FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

Título: **Análisis de la Gravedad de los Accidentes en Reino Unido entre 2005 y 2014**

***Alumno*: Sara Gayo Rodríguez**

***Tutor*: Javier Castro**

Junio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

Índice de Tablas

Índice de Figuras

1. Resumen

1.1. Resumen

Este Trabajo de Fin de Grado estudia los factores que influyen en la gravedad de los accidentes de tráfico en el Reino Unido (2005–2014) usando las bases de datos *Accidents*, *Vehicles* y *Casualties* para crear un conjunto de datos con información sobre las personas implicadas en el accidente, los vehículos y las condiciones y el entorno del accidente. Se seleccionaron 32 variables tras limpiar y agrupar datos, definiendo como variables a predecir *Accident_Severity* (leve, moderado, grave) y *Accident_Severity_Score* (suma ponderada). El análisis descriptivo univariante mostró que el 85.2 % de los siniestros fueron leves, los conductores tenían una edad media cercana a 40 años y predominaban los vehículos de 5–10 años de antigüedad, entre otros. En el estudio bivariante se hallaron fuertes vínculos entre las características de los accidentados y el grado de severidad, así como entre las variables sobre la edad. Con clústeres k-medias (k=5, 7) se agruparon los accidentes según sus características, formando grupos de accidentes similares. Para predecir la gravedad se usaron regresión logística ordinal y árboles de clasificación para *Accident_Severity*, y regresión lineal y árboles de regresión para *Accident_Severity_Score*, identificando que edades avanzadas en los accidentados, las condiciones lumínicas y meteorológicas y la velocidad de la vía, entre otros, son algunos de los factores más influyentes. En conclusión, los factores humanos (edad y número de implicados) pesan más que los del entorno, aunque la combinación de ambos mejora las predicciones y puede guiar campañas de prevención centradas en motoristas, personas mayores y zonas con velocidad limitada a 15 mph, donde el riesgo de accidentes graves es mayor.

1.2. Abstract

This project studies the factors that influence the severity of traffic accidents in the United Kingdom (2005–2014) using the databases *Accidents*, *Vehicles*, and *Casualties* to create a dataset containing information about the people involved in the accident, the vehicles, and the conditions and environment of the accident. Thirty-two variables were selected after cleaning and grouping data, defining as target variables *Accident_Severity* (slight, serious, fatal) and *Accident_Severity_Score* (weighted sum). The univariate descriptive analysis showed that 85.2 % of incidents were slight, drivers had an average age close to 40 years, and vehicles aged 5–10 years predominated, among other findings. In the bivariate study, strong links were found between the characteristics of the accident victims and the degree of severity, as well as among age-related variables. With k-means clustering ($k = 5, 7$), accidents were grouped according to their characteristics, forming clusters of similar accidents. To predict severity, ordinal logistic regression and classification trees were used for *Accident_Severity*, and linear regression and regression trees were used for *Accident_Severity_Score*, identifying that advanced ages of the accident victims, lighting and weather conditions, and road speed, among others, are some of the most influential factors. In conclusion, human factors (age and number of people involved) weigh more than environmental factors, although the combination of both improves predictions and can guide prevention campaigns focused on motorcyclists, older people, and areas with a speed limit of 15 mph, where the risk of severe accidents is higher.

2. Introducción

2.1. Contexto

Los accidentes de tráfico siguen siendo uno de los principales problemas de seguridad vial en todo el mundo. Cada año causan millones de heridos y fallecidos. Según la Organización Mundial de la Salud (OMS), provocan más de 1.35 millones de muertes al año y son la principal causa de muerte entre jóvenes de 5 a 29 años [1].

En países como el Reino Unido se ha logrado reducir un 36 % las víctimas mortales en las últimas décadas: de más de 5.000 muertes anuales en los años 90 a menos de 1.800 en la actualidad, según su Departamento de Transporte [2].

El Departamento de Transporte británico recopila una gran extensión de datos sobre cada accidente de tráfico. Esta información se publica en bases de datos como STATS19, que recogen información como la fecha, lugar, tipo de vía, condiciones del entorno, vehículos implicados y personas afectadas, entre otros. Estos datos permiten hacer análisis muy detallados [3].

Distintos grupos de investigación han utilizado esta información para analizar los accidentes desde diferentes puntos de vista. Algunos trabajos se han centrado en las variables que influyen en la gravedad de los accidentes y otros han usado modelos estadísticos o de inteligencia artificial para hacer predicciones.

Por ejemplo, en un estudio realizado en 2023 [4] se combinaron técnicas como Random Forest, modelos econométricos y análisis de series temporales para predecir la gravedad de los accidentes. También aplicaron métodos de inteligencia artificial para saber qué variables eran más importantes. Entre ellas destacaban el tipo de carretera, la zona del accidente y el perfil del conductor.

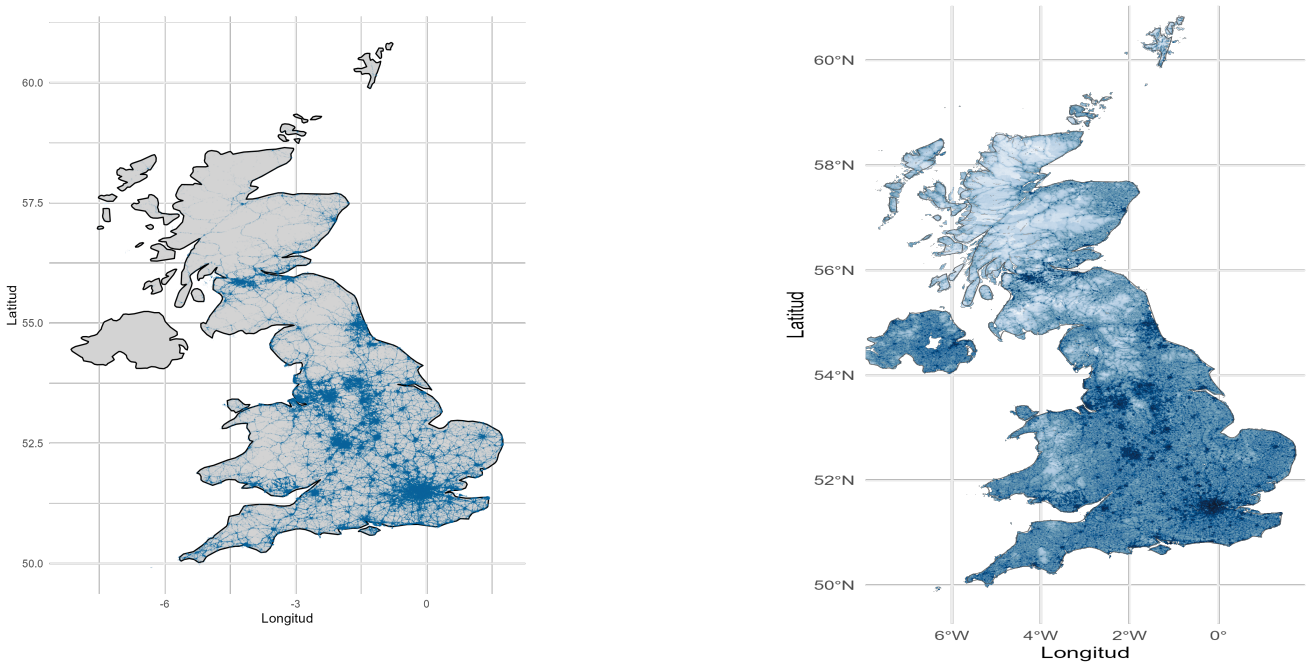
En 2024 [5] se examinaron casi 200 estudios publicados en los últimos cinco años centrados en predecir la gravedad del accidente y otras variables, destacando la efectividad de integrar diversas bases de datos y utilizar técnicas avanzadas de Machine Learning en el análisis de accidentes. Se llegó a la conclusión de que esto mejoraba de manera considerable la precisión de las predicciones. En el estudio se resalta la importancia de seguir mejorando estos sistemas como una herramienta para reducir el número de víctimas, en línea con los objetivos de la OMS para 2030 [1].

La mayoría de los estudios analizan un solo aspecto de los datos, como puede ser centrarse en los accidentes sin tener en cuenta los vehículos o las personas implicadas o hacer predicciones sin realizar un análisis descriptivo previo.

Este trabajo busca ofrecer una visión más completa. Para ello, se han unido tres bases de datos: una con información sobre los accidentes, otra sobre los vehículos implicados y una tercera sobre las personas afectadas [3]. A partir de ellas, se han creado nuevas variables, y los datos se han limpiado y preparado para facilitar su análisis. Se aplicarán tanto técnicas descriptivas como predictivas para entender mejor qué factores influyen en la gravedad de los accidentes.

También se han utilizado métodos de agrupación (como los clústers) para identificar tipos de accidentes similares, y modelos predictivos para anticipar cuáles pueden tener consecuencias más graves.

Antes de pasar al análisis detallado, conviene ver cómo se distribuyen los accidentes en el país. A continuación se muestran dos mapas: el primero con todos los accidentes ocurridos entre 2005 y 2014 y el otro la densidad de población en Reino Unido en 2020. En el primer mapa, cada punto azul representa un accidente y en el segundo, a mayor saturación en el color, más población.



(a) Mapa de accidentes (2005-2014)

(b) Mapa de densidad de población (2020)

Figura 1: Localización de los Accidentes y Densidad de Población en Reino Unido.

La mayoría de los accidentes se concentran en el sur de Inglaterra, especialmente en zonas como Londres y alrededores, donde la densidad de población es mayor. En cambio, en regiones como Escocia hay menos accidentes registrados, posiblemente por tener menos población y menos tráfico. Esto parece confirmar que las zonas urbanas, donde circulan más vehículos y personas, tienen más accidentes.

Al comparar las Figuras 1(a) y 1(b), se ve que las zonas más oscuras del mapa de población coinciden con las regiones donde hay más accidentes. Esto apoya la idea mencionada anteriormente sobre la relación entre la densidad de población y el número de accidentes.

En el mapa de accidentes se pueden observar algunas líneas que recorren el país, que coinciden con las principales carreteras o rutas que conectan las ciudades, lo cual sugiere que no solo importa cuánta gente vive en un sitio, sino también la presencia de una carretera importante que pasa por allí.

2.2. Objetivos y Preguntas de Investigación

El objetivo general de este trabajo es analizar los elementos que influyen en la gravedad de los accidentes de tráfico en el Reino Unido. A partir de este análisis, se busca identificar patrones que puedan servir como base para desarrollar futuras campañas de prevención.

De forma más concreta, los objetivos son:

- Analizar la distribución de las variables *Accident_Severity* (gravedad o severidad del accidente) y *Accident_Severity_Score* (puntuación de la gravedad del accidente), y estudiar cómo se relacionan con el resto de variables del conjunto de datos.
- Desarrollar modelos predictivos que permitan anticipar la gravedad de un accidente según sus características, a nivel humano y del entorno.
- Detectar cuáles son las variables más influyentes a la hora de determinar si un accidente será grave, moderado o leve.
- Identificar factores clave que puedan servir de apoyo para diseñar futuras medidas de prevención o campañas de actuación.

A partir de estos objetivos, se plantean las siguientes preguntas de investigación:

- ¿Qué variables están más asociadas a los accidentes graves?
- ¿Es posible predecir la gravedad de un accidente a partir de información básica como el tipo de vía, el vehículo implicado o las condiciones ambientales?
- ¿Qué perfiles de accidentados o zonas de accidente deberían ser prioritarios en campañas de prevención?
- ¿Hasta qué punto influyen los factores humanos (edad, sexo, tipo de conductor...) frente a los factores del entorno (hora, luz, clima, tipo de impacto...)?

2.3. Metodología

A la hora de llevar a cabo un proyecto de análisis de datos, es útil seguir una serie de pasos que ayuden a organizar bien el trabajo, desde el inicio hasta las conclusiones. Para eso existen metodologías que marcan el camino a seguir. Dos de las más conocidas son *SEMMA* y *CRISP-DM*. A continuación se explica en qué consiste cada una.

2.3.1. SEMMA

La metodología SEMMA fue desarrollada por el SAS Institute. Se centra tanto en el tratamiento técnico de los datos, como en la creación de modelos predictivos. Las cinco fases que componen su nombre son:

1. **Sample (Muestreo):** Se elige una muestra representativa de los datos para poder trabajar más rápido sin perder información importante.
2. **Explore (Explorar):** Se revisan los datos para entender qué contienen, cómo se relacionan las variables y si hay errores o datos atípicos.
3. **Modify (Modificar):** Se preparan los datos para que se puedan usar en modelos. Esto incluye limpiar, transformar y crear nuevas variables si hace falta.
4. **Model (Modelar):** Se aplican modelos que sirven para predecir o clasificar, dependiendo del objetivo del proyecto.
5. **Assess (Evaluar):** Se evalúan los resultados de los modelos para identificar cuál ofrece un mejor rendimiento y predicciones más precisas [6].

2.3.2. CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es una metodología muy usada en proyectos de análisis de datos que permite desde entender el problema hasta aplicar los resultados finales. Está dividida en seis fases:

1. **Comprensión del problema:** Se busca entender bien qué problema se quiere resolver y qué se espera conseguir con el análisis.
2. **Comprensión de los datos:** Se recopilan y revisan los datos disponibles para ver si son de buena calidad, detectar errores y conocer mejor su contenido.
3. **Preparación de los datos:** Se limpian y transforman para que se puedan usar en los modelos. Por ejemplo, se imputan valores missings, se ajustan formatos o se crean nuevas variables, entre otras.
4. **Modelado:** Se aplican distintos modelos estadísticos para resolver el problema y se prueban varias configuraciones para ver cuál funciona mejor.
5. **Evaluación:** Se revisan los resultados para comprobar si realmente ayudan a responder al problema planteado.
6. **Despliegue:** Se usan los resultados en la práctica. Puede ser a través de informes, gráficos o integrando el modelo en un sistema [7].

En este trabajo se ha decidido seguir la metodología CRISP-DM porque se ajusta bien a como está planteada la base de datos. Aunque normalmente empieza con una fase para entender el problema y otra para revisar los datos, en este caso, esas dos partes no han sido el foco principal. Lo más importante ha sido preparar bien los datos, ya que era necesario limpiarlos y transformarlos para poder analizarlos correctamente.

Además, antes de aplicar los modelos, se ha añadido una fase intermedia de análisis descriptivo, en la que se analizan los datos ya preparados para ver cómo se comportan y qué relaciones puede haber entre las variables. Esto ayuda a elegir mejor qué modelos aplicar y a entender mejor los resultados.

3. Obtención de la Base de Datos

La base de datos utilizada se ha obtenido a partir de la unión de tres tablas del mismo conjunto de datos disponible en Kaggle: *Accidents*, *Vehicles* y *Casualties* [3].

La tabla principal sobre la que se trabajará será *Accidents*, a la que se le han añadido distintas variables derivadas de las otras dos tablas.

Para la codificación de las variables categóricas, el propio sitio web incluía un archivo Excel con la descripción de todos los códigos utilizados en las tres bases de datos. Este archivo ha sido fundamental para interpretar y transformar adecuadamente dichas variables.

En los siguientes apartados se presentará un resumen de las tres bases de datos, con las variables nuevas creadas; el proceso que se ha hecho para unir la información y, finalmente, el tratamiento de los valores perdidos.

3.1. Bases de Datos

3.1.1. Accidents

Como se ha comentado anteriormente, la base de datos *Accidents* será la principal sobre la que se trabajará, aunque en una versión modificada para adaptarla al análisis estadístico. Cada fila representa un accidente individual, identificado mediante la variable *Accident_Index*. Para cada accidente se recogen diversas características técnicas, como la localización (latitud, longitud, distrito, distrito policial, etc.), la hora y la fecha.

De las dos variables que se van a predecir, una de ellas ya está incluida en esta base: *Accident_Severity*, cuya codificación es la siguiente:

- **Fatal (1):** accidente con consecuencias graves o mortales para alguno de los involucrados.
- **Serious (2):** accidente con consecuencias moderadas, como heridas con posibles secuelas a largo plazo.
- **Slight (3):** accidente con consecuencias leves, como esguinces o pequeñas fracturas sin secuelas importantes.

La interpretación de estas categorías es propia, ya que el conjunto de datos original solo incluía las etiquetas sin una definición detallada.

La otra variable es *Accident_Severity_Score*, que es una puntuación hecha con la suma ponderada de la gravedad de cada accidentado. Más adelante, en el apartado de Accidentados, se describirán las ponderaciones con mayor detalle.

Otras variables destacadas en esta base hacen referencia a las condiciones de la vía, el clima, la luz, o el número de vehículos y personas implicadas en cada accidente.

Las variables más importantes que se van a utilizar del conjunto *Accidents* son:

Tabla 1: Resumen Variables Más Importantes

Variable	Descripción
Accident_Index	Identificador del Accidente
Accident_Severity	Severidad del Accidente
Number_of_Vehicles	Nº Vehículos
Number_of_Casualties	Nº Accidentados
Date	Fecha del Accidente
Day_of_Week	Día de la Semana
Time	Hora del Accidente
Road_Type	Tipo de Carretera
Speed_limit	Límite de Velocidad
Light_Conditions	Visibilidad Lumínica
Weather_Conditions	Tiempo Meteorológico
Road_Surface_Conditions	Estado de la Carretera
Urban_or_Rural_Area	Área Rural o Urbana

3.1.2. Vehicles

Esta base de datos es más extensa que *Accidents*, ya que cada fila representa un vehículo implicado en algún accidente. Para identificar cada vehículo, se incluye el código del accidente al que pertenece, y además un número que distingue a cada vehículo dentro de ese accidente (1, 2, 3, etc.).

Para agregar la mayor parte de la información de los vehículos a la tabla de accidentes, se han creado las siguientes variables:

Tabla 2: Variables Creadas Base Vehicles

Variable	Descripción
Male_Driver	Nº Conductores (Hombres)
Female_Driver	Nº Conductoras
Avg_Driver_Age	Edad Media de los Conductores
Max_Driver_Age	Edad Máxima de los Conductores
Min_Driver_Age	Edad Mínima de los Conductores
Left_Hand_Drive	Presencia de Algún Conductor Zurdo
Car	Presencia de Coches o Furgonetas Personales
Two_Wheel_Vehicle	Presencia de Vehículos de 2 Ruedas
Trucks	Presencia de Camiones o Camionetas
Impact_Did_Not_Impact	Nº de Primeros Impactos
Impact_Front	Nº Impactos Delanteros
Impact_Back	Nº Impactos Traseros
Impact_Offside	Nº Impactos en Lado Opuesto al Conductor
Impact_Nearside	Nº Impactos en Lado del Conductor
Work_Purpose	Viaje por Trabajo

Tabla 3: Variables Creadas Base Vehicles (continuación)

Variable	Descripción
Education_Purpose	Viaje por Estudios
Avg_Vehicle_Age	Edad Media del Vehículo

3.1.3. Casualties

Esta es la base de datos más grande, ya que hay más personas implicadas que vehículos o accidentes. Contiene más de dos millones de filas, una por cada persona accidentada. Para identificar a cada individuo, se incluye el código del accidente al que pertenece, el identificador del vehículo con el que está relacionado, y un número adicional que indica qué persona es dentro del accidente (siguiendo la misma codificación que la utilizada para los vehículos).

Al igual que en la base de vehículos, se han creado variables que recopilan gran parte de la información de cada accidentado por accidente. Son las siguientes:

Tabla 4: Variables Creadas Base Accidentados

Variable	Descripción
Num_Drivers	Nº Conductores
Num_Passengers	Nº Pasajeros
Num_Pedestrians	Nº Peatones
Male_Casualties	Nº Hombres Accidentados
Female_Casualties	Nº Mujeres Accidentadas
Avg_Casualty_Age	Edad Media de los Accidentados
Max_Casualty_Age	Edad Máxima de los Accidentados
Min_Casualty_Age	Edad Mínima de los Accidentados
Accident_Severity_Score	Puntuación de la Gravedad del Accidente
Front_Seat_Passengers	Nº Copilotos
Rear_Seat_Passengers	Nº Pasajeros (No Copilotos)

En esta base se encuentra la otra variable que se va a predecir, aunque en este caso es una variable continua: *Accident_Severity_Score*. Esta variable ha sido creada a partir de *Casualty_Severity*, que utiliza la misma codificación que *Accident_Severity* de la primera base de datos.

Para construir *Accident_Severity_Score*, se ha realizado una suma ponderada de las puntuaciones de todas las personas implicadas en cada accidente, asignando los siguientes valores a cada categoría de gravedad:

- **Casualty_Severity = 1:** Se multiplica por 1.
- **Casualty_Severity = 2:** Se multiplica por 0.5.
- **Casualty_Severity = 3:** Se multiplica por 0.1.

3.2. Unión de Datos

Al decidir cómo trabajar con las tres bases de datos, se descartó realizar el producto cartesiano entre ellas, principalmente por limitaciones de capacidad computacional. En su lugar, se optó por crear nuevas variables a partir de las bases *Vehicles* y *Casualties*, y añadirlas a la base principal, *Accidents*.

Esta fase fue la más larga del proyecto, ya que muchas variables requirieron varias pruebas hasta encontrar una versión adecuada. En el caso de la variable objetivo continua, *Accident_Severity_Score*, inicialmente se planteó calcular la media de las puntuaciones de gravedad (de 1 a 3) de los accidentados. Sin embargo, este método no resultaba representativo: un accidente con una persona levemente herida (3) y otra en estado muy grave o que había fallecido (1) generaba una media engañosa de 2 (gravedad moderada), que no reflejaba la gravedad real del suceso.

Por ello, se optó por realizar una suma de las puntuaciones, de modo que la gravedad de cada persona quedara reflejada. Aun así, se observó que tener cinco personas con heridas leves no era equivalente a tener una persona fallecida. La solución final fue aplicar una suma ponderada, asignando más peso a los casos más graves (como se detalla en el apartado anterior). Este caso muestra cómo crear una sola variable puede llevar mucho tiempo, ya que fue necesario probar distintas opciones, valorar cuál tenía más sentido y volver a ajustar el código varias veces. A esto se suma que cada intento tardaba bastante en ejecutarse, lo que alargaba aún más el proceso. Estas situaciones se han dado con muchas variables, por eso el trabajo con la base de datos ha conllevado más del 50 % del tiempo total empleado.

Otro grupo de variables especialmente complejo fue el relacionado con el tipo de impacto. Inicialmente se consideró generar una variable por cada combinación posible de impactos (frontal, trasero, lateral, etc.), lo que implicaba más de diez nuevas variables. Esta opción fue descartada por resultar poco manejable. También se evaluó hacer una media de los impactos por accidente, pero su interpretación era confusa, porque no se podían interpretar los decimales de una media. Finalmente, se crearon cuatro variables, una por cada tipo de impacto, más una adicional con el número total de impactos registrados. Esto permitió no perder la información sin saturar la base de datos. Para las edades se crearon variables que resumieran la información utilizando la media, el máximo y el mínimo.

Respecto al motivo del viaje, se optó por crear dos variables resumen que agruparan las categorías más representativas sin llegar a una codificación excesiva. La variable *Journey_Purpose_of_Driver* incluía cinco categorías distintas, de las cuales dos estaban relacionadas con el trabajo (traslado al trabajo o viaje de trabajo), dos con el entorno educativo (estudiante yendo o siendo llevado a su lugar de estudio), y una que estaba sin especificar. Por ello, se crearon dos nuevas variables: *Work_Purpose* y *Education_Purpose*.

También se trabajó con la variable tipo de vehículo, que originalmente contaba con 19 categorías (coches, motocicletas de diversas cilindradas, camiones, vehículos agrícolas, entre otros). Con el objetivo de simplificar, se agruparon en tres variables principales: *Car*, *Two_Wheel_Vehicle* (que incluye motocicletas, bicicletas, scooters, etc.) y *Truck*. Cada una de estas nuevas variables agrupa varias categorías del conjunto original. Por ejemplo, *Two_Wheel_Vehicle* reúne 8 de las 19 categorías originales.

Tras todas estas transformaciones, y algunas adicionales, se pasó de tener 67 variables en total a 60 en la base final. De estas, solo 32 han sido utilizadas para el análisis descriptivo y los modelos predictivos porque había variables como las relacionadas con la localización que no se consideró que pudieran aportar al modelo, ya que algunas de ellas tenían más de 100 categorías.

3.3. Datos Missing

Antes de realizar el análisis descriptivo y predictivo, es necesario identificar qué variables contienen valores perdidos (missing), ya que esto puede afectar a la creación de los modelos.

Tabla 5: Resumen de Missings

Variable	Número Missings	Porcentaje
Avg_Driver_Age	63209	3.85 %
Max_Driver_Age	63209	3.85 %
Min_Driver_Age	63209	3.85 %
Left_Hand_Drive	12512	0.76 %
Avg_Vehicle_Age	247923	15.11 %
Location_Easting_OSGR	111	0.01 %
Location_Northing_OSGR	111	0.01 %
Longitude	112	0.01 %
Latitude	111	0.01 %
Time	133	0.01 %
X.1st_Road_Number	2	0.0001 %
Junction_Detail	18	0.001 %
Junction_Control	585322	36.68 %
X.2nd_Road_Class	676268	41.22 %
X.2nd_Road_Number	16118	0.98 %
Pedestrian_CrossingHuman_Control	21	0.001 %
Pedestrian_CrossingPhysical_Facilities	37	0.002 %
Weather_Conditions	161	0.01 %
Road_Surface_Conditions	2155	0.13 %
Special_Conditions_at_Site	18	0.001 %
Carriageway_Hazards	32	0.002 %
Did_Police_Officer_Attend_Scene_of_Accident	270	0.02 %

El resto de variables que no aparecen en la tabla no tienen valores perdidos.

Entre las variables de la tabla, hay dos que destacan por tener un porcentaje alto de valores perdidos: *Junction_Control* y *X.2nd_Road_Class*, ambas con más de un 35 % de missings. Como se verá más adelante en el gráfico 12, la primera de las variables no tenía casi relación con las variables dependientes. La segunda ni siquiera está incluida ya que, para una mejor visualización, se suprimieron algunas variables que estaban aisladas, siendo *X.2nd_Road_Class* una de ellas. Por este motivo, se ha decidido no trabajar con ellas, y por tanto no se imputarán sus valores.

Como ya se mencionó en el apartado de unión de los datos, variables relacionadas con la localización geográfica o el tiempo, tenían un rango muy amplio de valores: una longitud y latitud por cada accidente o la hora, minutos y segundos de cada uno (hasta 86.400 posibles combinaciones).

De forma similar a las variables anteriores, algunas aparecen en las correlaciones bivariantes con baja relación o directamente se suprimieron para hacer el gráfico más legible. Por estos motivos, las siguientes variables no se utilizarán en los modelos, por lo que no se imputarán sus valores:

- *X.2nd_Road_Number*
- *Pedestrian_CrossingHuman_Control*
- *Pedestrian_CrossingPhysical_Facilities*
- *Location_Easting_OSGR*
- *Location_Northing_OSGR*
- *Longitude*
- *Latitude*
- *Time*
- *X.1st_Road_Number*
- *Junction_Detail*
- *Special_Conditions_at_Site*
- *Carriageway_Hazards*
- *Did_Police_Officer_Attend_Scene_of_Accident*

La única variable que si tenía cierta relación con las variables dependientes y no se incluyó fue *Did_Police_Officer_Attend_Scene_of_Accident*. Esta se consideró eliminarla porque era un factor posterior al accidente y no algo que evitara que se diera o que cambiara su gravedad, ya que los accidentados ya estaban heridos.

Por tanto, las únicas variables que se tendrán en cuenta para imputar valores perdidos son:

Tabla 6: Resumen de Missings con Variables Seleccionadas

Variable	Número Missings	Porcentaje	Moda/Media
Avg_Driver_Age	63209	3.85 %	Me
Max_Driver_Age	63209	3.85 %	Me
Min_Driver_Age	63209	3.85 %	Me
Left_Hand_Drive	12512	0.76 %	Mo
Avg_Vehicle_Age	247923	15.11 %	Me
Weather_Conditions	161	0.01 %	Mo
Road_Surface_Conditions	2155	0.13 %	Mo

Como se verá más adelante en el análisis descriptivo bivariante, no se observan correlaciones altas entre la variable *Avg_Vehicle_Age* y otras variables que permitan construir un modelo de regresión lineal para imputar los valores perdidos. Por este motivo, se optará por imputarlos en esta variable utilizando la media.

Por simplicidad, al tener el resto de variables un porcentaje pequeño de missings, se ha optado por imputarlas directamente por la moda o la media.

4. Análisis Descriptivo

4.1. Análisis Descriptivo Univariante

En esta sección se procederá a hacer un análisis descriptivo univariante de las variables a predecir y algunas otras importantes.

Aunque en el apartado anterior se imputaron los missings para las variables con datos faltantes, se han utilizado para los análisis descriptivos únicamente los valores que no han sido imputados con el fin de ser lo más fieles posible a la información que se tenía.

4.1.1. Accident_Severity

El siguiente gráfico muestra la proporción de accidentes clasificados por su gravedad:

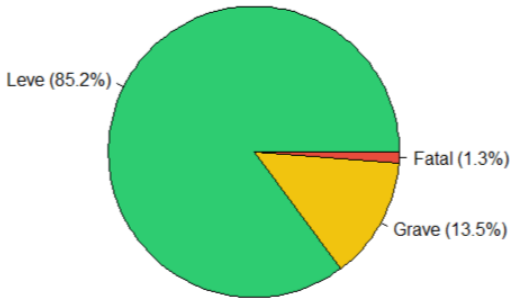


Figura 2: Gráfico de Sectores de Severidad del Accidente

En su mayoría, los accidentes se han categorizado como leves, teniendo solo un 1.3 % de ellos que son accidentes fatales. Al final, desde el punto de vista social y sanitario, son estos accidentes los más interesantes para el diseño de políticas de seguridad.

Aunque parece que podrían ser pocas observaciones, la ventaja de trabajar con una base de datos extensa es que este porcentaje se traduce en 21.327 observaciones aproximadamente. Aún así, esta fuerte desproporción entre clases habrá de tenerse en cuenta en la etapa de modelado predictivo, ya que podría sesgar los modelos hacia la clase mayoritaria.

4.1.2. Accident_Severity_Score

El siguiente gráfico muestra la proporción de accidentes que han tenido un rango de puntuación de gravedad en función del estado de las víctimas:

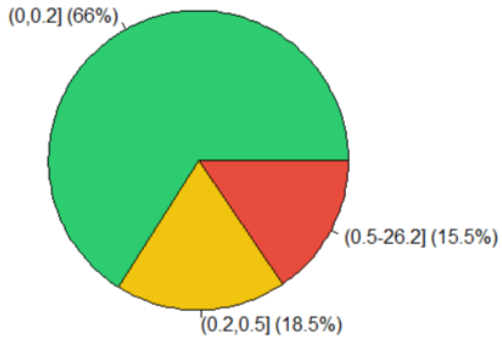


Figura 3: Gráfico de Sectores de Puntuación Severidad del Accidente

Al igual que en la variable categórica *Accident_Severity*, la mayor parte de las observaciones se concentran en los valores bajos de la variable continua *Accident_Severity_Score*. Más de la mitad de los accidentes presentan una puntuación inferior a 0.5, lo que implica que la gravedad de los accidentados ha sido leve (una puntuación de 0.1 indica un herido leve, 0.2 dos heridos leves, etc.). Una puntuación de 0.5 puede corresponder tanto a cinco heridos leves como a un herido moderado. Para facilitar la visualización, se ha ampliado la escala en el intervalo (0, 1], ya que la mayoría de las observaciones se acumulan en esta región. Los valores más altos, al igual que en el gráfico anterior, son mucho menos frecuentes.

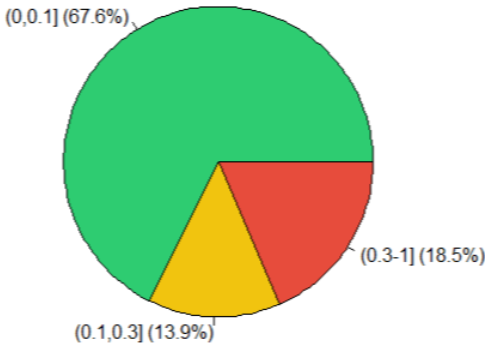


Figura 4: Gráfico de Sectores de Puntuación Severidad del Accidente Ampliación

Efectivamente, más de la mitad de los accidentes contienen solo a una persona cuyo estado ha sido considerado como leve. Esto confirma la tendencia observada en la variable categórica, reforzando la idea de que los accidentes graves o moderados son poco frecuentes dentro del conjunto de datos.

4.1.3. Number_of_Vehicles

En el siguiente gráfico se presenta el número de vehículos que se han visto implicados en el accidente:

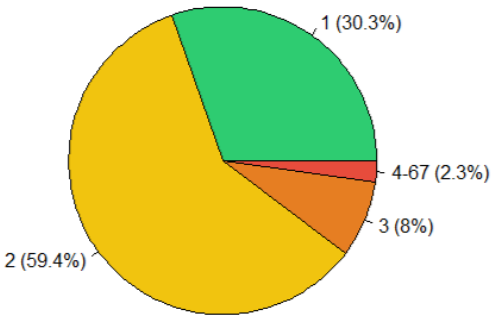


Figura 5: Gráfico de Sectores del Número de Vehículos Involucrados

El gráfico 5 muestra que en más del 89 % de los accidentes están involucrados solo uno o dos vehículos. Los accidentes con tres o más vehículos son muy pocos dentro del conjunto de datos. Esto indica que la mayoría de los casos analizados son situaciones simples, algo que puede influir en la gravedad del accidente y en cómo se analizan los datos más adelante. Además, al ser tan pocos los casos con muchos vehículos, no tendrán mucho peso en los modelos de predicción.

4.1.4. Number_of_Casualties

En el siguiente gráfico se presenta el número de personas, ya sean peatones, conductores o pasajeros, que han estado involucradas en el accidente:

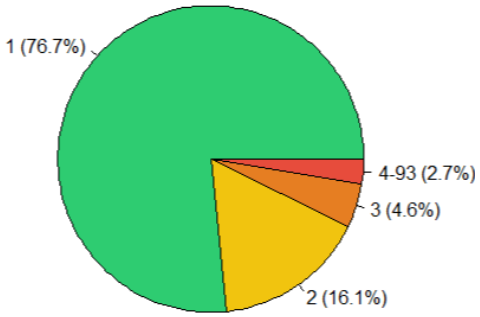


Figura 6: Gráfico de Sectores de Número de Personas Accidentadas

En el gráfico 6 se puede observar que en más de las tres cuartas partes de todos los accidentes solo hay una persona accidentada, que se asumirá que es el conductor. Los casos de dos o tres personas son un porcentaje menor y los de más de cuatro apenas son frecuentes. Este patrón indica que la mayoría de los siniestros involucran vehículos con pocos ocupantes, como turismos o motocicletas, y que, en general, no se trata de accidentes múltiples o de gran magnitud, como ya se había comprobado en el número de vehículos accidentados.

4.1.5. Male_Driver y Female_Driver

En los dos siguientes gráficos se presenta la distribución del número de conductores por accidente, en función del sexo.

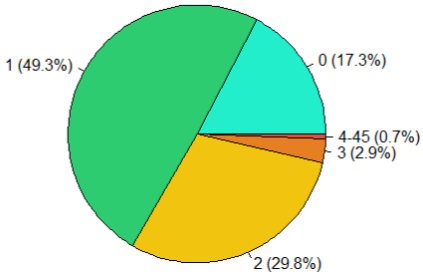


Figura 7: Gráfico de Sectores de Número de Conductores

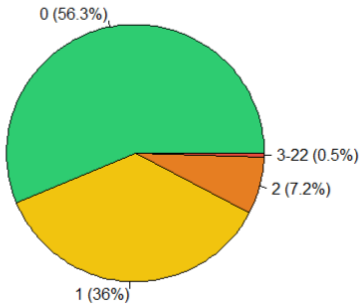


Figura 8: Gráfico de Sectores de Número de Conductoras

La diferencia más clara entre los dos gráficos está en el porcentaje de ceros. En el de conductores, el 0 aparece en algo más del 17 %, pero en el de conductoras supera el 56 %. Esto no quiere decir que el coche fuera sin conductor, sino que en la mayoría de esos casos no había ninguna mujer conduciendo, solo hombres. En general, hay muchas menos conductoras que conductores.

También se ve que entre los hombres es más común que haya uno o dos conductores, mientras que entre las mujeres, lo más habitual es que solo haya una o directamente ninguna. Las categorías con tres o más personas al volante son muy poco frecuentes en ambos casos, esperable por el número de vehículos.

4.1.6. Avg_Vehicle_Age

A continuación se presenta un gráfico sobre la edad media de todos los vehículos que se han visto involucrados en el accidente, en intervalos de 5 en 5 años:

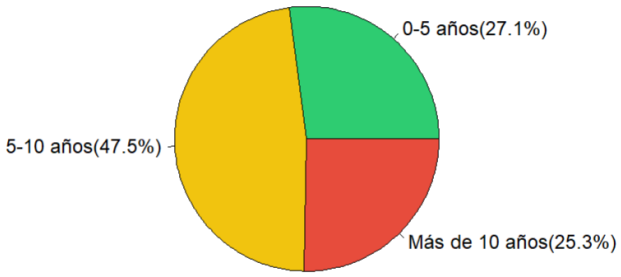


Figura 9: Gráfico de Sectores de Edad Media de los Vehículos

En el gráfico se observa que el grupo más numeroso corresponde a los vehículos con una antigüedad de entre 5 y 10 años, que representan el 47.5 % del total. A continuación, se encuentran los vehículos de entre 0 y 5 años (27.1 %), seguidos por los de más de 10 años (25.3 %). Esta distribución sugiere que los vehículos de edad entre 5 y 10 años son los que más se ven involucrados en los accidentes, aunque también existe una proporción considerable de vehículos muy nuevos y muy antiguos.

4.1.7. Avg_Driver_Age

A continuación se presenta un gráfico sobre la edad media de los conductores implicados en un accidente:

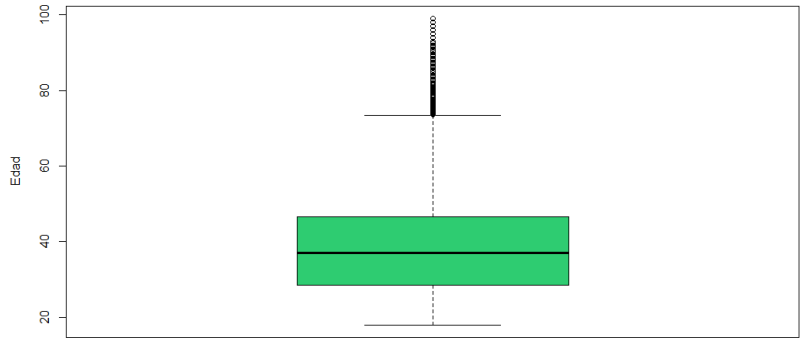


Figura 10: Diagrama de Caja de la Edad Media de los Conductores

El gráfico muestra que la mediana de edad de los conductores está en torno a los 40 años, y la mayoría se concentra entre los 30 y 45 años, lo que indica que los accidentes afectan sobre todo a adultos. Se observan valores atípicos por encima de los 70 e incluso algunos cercanos a los 100 años (posibles errores), y pocos casos por debajo de los 25. Esto refleja una alta variabilidad en la edad de los conductores implicados.

4.1.8. Avg_Casualty_Age

En el siguiente gráfico se presenta la distribución de la edad media de las víctimas:

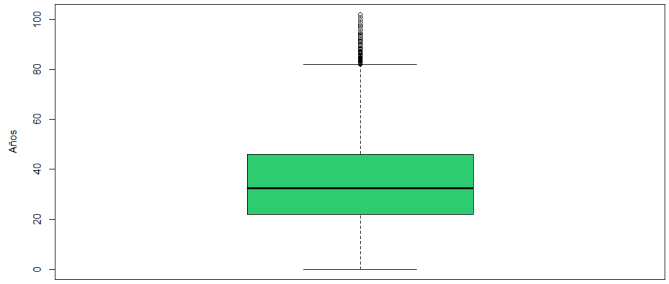


Figura 11: Diagrama de Caja de la Edad Media de los Accidentados

El gráfico muestra que la edad media de las víctimas en los accidentes se sitúa aproximadamente entre los 25 y los 45 años, con una mediana cercana a los 35 años. También se observa la presencia de algunos valores atípicos por encima de los 80 años, lo que indica que, aunque es menos común, también hay accidentes con víctimas de edad muy avanzada. En cambio, no se aprecian valores atípicos en edades muy bajas, lo que sugiere que los accidentes con niños como principales víctimas son frecuentes en el conjunto de datos.

Comparando los últimos dos gráficos, se observa que la edad media de los conductores tiende a ser algo más alta que la de las víctimas. Además, hay más valores extremos en el caso de los conductores, especialmente en edades avanzadas, lo que indica una mayor presencia de conductores mayores en comparación con víctimas de edad similar. En cambio, las víctimas tienden a concentrarse más en edades jóvenes-adultas.

4.2. Análisis Descriptivo Bivariante

En este apartado se estudiarán las relaciones entre pares de variables usando correlaciones. El objetivo es ver qué variables están más relacionadas entre sí y entender mejor cómo se conectan distintos aspectos de los accidentes, los vehículos y las personas implicadas. Para ver estas relaciones de forma visual, se ha creado un grafo donde cada nodo es una variable y las líneas indican la fuerza de la relación entre ellas.

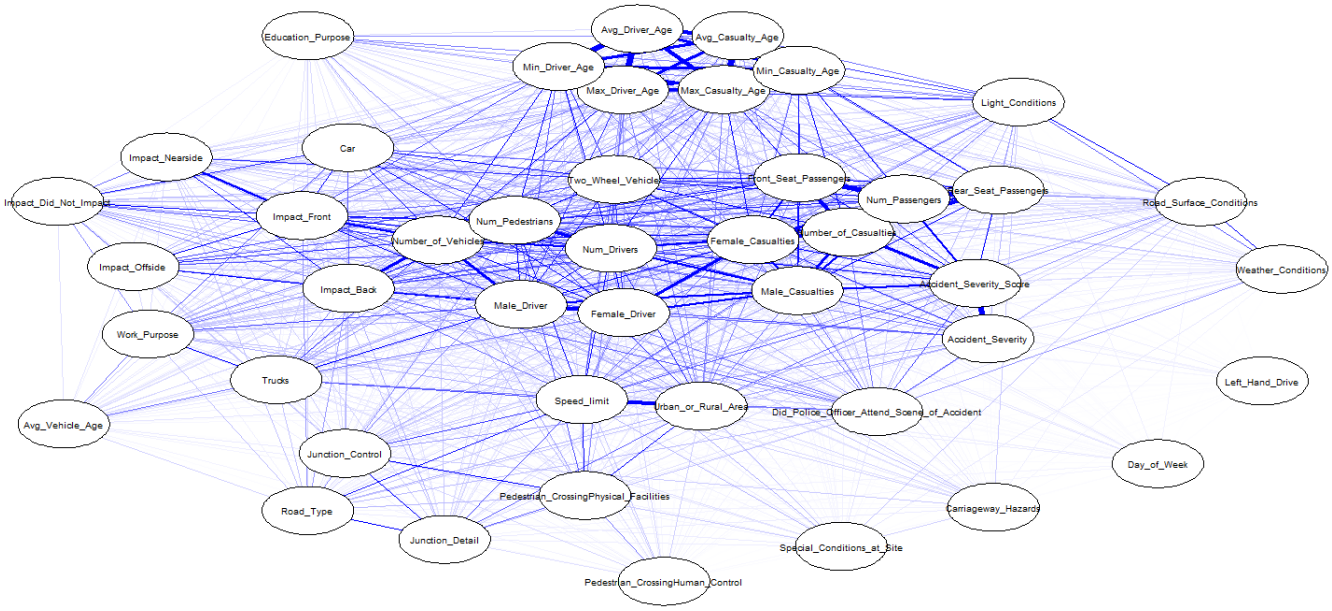


Figura 12: Grafo Correlaciones Bivariantes

Las correlaciones más destacadas se encuentran en la parte superior del gráfico, donde están las variables relacionadas con la edad. Esto es esperable, ya que se trata de medidas del mismo tipo (edad mínima, máxima o media), que aunque distintas en su definición, están claramente conectadas al hablar de una misma unidad: los conductores y accidentados.

También se observan algunas correlaciones individuales relevantes, como entre el número de impactos en el lado del conductor y los impactos frontales, lo cual tiene sentido en accidentes con choques directos. Asimismo, hay relaciones marcadas entre las variables que describen el número de pasajeros, ya sean totales, copilotos o pasajeros traseros, ya que están todas vinculadas al número de ocupantes del vehículo.

Otra relación lógica es la que aparece entre el límite de velocidad y si la zona es rural o urbana, ya que normalmente los límites son más bajos en zonas urbanas por la presencia de peatones y cruces.

Una asociación muy evidente es la que hay entre la severidad del accidente y la puntuación de severidad, ya que ambas variables miden lo mismo, pero con escalas distintas.

En el centro del gráfico se concentra un grupo de variables fuertemente conectadas entre sí, muchas de las cuales fueron generadas a partir de la base de datos de accidentados, por lo que comparten estructura y significado.

Por último, también es útil mencionar aquellas variables que apenas presentan correlación con el resto. Entre ellas destacan el día de la semana y si había algún conductor zurdo, entre otras.

4.3. Análisis Descriptivo Multivariante Clúster

Para el análisis multivariante, se aplicará la metodología clúster con el objetivo de agrupar los accidentes en distintos grupos según sus características. Se realizará el análisis para $k = 5$ y $k = 7$, y se incluirá el número de casos pertenecientes a cada clúster. Se han elegido estos dos números de grupos ya que son valores intermedios que permiten explorar distintas divisiones sin llegar a sobreajustar.

4.3.1. Marco Teórico

El análisis de clúster es una técnica que permite agrupar observaciones en función de sus similitudes, sin necesidad de tener previamente una variable objetivo. Es un método de aprendizaje no supervisado que busca identificar patrones en los datos, organizando las observaciones en grupos (clústeres) homogéneos internamente y heterogéneos entre los conjuntos [8].

Existen distintos tipos de métodos de clúster, como los jerárquicos, que generan una estructura en forma de árbol; y los particionales, entre los que destaca el algoritmo k-medias. Este último es uno de los más utilizados, ya que es bastante sencillo. Consiste en asignar cada observación al grupo con el centroide más cercano, repitiendo el proceso hasta que los grupos dejan de cambiar. El centroide es el punto que representa el "centro" del grupo, calculado como el promedio de todas las observaciones asignadas a ese clúster.

El número de clústeres k tiene que ser fijado antes, por lo que habitualmente se utilizan criterios como el método del codo o la interpretación sustantiva para seleccionar un número adecuado de grupos. En este análisis se utilizará el algoritmo k-medias con dos valores diferentes, 5 y 7.

4.3.2. Desarrollo del Modelo

Se analizarán las similitudes dentro de los clústeres (centroides) y se creará un ejemplo de árbol explicativo para el clúster $k = 5$ y otro para el $k = 7$.

4.3.3. Clúster k=5

En este apartado se analizará la metodología clúster con $k = 5$ medias, incluyendo el número de observaciones que han caído en cada clúster. Esto último queda reflejado en la siguiente tabla:

Tabla 7: Número de casos en cada clúster

Clúster	Casos
1	230226.000
2	501128.000
3	225564.000
4	246159.000
5	437520.000

Los clúster que se han llevado más de la mitad de la muestra han sido los clústers 2 y 4, sumando casi un millón de observaciones. Y el resto, alrededor de 12 % cada uno.

Tabla 8: Centros de Clústeres Finales $k = 5$

	Clúster				
Variable	**1**	**2**	**3**	**4**	**5**
Male_Driver	1	1	1	1	1
Female_Driver	1	0	0	0	1
Avg_Driver_Age	40.583	24.749	56.053	43.196	40.827
Max_Driver_Age	47.77	27.31	62.93	52.29	46.44
Min_Driver_Age	33.56	22.22	49.27	34.29	35.27
Left_Hand_Drive	1	1	1	1	1
Car	1	1	1	1	1
Two_Wheel_Vehicle	0	0	0	0	0
Trucks	0	0	0	0	0
Impact_Front	1	1	1	1	1
Impact_Back	0	0	0	0	0
Impact_Offside	0	0	0	0	0
Impact_Nearside	0	0	0	0	0
Work_Purpose	0	0	0	0	0
Education_Purpose	0	0	0	0	0
Avg_Vehicle_Age	6.97	7.50	7.24	7.20	7.21
Accident_Severity	3	3	3	3	3
Number_of_Vehicles	2	2	2	2	2
Number_of_Casualties	2	1	1	1	1
Speed_limit	63	38	39	33	31
Light_Conditions	1	1	1	1	1
Weather_Conditions	2	2	2	2	2
Road_Surface_Conditions	1	1	1	1	1
Num_Passengers	0	0	0	0	0
Num_Pedestrians	0	0	0	0	0
Male_Casualties	1	1	1	1	1
Female_Casualties	1	1	1	1	1
Avg_Casualty_Age	38.81	22.11	66.24	18.67	42.22
Max_Casualty_Age	43.26	23.25	67.94	20.04	44.91
Min_Casualty_Age	34.55	21.06	64.46	17.41	39.63
Accident_Severity_Score	0.27	0.21	0.22	0.18	0.19
Front_Seat_Passengers	0	0	0	0	0
Rear_Seat_Passengers	0	0	0	0	0

Lo primero que hay que determinar es qué son los clústers finales. Las dos tablas muestran los valores medios (centroides) de cada variable dentro de los cinco grupos. Cada clúster representa un conjunto de observaciones con patrones similares, agrupadas automáticamente según sus características.

Los cinco clústeres obtenidos agrupan perfiles diferentes en función de la edad de los conductores, las víctimas, la gravedad del accidente y otras variables asociadas:

- **Clúster 1:** conductores de edad media (40.6 años), ambos sexos y mayor número de víctimas por

accidente (2). Destaca por su mayor severidad (0.27) y un límite de velocidad elevado, lo que podría estar relacionado con la mayor gravedad de los siniestros.

- **Clúster 2:** conductores muy jóvenes (24.7 años) y víctimas también de baja edad (22.1 años). Esto podría reflejar accidentes leves protagonizados por jóvenes conductores sin acompañantes.
- **Clúster 3:** conductores de mayor edad (56 años), con víctimas también mayores (66.2 años).
- **Clúster 4:** las víctimas son las más jóvenes, lo que indica pasajeros (o peatones) menores de edad o jóvenes adultos.
- **Clúster 5:** conductores adultos (40.8 años) y víctimas también en torno a los 42 años. No destaca especialmente en ninguna variable, lo que sugiere que podría representar accidentes más comunes de la base.

Como ya se ha mencionado, los clústers presentan diferencias en variables como la edad de los conductores y de las víctimas, pero hay en algunas de ellas en las que estos grupos son muy homogéneos (en cuanto al centroide).

En primer lugar, en la mayoría de los accidentes hubo más de un coche involucrado, ya que vemos que en la variable *Number_of_Vehicles* todos los centroides son iguales a 2. Además, si unimos esta información con el tipo que fue el primer impacto, observamos que todos los centroides en la variable *Impact_Front* son 1, lo que puede interpretarse como que lo más común en los accidentes es un choque entre dos coches (fijándonos en la variable *Car*) con un primer impacto frontal en ambos.

Respecto a los conductores, como mínimo hay uno, pero no en todos ha habido una conductora. El número de accidentados medios por accidente es 1, excepto en el clúster 2, que hay dos accidentados. Aunque hay una pequeña diferencia en los grupos, lo notable es que este número no pasa de 2 o 3 personas, por lo que estamos ante accidentes con pocas personas involucradas.

Para analizar las similitudes en cuanto a las condiciones de la carretera, lumínicas y meteorológicas, hay que definir antes las distintas categorías para cada variable:

Tabla 9: Categorías Variables

Categoría	Road_Surface_Conditions	Weather_Conditions	Light_Conditions
1	Seco	Despejado sin viento fuerte	Luz diurna
2	Húmedo / Mojado	Lluvia sin viento fuerte	Oscuridad y luces encendidas
3	Nieve	Nieve sin viento fuerte	Oscuridad y luces apagadas
4	Escarcha / Hielo	Despejado con viento fuerte	Oscuridad sin iluminación
5	Inundación >3 cm	Lluvia con viento fuerte	Oscuridad e iluminación desconocida
6	Aceite / Gasolina	Nieve con viento fuerte	_____
7	Barro	Niebla	_____
8	_____	Otro	_____
9	_____	Desconocido	_____

Nos encontramos con que, en su mayoría, hablamos de una calzada seca (sin lluvia o hielo), con lluvia pero sin viento fuerte y de día.

No se encuentran entre los accidentados pasajeros o peatones, y en función del sexo, uno de cada. Además, si nos fijamos en la variable *Accident_Severity*, todos los clúster muestran como centroides una gravedad leve del accidente. Esto no es algo de extrañar si se tiene en cuenta que en el gráfico 2 se observaba un 85.2 % de accidentes leves.

La variable *Accident_Severity_Score* también concuerda con los gráficos 3 y 4, donde se veía que un 66 % de los accidentes caían en el intervalo (0,0.2]. Todos los centroides de esta variable caen cerca del 0.2, excepto el primer clúster que se aleja más de esta cifra.

Árbol Explicativo Clúster k=5

A continuación se presenta un árbol explicativo del clúster con una posible estructura, hecha a partir de los clúster, que explique la división de los grupos:

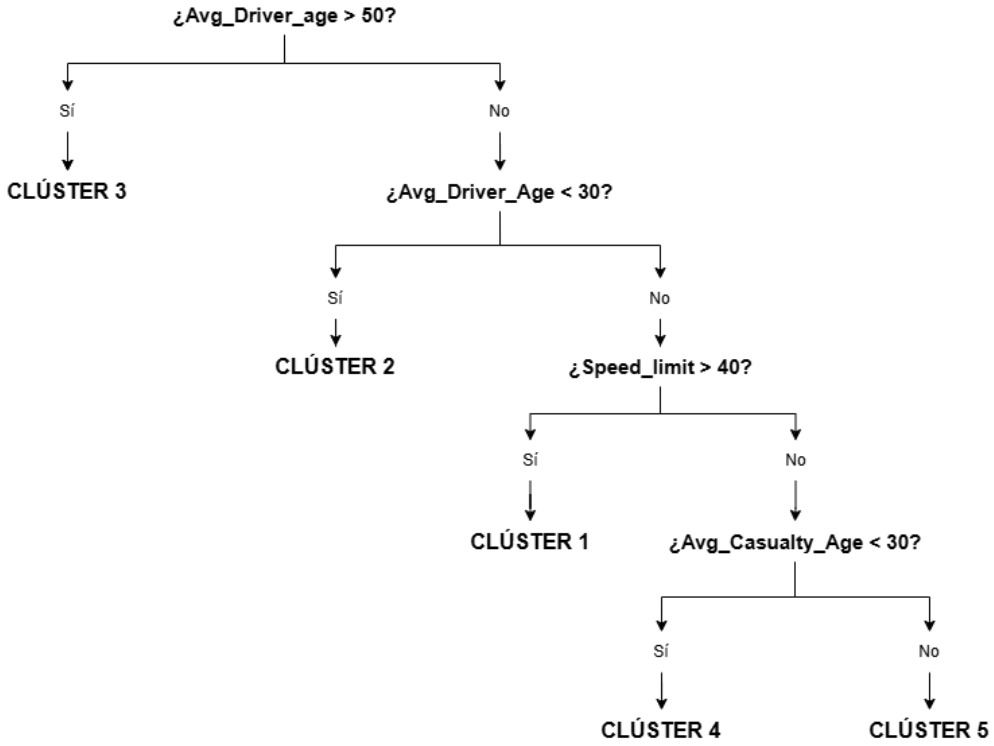


Figura 13: Árbol de Decisión del Clúster k=5

El árbol muestra que los accidentes con conductores mayores de 50 años van al clúster 3. Si los conductores son más jóvenes, la edad del accidentado y la velocidad ayudan a distinguir el resto. Los conductores más jóvenes pertenecen generalmente al clúster 2. Si están entre los 30 y los 50 y la velocidad de la vía era mayor a 40 mph, serán clasificados en el clúster 1. Para distinguir entre el clúster 4 y el 5, la principal diferencia será la edad media del accidentado; si es menor de 30, irá al 4 y sino, al 5.

4.3.4. Clúster k=7

Para ampliar la información de los clúster realizados en el apartado anterior, se ha ampliado el número permitido de clúster a $k = 7$. El número de observaciones por grupo es el siguiente:

Tabla 10: Número de casos en cada clúster

Clúster	Casos
1	233322.000
2	6015.000
3	218984.000
4	113087.000
5	204103.000
6	391663.000
7	473423.000

Esta división también se ha realizado de manera poco equitativa, ya que hay un clúster que tiene casi medio millón de observaciones y otro que no llega a diez mil.

Tabla 11: Centros de Clústeres Finales $k = 7$

	Clúster						
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Male_Driver	1	1	1	1	1	1	1
Female_Driver	0	1	1	0	1	1	0
Avg_Driver_Age	43.141	46.672	45.861	64.097	43.639	37.765	24.395
Max_Driver_Age	52.41	58.35	52.46	70.43	52.62	42.26	26.91
Min_Driver_Age	34.05	35.56	39.30	57.87	34.89	33.30	21.91
Left_Hand_Drive	1	1	1	1	1	1	1
Car	1	1	1	1	1	1	1
Two_Wheel_Vehicle	0	0	0	0	0	0	0
Trucks	0	0	0	0	0	0	0
Impact_Front	1	1	1	1	1	1	1
Impact_Back	0	1	0	0	1	0	0
Impact_Offside	0	0	0	0	0	0	0
Impact_Nearside	0	0	0	0	0	0	0
Work_Purpose	0	0	0	0	0	0	0
Education_Purpose	0	0	0	0	0	0	0
Avg_Vehicle_Age	7.20	7.40	7.19	7.33	6.97	7.20	7.50
Accident_Severity	3	3	3	3	3	3	3
Number_of_Vehicles	2	2	2	2	2	2	2
Number_of_Casualties	1	5	1	1	2	1	1
Speed_limit	33	38	35	41	60	32	40
Light_Conditions	1	1	1	1	1	1	1
Weather_Conditions	2	1	2	2	2	2	2
Road_Surface_Conditions	1	1	1	1	1	1	1
Num_Passengers	0	3	0	0	0	0	0
Num_Pedestrians	0	0	0	0	0	0	0
Male_Casualties	1	2	1	1	1	1	1
Female_Casualties	1	3	1	1	1	1	1
Avg_Casualty_Age	17.73	34.31	57.33	69.95	40.63	37.57	21.56

Tabla 12: Centros de Clústeres Finales $k = 7$ (continuación)

	Clúster						
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Max_Casualty_Age	18.80	67.33	58.53	71.54	46.97	39.58	22.66
Min_Casualty_Age	16.75	7.10	56.11	68.28	34.44	35.65	20.53
Accident_Severity_Score	0.18	0.73	0.19	0.23	0.29	0.18	0.21
Front_Seat_Passengers	0	1	0	0	0	0	0
Rear_Seat_Passengers	0	1	0	0	0	0	0

Los perfiles generales que definen cada clúster son los siguientes:

- **Clúster 1:** conductores adultos (43.1 años), sin acompañantes, con accidentados muy jóvenes (17.7 años). La puntuación de gravedad es de las más bajas, de apenas dos accidentados leves (0.18).
- **Clúster 2:** es el más pequeño del conjunto, con solo 6.000 casos. Destaca por tener el mayor número de víctimas (5 por accidente) y la puntuación de severidad más alta (0.73). Además, presenta la única combinación con pasajeros tanto delante como detrás, y con edades de víctimas muy variables (de 7 a 67 años). Puede tratarse de accidentes más graves, con vehículos ocupados por familias.
- **Clúster 3:** perfil de adultos de mediana edad (45.8 años) con víctimas mayores (57.3 años). No destaca en severidad ni velocidad, y tampoco presenta acompañantes. Puede reflejar siniestros de baja intensidad pero con población de riesgo.
- **Clúster 4:** agrupa conductores de edad muy avanzada (64.1 años) y víctimas también mayores (69.9 años). Tiene la velocidad más alta (41 mph entre los moderados) y una severidad moderada (0.23). Representa con claridad un grupo de personas mayores en trayectos comunes.
- **Clúster 5:** conductores adultos (43.6 años), con 2 víctimas por accidente. Tiene una velocidad notablemente más alta (60 mph), lo que puede estar relacionado con su puntuación de severidad algo elevada (0.29). Se trata posiblemente de un grupo donde la velocidad ha influido en la gravedad.
- **Clúster 6:** conductores algo más jóvenes (37.7 años), con víctimas adultas (edad media: 37.6) y sin pasajeros. Severidad baja (0.18) y velocidad algo inferior (32 mph). Puede representar accidentes de media intensidad entre adultos.
- **Clúster 7:** el grupo más numeroso. Conductores muy jóvenes (24.4 años) y víctimas también jóvenes (21.5). Aunque tiene baja severidad (0.21), es destacable por el tamaño del grupo y por representar probablemente a la mayoría de siniestros leves entre jóvenes.

Pese a las diferencias en edad, gravedad y número de víctimas, hay variables que se mantienen constantes en todos los clústeres. Al igual que en la clasificación con $k = 5$, todos los centroides incluyen la presencia de un conductor. Sin embargo, ha aumentado el número de clústeres con conductoras: en la tabla 8 solo había dos clústeres con alguna conductora, mientras que ahora 4 de los 7 clústeres presentan al menos una.

Además, como en el apartado anterior, se observa que en todos los clústeres está presente al menos un conductor zurdo. De nuevo, los coches son los únicos vehículos que aparecen como presentes.

En cuanto a los impactos, aunque en todos los clústeres se mantiene el impacto frontal, en dos de ellos (clústeres 2 y 5) también aparece un impacto trasero. Si se interpretan ambos impactos en conjunto, puede decirse que estos dos clústeres corresponden a accidentes de tipo alcance trasero.

En lo relativo a las condiciones lumínicas y de la carretera, se mantiene la categoría 1 en todos los clústeres para ambos, que corresponde con luz diurna y calzada seca. En el clúster 2 para las condiciones meteorológicas sí que vemos un cambio, porque en este se describe un tiempo despejado sin viento fuerte y en el resto lluvia sin viento fuerte.

Lo más destacable es que el clúster más pequeño resulta ser el más grave, mientras que el más numeroso agrupa los siniestros más leves, lo que puede estar relacionado con la frecuencia esperada de los distintos tipos de accidentes.

Árbol Explicativo Clúster k=7

A continuación se presenta un árbol explicativo del clúster con una posible estructura, hecha a partir de los clústers, que explica la división de los siete grupos:

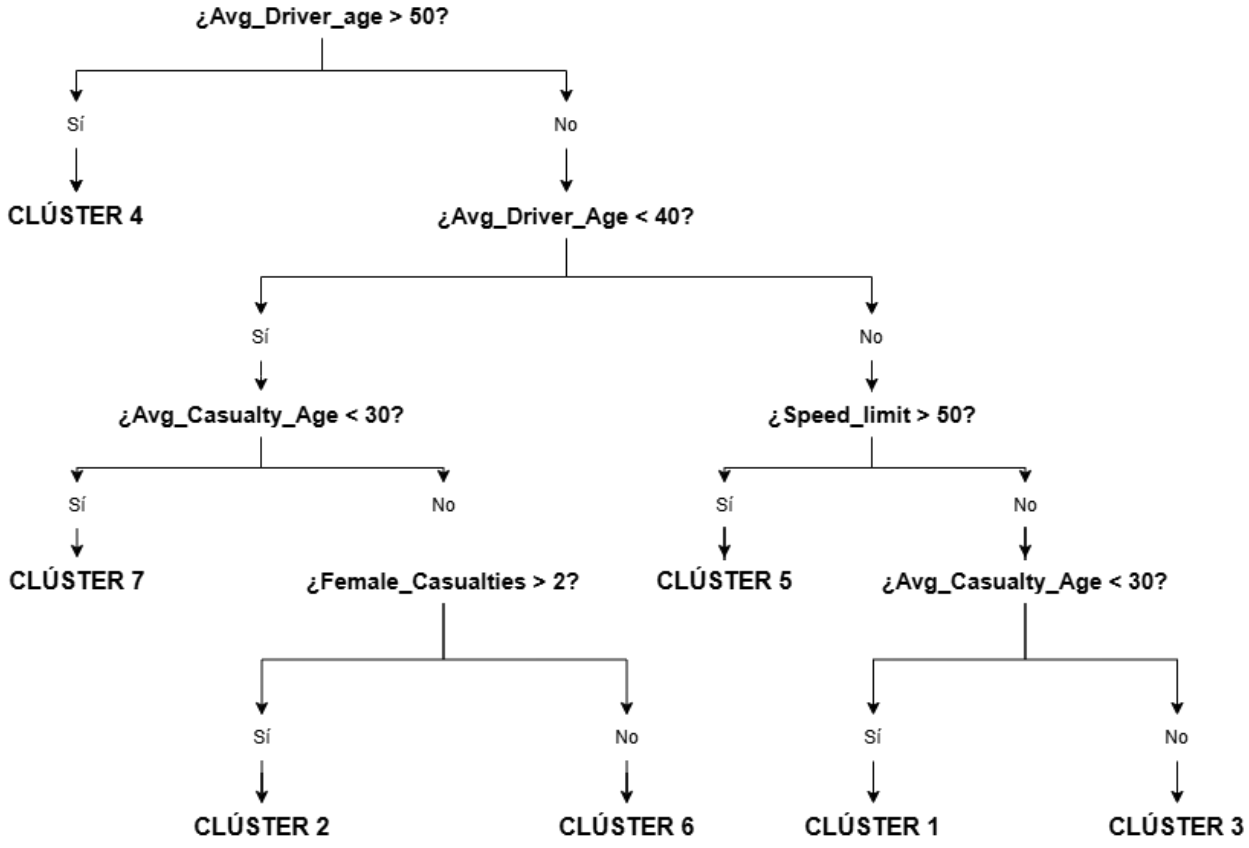


Figura 14: Árbol de Decisión del Clúster k=7

El árbol muestra que los accidentes con conductores mayores de 50 años van al clúster 4, como pasaba en

el árbol de decisión para clúster con $k = 5$ en el gráfico 13. Si los conductores son más jóvenes, la edad del accidentado y la velocidad ayudan a distinguir algunos otros grupos. Los conductores más jóvenes pertenecen generalmente al clúster 7, si tienen menos de 30 años. Si no, habrá que hacer una distinción por el sexo de los accidentados. Si hay más de dos mujeres involucradas, esa observación caerá en el clúster 2 y si no, en el 6.

Si la edad media de los conductores está entre los 30 y 40 años y la velocidad máxima de la vía era mayor a 50, la observación pertenecerá al clúster 5, y si no se clasificará en función de la edad media de los accidentados otra vez. Si son menores de 30 años, al clúster 1, sino al 3.

5. Técnicas Predictivas

En este apartado se han aplicado dos modelos distintos para cada una de las variables dependientes, con el fin de comparar su capacidad predictiva. En el caso de la variable categórica, se ha utilizado una regresión logística ordinal y un árbol de clasificación. Para la variable continua, se ha empleado una regresión lineal múltiple y un árbol de regresión.

Algunas de las variables utilizadas en los modelos predictivos no coinciden con las empleadas en el análisis clúster, y viceversa. Esta diferencia se debe a que ciertas variables resultaban poco interpretables o no aportaban una información clara, como es el caso del clúster, mientras que sí eran útiles y comprensibles en los modelos de regresión y clasificación. Por ello, la selección de variables se adaptó para los descriptivos multivariantes y para las técnicas predictivas.

5.1. Regresión Logística Ordinal

5.1.1. Marco Teórico

La regresión logística ordinal es una técnica que se utiliza cuando la variable que se quiere predecir es cualitativa y sus categorías siguen un orden lógico. En este caso, se ha elegido este modelo porque la gravedad de los accidentes se clasifica como leve, moderada o grave, lo que implica un nivel creciente de severidad [9].

Para poder incluir variables cualitativas como predictoras, es necesario transformarlas en variables numéricas. En concreto, se crean variables *dummy*, es decir, una columna por cada categoría menos una, que actúa como referencia.

Los parámetros del modelo se estiman mediante máxima verosimilitud, y permiten interpretar el efecto de cada variable sobre la probabilidad de que el accidente se clasifique en una categoría más leve. En particular, se utilizan los *odds ratio acumulados*, que indican cómo cambian esas probabilidades acumuladas cuando una variable aumenta en una unidad, manteniendo el resto constantes.

Una ventaja de este tipo de modelo es que introduce el uso del kappa ponderado, una medida que penaliza más los errores cuanto mayor sea la diferencia entre la categoría predicha y la real. Por ejemplo, clasificar un accidente leve como grave es peor que hacerlo como moderado [9], [10]. Además, para tener una visión completa del rendimiento de los modelos se combinarán distintas métricas:

- **Accuracy**: es la proporción de todas las predicciones correctas.
- **Sensibilidad**: es la fracción de casos reales de cada categoría clasificados correctamente.
- **Especificidad**: es la fracción de casos que no pertenecen a una categoría y que el modelo clasifica correctamente como no pertenecientes.
- **Kappa no ponderado**: indica cuánto coinciden las predicciones y la realidad sin tener en cuenta las coincidencias que han sido por pura aleatoriedad
- **Kappa ponderado**: es parecido al kappa anterior solo que penaliza la distancia entre lo predicho y lo real (como en el ejemplo del párrafo anterior).

Las tres primeras métricas muestran el rendimiento general del modelo, y los kappa miden el nivel de coincidencia real descontando el azar, penalizando o no la distancia entre real y predicción. [9], [10]

5.1.2. Desarrollo del Modelo

Para aplicar esta técnica, lo primero fue transformar las variables categóricas en factores, para facilitar su uso en el modelo. Después, se construyó un modelo inicial dividiendo el conjunto de datos en dos partes: un 10 % para entrenamiento (train) y un 90 % para prueba (test). El motivo de usar esta división, y no la más común de 70 % y 30 %, fue por una limitación computacional porque, en el entrenamiento del modelo, no admitía más de un 10 % de la muestra.

Una vez ajustado el primer modelo, se aplicó un análisis ANOVA de tipo II para detectar qué variables no eran estadísticamente significativas. Estas variables se eliminaron y, con las restantes, se construyó un segundo modelo más simplificado. Para asegurar que solo permanecían las variables relevantes, se volvió a realizar el ANOVA sobre este nuevo modelo.

Finalmente, se evaluó el rendimiento del modelo calculando la sensibilidad, especificidad, precisión (*Accuracy*), coeficiente Kappa y Kappa ponderado, tanto para el conjunto de entrenamiento como para el de prueba.

5.1.3. Resultados y Evaluación del Modelo

El primer modelo se construyó a partir del conjunto de entrenamiento con todas las variables disponibles, en total 30, siendo una de ellas la variable dependiente *Accident_Severity*. De este modelo inicial se obtuvieron 58 parámetros, de los cuales 12 no resultaron ser estadísticamente significativos.

Para evaluar la relevancia individual de cada variable, se aplicó un análisis de varianza (ANOVA) de tipo II, cuyos resultados se muestran en la siguiente tabla:

Tabla 13: ANOVA Tipo II Modelo Inicial

Variable	χ^2	**G.L.**	**p-valor**
Male_Driver	30.1	1	4.05e-08 ***
Female_Driver	5.0	1	0.0249 *
Avg_Driver_Age	7.6	1	0.0060 **
Car	73.5	1	<2.2e-16 ***
Two_Wheel_Vehicle	3634.5	1	<2.2e-16 ***
Trucks	113.1	1	<2.2e-16 ***
Impact_Did_Not_Impact	2.0	1	0.1584
Impact_Front	2.0	1	0.1599
Impact_Back	0.0	1	0.8885
Impact_Offside	0.9	1	0.3360
Impact_Nearside	0.6	1	0.4252
Work_Purpose	0.5	1	0.4678
Education_Purpose	0.7	1	0.3742
Avg_Vehicle_Age	16.6	1	4.50e-05 ***
Number_of_Vehicles	3.6	1	0.0587 .
Number_of_Casualties	0.0	1	0.9127

Tabla 14: ANOVA Tipo II Modelo Inicial (continuación)

Variable	χ^2	**G.L.**	**p-valor**
Day_of_Week	128.8	6	<2.2e-16 ***
Road_Type	284.5	5	<2.2e-16 ***
Speed_limit	599.4	7	<2.2e-16 ***
Light_Conditions	585.1	4	<2.2e-16 ***
Weather_Conditions	71.3	8	2.71e-12 ***
Road_Surface_Conditions	20.2	4	0.00045 ***
Urban_or_Rural_Area	198.6	2	<2.2e-16 ***
Num_Drivers	222.7	1	<2.2e-16 ***
Num_Pedestrians	5824.3	1	<2.2e-16 ***
Male_Casualties	0.3	1	0.5634
Female_Casualties	0.0	1	0.9381
Avg_Casualty_Age	566.9	1	<2.2e-16 ***
Front_Seat_Passengers	83.5	1	<2.2e-16 ***
Rear_Seat_Passengers	72.7	1	<2.2e-16 ***

A partir de estos resultados, se identificaron aquellas variables que no alcanzaban un nivel de significación estadística (p-valor >0.05), es decir, aquellas sin asteriscos o con un punto. Por tanto, se decidió eliminar las siguientes variables del modelo por no aportar suficiente información relevante a la predicción de la gravedad del accidente:

- *Impact_Did_Not_Impact*
- *Impact_Front*
- *Impact_Back*
- *Impact_Offside*
- *Impact_Nearside*
- *Work_Purpose*
- *Education_Purpose*
- *Number_of_Vehicles*
- *Number_of_Casualties*
- *Female_Casualties*
- *Male_Casualties*

Con las variables que resultaron significativas en el análisis ANOVA, se construyó un segundo modelo más sencillo. En él solo se incluyeron aquellas variables que realmente aportaban información útil. Esto permite que el modelo sea más fácil de interpretar y evita añadir variables que no mejoran las predicciones, pero que sí aumentan su complejidad.

Por motivos de espacio, no se ha incluido la fórmula completa del modelo, ya que contiene muchas variables *dummy* y ocuparía demasiado. Lo que sí se puede mostrar son los estimadores de los parámetros y los puntos de corte que se utilizan para asignar una observación a una categoría u otra. Los parámetros son los siguientes:

Tabla 15: Coeficientes del Modelo

Variable	**Coeficiente**	**Variable**	**Coeficiente**
Male_Driver	0.1771	Female_Driver	0.4166
Avg_Driver_Age	-0.0025	Car1	0.2462
Two_Wheel_Vehicle1	-1.2705	Trucks1	-0.1968
Avg_Vehicle_Age	-0.0090	Day_of_Week2	0.2161
Day_of_Week3	0.2310	Day_of_Week4	0.2352
Day_of_Week5	0.2115	Day_of_Week6	0.1526
Day_of_Week7	0.0345	Road_Type2	-0.2236
Road_Type3	-0.4406	Road_Type4	-0.5676
Road_Type5	0.0181	Road_Type6	-0.2118
Speed_limit15	11.3219	Speed_limit20	2.4729
Speed_limit30	2.4380	Speed_limit40	2.0782
Speed_limit50	1.9815	Speed_limit60	1.7753
Speed_limit70	1.9888	Light_Conditions4	-0.4033
Light_Conditions5	-0.3258	Light_Conditions6	-0.5360
Light_Conditions7	-0.3208	Weather_Conditions2	0.1763
Weather_Conditions3	0.3430	Weather_Conditions4	-0.0145
Weather_Conditions5	0.1982	Weather_Conditions6	0.2649
Weather_Conditions7	0.2344	Weather_Conditions8	0.2404
Weather_Conditions9	0.2637	Road_Surface_Conditions2	-0.0289
Road_Surface_Conditions3	0.0319	Road_Surface_Conditions4	0.2030
Road_Surface_Conditions5	-0.1784	Urban_or_Rural_Area2	-0.3220
Urban_or_Rural_Area3	-1.8991	Num_Drivers	-0.5964
Num_Pedestrians	-1.6482	Avg_Casualty_Age	-0.0111
Front_Seat_Passengers	-0.3640	Rear_Seat_Passengers	-0.3695

Los puntos de corte se basan en el valor de η, que es la suma de todas las variables del modelo multiplicadas por sus respectivos coeficientes β_k.

$$\text{Accident\_Severity} = \begin{cases} \text{Leve} & \text{si } \eta < -4{,}074 \\ \text{Moderado} & \text{si } -4{,}074 \leq \eta < -1{,}360 \\ \text{Grave} & \text{si } \eta \geq -1{,}360 \end{cases}$$

En total se estimaron 48 parámetros, de los cuales 8 no fueron estadísticamente significativos. Al aplicar de nuevo el análisis ANOVA de tipo II al modelo reducido, se comprobó que todas las variables incluidas eran ahora significativas (p-valor < 0.05). Esto confirma que se eliminaron correctamente las variables que no aportaban información útil al modelo.

Para interpretar mejor los coeficientes obtenidos, se calcularon los *odds ratios acumulados*. Estos valores indican cómo afecta a la probabilidad de que un accidente sea más o menos grave cuando una variable aumenta una unidad, manteniendo las demás constantes. En la siguiente tabla se muestran los *odds ratios* de los parámetros del modelo que son significativos e interpretables:

Tabla 16: Odds Ratios Acumulados Modelo Logístico Ordinal

Variable	Odds Ratio	Variable	Odds Ratio
Avg_Casualty_Age	0.989	Avg_Driver_Age	0.998
Avg_Vehicle_Age	0.991	Car1	1.279
Day_of_Week2	1.241	Day_of_Week3	1.260
Day_of_Week4	1.265	Day_of_Week5	1.235
Day_of_Week6	1.165	Female_Driver	1.517
Front_Seat_Passengers	0.695	Light_Conditions2	0.668
Light_Conditions3	0.722	Light_Conditions4	0.585
Male_Driver	1.194	Num_Drivers	0.551
Num_Pedestrians	0.192	Rear_Seat_Passengers	0.691
Road_Type2	0.800	Road_Type3	0.644
Road_Type4	0.567	Road_Surface_Conditions4	1.225
Speed_limit15	19.954	Speed_limit20	11.857
Speed_limit30	11.450	Speed_limit40	7.990
Speed_limit50	7.254	Speed_limit60	5.902
Speed_limit70	7.307	Trucks1	0.821
Two_Wheel_Vehicle1	0.281	Urban_or_Rural_Area2	0.725
Weather_Conditions2	1.193	Weather_Conditions3	1.409
Weather_Conditions5	1.219	Weather_Conditions7	1.264

Antes de interpretar algunos *odds ratios*, es importante recordar que en este modelo, la categoría 1 corresponde al accidente más grave y el 3 al menos grave. Algunas de las interpretaciones son:

- **Avg_Casualty_Age (OR = 0.989):** por cada año más que tengan en media las personas accidentadas, la probabilidad de que el accidente sea leve disminuye un 1.1 %. Es decir, los accidentes con víctimas de mayor edad tienden a ser algo más graves que los que afectan a personas jóvenes.
- **Day_of_Week (2, 3, 4, 5, 6):** cuando un accidente ocurre entre semana (correspondientes a esos números), la probabilidad de que sea leve aumenta en comparación con el domingo (*Day_of_Week1*), que es la referencia. Una posible hipótesis sobre esto es que en fines de semana, la gente se suele desplazar más, en grupos más grandes o que sale de fiesta lo que hace que aumente la gravedad del accidente.
- **Light_Conditions4 (OR = 0.585):** que el accidente se de por la noche sin alumbrado público hace que la probabilidad de que el accidente sea leve disminuye en algo más de un 40 %.
- **Two_Wheel_Vehicle1 (OR = 0.281):** la presencia de una moto, por ejemplo, hace que sea algo más de un 70 % menos probable que el accidente sea leve, o en otras palabras, que sea más grave.

Unos *odds ratios* que llaman especialmente la atención son los de los parámetros de la variable *Speed_limit*, que toma como referencia el límite de 10 mph. Aunque podría parecer que velocidades tan bajas reducen la gravedad de los accidentes, el modelo sugiere lo contrario. Para visualizar mejor las conversiones de millas por hora a kilómetros por hora se incluye la siguiente tabla:

Tabla 17: Conversión de Límites de Velocidad de mph a km/h

Límite (mph)	km/h	Límite (mph)	km/h
10	16.09	40	64.37
15	24.14	50	80.47
20	32.19	60	96.56
30	48.28	70	112.65

Desde una interpretación subjetiva, podría suponerse que en tramos con límites de velocidad tan bajos los conductores no prestan atención a los límites y circulan bastante más rápido. Estos máximos suelen encontrarse en zonas como aparcamientos, zonas escolares o calles residenciales, donde, aunque la velocidad es reducida, los accidentes tienden a implicar a peatones o ciclistas, que al no tener la protección del vehículo pueden acarrear peores consecuencias, elevando la probabilidad de que el accidente sea moderado o grave.

Por otro lado, límites superiores como 70 mph suelen ser en vías amplias, con varios carriles, buena visibilidad y sin peatones.

En general, los accidentes tienden a ser más graves cuando hay peatones, motoristas, personas mayores o más de un pasajero. También influyen negativamente las malas condiciones de luz (por ejemplo, de noche sin alumbrado) y el mal tiempo. En cambio, los accidentes que ocurren en carreteras amplias, con buena visibilidad y límites de velocidad altos, suelen ser menos graves, probablemente porque implican solo vehículos y no personas vulnerables. Además, los días laborables se asocian con menor gravedad que los fines de semana.

Por último, se evaluó el rendimiento del modelo tanto en el conjunto de entrenamiento como en el conjunto de prueba. Los resultados se presentan a continuación:

Tabla 18: Matrices de Confusión para Train (Izda) y Test (Dcha)

Pred \ Real	Grave	Moderado	Leve
Grave	8	15	18
Moderado	201	713	1026
Leve	1930	21477	138674

Pred \ Real	Grave	Moderado	Leve
Grave	41	158	121
Moderado	1572	6274	9335
Leve	17630	193405	1247999

Las matrices de confusión muestran que el modelo tiende a subestimar la gravedad de los accidentes. En ambos conjuntos (train y test), muchos accidentes graves y moderados son clasificados como leves.

Por ejemplo, en el conjunto de prueba, el modelo predice casi 1.5 millones de accidentes como leves, incluyendo muchos que en realidad eran moderados (193.405) o graves (17.630). Esto indica una fuerte inclinación del modelo a predecir la categoría menos grave, algo lógico ya que es la que más porcentaje ocupa del total. Aunque hay un gran error en la clasificación, el test parece que clasifica mejor los graves y moderados que en el train.

Como se está trabajando con dos conjuntos de datos que no tienen el mismo número de observaciones, la siguiente tabla muestra las matrices de confusión en porcentajes:

Tabla 19: Matrices de Confusión en Porcentajes para Train (Izda) y Test (Dcha)

Pred \ Real	Grave	Moderado	Leve
Grave	0.005 %	0.009 %	0.011 %
Moderado	0.12 %	0.43 %	0.63 %
Leve	1.18 %	13.09 %	84.53 %

Pred \ Real	Grave	Moderado	Leve
Grave	0.003 %	0.011 %	0.008 %
Moderado	0.11 %	0.42 %	0.63 %
Leve	1.19 %	13.10 %	84.52 %

Tras convertir a porcentajes se confirma que el modelo acierta masivamente con la clase leve (84 % de 85.171 %), pero apenas detecta correctamente las más graves: solo el 0.42 % sobre el 13.529 % de los moderados y 0.003 % de los graves sobre el 1.305 %. Además, la ligera mejora aparente que se comentaba antes en test frente a train en la clasificación de Moderados y Graves resulta engañosa, ya que ambos conjuntos presentan porcentajes prácticamente idénticos.

A continuación se muestran las métricas del modelo para ambos conjuntos:

Tabla 20: Métricas del Modelo en los Conjuntos Train y Test

Métrica	Train	Test
Accuracy	0.850	0.849
Kappa (no ponderado)	0.043	0.041
Kappa ponderado	0.047	0.044
Sensibilidad clase 1 (Grave)	0.004	0.002
Sensibilidad clase 2 (Moderado)	0.032	0.031
Sensibilidad clase 3 (Leve)	0.993	0.992
Especificidad clase 1	0.999	0.999
Especificidad clase 2	0.991	0.991
Especificidad clase 3	0.038	0.037

La precisión del modelo es alta (alrededor del 85 %) tanto en el conjunto de entrenamiento como en el de prueba, lo que sugiere que el modelo predice bien a nivel general. Sin embargo, el coeficiente Kappa es muy bajo, lo que indica que gran parte de esa precisión tan alta se debe al gran número de casos en la clase de accidentes leves.

En cuanto a la sensibilidad, el modelo detecta muy bien los accidentes leves, pero tiene dificultades para identificar los moderados y, especialmente, los graves, con sensibilidades cercanas a cero.

La especificidad, en cambio, es muy alta en todas las clases, lo que implica que el modelo rara vez clasifica como grave o moderado un accidente que en realidad no lo es, lo cual es positivo porque evita alarmas innecesarias; aunque eso también significa que puede pasar por alto muchos de los que sí lo son.

A pesar de que el modelo alcanza una precisión global cercana al 85 %, el coeficiente Kappa, tanto ponderado como no ponderado, sugiere que gran parte de esta precisión puede atribuirse al desequilibrio en la distribución de clases más que a una buena capacidad predictiva del modelo. En concreto, la mayoría de los accidentes se clasifican como leves, lo que facilita una alta precisión simplemente favoreciendo la clase mayoritaria. En este contexto, el uso del Kappa ponderado resulta especialmente relevante, ya

que penaliza en mayor medida los errores más graves, proporcionando así una medida más informativa y realista del rendimiento del modelo.

A continuación se muestra la importancia de las 15 primeras variables más relevantes del modelo en escala logarítmica (por temas de visualización):

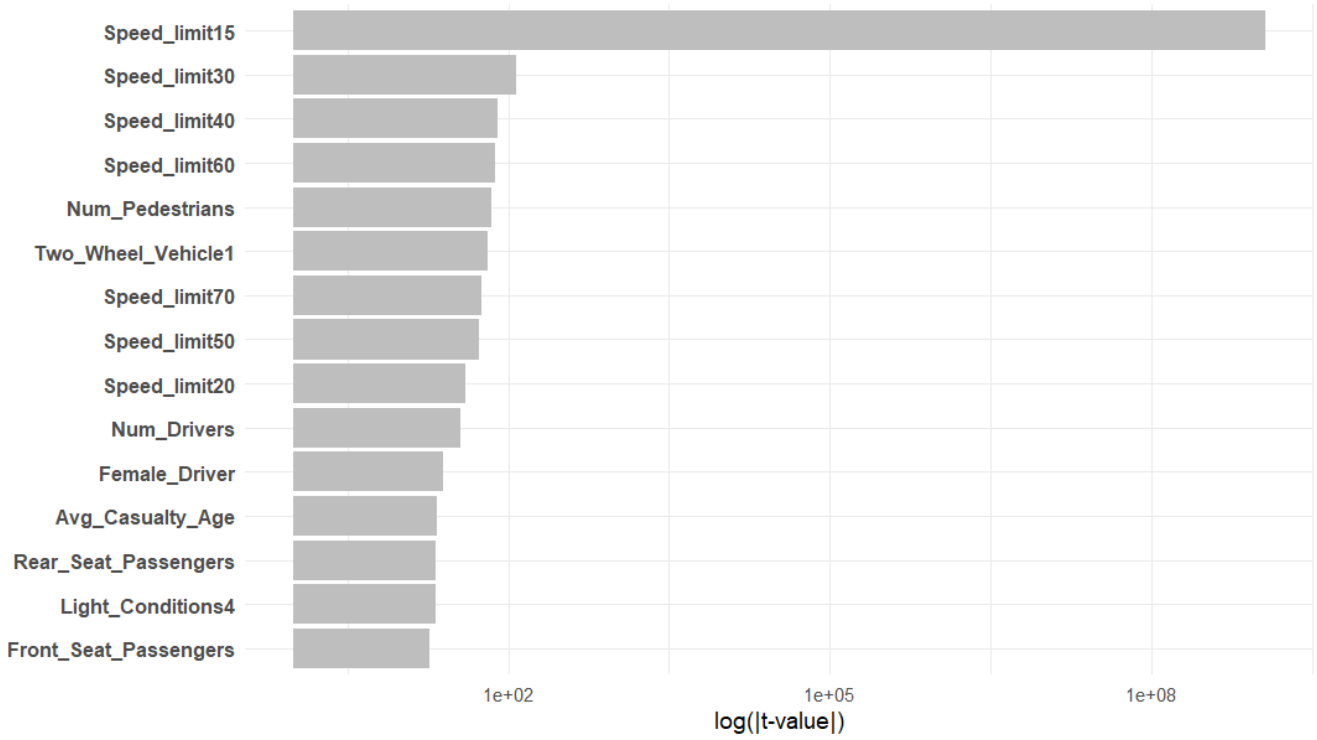


Figura 15: Importancia de Variables Logística Ordinal en Escala Logarítmica

La gráfica de importancia muestra que, con diferencia, el límite de velocidad de 15 mph es la variable que más aporta al modelo. A continuación destacan otros tramos de velocidad (30, 40, 60 y 70 mph), seguidos por el número de peatones y la presencia de vehículos de dos ruedas. Más abajo figuran variables relacionadas con las personas implicadas en el accidente (número de conductores, edad media de las víctimas, pasajeros en asientos delanteros y traseros) y, finalmente, las condiciones de luz (*Light_Conditions4*). En conjunto, estos resultados sugieren que los límites de velocidad y el volumen y tipo de usuarios en la vía son los principales determinantes de la severidad de los accidentes en nuestro modelo ordinal.

5.2. Regresión Lineal Múltiple

5.2.1. Marco Teórico

La regresión lineal múltiple es un tipo de modelización que permite analizar cómo algunas variables pueden explicar o predecir otra. La variable a predecir se llama respuesta o dependiente y las que la explican son aquellas llamadas predictoras o independientes. En este modelo la variable regresora será *Accident_Severity_Score* y algunas de las independientes serán *Number_of_Vehicles* o *Female_Casualties*

La diferencia con el modelo de regresión simple, que solo tiene una variable explicativa, es que en el modelo de regresión múltiple se usan dos o más al mismo tiempo, de ahí el *múltiple*. El modelo supone que existe una relación lineal entre la variable respuesta y cada una de las variables predictoras, es decir, que cada una tiene algún tipo de efecto constante sobre el resultado de la variable dependiente, si el resto se mantienen iguales.

El modelo se puede expresar matemáticamente así:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,, \quad \forall i = 1...n$$

siendo β_j los coeficientes que indican cuánto cambia la variable dependiente cuando cambia la variable explicativa j, y ε_i el error aleatorio [11] [12].

5.2.2. Desarrollo del Modelo

Para construir el modelo de regresión lineal múltiple, se dividieron los datos en dos partes: un 70 % para entrenar el modelo y un 30 % para probarlo. Se ha tomado esta división porque, como ya se comentó anteriormente en el modelo de regresión logística ordinal, es la más común. El modelo inicial se ajustó utilizando todas las variables disponibles como posibles explicativas.

Después de ver las métricas del R^2 y el RMSE, se procedió a eliminar una variable que no tenía ninguno de sus parámetros significativos. Se volvió a repetir todo el proceso, creando los datos train y test, calculando el nuevo modelo y comparando las métricas con el modelo anterior.

En ambos modelos se intentó reducir el número de variables aplicando diferentes métodos de selección automática: *forward* (añadir variables una a una), *backward* (eliminarlas una a una) y *stepwise* (combinación de ambas), usando como criterios AIC y BIC. Sin embargo, todos estos métodos devolvieron modelos casi iguales al original, con prácticamente el mismo número de variables, por lo que no resultaron útiles para simplificar el modelo.

5.2.3. Resultados y Evaluación del Modelo

El primer modelo se construyó con todas las variables, dividiendo la base de datos en un 70 % para train y 30 % para test. Se obtuvieron 61 parámetros de los cuales no eran significativos los siguientes:

- β_0: el intercept o término independiente.
- *Road_Type5:* carril de aceleración o deceleración.
- *Road_Type6:* desconocido.

- *Speed_Limit:* todos los parámetros de la variable son no signficativos.
- *Weather_Conditions4:* cielos despejados con viento fuerte.
- *Road_Surface_Conditions2:* calzada húmeda mojada.
- *Road_Surface_Conditions5:* inundación con mas de 3 cm.
- *Urban_or_Rural_Area3:* sin localización.

Los parámetros que sí son significativos se adjuntan en la siguiente tabla:

Tabla 21: Coeficientes Estimados del Modelo de Regresión Inicial

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Male_Driver	0.00704	0.00057	12.276	<2e-16
Female_Driver	0.00668	0.00063	10.591	<2e-16
Avg_Driver_Age	-0.00008	0.00002	-4.531	5.87e-06
Left_Hand_Drive1	-0.03179	0.00683	-4.652	3.28e-06
Car1	-0.02986	0.00074	-40.487	<2e-16
Two_Wheel_Vehicle1	0.06777	0.00051	132.702	<2e-16
Trucks1	0.01715	0.00060	28.768	<2e-16
Impact_Did_Not_Impact	-0.14080	0.00470	-29.984	<2e-16
Impact_Front	-0.13690	0.00468	-29.291	<2e-16
Impact_Back	-0.17590	0.00468	-37.609	<2e-16
Impact_Offside	-0.15120	0.00468	-32.293	<2e-16
Impact_Nearside	-0.15280	0.00469	-32.583	<2e-16
Work_Purpose1	-0.00422	0.00041	-10.272	<2e-16
Education_Purpose1	-0.01032	0.00125	-8.232	<2e-16
Avg_Vehicle_Age	0.00083	0.00005	16.616	<2e-16
Number_of_Vehicles	0.13920	0.00470	29.645	<2e-16
Number_of_Casualties	0.12160	0.00746	16.310	<2e-16
Day_of_Week2	-0.01434	0.00074	-19.459	<2e-16
Day_of_Week3	-0.01687	0.00073	-23.073	<2e-16
Day_of_Week4	-0.01678	0.00073	-22.994	<2e-16
Day_of_Week5	-0.01532	0.00073	-20.984	<2e-16
Day_of_Week6	-0.01453	0.00071	-20.352	<2e-16
Day_of_Week7	-0.00388	0.00074	-5.257	1.47e-07
Road_Type2	0.01165	0.00146	7.963	1.68e-15
Road_Type3	0.01884	0.00092	20.439	<2e-16
Road_Type4	0.02262	0.00075	30.308	<2e-16
Light_Conditions4	0.02116	0.00049	43.509	<2e-16
Light_Conditions5	0.01922	0.00269	7.134	9.72e-13
Light_Conditions6	0.04512	0.00087	52.139	<2e-16
Light_Conditions7	0.01125	0.00179	6.280	3.39e-10
Weather_Conditions2	-0.01133	0.00071	-15.885	<2e-16

Tabla 22: Coeficientes Estimados del Modelo de Regresión Inicial (continuación)

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Weather_Conditions3	-0.00999	0.00278	-3.594	0.000325
Weather_Conditions5	-0.01139	0.00160	-7.134	9.75e-13
Weather_Conditions6	-0.01839	0.00541	-3.399	0.000677
Weather_Conditions7	-0.00768	0.00247	-3.103	0.001914
Weather_Conditions8	-0.01435	0.00128	-11.185	<2e-16
Weather_Conditions9	-0.00999	0.00134	-7.439	1.02e-13
Road_Surface_Conditions3	-0.01607	0.00300	-5.366	8.07e-08
Road_Surface_Conditions4	-0.01821	0.00135	-13.516	<2e-16
Urban_or_Rural_Area2	0.01683	0.00055	30.805	<2e-16
Num_Drivers	0.00932	0.00060	15.406	<2e-16
Num_Pedestrians	0.09023	0.00074	121.562	<2e-16
Male_Casualties	0.03823	0.00745	5.131	2.88e-07
Female_Casualties	0.01659	0.00745	2.226	0.026009
Avg_Casualty_Age	0.00090	0.00001	67.992	<2e-16
Front_Seat_Passengers	0.00402	0.00066	6.063	1.34e-09
Rear_Seat_Passengers	0.00978	0.00066	14.733	<2e-16

El modelo nos permite detectar algunos patrones claros sobre cuándo un accidente tiende a ser más grave. En general, los accidentes con peatones, motoristas o personas mayores suelen ser más graves. También lo son cuando hay varios pasajeros o las condiciones de luz y clima son malas. Por el contrario, los accidentes en carreteras amplias, bien iluminadas y sin peatones suelen ser más leves. Estas son características que ya se habían observado en el modelo logístico ordinal.

Además, este modelo aporta otros detalles como, por ejemplo, se aprecia que los accidentes ocurridos entre semana, (2, 3, 4, 5, 6) tienden a ser menos graves que los de fin de semana, probablemente por ser en sábados y domingos, días en los que el ocio aumenta. A nivel del sexo de las víctimas, cuando aumenta el número, claramente aumenta la puntuación. En los hombres es algo mayor el efecto, pero no tiene por qué ser por un motivo concreto, sino porque hay más presencia de víctimas masculinas que femeninas.

Por otro lado, se observa que el tipo de vía (*Road_Type*) y las condiciones de la calzada en ciertas categorías (*Road_Surface_Conditions*) también aumentan la gravedad, aunque en menor medida que otras variables.

A continuación se muestran algunos parámetros con sus interpretaciones:

- **Two_Wheel_Vehicle1 ($\beta = 0{,}0678$):** la presencia de una moto, frente a no tenerla, aumenta la puntuación de gravedad 0.0678 puntos, si el resto de variables son constantes
- **Light_Conditions4 ($\beta = 0{,}0212$):** si el accidente ocurre de noche y sin luces de la calle, la puntuación se agrava un 0.02 y por tanto hace que el accidente sea algo más grave.
- **Num_Pedestrians ($\beta = 0{,}0902$):** cada peatón que se ha visto involucrado en el accidente aumenta un 0.09 la puntuación de gravedad. Este valor es bastante cercano a la ponderación otorgada a una persona en estado de gravedad leve.

Para ver la eficiencia del modelo, se calcularon algunas métricas entre las que se encuentran el coeficiente de determinación $R^2 = 0{,}327$ y el ajustado $R^2_{\text{adj}} = 0{,}3269$. Esto significa que el modelo explica cerca del 33 % de la variabilidad total. Aunque este valor parece bajo, hay que tener en cuenta que los accidentes son muy variables y, por tanto, difíciles de interpretar, lo que hace que este R^2 sea bajo. El error estándar residual (RSE) fue de 0.1943.

Como ya se ha comentado, hay que tener en cuenta el número de observaciones y variables.

Se volvió a calcular el modelo obteniendo 54 parámetros de los cuales solo 4 de ellos eran no significativos. De hecho, el término independiente en este modelo sí que es significativo. Los siguientes parámetros que se muestran son los que son significativos, por lo que se ha eliminado el p-valor:

Tabla 23: Coeficientes Estimados del Modelo de Regresión Final

Variable	**Estimate**	**Variable**	**Estimate**
(Intercept)	-0.03228	Road_Type3	0.02354
Male_Driver	0.00792	Road_Type6	0.02160
Female_Driver	0.00714	Road_Type7	0.00609
Avg_Driver_Age	-0.00009	Light_Conditions4	0.01829
Left_Hand_Drive1	-0.03264	Light_Conditions5	0.01758
Car1	-0.03038	Light_Conditions6	0.05488
Two_Wheel_Vehicle1	0.06550	Light_Conditions7	0.00945
Trucks1	0.01822	Weather_Conditions2	-0.01206
Impact_Did_Not_Impact	-0.13970	Weather_Conditions3	-0.01100
Impact_Front	-0.13640	Weather_Conditions5	-0.01157
Impact_Back	-0.17560	Weather_Conditions6	-0.01788
Impact_Offside	-0.15050	Weather_Conditions7	-0.00551
Impact_Nearside	-0.15280	Weather_Conditions8	-0.01505
Work_Purpose1	-0.00381	Weather_Conditions9	-0.01143
Education_Purpose1	-0.01233	Road_Surface_Conditions2	0.00211
Avg_Vehicle_Age	0.00083	Road_Surface_Conditions3	-0.01391
Number_of_Vehicles	0.13680	Road_Surface_Conditions4	-0.01488
Number_of_Casualties	0.12160	Urban_or_Rural_Area2	0.04105
Day_of_Week2	-0.01511	Num_Drivers	0.01221
Day_of_Week3	-0.01764	Num_Pedestrians	0.08735
Day_of_Week4	-0.01749	Male_Casualties	0.03773
Day_of_Week5	-0.01602	Female_Casualties	0.01600
Day_of_Week6	-0.01517	Avg_Casualty_Age	0.00091
Day_of_Week7	-0.00429	Front_Seat_Passengers	0.00569
Road_Type2	0.00831	Rear_Seat_Passengers	0.01069

En este caso la constante β_0 representa la gravedad estimada del accidente cuando todas las variables del modelo valen cero. Como muchas son dummies, este valor no tiene una interpretación práctica clara.

Al igual que en el modelo original, las variables relacionadas con el primer impacto de cada coche salen con coeficientes negativos. Esto llama la atención, ya que significa que a mayor número de impactos menor va a ser la puntuación de gravedad. Que haya más impactos no implica necesariamente mayor gra-

vedad porque estos accidentes pueden ser situaciones con varios coches con golpes leves (por ejemplo, golpes en cadena o en ciudad) donde el riesgo para las personas dentro del vehículo es bajo. Este puede ser el motivo de que el modelo asigne un coeficiente negativo, ya que la gravedad del accidente tiende a ser menor.

Como es lógico, a mayor número de vehículos y personas accidentadas, mayor será la puntuación de gravedad. Esto no quiere decir que si estos números son altos, el accidente haya sido en general mucho más grave, sino que al final se van sumando las puntuaciones y por cada coche se aumenta 0.14 la gravedad y por cada accidentado 0.12.

Al igual que en el modelo anterior, se calcularon medidas de eficiencia para evaluar el ajuste del modelo. En este caso, el coeficiente de determinación fue $R^2 = 0{,}3236$, y el ajustado fue exactamente el mismo, lo que indica que el número de variables no parece penalizar mucho el modelo.

Si se comparan estos valores con los del modelo anterior ($R^2 = 0{,}327$ y $R^2_{\text{adj}} = 0{,}3269$), se observa que las diferencias son mínimas. Esto sugiere que, aunque se ha simplificado el modelo, el ajuste a los datos sigue siendo prácticamente igual a pesar de reducir la complejidad.

Para ambos modelos se probaron los métodos automáticos de selección de variables (*forward*, *backward* y *stepwise*) usando los criterios AIC y BIC. Para el modelo original se obtuvieron en todos los modelos AIC 61 parámetros (los mismos que el original) y para los BIC 60. Pasó lo mismo con el modelo reducido final, para el AIC se obtuvieron 54 parámetros y para BIC 53. Como ninguno redujo el número de parámetros y complicaba el proceso, se optó por no utilizarlos.

Como en el modelo de regresión logística ordinal, se presenta a continuación un gráfico con las 15 variables más importantes del modelo final:

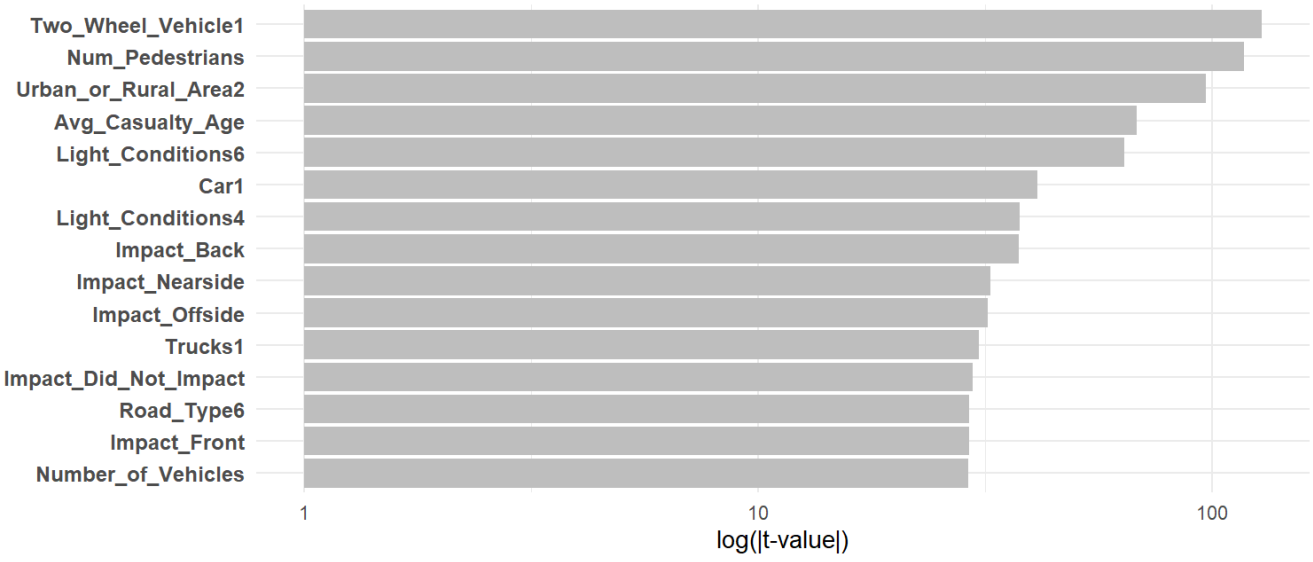


Figura 16: Importancia de Variables Regresión Lineal Múltiple en Escala Logarítmica

En el gráfico se observa que las variables sobre los accidentados (número de peatones, vehículos de dos ruedas y edad de las víctimas) y el entorno (urbano vs. rural y condiciones lumínicas), dominan la

importancia, seguidas de las condiciones del choque (tipo de impacto). El límite de velocidad y otras características de la vía tienen un peso menor en este modelo.

En comparación, el modelo logístico ordinal prioriza los tramos de velocidad (especialmente 15 mph) como principales variables, mientras que la regresión lineal refleja más la cantidad y tipo de usuarios y las circunstancias del impacto. Esto sugiere que, al predecir la gravedad en forma continua, los accidentados y sus condiciones toman más relevancia que los límites de velocidad.

5.3. Árboles de Clasificación y Regresión

5.3.1. Marco Teórico

Árboles de Clasificación y Regresión

Los árboles de clasificación y regresión construyen divisiones óptimas para predecir una variable, utilizando toda la base de datos y teniendo en cuenta la información de todas las variables. Se ha optado por añadir esta técnica ya que es muy visual y ayuda a entender de manera rápida los datos sin tener que saber interpretar parámetros u *odds ratios*.

Estos árboles funcionan dividiendo los datos en ramas, buscando en cada paso la mejor forma de separar los casos según las variables disponibles y los puntos de corte óptimos. Así se van creando nodos hasta llegar a hojas donde se hace la predicción. En los árboles de regresión se predice un valor numérico (continuo), y en los de clasificación una categoría [9], [10].

Bagging y Random Forest

Random Forest es una técnica que mejora los árboles individuales generando muchos árboles distintos y combinando sus resultados. A diferencia de un único árbol, que puede variar mucho con pequeños cambios en los datos, Random Forest es más estable y suele ofrecer mejores resultados.

Su gran ventaja es que introduce dos fuentes de aleatoriedad:

- **Bagging (Bootstrap Aggregating):** cada árbol se entrena con un subconjunto aleatorio de las observaciones (muestra con reemplazamiento). De este modo, los árboles ven datos ligeramente distintos unos de otros y no todos aprenden los mismos detalles del conjunto original.
- **Selección aleatoria de variables:** en cada nodo, en lugar de evaluar todas las variables para la división, se elige al azar un pequeño grupo. Esto evita que las características muy fuertes dominen siempre el proceso y fomenta la diversidad entre los árboles.

Al combinar (promediar o votar) los resultados de un número elevado de árboles, Random Forest puede:

- **Reducir la varianza:** el promedio de muchos árboles con errores no correlacionados tiende a acercarse al valor esperado verdadero, suavizando fluctuaciones aleatorias.
- **Mantener el sesgo bajo:** cada árbol individual es un modelo potente (bajo sesgo), y al agregarlos conservamos esa capacidad.
- **Proporcionar estimaciones internas de error:** usando las observaciones (OOB), podemos evaluar sin necesidad de test externo.
- **Calcular medidas de importancia de variables:** analiza cómo cambia el error cuando permutamos una característica, identificando las más influyentes.

En resumen, Bagging aporta robustez frente a la variabilidad de los datos y la selección aleatoria de variables fomenta la diversidad. Esta combinación hace de Random Forest uno de los métodos más fiables y versátiles en problemas de clasificación y regresión [13] [14].

Importancia de las Variables

Una de las utilidades de los árboles es que permiten conocer qué variables son más importantes para la predicción. Para medirlo, se puede alterar una variable en los datos, por ejemplo, mezclando aleatoriamente sus valores, y observar cuánto cambia el error del modelo. Si el error aumenta significativamente, eso implica que esa variable tenía un papel importante a la hora de predecir en el modelo. [9], [10].

5.3.2. Desarrollo del Modelo

Árbol de Clasificación

Para el árbol de clasificación se dividió el conjunto de datos en dos partes, un 70 % para train y un 30 % para test. Sin embargo, se observó que las clases estaban desbalanceadas, ya que había un número muy superior de leves que de graves.

Para evitar que el modelo se centrara solo en la clase mayoritaria, se balanceó el conjunto de entrenamiento. Esto se hizo dejando un mismo número aproximado de casos en cada clase, igualando todas al tamaño menos frecuente, haciendo una selección aleatoria dentro de cada grupo. A este tipo de técnica de balance se le da el nombre de *undersampling* o submuestreo. De este modo, el modelo tiene más posibilidades de aprender bien a distinguir entre todas las categorías de gravedad del accidente y no solo la más común.

Una vez hecho esto, se construyó el modelo ajustando los parámetros para tener un mínimo de 20 observaciones para realizar divisiones, un mínimo de 35 por hoja y un parámetro de complejidad de 0.005 para evitar sobreajuste. Todos estos parámetros se modificaron varias veces para obtener un resultado con bastante información, pero que siguiera siendo visible en la imagen.

Posteriormente, se entrenó un modelo de Random Forest a partir del conjunto de entrenamiento original, sin balancear, ya que este algoritmo lo hace de forma automática. Se construyeron 100 árboles y, tras el entrenamiento, se realizaron las predicciones sobre el conjunto de prueba y se obtuvo un gráfico de importancia de las variables.

Árbol de Regresión

Para el caso del árbol de regresión, el procedimiento fue bastante similar, sin hacer claramente ese balanceo en las categorías. En este caso, los parámetros se fijaron en un mínimo de 10 observaciones por nodo y por hoja, y un valor de complejidad de 0.002.

Después, se entrenó un modelo de *Random Forest* de regresión sobre el mismo conjunto de entrenamiento, empleando de nuevo 100 árboles. Al igual que en el ordinal, se realizaron predicciones sobre el conjunto de prueba y se generó el gráfico de importancia de las variables.

5.3.3. Resultados y Evaluación del Modelo

Árbol de Clasificación

A continuación se presenta el árbol de clasificación ordinal:

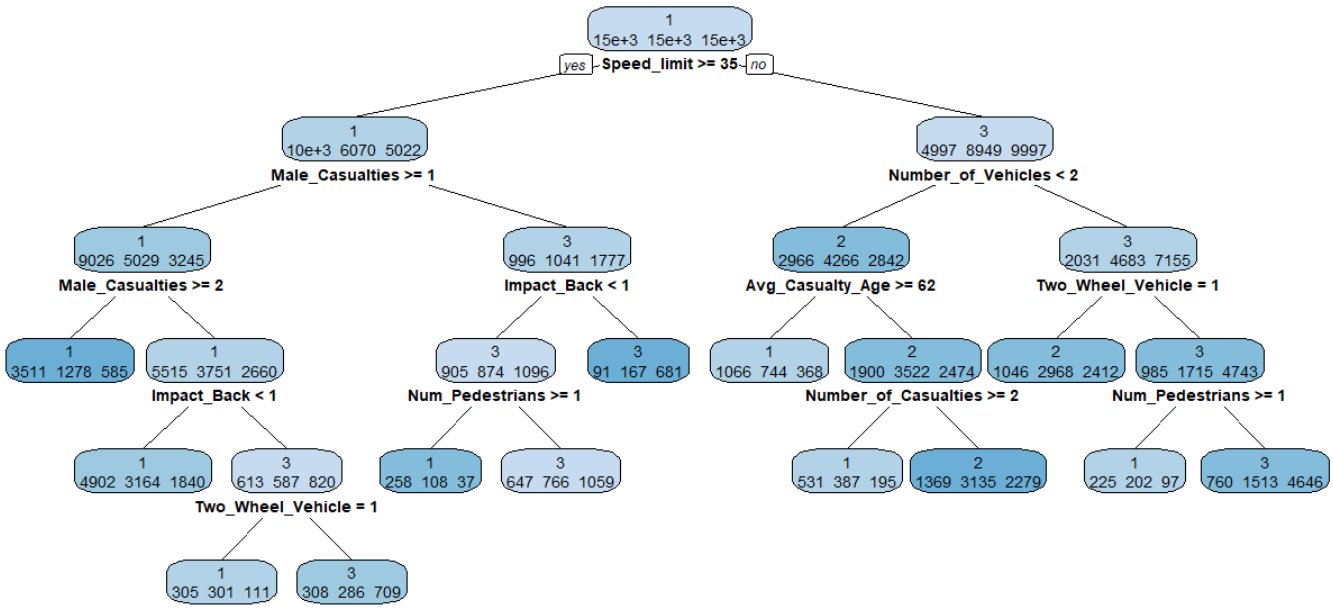


Figura 17: Árbol Individual de Clasificación

El árbol tiene 12 hojas y de profundidad 6 nodos como máximo. La división es coherente, dado que en la mayoría de las hojas hay diferencia entre las categorías.

La primera variable que divide el árbol es el *Speed_limit*. Si el límite de velocidad es mayor o igual a 35 mph, el modelo tiende a clasificar el accidente como menos grave. En cambio, si el límite es más bajo, el modelo predice más casos graves. Esto es algo que ya se había observado en el modelo de logística ordinal, donde se dio una posible explicación.

En la rama de la izquierda se puede observar que una de las variables más importantes para clasificar un accidente como grave, es la de accidentados masculinos, especialmente si hay más de 2, ya se clasifica como tal. Si hay solo uno y no hay primeros impactos en la parte trasera también irá a esa categoría. A partir de ahí, en esa rama ya no clasifica tan bien los accidentes graves, ya que en la tercera hoja empezando por la izquierda hay una proporción muy similar de accidentes graves o moderados. Lo bueno, al menos, es que sí que hace una gran distinción con los accidentes leves.

Es importante remarcar la repetida aparición de las motos, que indica que si hay presencia de este tipo de vehículos, el accidente será peor. También aparece la variable de número de peatones accidentados. La interpretación es parecida a la de las motos, si hay más de un peatón el accidente será grave. De hecho, este caso lo vemos dos veces. Estos resultados son lógicos ya que las motos son vehículos menos protegidos (en comparación a un coche) y los peatones no tienen protección ninguna frente a un choque.

De la rama de la derecha ya se han comentado algunas de las divisiones y la única variable a destacar es la edad media del accidentado. Si este es mayor de 62 años, la gravedad del accidente será alta, algo entendible, ya que la recuperación es peor que la de una persona joven.

Aunque se han obtenido resultados parecidos, el árbol aporta una información mucho más manejable, como la de la última variable mencionada. No es lo mismo decir personas mayores, que personas mayores de 62 años, es mucho más concreto.

Por último, se evaluó el rendimiento del modelo tanto en el conjunto de entrenamiento como en el conjunto de prueba. Las matrices de confusión, ya desbalanceadas, se muestran a continuación:

Tabla 24: Matrices de Confusión para Train (Izda) y Test (Dcha)

Pred \ Real	Grave	Moderado	Leve
Grave	10798	64100	210392
Moderado	2415	63020	304290
Leve	1806	28076	463520

Pred \ Real	Grave	Moderado	Leve
Grave	4574	27653	90626
Moderado	1028	27081	130632
Leve	761	12112	197713

Como en el modelo de logística ordinal se muestran también los porcentajes para una mejor visualización:

Tabla 25: Matrices de Confusión en Porcentajes para Train (Izda) y Test (Dcha)

Pred \ Real	Grave	Moderado	Leve
Grave	0.94 %	5.58 %	18.32 %
Moderado	0.21 %	5.49 %	26.50 %
Leve	0.16 %	2.44 %	40.36 %

Pred \ Real	Grave	Moderado	Leve
Grave	0.93 %	5.62 %	18.41 %
Moderado	0.21 %	5.50 %	26.54 %
Leve	0.15 %	2.46 %	40.17 %

Las tablas muestran un comportamiento muy similar en train y test, lo que indica que el modelo no está sobreajustando. Solo alrededor del 1 % de los accidentes graves (sobre el total del 1.31 %) y del 5,5 % de los moderados (sobre el 13.75 %) se clasifican correctamente; el resto suele caer en la categoría leve, que acumula casi el 40 % de los casos. Esto confirma el fuerte sesgo hacia la clase mayoritaria y sugiere la necesidad de ajustar pesos o umbrales para mejorar la detección de accidentes graves y moderados.

Comparando estos resultados con el modelo de regresión logística ordinal, el árbol de clasificación presenta un patrón muy similar en la estimación de casos graves y moderados, con la mayoría de predicciones concentradas en la categoría leve. No obstante, la regresión logística alcanza una detección ligeramente superior de moderados y graves, mientras que el árbol tiende a exagerar aún más el sesgo hacia leve. En ambos casos, resulta imprescindible ajustar pesos o umbrales para mejorar la identificación de accidentes de mayor severidad. A continuación se muestran las métricas del modelo para ambos conjuntos:

Tabla 26: Métricas del Modelo en los Conjuntos Train y Test

Métrica	Train	Test
Accuracy	0.4679	0.466
Kappa (no ponderado)	0.094	0.093
Kappa ponderado	0.1231	0.1221
Sensibilidad clase 1 (Grave)	0.7190	0.7188
Sensibilidad clase 2 (Moderado)	0.4061	0.4051
Sensibilidad clase 3 (Leve)	0.4738	0.4719
Especificidad clase 1	0.7578	0.7565
Especificidad clase 2	0.6912	0.6905
Especificidad clase 3	0.8244	0.8242

Los resultados de la tabla muestran métricas casi idénticas en train y test, lo que indica que no hay mucho sobreajuste. La precisión es baja y los kappa revelan un acuerdo apenas superior al azar (0.094 sin ponderar, 0.123 ponderado), aunque este último refleja que el modelo capta algo del orden entre clases. La sensibilidad para accidentes graves es alta, mientras que para moderados (40 %) y leves (47 %) es moderada, y la especificidad es notablemente mejor para graves y leves. En conjunto, el modelo identifica bien los casos más severos, pero requiere ajustes (pesos de clase o umbrales) para mejorar la detección de los accidentes moderados y leves.

En comparación con el modelo de regresión logística ordinal, presentaba menor precisión, pero detectaba mejor los accidentes graves y moderados.

La siguiente imagen es una representación visual de la importancia de las variables en el modelo de Random Forest:

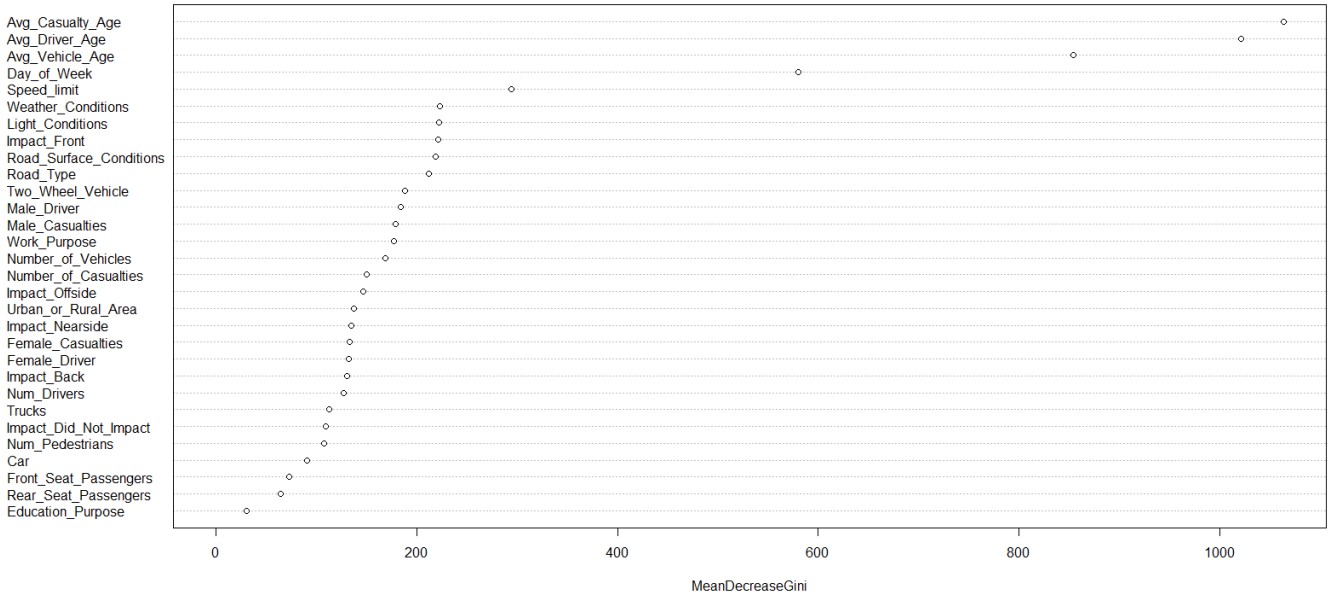


Figura 18: Importancia de Variables Random Forest Clasificación

Las variables más importantes en la iteración de la creación de árboles son aquellas relacionadas con la edad, ya sea del accidentado (*Avg_Casualty_Age*) o conductor (*Avg_Driver_Age*) como la del coche (*Avg_Vehicle_Age*). Después destacan variables como el día de la semana (*Day_of_Week*), la velocidad límite (*Speed_Limit*) o las condiciones meteorológicas o lumínicas (*Weather_Conditions* / *Light_Conditions*).

Con estas medidas se muestra la importancia de introducir la aleatoriedad en los modelos, ya que solo dos variables de las mencionadas se encontraban incluidas en el árbol individual. Aún así, no hay ninguna variable que presente algo sorprendente o fuera de lugar, ya que todas han ido apareciendo en los modelos anteriores.

Árbol de Regresión

A continuación se presenta el árbol individual de regresión sobre la variable *Accident_Severity_Score*:

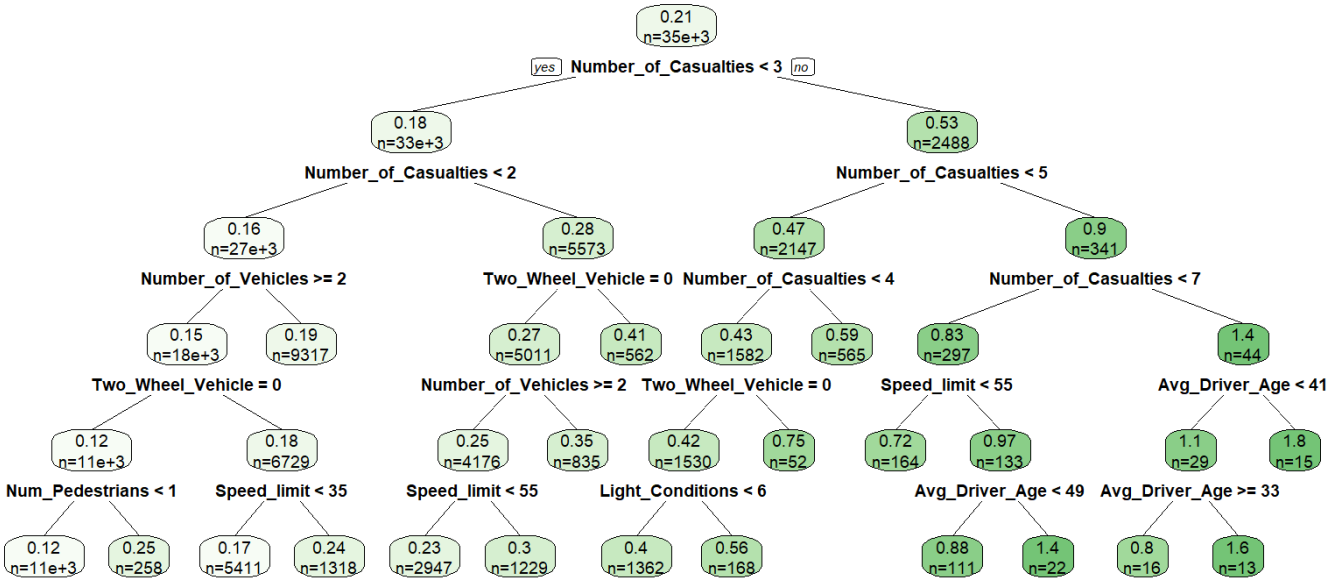


Figura 19: Árbol Individual de Regresión

Algunas de las variables que más destacan en el árbol son el número de personas accidentadas (*Number_of_Casualties*), la edad media de los conductores (*Avg_Driver_Age*) y el límite de velocidad de la vía (*Speed_limit*). La variable *Number_of_Casualties* aparece hasta cinco veces, mientras que las otras dos se repiten tres veces cada una. Esto muestra su gran relevancia para explicar la gravedad de los accidentes, al ser seleccionadas en varios nodos del árbol como el mejor punto de división. La parte negativa es que no permite incluir a otras variables que son menos óptimas para que aporten información.

El valor que aparece en cada nodo del árbol representa la predicción promedio de la variable dependiente en ese subconjunto de datos, algo parecido al clúster. Por ejemplo, el nodo raíz (el más alto) tiene un valor de 0.21, lo que indica que el valor medio general del conjunto es bajo, algo esperable, ya que la mayoría de los accidentes son leves. Sin embargo, hay ramas del árbol que alcanzan valores mayores, como 1.8 o 1.6, lo que representa grupos donde la gravedad promedio es varias veces la del primer nodo.

Se puede observar que los accidentes son más graves cuando hay muchos accidentados, el límite de velocidad es alto o los conductores implicados son de mayor edad. A la inversa, valores bajos en esas variables suelen llevar a nodos con menor puntuación de gravedad.

En Random Forest se obtuvo un RMSE de 0.2019, frente a 0.1934 del modelo lineal, y un R^2 de 0.3134, ligeramente inferior al 0.3236 de la regresión lineal. Esto indica que explica un poco peor la variabilidad de los datos y presenta un error medio algo mayor, por lo que su precisión media es ligeramente menor.

De nuevo, a continuación se muestra el gráfico de importancia de variables de Random Forest:

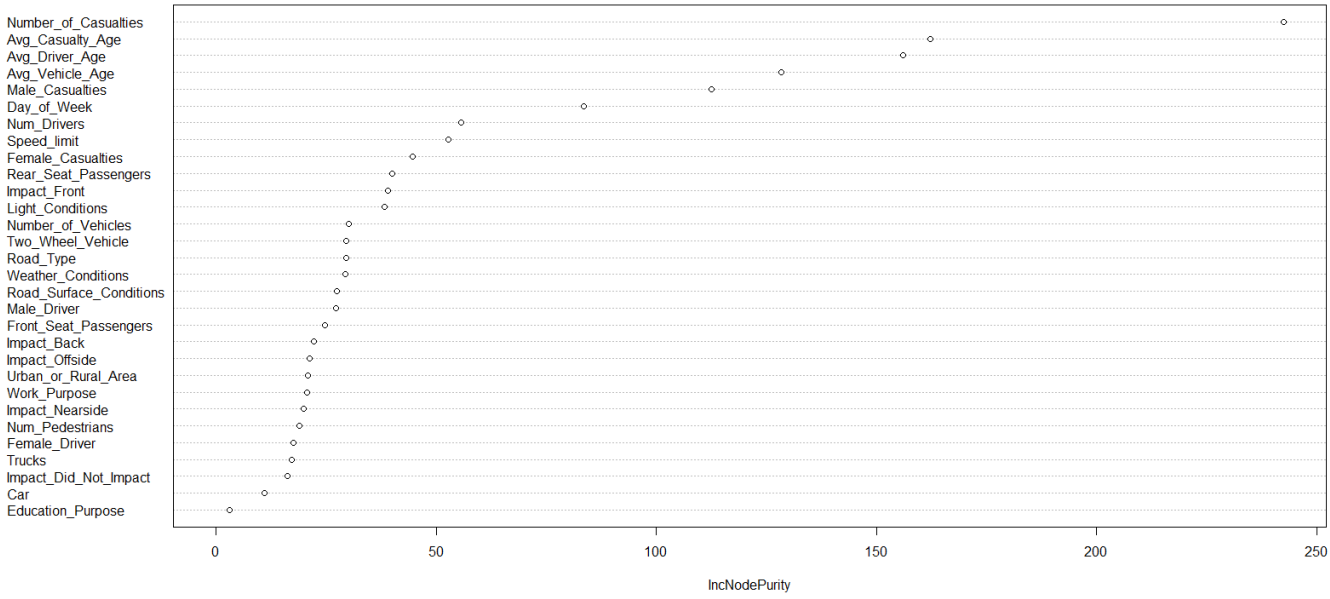


Figura 20: Importancia de Variables Random Forest Regresión

Aunque las variables de edad siguen presentes entre el grupo de las más importantes, la variable *Number_of_Casualties* toma el primer puesto. Es normal, ya que la variable está creada a partir de los accidentados y, si un accidente tiene muchas personas involucradas, aunque sea leve, su puntuación va a ser alta.

Otra de las variables que ha experimentado un aumento de importancia son los accidentados masculinos. Aunque sí se ha observado una cierta relación en otros modelos con la gravedad del accidente, hay que tener presente que solo había un 17.3 % de los accidentes que no tenían conductores, frente al 56.3 % de conductoras. Esto hace que haya una mayor presencia masculina en los accidentes y que pueda influir en esta importancia de variables.

6. Conclusiones

6.1. Resultados obtenidos

Este Trabajo de Fin de Grado estudia los factores que influyen en la gravedad de los accidentes de tráfico en el Reino Unido (2005–2014). Para ello, se unieron las bases de datos *Accidents*, *Vehicles* y *Casualties* con el fin de hacer el trabajo con los datos más manejables sin perder información. Tras el proceso de limpieza y agrupación de categorías, se seleccionaron 32 variables clave que aportan información sobre las personas implicadas, los vehículos y las condiciones del entorno.

En el análisis descriptivo univariante se encontró que el 85.2 % de los siniestros fueron clasificados como leves, con conductores cuya edad media rondaba los 40 años y vehículos de 5–10 años de antigüedad. Se crearon gráficos de sectores y diagramas de caja para algunas de las variables seleccionadas, revelando que la mayoría de los accidentes involucraba uno o dos vehículos y que más de la mitad de estos tenían una antigüedad de entre 5 y 10 años. El estudio bivariante mostró asociaciones significativas entre las variables sobre la edad de conductores y víctimas, así como entre los números de pasajeros y conductores con la gravedad del accidente.

Con clústeres k-medias (k=5,7) se agruparon los accidentes según sus características, formando subconjuntos con perfiles similares. Este paso permitió clasificar accidentes atendiendo a combinaciones de variables como edad promedio de los implicados, número de personas afectadas, tipo de vehículo y condiciones de la vía, sin llegar a predecir, sino identificando subconjuntos homogéneos para su posterior análisis.

Para la predicción de la variable categórica *Accident_Severity* se emplearon modelos de regresión logística ordinal y árboles de clasificación. Los resultados mostraron que, para el primero, las variables más influyentes en la probabilidad de que un accidente sea grave, moderado o leve fueron el límite de la velocidad de la vía, la presencia de peatones o motos, el límite de velocidad de la vía y la edad media de los implicados. El segundo modelo también incluía las ya mencionadas y añade las condiciones meteorológicas y lumínicas. La precisión alcanzada fue baja, debido a la dificultad de predecir los accidentes.

En cuanto a la predicción de la variable continua *Accident_Severity_Score*, se aplicó un modelo de regresión lineal múltiple y árboles de regresión. Los coeficientes de la regresión lineal mostraron que edades avanzadas de los accidentados, condiciones lumínicas deficientes, el número de vehículos y el de accidentados incrementan significativamente la puntuación de gravedad. Los árboles de regresión permitieron identificar umbrales concretos, por ejemplo, un límite de velocidad superior a 55 mph derivaba en valores de *Accident_Severity_Score* más altos.

Una vez analizados todos los modelos, podemos dar respuesta a las preguntas de investigación planteadas al principio de este trabajo.

Las variables que más se relacionan con la gravedad de un accidente son el número de personas accidentadas, la presencia de peatones o motos, la velocidad límite de la vía y la edad media de los implicados. Otras variables relacionadas son las condiciones lumínicas, el tipo y estado de la vía o el clima. Muchas de estas variables han aparecido de forma repetida en diferentes modelos de tal manera que hacían el accidente más grave.

Con información básica como el tipo de vehículo, las condiciones ambientales o el número de personas implicadas, es posible hacer una estimación razonable de la gravedad de un accidente. Aún así, a pesar de que ha habido coincidencias en los modelos, no se ha obtenido una precisión muy alta, posiblemente debido al número de variables y observaciones.

Los perfiles de accidente que deberían priorizarse en las campañas de prevención son los motoristas y las personas mayores. En lo que respecta a las condiciones de la vía, se debe focalizar la protección del peatón en zonas donde el límite de velocidad sea menor a 15 millas por hora, debido a la alta probabilidad de que el accidente sea más grave. Otra característica de la vía a reforzar es el alumbrado en zonas donde la visibilidad es reducida.

Por último, los factores humanos han resultado ser más importantes que los del entorno en los modelos. Variables como la edad o el número de personas implicadas aportan más información que otras como la hora del día o el tipo de impacto. Aun así, combinar ambos tipos de variables da mejores resultados, ya que permite obtener una información más completa.

6.2. Dificultades y Posibles Mejoras

A continuación se describen las limitaciones y consideraciones pendientes en la fase de modelización, así como algunas propuestas de trabajo futuro para profundizar en el análisis de las variables y validar los resultados obtenidos.

En los modelos de logística ordinal se podría haber hecho el balance de categorías para que fuera más representativo. No se hizo por dos motivos principalmente. El primero es que el tiempo de compilación podía llegar a ser de casi 24 horas para realizar los dos modelos y las dos ANOVAS. El segundo fue porque se quería poder ver la comparación entre hacer el balance o no.

A nivel general se podrían realizar análisis descriptivos focalizados únicamente en los accidentes graves para afinar la distribución del resto de variables en esta categoría.

Faltaría comparar las proporciones obtenidas en el análisis descriptivo con los datos reales. Por ejemplo, en el gráfico 9 se observa que el mayor porcentaje de vehículos accidentados son los que tienen entre 5 y 10 años, pero no se tiene en cuenta que muy probablemente representan el mayor porcentaje de los vehículos.

7. Bibliografía

[1] W. H. Organization, "Global Status Report On Road Safety," inf. téc., 2018.

[2] *Road safety statistics - GOV.UK*.
Url: https://www.gov.uk/government/collections/road-accidents-and-safety-statistics.

[3] *UK Accidents 10 years history with many variables*.
Url: https://www.kaggle.com/datasets/benoit72/uk-accidents-10-years-history-with-many-variables.

[4] M. A. Sufian y J. Varadarajan, *Enhancing Prediction and Analysis of UK Road Traffic Accident Severity Using AI*.
Url: https://arxiv.org/pdf/2309.13483.

[5] N. Behboudi, S. Moosavi y R. Ramnath, "Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques,"

[6] *SAS Help Center: Introduction to SEMMA*.
Url: https://documentation.sas.com/doc/en/emref/15.2/n061bzurmej4j3n1jnj8bbjjm1a2.htm.

[7] *Metodologías de Minería de Datos: KDD, CRISP-DM y SEMMA*.
Url: https://www.linkedin.com/pulse/comparativa-de-metodolog%C3%ADas-miner%C3%ADa-datos-kdd-crisp-dm-y-semma-explicadas-3uhwe/.

[8] S. Gayo, "Apuntes Propios Técnicas Estadísticas Multivariantes I," 2023.

[9] S. Gayo, "Apuntes Propios Técnicas de Segmentación y Tratamiento de Encuestas," 2024.

[10] A. Calviño, "Material Didáctico de la Asignatura de Técnicas de Segmentación y Tratamiento de Encuestas," 2024.

[11] J. Amador, "Material Didáctico de la Asignatura de Métodos de Predicción Lineal," 2023.

[12] S. Gayo, "Apuntes Propios Métodos de Predicción Lineal," 2023.

[13] S. Gayo, "Apuntes Propios Técnicas Avanzadas de Predicción," 2025.

[14] J. Portela, "Material Didáctico de la Asignatura de Técnicas Avanzadas de Predicción," 2023.

[15] R. Wirth y J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining,"