COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

ANATOMÍA DE LA SEVERIDAD VIAL: SEGMENTACIÓN DE ACCIDENTES VIALES Y ANÁLISIS DE SU GRAVEDAD A PARTIR DE VARIABLES TEMPORALES, CONTEXTUALES Y PERSONALES

Clave: 202101547

MADRID | Abril 2026

Declaración de uso de herramientas de inteligencia artificial

Por la presente, yo, Jimena Crespo Ayuso, estudiante de 5º E3-Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Anatomía de la severidad vial: segmentación de accidentes viales y análisis de su gravedad a partir de variables temporales, contextuales y personales", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

I. Crítico: Para encontrar contraargumentos a una tesis específica que pretendo defender.

II. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.

III. Interpretador de código: Para realizar análisis de datos preliminares.

IV. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

V. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

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

Fecha: 20 de abril de 2026

Firma: 

Resumen

La siniestralidad vial constituye un problema de gran relevancia social y sanitaria, cuya gravedad depende de la interacción de diversos factores temporales, contextuales y personales. Sobre esta base, el presente Trabajo de Fin de Grado desarrolla un modelo analítico orientado a segmentar los accidentes de tráfico y estudiar los factores asociados a su gravedad.

Para ello, se construye un *dataset* integrado a partir de microdatos oficiales de la Dirección General de Tráfico correspondientes al período 2021-2023, complementados con variables personales y de consumo de sustancias incorporadas mediante procedimientos de estimación sintética. Sobre este *dataset*, se aplican técnicas de *clustering* para identificar tipologías de siniestro, así como modelos de regresión logística para evaluar la relación entre dichas tipologías y la probabilidad de gravedad del siniestro.

Los resultados muestran que, si bien la segmentación permite distinguir tipologías diferenciadas de accidentalidad, la pertenencia a un clúster no muestra por sí sola diferencias estadísticamente significativas en la severidad del accidente. No obstante, el análisis sí permite identificar factores especialmente explicativos de la gravedad dentro de cada grupo, destacando la conducción nocturna sin iluminación artificial y determinadas restricciones de visibilidad.

Así las cosas, el trabajo aporta una herramienta útil para comprender mejor los patrones asociados a la severidad vial y ofrece una base analítica aplicable al diseño de actuaciones preventivas y políticas públicas de seguridad vial.

Palabras clave: accidentes de tráfico, severidad vial, gravedad de los accidentes, segmentación, regresión logística, análisis de datos, seguridad vial.

Abstract

Traffic accidents represent a major social and public health issue, and their severity is shaped by the interaction of multiple temporal, contextual, and personal factors. Within this framework, this study develops an analytical model aimed at segmenting traffic accidents and studying the factors associated with their severity.

To do so, an integrated dataset is constructed using official microdata provided by the Spanish Traffic Authority (*Dirección General de Tráfico*) for the period 2021–2023, complemented with personal and substance use variables incorporated through synthetic estimation procedures. On this dataset, clustering techniques are applied to identify accident typologies, as well as logistic regression models are used to assess the relationship between these typologies and the probability of severe accidents.

The results show that, while segmentation allows for the distinction of different accident typologies, belonging to a cluster does not, on its own, show statistically significant differences in accident severity. However, the analysis does make it possible to identify factors that are particularly explanatory of severity within each group, especially nighttime driving without artificial lighting and certain visibility restrictions.

Overall, this study provides a useful tool for better understanding the patterns associated with traffic severity and offers an analytical framework that can be applied to the design of preventive measures and road safety public policies.

Keywords: traffic accidents, traffic severity, accident severity, clustering, logistic regression, data analysis, road safety.

Índice

Índice de Tablas

Índice de Figuras

CAPÍTULO 1. INTRODUCCIÓN

1.1. Justificación e interés del tema

La siniestralidad vial constituye uno de los principales problemas de salud pública y desarrollo social tanto a nivel global como en España. A escala mundial, los accidentes de tráfico causan aproximadamente 1,19 millones de muertes anuales (WHO, 2023) y, en el caso de España, durante el año 2023 se registraron más de 100.000 siniestros viales, con un balance de 1.806 personas fallecidas y 133.531 heridos (DGT, informe publicado en 2025 con datos cerrados correspondientes a 2023). Este fenómeno, por tanto, refleja una tragedia humana que conlleva consecuencias devastadoras tanto en las víctimas como en el sistema en su totalidad.

Por un lado, en relación con las víctimas, los accidentes de tráfico provocan un amplio espectro de lesiones físicas y psicológicas. Las primeras, de diversa índole, van desde fracturas óseas y traumatismos craneoencefálicos, hasta daños vertebrales y medulares, e incluso la muerte en los casos más críticos (Collado, 2020). Las segundas se reflejan en la aparición de secuelas como el trastorno de estrés postraumático, la depresión, la ansiedad y los cambios de personalidad, que alteran profundamente el día a día de las víctimas y sus familiares (Echeburúa y Esbec, 2015).

Por otro lado, en lo que respecta al sistema, la siniestralidad vial provoca un notable impacto económico y sanitario. El coste económico asociado a los accidentes de tráfico se estima entre el 1% y el 3% del PIB mundial, teniendo en cuenta los gastos derivados de atención médica, pérdida de productividad y daños materiales (Alcaide, 2023). Asimismo, en España, los accidentes de tráfico constituyen la principal causa de ingreso en las Unidades de Cuidados Intensivos (UCI), concentrando el 41% de los ingresos por lesiones graves (Chico et al., 2023).

Así las cosas, considerando la magnitud y repercusión de la siniestralidad vial, resulta evidente la necesidad de desarrollar herramientas analíticas que permitan identificar los factores asociados a la gravedad de los accidentes de tráfico y analizar cómo varía el efecto de estos factores según el contexto en que se producen. En este sentido, el presente trabajo propone la construcción de un modelo de análisis de la severidad vial basado en técnicas de segmentación y regresión.

El modelo planteado pretende integrar información relativa al entorno del accidente (como las condiciones meteorológicas o tipo de vía), el momento temporal del mismo (como la hora del día o día de la semana) y las características personales del conductor (como la edad o el sexo). Analizando este tipo de datos, se busca identificar tipologías diferenciadas de accidente y estudiar cómo varía la probabilidad de severidad en función de esas tipologías observadas.

Desde una perspectiva colectiva, el desarrollo de este modelo podría representar una herramienta clave para mejorar el diseño de estrategias de prevención vial, permitiendo priorizar campañas de sensibilización en función de los perfiles de mayor riesgo, segmentar los distintos escenarios asociados a una mayor severidad y apoyar la toma de decisiones en materia de seguridad vial, contribuyendo así a una mejor asignación de los recursos destinados al control del tráfico.

Todo ello resulta particularmente relevante en el marco de la Visión Cero de la Unión Europea, que persigue, como principal objetivo, reducir a la mitad las muertes y lesiones graves por tráfico para 2030 y eliminarlas para 2050 (Sánchez, 2024). Si bien el número de muertes en carretera en España es menor conforme pasan los años (Llaneza, 2024), el ritmo de mejora es insuficiente para cumplir con estos objetivos, por lo que son especialmente necesarias herramientas como la planteada en este trabajo.

En definitiva, el interés del estudio radica en su relevancia social y económica, al abordar un fenómeno que sigue generando un elevado coste humano y sanitario. Se pretende aportar conocimiento útil para una comprensión más precisa del problema y para el diseño de respuestas institucionales más eficaces. En última instancia, los resultados del trabajo pueden contribuir a la reducción de la mortalidad y morbilidad en carretera.

1.2. Objetivos

El propósito principal del estudio es desarrollar un modelo analítico de severidad vial que permita identificar tipologías de accidente y analizar cómo varía la probabilidad de que el accidente sea grave en cada tipología identificada.

Para ello, el trabajo plantea los siguientes objetivos específicos:

I. Construir una base de datos analítica que contenga más de 100.000 observaciones y que integre variables contextuales, temporales y personales.

II. Identificar al menos tres clústeres de accidentes a partir de la base de datos construida, segmentando los accidentes en tipologías diferenciadas.

III. Comparar las tasas de severidad entre los clústeres mediante medidas estadísticas de asociación, así como analizar si pertenecer a un determinado clúster influye en la probabilidad de que el accidente sea grave, utilizando para ello un modelo de regresión logística.

IV. Detectar, dentro de cada tipología de accidente, las cinco variables con mayor incidencia sobre la probabilidad de severidad.

V. Desarrollar visualizaciones analíticas, recogidas en una infografía, que plasmen los principales resultados obtenidos en el análisis.

En última instancia, el trabajo busca aportar una base analítica que contribuya a una mejor comprensión de las características asociadas a la severidad de los accidentes de tráfico y que pueda servir de apoyo para el diseño de estrategias preventivas más eficaces en materia de seguridad vial.

1.3. Concepto de gravedad y estado de la cuestión

Para la realización del Trabajo de Fin de Grado se han revisado numerosos estudios sobre siniestralidad vial centrados en la lesividad de los accidentes de tráfico. Esta literatura resulta especialmente relevante para el enfoque del trabajo, pues proporciona tanto un marco conceptual para entender la gravedad de los siniestros como posibles aproximaciones metodológicas para analizarla a partir de un conjunto de datos.

Con carácter previo, es importante delimitar qué se entiende por gravedad o severidad del accidente. Siguiendo las definiciones de la Orden INT/2223/2014, de 27 de octubre aplicadas por la Dirección General de Tráfico (en adelante, DGT), se entenderá por accidente mortal aquel siniestro con víctimas en el que, al menos una de ellas, resulte fallecida, considerando fallecida a toda persona que, como consecuencia del accidente, muera en el acto o dentro de los treinta días siguientes.

Con base en la misma fuente, se entenderá como accidente grave aquel siniestro con víctimas no mortales en el que, al menos, una de las personas implicadas resulte herida con hospitalización superior a veinticuatro horas como consecuencia de dicho siniestro.

En este sentido, se entenderá la gravedad o severidad vial como una medida del nivel de daño personal derivado del accidente, obtenida a partir de la presencia de fallecidos y de heridos graves.

Por su parte, en cuanto a la literatura revisada, en primer lugar destaca la tesis doctoral de Úbeda González (Universidad Miguel Hernández, 2017), centrada en la predicción de la severidad de las lesiones por ocupante a partir de variables del conductor, vehículo y entorno. Su principal aportación para este trabajo radica en demostrar que la severidad vial puede analizarse integrando variables personales y contextuales dentro de un mismo modelo, así como en identificar predictores relevantes como la edad, la maniobra realizada, la hora del día, la luminosidad o el tipo de vía.

En segundo lugar, resulta especialmente útil el trabajo de fin de grado de Herrera Briones (Universidad Politécnica de Madrid, 2021), centrado en la lesividad de colisiones interurbanas entre dos vehículos. Su interés reside en mostrar que el peso explicativo de los factores asociados a la severidad puede variar según el escenario analizado. En particular, destaca la importancia del tipo de colisión y de las infracciones del conductor, al tiempo que encuentra una menor incidencia de la luminosidad y de los factores atmosféricos.

También resulta relevante la tesis doctoral de Febres Eguiguren (Universidad de Burgos, 2021), centrada en la modelización de la lesividad de los accidentes de tráfico en España mediante Redes Bayesianas. Entre sus resultados, destacan el efecto protector del cinturón de seguridad y el aumento de lesividad asociado a la conducción sin licencia o a la distracción tecnológica, así como la conveniencia de profundizar en el impacto del consumo de alcohol y drogas.

Por último, debe mencionarse el artículo de Sanjurjo-de-No, Arenas-Ramírez, Mira y Aparicio-Izquierdo (2020), que identifica patrones de conducta en accidentes de tráfico en España mediante *self-organizing maps.* Su principal aportación radica en respaldar la utilidad de las técnicas no supervisadas para descubrir perfiles diferenciados dentro de la siniestralidad vial, demostrando así que la segmentación resulta útil para identificar tipologías de accidente y distinguir contextos con severidad potencialmente distinta.

En conjunto, la literatura revisada confirma que la severidad de los accidentes viales constituye un fenómeno complejo en el que intervienen factores de diversa índole y que puede abordarse mediante metodologías supervisadas y no supervisadas complementarias.

1.4. Introducción a la metodología

La metodología del trabajo se articula en tres fases sucesivas y complementarias: primero, la recopilación y preparación de datos; segundo, la modelización de la severidad; y tercero, la visualización de resultados.

En la primera fase, se partirá de los microdatos contextuales y temporales de accidentes viales de la DGT, a los que se les integrará determinadas características personales del conductor. A continuación, se realizará un análisis exploratorio de datos (EDA) orientado a evaluar la calidad y consistencia del conjunto de datos, depurar registros, tratar valores ausentes, unificar codificaciones y generar variables útiles para el posterior análisis.

En la segunda fase, se aplicarán herramientas de aprendizaje no supervisado y supervisado para llevar a cabo la modelización. Por un lado, se utilizarán técnicas de *clustering* para identificar tipologías de accidente. Por otro lado, se emplearán medidas estadísticas de asociación y modelos de regresión logística con el objetivo de evaluar el efecto de la pertenencia a cada clúster sobre la probabilidad de severidad del accidente. Adicionalmente, se identificarán los principales *drivers* de severidad dentro de cada tipología vial y, por último, se realizará un estudio descriptivo de la variable de consumo de sustancias con el fin de analizar su prevalencia y distribución en siniestros mortales.

En la tercera fase, se elaborará una infografía analítica que facilite la comprensión de los patrones detectados.

1.5. Estructura

El presente trabajo se compone de seis capítulos.

El primero, que concluye con este epígrafe, introduce el estudio: justificación e interés del tema, objetivos generales y específicos, estado del arte, visión de la metodología y estructura.

El segundo desarrolla el marco metodológico, describiendo las fuentes de información, el proceso de preparación de los datos, la modelización de la severidad vial y la visualización analítica de los resultados.

El tercero recoge el desarrollo técnico del análisis, incluyendo la construcción del *dataset*, el preprocesamiento, la segmentación de accidentes, el estudio de la severidad por clúster, la identificación de sus principales *drivers* y el análisis del consumo de sustancias en siniestros mortales.

El cuarto presenta la discusión final, las conclusiones, las limitaciones del estudio y sus posibles líneas futuras.

Por último, el quinto capítulo contiene la bibliografía y el sexto los anexos.

CAPÍTULO 2. DESARROLLO METODOLÓGICO

2.1. Recopilación y preparación de datos

2.1.1. Recolección de datos

Con el objetivo de construir un modelo analítico de severidad vial basado en factores humanos y del entorno, el primer paso consiste en la recolección de los datos necesarios para el análisis. Para ello, se buscan fuentes de información fiables, completas y relevantes que permitan describir las condiciones en las que se producen los siniestros. La calidad, cobertura y coherencia de los datos recolectados resulta imprescindible para la posterior validez y veracidad de los resultados obtenidos.

La elección de las fuentes se fundamenta en dos criterios fundamentales: la fiabilidad y cobertura de las fuentes, que garantizan la procedencia de organismos oficiales, la continuidad temporal de los registros y la homogeneidad geográfica de los datos; y la relevancia analítica, que asegura que las variables incluidas reflejan los factores que inciden en la siniestralidad vial.

De acuerdo con estos criterios, las fuentes de información empleadas, que se explicarán en el apartado correspondiente del desarrollo técnico, son las siguientes:

I. DGT – "Ficheros de microdatos de accidentes con víctimas."

II. DGT – "Tablas estadísticas de accidentes."

III. Ministerio de Justicia, Gobierno de España – "Hallazgos toxicológicos en víctimas mortales de accidentes de tráfico."

Asimismo, durante la fase de exploración de fuentes adicionales, se analizó otro *dataset* que finalmente fue descartado por cuestiones de pertinencia y calidad de la información. Este *dataset*, también explicado en el apartado correspondiente del desarrollo técnico, es el siguiente: DGT – "Tablas estadísticas de accidentes relativas a la conducción sin permiso y a la comisión de infracciones."

Finalmente, en cuanto a la temporalidad de la información utilizada, los datos recolectados abarcan el periodo desde 2021 hasta 2023, al tratarse del intervalo temporal disponible más reciente en las tres fuentes empleadas.

Desde un punto de vista analítico, este rango ofrece un volumen de observaciones suficiente para el modelado, al incluir más de 250.000 registros de siniestros. Desde una perspectiva práctica, este periodo es adecuado por reflejar un parque automovilístico, un estado de la red viaria y unos patrones de movilidad comparables a los actuales, evitando distorsiones por cambios normativos o tecnológicos significativos.

2.1.2. Características y fusión de los datos

La construcción del *dataset* final requiere integrar tres fuentes de información de naturaleza y granularidad distintas. Desde un punto de vista metodológico, esta integración sirve para capturar de forma conjunta las dimensiones temporal, contextual y humana de los siniestros, dado que ninguna de las fuentes disponibles permite, por sí sola, abarcar todos los factores relevantes para la modelización del riesgo. Cada una de ellas aporta un conocimiento necesario, por lo que su combinación constituye un paso crítico dentro de la metodología del trabajo.

En relación con la primera fuente, los **ficheros de microdatos de la DGT** conforman la base estructural del *dataset*, al proporcionar la unidad mínima de análisis del estudio: el siniestro vial individual ocurrido en España. Estos datos, que se presentan en formato individualizado, permiten describir las circunstancias del accidente, otorgando un nivel de detalle imprescindible para la modelización y una cobertura completa del territorio y del período temporal analizado.

Estos ficheros incluyen 74 variables relativas al entorno vial y al contexto de ocurrencia del siniestro, resumidas en la siguiente tabla:

Tabla 1: Variables incluidas en los microdatos de la DGT

Tipo de variables	Ejemplos de variables
Momento temporal	Año, mes, día de la semana, hora del siniestro
Localización	Provincia, municipio, isla, zona, carretera, punto kilométrico
Tipología de la vía y del accidente	Tipo y titularidad de la vía, configuración de la intersección, maniobra previa, tipo de colisión, situación del siniestro.
Condiciones físico-ambientales de la vía	Estado del firme, condiciones de circulación, factores meteorológicos (lluvia, niebla, viento), visibilidad y condiciones de iluminación
Consecuencias del accidente	Número total de víctimas diferenciando su gravedad (ilesos, heridos leves, heridos graves y fallecidos).

Fuente: Elaboración propia

Desde el punto de vista técnico, la única transformación necesaria para su preparación consiste en integrar los ficheros correspondientes 2021, 2022 y 2023, ya que todos ellos comparten idéntica estructura y codificación.

Por su parte, en relación con la segunda fuente, **las tablas estadísticas de la DGT** no presentan datos individualizados, sino distribuciones agregadas de las características personales de los implicados en siniestros viales. Entre ellas, estas tablas recogen información relativa a tres variables que la literatura identifica como especialmente relevantes en el análisis de severidad de los accidentes de tráfico: la edad, el sexo y la antigüedad del permiso de conducir del conductor.

Dado que estas variables no aparecen desagregadas, su incorporación requiere un proceso de generación de microdatos sintéticos que permita asignar a cada observación individual un perfil personal coherente con las distribuciones reales publicadas.

De esta manera, si bien el uso de estas variables sintéticas permite incorporar en el análisis factores personales relevantes no disponibles en los microdatos originales, también puede introducir determinados riesgos metodológicos que deben mencionarse y considerarse en la interpretación de los resultados del análisis. La Figura 1 resume las principales implicaciones derivadas de su utilización en el presente trabajo.

Figura 1: Alcance metodológico del uso de variables sintéticas

Fuente: Elaboración propia

Por último, en relación con la tercera fuente, el *dataset* de **"La justicia en datos",** que recoge información sobre los siniestros mortales ocurridos desde 2021 hasta 2023, se utiliza para introducir una variable específica relativa al consumo de alcohol, drogas u otras sustancias psicoactivas por parte del conductor.

Si bien los datos disponibles están individualizados, su tamaño muestral, de 3.483 observaciones, es considerablemente inferior al de los microdatos de la DGT, pues este *dataset* únicamente contiene datos de los siniestros mortales. Es por ello por lo que no es metodológicamente viable realizar una integración a nivel de observación individual.

En su lugar, considerando que ambas fuentes comparten un conjunto de variables relevantes tanto del entorno como del conductor, se estiman patrones agregados de consumo condicionados a estas variables comunes, incorporando esta información al *dataset* final. En concreto, las variables comunes utilizadas para este propósito son el año del accidente, el tipo de vía, el sexo y el rango de edad del conductor.

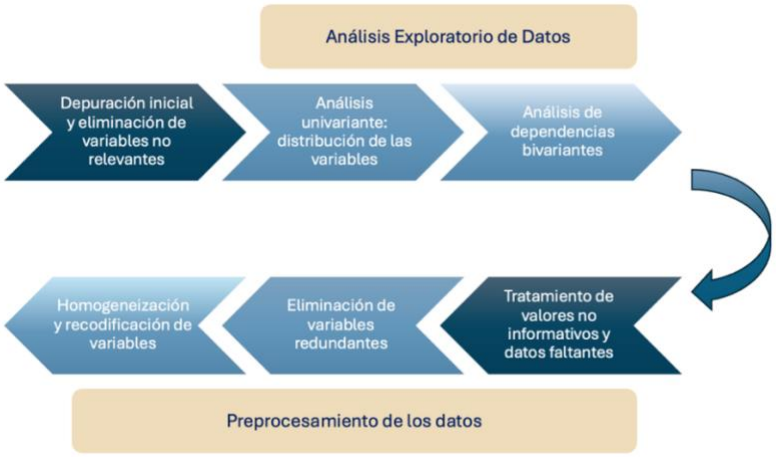
No obstante, debe señalarse que esta variable no procede del mismo universo estadístico que la base principal y ha sido incorporada mediante un procedimiento de estimación sintética, lo que introduce posibles sesgos y limitaciones en términos de representatividad. Por este motivo, si bien su inclusión se justifica por la especial relevancia que el consumo de sustancias presenta como factor asociado a la severidad vial, sus resultados deben interpretarse con cautela y no permiten extraer conclusiones determinantes en el análisis posterior.

2.1.3. Preprocesamiento y análisis exploratorio de los datos

Una vez integrado el *dataset* final, se lleva a cabo una fase de preprocesamiento y análisis exploratorio de los datos. Este apartado tiene como objetivos obtener una primera comprensión de la estructura, calidad y patrones del conjunto de datos y, simultáneamente, preparar dicho conjunto para su posterior utilización en el modelo.

El procedimiento seguido se articula en los seis pasos enumerados a continuación.

Figura 2: EDA y preprocesamiento de los datos



Fuente: Elaboración propia

I. **Depuración inicial y eliminación de variables** que no aportan información significativa para la posterior modelización.

II. **Análisis univariante** con el fin de describir la distribución de cada variable de forma individual. Así, se verifica la coherencia de los rangos, se detectan posibles valores anómalos y se comprende la estructura general de los datos.

III. **Análisis exploratorio de dependencias** entre pares de variables con el objetivo de explorar relaciones básicas y evaluar la coherencia interna del *dataset*. Ello permite detectar posibles inconsistencias en los datos y validar su calidad, asegurando que no se presenten combinaciones incompatibles (por ejemplo, rango de edad de 18-24 años y más de 10 años de antigüedad del permiso de conducir).

IV. **Tratamiento de valores no informativos y datos faltantes.**

V. **Detección y eliminación de variables redundantes,** que no aportan información adicional y pueden generar problemas de multicolinealidad.

VI. **Homogeneización y recodificación de variables.** Se verifica que cada variable presente un tipo consistente con su significado; las variables binarias se mantienen en formato 0/1, las ordinales se conservan con una codificación ordenada y las categóricas nominales se preparan para una codificación posterior que evite la imposición de órdenes artificiales entre categorías.

2.2. Modelización de la severidad vial

La segunda fase del trabajo tiene como finalidad identificar tipologías diferenciadas de siniestro, estudiar si dichas tipologías presentan diferencias en términos de gravedad y, seguidamente, profundizar en los factores que explican la severidad dentro de cada contexto identificado. Para ello, se combinan técnicas de aprendizaje no supervisado y modelos supervisados.

2.2.1. Segmentación de los accidentes de tráfico

El primer bloque de la modelización se basa en el uso de técnicas de aprendizaje no supervisado, en particular, *clustering*, para descubrir grupos de accidentes con características similares y distinguir tipologías homogéneas de siniestro a partir de las variables previamente tratadas. Para ello, se parte del *dataset* completo preprocesado, compuesto por 45 variables.

Antes de la segmentación, se escalan las variables mediante *StandardScaler()* y se reduce la dimensionalidad mediante Análisis de Componentes Principales (PCA), reteniendo aproximadamente el 85% de la varianza explicada. Así, se evitan redundancias entre variables y se mejora la estabilidad y eficiencia del análisis.

Sobre esta representación reducida se aplica en primer lugar un *clustering* particional mediante *K-means*, seleccionando el número de clústeres a partir del método del codo y del índice *silhouette*. Finalmente, se adopta una solución de 4 clústeres, que permite segmentar el conjunto de accidentes en escenarios diferenciados e interpretables.

Una vez completado el *clustering* general y dado que uno de los grupos presenta un tamaño muy superior al resto, se lleva a cabo una segunda segmentación centrada en el clúster mayoritario, interpretado como aquel que representa las condiciones normales de circulación.

2.2.2. Caracterización de la severidad por clúster

A partir de los clústeres ya identificados y perfilados, se analiza la severidad de los accidentes en cada grupo con el objetivo de evaluar si la segmentación obtenida a partir de variables temporales, contextuales y personales presenta también diferencias en términos de gravedad. Este análisis, de carácter descriptivo, permite establecer una primera aproximación comparativa entre los distintos clústeres.

Para ello, la información relativa al número de fallecidos y heridos graves se integra en la estructura original del microdato. A partir de estas nuevas variables, se construyen distintos indicadores de severidad: primero, se definen dos indicadores binarios, accidente mortal y accidente grave, en coherencia con las definiciones previamente establecidas en el capítulo introductorio; segundo, se crea un índice sintético de severidad que asigna a las muertes un peso cinco veces superior al de los heridos graves, al considerarse que representan un desenlace sustancialmente más grave y, por tanto, deben tener una mayor incidencia en la medida final.

Con base en estas variables, el análisis se articula mediante tablas de indicadores por clúster. Para cada grupo, se calcula el número total de accidentes, el número de fallecidos y heridos graves, la proporción de accidentes mortales y graves, así como la media de estos resultados por accidente y el valor medio del índice sintético de severidad.

Esta metodología permite comparar la gravedad entre clústeres desde una doble perspectiva: por un lado, en términos absolutos, a través del volumen total de víctimas; y, por otro, en términos relativos, mediante medidas promedio que corrigen las diferencias en el tamaño de los grupos analizados.

Tabla 2: Indicadores de severidad de los accidentes por clúster

CLUSTER	Nº de accidentes	Nº total de fallecidos	Nº total de heridos graves	Accidentes mortales (%)	Accidentes graves (%)	Media de fallecidos por accidente	Media de heridos graves por accidente	Índice medio de severidad
0								
1								
2								
3								

Fuente: Elaboración propia (con Python)

Este mismo procedimiento se replica en el *subclustering* del grupo de condiciones normales para analizar si, incluso dentro de un mismo contexto ambiental, la severidad cambia en función de la composición de los subclústeres.

En síntesis, este análisis descriptivo permite vincular los tipos de siniestros con sus resultados en términos de gravedad, constituyendo la base para el posterior análisis sobre la relación entre pertenencia al clúster y probabilidad de severidad.

2.2.3. Efecto del clúster sobre la probabilidad de severidad

A continuación, se desarrolla una fase de análisis inferencial orientada a evaluar si la pertenencia a un determinado grupo se asocia con una mayor probabilidad de gravedad del accidente.

Para ello, se construye una variable binaria de gravedad, denominada SEVERIDAD, que toma el valor 1 cuando el accidente presenta al menos un fallecido o herido grave, y el valor 0 en caso contrario. A partir de esta variable, se calcula la tasa de severidad para cada clúster y subclúster, entendida como la proporción de accidentes graves o mortales dentro de cada grupo.

Después, se obtienen intervalos de confianza del 95% y se construyen tablas de contingencia entre las variables CLUSTER (y SUBCLUSTER) y SEVERIDAD, aplicando una prueba Chi-cuadrado de independencia para contrastar la existencia de asociación estadísticamente significativa entre la pertenencia al clúster y la probabilidad de accidente grave.

Complementariamente, se estima un modelo de regresión logística para evaluar el efecto de la pertenencia a cada clúster sobre la probabilidad de severidad y cuantificar la contribución relativa de los distintos escenarios identificados al riesgo de accidente grave.

Este procedimiento permite verificar si la segmentación obtenida resulta relevante no solo desde una perspectiva descriptiva, sino también en términos probabilísticos y explicativos de la gravedad del siniestro.

2.2.4. Identificación de los determinantes de severidad en cada clúster

Una vez evaluado el efecto de la pertenencia a los distintos clústeres sobre la probabilidad de severidad, se desarrolla un análisis orientado a detectar los principales factores o *drivers* asociados a la ocurrencia de accidentes graves o mortales dentro de cada tipología.

Para ello, se estiman modelos de regresión logística con penalización L1, útiles para la selección de variables en presencia de un elevado número de predictores binarios.

A partir de los coeficientes estimados, se calculan los correspondientes *odds ratio*, así como indicadores adicionales de prevalencia, intensidad e impacto. En concreto, el *odds ratio* mide cómo varía la probabilidad de severidad cuando la variable está presente; la prevalencia recoge la frecuencia relativa con la que dicha característica aparece dentro del grupo analizado; la intensidad refleja la magnitud del efecto estimado, medida a través del valor absoluto del logaritmo del *odds ratio*; y el impacto combina prevalencia e intensidad, permitiendo priorizar como *drivers* aquellos factores que combinan mayor frecuencia de aparición con mayor capacidad explicativa de severidad.

Adicionalmente, se calcula el efecto marginal de los principales *drivers* sobre la probabilidad de severidad dentro de cada clúster y subclúster, comparando la probabilidad estimada de accidente grave en presencia y ausencia de cada característica.

Para terminar, todo ello se completa con una simulación *what-if* orientada a estimar cómo variaría la severidad esperada ante la hipotética reducción en la prevalencia del factor más explicativo de la severidad. De este modo, se obtiene una aproximación cuantitativa del potencial impacto que podrían tener determinadas medidas de prevención y seguridad vial.

2.2.5. Análisis del consumo de sustancias en siniestros mortales

Como último bloque, se realiza un análisis descriptivo de la prevalencia del consumo de sustancias en conductores implicados en siniestros mortales, con el objetivo de identificar su distribución en función de variables como el sexo, grupo de edad y tipo de vía.

A diferencia de las secciones previas, en este apartado no se utiliza el *dataset* empleado hasta ahora, sino el *dataset* original procedente del portal oficial "Justicia en datos", del que se extrajo previamente la información necesaria para construir la variable sintética de consumo de sustancias.

Aunque este *dataset* presenta un tamaño muestral inferior, con 3.483 observaciones, su utilización resulta metodológicamente más adecuada; se trata de una fuente institucional primaria que recoge los resultados toxicológicos asociados a siniestros mortales, lo que garantiza la fiabilidad y trazabilidad de la información analizada y reduce posibles distorsiones derivadas del uso de variables sintéticas.

Si bien se es consciente de las limitaciones derivadas del tamaño de la muestra y de la ausencia de acceso a registros más amplios de esta naturaleza, se considera que, dentro del alcance del presente trabajo, esta constituye la fuente pública más adecuada disponible para abordar el análisis de esta variable.

En este apartado, por tanto, se construye una variable binaria de consumo positivo a partir del resultado toxicológico para estimar la prevalencia global y su distribución por provincia, tipo de vía, sexo y grupo de edad. Así las cosas, se identifican perfiles combinados de riesgo mediante el análisis conjunto de estas variables.

2.3. Visualización analítica de los patrones de severidad

Para terminar, se desarrolla, mediante la herramienta *Canva,* una infografía que sintetiza de forma clara e integrada los principales resultados obtenidos a lo largo del estudio, favoreciendo así su lectura conjunta y su comprensión global.

Esta infografía se plantea como una herramienta dirigida a administraciones públicas y organismos responsables de la seguridad vial, pues pretende dar apoyo en la planificación de políticas públicas orientadas a la reducción de la siniestralidad vial grave y mortal. De esta forma, esta fase final no solo cumple una función de síntesis expositiva dentro del trabajo, sino que refuerza la utilidad práctica del análisis realizado.

CAPÍTULO 3. DESARROLLO TÉCNICO

3.1. Recopilación y preparación de datos

3.1.1. Recolección de datos

De acuerdo con los criterios metodológicos definidos en el capítulo anterior, el presente trabajo se apoya en tres fuentes de información que permiten capturar tanto las características del entorno vial como determinados factores humanos relevantes para el análisis de su severidad. Estas son:

I. **DGT – Ficheros de microdatos de accidentes con víctimas.** Contiene registros individualizados de siniestros viales en España, con más de cuarenta variables relativas a las condiciones del entorno, ya resumidas en la Tabla 1. Ha sido elegida por su fiabilidad institucional y por su nivel de detalle.

II. **DGT – Tablas estadísticas de accidentes (datos agregados).** Aporta información centrada en los factores humanos, incluyendo variables como el sexo y rango de edad del individuo y la antigüedad de su permiso de conducir.

III. **Ministerio de Justicia, Gobierno de España – "Hallazgos toxicológicos en víctimas mortales de accidentes de tráfico".** Se utiliza como fuente auxiliar para incorporar información relativa al consumo de sustancias en accidentes con víctimas mortales, permitiendo aproximar la influencia de la conducción bajo los efectos de sustancias en la gravedad de la siniestralidad vial.

Tabla 3: Resumen de las fuentes de información empleadas en el estudio

Fuente	Organismo	Información aportada	Tipo de variables
Microdatos de accidentes con víctimas	Dirección General de Tráfico (DGT)	Variables del entorno en registros individualizados	Año, mes, día de la semana y hora; localización geográfica; tipo y características de la vía; condiciones meteorológicas, de iluminación y del firme
Tablas estadísticas de accidentes	Dirección General de Tráfico (DGT)	Variables humanas en registros agregados	Sexo; rango de edad; tipo de vehículo; antigüedad del permiso de conducir
Hallazgos toxicológicos en víctimas mortales de accidentes de tráfico	Ministerio de Justicia	Resultados toxicológicos en víctimas mortales de siniestros viales	Presencia de alcohol, drogas u otras sustancias psicoactivas

Fuente: Elaboración propia

Por su parte, también se evaluaron otras fuentes de información, que finalmente fueron descartadas por no resultar adecuadas para los objetivos del análisis. A continuación, se detalla una de ellas y los motivos que justifican su exclusión.

I. **DGT – Tablas estadísticas de accidentes relativas a la conducción sin permiso y a la comisión de infracciones.** Inicialmente, se pretendía extraer información sobre la conducción sin permiso y comisión de infracciones, para valorar su influencia en la gravedad del siniestro vial, que Febres Eguiguren consideraba de especial relevancia en su tesis doctoral. Sin embargo, la primera variable se descartó por su escasa relevancia estadística, ya que los accidentes en los que interviene un conductor sin permiso no llegan a representar el 2% del total, según los datos del año 2023; y la segunda, relativa a la comisión de infracciones, fue igualmente excluida por presentar un alto grado de sesgo, dado que, también en 2023, en el 45,6% de casos la categoría registrada es "se desconoce", lo que compromete la validez de cualquier análisis posterior.

Tabla 4: Edad y sexo de los conductores implicados en accidentes con víctimas en función de la antigüedad de su permiso de conducir, 2023

EN VÍAS INTERURBANAS	SEXO	De 0 a 13 años	De 14 a 17 años	De 18 a 24 años	De 25 a 44 años	De 45 a 64 años	De 65 a 74 años	Más de 74 años	Se desconoce	Total	
Sin permiso	Hombre	0	26	67	223	88	9	4	1	418	
	Mujer	0	0	10	24	16	0	2	0	52	
	Se desconoce	0	0	0	0	0	0	0	1	1	
	Total	0	26	77	247	104	9	6	2	471	0,77% del total
Con permiso o desconocido	Hombre	27	397	4.664	16.523	17.300	3.309	1.948	215	44.383	
	Mujer	6	62	2.042	7.460	5.423	732	255	45	16.025	
	Se desconoce	0	0	4	1	8	0	0	294	307	
	Total	33	459	6.710	23.984	22.731	4.041	2.203	554	60.715	
EN VÍAS URBANAS	SEXO	De 0 a 13 años	De 14 a 17 años	De 18 a 24 años	De 25 a 44 años	De 45 a 64 años	De 65 a 74 años	Más de 74 años	Se desconoce	Total	
Sin permiso	Hombre	2	45	345	940	593	88	49	119	2.181	
	Mujer	2	4	117	326	208	18	6	20	701	
	Se desconoce	0	2	0	4	1	0	0	1	8	
	Total	4	51	462	1.270	802	106	55	140	2.890	2,65% del total
Con permiso o desconocido	Hombre	267	1.611	9.214	28.375	26.554	4.407	2.373	2.458	75.259	
	Mujer	81	388	3.404	13.450	9.703	1.076	421	902	29.425	
	Se desconoce	6	11	46	99	63	4	9	1.312	1.550	
	Total	354	2.010	12.664	41.924	36.320	5.487	2.803	4.672	106.234	
EN VÍAS URBANAS E INTERURBANAS											
Con permiso o desconocido										166.949	
Sin permiso										3.361	1,97% del total

Fuente: Dirección General de Tráfico (DGT, 2024)

Tabla 5: Infracciones de conductores implicados en los accidentes con víctimas, 2023

TIPO DE INFRACCIÓN	Total	
Infracción de velocidad	4.541	
Marcha lenta entorpeciendo la circulación	42	
Ninguna infracción de velocidad	28.098	
Se desconoce	28.367	46,47% del total
Total infracciones de velocidad	**61.048**	
No respetar señal de STOP	1.188	
No respetar paso para peatones	145	
No respetar otra regulación de prioridad	2.710	
Circular en sentido contrario o por lugar prohibido	245	
Invadir parcialmente el sentido contrario	1.688	
Adelantar antirreglamentariamente	558	
No mantener el intervalo de seguridad	5.910	
Otra infracción	1.105	
Ninguna infracción	23.305	
Se desconoce	24.194	39,63% del total
Total infracciones del conductor	**61.048**	
Apertura de puertas sin precaución	9	
Ninguna infracción	32.124	
Se desconoce	28.915	47,36% del total
Total infracciones de apertura de puerta	**61.048**	
Incorrecta utilización del alumbrado	85	
Ninguna infracción	32.124	
Se desconoce	28.839	47,24% del total
Total infracciones de alumbrado	**61.048**	
Exceso, mal acondicionamiento o desprendimiento de la carga	77	
Ninguna infracción	32.124	
Se desconoce	28.847	47,25% del total
Total infracciones de carga del vehículo	**61.048**	
Promedio de observaciones en "Se desconoce"	**27.832**	**45,59% del total**

Fuente: Dirección General de Tráfico (DGT, 2024)

3.1.2. Características y fusión de los datos

Una vez seleccionadas las fuentes, se construye un *dataset* único que combine información contextual procedente de los microdatos de la DGT con variables personales generadas de forma sintética a partir de las distribuciones estadísticas oficiales.

(i) Integración de los ficheros de microdatos de la DGT

En primer lugar, se procede a la integración de los ficheros de microdatos de accidentes con víctimas correspondientes a los años 2021, 2022 y 2023. Estos ficheros presentan idéntica estructura, codificación y definición de variables, lo que permite su concatenación directa sin necesidad de transformaciones previas.

Desde el punto de vista técnico, el proceso consiste en la carga individual de cada fichero y su posterior concatenación en un único conjunto de datos. Como resultado, se obtiene una base consolidada con 289.084 registros individuales, cada uno de los cuales representa un siniestro vial ocurrido en el periodo analizado.

El código desarrollado en Python para efectuar este primer paso se incluye en el Anexo 1 del presente trabajo.

(ii) Creación e integración de variables personales sintéticas

Dado que los microdatos de la DGT no incluyen información individualizada sobre las características personales del conductor, se procede, a partir de tablas estadísticas agregadas, a la generación de variables sintéticas relativas a la antigüedad del permiso de conducir, edad y sexo de los conductores implicados en siniestros viales, cuyo alcance y limitaciones ya han sido explicados en la sección metodológica.

En este sentido, se utilizan, para cada año (2021, 2022 y 2023), dos tablas de la DGT, una relativa a vías urbanas y otra a interurbanas, tituladas "Edad y sexo de los conductores implicados en accidentes con víctimas en función de la antigüedad del permiso de conducir". Para ilustrar la estructura de estas tablas, a continuación se muestra un extracto correspondiente a la tabla del año 2023 en vías urbanas:

Tabla 6: Edad y sexo de los conductores implicados en accidentes con víctimas en función de la antigüedad de su permiso de conducir, en vías urbanas. Año 2023.

ANTIGÜEDAD PERMISO	SEXO	De 0 a 13 años	De 14 a 17 años	De 18 a 24 años	De 25 a 44 años	De 45 a 64 años	De 65 a 74 años	Más de 74 años	Se desconoce	Total
1 año o menos	Hombre	0	240	731	385	85	10	5	0	1.456
	Mujer	0	38	246	127	44	1	0	0	456
	Se desconoce	0	0	2	2	2	0	0	0	6
	Total	**0**	**278**	**979**	**514**	**131**	**11**	**5**	**0**	**1.918**
1 año a 2 años	Hombre	0	316	1.356	732	132	7	2	0	2.545
	Mujer	0	44	392	215	42	2	0	0	695
	Se desconoce	0	0	1	2	1	0	0	0	4
	Total	**0**	**360**	**1.749**	**949**	**175**	**9**	**2**	**0**	**3.244**
2 años a 3 años	Hombre	0	105	1.170	762	146	5	1	0	2.189
	Mujer	0	25	415	275	48	2	1	0	766
	Se desconoce	0	0	3	2	0	0	0	0	5
	Total	**0**	**130**	**1.588**	**1.039**	**194**	**7**	**2**	**0**	**2.960**
3 años a 4 años	Hombre	0	18	686	576	140	3	2	0	1.425
	Mujer	0	7	270	220	24	3	1	0	525
	Se desconoce	0	0	1	1	1	0	0	0	3
	Total	**0**	**25**	**957**	**797**	**165**	**6**	**3**	**0**	**1.953**
4 años a 5 años	Hombre	0	0	696	668	145	7	3	0	1.519
	Mujer	0	0	284	296	41	2	1	0	624
	Se desconoce	0	0	1	0	0	0	0	0	1
	Total	**0**	**0**	**981**	**964**	**186**	**9**	**4**	**0**	**2.144**
5 años a 6 años	Hombre	0	0	556	690	159	9	3	0	1.417
	Mujer	0	0	209	330	57	1	0	0	597
	Se desconoce	0	0	0	0	0	0	0	0	0
	Total	**0**	**0**	**765**	**1.020**	**216**	**10**	**3**	**0**	**2.014**
6 años a 7 años	Hombre	0	0	217	726	129	5	1	0	1.078
	Mujer	0	0	89	333	36	1	1	0	460
	Se desconoce	0	0	1	1	0	0	0	0	2
	Total	**0**	**0**	**307**	**1.060**	**165**	**6**	**2**	**0**	**1.540**
7 años a 8 años	Hombre	0	0	79	856	92	2	0	0	1.029
	Mujer	0	0	23	390	41	2	0	0	456
	Se desconoce	0	0	0	0	0	0	0	0	0
	Total	**0**	**0**	**102**	**1.246**	**133**	**4**	**0**	**0**	**1.485**
8 años a 9 años	Hombre	0	0	24	818	74	4	1	0	921
	Mujer	0	0	2	400	42	3	2	0	449
	Se desconoce	0	0	0	2	0	0	0	0	2
	Total	**0**	**0**	**26**	**1.220**	**116**	**7**	**3**	**0**	**1.372**
9 años a 10 años	Hombre	0	0	12	801	103	9	2	0	927
	Mujer	0	0	1	428	44	1	1	0	475
	Se desconoce	0	0	0	0	1	0	0	0	1
	Total	**0**	**0**	**13**	**1.229**	**148**	**10**	**3**	**0**	**1.403**
Más de 10 años	Hombre	0	0	2	11.231	16.346	2.831	1.531	0	31.941
	Mujer	0	0	0	5.849	6.310	734	291	0	13.184
	Se desconoce	0	0	0	21	17	2	5	0	45
	Total	**0**	**0**	**2**	**17.101**	**22.673**	**3.567**	**1.827**	**0**	**45.170**
Sin especificar	Hombre	269	977	4.030	11.070	9.596	1.603	871	2.577	30.993
	Mujer	83	278	1.590	4.913	3.182	342	129	922	11.439
	Se desconoce	6	13	37	72	42	2	4	1.313	1.489
	Total	**358**	**1.268**	**5.657**	**16.055**	**12.820**	**1.947**	**1.004**	**4.812**	**43.921**
Total Antigüedad	Hombre	269	1.656	9.559	29.315	27.147	4.495	2.422	2.577	77.440
	Mujer	83	392	3.521	13.776	9.911	1.094	427	922	30.126
	Se desconoce	6	13	46	103	64	4	9	1.313	1.558
	Total	**358**	**2.061**	**13.126**	**43.194**	**37.122**	**5.593**	**2.858**	**4.812**	**109.124**

Fuente: Dirección General de Tráfico (DGT, 2024)

En una primera fase, estas tablas fueron integradas en una única hoja de trabajo, homogeneizando las categorías y eliminando los registros catalogados como desconocidos. En cuanto a estos últimos, su peso relativo fue analizado para determinar el impacto de su exclusión.

En las variables sexo y edad, los valores desconocidos representan únicamente entre el 1 % y 3 % del total anual, de tal forma que su eliminación no modifica significativamente las distribuciones. En cambio, en la variable antigüedad del permiso, los valores sin especificar alcanzan un porcentaje superior: el 28% de los datos. En este sentido, si bien se planteó la posibilidad de conservarlos, la alternativa sería redistribuirlos proporcionalmente entre las categorías conocidas, lo que, al trabajar con probabilidades normalizadas, produciría idénticos resultados en la tabla final. Por ello, la eliminación de los registros sin especificar se consideró metodológicamente coherente para el análisis.

Figura 3: Análisis del peso de los valores desconocidos en las variables personales (2021-2023)

Fuente: Elaboración propia

A continuación, se calculan las probabilidades conjuntas de cada combinación de antigüedad del permiso, sexo y tramo de edad, condicionadas al tipo de vía (urbana/interurbana) y al año correspondiente.

Como resultado, se obtiene una tabla final de probabilidades que recoge, para cada año y tipo de vía, la distribución empírica completa de las tres variables personales.

Tabla 7: Extracto de la tabla de probabilidades conjuntas de edad, sexo y antigüedad del permiso por tipo de vía y año (2021-2023) – Primeras 25 filas

Año	Antigüedad a	Sexo s	Edad e	N(urbana,a,s,e)	P(e | a,s,urbana)	N(interurbana,a,s,e)	P(e | a,s,interurbana)
2021	1 año o menos	Hombre	0-13 años	0	**0,0000**	0	**0,0000**
2021	1 año o menos	Hombre	14-17 años	276	**0,1021**	73	**0,0551**
2021	1 año o menos	Hombre	18-24 años	1.091	**0,4035**	562	**0,4242**
2021	1 año o menos	Hombre	24-44 años	545	**0,2016**	280	**0,2113**
2021	1 año o menos	Hombre	45-64 años	98	**0,0362**	41	**0,0309**
2021	1 año o menos	Hombre	65-74 años	8	**0,0030**	2	**0,0015**
2021	1 año o menos	Hombre	Más de 74 años	1	**0,0004**	0	**0,0000**
2021	1 año o menos	Mujer	0-13 años	0	**0,0000**	0	**0,0000**
2021	1 año o menos	Mujer	14-17 años	52	**0,0192**	20	**0,0151**
2021	1 año o menos	Mujer	18-24 años	395	**0,1461**	227	**0,1713**
2021	1 año o menos	Mujer	24-44 años	193	**0,0714**	104	**0,0785**
2021	1 año o menos	Mujer	45-64 años	36	**0,0133**	14	**0,0106**
2021	1 año o menos	Mujer	65-74 años	4	**0,0015**	0	**0,0000**
2021	1 año o menos	Mujer	Más de 74 años	0	**0,0000**	2	**0,0015**
2021	1 a 2 años	Hombre	0-13 años	0	**0,0000**	0	**0,0000**
2021	1 a 2 años	Hombre	14-17 años	310	**0,0892**	114	**0,0684**
2021	1 a 2 años	Hombre	18-24 años	1.266	**0,3642**	626	**0,3758**
2021	1 a 2 años	Hombre	24-44 años	857	**0,2465**	386	**0,2317**
2021	1 a 2 años	Hombre	45-64 años	165	**0,0475**	76	**0,0456**
2021	1 a 2 años	Hombre	65-74 años	8	**0,0023**	2	**0,0012**
2021	1 a 2 años	Hombre	Más de 74 años	5	**0,0014**	0	**0,0000**
2021	1 a 2 años	Mujer	0-13 años	0	**0,0000**	0	**0,0000**
2021	1 a 2 años	Mujer	14-17 años	65	**0,0187**	17	**0,0102**
2021	1 a 2 años	Mujer	18-24 años	481	**0,1384**	307	**0,1843**
2021	1 a 2 años	Mujer	24-44 años	269	**0,0774**	122	**0,0732**

Fuente: Elaboración propia

En una segunda fase, esta tabla consolidada de probabilidades fue importada a Python para ejecutar el proceso de asignación sintética. Así, para cada registro de microdatos se seleccionó una combinación de antigüedad, sexo y edad mediante muestreo aleatorio ponderado, utilizando como pesos las probabilidades previamente calculadas para el año y la vía asociados al accidente.

(iii) Integración de la variable de consumo de sustancias

Por último, se incorpora al *dataset* la variable relativa al consumo de alcohol, drogas u otras sustancias psicoactivas, cuyas limitaciones metodológicas ya han sido explicadas con anterioridad y deben tenerse en cuenta en la interpretación del análisis posterior.

Si bien la diferencia de tamaños muestrales no permite una integración directa observación a observación, ambas fuentes comparten un conjunto de variables comunes, como son el año del accidente, el tipo de vía, el sexo y el rango de edad del conductor, por lo que se realiza una integración indirecta basada en la estimación de probabilidades condicionadas a dichas características.

Así las cosas, se construyen distribuciones empíricas de probabilidad que recogen, para cada combinación de año, tipo de vía, sexo y tramo de edad, la frecuencia observada de consumo de sustancias.

Tabla 8: Tabla de probabilidades de consumo de sustancias según año, tipo de vía, sexo y rango de edad

Tabla de probabilidades: variable consumo de sustancias						
Año	Tipo de vía	Sexo	Edad	Nº casos	Nº positivos	Probabilidad consumo
2021	Interurbana	Hombre	18-24 años	74	47	63,51%
2021	Interurbana	Hombre	24-44 años	197	118	59,90%
2021	Interurbana	Hombre	45-64 años	223	120	53,81%
2021	Interurbana	Hombre	Más de 65 años	82	29	35,37%
2021	Interurbana	Mujer	18-24 años	19	9	47,37%
2021	Interurbana	Mujer	24-44 años	31	14	45,16%
2021	Interurbana	Mujer	45-64 años	35	11	31,43%
2021	Interurbana	Mujer	Más de 65 años	22	9	40,91%
2021	Urbana	Hombre	18-24 años	14	9	64,29%
2021	Urbana	Hombre	24-44 años	40	28	70,00%
2021	Urbana	Hombre	45-64 años	53	28	52,83%
2021	Urbana	Hombre	Más de 65 años	29	6	20,69%
2021	Urbana	Mujer	18-24 años	1	1	100,00%
2021	Urbana	Mujer	24-44 años	3	1	33,33%
2021	Urbana	Mujer	45-64 años	8	4	50,00%
2021	Urbana	Mujer	Más de 65 años	13	4	30,77%
2022	Interurbana	Hombre	18-24 años	57	26	45,61%
2022	Interurbana	Hombre	24-44 años	233	145	62,23%
2022	Interurbana	Hombre	45-64 años	232	119	51,29%
2022	Interurbana	Hombre	Más de 65 años	106	42	39,62%
2022	Interurbana	Mujer	18-24 años	18	11	61,11%
2022	Interurbana	Mujer	24-44 años	29	16	55,17%
2022	Interurbana	Mujer	45-64 años	36	17	47,22%
2022	Interurbana	Mujer	Más de 65 años	35	10	28,57%
2022	Urbana	Hombre	18-24 años	22	13	59,09%
2022	Urbana	Hombre	24-44 años	53	41	77,36%
2022	Urbana	Hombre	45-64 años	36	23	63,89%
2022	Urbana	Hombre	Más de 65 años	46	17	36,96%
2022	Urbana	Mujer	18-24 años	0	0	62,5%*
2022	Urbana	Mujer	24-44 años	9	3	33,33%
2022	Urbana	Mujer	45-64 años	12	5	41,67%
2022	Urbana	Mujer	Más de 65 años	26	9	34,62%
2023	Interurbana	Hombre	18-24 años	59	28	47,46%
2023	Interurbana	Hombre	24-44 años	185	120	64,86%
2023	Interurbana	Hombre	45-64 años	237	132	55,70%
2023	Interurbana	Hombre	Más de 65 años	111	54	48,65%
2023	Interurbana	Mujer	18-24 años	15	4	26,67%
2023	Interurbana	Mujer	24-44 años	40	20	50,00%
2023	Interurbana	Mujer	45-64 años	50	10	20,00%
2023	Interurbana	Mujer	Más de 65 años	23	9	39,13%
2023	Urbana	Hombre	18-24 años	19	10	52,63%
2023	Urbana	Hombre	24-44 años	49	36	73,47%
2023	Urbana	Hombre	45-64 años	44	29	65,91%
2023	Urbana	Hombre	Más de 65 años	37	17	45,95%
2023	Urbana	Mujer	18-24 años	4	1	25,00%
2023	Urbana	Mujer	24-44 años	7	5	71,43%
2023	Urbana	Mujer	45-64 años	7	4	57,14%
2023	Urbana	Mujer	Más de 65 años	19	10	52,63%

* Debido a la ausencia de observaciones correspondientes a mujeres de entre 18 y 24 años en vías urbanas durante el año 2022, la probabilidad de consumo para esta categoría se ha estimado como el promedio de las probabilidades observadas en la misma combinación de variables para los años 2021 y 2023.

Fuente: Elaboración propia

Antes del proceso de imputación, fue necesario armonizar las categorías de edad entre ambas fuentes, pues los rangos no coincidían exactamente. En particular, se eliminaron los registros correspondientes a menores de edad (14–17 años) y se unificaron los tramos de edad superiores (65–74 años y más de 74 años) en una única categoría de "más de 65 años".

Por un lado, la eliminación de los datos de menores de edad se justifica por la falta de información sobre consumo de sustancias para este grupo, imposibilitando la estimación de probabilidades fiables, así como por su escasa relevancia analítica, pues se trata de un colectivo para el que cualquier consumo resulta ilegal, lo que dificulta la comparación con el resto de los conductores. Por otro lado, la agregación de los tramos de edad más elevados responde a la necesidad de garantizar la coherencia entre ambas fuentes y de evitar problemas derivados de tamaños muestrales reducidos en los grupos de mayor edad. Dado que estos tramos presentan patrones de consumo relativamente homogéneos y un peso reducido en el conjunto total, su unificación implica una pérdida mínima de información analítica.

Una vez estimadas las probabilidades, la variable de consumo de sustancias se genera mediante un procedimiento de imputación condicionada. Para cada observación del *dataset* de microdatos, se identifica la probabilidad asociada a su perfil y se asigna un valor binario (consumo positivo o negativo), garantizando que, a nivel agregado, la distribución resultante reproduzca fielmente los patrones observados en la fuente toxicológica.

3.1.3. Preprocesamiento y análisis exploratorio de los datos

Tras integrar todas las fuentes de información en un único conjunto de datos, se lleva a cabo el preprocesamiento y análisis exploratorio de los datos, siguiendo el orden metodológico definido en el apartado anterior.

(i) Depuración inicial y eliminación de variables no relevantes

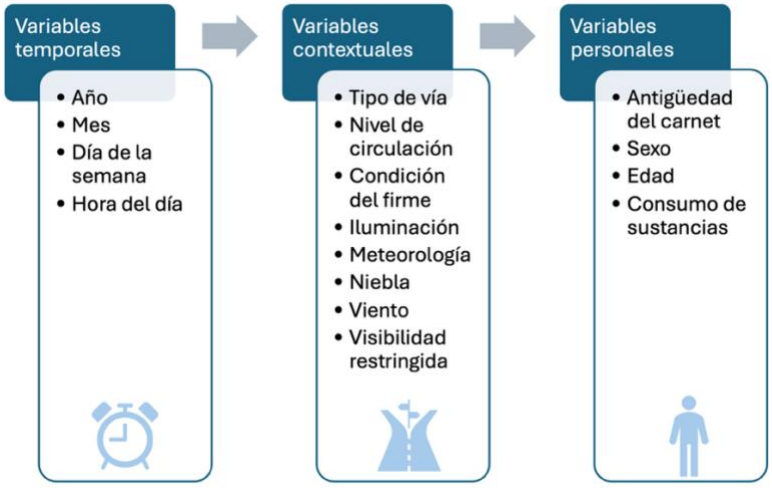
Como primer paso, se identifican y eliminan aquellas variables que no resultan relevantes para el análisis de la gravedad del siniestro vial.

Para ello, se elabora un diccionario estructurado de variables, en el que se recopila información de cada una de las variables inicialmente disponibles en el *dataset*. Para cada una, se documenta su tipo, las categorías o rangos asociados, una descripción funcional y la justificación de su utilidad o no para el análisis de la severidad. Un extracto representativo de dicho diccionario se incluye en el Anexo 2 del trabajo.

A partir de esta revisión, se excluyen del análisis las variables puramente identificativas, aquellas de carácter administrativo o descriptivo sin contenido analítico, así como las variables que recogen consecuencias posteriores al siniestro (número de heridos o fallecidos), que se introducirán en una fase posterior del trabajo, una vez completado el *clustering*. Estas últimas variables no describen las condiciones en que se produce el accidente, por lo que su inclusión en la segmentación sesgaría la formación de los clústeres al incorporar información del desenlace que precisamente se pretende estudiar después.

Como resultado de este proceso, se selecciona un subconjunto de variables que describen el contexto temporal, ambiental y humano de los siniestros viales estudiados.

Figura 4: Variables seleccionadas para la modelización



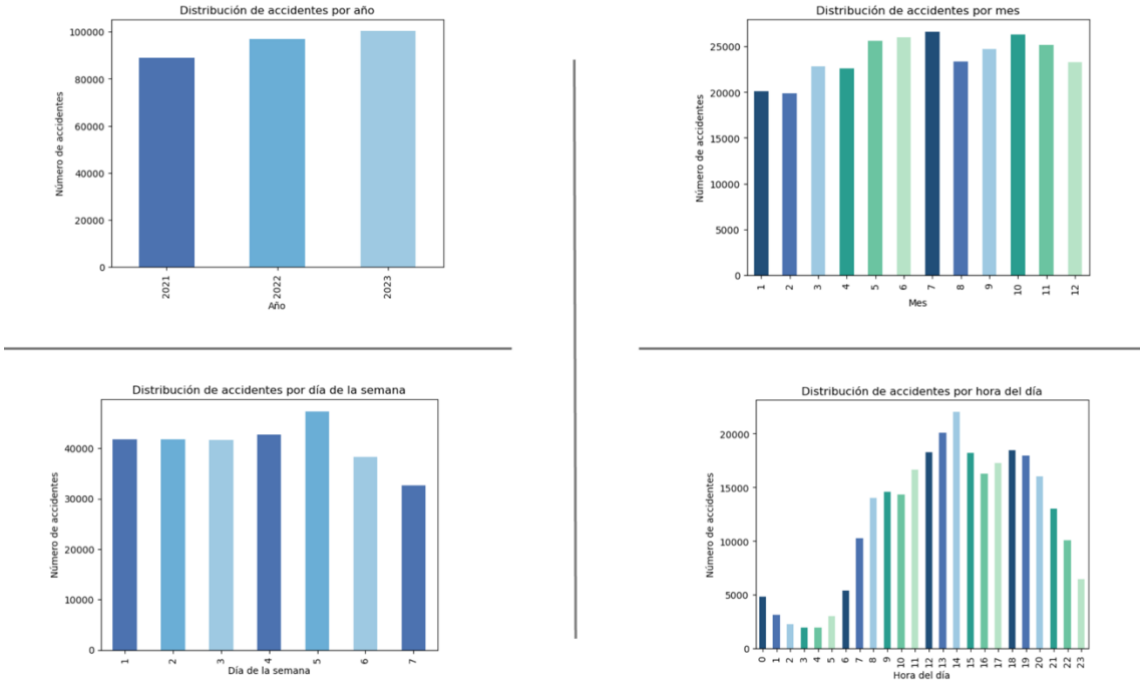
Fuente: Elaboración propia

(ii) Análisis univariante: distribución de las variables

Utilizando sólo las variables seleccionadas para el estudio, se lleva a cabo un análisis univariante para describir la distribución de cada variable y evaluar su coherencia interna.

Primero, se analizan las distribuciones de las **variables temporales** para comprobar la correcta codificación de los valores y la ausencia de rangos inválidos o valores fuera de dominio.

Figura 5: Distribuciones de las variables temporales



Fuente: Elaboración propia (con Python)

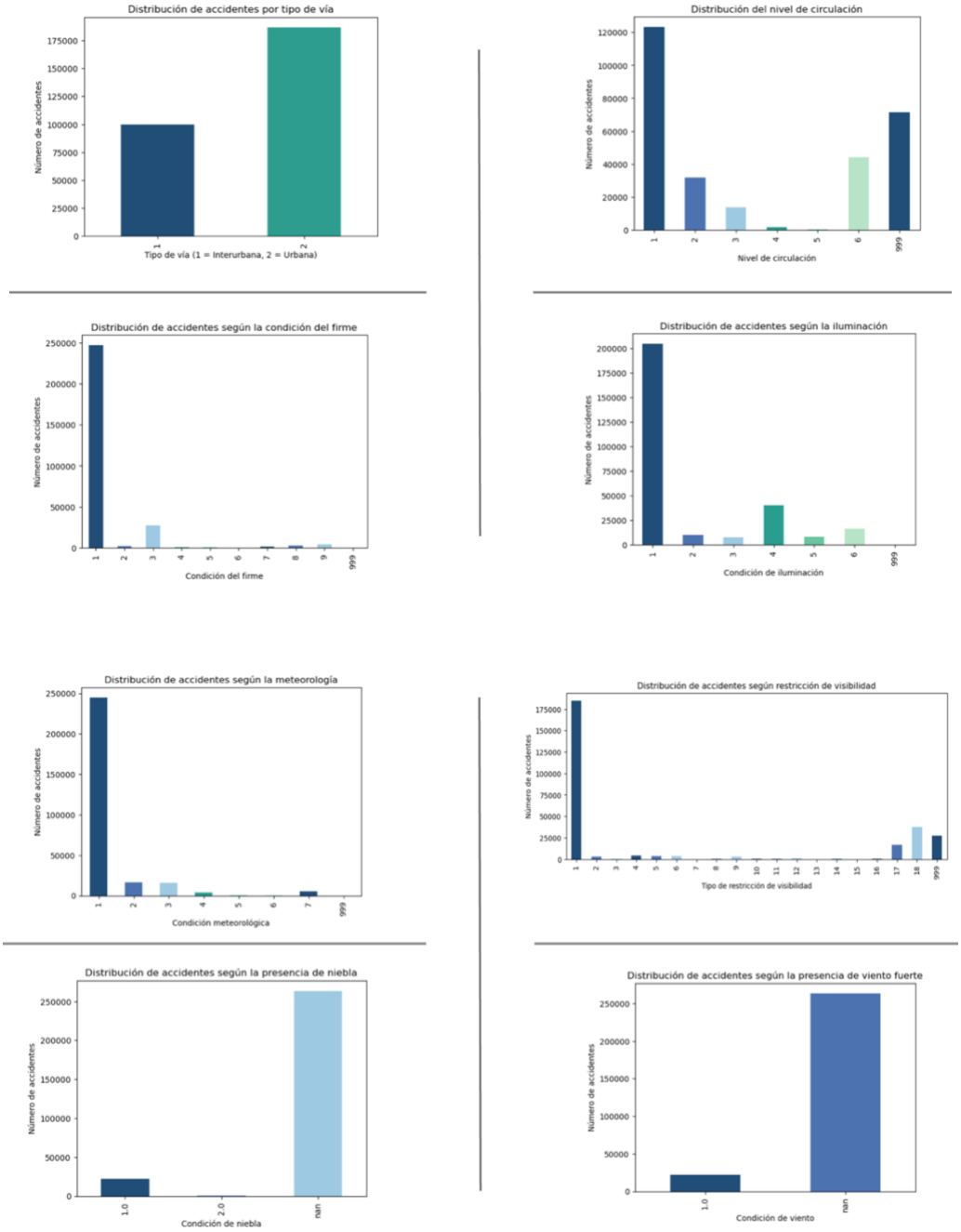
La distribución por año muestra una cobertura completa del periodo comprendido entre 2021 y 2023, con una ligera tendencia creciente a lo largo de los años. Por mes, los accidentes se reparten de forma relativamente homogénea a lo largo del año, con ligeras variaciones que reflejan posibles efectos estacionales.

En relación con el día de la semana, los días laborables concentran un mayor número de siniestros, mientras que durante el fin de semana se observa cierta reducción. La hora del día, por su parte, presenta una distribución no uniforme, con mayores frecuencias durante las mañanas y tardes y valores bajos en las horas nocturnas.

Todo ello confirma la coherencia interna del *dataset* y muestra patrones temporales potencialmente relevantes para el posterior análisis de la gravedad de los accidentes.

Segundo, en cuanto a las **variables contextuales**, se calcula igualmente la distribución de los siniestros en función de las condiciones ambientales en las que se producen.

Figura 6: Distribuciones de las variables contextuales



Fuente: Elaboración propia (con Python)

La distribución por tipo de vía muestra una mayor concentración de accidentes en vías urbanas que en interurbanas, lo que resulta coherente con la mayor densidad de tráfico, cruces e interacción entre vehículos y peatones en entornos urbanos.

En cuanto al nivel de circulación, predominan claramente los niveles bajos de circulación: el nivel blanco (1) concentra la mayoría de accidentes, seguido por el nivel verde (2) y nivel amarillo (3). Los niveles de alta congestión (rojo 4 y negro 5) son muy poco frecuentes, lo que indica que la mayoría de los accidentes no ocurren en situaciones de tráfico extremo.

Asimismo, se observa una proporción relevante de registros correspondientes a las categorías "Se desconoce" (6) y "Sin especificar" (999), que representan el 15,4% y el 24,93% de los casos, respectivamente. Dado que la distribución del nivel de circulación está claramente dominada por una única categoría, se opta por imputar ambos valores a la moda de la variable, por considerarse la aproximación más probable a falta de información directa. Esta decisión tiene sentido al tratarse de una variable ordinal, para la que la imputación modal constituye una estrategia conservadora que permite preservar su orden lógico sin introducir variabilidad artificial adicional.

Aunque la imputación proporcional constituye una alternativa posible, su aplicación implicaría asumir que los valores ausentes reproducen exactamente la distribución observada en los datos disponibles, lo que introduce un supuesto difícilmente verificable en este contexto. Por ello, la imputación mediante la moda se considera la opción más prudente.

Respecto a la condición del firme, la gran mayoría de los accidentes se producen con firme seco y limpio (1). En segundo lugar, aparece firme mojado (3), mientras que el resto de las categorías adversas (barro, inundado, hielo, nieve, aceite) tienen una presencia muy reducida y residual. Los valores "Se desconoce" (9) y "Sin especificar" (999) se consideran no informativos y, dada la clara dominancia de la categoría principal y la justificación previamente expuesta, se imputan igualmente a la moda.

En relación con las condiciones de iluminación, la mayoría de los accidentes se producen con luz natural diurna (1), lo que responde a una mayor exposición al tráfico durante las horas de día. En segundo lugar, destacan los accidentes nocturnos con iluminación artificial encendida (4), lo que indica que, aun existiendo alumbrado, el contexto nocturno sigue siendo relevante. La categoría "Sin especificar" (999) es residual y no aporta información interpretable, por lo que, por lo explicado anteriormente, se imputa a la moda.

Por su parte, la distribución de las condiciones meteorológicas muestra la meteorología despejada (1) como el escenario de circulación más frecuente. Después, aparecen el cielo nublado (2) y la lluvia débil (3), con una presencia claramente inferior, mientras que los episodios más adversos (lluvia fuerte, granizo o nieve) son residuales. Las categorías "Se desconoce" (7) y "Sin especificar" (999), que representan un porcentaje muy reducido y no aportan información, igualmente se imputarán a la moda, permitiendo conservar las observaciones sin introducir ruido en el posterior análisis.

En cuanto a la restricción de visibilidad, la mayoría de los accidentes se producen en condiciones de buena visibilidad (1). El resto de las categorías (edificios, factores atmosféricos, deslumbramientos, obras u otros elementos) presentan una presencia mucho menor. Destaca, no obstante, un volumen relevante de registros clasificados como "Se desconoce" (18) y "Sin especificar" (999), que, al no aportar información concreta sobre la causa de la restricción de visibilidad y teniendo en cuenta el claro predominio de la categoría de buena visibilidad, se imputarán también a la moda.

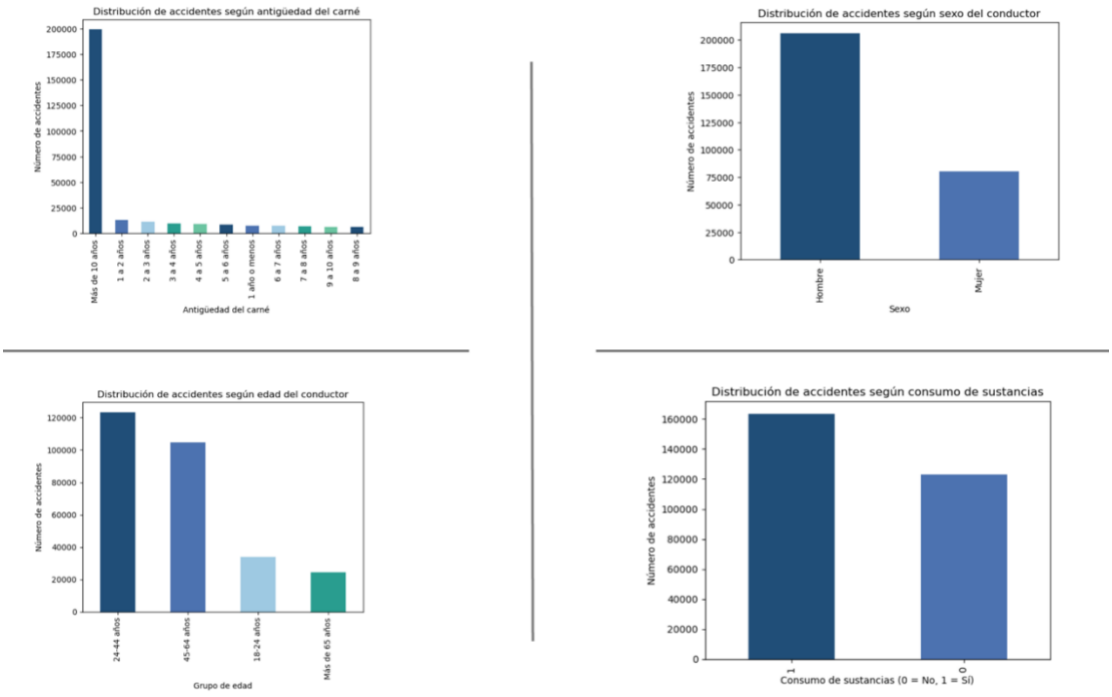
Finalmente, la interpretación de las dos últimas variables contextuales (niebla y viento) difiere del resto, pues en ambos casos la ausencia de valor no se asocia a información desconocida, sino a la ausencia del fenómeno.

En la condición de niebla, el diccionario de variables de la DGT únicamente recoge situaciones en las que existe niebla, distinguiendo entre niebla ligera (1) y niebla intensa (2), sin contemplar una categoría explícita para su ausencia. Dado que la mayoría de las observaciones no presentan valor registrado y que, en términos prácticos, la ausencia de niebla es la situación más habitual, los valores faltantes se interpretan razonablemente como inexistencia de niebla en el momento del accidente. En consecuencia, estos valores se recodificarán como una categoría específica que representa la ausencia del fenómeno y que tomará el valor 0.

De forma análoga, en la condición de viento, la variable distingue entre presencia de viento fuerte (1) y ausencia del mismo, si bien en el *dataset* no aparece registrada explícitamente esta última categoría. La elevada proporción de valores faltantes permite interpretar operativamente la ausencia de registro como ausencia de viento fuerte. Por ello, los valores faltantes también se recodificarán como 0, manteniendo el valor 1 para los casos en los que sí se observa viento fuerte.

Tercero, en cuanto a las **variables personales**, también se analizaron sus rangos y distribuciones.

Figura 7: Distribuciones de las variables personales



Fuente: Elaboración propia (con Python)

La Figura 7 revela patrones consistentes con la estructura del *dataset*. En primer lugar, la antigüedad del permiso de conducir muestra una clara concentración en conductores con más de 10 años de experiencia, lo que refleja la mayor presencia de conductores experimentados en la población general.

En segundo lugar, atendiendo al sexo, se observa un mayor número de accidentes protagonizados por hombres. En cuanto a la edad, los siniestros se concentran principalmente en los grupos de 24–44 y 45–64 años, mientras que los conductores más jóvenes y los de mayor edad presentan una frecuencia menor.

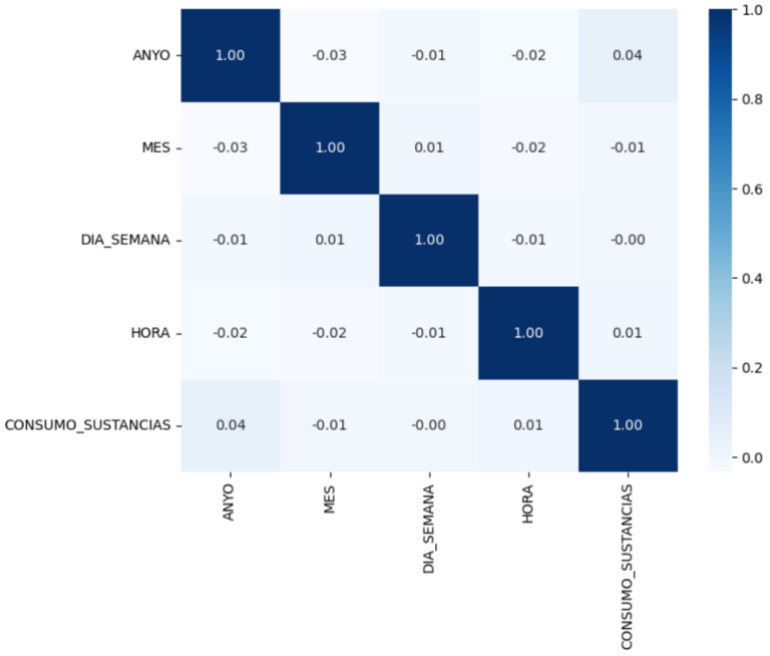
Por último, respecto al consumo de sustancias, la mayoría de los accidentes corresponden a conductores sin consumo, si bien la proporción de casos con consumo positivo es relevante. Ello justifica su consideración como un factor potencialmente determinante en el análisis de la severidad, sin perjuicio de las limitaciones metodológicas ya explicadas asociadas al alcance de esta variable.

(iii) Análisis de dependencias bivariantes

Una vez analizadas las variables de forma individual, se examinan relaciones entre pares de variables para evaluar la coherencia interna del *dataset* y la posible existencia de asociaciones relevantes.

Por un lado, para las variables numéricas, se calcula una matriz de correlaciones que identifique relaciones lineales entre magnitudes temporales y personales.

Figura 8: Matriz de correlaciones entre variables numéricas

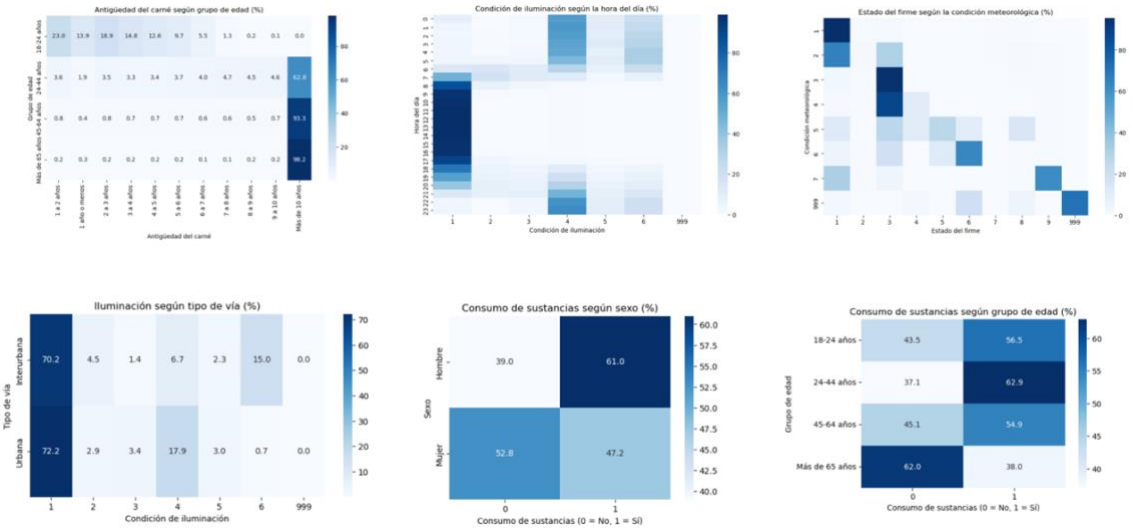


Fuente: Elaboración propia (con Python)

Como refleja la Figura 8, no existen relaciones lineales significativas entre las variables temporales ni entre estas y la variable de consumo de sustancias, pues en todos los casos los coeficientes de correlación presentan valores muy próximos a cero. Este resultado sugiere que, si existen relaciones entre estos factores, no se manifiestan de forma lineal, reforzando la idoneidad de emplear modelos no lineales para capturar posibles patrones complejos en el análisis de severidad de la siniestralidad vial.

Por otro lado, para las variables categóricas, se construyeron tablas de contingencia entre pares seleccionados de variables. Así, se verifica la coherencia lógica entre variables que, por su naturaleza, deben estar relacionadas, y se exploran asociaciones potencialmente relevantes desde el punto de vista analítico.

Figura 9: Tablas de contingencia entre variables categóricas



Fuente: Elaboración propia (con Python)

Las tres primeras representaciones analizan relaciones que, por definición, deben ser consistentes: la antigüedad del permiso de conducir en relación con el grupo de edad, la condición de iluminación en función de la hora del día y el estado del firme según la condición meteorológica.

Los resultados confirman patrones lógicos y esperables, como la imposibilidad de que conductores jóvenes presenten una elevada antigüedad del permiso, la correspondencia entre luz diurna y horas centrales del día o la asociación entre episodios de lluvia y firme mojado.

Las tres últimas tablas, por su parte, se orientan a explorar asociaciones potencialmente significativas para el análisis del riesgo de siniestro. En primer lugar, la relación entre el tipo de vía y la iluminación refleja diferencias claras entre vías urbanas e interurbanas, con una mayor presencia de iluminación artificial en vías urbanas y una mayor dependencia de la luz natural en las interurbanas. En segundo lugar, el análisis del consumo de sustancias revela patrones diferenciados según variables personales, con una mayor proporción de consumo positivo entre los hombres y entre los conductores de 24 a 44 años. En este sentido, estas asociaciones, sin implicar causalidad, aportan información sobre interacciones entre variables contextuales y personales que resultan de interés para su posterior incorporación en el análisis de la gravedad del siniestro vial.

(iv) Tratamiento de valores no informativos o faltantes

Una vez explicada en la fase anterior las estrategias de tratamiento de los valores no informativos y faltantes, se procede a su implementación efectiva sobre el *dataset*.

Por un lado, se identifican aquellas variables que incorporan categorías destinadas a codificar información desconocida o no especificada, tales como el valor 999 u otros códigos equivalentes. De acuerdo con lo establecido, estos valores se imputan a la moda, al tratarse de variables categóricas u ordinales con una distribución claramente dominada por una categoría principal. De esta forma, se conserva el tamaño muestral, se evita la pérdida de información y se preserva la estructura global de las distribuciones.

Por otro lado, se aborda el tratamiento de los valores faltantes en las variables relativas a la presencia de niebla y viento fuerte. Tal y como se razonó en el EDA, en estos casos la ausencia de registro no se interpreta como información desconocida, sino como ausencia del fenómeno. En consecuencia, los valores faltantes se recodifican como una categoría específica que representa dicha ausencia, asignándoles el valor 0, mientras que la presencia del fenómeno mantiene su codificación original.

(v) Eliminación de variables redundantes

Durante el proceso de depuración, se identifica una redundancia informativa entre las variables ZONA_AGRUPADA y VIA, pues ambas describen el carácter urbano o interurbano de la vía, aunque mediante codificaciones distintas. Por ello, considerando que la inclusión de ambas no aporta información adicional, se elimina la variable VIA y conserva ZONA_AGRUPADA, al tratarse de una variable original del microdato y estar ya codificada numéricamente.

Además, para mejorar la interpretabilidad del *dataset*, ZONA_AGRUPADA se renombra como TIPO_VIA, sin alterar su contenido informativo.

(vi) Homogeneización y recodificación de variables

Esta última fase del preprocesamiento pretende garantizar que todas las variables presenten un tipo coherente con su significado analítico, diferenciando entre variables numéricas, binarias, categóricas ordinales y categóricas nominales.

- Las variables temporales se mantienen sin modificaciones, al ser numéricas con significado real.

- TIPO_VIA se conserva como categórica codificada, sin orden inherente.

- Las variables CONDICION_NIVEL_CIRCULA y CONDICION_NIEBLA se tratan como categóricas ordinales, ya que sus categorías representan niveles crecientes de congestión y peligrosidad, respectivamente.

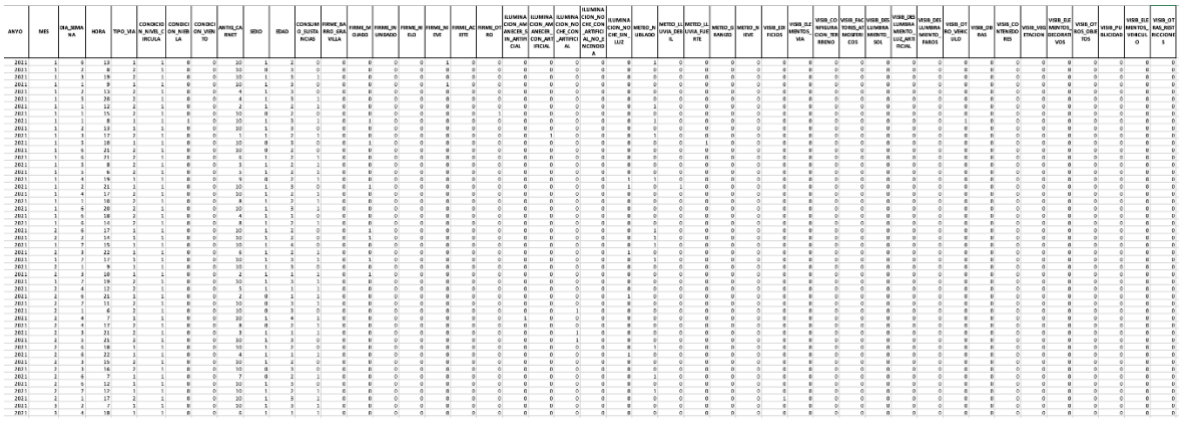
- CONDICION_VIENTO se mantiene como binaria (0/1).

- En cuanto a las variables personales, ANTIG_CARNET y EDAD se recodifican como ordinales, reflejando niveles crecientes de experiencia y etapas vitales, mientras que la variable SEXO se recodifica como binaria. La variable CONSUMO_SUSTANCIAS ya se encuentra correctamente codificada.

- Las variables CONDICION_FIRME, CONDICION_ILUMINACION, CONDICION_METEOROLOGICA y VISIBILIDAD_RESTRINGIDA_POR se consideran variables categóricas nominales, ya que sus categorías no presentan un orden natural entre sí. En consecuencia, se transforman mediante *one-hot encoding*, generando variables binarias que indican la presencia o ausencia de cada condición específica. Ello evita imponer relaciones de orden artificiales.

Llegados a este punto, se presenta a continuación un extracto del *dataset* final resultante, que servirá como base para el análisis de la severidad en carretera.

Tabla 9: Extracto del dataset final preprocesado



Fuente: Elaboración propia (con Python)

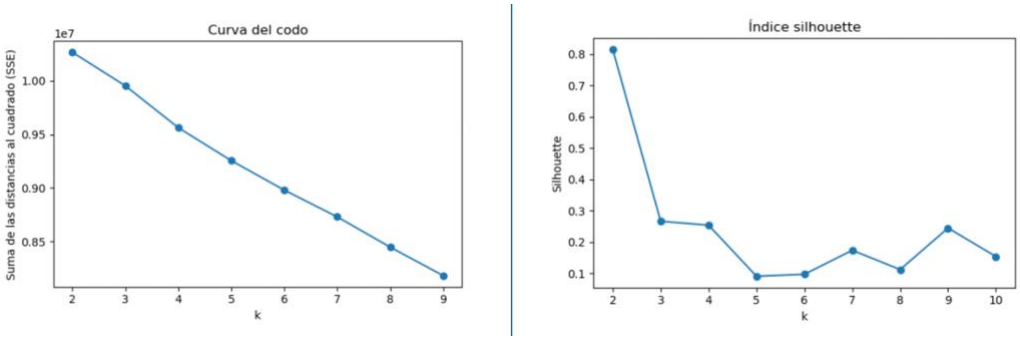
3.2. Modelización de la severidad vial

3.2.1. Segmentación de los accidentes de tráfico

El primer paso de la modelización de la severidad vial tiene como objetivo identificar tipologías homogéneas de accidente en función de sus características contextuales, temporales y personales. En este sentido, una vez escaladas las variables y reducida la dimensionalidad del conjunto de datos, se implementa un algoritmo de *clustering* mediante *K-means*.

Dado que *K-means* requiere especificar previamente el número de clústeres (*k*), es necesario determinar su valor óptimo antes de realizar la segmentación. Para ello, se emplean dos criterios complementarios: el método del codo, basado en la evolución de la suma de cuadrados intra-clúster, y el índice *silhouette*, que evalúa simultáneamente la cohesión interna de los clústeres y su separación respecto de los restantes grupos.

Figura 10: Métodos de selección del número de clústeres k



Fuente: Elaboración propia (con Python)

Por una parte, la curva del codo muestra una reducción progresiva de la variabilidad intra-clúster a medida que aumenta el número de grupos, sin apreciarse un punto de inflexión definido que permita identificar el número óptimo de clústeres. En condiciones ideales, se esperaría observar una caída brusca de la variabilidad hasta el valor óptimo de *k,* seguida de un aplanamiento claro de la curva. Sin embargo, en este caso, el descenso es gradual, sin un "codo" claramente marcado, lo que resulta habitual en *datasets* de alta dimensionalidad y con grupos parcialmente solapados, como ocurre con este fenómeno de la siniestralidad vial. La ausencia de un punto de inflexión claro hace necesario complementar el análisis con el índice *silhouette*.

Por su parte, el índice *silhouette* alcanza su valor máximo en k=2, lo que sugiere una partición con sólo dos grupos muy separados entre sí. No obstante, esta solución conduciría a una segmentación excesivamente agregada y poco informativa, al reducir la heterogeneidad del fenómeno a únicamente dos tipologías de accidente. En cambio, para k=3 y k=4, los valores del índice siguen siendo aceptables y permiten obtener una estructura de segmentación más rica y analíticamente útil.

En consecuencia, atendiendo al equilibrio entre calidad estadística e interpretabilidad de los resultados, se opta por una solución de cuatro clústeres.

Aplicado el algoritmo *K-means*, cuyo código se recoge en el Anexo 3, se obtiene una partición del *dataset* en cuatro grupos diferenciados de accidentes. A partir de este resultado, se analizan los centroides de cada clúster para identificar las variables que presentan valores más elevados o distintivos en cada grupo y, sobre esa base, perfilar cada uno de ellos como una tipología específica de siniestro vial.

Tabla 10: Valores medios de las variables por clúster (centroides del modelo)

	0	1	2	3
ANYO	2021,990	2022,056	2022,010	2021,668
MES	7,313	6,643	6,614	3,460
DIA_SEMANA	3,848	3,906	3,901	4,077
HORA	13,547	13,875	13,839	13,433
TIPO_VIA	1,513	1,660	1,681	1,418
CONDICION_NIVEL_CIRCULA	1,278	1,227	1,220	1,404
CONDICION_NIEBLA	0,073	0,081	0,082	0,065
CONDICION_VIENTO	0,062	0,080	0,083	0,050
ANTIG_CARNET	8,665	9,812	3,617	8,688
SEXO	0,722	0,721	0,716	0,757
EDAD	2,478	2,695	1,510	2,409
CONSUMO_SUSTANCIAS	0,559	0,559	0,615	0,561
FIRME_BARRO_GRAVILLA	0,003	0,009	0,008	0,003
FIRME_MOJADO	0,940	0,009	0,022	0,134
FIRME_INUNDADO	0,017	0,000	0,000	0,000
FIRME_HIELO	0,003	0,001	0,001	0,077
FIRME_NIEVE	0,000	0,000	0,000	0,742
FIRME_ACEITE	0,001	0,005	0,004	0,000
FIRME_OTRO	0,003	0,011	0,011	0,003
ILUMINACION_AMANECER_SIN_ARTIFICIAL	0,053	0,032	0,033	0,033
ILUMINACION_AMANECER_CON_ARTIFICIAL	0,045	0,025	0,025	0,021
ILUMINACION_NOCHE_CON_ARTIFICIAL	0,223	0,132	0,132	0,184
ILUMINACION_NOCHE_CON_ARTIFICIAL_NO_ENCENDIDA	0,029	0,028	0,026	0,021
ILUMINACION_NOCHE_SIN_LUZ	0,106	0,052	0,050	0,128
METEO_NUBLADO	0,179	0,043	0,049	0,053
METEO_LLUVIA_DEBIL	0,601	0,000	0,002	0,024
METEO_LLUVIA_FUERTE	0,142	0,000	0,000	0,006
METEO_GRANIZO	0,004	0,000	0,000	0,047
METEO_NIEVE	0,000	0,000	0,000	0,703
VISIB_EDIFICIOS	0,006	0,009	0,009	0,006
VISIB_ELEMENTOS_VIA	0,002	0,002	0,002	0,000
VISIB_CONFIGURACION_TERRRENO	0,033	0,014	0,014	0,012
VISIB_FACTORES_ATMOSFERICOS	0,130	0,000	0,000	0,282
VISIB_DESLUMBRAMIENTO_SOL	0,001	0,014	0,012	0,006
VISIB_DESLUMBRAMIENTO_LUZ_ARTIFICIAL	0,000	0,000	0,000	0,000
VISIB_DESLUMBRAMIENTO_FAROS	0,001	0,000	0,000	0,000
VISIB_OTRO_VEHICULO	0,010	0,011	0,011	0,003
VISIB_OBRAS	0,001	0,001	0,001	0,000
VISIB_CONTENEDORES	0,001	0,001	0,001	0,003
VISIB_VEGETACION	0,004	0,004	0,004	0,000
VISIB_ELEMENTOS_DECORATIVOS	0,000	0,000	0,000	0,000
VISIB_OTROS_OBJETOS	0,001	0,001	0,001	0,000
VISIB_PUBLICIDAD	0,000	0,000	0,000	0,000
VISIB_ELEMENTOS_VEHICULO	0,001	0,001	0,001	0,003
VISIB_OTRAS_RESTRICCIONES	0,034	0,062	0,061	0,024

Fuente: Elaboración propia (con Python)

Con base en la tabla anterior, la segmentación obtenida permite distinguir cuatro tipologías diferenciadas de siniestro vial, cada una con las características presentadas en la Figura 11.

Figura 11: Perfiles interpretativos de los clústeres

Clúster 0: Accidentes con un perfil meteorológico adverso	Clúster 1 (MAYORITARIO): Accidentes en condiciones normales de circulación
- Tamaño: 25.012 observaciones (9% del total). - Agrupa accidentes caracterizados por la presencia de lluvia y firme mojado. - Presenta valores significativamente superiores en las variables relacionadas con precipitaciones: METEO_LLUVIA_DEBIL, METEO_LLUVIA_FUERTE y FIRME_MOJADO.	- Tamaño: 196.561 observaciones (70,5% del total). - Agrupa accidentes ocurridos en condiciones normales de circulación. - Predominan valores bajos en las variables asociadas a fenómenos meteorológicos adversos y alteraciones del firme.
Clúster 2: **Accidentes con un perfil de conductores más jóvenes**	**Clúster 3:** **Accidentes en condiciones extremas (hielo o nieve)**
- Tamaño: 56.935 observaciones (20,4% del total). - Agrupa accidentes asociados principalmente a conductores de menor edad y menor antigüedad del permiso de conducción. Es un perfil vinculado a factores individuales del conductor.	- Tamaño: 337 observaciones (0,1% del total). - Agrupa accidentes producidos en condiciones meteorológicas especialmente adversas o extremas, caracterizadas por la presencia de nieve, hielo y limitaciones severas de visibilidad. - Presenta valores elevados en variables como FIRME_NIEVE, FIRME_HIELO y METEO_NIEVE.

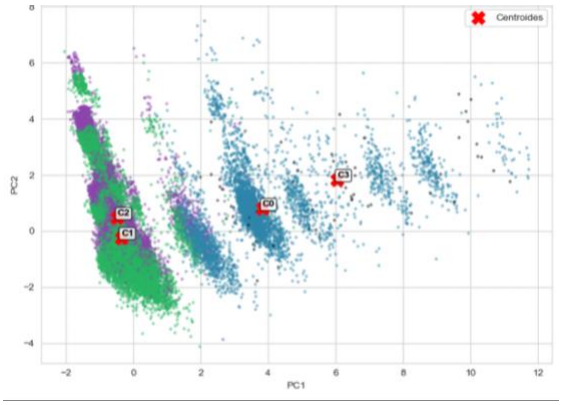
Fuente: Elaboración propia

El perfilado anterior se complementa con la Figura 12, que muestra la proyección de los clústeres en el espacio definido por las dos primeras componentes principales.

La representación aprecia cierto solapamiento entre algunos grupos, especialmente entre los clústeres 1 y 2. Ello resulta coherente por reducir a dos dimensiones un fenómeno originalmente descrito en un espacio de alta dimensionalidad como el de la siniestralidad vial. Así las cosas, la proximidad visual no implica necesariamente una separación deficiente entre clústeres, sino que refleja que parte de su diferenciación se produce en dimensiones no representadas en el gráfico.

Por su parte, el clúster 3 sí aparece relativamente aislado, por ser un perfil más extremo asociado a condiciones meteorológicas adversas.

Figura 12: Clústeres proyectados en PCA



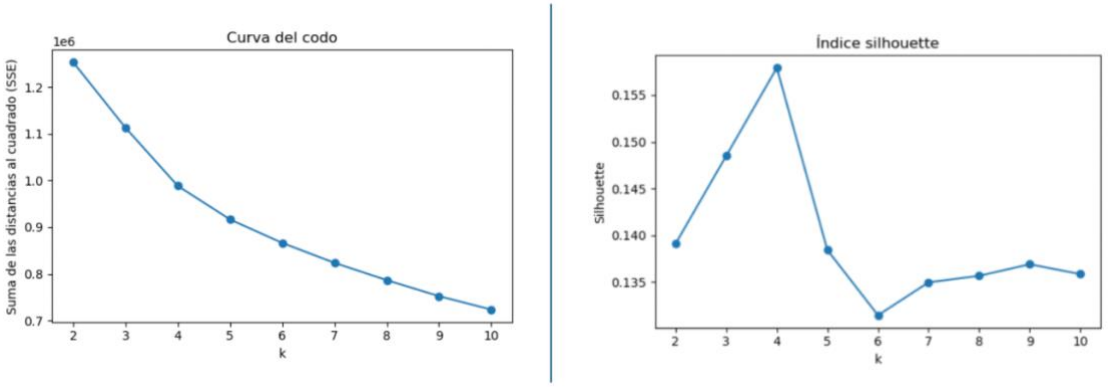
Fuente: Elaboración propia (con Python)

Llegados a este punto, si bien el *clustering* ofrece una primera estructura del fenómeno de la siniestralidad vial sobre la que desarrollar el posterior análisis de severidad, los resultados muestran que el clúster 1 concentra una proporción muy elevada de las observaciones del *dataset*. Ello sugiere que, dentro de este grupo aparentemente homogéneo, pueden seguir existiendo subperfiles diferenciados que la segmentación inicial no llega a capturar por completo.

Por ello, y teniendo en cuenta que este clúster ha sido perfilado como los accidentes en condiciones normales de circulación, se procede a una segunda segmentación aplicada exclusivamente sobre dicho grupo. En esta nueva fase, se eliminan las variables contextuales que habían sido determinantes en la clasificación inicial y se seleccionan únicamente variables temporales y personales del conductor, con el objetivo de detectar subestructuras internas dentro del contexto más frecuente de accidentalidad.

Al igual que en la segmentación general, se emplean conjuntamente el método del codo y el índice *silhouette* para determinar el número óptimo de subclústeres. Tal y como se aprecia en la Figura 13, la curva del codo muestra una reducción progresiva de la variabilidad intra-grupo sin un punto de inflexión claramente definido. Por su parte, el índice *silhouette* alcanza su valor máximo en k = 4, aunque los valores obtenidos son relativamente bajos, en torno a 0,135 y 0,155, lo que indica que la separación entre subgrupos es ciertamente débil. Ello tiene sentido por la elevada homogeneidad interna del clúster mayoritario analizado. En línea con lo anterior, y asumiendo que los subclústeres son interpretables pero parcialmente solapados en el espacio de variables, se adopta la solución óptima: cuatro subclústeres.

Figura 13: Métodos de selección del número de subclústeres k



Fuente: Elaboración propia (con Python)

Aplicado nuevamente el algoritmo *K-means* sobre la submuestra correspondiente al clúster mayoritario, se obtiene una nueva partición interna cuyos centroides se recogen en la Tabla 11.

Tabla 11: Valores medios de las variables por subclúster (centroides del modelo)

	0	1	2	3
ANYO	2022,09	2022,07	2022,04	2022,04
MES	6,71	6,62	6,64	6,66
DIA_SEMANA	3,93	3,90	3,90	3,91
HORA	13,96	13,90	13,81	13,89
ANTIG_CARNET	7,00	9,96	9,99	9,95
SEXO	0,73	1,00	0,92	0,24
EDAD	2,36	2,52	3,45	2,37
CONSUMO_SUSTANCIAS	0,51	1,00	0,17	0,35

Fuente: Elaboración propia (con Python)

A partir de estos valores medios, es posible identificar los rasgos diferenciales de cada subclúster y construir perfiles interpretativos más precisos del tipo de conductor implicado en los accidentes ocurridos bajo condiciones normales de circulación.

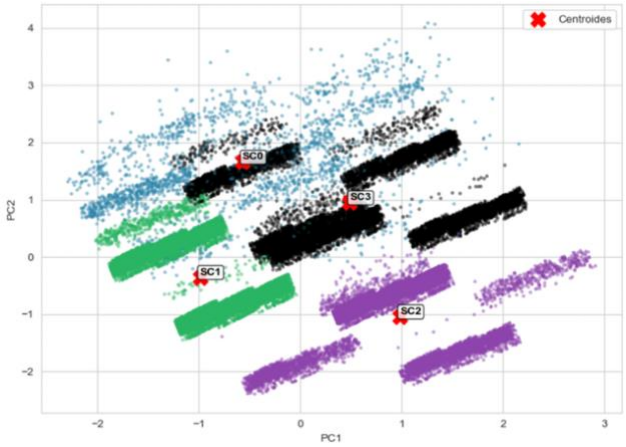
Figura 14: Perfiles interpretativos de subclústeres

Subclúster 0: Accidentes con conductores jóvenes con menos experiencia	Subclúster 1 (MAYORITARIO): Accidentes con conductores hombres con consumo de sustancias
- Tamaño: 10.321 observaciones (5,25% del total). - Agrupa conductores con menor experiencia al volante. - Presenta los valores más bajos en ANTIG_CARNET y en EDAD. - Muestra un nivel intermedio en CONSUMO_SUSTANCIAS.	- Tamaño: 73.986 observaciones (37,64% del total). - Agrupa conductores hombres bajo los efectos del alcohol o de las drogas. - Presenta el valor máximo en CONSUMO_SUSTANCIAS (1,00), evidenciando que todos los conductores han consumido. - Registra el valor máximo en la variable SEXO (1,00), reflejando que todos los conductores son hombres.
Sublúster 2: **Accidentes con conductores hombres con mayor experiencia y menor consumo de sustancias**	**Sublúster 3:** **Accidentes con conductoras mujeres con consumo de sustancias moderado**
- Tamaño: 48.845 observaciones (24,85% del total). - Agrupa conductores con experiencia al volante que, en términos generales, no están bajo los efectos del alcohol o de las drogas. - Presenta los valores más altos en EDAD y en ANTIG_CARNET. - Muestra el valor más bajo en CONSUMO_SUSTANCIAS.	- Tamaño: 63.409 observaciones (32,26% del total). - Agrupa mujeres al volante con consumo moderado; el 35% de ellas han consumido alguna sustancia. - Presenta el valor más bajo en SEXO (0,24), reflejando que la mayoría de los conductores que forman este subclúster son mujeres. - Registra valores bajos en CONSUMO_SUSTANCIAS (0,35).

Fuente: Elaboración propia

Al igual que en el *clustering*, la Figura 15 muestra la proyección de los subclústeres en el espacio definido por las dos primeras componentes principales. En este caso, también se observa un solapamiento entre los subgrupos, a excepción del subclúster 2 formado por conductores jóvenes, que presenta mayor separación. La representación, por tanto, revela que la separación entre subclústeres sí existe pero es moderada, en coherencia con los valores débiles del índice *silhouette.*

Figura 15: Subclústeres proyectados en PCA



Fuente: Elaboración propia (con Python)

En conjunto, la subsegmentación del clúster mayoritario confirma que, incluso dentro de las condiciones normales de circulación, persiste una heterogeneidad interna explicada fundamentalmente, no por variables temporales, sino por las variables personales del conductor.

3.2.2. Caracterización de la severidad por clúster

Una vez identificadas las tipologías de accidente y como primera aproximación al estudio de la severidad, se lleva a cabo un análisis descriptivo que determina el nivel de severidad en cada clúster. Para ello, se incorporan al *dataset* las variables relativas tanto al número de fallecidos como al número de heridos graves.

A partir de estas variables, se construyen los tres indicadores ya definidos en el apartado del desarrollo metodológico: el indicador de fallecidos, el de heridos graves y el índice sintético de severidad, consistente en una combinación ponderada del número de fallecidos y heridos graves por accidente.

Sobre esta base, se analizan los niveles de severidad asociados a cada uno de los clústeres obtenidos tanto en la segmentación general como en el *subclustering* del clúster mayoritario.

Tabla 12: Indicadores de severidad de los accidentes por clúster

CLUSTER	Nº de accidentes	Nº total de fallecidos	Nº total de heridos graves	Accidentes mortales (%)	Accidentes graves (%)	Media de fallecidos por accidente	Media de heridos graves por accidente	Índice medio de severidad
0	25012	1126	6462	4.086	21.274	0.045	0.258	0.483
1	196561	8670	51599	4.026	21.955	0.044	0.263	0.483
2	56935	2469	14964	3.978	22.097	0.043	0.263	0.480
3	337	12	95	3.561	22.552	0.036	0.282	0.460

Fuente: Elaboración propia (con Python)

Tabla 13: Indicadores de severidad de los accidentes por subclúster

SUBCLUSTER_CONDICIONES_NORMALES	Nº de accidentes	Nº total de fallecidos	Nº total de heridos graves	Accidentes mortales (%)	Accidentes graves (%)	Media de fallecidos por accidente	Media de heridos graves por accidente	Índice medio de severidad
0	10321	472	2725	4.215	22.343	0.046	0.264	0.493
1	73986	3204	19258	3.933	21.820	0.043	0.260	0.477
2	48845	2244	12951	4.179	22.150	0.046	0.265	0.495
3	63409	2750	16665	3.985	21.898	0.043	0.263	0.480

Fuente: Elaboración propia (con Python)

Según muestra la Tabla 12, se da un comportamiento significativamente homogéneo en el nivel de severidad entre los distintos clústeres obtenidos en la segmentación general. En concreto, los valores del índice medio de severidad se sitúan en niveles muy próximos, entre 0,480 y 0,483, en los clústeres 0, 1 y 2. Así, pese a las diferencias existentes en términos de volumen de accidentes o características contextuales y personales de cada grupo, la gravedad media de los siniestros es similar entre ellos. El clúster 3, por su parte, presenta un nivel ligeramente inferior de severidad media, pero su reducido tamaño limita la relevancia interpretativa de este resultado dentro del análisis.

Por otro lado, el análisis del *subclustering* sobre el clúster mayoritario asociado a condiciones normales de circulación muestra un patrón parecido en la Tabla 13: los subclústeres presentan diferencias mínimas de severidad entre sí, con los valores del índice oscilando entre 0,477 y 0,495, lo que confirma la existencia de una elevada homogeneidad también dentro de este grupo.

En conclusión, los resultados sugieren que la segmentación obtenida permite identificar tipologías de accidente en términos de sus características contextuales y personales, pero no genera, desde una perspectiva descriptiva, diferencias sustanciales en términos de gravedad media de los siniestros entre grupos.

No obstante, estos resultados no implican que las variables contextuales o personales no influyan en la severidad de los accidentes, sino que la severidad de dichos accidentes parece depender más de combinaciones específicas de factores dentro de cada contexto que de la pertenencia global a una tipología concreta de siniestro.

3.2.3. Efecto del clúster sobre la probabilidad de severidad

Si bien se muestra homogeneidad en los indicadores medios de severidad por clúster, ello no excluye la posible existencia de diferencias estadísticamente significativas en la probabilidad de que un accidente sea grave o mortal entre los distintos clústeres. Por lo tanto, se desarrolla en este apartado un análisis inferencial orientado a evaluar si la pertenencia a las distintas tipologías se asocia a la probabilidad de que un accidente presente consecuencias graves.

Para ello, se emplean dos aproximaciones complementarias. Primero, para evaluar la significación estadística de las diferencias observadas, se analiza la probabilidad de accidente severo en cada clúster mediante la estimación de intervalos de confianza del 95% y la prueba Chi-cuadrado. Segundo, para comprobar si la pertenencia a un clúster concreto modifica la probabilidad de severidad del accidente, se estima un modelo de regresión logística.

En relación con la primera aproximación, centrada en la **comparación de la probabilidad estimada de accidente grave entre clústeres**, los resultados obtenidos muestran valores muy similares entre los distintos grupos identificados.

Tabla 14: Probabilidad estimada de accidente grave por clúster e intervalos de confianza del 95%

	CLUSTER	Nº accidentes	Probabilidad de accidente severo	IC inferior (95%)	IC superior (95%)
0	0	25012	0.091356	0.087848	0.094990
1	1	196561	0.093727	0.092446	0.095023
2	2	56935	0.095196	0.092813	0.097634
3	3	337	0.077151	0.053191	0.110643

Fuente: Elaboración propia (con Python)

Aunque el clúster 2 registra la mayor tasa de severidad y el clúster 3 la menor, los intervalos de confianza del 95% asociados a estas estimaciones se solapan ampliamente entre sí, sugiriendo la ausencia de diferencias estadísticamente significativas relativas a la severidad entre clústeres.

A su vez, esta conclusión se confirma mediante la aplicación de la prueba Chi-cuadrado de independencia entre las variables CLUSTER y SEVERIDAD, cuyo resultado es el siguiente:

- $\chi^2 = 4{,}17$;
- p-valor = 0,243.

De esta manera, al ser el p-valor mayor a 0,05, no se puede rechazar la hipótesis nula de independencia entre ambas variables, por lo que no se observa evidencia estadística suficiente para afirmar que la pertenencia a una tipología de accidente se asocie con una mayor o menor probabilidad de que dicho accidente presente consecuencias graves.

Este mismo análisis se replica en el *subclustering* correspondiente a las condiciones normales de circulación. Sin embargo, los resultados obtenidos muestran nuevamente una elevada homogeneidad entre subclústeres, con tasas de severidad muy próximas entre sí y con intervalos de confianza solapados.

Tabla 15: Probabilidad estimada de accidente grave por subclúster e intervalos de confianza del 95%

	Subcluster	Nº accidentes	Probabilidad de accidente severo	IC inferior (95%)	IC superior (95%)
0	0	10321	0.097374	0.091804	0.103245
1	1	73986	0.093612	0.091534	0.095732
2	2	48845	0.093479	0.090930	0.096093
3	3	63409	0.093457	0.091216	0.095747

Fuente: Elaboración propia (con Python)

En línea con lo anterior, la prueba Chi-cuadrado aplicada a la relación entre SUBCLUSTER_CONDICIONES_NORMALES y SEVERIDAD muestra los siguientes resultados:

- $\chi^2 = 1{,}72$;
- p-valor = 0,633.

Así, se confirma la ausencia de asociación estadísticamente significativa entre la pertenencia a las subtipologías identificadas y la probabilidad de severidad del accidente.

Por todo ello, los resultados indican que, si bien la segmentación mediante técnicas de *clustering* permite identificar perfiles diferenciados de accidentalidad, la pertenencia a uno u otro grupo no constituye un factor determinante en la explicación de la gravedad del siniestro.

No obstante, para completar este análisis, se estima un **modelo de regresión logística** que contraste esta relación, cuyo código se recoge en el Anexo 4. En él, la variable dependiente es la severidad del accidente, ya definida como una variable binaria con valor 1 si el siniestro presenta fallecidos o heridos graves y con valor 0 si no lo hace.

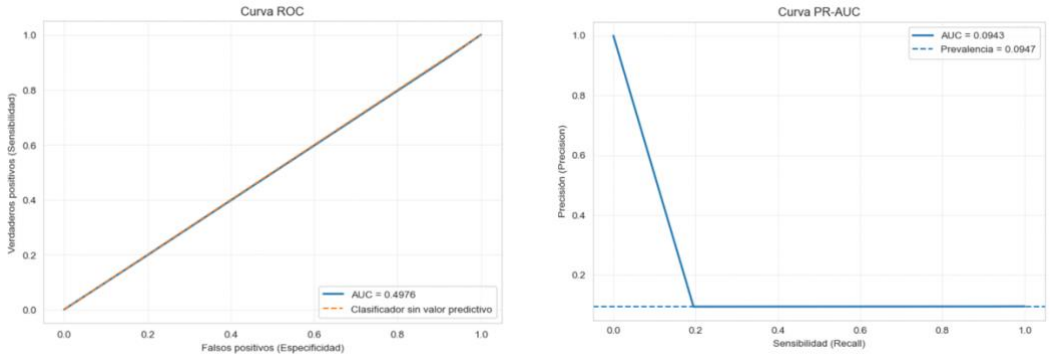
Para evaluar adecuadamente la capacidad predictiva del modelo, se adopta un enfoque de validación temporal basado en la separación de la muestra en conjuntos de entrenamiento (*train)* y validación (*test)*. En particular, los accidentes correspondientes a 2021 y 2022 se emplean para entrenar el modelo, mientras que los accidentes registrados en 2023 se reservan como conjunto de validación, de tal forma que se permite analizar la capacidad de generalización del modelo a un periodo posterior distinto del utilizado para su estimación.

En este sentido, se estima el modelo utilizando el conjunto de entrenamiento e introduciendo la variable de pertenencia al clúster como factor explicativo. Dado el carácter categórico de esta variable, el modelo toma uno de los clústeres, el clúster 1 y mayoritario, como categoría de referencia, y estima el efecto relativo del resto sobre la probabilidad de severidad del accidente. Así, los coeficientes obtenidos permiten interpretar la variación en términos de *odds ratio*, es decir, en términos del cambio relativo en la probabilidad de que un accidente sea severo asociado a la pertenencia a cada clúster respecto del grupo de referencia.

Una vez estimado el modelo, su capacidad predictiva se evalúa sobre el conjunto de validación, para lo que se utilizan dos métricas estándar ampliamente aceptadas en problemas de clasificación binaria: el área bajo la curva ROC (ROC-AUC) y el área bajo la curva precisión–*recall* (PR-AUC o *average precision*).

Por un lado, la curva ROC-AUC mide la capacidad global del modelo para discriminar entre accidentes severos y no severos a distintos umbrales de clasificación. Por su parte, la PR-AUC resulta especialmente informativa cuando la variable dependiente presenta desbalance entre clases, como ocurre en este caso, donde los accidentes severos representan una proporción reducida del total de siniestros.

Figura 16: Curvas ROC-AUC y PR-AUC del modelo para el clustering general



Fuente: Elaboración propia (con Python)

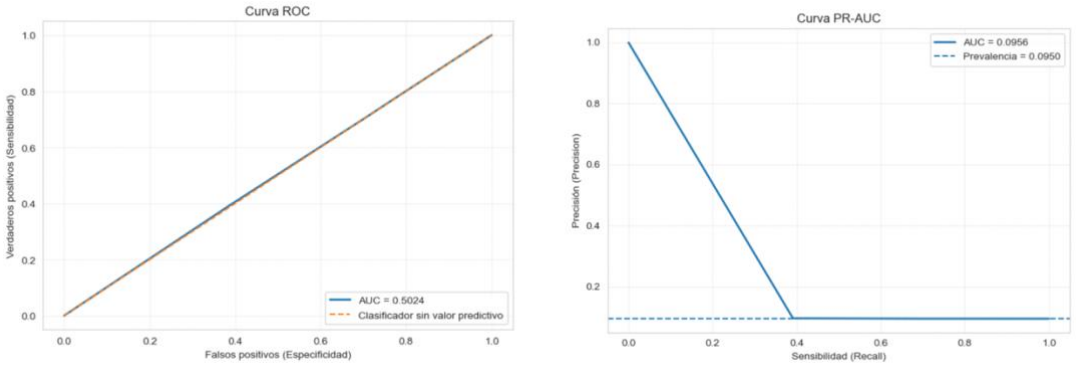
Los resultados obtenidos muestran valores de ROC-AUC próximos a 0,5 y valores de PR-AUC cercanos a la prevalencia observada de la variable de severidad, demostrando que el modelo no mejora respecto a una clasificación aleatoria según la frecuencia observada de accidentes severos. En consecuencia, la segmentación obtenida mediante el *clustering* no parece aportar una mejora en la capacidad predictiva del modelo en relación con la gravedad del accidente.

A continuación, se replica el mismo procedimiento en el *subclustering* correspondiente al grupo de siniestros mayoritario. En este caso, la variable explicativa utilizada en el modelo es la pertenencia a los distintos subclústeres identificados dentro de este grupo, manteniéndose la misma estrategia de validación temporal.

Los resultados obtenidos muestran que, tomando como referencia el subclúster 1, grupo mayoritario dentro de las condiciones normales de circulación, ninguno de los restantes subgrupos presenta coeficientes estadísticamente significativos. Los valores estimados de los *odds ratios* son próximos a la unidad y sus intervalos de confianza incluyen el valor 1, lo que indica la ausencia de diferencias estadísticamente relevantes en la probabilidad de severidad respecto al grupo de referencia.

Asimismo, la evaluación del modelo sobre el conjunto de validación muestra valores de ROC-AUC próximos a 0,5 y valores de PR-AUC prácticamente coincidentes con la prevalencia observada de la variable de severidad.

Figura 17: Curvas ROC-AUC y PR-AUC del modelo para el subclustering en condiciones normales de circulación



Fuente: Elaboración propia (con Python)

Consecuentemente, la pertenencia a los distintos subclústeres tampoco aporta capacidad discriminatoria relevante en términos de predicción de la gravedad del accidente.

En definitiva, los resultados obtenidos permiten concluir que, si bien la segmentación realizada es útil para identificar perfiles diferenciados de accidentalidad, no contribuye, si se considera de forma aislada, a la explicación predictiva de la severidad del siniestro.

En esta línea, es probable que la gravedad del accidente dependa en mayor medida de otros factores no incorporados en este modelo, como puede ser el tipo de colisión y las infracciones del conductor (como sostiene Herrera Briones en su trabajo de fin de grado), o el uso de medidas protectoras como el cinturón de seguridad (como defiende Febres Eguiguren en su tesis doctoral). La incorporación de estos factores constituye, por tanto, una posible línea de ampliación del análisis en trabajos futuros.

3.2.4. Identificación de los determinantes de severidad en cada clúster

Este apartado, que tiene como objetivo detectar qué variables contribuyen en mayor medida a explicar la probabilidad de que un siniestro sea grave en cada clúster y subclúster, se estructura en tres fases:

I. En primer lugar, se identifican los cinco *drivers* principales de severidad en cada clúster y subclúster.

II. En segundo lugar, se analiza el impacto del *driver* más relevante en la probabilidad de severidad dentro de cada grupo.

III. En tercer lugar, se realiza una simulación *what-if* que estima la reducción potencial de la severidad si disminuyera la incidencia del principal factor explicativo de la gravedad de los accidentes de tráfico.

En la primera fase, relativa a la **identificación de los cinco *drivers* principales de severidad por clúster y subclúster**, se estiman modelos de regresión logística binaria con penalización L1 para cada grupo. Dicha penalización reduce a cero los coeficientes asociados a variables con escasa capacidad explicativa, de tal forma que se seleccionan únicamente los factores con mayor contribución a la predicción de la severidad, favoreciendo modelos más interpretables.

En este contexto, para cada clúster se construye un subconjunto de datos que sólo incluye los siniestros pertenecientes a dicho clúster y, sobre ese grupo, se estima un modelo utilizando como variable dependiente el indicador de severidad y como variables explicativas las características temporales, contextuales y personales del accidente.

A partir de los coeficientes estimados, se identifican los *drivers* más relevantes en cada clúster, que se ordenan según su contribución al modelo, seleccionándose los cinco factores con mayor relevancia explicativa.

Además de los coeficientes estimados, que dan lugar a los *odds ratios* y permiten medir cuánto cambia la probabilidad de severidad cuando aparece cada variable, también se calculan indicadores adicionales que permiten interpretar la importancia de cada factor dentro del clúster.

En particular, se calcula la prevalencia de la variable, que muestra el porcentaje de accidentes del clúster en los que está presente; la intensidad de su asociación con la severidad, calculada como |log(OR)|, que mide la magnitud del efecto; y el impacto global, obtenido como el producto entre prevalencia e intensidad, que sirve para aproximar la relevancia práctica conjunta de cada factor y facilitar la comparación de su contribución relativa a la severidad con respecto a las demás variables.

Los resultados se muestran en la siguiente tabla:

Tabla 16: Top 5 drivers de severidad por clúster

	Cluster	Variable	Odds Ratio	Prevalencia	Intensidad	Impacto
0	0	SEXO	1.062	0.722	0.060	0.043
1	0	ILUMINACION_NOCHE_SIN_LUZ	1.237	0.106	0.213	0.023
2	0	VISIB_FACTORES_ATMOSFERICOS	1.035	0.130	0.035	0.005
3	0	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.005	0.223	0.005	0.001
4	1	ILUMINACION_NOCHE_SIN_LUZ	1.353	0.052	0.302	0.016
5	1	SEXO	1.020	0.721	0.020	0.014
6	1	VISIB_OTRAS_RESTRICCIONES	1.153	0.062	0.143	0.009
7	1	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.051	0.132	0.049	0.007
8	1	CONSUMO_SUSTANCIAS	1.011	0.559	0.011	0.006
9	2	CONSUMO_SUSTANCIAS	1.029	0.615	0.028	0.018
10	2	VISIB_OTRAS_RESTRICCIONES	1.256	0.061	0.228	0.014
11	2	ILUMINACION_NOCHE_SIN_LUZ	1.297	0.050	0.260	0.013
12	2	METEO_NUBLADO	1.146	0.049	0.136	0.007
13	2	ILUMINACION_AMANECER_SIN_ARTIFICIAL	1.059	0.033	0.057	0.002
14	3	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.348	0.184	0.299	0.055

Fuente: Elaboración propia (con Python)

Este procedimiento se replica de forma análoga para los subclústeres identificados dentro del grupo de condiciones normales, cuyos resultados se muestran en la Tabla 17.

Tabla 17: Top drivers de severidad por subclúster

	Subcluster	Variable	Odds Ratio	Prevalencia	Intensidad	Impacto
0	0	SEXO	1.067	0.734	0.065	0.048
1	0	ILUMINACION_NOCHE_SIN_LUZ	1.345	0.053	0.296	0.016
2	1	ILUMINACION_NOCHE_SIN_LUZ	1.306	0.048	0.267	0.013
3	1	VISIB_OTRAS_RESTRICCIONES	1.123	0.063	0.116	0.007
4	1	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.043	0.134	0.042	0.006
5	1	ILUMINACION_AMANECER_SIN_ARTIFICIAL	1.189	0.032	0.173	0.005
6	1	METEO_NUBLADO	1.083	0.043	0.080	0.003
7	2	SEXO	1.069	0.921	0.067	0.061
8	2	ILUMINACION_NOCHE_SIN_LUZ	1.335	0.058	0.289	0.017
9	2	VISIB_OTRAS_RESTRICCIONES	1.177	0.057	0.163	0.009
10	2	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.045	0.128	0.044	0.006
11	2	CONSUMO_SUSTANCIAS	1.032	0.170	0.031	0.005
12	3	CONSUMO_SUSTANCIAS	1.077	0.353	0.074	0.026
13	3	ILUMINACION_NOCHE_SIN_LUZ	1.403	0.052	0.338	0.017
14	3	VISIB_OTRAS_RESTRICCIONES	1.212	0.064	0.192	0.012
15	3	ILUMINACION_NOCHE_CON_ARTIFICIAL	1.074	0.133	0.071	0.009
16	3	METEO_NUBLADO	1.202	0.041	0.184	0.007

Fuente: Elaboración propia (con Python)

Los resultados muestran que los principales determinantes de la severidad presentan patrones consistentes entre clústeres y subclústeres, destacando especialmente dos factores contextuales recurrentes: las condiciones de iluminación nocturna y las restricciones de visibilidad. En particular, la iluminación nocturna sin luz artificial aparece de forma sistemática entre los factores con mayor intensidad en varios grupos, mientras que la variable relacionada con otras restricciones de visibilidad también muestra efectos relevantes.

Por su parte, si bien la variable sexo aparece en varios grupos como *driver*, su presencia se explica principalmente por su elevada prevalencia dentro de la muestra, pues la mayor parte de los accidentes corresponden a conductores hombres. Sin embargo, sus *odds ratios* se sitúan muy próximos a 1, lo que indica que su efecto real sobre la probabilidad de severidad es reducido. Por este motivo, su aparición entre los *drivers* principales responde más a su frecuencia que a su capacidad explicativa, de tal manera que no se considera un determinante sustantivo de la severidad en el análisis posterior.

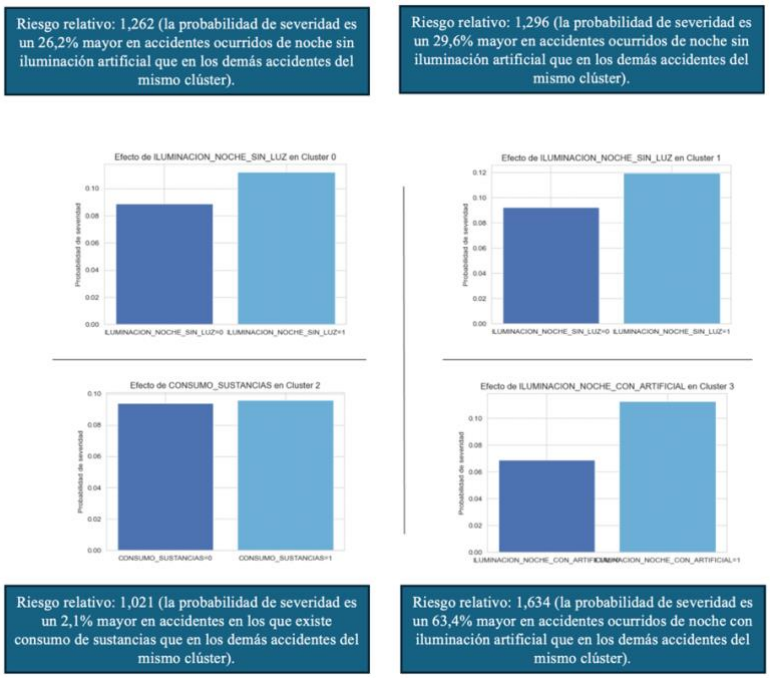
Por último, aparece el consumo de sustancias como predictor relevante en algunos grupos específicos, especialmente en el clúster 2 (siniestros de conductores jóvenes) y en el subclúster 3 (mujeres con consumo moderado), lo que resulta coherente con el perfil de riesgo asociado a estos grupos. En todo caso, la interpretación de este resultado debe realizarse con cautela, por haber sido construida la variable a partir de información procedente de un *dataset* complementario, como ya se ha explicado anteriormente.

No obstante, estas tablas únicamente permiten comparar la relevancia relativa de los distintos factores dentro de cada grupo, pero sin cuantificar directamente en qué medida cada factor aumenta la probabilidad de severidad. Por ello, a continuación, se aborda la segunda fase relativa al **análisis del impacto del *driver* más relevante en la probabilidad de severidad dentro de cada grupo**.

En esta fase, se estima directamente cuánto cambia la probabilidad de severidad cuando el factor está presente frente a cuando no lo está dentro de cada grupo. Para ello, se calcula la probabilidad observada de severidad en los accidentes en los que la variable está presente y se compara con la probabilidad correspondiente en los casos en los que dicha característica no aparece.

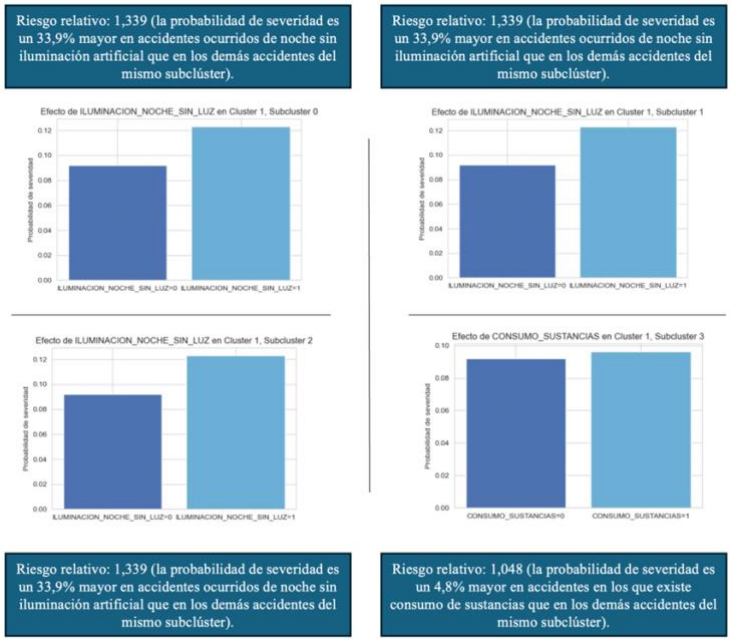
A partir de esta comparación, se calcula el riesgo relativo, que indica en qué medida aumenta la probabilidad de severidad cuando el factor está presente frente a cuando está ausente dentro del mismo clúster. Los resultados se presentan en las figuras siguientes.

Figura 18: Variación de la probabilidad de severidad asociada al principal driver de cada clúster



Fuente: Elaboración propia (con Python)

Figura 19: Variación de la probabilidad de severidad asociada al principal driver de cada subclúster



Fuente: Elaboración propia (con Python)

Estos resultados muestran que las condiciones de iluminación nocturna, especialmente en ausencia de luz artificial, constituyen el *driver* más importante del aumento de la severidad, con incrementos relativos de la probabilidad de severidad cercanos al 20% en grupos del *clustering* y al 34% en grupos del *subclustering*.

Asimismo, en el clúster 3, asociado a contextos de clima extremo con presencia de hielo o nieve, destaca el efecto de la iluminación nocturna con luz artificial, lo que sugiere que la combinación de condiciones ambientales desfavorables y circulación nocturna puede intensificar significativamente la gravedad de los siniestros.

Por el contrario, el consumo de sustancias, aunque aparece como determinante en determinados perfiles, presenta un impacto mucho más reducido sobre la probabilidad de severidad, lo que confirma su menor capacidad explicativa en comparación con los factores relacionados con las condiciones de visibilidad.

En conclusión, los resultados ponen de manifiesto que los factores asociados a la iluminación nocturna, especialmente en ausencia de luz artificial, constituyen el principal determinante del aumento de la severidad dentro de la mayor parte de los perfiles de siniestro.

Finalmente, en la tercera fase del apartado y en línea con lo anterior, se incorpora una **simulación *what-if* para estimar la reducción potencial de la severidad de los siniestros** en caso de que disminuyera la incidencia del principal factor explicativo de la gravedad de los accidentes. Este ejercicio permite trasladar los resultados a un plano aplicado y resulta útil para orientar decisiones de política pública en materia de seguridad vial, pues refleja el impacto potencial de actuaciones dirigidas a reducir la presencia de los principales factores de riesgo identificados.

En este contexto, se analiza, dentro de los clústeres 0 y 1 y de los subclústeres 0, 1 y 2, el efecto que tendría una reducción del 50% en la prevalencia de la variable de iluminación nocturna sin luz artificial, identificada como el *driver* más relevante en esos grupos.

Para ello, se parte de las probabilidades observadas de severidad cuando la variable está ausente y cuando está presente dentro de cada grupo.

A partir de estas probabilidades, se define, por un lado, la prevalencia observada del factor de riesgo y, por otro, una prevalencia simulada, obtenida al reducir en un 50% su frecuencia de aparición.

Con estos valores, se estima la severidad actual y la severidad simulada de cada grupo como media ponderada de las probabilidades de severidad en presencia y ausencia del factor, utilizando como pesos, respectivamente, la prevalencia observada y la simulada.

A continuación, se compara la severidad actual con la severidad simulada para obtener el cambio absoluto asociado a la reducción hipotética del *driver*. Además, con el fin de dotar de una interpretación más aplicada a los resultados, esa diferencia se traduce en un número estimado de accidentes severos reducidos, calculado multiplicando la reducción de severidad por el tamaño muestral de cada grupo analizado.

Tabla 18: Reducción estimada de la severidad en el escenario what-if por clúster y subclúster

	Cluster	Prevalencia actual	Prevalencia nueva	Severidad actual	Severidad simulada	Cambio absoluto	Número de accidentes severos reducidos
0	0	0.1059	0.0529	0.0914	0.0901	-0.0012	30.8058
1	1	0.0520	0.0260	0.0937	0.0930	-0.0007	139.9164

	Cluster	Subcluster	Prevalencia actual	Prevalencia nueva	Severidad actual	Severidad simulada	Cambio absoluto	Número de accidentes severos reducidos
0	1	0	0.0527	0.0264	0.0974	0.0965	-0.0009	8.9879
1	1	1	0.0482	0.0241	0.0936	0.0930	-0.0006	42.6447
2	1	2	0.0582	0.0291	0.0935	0.0927	-0.0008	37.2895

Fuente: Elaboración propia (con Python)

Los resultados muestran que, bajo este escenario, la reducción del 50% en la presencia de accidentes ocurridos de noche sin iluminación artificial produciría una disminución apreciable, aunque no muy elevada, de la severidad esperada en todos los grupos considerados.

Si bien la estimación apunta a que podrían evitarse alrededor de 260 accidentes severos, lo que en términos humanos supone un impacto muy relevante al implicar una disminución en 260 familias afectadas por siniestros graves, la magnitud de la reducción resulta limitada si se compara con el tamaño total del *dataset*.

Ello probablemente se debe a que la severidad de los accidentes es un fenómeno inherentemente multifactorial, condicionado simultáneamente por diversas variables como la velocidad de circulación, el tipo de vía, las condiciones meteorológicas o el comportamiento del conductor. De esta forma, aunque la conducción nocturna sin iluminación artificial actúe como un determinante relevante dentro de los grupos analizados, su reducción solo puede generar mejoras parciales, siendo necesario actuar de forma conjunta sobre todos los factores de riesgo para lograr descensos más trascendentes en la siniestralidad vial grave.

Además, esta aparente limitación del efecto agregado puede explicarse por el hecho de que la falta de iluminación presenta un impacto elevado sobre la severidad individual del siniestro, como refleja su alto *odds ratio*, pero afecta únicamente a una proporción relativamente reducida del total de accidentes. Se trata, por tanto, de un factor de riesgo muy intenso pero poco frecuente, lo que contribuye a explicar que su reducción produzca mejoras relevantes en términos individuales, pero más moderadas cuando se analizan sobre el conjunto global de los accidentes de tráfico.

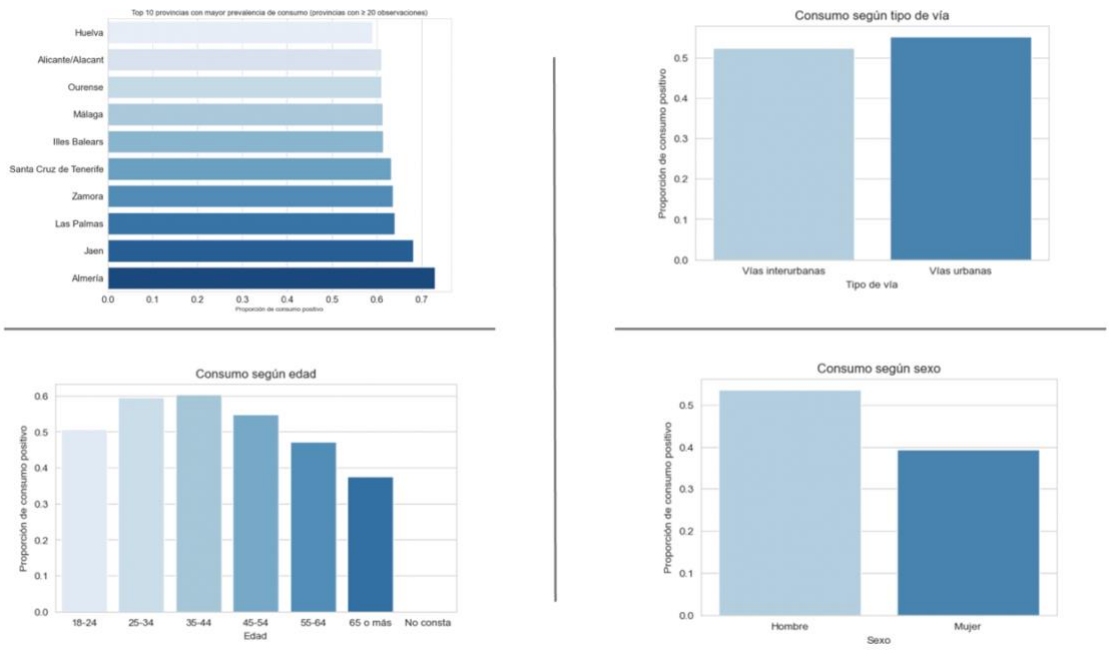
3.2.5. Análisis del consumo de sustancias en siniestros mortales

Para cerrar el capítulo, se incorpora un análisis específico de la prevalencia del consumo de sustancias en conductores implicados en siniestros mortales, a partir del *dataset* original procedente del portal "Justicia en datos". Este análisis permite complementar la caracterización de los factores asociados a la severidad mediante la incorporación de una variable especialmente relevante en la siniestralidad mortal.

En línea con el desarrollo metodológico previamente descrito, se construye una variable binaria de resultado toxicológico positivo que permite estimar, en primer lugar, la prevalencia global del consumo y, posteriormente, su distribución según distintas dimensiones relevantes, como la provincia, el tipo de vía, el sexo y el grupo de edad.

Finalmente, el análisis conjunto de estas variables permite extraer perfiles combinados de mayor riesgo, con el objetivo de caracterizar los contextos y colectivos en los que la presencia de consumo resulta más frecuente dentro de los siniestros mortales analizados. El código para la elaboración de este análisis se encuentra recogido en el Anexo 5 y los principales resultados se muestran a continuación.

Figura 20: Distribución de la prevalencia del consumo de sustancias en siniestros mortales según provincia, tipo de vía, edad y sexo



Fuente: Elaboración propia (con Python)

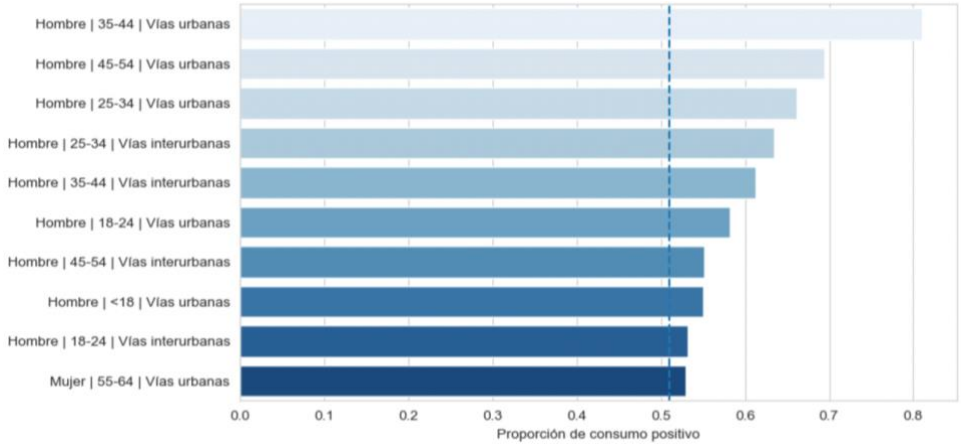
Desde el punto de vista territorial, se observan diferencias apreciables entre provincias españolas, con prevalencias especialmente elevadas en Almería, Jaén, Las Palmas o Zamora. Sin embargo, estos resultados deben interpretarse con cuidado por la heterogeneidad en el tamaño muestral provincial.

Desde el punto de vista del tipo de vía, la prevalencia de consumo resulta ligeramente superior en vías urbanas (55,1%) que en vías interurbanas (52,4%), lo que sugiere una mayor presencia de conductores con resultado positivo en ciudad.

Desde el punto de vista del análisis sociodemográfico, se presentan diferencias más marcadas. En concreto, la prevalencia de consumo es claramente superior en hombres (53,5%) que en mujeres (39,3%), lo que confirma la existencia de un patrón diferencial por sexo ya observado en la literatura sobre siniestralidad vial grave. Por grupos de edad, el consumo presenta sus valores más elevados en los intervalos de 25 a 34 años (59,5%) y de 35 a 44 años (60,3%), descendiendo progresivamente en los grupos de mayor edad hasta situarse en torno al 37,5% en mayores de 65 años.

Finalmente, el análisis combinado de sexo, edad y tipo de vía permite identificar perfiles con mayor prevalencia de consumo de sustancias.

Figura 21: Top 10 perfiles con mayor prevalencia de consumo



Fuente: Elaboración propia (con Python)

Con base en la Figura 21, destacan los hombres de entre 35 y 44 años en vías urbanas, con prevalencias superiores al 80%, seguidos por otros perfiles masculinos en edades intermedias tanto en vías urbanas como interurbanas.

En consecuencia, la identificación de estos perfiles permite orientar con mayor precisión las actuaciones preventivas en materia de seguridad vial, facilitando a las autoridades la priorización de campañas de control y sensibilización dirigidas a los colectivos con mayor prevalencia de consumo. En este sentido, la concentración de valores elevados en hombres de edad adulta y en entornos urbanos sugiere la conveniencia de reforzar las medidas de vigilancia, los controles preventivos y las estrategias de concienciación en estos contextos específicos.

3.3. Visualización analítica de los patrones de severidad

Para terminar, con el objetivo de sintetizar los principales hallazgos del análisis y trasladar las conclusiones más relevantes a los agentes públicos implicados en el diseño de políticas de seguridad vial, se integran los resultados obtenidos en una infografía final.

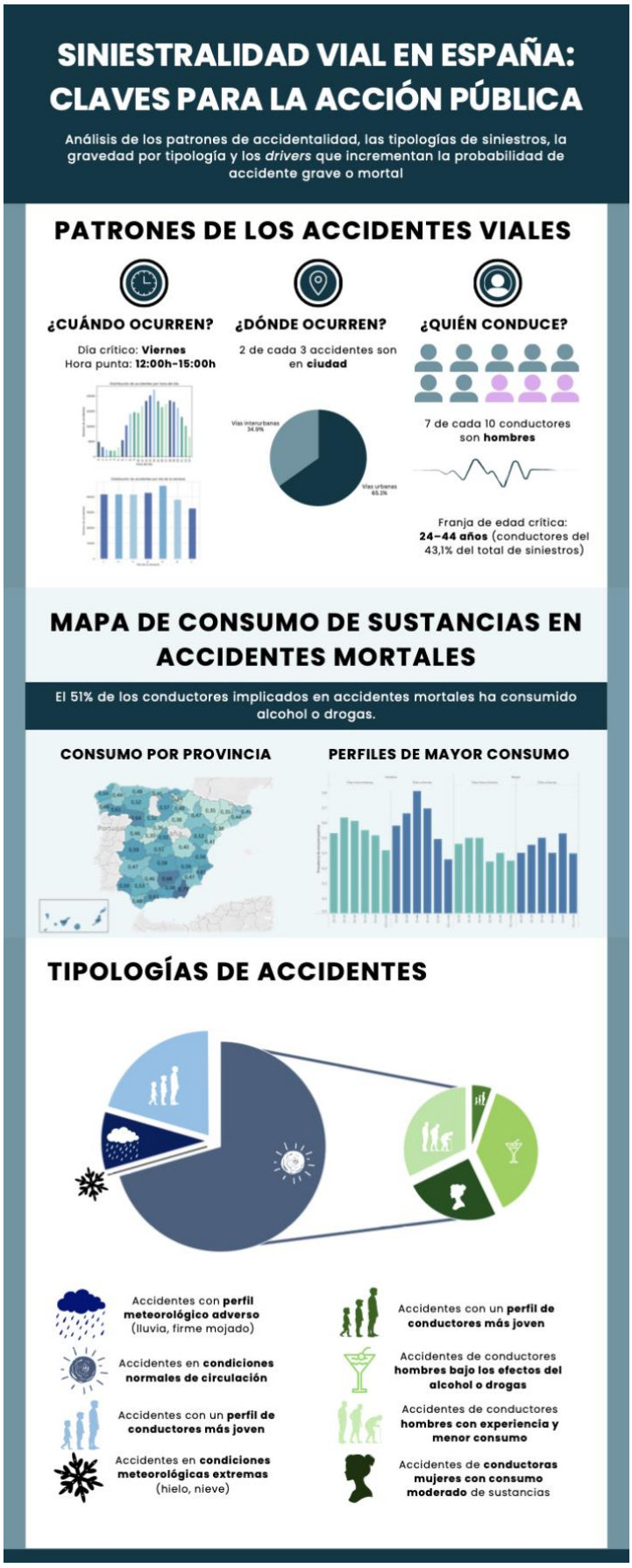
La infografía, realizada en la plataforma *Canva*, presenta de forma estructurada los principales ejes del análisis: los patrones generales de accidentalidad, la distribución del consumo de sustancias en accidentes mortales, la segmentación de tipologías de siniestros

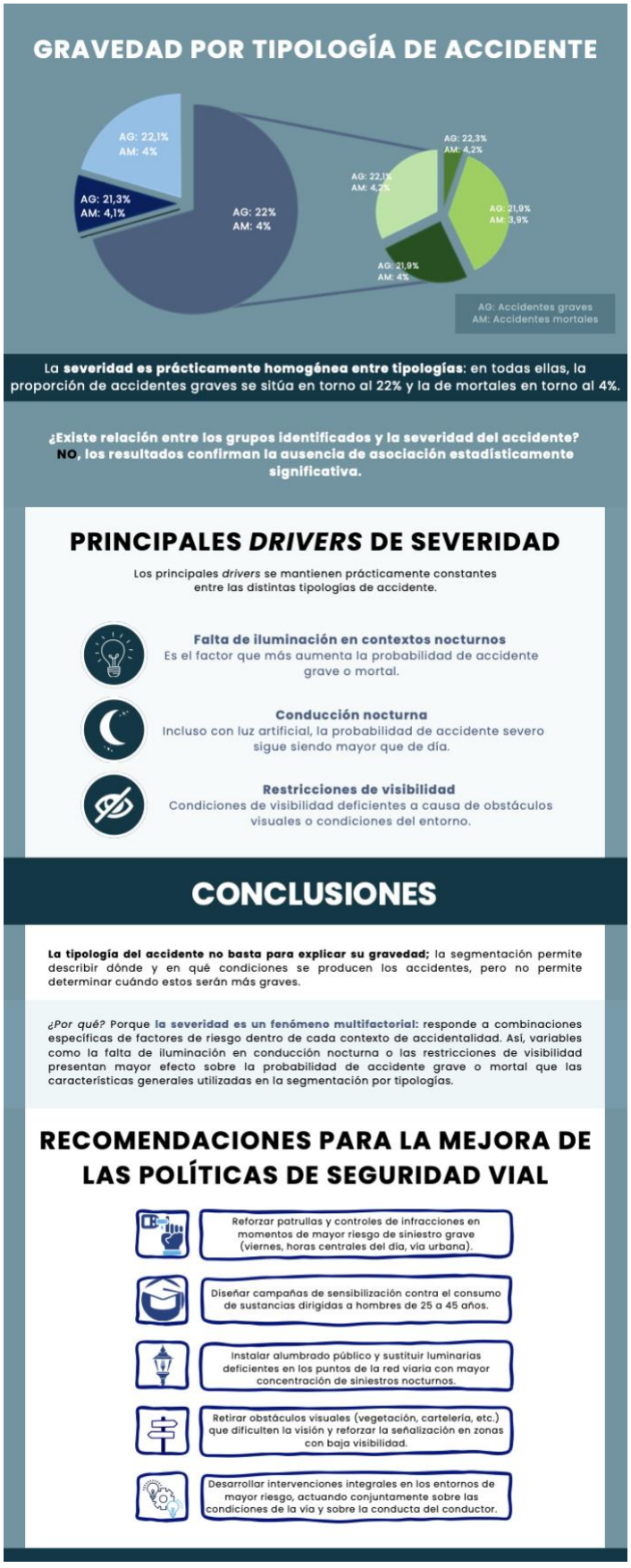
identificadas mediante el *clustering*, la comparación de la severidad entre tipologías, la identificación de los principales *drivers* asociados a la severidad vial y, finalmente, un conjunto de recomendaciones orientadas a la intervención pública.

Desde el punto de vista técnico, los elementos gráficos que la integran proceden de distintas herramientas de visualización utilizadas a lo largo del análisis; algunos gráficos se han generado directamente en *Canva*, otros se han elaborado en *PowerPoint* y otros se han desarrollado en *Tableau*. Esta combinación de herramientas ha permitido optimizar tanto la precisión analítica de los gráficos como su claridad comunicativa.

La elaboración de la infografía responde, por tanto, a un doble objetivo: por un lado, integrar en un único soporte los resultados más relevantes del análisis; y, por otro, apoyar al diseño de intervenciones prioritarias en materia de seguridad vial.

Figura 22: Infografía de síntesis de resultados





Fuente: Elaboración propia (con Canva)

CAPÍTULO 4. RESULTADOS, DISCUSIÓN Y CONCLUSIONES

4.1. Conclusiones del estudio

4.1.1. Conclusiones en comparación con los objetivos planteados

Para evaluar el cumplimiento de los objetivos establecidos al inicio del trabajo, se presenta a continuación una síntesis comparativa entre los objetivos específicos definidos en el apartado 1.2 y los resultados efectivamente alcanzados.

Tabla 19: Comparación entre objetivos y resultados

OBJETIVOS PLANTEADOS	RESULTADOS OBTENIDOS
Construir una base de datos con más de 100.000 observaciones que integre variables contextuales, temporales y personales.	Se ha construido una base de datos analítica integrada y depurada con más de 250.000 registros. Si bien, inicialmente, el *dataset* no contenía variables personales, estas han sido incorporadas mediante procesos de integración.
Identificar al menos tres clústeres de siniestros viales, segmentando los accidentes en tipologías diferenciadas.	Se han identificado cuatro tipologías diferenciadas de siniestros y, dentro del grupo mayoritario, se ha realizado una segmentación adicional que ha permitido distinguir cuatro subtipologías específicas.
Comparar las tasas de severidad entre los clústeres, así como evaluar el efecto de la pertenencia a cada clúster sobre la probabilidad de severidad.	Se ha realizado un análisis estadístico mediante medidas de asociación y regresión logística, que muestra que la pertenencia a los clústeres no presenta una asociación estadísticamente significativa con la severidad.
Identificar, dentro de cada clúster, las cinco variables con mayor incidencia sobre la probabilidad de severidad, para detectar los principales drivers de gravedad en cada tipología de accidente.	Se han identificado los cinco *drivers* principales de severidad en cada clúster y subclúster; se ha analizado el impacto del *driver* más relevante en la probabilidad de severidad dentro de cada grupo; y se ha realizado una simulación *what-if* que estima la reducción potencial de la severidad si disminuyera la incidencia del principal factor explicativo.
Desarrollar visualizaciones analíticas que plasmen los principales resultados obtenidos a lo largo del análisis.	Se ha diseñado una infografía que sintetiza los principales hallazgos del estudio y facilita la interpretación de los resultados desde una perspectiva orientada a la toma de decisiones.

Fuente: Elaboración propia

En términos generales, los resultados permiten afirmar que los objetivos definidos al inicio del trabajo han sido alcanzados y, de igual manera, que el objetivo general del trabajo, consistente en desarrollar un modelo analítico de severidad vial capaz de identificar tipologías de accidente y analizar cómo varía la severidad de cada tipo, también se ha cumplido.

4.1.2. Conclusiones en comparación con la literatura previa

Para evaluar la coherencia de los resultados obtenidos y extraer posibles aportaciones del trabajo, se presenta una segunda comparación entre los hallazgos del presente análisis y las conclusiones recogidas en la literatura académica analizada en el capítulo introductorio.

Tabla 20: Comparación entre la literatura previa y los resultados del análisis

CONCLUSIONES DE LA LITERATURA	CONCLUSIONES DEL PRESENTE ESTUDIO
Úbeda González (Universidad Miguel Hernández): La gravedad de los siniestros se asocia con variables como la edad del conductor, la hora del día, las condiciones de luminosidad o el estado de la calzada.	El presente análisis confirma la relevancia de factores relacionados con la luminosidad, pero da menor importancia a la edad del conductor y a la hora del día. El estado de la calzada, por su parte, no ha sido analizado.
Herrera Briones (Universidad Politécnica de Madrid): Los principales factores asociados a la severidad son el tipo de colisión y las infracciones del conductor, mientras que la influencia de la luminosidad y las condiciones atmosféricas resulta menor en algunos contextos.	Si bien el presente estudio no analiza ni el tipo de colisión ni las infracciones del conductor, sí afirma la falta de relevancia de las condiciones atmosféricas, aunque defiende la influencia de la luminosidad. En todo caso, ambos estudios muestran que la severidad constituye un fenómeno afectado por una gran variedad de factores explicativos.
Febres Eguiguren (Universidad de Burgos): La severidad aumenta en casos de conducción sin licencia o distracciones tecnológicas, así como disminuye con el uso del cinturón de seguridad. Asimismo, recalca la necesidad del estudio del impacto del consumo de sustancias al volante.	El presente estudio, aunque no analiza la conducción sin licencia, las distracciones tecnológicas o el uso de cinturón, confirma la relevancia del consumo de alcohol y drogas en la severidad de la siniestralidad vial.

Fuente: Elaboración propia

Como puede observarse, este Trabajo de Fin de Grado resulta complementario y, en determinados aspectos, consistente con los principales hallazgos de la literatura previa. En algunos casos, los resultados obtenidos coinciden con los identificados por los trabajos analizados, mientras que en otros aportan matices adicionales derivados de las variables consideradas y del enfoque metodológico empleado.

En todo caso, esta comparación refuerza la idea de que la severidad de los siniestros viales constituye un fenómeno multifactorial, condicionado por la interacción de variables personales, contextuales, infraestructurales y conductuales. En este sentido, cada una de las fuentes identifica factores explicativos parcialmente distintos, y no siempre coincidentes entre sí, reflejando la complejidad del fenómeno y la necesidad de abordarlo desde enfoques integradores.

4.2. Limitaciones del estudio y líneas futuras de análisis

Pese a la utilidad analítica de los resultados obtenidos, este estudio presenta una serie de limitaciones que conviene señalar, tanto para interpretar adecuadamente sus conclusiones como para orientar posibles líneas futuras del trabajo.

Estas limitaciones son las siguientes:

I. **El uso de variables personales y de consumo sintéticas**. La información relativa al conductor se incorporó mediante un procedimiento de asignación sintética basado en distribuciones agregadas oficiales. Si bien ello permite enriquecer el análisis al incorporar dimensiones personales, también obliga a interpretar con cautela la generalización de algunos resultados.

    a. Línea futura: establecer colaboración con la DGT para acceder a bases de microdatos completas, que incorporen variables personales del conductor.

II. **El alcance temporal, acotado al período entre 2021 y 2023.** Aunque este intervalo se seleccionó adecuadamente por su actualidad y homogeneidad, también limita la posibilidad de analizar cambios a largo plazo.

    a. Línea futura: extender el análisis a series temporales más largas.

III. **La presencia de valores desconocidos y la necesidad de imputación modal.** En varias variables del *dataset* original existían categorías no informativas, que se imputaron a la moda para evitar perder observaciones y preservar el tamaño muestral. Si bien esta decisión es metodológicamente razonable y conservadora, puede reducir la variabilidad real de los datos y afectar a la precisión del análisis.

    a. Línea futura: incentivar una recogida más completa por parte de la autoridad competente de toda la información relevante en el momento en que se produce el accidente.

IV. **La ausencia de variables potencialmente explicativas.** El presente trabajo concluye que la severidad vial es un fenómeno multifactorial, pero diversas variables que la literatura identifica como relevantes no estaban disponibles, como el estado de la calzada, las infracciones del conductor, el uso del cinturón de seguridad, la velocidad o determinadas distracciones al volante.

    a. Línea futura: construir modelos que combinen distintas fuentes e integren variables de comportamiento, protección y mecánica del siniestro.

V. **La separación débil entre subclústeres**. En el *subclustering* aplicado al grupo mayoritario, la curva del codo no mostraba un punto de inflexión claro y los valores del índice *silhouette* eran bajos, indicando que la separación entre subgrupos era limitada y que existía cierto solapamiento entre ellos.

    a. Línea futura: complementar el análisis con técnicas alternativas de segmentación, como el *clustering* jerárquico o los métodos basados en densidad.

VI. **La capacidad explicativa limitada de la pertenencia al clúster sobre la severidad**. El estudio concluye que las tasas de severidad son muy homogéneas entre tipologías y que no existe una asociación estadísticamente significativa entre pertenencia a clúster y severidad. Ello no invalida la utilidad de la segmentación, pero sí limita su valor como herramienta predictiva de la gravedad del siniestro.

    a. Línea futura: incorporar al *dataset* utilizado para el *clustering* variables adicionales con mayor capacidad explicativa de la severidad: variables de mecánica del accidente, de las medidas de protección utilizadas, etc.

En definitiva, si bien las limitaciones señaladas no invalidan los resultados del trabajo, delimitan su alcance y demuestran la necesidad de seguir avanzando hacia modelos más completos para el análisis de la severidad vial.

4.3. Aplicabilidad práctica del estudio

La principal utilidad práctica de este trabajo radica en que transforma un fenómeno complejo, como es la severidad de los accidentes de tráfico, en información analítica disponible para ser empleada en la toma de decisiones.

En este contexto, los resultados obtenidos pueden servir de apoyo al sector público para priorizar actuaciones en materia de seguridad vial, así como para diseñar campañas de educación y estrategias operativas más focalizadas y eficaces.

Además, esta aplicabilidad se proyecta en la infografía del apartado 3.3, que sintetiza los hallazgos técnicos del análisis en un formato claro y accesible para responsables públicos implicados en la prevención de la siniestralidad vial.

En última instancia, todo trabajo analítico aspira, de una u otra forma, a mejorar un beneficio o a minimizar un coste. En el caso de este estudio, su valor práctico reside precisamente en esto último: en ofrecer evidencia que permita minimizar el coste humano, social e institucional asociado a la severidad vial.

Si los resultados aquí obtenidos contribuyen, aunque sea parcialmente, a orientar mejores decisiones públicas y a reducir el número de vidas afectadas por los accidentes de tráfico, entonces el análisis habrá trascendido el plano académico para proyectarse sobre lo verdaderamente importante: la reducción del daño en carretera.

5. BIBLIOGRAFÍA

Alcaide, J. B. (2023). *Análisis del efecto de las políticas de intervención y descentralización en la seguridad vial: el caso de España en la Unión Europea* [Tesis de doctorado, Universidad de Sevilla]. Depósito de Investigación de la Universidad de Sevilla. https://idus.us.es/server/api/core/bitstreams/e0e49c81-36f6-4ad3-a4d5-e1a8d6d831a8/content

Chico Fernández, M., Llompart Pou, J. A., García Fuentes, C., Barea Mendoza, J. A., y Fernández Hervás, H. (2023). *Evolución temporal reciente e impacto de la pandemia COVID-19 en la enfermedad traumática grave (ETGE).* Fundación MAPFRE. https://documentacion.fundacionmapfre.org/documentacion/publico/es/media/group/1119716.do

Collado Tortosa, M. (2020). *Efectos en la salud de las víctimas de accidentes de tráfico* [Trabajo de Fin de Máster, Universidad Católica de Valencia]. Research Gate. https://www.researchgate.net/profile/Maria-Collado-Tortosa/publication/342819585_Los_efectos_en_la_salud_de_las_victimas_de_accidentes_de_trafico/links/5f076171299bf188160e95bb/Los-efectos-en-la-salud-de-las-victimas-de-accidentes-de-trafico.pdf

Dirección General de Tráfico. (2022). *Accidentes con víctimas – Tablas estadísticas 2021* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/qm691r

Dirección General de Tráfico. (2022). *Ficheros microdatos de accidentes con víctimas 2021* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/gu05v

Dirección General de Tráfico. (2024). *Ficheros microdatos de accidentes con víctimas 2022* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/b900n

Dirección General de Tráfico. (2024). *Ficheros microdatos de accidentes con víctimas 2023* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/chiiz

Dirección General de Tráfico. (2024). *Accidentes con víctimas – Tablas estadísticas 2022* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/1y5mn

Dirección General de Tráfico. (2024). *Accidentes con víctimas – Tablas estadísticas 2023* [Conjunto de datos]. DGT, Ministerio del Interior. https://n9.cl/plhw9

Dirección General de Tráfico. (2025). *Las principales cifras de la siniestralidad vial (2023).* Dirección General de Tráfico. https://www.dgt.es/menusecundario/dgt-en-cifras/dgt-en-cifras-resultados/dgt-en-cifras-detalle/Las-principales-cifras-de-la-siniestralidad-en-Espana-2023/

Echeburúa, E. y Esbec, E. (2015). Secuelas psíquicas en víctimas de accidentes de tráfico: un análisis clínico y pericial a la luz del nuevo baremo español de 2015. *Revista Española de Medicina Legal,* 41(4), 212-221. https://doi.org/10.1016/j.reml.2015.08.001

Febres Eguiguren, J. D. (2021). *Estudio de la lesividad de los accidentes de tráfico en España. Modelización de los factores técnicos y humanos*. [Tesis de doctorado, Universidad de Burgos]. Repositorio Institucional UBU. https://riubu.ubu.es/handle/10259/7786

Herrera Briones, J. (2021). *Análisis y predicción de la lesividad en accidentes de tráfico mediante la aplicación de Random Forest.* [Trabajo de Fin de Grado, Universidad Politécnica de Madrid]. Archivo Digital UPM. https://oa.upm.es/67548/

Llaneza Tabares, A. (2024). *Reducción de accidentes y mortalidad en la carretera. ¿Normativa o progreso técnico? Un análisis empírico.* [Trabajo de Fin de Grado, Universidad de Oviedo]. Repositorio Institucional de la Universidad de Oviedo. https://hdl.handle.net/10651/74471

Ministerio de Justicia (2025). *La justicia en datos: Accidentes de tráfico – Años 2021, 2022, 2023* [Conjunto de datos]. Ministerio de Justicia. https://n9.cl/lv56cn

Ministerio del Interior. (2014). *Orden INT/2223/2014, de 27 de octubre, por la que se regula la comunicación de la información al Registro Nacional de Víctimas de Accidentes de Tráfico. Boletín Oficial del Estado*, (289), 97785–97817.

Sánchez Sánchez, M. P. (2024). *Políticas Preventivas de Seguridad Vial y Criminología Ambiental: Hacia un esquema analítico para la toma de decisiones basadas en evidencias.* [Tesis de doctorado, Universidad Miguel Hernández de Elche]. Repositorio RediUMH. https://hdl.handle.net/11000/35626

Sanjurjo-De-No, A., Arenas-Ramírez, B., Mira, J., & Aparicio-Izquierdo, F. (2020). Driver pattern identification in road crashes in Spain. *IEEE Access,* 8, 182014-182025. https://ieeexplore.ieee.org/document/9210610

Úbeda González, D. (2017). *Predicción de la severidad de accidentes de tráfico en la Red de Carreteras de España y Reino Unido mediante modelos estadísticos basados en Random Forest y Regresión Logística.* [Tesis de doctorado, Universidad Miguel Hernández de Elche]. Repositorio RediUMH. http://hdl.handle.net/11000/4536

World Health Organization. (2023). *Global status report on road safety 2023.* World Health Organization (WHO). https://iris.who.int/server/api/core/bitstreams/46275f9f-ef66-4892-8ddd-a496ef8c1b74/content

6. ANEXOS

Anexo 1: Código en Python para la carga e integración de los microdatos de la DGT

```
#Cargar los excels
dgt_2021 = pd.read_excel("TABLA_ACCIDENTES_21.xlsx")
dgt_2022 = pd.read_excel("TABLA_ACCIDENTES_22.xlsx")
dgt_2023 = pd.read_excel("TABLA_ACCIDENTES_23.xlsx")

#Integrar los excels
accidentes = pd.concat([dgt_2021, dgt_2022, dgt_2023], ignore_index=True)

len(accidentes)
```

```
289084
```

```
accidentes.head()
```

	ID_ACCIDENTE	ANYO	MES	DIA_SEMANA	HORA	COD_PROVINCIA	COD_MUNICIPIO	ISLA	ZONA	ZONA_AGRUPADA	CARRETERA	KM	SENTIDO_1F
0	1	2021	1	6	13	1	0	NaN	1	1	A-625	365.0	2
1	2	2021	1	2	8	1	1059	NaN	3	2	No inventariada	0.0	4
2	3	2021	1	3	19	1	1059	NaN	3	2	No inventariada	0.0	4
3	4	2021	1	1	9	1	0	NaN	1	1	A-3012	15.1	1
4	5	2021	1	2	13	1	1059	NaN	3	2	No inventariada	0.0	4

Fuente: Elaboración propia (con Python)

Anexo 2: Extracto del diccionario de variables del dataset integrado

	Variable	Descripción	Tipo	Categorías	Uso en el análisis
1	ID_ACCIDENTE	Identificador de cada accidente	Numérica discreta	Código único por accidente	No se utiliza; no aporta capacidad explicativa para el análisis de severidad de accidente vial.
2	ANYO	Año en el que ocurrió el accidente	Numérica discreta	2021, 2022, 2023	Sí se utiliza; permite analizar la evolución temporal de la gravedad
3	MES	Mes en el que ocurrió el accidente	Categórica codificada	1-12 (enero-diciembre)	Sí se utiliza; permite analizar la estacionalidad y patrones temporales de la severidad vial.
4	DIA_SEMANA	Día de la semana en el que ocurrió el accidente	Categórica codificada	1-7 (lunes-domingo)	Sí se utiliza; permite analizar diferencias de severidad entre días laborables y fines de semana.
5	HORA	Hora en la que ocurrió el accidente	Numérica discreta	0–23	Sí se utiliza; permite identificar franjas horarias de mayor severidad (noche, madrugada, etc.).
6	COD_PROVINCIA	Provincia donde ocurrió el accidente	Categórica codificada	1 = Araba/Álava; 2 = Albacete; 3 = Alicante/Alacant; 4 = Almería; 5 = Ávila; 6 = Badajoz; 7 = Balears, Illes; 8 = Barcelona; 9 = Burgos; 10 = Cáceres; 11 = Cádiz; 12 = Castellón/Castelló; 13 = Ciudad Real; 14 = Córdoba; 15 = Coruña, A; 16 = Cuenca; 17 = Girona; 18 = Granada; 19 = Guadalajara; 20 = Gipuzkoa; 21 = Huelva; 22 = Huesca; 23 = Jaén; 24 = León; 25 = Lleida; 26 = Rioja, La; 27 = Lugo; 28 = Madrid; 29 = Málaga; 30 = Murcia; 31 = Navarra; 32 = Ourense; 33 = Asturias; 34 = Palencia; 35 = Palmas, Las; 36 = Pontevedra; 37 = Salamanca; 38 = Santa Cruz de Tenerife; 39 = Cantabria; 40 = Segovia; 41 = Sevilla; 42 = Soria; 43 = Tarragona; 44 = Teruel; 45 = Toledo; 46 = Valencia/València; 47 = Valladolid; 48 = Bizkaia; 49 = Zamora; 50 = Zaragoza; 51 = Ceuta; 52 = Melilla	No se utiliza; variable administrativa con alto riesgo de sobreajuste.
7	COD_MUNICIPIO	Municipio donde ocurrió el accidente	Categórica codificada	Municipio de menos de 5000 habitantes = 0; Municipio de 5000 habitantes o más = Código de municipio normalizado por el INE (5 dígitos alfanuméricos, 2 para provincia+3 para municipio)	No se utiliza; excesiva granularidad y escasa generalización del modelo.
8	ISLA	Isla donde ocurrió el accidente	Categórica codificada	Vacío = No aplica; 1 = Mallorca; 2 = Menorca; 3 = Ibiza; 4 = Formentera; 5 = Cabrera; 6 = Conejera; 7 = Gran Canaria; 8 = Fuerteventura; 9 = Lanzarote; 10 = Tenerife; 11 = La Palma; 12 = Gomera; 13 = Hierro	No se utiliza; información geográfica específica no relevante para el análisis de severidad general.
9	ZONA	Zona en la que ocurrió el accidente.	Categórica codificada	1 = Carretera; 2 = Travesía; 3 = Calle; 4 = Autopista o autovía urbana	No se utiliza; se descarta en favor de ZONA_AGRUPADA para reducir la complejidad y mejorar la generalización del modelo.
10	ZONA_AGRUPADA	Zona en la que ocurrió el accidente. (Agregación de la Zona en 2 valores: vías interurbanas y vías urbanas)	Categórica binaria	1 = Vías interurbanas; 2 = Vías urbanas	Sí se utiliza; permite diferenciar la gravedad de los siniestros viales en función del contexto viario general.
11	CARRETERA	Denominación de la carretera en la que ocurrió el accidente	Categórica	Código de carreteras (A-x, N-x, etc.)	No se utiliza; variable excesivamente específica que dificulta la generalización del modelo.
12	KM	Punto quilométrico en el que ocurrió el accidente	Numérica discreta	Valores numéricos (km)	No se utiliza; información muy localizada sin capacidad analítica general.

Fuente: Elaboración propia

Anexo 3: Código en Python para el clustering mediante el algoritmo K-Means

```python
FINAL_K = 4

kmeans_final = KMeans(n_clusters=FINAL_K, n_init=20, random_state=RANDOM_STATE)
clusters = kmeans_final.fit_predict(X_model)

df["CLUSTER"] = clusters
print("\nDistribución de clústeres:")
print(df["CLUSTER"].value_counts().sort_index())
```

```
Distribución de clústeres:
0     25012
1    196561
2     56935
3       337
```

Fuente: Elaboración propia (con Python)

Anexo 4: Código en Python de los modelos de regresión logística

Clustering general

```python
datos = df.copy()

datos["ANYO"] = datos["ANYO"].astype(int)
datos["SEVERIDAD"] = datos["SEVERIDAD"].astype(int)
datos["CLUSTER"] = datos["CLUSTER"].astype(int)
```

1. Split entrenamiento y validación (Train: 2021-2022 / Test: 2023)

```python
train = datos[datos["ANYO"].isin([2021, 2022])].copy()
test  = datos[datos["ANYO"].isin([2023])].copy()

print("Observaciones entrenamiento:", len(train))
print("Observaciones validación:", len(test))

print("\nTasa de severidad en train:", round(train["SEVERIDAD"].mean(), 4))
print("Tasa de severidad en test:", round(test["SEVERIDAD"].mean(), 4))
```

```
Observaciones entrenamiento: 180937
Observaciones validación: 97908

Tasa de severidad en train: 0.0933
Tasa de severidad en test: 0.0947
```

2. Ajuste del modelo logístico

```python
modelo = smf.logit("SEVERIDAD ~ C(CLUSTER)", data=train).fit()

print(modelo.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.310113
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:              SEVERIDAD   No. Observations:               180937
Model:                          Logit   Df Residuals:                   180933
Method:                           MLE   Df Model:                            3
Date:                Tue, 14 Apr 2026   Pseudo R-squ.:               8.678e-05
Time:                        19:21:50   Log-Likelihood:                -56111.
converged:                       True   LL-Null:                       -56116.
Covariance Type:            nonrobust   LLR p-value:                    0.02092
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         -2.3387      0.027    -87.021      0.000      -2.391      -2.286
C(CLUSTER)[T.1]    0.0648      0.029      2.268      0.023       0.009       0.121
C(CLUSTER)[T.2]    0.0952      0.032      2.968      0.003       0.032       0.158
C(CLUSTER)[T.3]   -0.1462      0.240     -0.608      0.543      -0.617       0.325
==============================================================================
```

3. Odds ratios e intervalos de confianza

```python
coeficientes = modelo.params
intervalos = modelo.conf_int()
intervalos.columns = ["LI_logodds", "LS_logodds"]
```

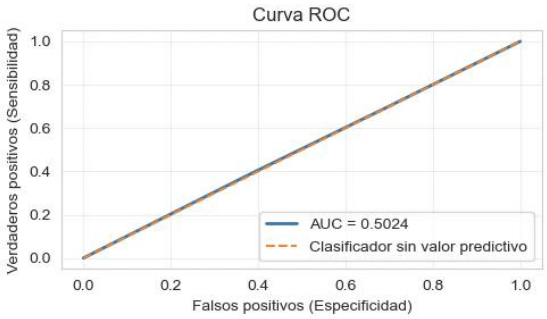
```
coeficientes = modelo.params
intervalos = modelo.conf_int()
intervalos.columns = ["LI_logodds", "LS_logodds"]

tabla_or = pd.DataFrame({
    "Odds Ratios": np.exp(coeficientes),
    "IC 2,5%": np.exp(intervalos["LI_logodds"]),
    "IC 97,5%": np.exp(intervalos["LS_logodds"]),
    "P-valor": modelo.pvalues
})

tabla_or = tabla_or.reset_index().rename(columns={"index": "Variable"})

print("\nOdds Ratios:")
print(or_table)
```

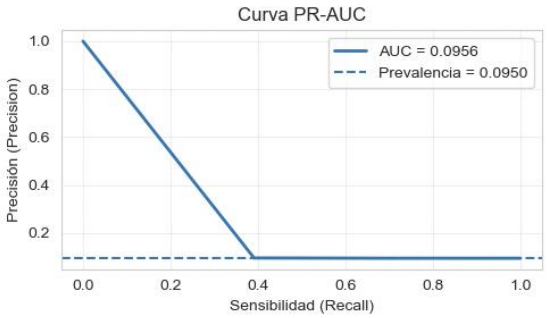
```
Odds Ratios:
          Variable  Odds Ratios   IC 2,5%  IC 97,5%   P-valor
0        Intercept     0.106390  0.101102  0.111955  0.000000
1  C(CLUSTER)[T.1]     0.973898  0.922587  1.028063  0.338193
2  C(CLUSTER)[T.2]     0.997535  0.938226  1.060593  0.937098
3  C(CLUSTER)[T.3]     1.087533  0.623922  1.895632  0.767239
```

4. Predicciones en test y métricas de evaluación en test (curvas ROC-AUC y PR-AUC)

```
test = test.copy()
test["y_true"] = test["SEVERIDAD"]
test["y_pred_prob"] = modelo.predict(test)
```

```
roc_auc = roc_auc_score(test["y_true"], test["y_pred_prob"])
pr_auc = average_precision_score(test["y_true"], test["y_pred_prob"])
brier = brier_score_loss(test["y_true"], test["y_pred_prob"])

falsos_positivos, verdaderos_positivos, umbrales_roc = roc_curve(test["y_true"], test["y_pred_prob"])
precision, recall, umbrales_pr = precision_recall_curve(test["y_true"], test["y_pred_prob"])

roc_df = pd.DataFrame({
    "falsos_positivos": falsos_positivos,
    "verdaderos_positivos": verdaderos_positivos
})

pr_df = pd.DataFrame({
    "precision": precision,
    "recall": recall
})


# Curva ROC
plt.figure(figsize=(5, 3))
plt.plot(falsos_positivos, verdaderos_positivos, linewidth=2, label=f"AUC = {roc_auc:.4f}")
plt.plot([0, 1], [0, 1], linestyle="--", linewidth=1.5, label="Clasificador sin valor predictivo")
plt.xlabel("Falsos positivos (Especificidad)")
plt.ylabel("Verdaderos positivos (Sensibilidad)")
plt.title("Curva ROC")
plt.legend(loc="lower right")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()

# Curva PR-AUC
baseline = test["y_true"].mean()

plt.figure(figsize=(5, 3))
plt.plot(recall, precision, linewidth=2, label=f"AUC = {pr_auc:.4f}")
plt.axhline(baseline, linestyle="--", linewidth=1.5, label=f"Prevalencia = {baseline:.4f}")
plt.xlabel("Sensibilidad (Recall)")
plt.ylabel("Precisión (Precision)")
plt.title("Curva PR-AUC")
plt.legend(loc="upper right")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()
```





Subclustering de las condiciones normales de circulación

```
data = df_sub.copy()
data.loc[:, "SUBCLUSTER_CONDICIONES_NORMALES"] = df_normal["SUBCLUSTER_CONDICIONES_NORMALES"]

data = data.dropna(subset=["ANYO", "SEVERIDAD", "SUBCLUSTER_CONDICIONES_NORMALES"]).copy()

data["ANYO"] = data["ANYO"].astype(int)
data["SEVERIDAD"] = data["SEVERIDAD"].astype(int)
data["SUBCLUSTER_CONDICIONES_NORMALES"] = data["SUBCLUSTER_CONDICIONES_NORMALES"].astype(int)
```

1. Split entrenamiento y validación (Train: 2021-2022 / Test: 2023)

```
train = data[data["ANYO"].isin([2021, 2022])].copy()
test  = data[data["ANYO"].isin([2023])].copy()

print("Observaciones train:", len(train))
print("Observaciones test:", len(test))

print("\nTasa de severidad en train:", round(train["SEVERIDAD"].mean(), 4))
print("Tasa de severidad en test:", round(test["SEVERIDAD"].mean(), 4))
```

```
Observaciones train: 85222
Observaciones test: 48218

Tasa de severidad en train: 0.0921
Tasa de severidad en test: 0.095
```

2. Ajuste del modelo logístico

```
modelo_sub = smf.logit(
    "SEVERIDAD ~ C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(reference=1))", data=train).fit()

print(modelo_sub.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.307396
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:              SEVERIDAD   No. Observations:                85222
Model:                          Logit   Df Residuals:                    85218
Method:                           MLE   Df Model:                            3
Date:                Tue, 14 Apr 2026   Pseudo R-squ.:               5.467e-06
Time:                        19:22:23   Log-Likelihood:                -26197.
converged:                       True   LL-Null:                       -26197.
Covariance Type:            nonrobust   LLR p-value:                    0.9626
================================================================================================================================
                                                                   coef    std err          z      P>|z|
[0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------------
Intercept                                                       -2.2815      0.020   -116.992      0.000      -
2.320      -2.243
C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(reference=1))[T.0]    -0.0112      0.056     -0.201      0.841      -
0.121       0.099
C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(reference=1))[T.2]    -0.0161      0.031     -0.527      0.599      -
0.076       0.044
C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(reference=1))[T.3]    -0.0059      0.028     -0.207      0.836      -
0.062       0.050
================================================================================================================================
```

3. Odds ratios e intervalos de confianza

```python
coeficientes = logit_model_sub.params
intervalos = logit_model_sub.conf_int()
intervalos.columns = ["LI_logodds", "LS_logodds"]

tabla_or_sub = pd.DataFrame({
    "Odds Ratios": np.exp(coeficientes),
    "IC 2,5%": np.exp(intervalos["LI_logodds"]),
    "IC 97,5%": np.exp(intervalos["LS_logodds"]),
    "P-valor": logit_model_sub.pvalues
})

tabla_or_sub = tabla_or_sub.reset_index().rename(columns={"index": "Variable"})

print("\n--- Odds Ratios ---")
print(or_table_sub)
```

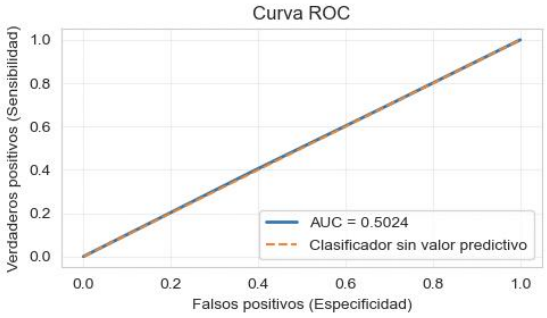
```
--- Odds Ratios ---
                                      variable  Odds Ratios   IC 2,5%  IC 97,5%   P-valor
0                                    Intercept     0.105054  0.101384  0.108857  0.000000
1  C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(r...     0.974343  0.878432  1.080725  0.622991
2  C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(r...     0.969477  0.915808  1.026292  0.286056
3  C(SUBCLUSTER_CONDICIONES_NORMALES, Treatment(r...     0.976876  0.926632  1.029844  0.385166
```

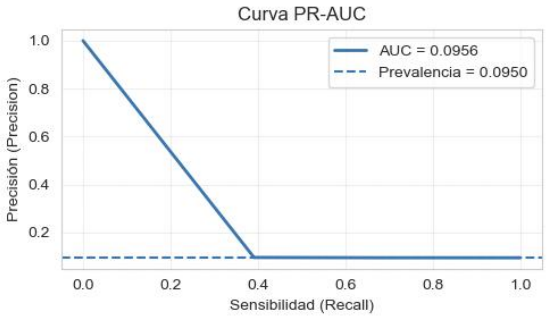
4. Predicciones en test y métricas de evaluación en test (curvas ROC-AUC y PR-AUC)

```python
test = test.copy()
test["y_true"] = test["SEVERIDAD"]
test["y_pred_prob"] = logit_model_sub.predict(test)


test = test.dropna(subset=["y_true", "y_pred_prob"]).copy()
test["y_pred"] = (test["y_pred_prob"] >= 0.5).astype(int)
```

```python
roc_auc = roc_auc_score(test["y_true"], test["y_pred_prob"])
pr_auc = average_precision_score(test["y_true"], test["y_pred_prob"])
brier = brier_score_loss(test["y_true"], test["y_pred_prob"])

falsos_positivos, verdaderos_positivos, umbrales_roc = roc_curve(
    test["y_true"], test["y_pred_prob"]
)
precision, recall, umbrales_pr = precision_recall_curve(
    test["y_true"], test["y_pred_prob"]
)
```

```
)

roc_df = pd.DataFrame({
    "falsos_positivos": falsos_positivos,
    "verdaderos_positivos": verdaderos_positivos
})

pr_df = pd.DataFrame({
    "precision": precision,
    "recall": recall
})

# Curva ROC
plt.figure(figsize=(5, 3))
plt.plot(falsos_positivos, verdaderos_positivos, linewidth=2, label=f"AUC = {roc_auc:.4f}")
plt.plot([0, 1], [0, 1], linestyle="--", linewidth=1.5, label="Clasificador sin valor predictivo")
plt.xlabel("Falsos positivos (Especificidad)")
plt.ylabel("Verdaderos positivos (Sensibilidad)")
plt.title("Curva ROC")
plt.legend(loc="lower right")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()

# Curva PR-AUC
baseline = test["y_true"].mean()

plt.figure(figsize=(5, 3))
plt.plot(recall, precision, linewidth=2, label=f"AUC = {pr_auc:.4f}")
plt.axhline(baseline, linestyle="--", linewidth=1.5, label=f"Prevalencia = {baseline:.4f}")
plt.xlabel("Sensibilidad (Recall)")
plt.ylabel("Precisión (Precision)")
plt.title("Curva PR-AUC")
plt.legend(loc="upper right")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.show()
```





Anexo 5: Código en Python del análisis del consumo de sustancias en siniestros viales

Carga del dataset original

```
df_analisis = pd.read_excel("DATASET_CONSUMO_SUSTANCIAS.xlsx")

display(df_analisis.head())

print("\nDimensión del dataset:")
print(df_analisis.shape)
```

	Año	Provincia	Zona	Rango edad	Sexo	Resultado toxicológico
0	2021	Alicante/Alacant	Vías interurbanas	55-64	Hombre	Positivo
1	2021	Alicante/Alacant	Vías interurbanas	55-64	Hombre	Positivo
2	2021	Illes Balears	Vías interurbanas	55-64	Mujer	Negativo
3	2021	Alicante/Alacant	Vías urbanas	45-54	Hombre	Positivo
4	2021	Castellón/Castelló	Vías interurbanas	55-64	Hombre	Positivo

```
Dimensión del dataset:
(3483, 6)
```

Creación de la variable binaria de consumo

```
df_analisis["consumo_positivo"] = (
    df_analisis["Resultado toxicológico"]
    .str.strip()
    .str.lower()
    .eq("positivo")
    .astype(int))


prevalencia_global = df_analisis["consumo_positivo"].mean()
print("Prevalencia global consumo:", round(prevalencia_global, 5))
```

```
Prevalencia global consumo: 0.51019
```

Análisis territorial

```
df_territorial = df_analisis[df_analisis["Resultado toxicológico"].isin(["Positivo", "Negativo"])].copy()

territorial = (df_territorial.groupby("Provincia")
        .agg(
        n_casos=("consumo_positivo", "size"),
        positivos=("consumo_positivo", "sum"),
        prevalencia=("consumo_positivo", "mean"))
    .sort_values("prevalencia", ascending=False)
    .reset_index())

print("Análisis territorial:")
print(territorial.head())

territorial_filtrado = territorial[territorial["n_casos"] >= 20]

top10_provincias = territorial_filtrado.sort_values("prevalencia", ascending=False).head(10)

plt.figure(figsize=(10,8))
sns.barplot(
    data=top10_provincias.sort_values("prevalencia"),
    x="prevalencia",
    y="Provincia",
    palette="Blues")
plt.title("Top 10 provincias con mayor prevalencia de consumo (provincias con ≥ 20 observaciones)")
plt.xlabel("Proporción de consumo positivo")
plt.ylabel(" ")
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.show()
```

```
Análisis territorial:
    Provincia  n_casos  positivos  prevalencia
0    Palencia       10          9     0.900000
1     Almería       37         27     0.729730
2        Jaen       44         30     0.681818
3  Las Palmas       50         32     0.640000
4      Zamora       22         14     0.636364
```

Análisis por tipo de vía

```
via = (
    df_analisis.groupby("Zona")
    .agg(
        n_casos=("consumo_positivo", "size"),
        positivos=("consumo_positivo", "sum"),
        prevalencia=("consumo_positivo", "mean"))
    .reset_index())

print("\nAnálisis por tipo de vía:")
print(via)


via_grafica = via[via["Zona"] != "No consta"]


plt.figure(figsize=(6,4))
sns.barplot(
    data=via_grafica,
    x="Zona",
    y="prevalencia",
    palette="Blues")
plt.title("Consumo según tipo de vía")
plt.xlabel("Tipo de vía")
plt.ylabel("Proporción de consumo positivo")
plt.show()
plt.show()
```

```
Análisis por tipo de vía:
                Zona  n_casos  positivos  prevalencia
0          No consta      715        310     0.433566
1  Vías interurbanas     2191       1149     0.524418
2       Vías urbanas      577        318     0.551127
```

Análisis sociodemográfico: en función del sexo y de la edad

```
sexo = (
    df_analisis.groupby("Sexo")
    .agg(
        n_casos=("consumo_positivo", "size"),
        positivos=("consumo_positivo", "sum"),
        prevalencia=("consumo_positivo", "mean")
    )
    .reset_index())

print("\nConsumo según sexo:")
print(sexo)

sexo_plot = sexo[sexo["Sexo"] != "No consta"]

plt.figure(figsize=(6,4))
sns.barplot(
    data=sexo_plot,
    x="Sexo",
    y="prevalencia",
    palette="Blues")
plt.title("Consumo según sexo")
plt.xlabel("Sexo")
plt.ylabel("Proporción de consumo positivo")
plt.show()



df_analisis["Rango edad"] = df_analisis["Rango edad"].fillna("No consta")
df_analisis["Rango edad"] = df_analisis["Rango edad"].replace(["nan", "NaN", ""],"No consta")


orden_edad = [
    "18-24",
    "25-34",
    "35-44",
    "45-54",
    "55-64",
    "65 o más",
    "No consta"]


edad = (
    df_analisis.groupby("Rango edad")
    .agg(
        n_casos=("consumo_positivo", "size"),
        positivos=("consumo_positivo", "sum"),
        prevalencia=("consumo_positivo", "mean"))
    .reset_index())
```

```
edad["Rango edad"] = pd.Categorical(
    edad["Rango edad"],
    categories=orden_edad,
    ordered=True)


edad = edad.sort_values("Rango edad")


print("\nConsumo según edad:")
print(edad)

edad_plot = edad[edad["Rango edad"] != "No consta"]

plt.figure(figsize=(7,4))
sns.barplot(
    data=edad_plot,
    x="Rango edad",
    y="prevalencia",
    palette="Blues")
plt.title("Consumo según edad")
plt.xlabel("Edad")
plt.ylabel("Proporción de consumo positivo")
plt.show()
```

```
Consumo según sexo:
        Sexo  n_casos  positivos  prevalencia
0     Hombre     2853       1527     0.535226
1      Mujer      626        246     0.392971
2  No consta        4          4     1.000000



Consumo según edad:
  Rango edad  n_casos  positivos  prevalencia
0      18-24      375        190     0.506667
1      25-34      499        297     0.595190
2      35-44      597        360     0.603015
3      45-54      643        352     0.547434
4      55-64      575        272     0.473043
5   65 o más      686        257     0.374636
7  No consta       36         21     0.583333
6        NaN       72         28     0.388889
```

Análisis combinado: tipo de vía + edad + sexo

```
df_perfil = df_analisis[
    (df_analisis["Sexo"] != "No consta") &
    (df_analisis["Rango edad"] != "No consta") &
    (df_analisis["Zona"] != "No consta")].copy()


perfil = (
    df_perfil.groupby(["Sexo","Rango edad","Zona"])
    .agg(
        n_casos=("consumo_positivo","size"),
        positivos=("consumo_positivo","sum"),
        prevalencia=("consumo_positivo","mean"))
    .reset_index())


perfil = perfil.sort_values(
    ["prevalencia","n_casos"],
    ascending=False)


print("\nPerfiles con mayor proporción de consumo:")
print(perfil.head(15))


perfil_filtrado = perfil[perfil["n_casos"] >= 15]


top10 = perfil_filtrado.head(10).copy()

top10["perfil"] = (
    top10["Sexo"]+ " | "+ top10["Rango edad"]+ " | "+ top10["Zona"])

plt.figure(figsize=(9,5))
sns.barplot(
    data=top10,
    x="prevalencia",
    y="perfil",
    palette="Blues")
plt.axvline(prevalencia_global, linestyle="--")
plt.title("Top perfiles con mayor prevalencia de consumo")
plt.xlabel("Proporción de consumo positivo")
plt.show()
```

Perfiles con mayor proporción de consumo:

	Sexo	Rango edad	Zona	n_casos	positivos	prevalencia
5	Hombre	35–44	Vías urbanas	74	60	0.810811
7	Hombre	45–54	Vías urbanas	72	50	0.694444
27	Mujer	<18	Vías urbanas	3	2	0.666667
3	Hombre	25–34	Vías urbanas	68	45	0.661765
2	Hombre	25–34	Vías interurbanas	274	174	0.635036
4	Hombre	35–44	Vías interurbanas	341	209	0.612903
1	Hombre	18–24	Vías urbanas	55	32	0.581818
6	Hombre	45–54	Vías interurbanas	375	207	0.552000
13	Hombre	<18	Vías urbanas	20	11	0.550000
0	Hombre	18–24	Vías interurbanas	190	101	0.531579
23	Mujer	55–64	Vías urbanas	17	9	0.529412
8	Hombre	55–64	Vías interurbanas	317	164	0.517350
18	Mujer	35–44	Vías interurbanas	54	27	0.500000
16	Mujer	25–34	Vías interurbanas	46	23	0.500000
19	Mujer	35–44	Vías urbanas	8	4	0.500000