decisión, pero ninguna de ellas logró una mejora significativa de los resultados. Lejos de ser un fracaso, este hallazgo resulta en sí mismo importante: muestra que la limitación no está tanto en la técnica empleada como en la propia información disponible, ya que los factores que más influyen en la gravedad de un accidente solo se conocen una vez ya ha pasado. Esta idea se vio reforzada por el análisis de interpretabilidad con SHAP, que confirmó que las variables más asociadas a las lesiones son las que describen el propio accidente.

En conjunto, el trabajo demuestra que la analítica de datos permite extraer información útil sobre la siniestralidad vial, aunque también muestra las dificultades de predecir un fenómeno tan complejo con los datos disponibles. Por encima del rendimiento del modelo final, el principal valor de este trabajo reside en el proceso seguido: aprender a manejar datos reales, a elegir y justificar las métricas adecuadas, a probar de forma

crítica distintas técnicas y a interpretar los resultados con honestidad, entendiendo tanto las posibilidades como los límites de la analítica de datos aplicada a un problema real.

6.2. Limitaciones del estudio

Más allá de los resultados obtenidos, conviene señalar algunas limitaciones relativas al alcance del propio estudio y a las conclusiones que pueden extraerse de él.

La primera limitación tiene que ver con la definición de la variable objetivo. Por las razones explicadas, el problema se planteó como una clasificación binaria entre accidentes con y sin lesión. Esto implica que el modelo no permite determinar la gravedad concreta de un accidente, sino únicamente si la persona implicada resultó o no lesionada. Es decir, no se distingue entre una lesión leve, una grave o un fallecimiento, lo que reduce el nivel de detalle de las conclusiones. Aunque esta simplificación estaba justificada por el fuerte desbalanceo de las categorías más graves, supone una limitación a la hora de interpretar los resultados en términos de gravedad.

En segundo lugar, la capacidad predictiva del modelo es limitada, por lo que sus resultados no deberían utilizarse como base para la toma de decisiones reales sin un análisis más profundo. El trabajo permite identificar tendencias y factores asociados a la siniestralidad, pero no proporciona una herramienta de predicción fiable en un contexto operativo y real.

Por último, el análisis de interpretabilidad realizado con SHAP debe entenderse también con cautela. Esta técnica explica qué variables utiliza el modelo para tomar sus decisiones, pero no permite establecer relaciones de causa y efecto. Por tanto, aunque el análisis señala qué factores se asocian a una mayor probabilidad de lesión, no puede afirmarse que dichos factores sean la causa real de estas.

En conjunto, estas limitaciones no invalidan el trabajo, pero sí delimitan el alcance de sus conclusiones, que deben interpretarse como una primera aproximación al problema más que como resultados definitivos.

6.3. Futuras líneas de investigación

A partir del trabajo realizado y de las limitaciones señaladas, se plantean varias líneas de mejora que podrían desarrollarse en el futuro.

Una primera línea sería enriquecer el conjunto de datos con nuevas variables que aporten información más relevante para la predicción. En particular, sería interesante incorporar un mayor número de variables numéricas, ya que el *dataset* utilizado es mayoritariamente categórico, lo que condiciona tanto el entrenamiento como la interpretación de los modelos. Además, como se ha visto, una de las principales limitaciones es que los factores más determinantes solo se conocen después del accidente, por lo que incorporar datos como el volumen de tráfico, la velocidad real en el momento del impacto o información meteorológica más detallada podría mejorar la capacidad predictiva. En esta línea, y tal como se recogía en el estado del arte, resultaría útil combinar estos datos con otros conjuntos de datos o fuentes externas, como imágenes del entorno, datos de tráfico en tiempo real o información procedente de dispositivos *IoT*.

Relacionado con lo anterior, otra posible mejora sería trabajar con conjuntos de datos más completos o mejor estructurados que el utilizado en este trabajo. El *dataset* empleado, al ser de carácter preliminar y proceder de informes policiales, presentaba numerosas inconsistencias y valores desconocidos; partir de fuentes más refinadas o verificadas permitiría reducir el ruido y obtener resultados más fiables.

Una tercera línea consistiría en reformular el problema como una clasificación multiclase, prediciendo el nivel concreto de gravedad (leve, grave o mortal) en lugar de una variable binaria. Esto permitiría obtener conclusiones más detalladas, aunque exigiría disponer de más casos de las categorías más graves para tratar correctamente su escasez.

También sería interesante probar otros algoritmos más avanzados, como modelos de *gradient boosting* (por ejemplo, CatBoost, especialmente indicado para variables categóricas) o incluso redes neuronales, para comprobar si consiguen mejorar los resultados. Del mismo modo, ampliar el análisis a otros condados o regiones, más allá de Montgomery, permitiría comparar patrones de siniestralidad y valorar hasta qué punto las conclusiones son generalizables.

Por último, una línea más orientada a la aplicación práctica sería desarrollar y ampliar el panel de visualización interactivo, de manera que pudiera servir como herramienta de apoyo a la toma de decisiones para responsables de seguridad vial y administraciones públicas, facilitando la identificación de zonas y factores de riesgo.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

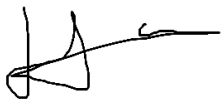
ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Carmen Humanes Fernández, estudiante de E2 + Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis exploratorio y modelización de lesiones en accidentes de tráfico en el condado de Montgomery" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código solo en el contexto de las actividades descritas a:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.
6. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
8. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
9. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
10. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
11. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
12. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
13. 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: Junio 2026

Firma: 

7. Bibliografía

American Association of State Highway and Transportation Officials. (2010). *Highway Safety Manual*. AASHTO.

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). *Real-time crash prediction in an urban expressway using disaggregated data*. Transportation Research Part C: Emerging Technologies, 86, 303–318. https://www.sciencedirect.com/science/article/pii/S0968090X17303315

Bao, J., Liu, P., & Ukkusuri, S. V. (2019). *A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data*. Accident Analysis & Prevention, 122, 239–254. https://www.sciencedirect.com/science/article/abs/pii/S0001457518303877

Çodur, M. Y., & Tortum, A. (2015). *An artificial neural network model for highway accident prediction: A case study of Turkey*. Promet – Traffic & Transportation, 27(3), 217–225. https://hrcak.srce.hr/file/207793

Elvik, R. (2007). *State-of-the-art approaches to road accident black spot management and safety analysis*. Institute of Transport Economics (TØI), Oslo, Norway. https://www.toi.no/getfile.php/139022-1210254330/Publikasjoner/T%C3%98I%20rapporter/2007/883-2007/883-2007-nett.pdf

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, 35(2), 273–285. https://doi.org/10.1016/S0001-4575(02)00005-2

Huenupi, M., Montero, G., & Thomas, A. (2018). *Estimación de modelos predictivos de accidentes en ámbito urbano para la incorporación en la evaluación social de proyectos en Chile*. Revista de la Sociedad Chilena de Transporte (SOCHITRAN). https://estudiosdetransporte.org/sochitran/article/view/186#:~:text=Elvik%2C%20R.%20%282007%29.%20State,Oslo%2C%20Noruega

Hossain, M., & Muromachi, Y. (2012). *A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways*.

Accident Analysis & Prevention, 45, 373–381. https://www.sciencedirect.com/science/article/pii/S0001457511002247

Marcillo, P., Valdivieso Caraguay, Á. L., & Hernández-Álvarez, M. (2022*). A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources.* Applied Sciences, 12(9), 4529. https://www.mdpi.com/2076-3417/12/9/4529

Maryland Highway Safety Office. (2024). *Dashboard Data Definitions*. Zero Deaths Maryland. https://zerodeathsmd.gov/resources/crashdata/crashdashboard/dashboard-data-definitions/

Montgomery County of Maryland. (2025, 11 de octubre). *Crash Reporting – Drivers Data* [Conjunto de datos]. Catalog (data.gov). https://catalog.data.gov/dataset/crash-reporting-drivers-data

Parra, C., Ponce, C., & Rodrigo, S. F. (2020). *Evaluating the performance of explainable machine learning models in traffic accidents prediction in California*. In Proceedings of the 39th International Conference of the Chilean Computer Science Society (SCCC 2020) (pp. 1–8). IEEE. https://ieeexplore.ieee.org/abstract/document/9281196

Ren, H., Song, Y., Wang, J., Hu, Y., & Lei, J. (2018). *A deep learning approach to the citywide traffic accident risk prediction*. IEEE Access, 6, 68793–68801. https://ieeexplore.ieee.org/abstract/document/8569437

Sohail, A., Cheema, M. A., Ali, M. E., Toosi, A. N., & Rakha, H. A. (2023). Data-driven approaches for road safety: A comprehensive systematic literature review. Safety Science, 158, 105949. https://www.adelnadjarantoosi.info/pdf/SS2022.pdf

Verma, R., & Agarwal, M. M. (2025). *A systematic review on road accident prediction: A special attention toward machine learning and deep learning approaches*. Journal of Circuits, Systems and Computers, 35(02), 2530009. https://www.worldscientific.com/doi/10.1142/S0218126625300090

Vision Zero Network. (2024, February 7). *Big cost savings & equity benefits highlighted in new Vision Zero study.* https://visionzeronetwork.org/big-cost-savings-equity-benefits-highlighted-in-new-vision-zero-study/

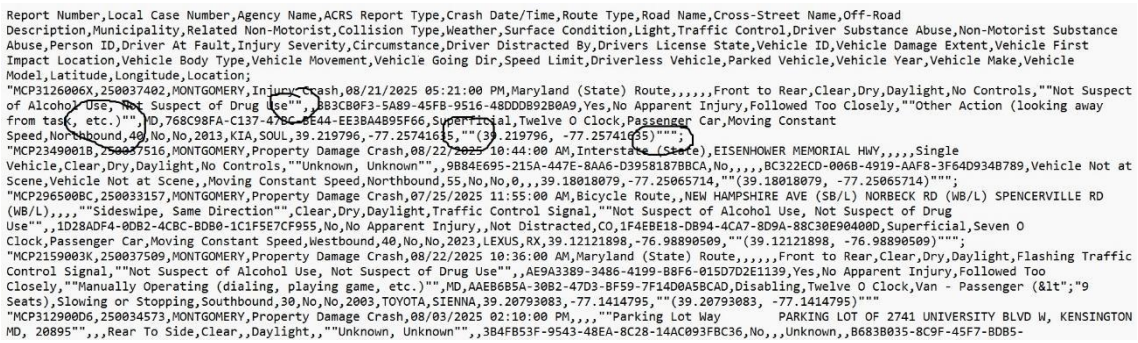
World Bank. (2023, February 21). *Unleashing the power of data to tackle traffic congestion and promote road safety.* Development Data Partnership. https://datapartnership.org/updates/unleashing-the-power-of-data-to-tackle-traffic-congestion-and-promote-road-safety/

World Health Organization. (2023, December 7). *Road traffic injuries*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

8. Anexo

Anexo 1: Estado inicial del conjunto de datos antes del proceso de limpieza

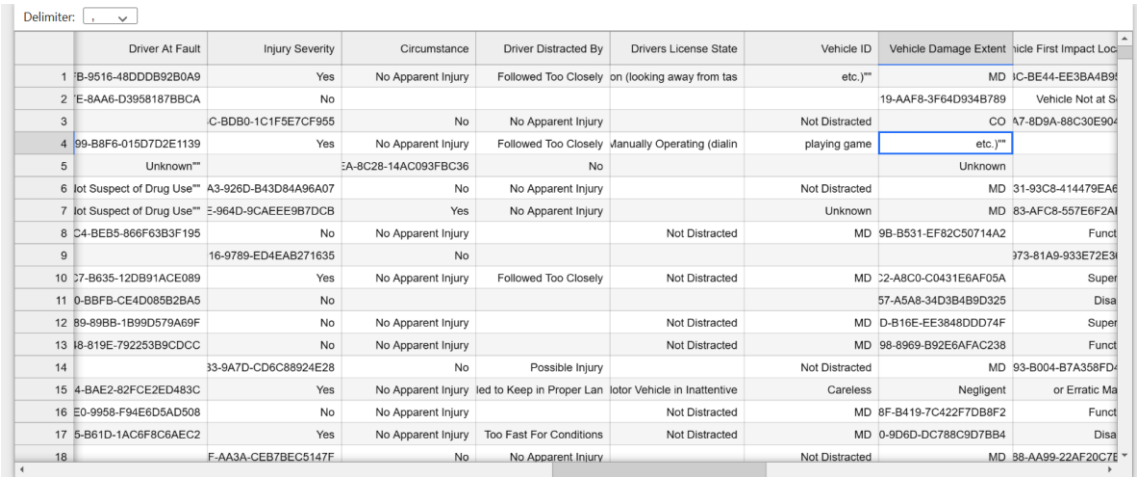
Figura A.1.1. Estructura original del archivo CSV, con presencia de comillas dobles y triples y registros mal delimitados.



```
Report Number,Local Case Number,Agency Name,ACRS Report Type,Crash Date/Time,Route Type,Road Name,Cross-Street Name,Off-Road
Description,Municipality,Related Non-Motorist,Collision Type,Weather,Surface Condition,Light,Traffic Control,Driver Substance Abuse,Non-Motorist Substance
Abuse,Person ID,Driver At Fault,Injury Severity,Circumstance,Driver Distracted By,Drivers License State,Vehicle ID,Vehicle Damage Extent,Vehicle First
Impact Location,Vehicle Body Type,Vehicle Movement,Vehicle Going Dir,Speed Limit,Driverless Vehicle,Parked Vehicle,Vehicle Year,Vehicle Make,Vehicle
Model,Latitude,Longitude,Location;
"MCP3126006X,250037402,MONTGOMERY,Injury Crash,08/21/2025 05:21:00 PM,Maryland (State) Route,,,,,,Front to Rear,Clear,Dry,Daylight,No Controls,""Not Suspect
of Alcohol Use, Not Suspect of Drug Use"",,3B3CB0F3-5A89-45FB-9516-48DDDB92B0A9,Yes,No Apparent Injury,Followed Too Closely,""Other Action (looking away
from tas(, etc.)"",,MD,768C98FA-C137-47BC-8E44-EE3BA4B95F66,Superficial,Twelve O Clock,Passenger Car,Moving Constant
Speed,Northbound,40,No,No,2013,KIA,SOUL,39.219796,-77.25741615,""(39.219796, -77.25741615)""";
"MCP2349001B,250037516,MONTGOMERY,Property Damage Crash,08/22/2025 10:44:00 AM,Interstate (State),EISENHOWER MEMORIAL HWY,,,,,Single
Vehicle,Clear,Dry,Daylight,No Controls,""Unknown, Unknown"",,9B84E695-215A-447E-8AA6-D39581878BCA,No,,,,,BC322ECD-006B-4919-AAF8-3F64D934B789,Vehicle Not at
Scene,Vehicle Not at Scene,,Moving Constant Speed,Northbound,55,No,No,0,,,39.18018079,-77.25065714,""(39.18018079, -77.25065714)""";
"MCP296500BC,250033157,MONTGOMERY,Property Damage Crash,07/25/2025 11:55:00 AM,Bicycle Route,,NEW HAMPSHIRE AVE (SB/L) NORBECK RD (WB/L) SPENCERVILLE RD
(WB/L),,,,,""Sideswipe, Same Direction"",Clear,Dry,Daylight,Traffic Control Signal,""Not Suspect of Alcohol Use, Not Suspect of Drug
Use"",,1D28ADF4-0DB2-4CBC-BDB0-1C1F5E7CF955,No,No Apparent Injury,,Not Distracted,CO,1F4EBE18-DB94-4CA7-8D9A-88C30E90400D,Superficial,Seven O
Clock,Passenger Car,Moving Constant Speed,Westbound,40,No,No,2023,LEXUS,RX,39.12121898,-76.98890509,""(39.12121898, -76.98890509)""";
"MCP2159003K,250037509,MONTGOMERY,Property Damage Crash,08/22/2025 10:36:00 AM,Maryland (State) Route,,,,,,Front to Rear,Clear,Dry,Daylight,Flashing Traffic
Control Signal,""Not Suspect of Alcohol Use, Not Suspect of Drug Use"",,AE9A3389-3486-4199-B8F6-015D7D2E1139,Yes,No Apparent Injury,Followed Too
Closely,""Manually Operating (dialing, playing game, etc.)"",MD,AAEB6B5A-30B2-47D3-BF59-7F14D0A5BCAD,Disabling,Twelve O Clock,Van - Passenger (&lt;"9
Seats),Slowing or Stopping,Southbound,30,No,No,2003,TOYOTA,SIENNA,39.20793083,-77.1414795,""(39.20793083, -77.1414795)"""
"MCP312900D6,250034573,MONTGOMERY,Property Damage Crash,08/03/2025 02:10:00 PM,,,,""Parking Lot Way        PARKING LOT OF 2741 UNIVERSITY BLVD W, KENSINGTON
MD, 20895"",,,Rear To Side,Clear,,Daylight,,""Unknown, Unknown"",,3B4FB53F-9543-48EA-8C28-14AC093FBC36,No,,,Unknown,,B683B035-8C9F-45F7-8DB5-
```

Fuente: Elaboración propia con fichero de texto

Figura A.1.2. Visualización de los datos originales, donde se observa el desajuste entre columnas y la asignación incorrecta de valores.

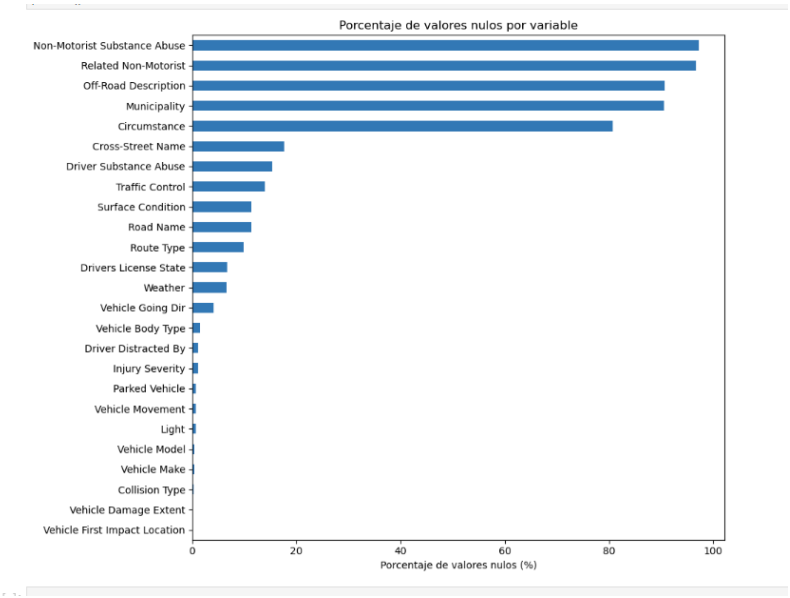


Delimiter: ,

	Driver At Fault	Injury Severity	Circumstance	Driver Distracted By	Drivers License State	Vehicle ID	Vehicle Damage Extent	hicle First Impact Loc
1	'B-9516-48DDDB92B0A9	Yes	No Apparent Injury	Followed Too Closely	on (looking away from tas	etc.)""	MD	)C-BE44-EE3BA4B9
2	'E-8AA6-D3958187BBCA	No					19-AAF8-3F64D934B789	Vehicle Not at S
3		C-BDB0-1C1F5E7CF955	No	No Apparent Injury		Not Distracted	CO	A7-8D9A-88C30E90
4	99-B8F6-015D7D2E1139	Yes	No Apparent Injury	Followed Too Closely	Manually Operating (dialin	playing game	etc.)""	
5	Unknown""		EA-8C28-14AC093FBC36	No			Unknown	
6	lot Suspect of Drug Use""	A3-926D-B43D84A96A07	No	No Apparent Injury		Not Distracted	MD	31-93C8-414479EA6
7	lot Suspect of Drug Use""	E-964D-9CAEEE9B7DCB	Yes	No Apparent Injury		Unknown	MD	83-AFC8-557E6F2A
8	C4-BEB5-866F63B3F195	No	No Apparent Injury		Not Distracted	MD	9B-B531-EF82C50714A2	Funct
9		16-9789-ED4EAB271635	No					973-81A9-933E72E3
10	7-B635-12DB91ACE089	Yes	No Apparent Injury	Followed Too Closely	Not Distracted	MD	2-A8C0-C0431E6AF05A	Super
11	0-BBFB-CE4D085B2BA5	No					57-A5A8-34D3B4B9D325	Disa
12	89-89BB-1B99D579A69F	No	No Apparent Injury		Not Distracted	MD	D-B16E-EE3848DDD74F	Super
13	8-819E-792253B9CDCC	No	No Apparent Injury		Not Distracted	MD	98-8969-B92E6AFAC238	Funct
14		3-9A7D-CD6C88924E28	No	Possible Injury		Not Distracted	MD	93-B004-B7A358FD
15	4-BAE2-82FCE2ED483C	Yes	No Apparent Injury	led to Keep in Proper Lan	lotor Vehicle in Inattentive	Careless	Negligent	or Erratic Ma
16	E0-9958-F94E6D5AD508	No	No Apparent Injury		Not Distracted	MD	8F-B419-7C422F7DB8F2	Funct
17	5-B61D-1AC6F8C6AEC2	Yes	No Apparent Injury	Too Fast For Conditions	Not Distracted	MD	0-9D6D-DC788C9D7BB4	Disa
18		F-AA3A-CEB7BEC5147F	No	No Apparent Injury		Not Distracted	MD	88-AA99-22AF20C7

Fuente: Elaboración propia

Anexo 2: Porcentaje de valores nulos por variable



Fuente: Elaboración propia

Anexo 3: Variable multietiqueta 'Driver Substance Abuse'

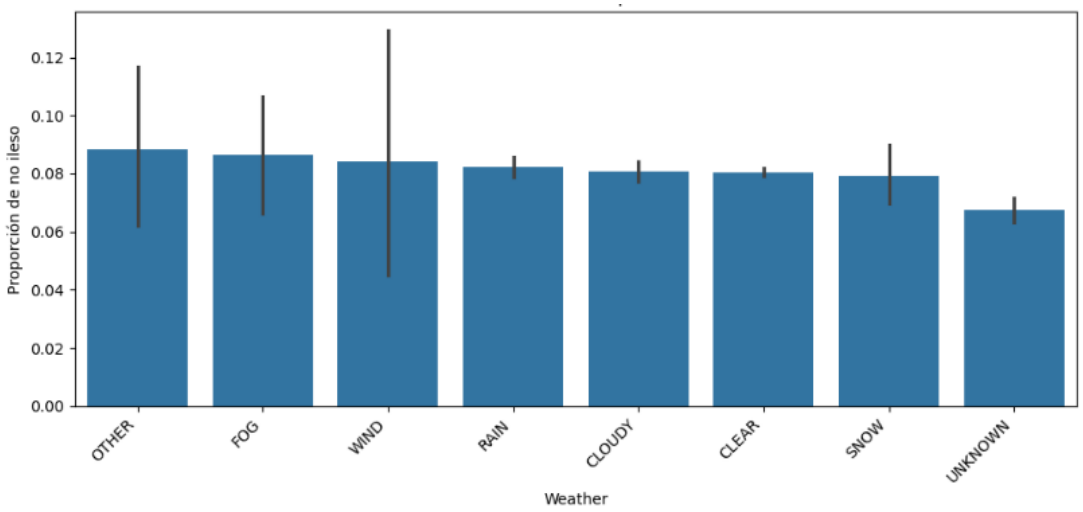
4]:

	Driver Substance Abuse
0	Not Suspect of Alcohol Use, Not Suspect of Dru...
1	Unknown, Unknown
2	Not Suspect of Alcohol Use, Not Suspect of Dru...
3	Not Suspect of Alcohol Use, Not Suspect of Dru...
4	Unknown, Unknown
5	Not Suspect of Alcohol Use, Not Suspect of Dru...
6	Not Suspect of Alcohol Use, Not Suspect of Dru...
7	Not Suspect of Alcohol Use, Not Suspect of Dru...
8	Unknown, Unknown
9	Not Suspect of Alcohol Use, Not Suspect of Dru...

Fuente: Elaboración propia con Python

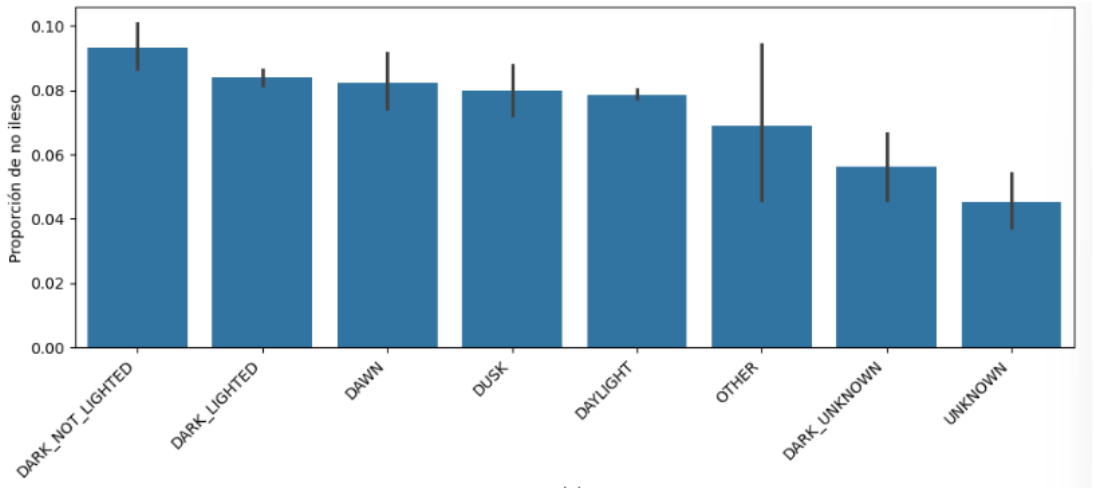
Anexo 4: Análisis bivariante entre las variables explicativas y la variable objetivo

Figura A.4.1. Proporción de lesionados según la variable Weather



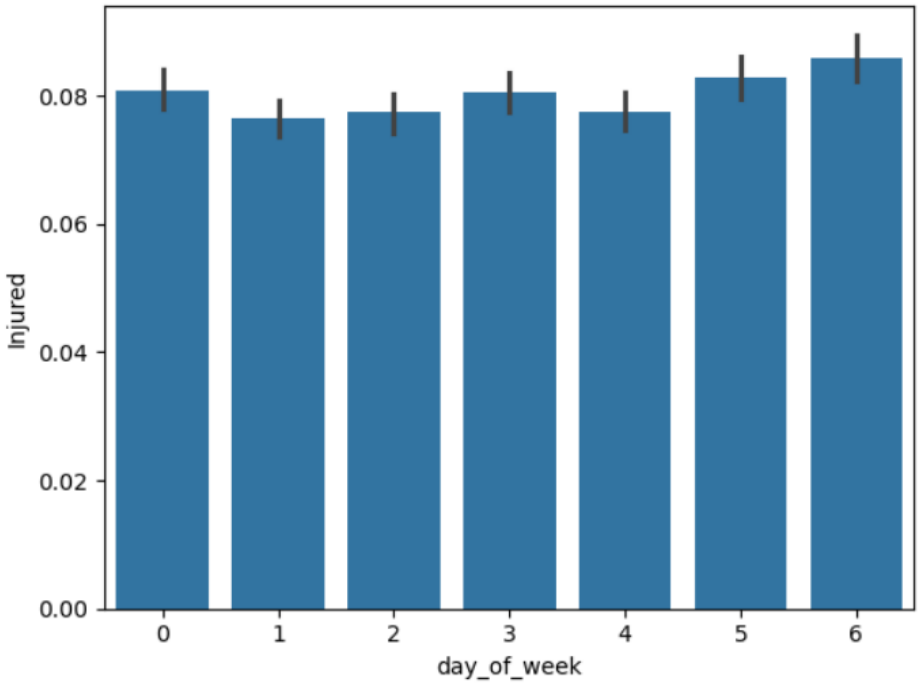
Fuente: Elaboración propia con Python

Figura A.4.2. Proporción de lesionados según la variable Light



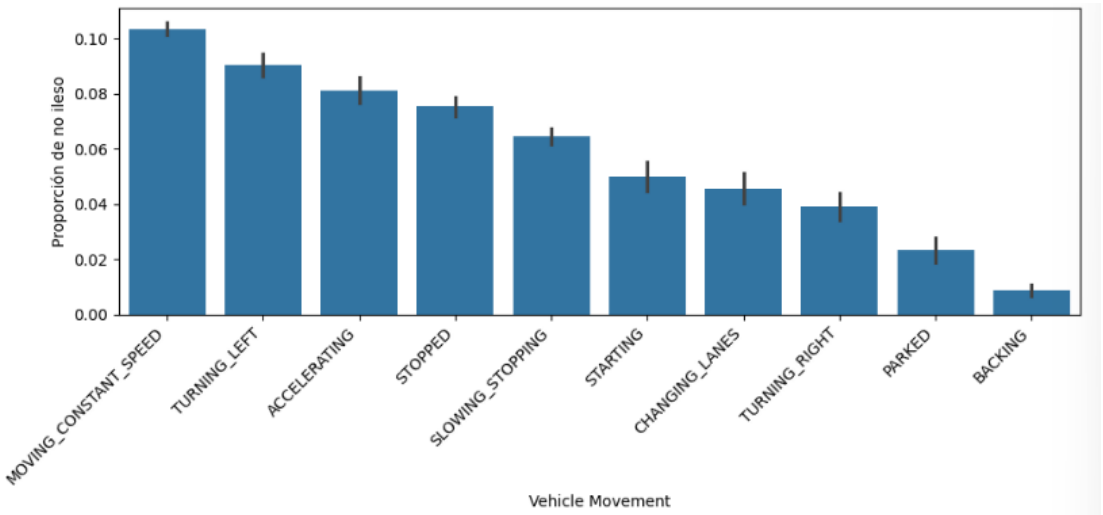
Fuente: Elaboración propia con Python

Figura A.4.3. Proporción de lesionados según la variable 'day_of_week'



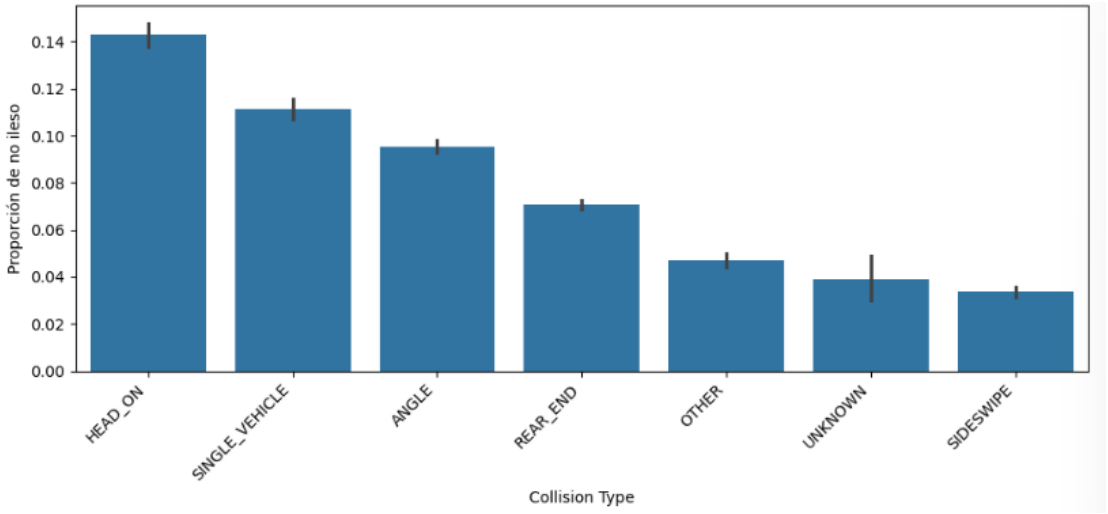
Fuente: Elaboración propia con Python

Figura A.4.4. Proporción de lesionados según la variable 'Vehicle Movement'



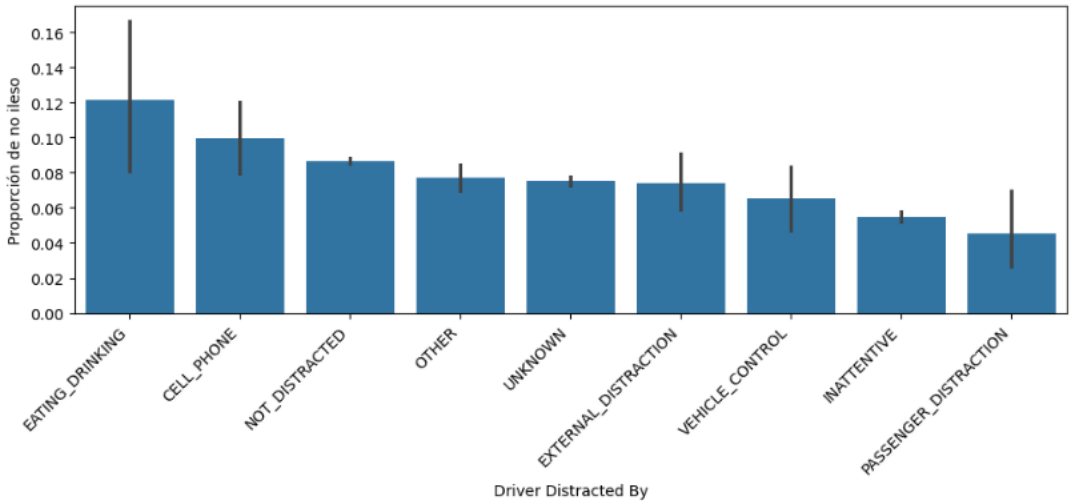
Fuente: Elaboración propia con Python

Figura A.4.5. Proporción de lesionados según la variable 'Collision Type'



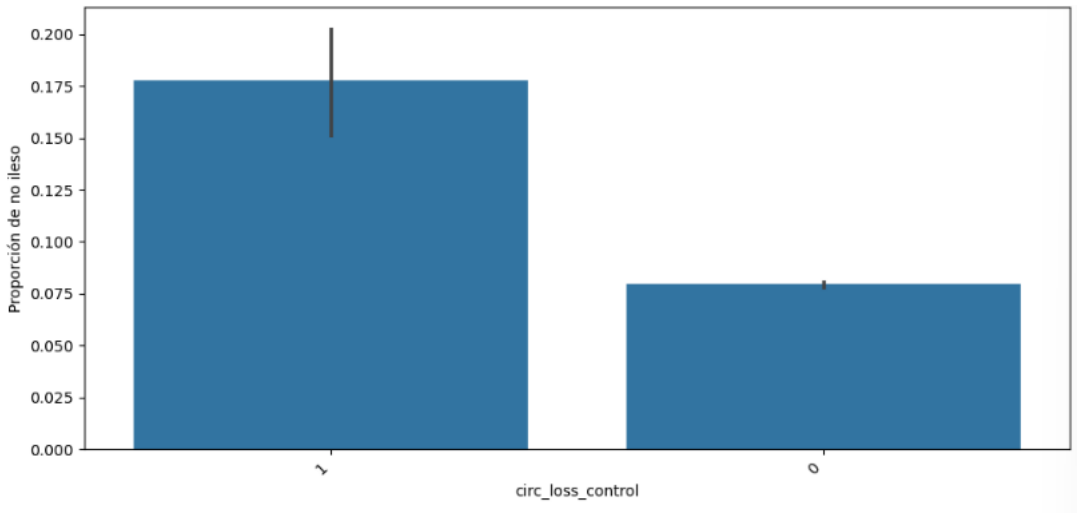
Fuente: Elaboración propia con Python

Figura A.4.6. Proporción de lesionados según la variable 'Driver Distracted By'



Fuente: Elaboración propia con Python

Figura A.4.7. Proporción de lesionados según la variable 'circ_loss_control'



Fuente: Elaboración propia con Python

Tabla A.4.8. Distribución de la variable objetivo (con/sin lesión) según distintas variables de comportamiento y circunstancias del conductor, en porcentaje

Variable	Categoría	Sin lesión (%)	Con lesión (%)
Driver_Substance_Abuse	No (0)	92,10	7,90
	Sí (1)	89,85	10,15
circ_driver_behavior	No (0)	92,06	7,94
	Sí (1)	87,87	12,13
circ_traffic_violation	No (0)	92,06	7,94
	Sí (1)	88,83	11,17

Fuente: Elaboración propia

Anexo 5: Código de Python

El código completo desarrollado para este trabajo está disponible públicamente en el siguiente repositorio de GitHub: https://github.com/carmenhfdz/TFG-Analytics---Accidentes-Montgomery