UNIVERSIDAD NEBRIJA

MÁSTER DE ESTADÍSTICA APLICADA CON R SOFTWARE

ANÁLISIS DE FACTORES DE LESIVIDAD EN LOS ACCIDENTES DE TRÁFICO EN ESPAÑA

AUTOR: Marcelo Moreno Porras

DIRECTOR: Juan Luis López Garrancho

FECHA: 31/08/2021

ENTIDAD COLABORADORA: Máxima Formación

RESUMEN

Con el objetivo de hallar una relación entre la lesividad de los individuos implicados en accidentes de tráfico en España y las condiciones que se dieron en el accidente, se creó un modelo de regresión logística usando los datos de accidentes de la Dirección General de Tráfico del año 2015.

Los resultados del modelo de regresión logística mostraron que dicha relación existía y, que los factores determinantes de la lesividad de un individuo eran el número de individuos implicados en el mismo accidente, la posición del individuo en el momento del accidente y el número de vehículos implicados.

A través del modelo de regresión logística se obtuvo un buen clasificador de los individuos entre los grupos de ilesos y no ilesos. Sin embargo, los resultados presentaban margen de mejora y, el disponer de un mayor número de variables relativas al accidente como el consumo de drogas o el tipo de vehículo podrían ayudar a mejorar el ajuste.

Dedico a:

Todos aquellos docentes e investigadores que no han decaído en interés sobre su área de estudio, siempre apasionados por su trabajo y comprometidos con su labor de compartir conocimiento con el resto del mundo de la mejor forma posible.

Agradecimiento a:

Juan Luis López y Rosana Ferrero, profesores del Máster en Estadística Aplicada con R. Gracias por estar siempre disponibles para cualquier duda o problema, demostrar amor por vuestro campo de estudio y por compartir el conocimiento más puntero.

ÍNDICE

INTRODUCCIÓN

Cada año, cientos de miles de personas se ven implicadas en accidentes de tráfico en España, la mayoría de ellas no resultan ilesas del accidente. Conocer los factores que controlan o determinan la lesividad de los individuos podría ayudar a focalizar esfuerzos, tanto legislativos como económicos para minimizar el porcentaje de no ilesos (y reducir la gravedad de las heridas) en carreteras españolas.

La autoridad española encargada de velar por la seguridad vial, la Dirección General de Tráfico (DGT), publica anualmente datos sobre los accidentes de tráfico ocurridos durante el año seguido, usualmente de una breve nota de prensa con algunas de las conclusiones principales extraídas de un análisis exploratorio de los datos, por lo que, un análisis estadístico en profundidad, modelizando relaciones entre variables podría dar una foto más completa sobre la situación de seguridad vial en el país y hacer al público conocedor de conclusiones que puedan ser relevantes para reducir al mínimo el número de individuos no ilesos en accidentes de tráfico.

Para realizar el presente trabajo de investigación se usaron los datos de 2015 de accidentes de tráfico en España extraídos del portal estadístico de la DGT con el fin de encontrar una relación entre la lesividad de un individuo y los factores que se dieron en un accidente de tráfico. El número de observaciones y de variables era muy amplio, por lo que dichos datos pasaron por una serie de filtros y análisis exploratorios para poder utilizarlos en la modelización de las relaciones de interés y obtener los mejores resultados.

MATERIAL Y MÉTODOS

I. DESCARGA DE DATOS Y CRUCE DE TABLAS

Los datos usados para la presente investigación fueron descargados del portal estadístico de la Dirección General de Tráfico (DGT). El fichero de datos correspondía al año 2015. La razón de elegir datos de 2015 es que eran los únicos datos más reciente que contaban con una granularidad que a nivel de individuo implicado en el accidente (y a nivel de accidente).

El fichero de datos estaba compuesto por tres tablas: "TABLA_ACCVICT_2015", "TABLA_PERS_2015", "TABLA_VEHIC_2015". La tabla "TABLA_ACCVICT_2015" contenía datos a nivel de accidente sobre el año, el mes, la hora, el día de la semana, la provincia, la comunidad autónoma, el número de vehículos implicados, etc. La tabla "TABLA_ACCVICT_2015" contenía datos a nivel de individuo implicado en un accidente sobre la edad, el estado tras el accidente, el tipo de infracción cometida, posición, etc. La tabla "TABLA_VEHIC_2015" contenía datos a nivel de vehículo implicado en un accidente sobre ciertas características técnicas del mismo.

El portal estadístico de la DGT también proporcionaba el diseño de registro de los datos, guía para el entendimiento y codificación de las variables. Sin embargo, el diseño de registro que proporcionaba el portal estadístico ocupaba desde 2011 hasta 2013, por lo que durante la preparación de los datos se encontraron algunas variables que no se recogen o cuya codificación no coincide con la presentada el diseño de registro de los datos.

La tabla "TABLA_VEHIC_2015" fue desechada de la investigación debido a que todas sus variables no venían explicadas o codificadas correctamente en el diseño de registro.

La estructura de datos objetivo para proceder al análisis consistía en una tabla que fuese un cruce entre la tabla "TABLA_ACCVICT_2015" y "TABLA_PERS_2015", de modo que cada fila de la tabla constituyese un individuo que había estado implicado en un accidente. Para lograr esto, se realizó una consulta en lenguaje SQLite sobre las tablas "TABLA_ACCVICT_2015" y "TABLA_PERS_2015", donde la variable que relacionaba ambas tablas era "ID_ACCIDENTE".

```
SELECT * FROM ACCVICT LEFT JOIN PERS ON ACCVICT.ID_ACCIDENTE =
PERS.ID_ACCIDENTE
UNION ALL
SELECT * FROM PERS LEFT JOIN ACCVICT ON ACCVICT.ID_ACCIDENTE =
PERS.ID_ACCIDENTE WHERE ACCVICT.ID_ACCIDENTE IS NULL
```

La tabla "TABLA_ACCVICT_2015" contenía 97756 observaciones (accidentes) y 39 variables. La tabla "TABLA_PERS_2015" contenía 238476 observaciones (individuos implicados en un accidente) y 31 variables. El resultado de cruzar las dos tablas fue una única tabla con 238476 observaciones (individuos implicados en un accidente) y 70 variables.

A destacar, el funcionamiento de la tabla obtenida junto con una de sus variables principales, "ID_ACCIDENTE":

- Cada fila corresponde a un individuo implicado en un accidente.
- La variable "ID_ACCIDENTE" es un identificador único del accidente 2015XXXXXXXXX.
- Puede haber varios individuos implicados en un mismo accidente (tienen el mismo "ID_ACCIDENTE").
- Pueden haber ocurrido múltiples accidentes un mismo día, pero en localizaciones distintas, por lo que el "ID_ACCIDENTE" es distinto.

II. FILTRADO DE VARIABLES

De las 70 variables obtenidas tras el cruce de tablas, se procedió a realizar un filtrado de estas debido a las diversas razones que se muestran a continuación, especificando las variables desechadas:

- Eliminación por no estar incluidas en el diseño de registro:
  - COD_MUNICIPIO
  - USO_CINTURON
  - USO_SRI
  - USO_CASCO
  - DICCIONARIO_MANIOBRAS
  - INFRACC_COND
  - INFRACC_APERTURA
  - INFRACC_ALUMBRADO
  - INFRACC_CARGA_VEHICULO
  - INFRACC_RESUMEN
  - DICCIONARIO_ACCION_PEATON
- Eliminación de variables cuyo rango de valores no coincidían con los del diseño de registro:
  - TIPO_ACCIDENTE
  - MANIOBRAS
  - SEXO
  - ACCION_PEATON
- Eliminación de variables superfluas y duplicadas:
  - ANIO (el único valor que toma es 2015)
  - ANIO [1] (duplicado de ANIO)
  - ID_ACCIDENTE [1] (duplicado de ID_ACCIDENTE)
- Otros motivos:
  Solo interesaban las cifras y estado de víctimas, muertos o heridos graves que estuviesen consolidadas (30D), por lo que se eliminaron las variables con la cifra no consolidada:
  - TOT_VICTIMAS
  - TOT_MUERTOS
  - TOT_HERIDOS_GRAVES
  - TOT_HERIDOS_LEVES
  - MUERTO_24H
  - HERIDO_GRAVE_24H

- HERIDO_LEVE_24H

Se añadió una nueva variable "PRIORIDAD", codificada a partir de los valores binarios de las variables de prioridad. Se eliminaron las variables originales:

- PRIORIDAD_AGENTE
- PRIORIDAD_SEMAFORO
- PRIORIDAD_STOP
- PRIORIDAD_CEDA
- PRIORIDAD_MARCAS
- PRIORIDAD_PASO
- PRIORIDAD_OTRA

De las 70 variables que se tenían inicialmente, tras el filtrado y eliminación de variables, la cifra bajó a 39. Adicionalmente, se creó una nueva variable, "ILESO_30D", a partir de las variables de estado (herido, grave, muerto) consolidadas (30D). El número de variables con la introducción de "ILESO_30D" resultó ser de 40.

Para evitar problemas de multicolinealidad se descartaron ciertas variables que eran combinaciones lineales de otras: "TOT_MUERTOS30D", "TOT_HERIDOS_GRAVES30D", "TOT_HERIDOS_LEVES30D", "MUERTO_30D", "HERIDO_GRAVE30D", "HERIDO_LEVE30D".

La reducción de dimensiones tras esta etapa fue considerable, se ha pasó de tener 70 variables a 34.

III. ANÁLISIS E IMPUTACIÓN DE VALORES PERDIDOS

El análisis e imputación de valores perdidos es una etapa de la preparación de los datos debe realizarse con suma precaución, la imputación de valores perdidos en variables que tienen un alto porcentaje de estos puede introducir gran cantidad de ruido en la muestra.[1] En los datos, muchas de las variables que tenían valores perdidos tenían otras que cumplían una función similar o que tenían más o menos granularidad que estas (por ejemplo, "MUNICIPIO" y "PROVINCIA"), por lo que, por este motivo y por evitar la introducción de ruido en la muestra, se decidió eliminar todas aquellas variables con más de un 5% de valores perdidos.[2]

Tabla 1. Resumen de los valores perdidos del conjunto de datos de accidentados

Variable	Perdidos	%
ACERAS	211590	89
VISIBILIDAD_RESTRINGIDA	210640	88
PRIORIDAD	185427	78
CARRETERA	145230	61
ANIO_PERMISO	39947	17
TRAZADO_NO_INTERSEC	36853	15
MUNICIPIO	25763	11

[1] INSTITUTO NACIONAL DE ESTADÍSTICA: "Métodos de inferencia estadística con datos faltantes", ESTADÍSTICA ESPAÑOLA, *Estudio de simulación sobre los efectos en las estimaciones*, 2006.
[2] SCHAFER, J.: "Multiple imputation: a primer", SAGE Publications Ltd, *Statistical Methods in Medical Research*, Volume 8 Issue 1, febrero 1999. Disponible en Web: https://doi.org/10.1177%2F096228029900800102.

FACTORES_ATMOSFERICOS	11248	5
EDAD	8601	4
SUPERFICIE_CALZADA	7701	3
POSICION	172	0

Tras la eliminación de las variables con más de un 5% de valores perdidos, el número de variables del conjunto de datos se redujo a 27.

Para las variables restantes con un 5% de valores perdidos o menos, se procedió a predecir los valores faltantes mediante la función *aregImpute* del paquete *Hmisc*. Esta función hizo una imputación múltiple usando Booststrap y PMM (Predictive Mean Matching), identificando automáticamente el tipo de variable y tratándolas correctamente según su tipología de cara a la predicción de los valores faltantes.

Tabla 2. Resumen de la predicción de los valores faltantes con la función aregImpute

Multiple Imputation using Bootstrap and PMM			
formula: ~FACTORES_ATMOSFERICOS + EDAD + SUPERFICIE_CALZADA + POSICION			
n: 238476	p: 4	Imputations: 5	nk: 3
Number of Nas:			
FACTORES_ATMOSFERICOS	EDAD		
11248	8601		
SUPERFICIE_CALZADA	POSICION		
7701	172		
Transformation of Target Variables Forced to be Linear			
R-Squares for Predicting Non-Missing Values for Each Variable			
Using Last Imputations of Predictors			
FACTORES_ATMOSFERICOS	EDAD		
0.690	0.078		
SUPERFICIE_CALZADA	POSICION		
0.687	0.196		

IV. ANÁLISIS Y GESTIÓN DE VALORES ATÍPICOS A NIVEL MULTIVARIANTE

Tras la imputación de los valores perdidos se procedió al análisis y gestión de valores atípicos a nivel multivariante. Esta fue una etapa de suma importancia, ya que, según la teoría estadística, una sola observación atípica puede distorsionar arbitrariamente la estimación de los parámetros poblacionales.

Para la detección de valores atípicos a nivel multivariante se hizo uso de la distancia de Mahalanobis. Con este método, el número de valores atípicos identificados ascendía a los 3927, un 1.64% del total de datos. La mayoría de estos valores correspondían a personas implicadas en accidentes con un gran número de víctimas o vehículos.

Finalmente, se consideró que el porcentaje de datos atípicos a nivel multivariante era tan pequeño, que la mejor solución era omitirlos del conjunto de datos.

V. ANÁLISIS GRÁFICO Y DESCRIPTIVO DE LAS VARIABLES CON RESPECTO A LA VARIABLE PRINCIPAL

En esta etapa se analizaron una a una todas las variables de relevante importancia con respecto a la variable principal del estudio "ILESO_30D". Se usaron múltiples técnicas gráficas para el análisis según el tipo de variable (gráfico de caja para variables continuas, gráficos de barras para variables categóricas, etcétera) y tablas cruzadas en ciertos casos.

Según varios informes de la DGT, la causa número uno de accidente en los últimos años han sido las distracciones, le siguen la velocidad excesiva o inadecuada y el alcohol. Existían, además, otros factores importantes como la disminución de la capacidad para prestar atención debida al entorno.[3] [4]

En cuanto a la lesividad, investigaciones han hallado evidencia de que los daños que sufría un individuo en un accidente dependían, en gran medida, de la posición en la que se encontraba este dentro del vehículo (dando un resultado más favorable para los individuos que se encontraban en los asientos traseros).[5] Los resultados de esta etapa ayudaron, junto con la teoría, a elegir las variables de las que se compondría el modelo.

VI. PREPARACIÓN DEL MODELO

Elección del tipo de modelo

Para poder modelizar la relación entre las condiciones que se dan en un accidente y la lesividad o no lesividad de los individuos implicados, se procedió a crear un modelo cuya variable dependiente sería "ILESO_30D". La variable dependiente era una variable binaria que podía tomar los valores "No ileso" o "Ileso", y el modelo que mejor se ajusta a este tipo de variable dependiente es un modelo de regresión logística. De modo que:

$$y_i = 0\ (No\ ileso)$$

$$y_i = 1\ (Ileso)$$

Donde la variable respuesta tomaría el valor 1 con probabilidad p_i y el valor 0 con probabilidad $1 - p_i$. Para relacionar la probabilidad directamente con los predictores sin que el resultado se salga del intervalo 0, 1, se recurrió a transformar la probabilidad con la función logit o log-odds para relacionarla con los predictores (el modelo de regresión logística asume que el logit de la probabilidad sigue un modelo lineal). De este modo:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_n * x_n$$

[3] DIRECCIÓN GENERAL DE TRÁFICO: *Las distracciones son la causa de uno de cada cuatro accidentes* [en línea], Notas de prensa DGT, 16 de septiembre 2019 [ref. 18 de agosto 2021]. Disponible en Web: https://www.dgt.es/es/prensa/notas-de-prensa/2019/Las-distracciones-son-la-causa-de-uno-de-cada-cuatro-accidentes.shtml.

[4] DIRECCIÓN GENERAL DE TRÁFICO: *Anuario estadístico de accidentes 2015* [en línea], Ministerio del interior, 2015 [ref. 19 de agosto 2021]. Disponible en Web: https://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/anuario-estadistico-de-accidentes/anuario-accidentes-2015.pdf

[5] SMITH, KSCHAFER, J.: "Passenger seating position and the risk of passenger death or injury in traffic crashes", PubMed, *Accident Analysis & Prevention*, Volume 36 Issue 2, marzo 2004. Disponible en Web: https://doi.org/10.1016/S0001-4575(03)00002-2.

La regresión logística presentó una serie de ventajas sobre otro tipo de regresiones como la lineal: no requería una distribución binomial y no requería de homocedasticidad, pero sí que existía el riesgo de que apareciesen problemas de sub o sobre dispersión, aunque estos problemas solían desaparecer cuando la cantidad de datos disponible era muy grande. Para lograr un buen ajuste del modelo de regresión logística, se requería de ausencia de multicolinealidad, independencia de observaciones y ausencia de valores atípicos.[6]

Preparación de la evaluación del modelo

Con motivo de testear las capacidades del modelo creado a través de validación cruzada, los datos se dividieron en datos de entrenamiento y datos de prueba. El conjunto de datos de entrenamiento conformó el 75% de las observaciones (175913) y se utilizó para crear el modelo de regresión logística. El conjunto de datos de prueba conformó el 25% de las observaciones (58636), y se utilizó para testear la capacidad predictiva del modelo.

Creación del modelo

A la hora de ajustar el modelo de regresión logística, inicialmente se incluyeron todas las variables y se usó la función *step()* del paquete *stats* para crear un conjunto de modelos del que la función extrajo automáticamente el modelo con mejor AIC (Akaike Information Criterion) que pudo encontrar. Sin embargo, el modelo obtenido, a pesar de tener el mejor AIC, sufría de severos problemas de multicolinealidad.

Tras obtener un mal resultado con la creación automática del modelo, se siguió la teoría y se tuvo en cuenta algunos de los resultados del modelo creado para crear un nuevo modelo algo más contenido en variables, con un AIC no tan bueno, pero con ausencia de multicolinealidad y con mucha mayor cantidad de coeficientes con significación individual.

Como se comentó anteriormente, según la teoría, la velocidad era uno de los factores determinantes de la causa de accidentes, por lo que, a priori, la inclusión de la variable "INFRACC_VELOCIDAD" en el modelo hubiese parecido necesaria. Sin embargo, la inclusión de dicha variable provocaba un problema de multicolinealidad con la variable "POSICION", por lo que se realizaron dos modelos, uno con "POSICION" y otro con "INFRACC_VELOCIDAD", y los mejores resultados se obtuvieron con el modelo con la variable "POSICION". Por este motivo, la variable "INFRACC_VELOCIDAD" fue descartada del modelo final.

Otra de las variables que sí parecía tener relevancia era "FACTORES_ATMOSFERICOS", pero del mismo modo que la infracción de velocidad, esta variable creaba un problema de multicolinealidad con "SUPERFICIE_CALZADA" y, después de realizar un modelo con cada una de las variables, se determinó que el modelo con "SUPERFICIE_CALZADA" arrojaba mejores resultados y se descartó la variable "FACTORES_ATMOSFERICOS" del modelo final.

El modelo final fue el mejor modelo con las mejores características que se pudo encontrar. Requirió 11 variables predictoras:

```
ILESO_30D ~ HORA + DIASEMANA + COMUNIDAD_AUTONOMA +
TOT_VEHICULOS_IMPLICADOS + TOT_VICTIMAS30D + TIPO_VIA + TIPO_INTERSEC
+ LUMINOSIDAD + SUPERFICIE_CALZADA + EDAD + POSICION
```

[6] JAMES, G.; WITTEN, D.; HASTIE, T. y TIBSHIRANI, R.: *An Introduction to Statistical Learning*. 6ª edición, Nueva York, Springer Science + Business Media, 2015, pp. 130-137.

RESULTADOS

I. ANÁLISIS DE LAS VARIABLES CON RESPECTO A LA VARIABLE PRINCIPAL

Durante el análisis detallado de las variables se obtuvieron muchas pistas y conclusiones interesantes que ayudarían a crear un modelo con la mejor selección de variables posible. Todas las variables se compararon contra la variable respuesta "ILESO_30D".

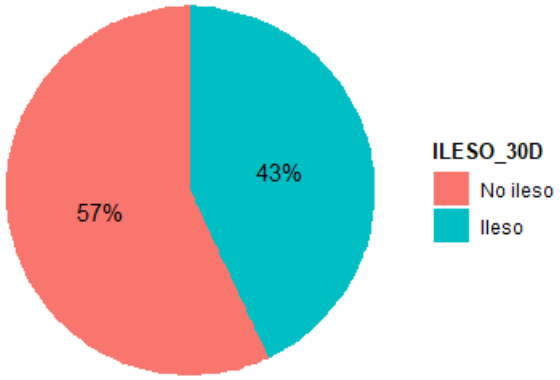

Ilustración 1. Porcentaje de individuos accidentados según la lesividad. La mayoría de los individuos implicados en un accidente de tráfico durante el año 2015 en España no salieron ilesos del mismo.

La variable "HORAS" era una variable categórica que representaba la hora del día en la que ocurrió el accidente en el que el individuo estuvo implicado. La variable presentaba una mayor cantidad de individuos accidentados en las horas de mayor actividad económica (desde las 8-9 hasta las 20-21), con repuntes a las horas de inicio y fin de la jornada laboral. Por la noche, el número de accidentes se reducía considerablemente, sin embargo, al analizar la variable "LUMINOSIDAD", se pudo ver que la mayoría de los accidentes con no ilesos ocurrieron en situaciones de baja luminosidad (de noche).

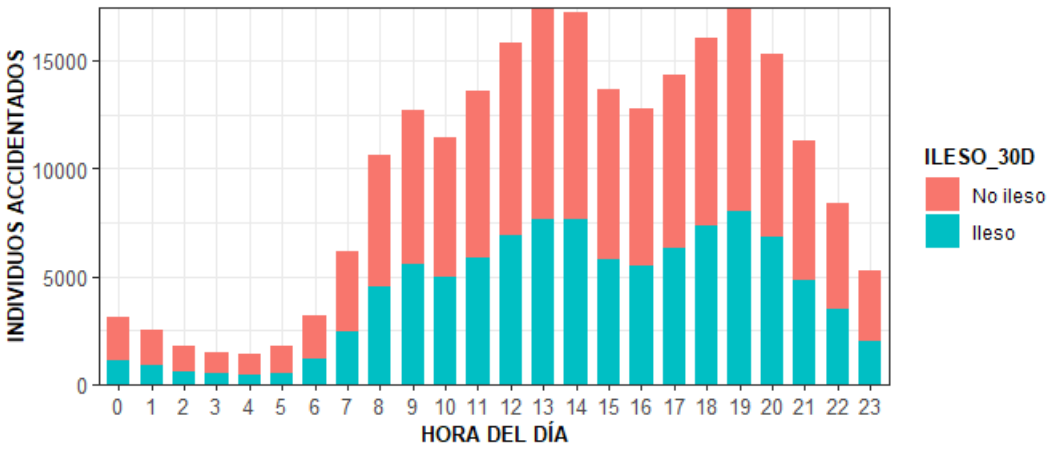

Ilustración 2. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la hora del día. Las horas del día a las que solían haber más accidentados se encontraba entre las 8 y las 21 horas, con un repunte entre las 13-14 y a las 19 horas.

La variable "DIASEMANA" era una variable categórica que representaba el día de la semana en el que ocurrió el accidente en el que el individuo estuvo implicado. Los días laborables (lunes a viernes) tenían una mayor cantidad de accidentados que los fines de semana. El motivo de esto podría ser similar al de la variable "HORA", los accidentados se concentraban en los días con mayores desplazamientos debido al desarrollo de actividades económicas. Existía un repunte de individuos accidentados el viernes, esto podría ser por motivo de que suele ser el día por excelencia en el que se producen desplazamientos por ocio e inicio de viajes.

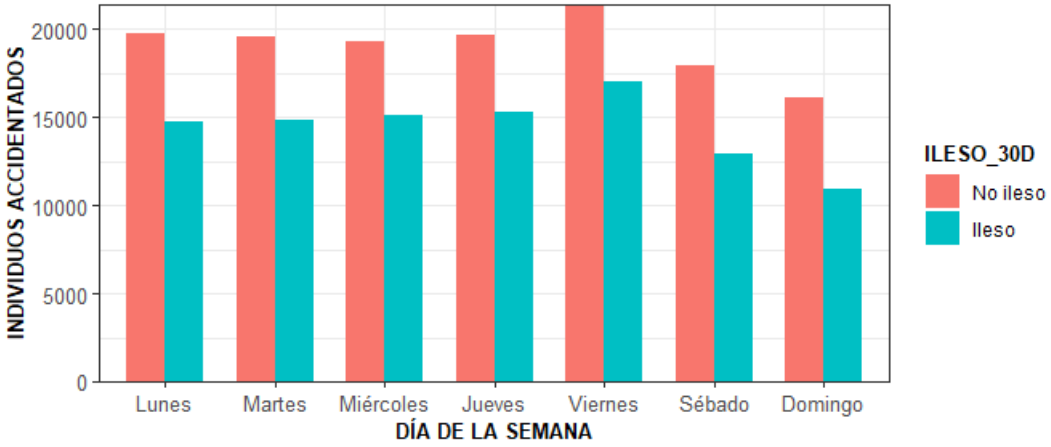

Ilustración 3. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según el día de la semana. Los días de la semana con mayor número de accidentados fueron los viernes. La cantidad de ilesos vs. no ilesos se mantuvo más o menos constante durante todos los días de la semana.

La variable "COMUNIDAD_AUTONOMA" era una variable categórica que representaba la Comunidad Autónoma en la que ocurrió el accidente en el que el individuo estuvo implicado. Las Comunidades Autónoma con una mayor incidencia eran las que más habitantes y mayor actividad económica tenían: Cataluña, seguido de Andalucía y de la Comunidad de Madrid.

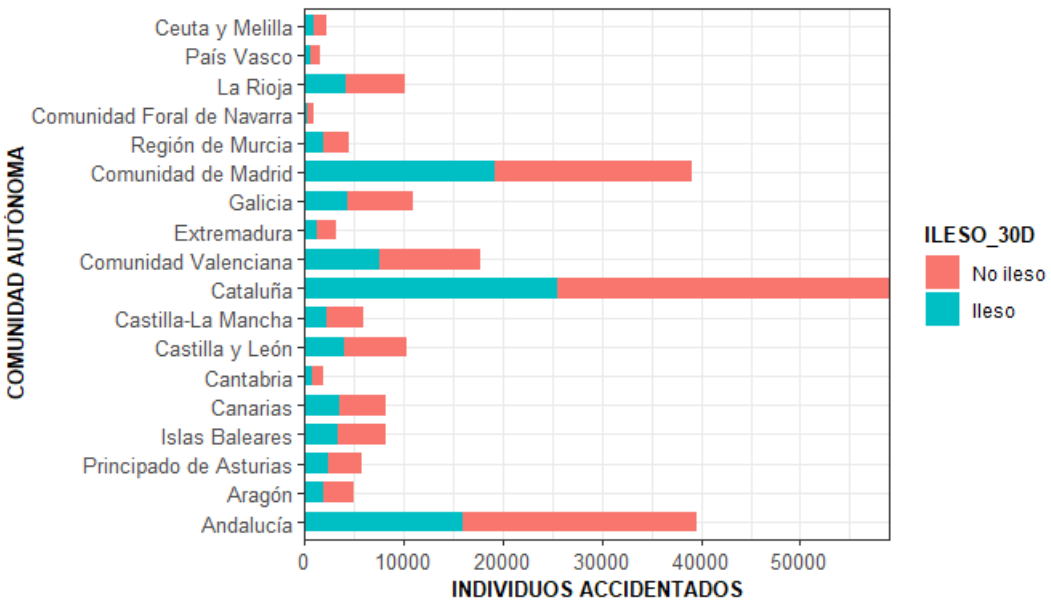

Ilustración 4. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la Comunidad Autónoma. Las Comunidades Autónomas con mayor número de individuos accidentados eran Cataluña, Andalucía y la Comunidad de Madrid.

La variable "TOT_VEHICULOS_IMPLICADOS" era una variable numérica que representaba el número de vehículos implicados en el accidente en el que estuvo el individuo. Los individuos que resultaban ilesos se concentraban en accidentes donde el número total de vehículos implicados solía ser de 2, mientras que los individuos que no resultaban ilesos se concentraban en accidentes donde el número total de vehículos implicados solía ser de 1-2.

Tabla 3. Resumen del total de vehículos implicados en un accidente según la lesividad del individuo accidentado

ILESO_30D	Media	Desviación estándar
No ileso	1.827489	0.7335200
Ileso	2.090236	0.7764459

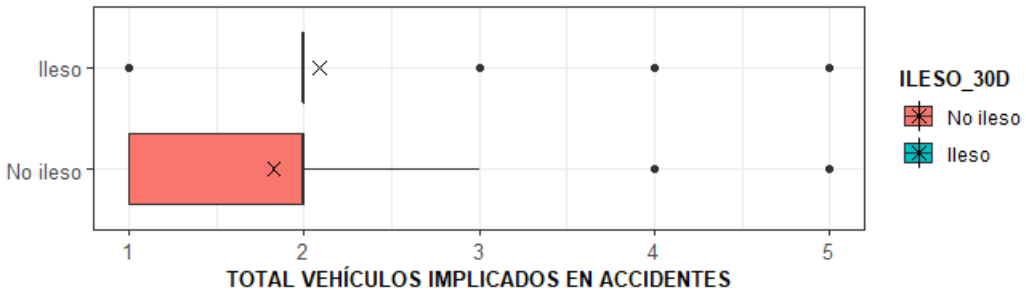

Ilustración 5. Total de vehículos implicados en accidentes de tráfico según la lesividad del individuo accidentado. Las medias entre los dos grupos diferían ligeramente, pero no las medianas, que se situaban en 2 vehículos en ambos grupos.

La variable "TOT_VICTIMAS30D" era una variable numérica que representaba el número total de individuos implicados en el mismo accidente que el individuo (incluyéndose). Los individuos ilesos se concentraban en accidentes en los que sólo ellos están implicados, mientras que los individuos que no resultaban ilesos se concentraban en accidentes con mayor número total de individuos implicados.

Tabla 4. Resumen del total de individuos implicados en un accidente según la lesividad de estos

ILESO_30D	Media	Desviación estándar
No ileso	1.827878	1.1865822
Ileso	1.373566	0.7937671

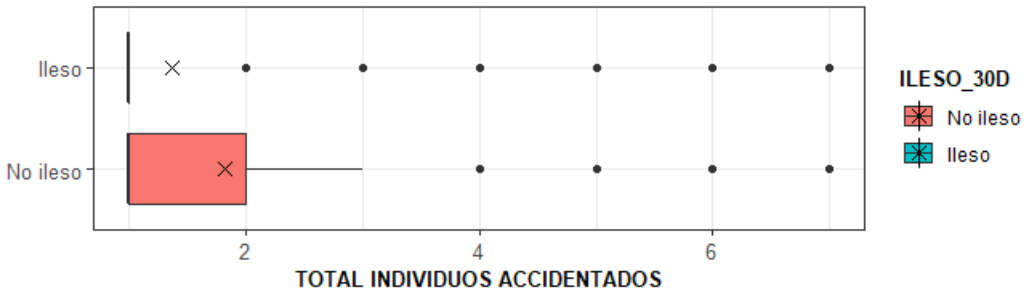

Ilustración 6. Total de individuos implicados en accidentes de tráfico según la lesividad de estos. Las medias entre los dos grupos diferían ligeramente, pero no las medianas, que se situaban en 1 individuo en ambos grupos.

La variable "ZONA" era una variable categórica que representaba la zona en la que ocurrió el accidente en el que el individuo estuvo implicado. Era de suponer que las zonas urbanas concentrarían el mayor número de accidentados, ya que, según la teoría, son las zonas con mayor concentración de vehículos y donde se producen miles de desplazamientos cortos cada día. Además, es más común encontrar peatones en zonas urbanas.

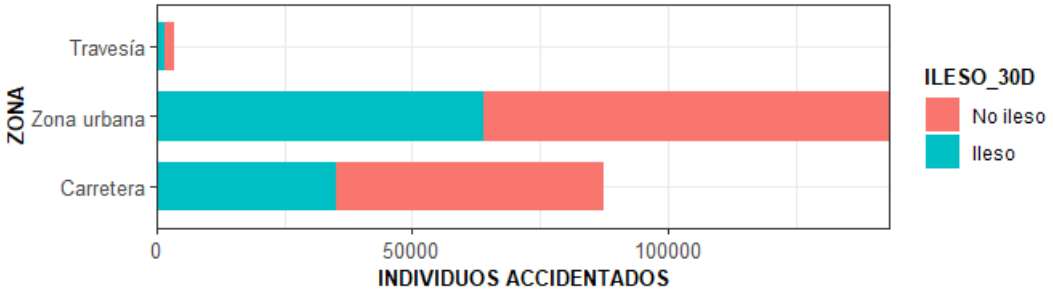

Ilustración 7. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la zona en la que se produjo el accidente. La mayoría de los individuos accidentados se concentraba en zonas urbanas, seguido de carreteras, y por último, con bastante diferencia, las travesías.

La variable "TIPO_VIA" era una variable categórica que representaba el tipo de vía en la que ocurrió el accidente en el que el individuo estuvo implicado. Los ramales de enlace eran, con diferencia, el tipo de vía en la que más accidentados se concentraban, esto podría ser debido a que se trata de uniones de vías (principales con secundarias) en las que los vehículos han de prestar especial atención en el tráfico de su alrededor, señalizar y girar para incorporarse o salir de la vía principal y, en el caso de incorporaciones o salidas de vías de gran velocidad el tiempo de reacción que tiene el conductor se reduce.

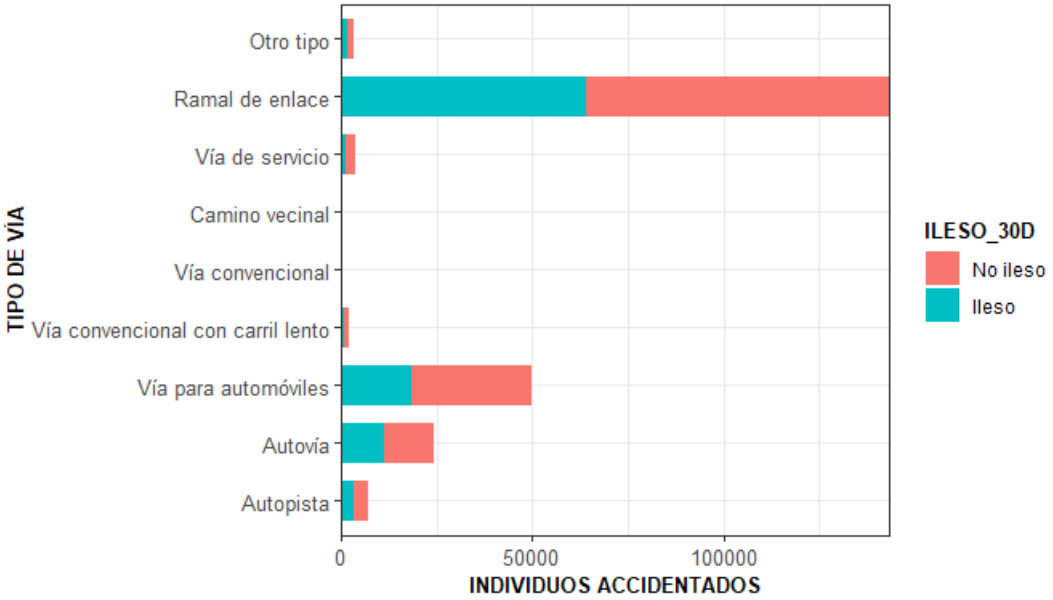

Ilustración 8. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según el tipo de vía en la que se produjo el accidente. La gran mayoría de los individuos accidentados se concentraba en ramales de enlace, seguido de vías para automóviles, autovías y autopistas.

La variable "TIPO_INTERSEC" era una variable categórica que representaba el tipo de intersección en la que ocurrió el accidente en el que el individuo estuvo implicado.

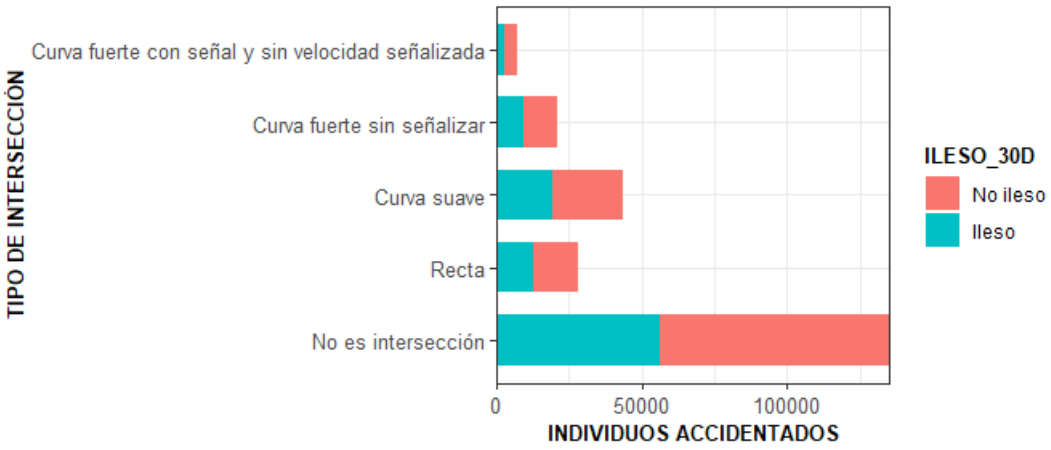

Ilustración 9. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según el tipo de intersección en la que se produjo el accidente. La mayoría de los implicados en accidentes de tráfico no se concentraba en intersecciones.

La variable "LUMINOSIDAD" era una variable categórica que representaba la luminosidad de la vía en el momento en el que ocurrió el accidente en el que el individuo estuvo implicado. La mayoría de los individuos accidentados se registraron a plena luz del día. Sin embargo, en cuanto al porcentaje de ilesos/no ilesos, hubo una mayor proporción de individuos no ilesos en accidentes en condiciones de luminosidad insuficiente que en el resto de las condiciones.

Tabla 5. Tabla cruzada entre las condiciones de luminosidad en las que se produjo el accidente y la lesividad

Luminosidad de la vía	No ileso		Ileso		Total	
Pleno día	94979	(56.2%)	73906	(43.8%)	168885	(100.0%)
Crepúsculo	7329	(58.0%)	5316	(42.0%)	12645	(100.0%)
Noche: iluminación suave	22098	(56.8%)	16830	(43.2%)	38928	(100.0%)
Noche: iluminación insuficiente	9296	(66.0%)	4795	(34.0%)	14091	(100.0%)
Total	133702	(57.0%)	100847	(43.0%)	234549	(100.0%)

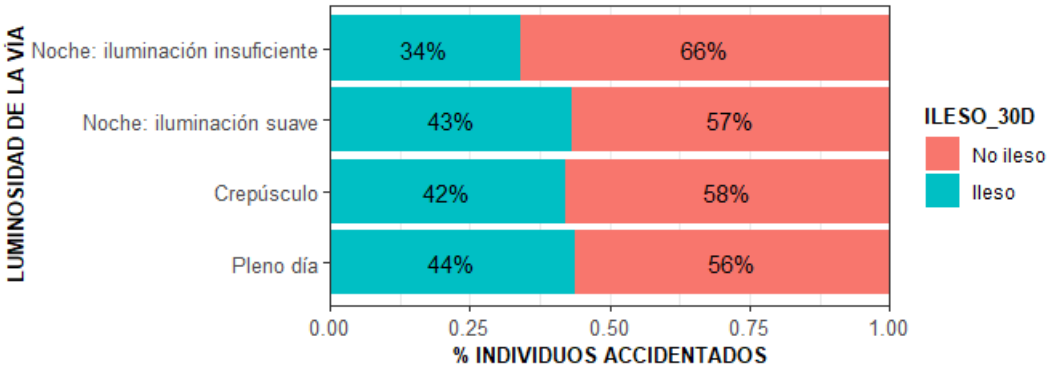

Ilustración 10. Porcentaje de individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según las condiciones de luminosidad de la vía en el momento en el que se produjo el accidente. Todos los porcentajes eran similares a excepción de las condiciones de noche con iluminación insuficiente, donde el porcentaje de individuos ilesos era algo menor.

La variable "SUPERFICIE_CALZADA" era una variable categórica que representaba las condiciones de la calzada en el momento en el que ocurrió el accidente en el que el individuo estuvo implicado. La mayoría de los accidentados se concentraba en condiciones en las que la calzada estaba seca y limpia.

Tabla 6. Tabla cruzada entre la superficie de la calzada en el momento en el que se produjo el accidente y la lesividad

Superficie de la calzada	No ileso		Ileso		Total	
Seca y limpia	115670	(56.1%)	90569	(43.9%)	206239	(100.0%)
Umbría	899	(74.1%)	314	(25.9%)	1213	(100.0%)
Mojada	14113	(62.1%)	8600	(37.9%)	22713	(100.0%)
Helada	227	(72.1%)	88	(27.9%)	315	(100.0%)
Nevada	161	(57.3%)	120	(42.7%)	281	(100.0%)
Barrillo	416	(76.1%)	131	(23.9%)	547	(100.0%)
Gravilla suelta	2216	(68.4%)	1025	(31.6%)	3241	(100.0%)
Total	133702	(57.0%)	100847	(43.0%)	234549	(100.0%)

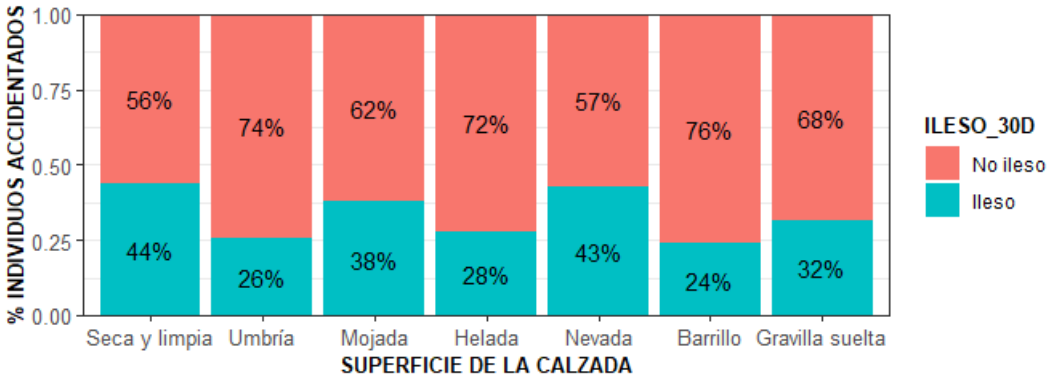

Ilustración 11. Porcentaje de individuos accidentados que resultaron ilesos o no ilesos en accidentes según la superficie de la calzada en el momento en el que se produjo el accidente. Los porcentajes de individuos ilesos eran notablemente menores respecto al resto en las condiciones en las que la calzada estaba umbría, helada o con barrillo.

La variable "FACTORES_ATMOSFERICOS" era una variable categórica que representaba las condiciones atmosféricas de la vía en el momento en el que ocurrió el accidente en el que el individuo estuvo implicado.

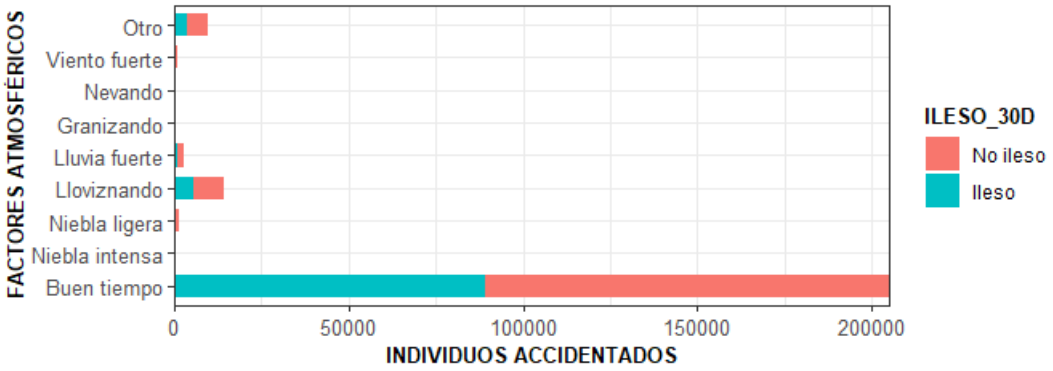

Ilustración 12. Individuos accidentados que resultaron ilesos o no en accidentes de tráfico según las condiciones atmosféricas en las que se produjo el accidente. La gran mayoría de individuos tuvieron un accidente en condiciones de buen tiempo.

La variable "INFRACC_VELOCIDAD" era una variable categórica que representaba la infracción de velocidad cometida (si aplica) por el individuo implicado en el accidente.

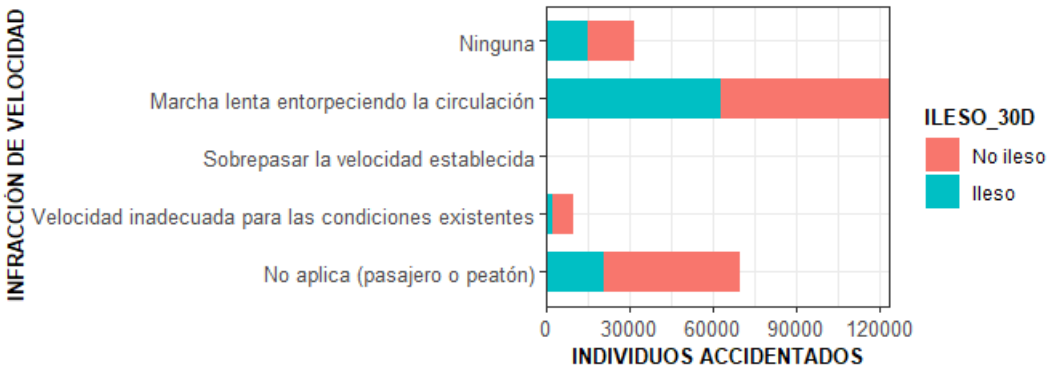

Ilustración 13. Individuos accidentados que resultaron ilesos o no en accidentes de tráfico según la infracción de velocidad cometida (si aplica). La mayoría de los accidentados cometieron una infracción de marcha lenta, entorpeciendo la circulación del tráfico.

La variable "EDAD" era una variable numérica que representaba la edad que tenía el individuo implicado en el accidente. La media de edad para los no ilesos se situaba en los 38.88 años, mientras que para los ilesos se situaba en los 40.78 años.

Tabla 7. Resumen de la edad de los individuos implicados en un accidente según la lesividad

ILESO_30D	Media	Desviación estándar	n
No ileso	38.88754	17.68170	133702
Ileso	40.78905	16.53492	100847

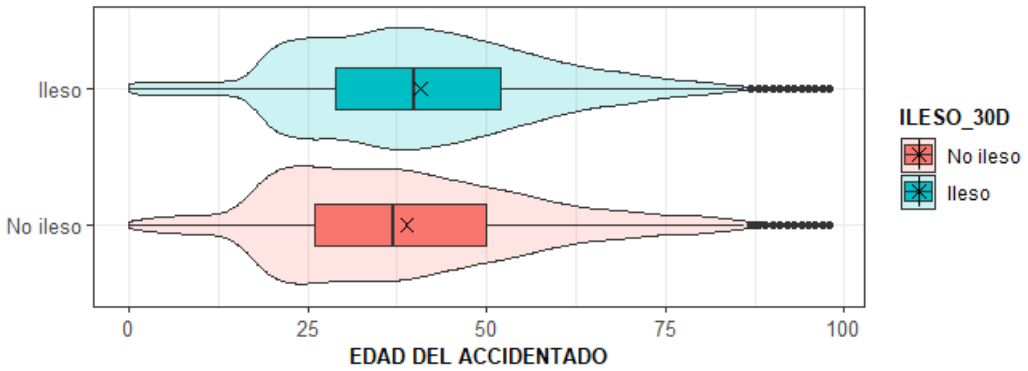

Ilustración 14. Gráfico de violín + gráfico de caja de la edad del individuo implicado en un accidente diferenciando entre ileso y no ileso. Gráficamente, las diferencias de edad entre ilesos y no ilesos parecían ser pequeñas. La mayoría de los individuos accidentados se concentraban en torno a las edades de 20 y 40 años.

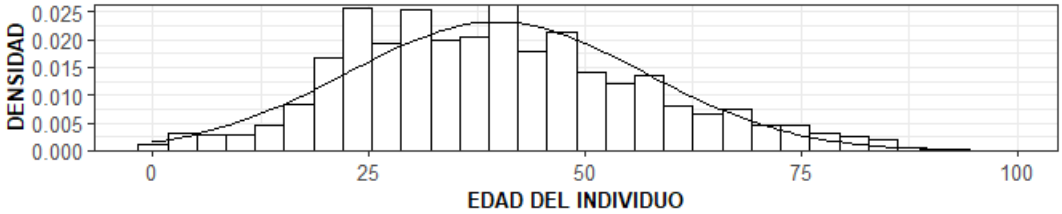

Ilustración 15. Gráfico de densidad de la edad de los individuos implicados en un accidente vs. distribución normal. Gráficamente, la distribución de la edad era parecida a una distribución normal.

La variable "POSICIÓN" era una variable categórica que representaba la posición del individuo en el momento en el que ocurrió el accidente. La mayoría de los individuos implicados en un accidente eran conductores, esto tiene su sentido en que muchos de los accidentes se producen cuando uno viaja solo y, contrastando con lo comentado anteriormente, en trayectos de corta distancia y en zonas urbanas, por ejemplo, ir al trabajo, ocio, etc. Los resultados del análisis de esta variable también arrojaron que el no tener una carrocería que defienda al individuo del exterior suponía un factor de gran importancia en la lesividad. Así, individuos que viajaban en motos o peatones tenían una mucha mayor tasa de lesividad en comparación con el resto de los grupos. Estar sentado en el momento del accidente también era relevante, ya que las tasas de lesividad eran mucho mayores cuando el individuo estaba de pie.

Tabla 8. Tabla cruzada entre la posición del individuo en el momento del accidente según la lesividad

Posición del individuo	No ileso		Ileso		Total	
Peatón	14458	(96.9%)	468	(3.1%)	14926	(100.0%)
Conductor vehículo	47079	(37.9%)	77116	(62.1%)	124195	(100.0%)
Pasajero delantero izquierdo	17892	(63.4%)	10315	(36.6%)	28207	(100.0%)
Pasajero trasero izquierdo	4180	(57.0%)	3153	(43.0%)	7333	(100.0%)
Pasajero trasero derecho	5322	(59.3%)	3648	(40.7%)	8970	(100.0%)
Pasajero trasero central	1507	(57.6%)	1111	(42.4%)	2618	(100.0%)
Conductor vehículo de dos ruedas	37624	(92.0%)	3253	(8.0%)	40877	(100.0%)
Pasajero vehículo de dos ruedas	3345	(92.1%)	288	(7.9%)	3633	(100.0%)
Otros pasajeros sentados	1894	(56.5%)	1456	(43.5%)	3350	(100.0%)
Otros pasajeros de pie	401	(91.1%)	39	(8.9%)	440	(100.0%)
Total	133702	(57.0%)	100847	(43.0%)	234549	(100.0%)

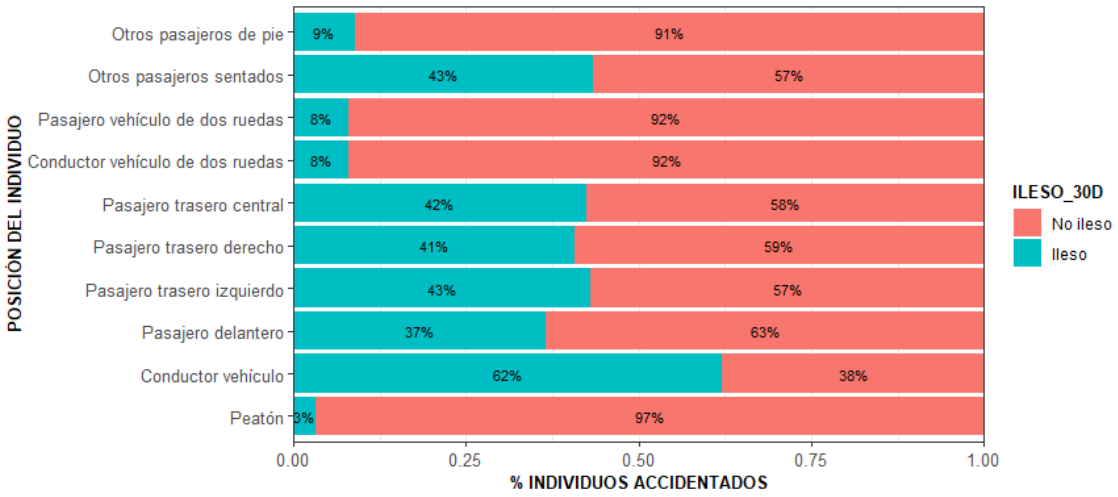

Ilustración 16. Porcentaje de individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la posición en la que se encontraban cuando se produjo el accidente. Los pasajeros que iban de pie, los peatones y los individuos que viajaban en un vehículo de dos ruedas tenían una tasa de lesividad mucho mayor que los individuos que iban en otras posiciones.

Al analizar más de dos variables de manera conjunta, se obtuvieron indicios de relaciones más complejas. Los conductores se concentraban en rangos de edad más cortos que el resto de las posiciones, entre los 25 y 50 años para los no ilesos y ligeramente más alto para los ilesos. La variabilidad en las edades de los peatones que resultaban no ilesos

era notablemente mayor comparada con la de los ilesos, ampliándose hacia mayores edades, esto podría ser debido a que personas de mayor edad tenían unas condiciones de salud más frágiles y no solían conducir tanto como las personas más jóvenes.

Tabla 9. Estadísticas sobre las edades de los individuos accidentados agrupados por lesividad y posición

ILESO_30D	ID_PERSONA	Media	Desviación estándar
No ileso	Conductor	39.36336	14.58328
No ileso	Pasajero	34.82521	19.74625
No ileso	Peatón	45.8370	24.97289
Ileso	Conductor	42.82465	14.79524
Ileso	Pasajero	32.69153	20.16595
Ileso	Peatón	38.06897	19.99426

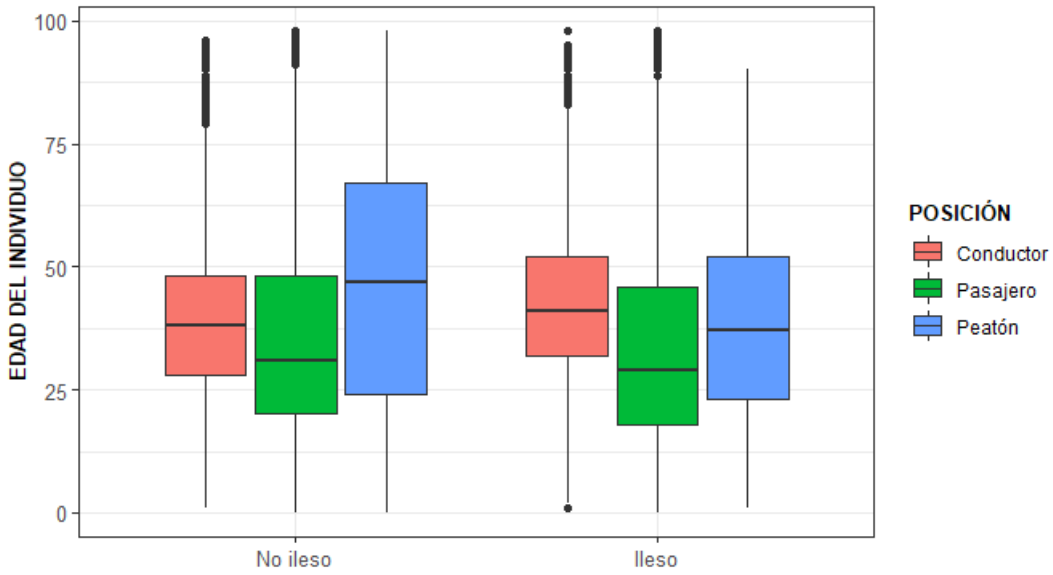

Ilustración 17. Edad de los individuos accidentados agrupada por posición y diferenciando entre ileso y no ileso. Individuos más jóvenes ocupaban posiciones como pasajeros. La variabilidad de edad en los peatones era muy alta en los no ilesos, mientras que la de los conductores era más pequeña. Los peatones no ilesos tenían una mayor variabilidad de edad que los ilesos. Las medianas de edad entre conductores y pasajeros ilesos y no ilesos eran parecidas.

II. ANÁLISIS DEL MODELO DE REGRESIÓN LOGÍSTICA

La relación que se deseaba estudiar era entre la lesividad o no lesividad de implicados en accidentes de tráfico en función de las condiciones que se dieron en el mismo. Se determinó que el mejor modelo que se podía realizar con los datos disponibles se trataba de un modelo de regresión logística con 11 variables predictoras.

Los resultados del modelo de regresión logística fueron, en general, buenos. En cuanto a significación individual de los coeficientes se puede decir que la gran mayoría de coeficiente resultaron ser individualmente significativos a un nivel de significación del 5%.

Para poder obtener los resultados del modelo de regresión logística de una forma más interpretable, se procedió a hacer la exponencial de cada coeficiente, obteniendo así el odds ratio (ratio o razón de probabilidades), que es el cociente entre el odds en el grupo con el factor y el odds en el grupo sin el factor.

Resultados de los predictores continuos

Variable "TOT_VICTIMAS30D":

- Por cada individuo adicional implicado en el accidente, el odds de resultar ileso fue 0.43 veces menor, es decir, cuantos más individuos implicados en un accidente, menor es la posibilidad de resultar ileso del mismo.

Variable "TOT_VEHICULOS_IMPLICADOS":

- Por cada vehículo adicional implicado en el accidente, el odds de resultar ileso fue 1.53 veces mayor, es decir, cuantos más vehículos implicados en un accidente, mayor es la posibilidad de resultar ileso del mismo.

Variable "EDAD":

- El odds ratio asociado a la variable "EDAD" es aproximadamente 1, por lo que el efecto de este predictor se determinó como nulo. La edad del individuo no tiene influencia directa sobre la lesividad del este. No hay diferencia entre las posibilidades de resultar ileso según la edad.

Resultados de los predictores categóricos

A la hora de interpretar los resultados de los predictores categóricos, es importante comprender el concepto de nivel o categoría de referencia, que es el primer nivel en el orden de la variable categórica sobre la que se calcula la diferencia entre los logaritmos de los odds, de este modo se puede evaluar la ventaja de las posibilidades del nivel 1 (del que se ha obtenido el coeficiente) respecto al nivel de referencia 0.

Variable "HORA". Categoría de referencia: 0 horas.

- Para los individuos que tuvieron un accidente en entre las 1 y las 7 horas, los odds de resultar ileso fueron entre 0.64 y 0.91 veces menor que para los que tuvieron un accidente a las 0 horas. En general, entre las 1 y las 7 horas, se obtuvieron coeficientes más bajos que indican que hay una menor probabilidad de salir ileso de un accidente durante esas horas que a las 0 horas.
- Para los individuos que tuvieron un accidente entre las 8 y las 23 horas, los odds de resultar ileso fueron entre 1.02 y 1.47 veces mayor que para los que tuvieron un accidente a las 0 horas. En general, entre las 8 y as 23 horas, se obtuvieron coeficientes más altos que indican que hay una mayor probabilidad de salir ileso de un accidente durante esas horas que las 0 horas.

Variable "DIASEMANA". Categoría de referencia: lunes.

- Para los individuos que tuvieron un accidente un martes, el odds de resultar ileso fue 1.02 veces mayor que para los que tuvieron un accidente un lunes.
- Para los individuos que tuvieron un accidente un miércoles, el odds de resultar ileso fue 1.03 veces mayor que para los que tuvieron un accidente un lunes.
- Para los individuos que tuvieron un accidente un jueves, el odds de resultar ileso fue 1.04 veces mayor que para los que tuvieron un accidente un lunes.

- Para los individuos que tuvieron un accidente un viernes, el odds de resultar ileso fue 1.08 veces mayor que para los que tuvieron un accidente un lunes.
- Para los individuos que tuvieron un accidente un sábado, el odds de resultar ileso fue 1.20 veces mayor que para los que tuvieron un accidente un lunes.
- Para los individuos que tuvieron un accidente un viernes, el odds de resultar ileso fue 1.23 veces mayor que para los que tuvieron un accidente un lunes.

Variable "COMUNIDAD_AUTONOMA". Categoría de referencia: Andalucía.

- Para individuos que tuvieron un accidente en las Comunidades Autónomas de Aragón, Canarias, Castilla y León, Castilla-La Mancha, Extremadura, Galicia, Comunidad Foral de Navarra, País Vasco o Ceuta y Melilla, el odds de resultar ileso fue entre 0.83 y 0.98 veces menor que para los que tuvieron un accidente en Andalucía.
- Para individuos que tuvieron un accidente en la comunidad de Islas Baleares, el odds de resultar ileso es aproximadamente 1, es decir, no hay diferencia entre las posibilidades de resultar ileso entre Islas Baleares y Andalucía.
- Para individuos que tuvieron un accidente en las Comunidades Autónomas de Principado de Asturias, Cantabria, Cataluña, Comunidad Valenciana, Comunidad de Madrid, Región de Murcia o La Rioja, el odds de resultar ileso fue entre 1.07 y 1.18 veces mayor que para los que tuvieron un accidente en Andalucía.

Variable "TIPO_VIA". Categoría de referencia: autopista.

- Para los individuos que tuvieron un accidente en una autovía, el odds de resultar ileso fue 1.04 veces mayor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en una vía para automóviles, el odds de resultar ileso fue 0.91 veces menor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en una vía convencional con carril lento, el odds de resultar ileso fue 0.89 veces menor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en una vía convencional, el odds de resultar ileso fue 0.76 veces menor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en un camino vecinal, el odds de resultar ileso fue 1.06 veces mayor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en una vía de servicio, el odds de resultar ileso fue 0.90 veces menor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en un ramal de enlace, el odds de resultar ileso fue 1.52 veces mayor que para los que tuvieron un accidente en la autopista.
- Para los individuos que tuvieron un accidente en otro tipo de vías, el odds de resultar ileso fue 1.16 veces mayor que para los que tuvieron un accidente en la autopista.

Variable "TIPO_INTERSEC". Categoría de referencia: no es intersección.

- Para los individuos que tuvieron un accidente en una recta, el odds de resultar ileso fue 1.25 veces mayor que para los que no tuvieron un accidente en una intersección.
- Para los individuos que tuvieron un accidente en una curva suave, el odds de resultar ileso fue 1.08 veces mayor que para los que no tuvieron un accidente en una intersección.
- Para los individuos que tuvieron un accidente en una curva fuerte sin señalizar, el odds de resultar ileso fue 1.16 veces mayor que para los que no tuvieron un accidente en una intersección.
- Para los individuos que tuvieron un accidente en una curva fuerte con señal y sin velocidad señalizada, el odds de resultar ileso fue 1.12 veces mayor que para los que no tuvieron un accidente en una intersección.

Variable "LUMINOSIDAD". Categoría de referencia: pleno día.

- Para los individuos que tuvieron un accidente durante el crepúsculo, el odds de resultar ileso fue 0.94 veces menor que para los que tuvieron un accidente con una iluminación de pleno día.
- Para los individuos que tuvieron un accidente durante una noche con iluminación suave, el odds de resultar ileso fue 1.09 veces mayor que para los que tuvieron un accidente con una iluminación de pleno día.
- Para los individuos que tuvieron un accidente durante una noche con iluminación insuficiente, el odds de resultar ileso fue 0.84 veces menor que para los que tuvieron un accidente con una iluminación de pleno día.

Variable "SUPERFICIE_CALZADA". Categoría de referencia: seca y limpia.

- Para los individuos que tuvieron un accidente en unas condiciones de calzada umbría, el odds de resultar ileso fue 0.88 veces menor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.
- Para los individuos que tuvieron un accidente en unas condiciones de calzada mojada, el odds de resultar ileso fue 0.75 veces menor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.
- Para los individuos que tuvieron un accidente en unas condiciones de calzada helada, el odds de resultar ileso fue 0.58 veces menor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.
- Para los individuos que tuvieron un accidente en unas condiciones de calzada nevada, el odds de resultar ileso fue 0.81 veces menor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.
- Para los individuos que tuvieron un accidente en unas condiciones de calzada con barrillo, el odds de resultar ileso fue 1.22 veces mayor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.
- Para los individuos que tuvieron un accidente en unas condiciones de calzada con gravilla suelta, el odds de resultar ileso fue 0.83 veces menor que para los que tuvieron un accidente con unas condiciones de calzada seca y limpia.

Variable "POSICION". Categoría de referencia: peatón.

- Para los individuos que tuvieron un accidente ocupando la posición de conductor de un vehículo (de más de dos ruedas), el odds de resultar ileso fue 58.44 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de pasajero delantero de un vehículo (de más de dos ruedas), el odds de resultar ileso fue 28.94 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de pasajero trasero izquierdo de un vehículo (de más de dos ruedas), el odds de resultar ileso fue 58.84 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de pasajero trasero derecho de un vehículo (de más de dos ruedas), el odds de resultar ileso fue 51.12 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de pasajero central de un vehículo (de más de dos ruedas), el odds de resultar ileso fue 58.74 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de conductor de un vehículo de dos ruedas, el odds de resultar ileso fue 1.98 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de pasajero de un vehículo de dos ruedas, el odds de resultar ileso fue 3.34 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de otros pasajeros sentados, el odds de resultar ileso fue 43.53 veces mayor que para los que tuvieron un accidente como peatones.
- Para los individuos que tuvieron un accidente ocupando la posición de otros pasajeros de pie, el odds de resultar ileso fue 3.56 veces mayor que para los que tuvieron un accidente como peatones.

Importancia relativa de los predictores

El análisis de la importancia relativa de las variables predictoras del modelo mostró que la variable más importante del modelo fue "TOT_VICTIMAS30D" seguido por algunas categorías de "POSICION" y "TOT_VEHICULOS_IMPLICADOS", seguido (con algo más de diferencia) del resto de variables.

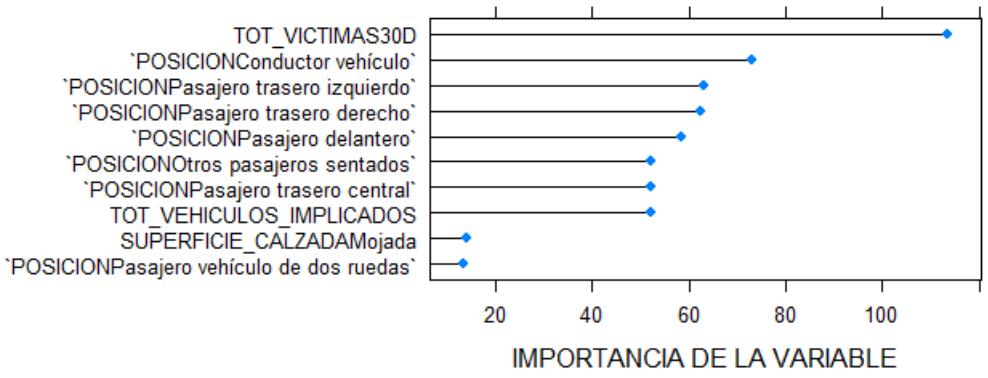

Ilustración 18. Importancia relativa de los predictores del modelo. Los predictores más importantes son "TOT_VICTIMAS30D", "POSICION" y "TOT_VEHICULOS_IMPLLICADOS".

Diagnóstico del modelo

El pseudo R-Cuadrado (R-Cuadrado de Tjur) calculado a partir del modelo final fue del 34.11%.

El test de inflación de la varianza (VIF) indica que no hubo fuertes indicios de multicolinealidad en el modelo, ya que todos los coeficientes fueron menores a 5.

Tabla 10. Factor de inflación de la varianza (VIF) de los predictores del modelo final

	GVIF	Df	GVIF^(1/(2*Df))
HORA	3.326.316	23	1.026.472
DIASEMANA	1.079.223	6	1.006.374
COMUNIDAD_AUTONOMA	2.079.613	17	1.021.768
TOT_VEHICULOS_IMPLICADOS	1.205.504	1	1.097.954
TOT_VICTIMAS30D	1.225.008	1	1.106.801
TIPO_VIA	2.281.416	8	1.052.902
TIPO_INTERSEC	1.448.454	4	1.047.401
LUMINOSIDAD	3.415.945	3	1.227.209
SUPERFICIE_CALZADA	1.353.051	6	1.025.517
EDAD	1.126.201	1	1.061.226
POSICION	1.330.946	9	1.016.010

El test del ajuste global del modelo, la prueba de Hosmer-Lemeshow indicó que, para un nivel de significación del 5%, existía evidencia en contra de la hipótesis nula de que el modelo se ajustaba adecuadamente a los datos. Es decir, el modelo no parecía encajar bien, ya que no había una diferencia significativa entre el modelo y los datos observados.

Tabla 11. Test de Hosmer-Lemeshow del ajuste global del modelo

Hosmer and Lemeshow goodness of fit (GOF) test
data: fit1$y, fitted(fit1)
X-squared = 1275.3, df = 8, p-value < 2.2e-16

Capacidad predictiva del modelo

En cuanto a capacidad predictiva del modelo de regresión logística se refiere, se usaron los datos de prueba para testearla (25% del conjunto de datos original). El modelo obtuvo una precisión del 76.78%, una sensibilidad del 74.55%, una especificidad del 79.74% y un valor Kappa de 0.53. La detección del modelo fue del 42.49% y clasificó correctamente al 82.98% de los no ilesos y al 70.26% de los ilesos.

Tabla 12. Matriz de confusión del modelo final

Confusion Matrix and Statistics			
	Reference		
Prediction	No ileso		Ileso
No ileso	24917		5109
Ileso	8508		20102
		Accuracy:	0.7678
		95% CI:	(0.7643, 0.7712)
		No Information Rate:	0.57

P-Value [Acc > NIR]:	<2.2e-16
Kappa:	0.534
Mcnemar's Test P-Value:	<2.2e-16
Sensitivity:	0.7455
Specificity:	0.7974
Pos Pred Value:	0.8298
Neg Pred Value:	0.7026
Prevalence:	0.5700
Detection Rate:	0.4249
Detection Prevalence:	0.5121
Balanced Accuracy:	0.7714
Positive' Class:	No ileso

El AUC (Area Under the Curve) de la curva ROC (Receiver Operating Characteristic) fue de 0.77. Como este valor se encontraba entre 0.6 y 0.8, por lo que hay indicios de que el modelo discriminó de forma adecuada.

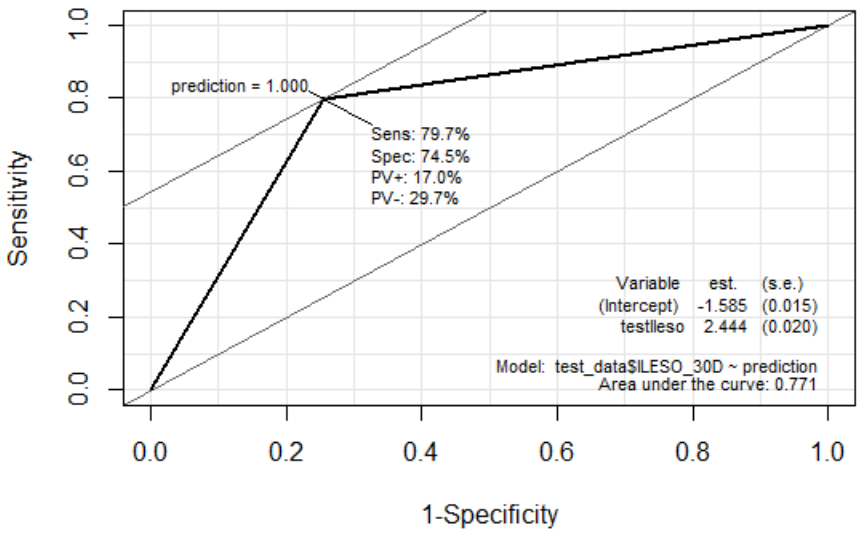

Ilustración 19. Curva ROC del modelo final con la predicción. El AUC de la curva ROC fue de 0.771.

DISCUSIÓN

En base a los resultados del modelo de regresión logística se podría decir que se encontró evidencia sobre la relación existente entre un accidente y las condiciones que se dieron en el mismo.

Los predictores que tomaron especial relevancia para explicar dicha relación a través del modelo fueron los relativos al número de individuos implicados en el accidente, la posición que ocupada el individuo y el total de vehículos implicados.

Se halló evidencia para concluir que, a mayor número de individuos implicados en el accidente, menor sería la probabilidad de resultar ileso del mismo. Lo contrario sucedía con el número de vehículos implicados, ya que a mayor número de vehículos implicados, la probabilidad de resultar ileso del accidente aumentaba.

En cuanto a la posición del individuo, se obtuvieron hallazgos interesantes. Individuos que tuvieron el accidente dentro de un vehículo de más de dos ruedas y estaban sentados obtuvieron una mucha mayor probabilidad de resultar ilesos que individuos que viajaban de pie, en vehículos de dos ruedas o como peatones. Los pasajeros traseros de un vehículo de más de dos ruedas obtuvieron mejores resultados que el conductor y el pasajero delantero, esto podría ser debido a que en la mayoría de las colisiones que se producen en un accidente, la parte delantera del vehículo es la que absorbe la mayoría del impacto. De este modo, se concluyó que tener un chasis que proteja al individuo de las condiciones exteriores es un factor de muy relevante importancia a la hora de salir ileso de un accidente. Los peatones fueron el grupo que salió peor parado, ya que tuvieron la menor probabilidad de resultar ilesos de entre todos los grupos (no hay nada que les proteja o se interponga entre ellos y el accidente).

Otros factores que explicaban la lesividad en un accidente de tráfico (pero que no eran tan importantes como los anteriormente nombrados) fueron la hora, el día, la Comunidad Autónoma, el tipo de vía, el tipo de intersección, la luminosidad y la superficie de la calzada.

Se halló que había mayor probabilidad de resultar ileso en un accidente que: tomase lugar entre las 8 y las 23 horas y/o en un fin de semana. Esto podría deberse a que entre las 0 y las 7 horas existe un mayor riesgo de accidente por la menor luminosidad de la vía (insuficiente en muchos casos); y que los fines de semana es cuando menor número de desplazamientos se producen en carretera, ya que durante estos días, la actividad económica está más relajada.

Las Comunidades Autónomas como Cataluña o Madrid, a pesar de concentrar los mayores números de accidentados resultaban tener más probabilidades de resultar ilesos en un accidente que otras comunidades con mucho menor número de accidentados.

Las vías para automóviles, vías de servicio y vías convencionales obtuvieron una menor probabilidad de resultar ileso que el resto. Accidentes en intersecciones obtuvieron una mayor probabilidad de resultar ileso.

En cuanto a la luminosidad se encontró evidencia de que tener un accidente con unas condiciones de luminosidad de crepúsculo tenía asociada una menor probabilidad de resultar ileso, esto puede ser debido a que los deslumbramientos producidos por el crepúsculo pueden provocar una distracción al volante. Conviene recordar que uno de los factores principales de causa de accidente eran las distracciones al volante derivadas del entorno.

En condiciones de luminosidad insuficiente, la probabilidad de resultar ileso fue menor que a pleno día. Esto podría ser debido a que el conductor del vehículo no tiene una visión del entorno todo lo completa que debería, y podría chocar más fácilmente contra objetos que no tienen una fuente de luz propia. En el caso de los peatones, éstos deberían de estar equipados con una señal luminosa en caso de estar en una vía con iluminación insuficiente para evitar ser atropellados.

Las calzadas cuyas superficies tienen agua, están heladas o nevadas suponen un mayor riesgo para el individuo (menor probabilidad de resultar ileso), probablemente porque en este tipo de superficies los vehículos son más susceptibles a perder el control.

El diagnóstico del modelo no arrojó resultados demasiado prometedores. A pesar de que no hubo indicios de multicolinealidad y la medida de bondad del ajuste (R-Cuadrado de Tjur) tuvo un nivel no demasiado bajo para el tipo de datos trabajados, el test del ajuste global del modelo (Hosmer-Lemeshow) indicó que el modelo no se ajustaba adecuadamente a los datos.

El modelo fue capaz de clasificar correctamente al 82.98% de los individuos que resultaron no ilesos, y al 70.26% de los individuos que resultaron ilesos. El resultado del AUC de la curva ROC se encontraba entre los valores que indicaban que el modelo discriminaba de forma adecuada, pero que podría ser mejorable.

Finalmente, se podría concluir que la relación inicial que se buscaba entre la lesividad de un individuo según las condiciones del accidente se halló, pero hubo indicios de que había factores que escapaban a los datos disponibles como, por ejemplo, si el individuo había estado distraído al volante por el uso del teléfono móvil, el uso de drogas, el tipo de vehículo, etc. La inclusión de dichas variables podría ayudar a mejorar notablemente la capacidad predictiva del modelo.

BIBLIOGRAFÍA

DIRECCIÓN GENERAL DE TRÁFICO: *Anuario estadístico de accidentes 2015* [en línea], Ministerio del interior, 2015 [ref. 19 de agosto 2021]. Disponible en Web: https://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/anuario-estadistico-de-accidentes/anuario-accidentes-2015.pdf

DIRECCIÓN GENERAL DE TRÁFICO: *Las distracciones son la causa de uno de cada cuatro accidentes* [en línea], Notas de prensa DGT, 16 de septiembre 2019 [ref. 18 de agosto 2021]. Disponible en Web: https://www.dgt.es/es/prensa/notas-de-prensa/2019/Las-distracciones-son-la-causa-de-uno-de-cada-cuatro-accidentes.shtml.

INSTITUTO NACIONAL DE ESTADÍSTICA: "Métodos de inferencia estadística con datos faltantes", ESTADÍSTICA ESPAÑOLA, *Estudio de simulación sobre los efectos en las estimaciones,* 2006*.*

JAMES, G.; WITTEN, D.; HASTIE, T. y TIBSHIRANI, R.: *An Introduction to Statistical Learning.* 6ª edición, Nueva York, Springer Science + Business Media, 2015, pp. 130-137.

SCHAFER, J.: "Multiple imputation: a primer", SAGE Publications Ltd, *Statistical Methods in Medical Research*, Volume 8 Issue 1, febrero 1999. Disponible en Web: https://doi.org/10.1177%2F096228029900800102.

SMITH, KSCHAFER, J.: "Passenger seating position and the risk of passenger death or injury in traffic crashes", PubMed, *Accident Analysis & Prevention*, Volume 36 Issue 2, marzo 2004. Disponible en Web: https://doi.org/10.1016/S0001-4575(03)00002-2.

ANEXO A

Ilustraciones y tablas adicionales obtenidas durante la investigación que no se han mostrado:

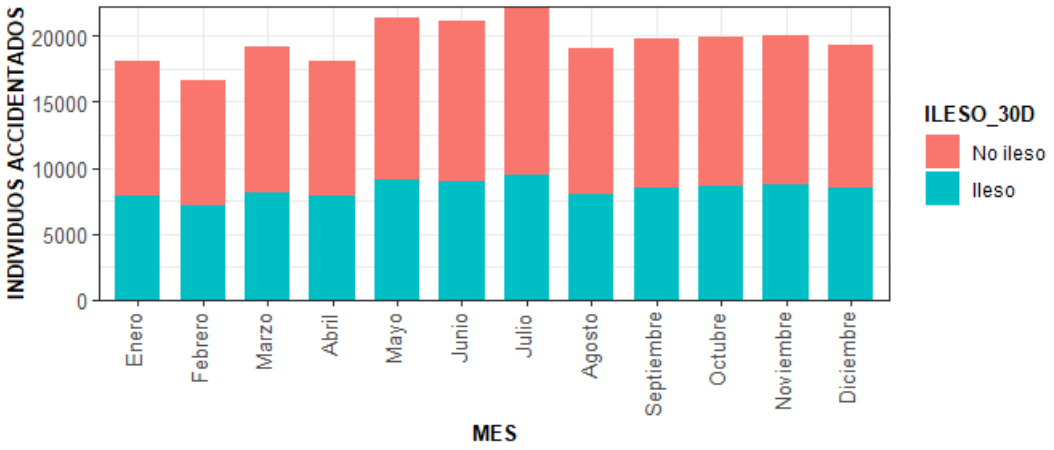

Ilustración 20. Individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la hora del día

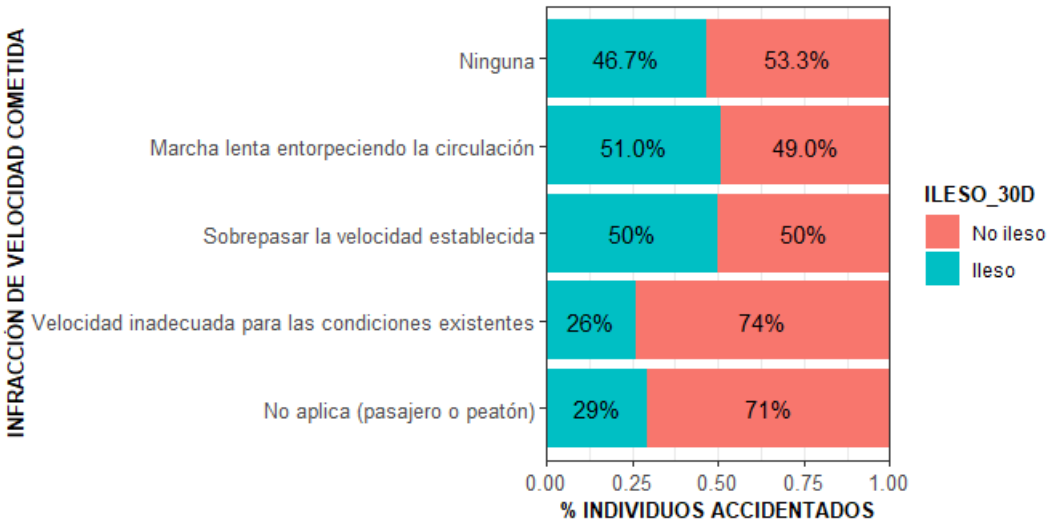

Ilustración 21. Porcentaje de individuos accidentados que resultaron ilesos o no ilesos en accidentes de tráfico según la infracción de velocidad cometida

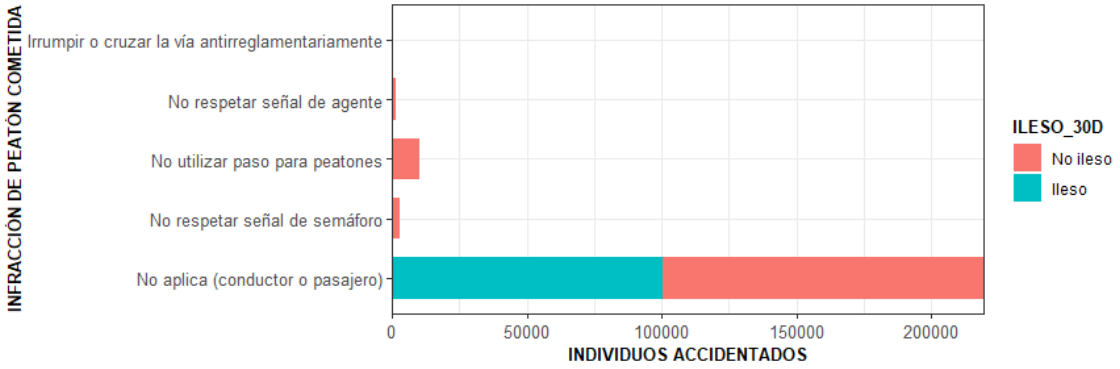

Ilustración 22. Individuos accidentados que resultaron ilesos o no en accidentes de tráfico según la infracción de peatón cometida

Tabla 13. Resumen del modelo final de regresión logística

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.8196330	0.0902153	-42.339	<2e-16	***
HORA1	-0.1028829	0.0776071	-1.326	0.184943	
HORA2	-0.2694636	0.0863128	-3.122	0.001797	**
HORA3	-0.2562950	0.0905449	-2.831	0.004646	**
HORA4	-0.4175987	0.0914629	-4.566	4.98e-06	***
HORA5	-0.4413000	0.0864485	-5.105	3.31e-07	***
HORA6	-0.2335518	0.0719935	-3.244	0.001178	**
HORA7	-0.0940702	0.0649508	-1.448	0.147524	
HORA8	0.1068812	0.0628036	1.702	0.088787	.
HORA9	0.1296399	0.0621916	2.085	0.037112	*
HORA10	0.1722319	0.0628078	2.742	0.006103	**
HORA11	0.1776259	0.0619033	2.869	0.004112	**
HORA12	0.2405603	0.0612520	3.927	8.59e-05	***
HORA13	0.1863551	0.0607691	3.067	0.002165	**
HORA14	0.1339162	0.0607788	2.203	0.027571	*
HORA15	0.0934419	0.0617518	1.513	0.130233	
HORA16	0.1192825	0.0620887	1.921	0.054711	.
HORA17	0.2601329	0.0612788	4.245	2.19e-05	***
HORA18	0.3472142	0.0593524	5.850	4.91e-09	***
HORA19	0.3849082	0.0580791	6.627	3.42e-11	***
HORA20	0.3558533	0.0581653	6.118	9.48e-10	***
HORA21	0.2767323	0.0590437	4.687	2.77e-06	***
HORA22	0.1921295	0.0607462	3.163	0.001562	**
HORA23	0.0225939	0.0654231	0.345	0.729830	
DIASEMANAMartes	0.0217086	0.0217934	0.996	0.319195	
DIASEMANAMiércoles	0.0276407	0.0217578	1.270	0.203949	
DIASEMANAJueves	0.0426333	0.0216686	1.968	0.049124	*
DIASEMANAViernes	0.0790349	0.0212284	3.723	0.000197	***
DIASEMANASábado	0.1787315	0.0226071	7.906	2.66e-15	***
DIASEMANADomingo	0.2098376	0.0236342	8.879	<2e-16	***
COMUNIDAD_AUTONOMAAragón	-0.0854080	0.0430315	-1.985	0.047169	*
COMUNIDAD_AUTONOMAPrincipado de Asturias	0.0736770	0.0402528	1.830	0.067196	.
COMUNIDAD_AUTONOMAIslas Baleares	-0.0023911	0.0353093	-0.068	0.946009	
COMUNIDAD_AUTONOMACanarias	-0.0829075	0.0340325	-2.436	0.014845	*
COMUNIDAD_AUTONOMACantabria	0.1642016	0.0660040	2.488	0.012855	*
COMUNIDAD_AUTONOMACastilla y León	-0.0231737	0.0316367	-0.732	0.463866	
COMUNIDAD_AUTONOMACastilla-La Mancha	-0.1840655	0.0392515	-4.689	2.74e-06	***
COMUNIDAD_AUTONOMACataluña	0.1185862	0.0190733	6.217	5.06e-10	***
COMUNIDAD_AUTONOMAComunidad Valenciana	0.0864658	0.0261316	3.309	0.000937	***
COMUNIDAD_AUTONOMAExtremadura	-0.1912357	0.0506835	-3.773	0.000161	***
COMUNIDAD_AUTONOMAGalicia	-0.0494324	0.0306198	-1.614	0.106442	
COMUNIDAD_AUTONOMAComunidad de Madrid	0.0934905	0.0208820	4.477	7.57e-06	***
COMUNIDAD_AUTONOMARegión de Murcia	0.1244886	0.0459333	2.710	0.006724	**
COMUNIDAD_AUTONOMAComunidad Foral de Navarra	-0.0274124	0.0913990	-0.300	0.764238	
COMUNIDAD_AUTONOMALa Rioja	0.0673352	0.0390255	1.725	0.084452	.
COMUNIDAD_AUTONOMAPaís Vasco	-0.1285467	0.0731968	-1.756	0.079058	.

COMUNIDAD_AUTONOMACeuta y Melilla	-0.1088266	0.0631419	-1.724	0.084794	.
TOT_VEHICULOS_IMPLICADOS	0.4526300	0.0086949	52.057	<2e-16	***
TOT_VICTIMAS30D	-0.8373998	0.0073750	-113.546	<2e-16	***
TIPO_VIAAutovía	0.0406450	0.0363083	1.119	0.262952	
TIPO_VIAVía para automóviles	-0.0948445	0.0346015	-2.741	0.006124	**
TIPO_VIAVía convencional con carril lento	-0.1157085	0.0722802	-1.601	0.109414	
TIPO_VIAVía convencional	-0.2728809	0.1618135	-1.686	0.091720	.
TIPO_VIACamino vecinal	0.0586354	0.1159520	0.506	0.613077	
TIPO_VIAVía de servicio	-0.1034767	0.0623909	-1.659	0.097212	.
TIPO_VIARamal de enlace	0.4215412	0.0339167	12.429	<2e-16	***
TIPO_VIAOtro tipo	0.1514619	0.0575224	2.633	0.008461	**
TIPO_INTERSECRecta	0.2209787	0.0192008	11.509	<2e-16	***
TIPO_INTERSECCurva suave	0.0789631	0.0172228	4.585	4.54e-06	***
TIPO_INTERSECCurva fuerte sin señalizar	0.1502046	0.0214700	6.996	2.63e-12	***
TIPO_INTERSECCurva fuerte con señal y sin velocidad señalizada	0.1099283	0.0377444	2.912	0.003586	**
LUMINOSIDADCrepúsculo	-0.0632888	0.0292504	-2.164	0.030488	*
LUMINOSIDADNoche: iluminación suave	0.0858289	0.0250017	3.433	0.000597	***
LUMINOSIDADNoche: iluminación insuficiente	-0.1751411	0.0316170	-5.539	3.03e-08	***
SUPERFICIE_CALZADAUmbría	-0.1229178	0.0949030	-1.295	0.195253	
SUPERFICIE_CALZADAMojada	-0.2811112	0.0202432	-13.887	<2e-16	***
SUPERFICIE_CALZADAHelada	-0.5493721	0.1598154	-3.438	0.000587	***
SUPERFICIE_CALZADANevada	-0.2075948	0.1551458	-1.338	0.180876	
SUPERFICIE_CALZADABarrillo	0.1998028	0.1513205	1.320	0.186703	
SUPERFICIE_CALZADAGravilla suelta	-0.1831457	0.0601023	-3.047	0.002310	**
EDAD	0.0023182	0.0003714	6.241	4.34e-10	***
POSICIONConductor vehículo	4.0680007	0.0557113	73.019	<2e-16	***
POSICIONPasajero delantero	3.3653694	0.0577146	58.310	<2e-16	***
POSICIONPasajero trasero izquierdo	4.0748524	0.0646132	63.065	<2e-16	***
POSICIONPasajero trasero derecho	3.9340791	0.0631235	62.324	<2e-16	***
POSICIONPasajero trasero central	4.0731335	0.0780230	52.204	<2e-16	***
POSICIONConductor vehículo de dos ruedas	0.6827406	0.0591053	11.551	<2e-16	***
POSICIONPasajero vehículo de dos ruedas	1.2055673	0.0906451	13.300	<2e-16	***
POSICIONOtros pasajeros sentados	3.7733857	0.0721549	52.296	<2e-16	***
POSICIONOtros pasajeros de pie	1.2710065	0.1981649	6.414	1.42e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

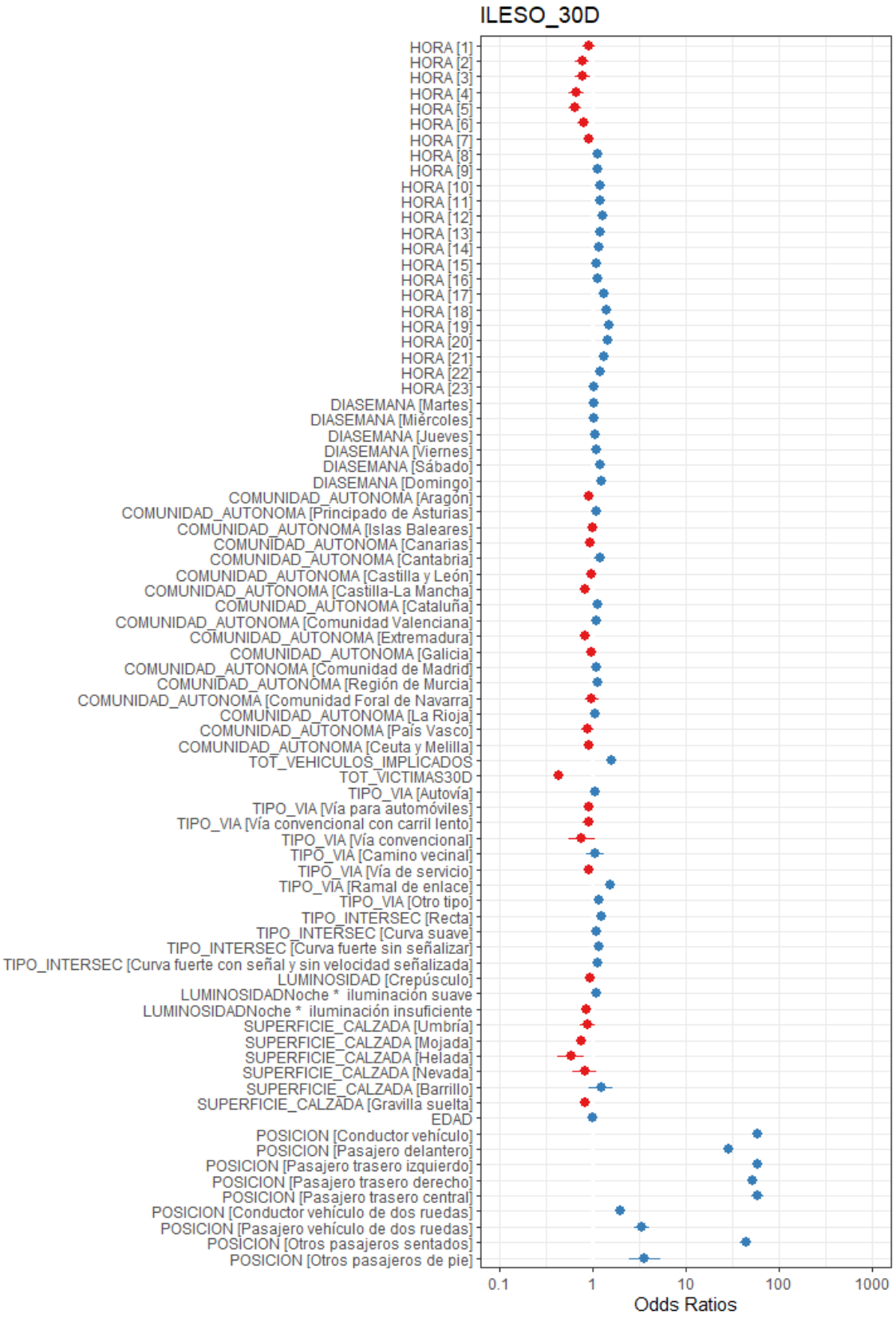

Ilustración 23. Odds ratios de los coeficientes del modelo final de regresión logística

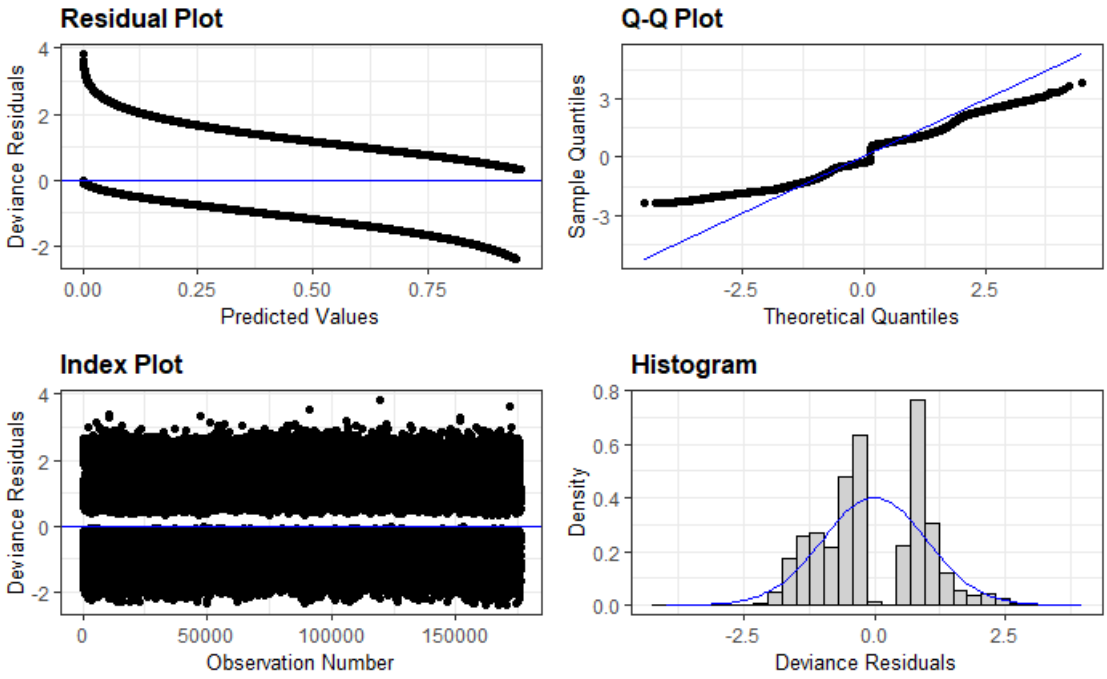

Ilustración 24. Análisis de residuos del modelo final de regresión logística

ANEXO B

Código en lenguaje R usado en la investigación:

```
# Los datos que se trabajan en el presente código han sido obtenidos de
# "https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/" -> Accidentes -> Accidentes
# -> Microdatos -> 2015 -> Descargar


## CARGA DE LOS DATOS

# Descompresión del archivo de datos descargado (se requiere que esté en el
# mismo directorio que el código):
un_files <- unzip("MICRODATOS_ACC_VICT_2015.zip",
                  exdir = "MICRODATOS_ACC_VICT_2015")

# Sólo se usan las tablas ACCVICT y PERS:
TABLA_ACCVICT_2015 <- read.csv2(un_files[2]) # cada línea es un accidente
TABLA_PERS_2015 <- read.csv2(un_files[3]) # cada línea es un individuo


## UNIÓN DE TABLAS

# Consulta en SQLite para la extracción de los datos cruzados de las
# tablas ACCVICT y PERS de modo que cada línea sea un individuo:
library(RSQLite)
library(DBI)
# Creación de una base de datos RSQLite efímera en memoria
con <- dbConnect(RSQLite::SQLite(), ":memory:")

# Carga de datos en la base de datos RSQLite creada:
dbWriteTable(con, "ACCVICT", TABLA_ACCVICT_2015)
dbWriteTable(con, "PERS", TABLA_PERS_2015)

# Ejecución de la consulta SQLite:
res <- dbSendQuery(con, "SELECT * FROM ACCVICT LEFT JOIN PERS ON
ACCVICT.ID_ACCIDENTE = PERS.ID_ACCIDENTE UNION ALL
SELECT * FROM PERS LEFT JOIN ACCVICT ON ACCVICT.ID_ACCIDENTE = PERS.ID_ACCIDENTE
WHERE ACCVICT.ID_ACCIDENTE IS NULL")
data <- dbFetch(res) # guardar datos en un objeto R

# Gestión de la consulta SQLite para liberar espacio en memoria:
dbClearResult(res) # limpiar los resultados
dbDisconnect(con) # desconectar la base de datos


## FILTRADO DE VARIABLES

# Eliminación de variables no incluidas en el diseño de registro que proporciona
# la Dirección General de Tráfico (DGT en adelante) sobre estos datos:
library(dplyr)
data$COD_MUNICIPIO <- NULL
data$USO_CINTURON <- NULL
data$USO_SRI <- NULL
data$USO_CASCO <- NULL
data$DICCIONARIO_MANIOBRAS <- NULL
data$INFRACC_COND <- NULL
data$INFRACC_APERTURA <- NULL
data$INFRACC_ALUMBRADO <- NULL
data$INFRACC_CARGA_VEHICULO <- NULL
data$INFRACC_RESUMEN <- NULL
data$DICCIONARIO_ACCION_PEATON <- NULL

# Eliminación de variables cuyo rango de valores no coinciden con los del diseño
# de registro de la DGT:
data$TIPO_ACCIDENTE <- NULL
data$MANIOBRAS <- NULL
data$SEXO <- NULL
data$ACCION_PEATON <- NULL

# Eliminación de variables superfluas y duplicadas:
data[2] <- NULL # ANIO, el único valor que toma es 2015
data[54] <- NULL # ANIO, el único valor que toma es 2015
data[37] <- NULL # duplicado de ID_ACCIDENTE
# solo importan las cifras consolidadas (30D)
data$TOT_VICTIMAS <- NULL
data$TOT_MUERTOS <- NULL
data$TOT_HERIDOS_GRAVES <- NULL
data$TOT_HERIDOS_LEVES <- NULL
data$MUERTO_24H <- NULL
data$HERIDO_GRAVE_24H <- NULL
data$HERIDO_LEVE_24H <- NULL
```

```
## CODIFICACIÓN DE VARIABLES

# Variable ID_ACCIDENTE:
data$ID_ACCIDENTE <- as.character(data$ID_ACCIDENTE)

# Variable MES:
data$MES <- factor(data$MES, labels = c("Enero",
                                        "Febrero",
                                        "Marzo",
                                        "Abril",
                                        "Mayo",
                                        "Junio",
                                        "Julio",
                                        "Agosto",
                                        "Septiembre",
                                        "Octubre",
                                        "Noviembre",
                                        "Diciembre"))

# Variable HORA:
data$HORA <- factor(data$HORA)

# Variable DIASEMANA:
data$DIASEMANA <- factor(data$DIASEMANA, labels = c("Lunes",
                                                    "Martes",
                                                    "Miércoles",
                                                    "Jueves",
                                                    "Viernes",
                                                    "Sábado",
                                                    "Domingo"))

# Variable PROVINCIA:
data$PROVINCIA <- factor(data$PROVINCIA, labels = c("Álava",
                                                    "Albacete",
                                                    "Alicante",
                                                    "Almería",
                                                    "Ávila",
                                                    "Badajoz",
                                                    "Islas Baleares",
                                                    "Barcelona",
                                                    "Burgos",
                                                    "Cáceres",
                                                    "Cádiz",
                                                    "Castelón",
                                                    "Ciudad Real",
                                                    "Córdoba",
                                                    "A Coruña",
                                                    "Cuenca",
                                                    "Girona",
                                                    "Granada",
                                                    "Guadalajara",
                                                    "Gipuzkoa",
                                                    "Huelva",
                                                    "Huesca",
                                                    "Jaén",
                                                    "León",
                                                    "Lleida",
                                                    "La Rioja",
                                                    "Lugo",
                                                    "Madrid",
                                                    "Málaga",
                                                    "Murcia",
                                                    "Navarra",
                                                    "Ourense",
                                                    "Asturias",
                                                    "Palencia",
                                                    "Las Palmas",
                                                    "Pontevedra",
                                                    "Salamanca",
                                                    "S.C. Tenerife",
                                                    "Cantabria",
                                                    "Segovia",
                                                    "Sevilla",
                                                    "Soria",
                                                    "Tarragona",
                                                    "Teruel",
                                                    "Toledo",
                                                    "Valencia",
                                                    "Valladolid",
                                                    "Bizkaia",
                                                    "Zamora",
                                                    "Zaragoza",
                                                    "Ceuta",
                                                    "Melilla"))

# Variable COMUNIDAD_AUTONOMA:
```

```
data$COMUNIDAD_AUTONOMA <- factor(data$COMUNIDAD_AUTONOMA,
                                  labels = c("Andalucía",
                                             "Aragón",
                                             "Principado de Asturias",
                                             "Islas Baleares",
                                             "Canarias",
                                             "Cantabria",
                                             "Castilla y León",
                                             "Castilla-La Mancha",
                                             "Cataluña",
                                             "Comunidad Valenciana",
                                             "Extremadura",
                                             "Galicia",
                                             "Comunidad de Madrid",
                                             "Región de Murcia",
                                             "Comunidad Foral de Navarra",
                                             "La Rioja",
                                             "País Vasco",
                                             "Ceuta y Melilla"))

# Variable ISLA:
library(tidyr)
data$ISLA <- replace_na(data$ISLA, 0) # NA == "No es isla"
data$ISLA <- factor(data$ISLA, labels = c("No es isla",
                                          "Mallorca",
                                          "Menorca",
                                          "Ibiza",
                                          "Formentera",
                                          "Gran Canaria",
                                          "Fuerteventura",
                                          "Lanzarote",
                                          "Tenerife",
                                          "La Palma",
                                          "Gomera",
                                          "Hierro"))

# Variable ZONA:
data$ZONA <- factor(data$ZONA, labels = c("Carretera", "Zona urbana", "Travesía"))

# Variable ZONA_AGRUPADA:
data$ZONA_AGRUPADA <- factor(data$ZONA_AGRUPADA, labels = c("Vías interurbanas",
                                                           "Vías urbanas"))

# Variable CARRETERA:
data$CARRETERA[data$CARRETERA == ""] <- NA

# Variable RED_CARRETERA:
data$RED_CARRETERA <- factor(data$RED_CARRETERA, labels = c("Titularidad Estatal",
                                                            "Titularidad Autonómica",
                                                            "Titularidad Provincial",
                                                            "Titularidad Municipal",
                                                            "Otras titularidades"))

# Variable TIPO_VIA:
data$TIPO_VIA <- factor(data$TIPO_VIA,
                        labels = c("Autopista",
                                   "Autovía",
                                   "Vía para automóviles",
                                   "Vía convencional con carril lento",
                                   "Vía convencional",
                                   "Camino vecinal",
                                   "Vía de servicio",
                                   "Ramal de enlace",
                                   "Otro tipo"))

# Variable TRAZADO_NO_INTERSEC:
data$TRAZADO_NO_INTERSEC <- replace_na(data$TRAZADO_NO_INTERSEC, 0)
# NA == "Es intersección"
data$TRAZADO_NO_INTERSEC[data$TRAZADO_NO_INTERSEC == 999] <- NA # 999 = NA
data$TRAZADO_NO_INTERSEC <- factor(data$TRAZADO_NO_INTERSEC,
                                   labels = c("Es intersección",
                                              "Recta",
                                              "Curva suave"))

# Variable TIPO_INTERSEC:
data$TIPO_INTERSEC <- replace_na(data$TIPO_INTERSEC, 0) # NA == "No es intersección"
data$TIPO_INTERSEC[data$TIPO_INTERSEC == 999] <- NA # 999 = NA
data$TIPO_INTERSEC <- factor(data$TIPO_INTERSEC,
                             labels = c("No es intersección",
                                        "Recta",
                                        "Curva suave",
                                        "Curva fuerte sin señalizar",
                                        "Curva fuerte con señal y sin velocidad señalizada"))

# Variable PRIORIDAD (nueva variable codificada a partir de las variables de
# PRIORIDAD: AGENTE, SEMÁFORO, STOP, CEDA, MARCAS, PASO, OTRA):
data <- mutate(data, PRIORIDAD = case_when(
```

```
  PRIORIDAD_AGENTE == 1 ~ 1,
  PRIORIDAD_SEMAFORO == 1 ~ 2,
  PRIORIDAD_STOP == 1 ~ 3,
  PRIORIDAD_CEDA == 1 ~ 4,
  PRIORIDAD_MARCAS == 1 ~ 5,
  PRIORIDAD_PASO == 1 ~ 6,
  PRIORIDAD_OTRA == 1 ~ 7
))
data$PRIORIDAD <- factor(data$PRIORIDAD, labels = c("Agente",
                                                    "Semáforo",
                                                    "Stop",
                                                    "Ceda el paso",
                                                    "Marcas viales",
                                                    "Paso de peatones",
                                                    "Otra"))
# eliminación de las variables que han sido agrupadas en PRIORIDAD
data$PRIORIDAD_AGENTE <- NULL
data$PRIORIDAD_SEMAFORO <- NULL
data$PRIORIDAD_STOP <- NULL
data$PRIORIDAD_CEDA <- NULL
data$PRIORIDAD_MARCAS <- NULL
data$PRIORIDAD_PASO <- NULL
data$PRIORIDAD_OTRA <- NULL

# Variable SUPERFICIE_CALZADA:
data$SUPERFICIE_CALZADA[data$SUPERFICIE_CALZADA == 999] <- NA
data$SUPERFICIE_CALZADA <- factor(data$SUPERFICIE_CALZADA, labels = c("Seca y limpia",
                                                                      "Umbría",
                                                                      "Mojada",
                                                                      "Helada",
                                                                      "Nevada",
                                                                      "Barrillo",
                                                                      "Gravilla suelta"))


# Variable LUMINOSIDAD:
data$LUMINOSIDAD <- factor(data$LUMINOSIDAD, labels = c("Pleno día",
                                                        "Crepúsculo",
                                                        "Noche: iluminación suave",
                                                        "Noche: iluminación insuficiente"))


# Variable FACTORES_ATMOSFERICOS:
data$FACTORES_ATMOSFERICOS[data$FACTORES_ATMOSFERICOS == 999] <- NA
data$FACTORES_ATMOSFERICOS <- factor(data$FACTORES_ATMOSFERICOS,
                                     labels = c("Buen tiempo",
                                                "Niebla intensa",
                                                "Niebla ligera",
                                                "Lloviznando",
                                                "Lluvia fuerte",
                                                "Granizando",
                                                "Nevando",
                                                "Viento fuerte",
                                                "Otro"))

# Variable VISIBILIDAD_RESTRINGIDA:
data$VISIBILIDAD_RESTRINGIDA[data$VISIBILIDAD_RESTRINGIDA == 999] <- NA
data$VISIBILIDAD_RESTRINGIDA <- factor(data$VISIBILIDAD_RESTRINGIDA,
                                       labels = c("Edificios",
                                                  "Configuración del terreno",
                                                  "Vegetación",
                                                  "Factores atmosféricos",
                                                  "Deslumbramiento",
                                                  "Polvo o humo",
                                                  "Otra causa"))

# Variable ACERAS:
data$ACERAS[data$ACERAS == 999] <- NA
data$ACERAS[data$ACERAS == 998] <- NA
data$ACERAS <- factor(data$ACERAS, labels = c("No hay acera", "Sí hay acera"))

# Variable ID_VEHICULO:
data$ID_VEHICULO <- replace_na(data$ID_VEHICULO, "No es vehículo (peatón)")
# NA == "No es vehículo (peatón)"

# Variable ID_PERSONA:
data$ID_PERSONA <- factor(data$ID_PERSONA, labels = c("Conductor", "Pasajero", "Peatón"))

# Variable ID_CONDUCTOR:
data$ID_CONDUCTOR <- replace_na(data$ID_CONDUCTOR, "No conductor (pasajero o peatón)")
# NA == "No conductor (pasajero o peatón)"

# Variable ID_PASAJERO:
data$ID_PASAJERO <- replace_na(data$ID_PASAJERO, "No pasajero (conductor o peatón)")
# NA == "No pasajero (conductor o peatón)"

# Variable ID_PEATON:
```

```
data$ID_PEATON <- replace_na(data$ID_PEATON, "No peatón (conductor o pasajero)")
# NA == "No peatón (conductor o pasajero)"

# Variable EDAD:
data$EDAD[data$EDAD == 999] <- NA

# Variable ANIO_PERMISO:
data$ANIO_PERMISO <- replace_na(data$ANIO_PERMISO, "No aplica (pasajero o peatón)")
data$ANIO_PERMISO[data$ANIO_PERMISO == "9999"] <- NA

# Variable POSICION:
data$POSICION <- replace_na(data$POSICION, 0) # NA == "Peatón"
data$POSICION[data$POSICION == 99] <- NA
data$POSICION <- factor(data$POSICION, labels = c("Peatón",
                                                  "Conductor vehículo",
                                                  "Pasajero delantero",
                                                  "Pasajero trasero izquierdo",
                                                  "Pasajero trasero derecho",
                                                  "Pasajero trasero central",
                                                  "Conductor vehículo de dos ruedas",
                                                  "Pasajero vehículo de dos ruedas",
                                                  "Otros pasajeros sentados",
                                                  "Otros pasajeros de pie"))

# Variable ILESO_30D (nueva variable codificada a partir de MUERTO_30D,
# HERIDO_GRAVE30D, HERIDO_LEVE30D):
data <- mutate(data, ILESO_30D = case_when(
  MUERTO_30D == 1 ~ 0,
  HERIDO_GRAVE30D == 1 ~ 0,
  HERIDO_LEVE30D == 1 ~ 0
))
data$ILESO_30D <- replace_na(data$ILESO_30D, 1)
data$ILESO_30D <- factor(data$ILESO_30D, labels = c("No ileso", "Ileso"))

# Variable MUERTO_30D:
data$MUERTO_30D <- factor(data$MUERTO_30D, labels = c("No fallecido", "Fallecido"))

# Variable HERIDO_GRAVE30D:
data$HERIDO_GRAVE30D <- factor(data$HERIDO_GRAVE30D, labels = c("No herido grave",
                                                                "Herido grave"))

# Variable HERIDO_LEVE30D:
data$HERIDO_LEVE30D <- factor(data$HERIDO_LEVE30D, labels = c("No herido leve",
                                                              "Herido leve"))

# Variable INFRACC_VELOCIDAD:
data$INFRACC_VELOCIDAD <- replace_na(data$INFRACC_VELOCIDAD, 0)
# NA == "No aplica (pasajero o peatón)"
data$INFRACC_VELOCIDAD <- factor(data$INFRACC_VELOCIDAD,
                                 labels = c("No aplica (pasajero o peatón)",
                                            "Velocidad inadecuada para las condiciones existentes",
                                            "Sobrepasar la velocidad establecida",
                                            "Marcha lenta entorpeciendo la circulación",
                                            "Ninguna"))

# Variable INFRACC_PEATON:
data$INFRACC_PEATON <- replace_na(data$INFRACC_PEATON, 0)
# NA == "No aplica (conductor o pasajero)"
data$INFRACC_PEATON <- factor(data$INFRACC_PEATON,
                              labels = c("No aplica (conductor o pasajero)",
                                         "No respetar señal de semáforo",
                                         "No utilizar paso para peatones",
                                         "No respetar señal de agente",
                                         "Irrumpir o cruzar la vía antirreglamentariamente"))


## CREACIÓN DE SUBSET DE DATOS SIN COMBINACIONES LINEALES

# Subset de los datos con ciertas variables de modo que no hayan problemas de que
# unas variables sean combinaciones lineales de otras:
data_subset <- data %>% select(!c(TOT_MUERTOS30D,
                                  TOT_HERIDOS_GRAVES30D,
                                  TOT_HERIDOS_LEVES30D,
                                  MUERTO_30D, HERIDO_GRAVE30D,
                                  HERIDO_LEVE30D))


## VALORES PERDIDOS

library(questionr)
head(freq.na(data_subset),11)

# Eliminación de variables con >5% de valores perdidos:
data_subset$ACERAS <- NULL
data_subset$VISIBILIDAD_RESTRINGIDA <- NULL
data_subset$PRIORIDAD <- NULL
data_subset$CARRETERA <- NULL
```

```
data_subset$ANIO_PERMISO <- NULL
data_subset$TRAZADO_NO_INTERSEC <- NULL
data_subset$MUNICIPIO <- NULL

# Guardado de datos en archivo csv
library(readr)
write_excel_csv2(data, "data.csv")

# Resumen de las variables pre-imputación valores perdidos:
na_variables <- data_subset %>% select(FACTORES_ATMOSFERICOS,
                                       EDAD,
                                       SUPERFICIE_CALZADA,
                                       POSICION)
summary(na_variables)

# Imputación de valores perdidos. Imputación múltiple usando Bootstrap y PMM:
set.seed(123)
library(Hmisc)
# puede tardar varios minutos
impute_arg <- aregImpute(~ FACTORES_ATMOSFERICOS + EDAD + SUPERFICIE_CALZADA + POSICION,
                         data = data_subset, n.impute = 5)
impute_arg
imputed <- impute.transcan(impute_arg, imputation = 1, data = data_subset,
                           list.out = TRUE, pr = FALSE, check = FALSE)

# Imputación de los valores obtenidos en las variables correspondientes:
data_subset$FACTORES_ATMOSFERICOS <- imputed$FACTORES_ATMOSFERICOS
data_subset$EDAD <- imputed$EDAD
data_subset$SUPERFICIE_CALZADA <- imputed$SUPERFICIE_CALZADA
data_subset$POSICION <- imputed$POSICION
head(freq.na(data_subset),11)

# Resumen de las variables post-imputación valores perdidos:
na_variables <- data_subset %>% select(FACTORES_ATMOSFERICOS, EDAD,
                                       SUPERFICIE_CALZADA, POSICION)
summary(na_variables)


## DATOS ATÍPICOS A NIVEL MULTIVARIANTE

library(rstatix)
(m_outliers <- data_subset %>%
    mahalanobis_distance() %>%
    filter(is.outlier == TRUE))
nrow(m_outliers)
(nrow(m_outliers)/nrow(data_subset))*100
m_dist <- mahalanobis(data_subset[, c(8,9,23)], colMeans(data_subset[, c(8,9,23)]),
                      cov(data_subset[, c(8,9,23)]))
data_subset$m_dist <- m_dist
data_subset$outlier_maha <- "No"
data_subset$outlier_maha[data_subset$m_dist > min(m_outliers$mahal.dist)] <- "Yes"
summary(factor(data_subset$outlier_maha))
outliers <- filter(data_subset, outlier_maha == "Yes")
outliers %>% select(TOT_VICTIMAS30D, TOT_VEHICULOS_IMPLICADOS, EDAD) %>% summary()

# Filtrado de outliers:
data_subset <- filter(data_subset, outlier_maha == "No")


## ANÁLISIS DE VARIABLES

library(summarytools)

# Variable ILESO_30D:
summary(data_subset$ILESO_30D)
#
nrows <- nrow(data_subset)
percentData <- data_subset %>% group_by(ILESO_30D) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/nrows))
ggplot(data_subset, aes(x = factor(1), fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  coord_polar("y", start = 0) +
  theme_void() +
  theme(legend.title = element_text(size = 10, face = "bold"))

# Variable MES:
ctable(data_subset$MES, data_subset$ILESO_30D)
#
ggplot(data_subset, aes(MES, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA)) +
  labs(x = "MES", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
```

```
        axis.text.x = element_text(size = 10, angle = 90, hjust = 0.95, vjust = 0.2),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))

# Variable HORA:
ctable(data_subset$HORA, data_subset$ILESO_30D)
#
ggplot(data_subset, aes(HORA, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA)) +
  labs(x = "HORA DEL DÍA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))

# Variable DIASEMANA:
ctable(data_subset$DIASEMANA, data_subset$ILESO_30D)
#
ggplot(data_subset, aes(x = DIASEMANA, fill = ILESO_30D)) +
  geom_bar(position = position_dodge(preserve = "single"), width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA)) +
  labs(x = "DÍA DE LA SEMANA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))

# Variable COMUNIDAD_AUTONOMA:
ctable(data_subset$COMUNIDAD_AUTONOMA, data_subset$ILESO_30D)
#
ggplot(data_subset, aes(x = COMUNIDAD_AUTONOMA, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA)) +
  labs(x = "COMUNIDAD AUTÓNOMA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable TOT_VICTIMAS30D:
data_subset %>%
  group_by(ILESO_30D) %>%
  summarise(mean = mean(TOT_VICTIMAS30D),
            sd = sd(TOT_VICTIMAS30D),
            n = n())
#
ggplot(data_subset, aes(x = ILESO_30D, y = TOT_VICTIMAS30D,
                        fill = ILESO_30D)) +
  geom_boxplot(width = 0.7) +
  stat_summary(fun.y="mean", shape = 4) +
  theme_bw() +
  labs(x = "", y = "TOTAL INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable TOT_VEHICULOS_IMPLICADOS:
data_subset %>%
  group_by(ILESO_30D) %>%
  summarise(mean = mean(TOT_VEHICULOS_IMPLICADOS),
            sd = sd(TOT_VEHICULOS_IMPLICADOS),
            n = n())
#
ggplot(data_subset, aes(x = ILESO_30D, y = TOT_VEHICULOS_IMPLICADOS,
                        fill = ILESO_30D)) +
  geom_boxplot(width = 0.7) +
  stat_summary(fun.y="mean", shape = 4) +
  theme_bw() +
  labs(x = "", y = "TOTAL VEHÍCULOS IMPLICADOS EN ACCIDENTES") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
```

```
# Variable ZONA:
ctable(data_subset$ZONA, data_subset$ILESO_30D)
#
ggplot(data_subset, aes(x = ZONA, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "ZONA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable TIPO_VIA:
ctable(data_subset$TIPO_VIA, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(TIPO_VIA) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = TIPO_VIA, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "TIPO DE VÍA", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))
#
ggplot(data_subset, aes(x = TIPO_VIA, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "TIPO DE VÍA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable SUPERFICIE_CALZADA:
ctable(data_subset$SUPERFICIE_CALZADA, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(SUPERFICIE_CALZADA) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = SUPERFICIE_CALZADA, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "SUPERFICIE DE LA CALZADA", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))
#
ggplot(data_subset, aes(x = SUPERFICIE_CALZADA, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "SUPERFICIE DE LA CALZADA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))

# Variable LUMINOSIDAD:
ctable(data_subset$LUMINOSIDAD, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(LUMINOSIDAD) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = LUMINOSIDAD, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "LUMINOSIDAD DE LA VÍA", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
```

```
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = LUMINOSIDAD, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "LUMINOSIDAD DE LA VÍA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable TIPO_INTERSEC:
ctable(data_subset$TIPO_INTERSEC, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(TIPO_INTERSEC) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = TIPO_INTERSEC, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "TIPO DE INTERSECCIÓN", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = TIPO_INTERSEC, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "TIPO DE INTERSECCIÓN", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable FACTORES_ATMOSFERICOS:
ctable(data_subset$FACTORES_ATMOSFERICOS, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(FACTORES_ATMOSFERICOS) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = FACTORES_ATMOSFERICOS, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "FACTORES ATMOSFÉRICOS", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10, angle = 90, hjust = 0.95, vjust = 0.2),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))
#
ggplot(data_subset, aes(x = FACTORES_ATMOSFERICOS, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "FACTORES ATMOSFÉRICOS", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable EDAD:
data_subset %>%
  group_by(ILESO_30D) %>%
  summarise(mean = mean(EDAD),
            sd = sd(EDAD),
            n = n())
#
ggplot(data_subset, aes(x = ILESO_30D, y = EDAD,
```

```
                          fill = ILESO_30D)) +
  geom_violin(alpha = 0.2) +
  geom_boxplot(width = 0.3) +
  stat_summary(fun.y="mean", shape = 4) +
  theme_bw() +
  labs(x = "", y = "EDAD DEL ACCIDENTADO") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = EDAD)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mean(data_subset$EDAD),
                                         sd = sd(data_subset$EDAD))) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA)) +
  labs(x = "EDAD DEL INDIVIDUO", y = "DENSIDAD") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))
# Anderson-Darling normality test (para muestras grandes):
library(nortest)
ad.test(data_subset$EDAD)

# Variable POSICION:
ctable(data_subset$POSICION, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(POSICION) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = POSICION, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5), size = 3) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "POSICIÓN DEL INDIVIDUO", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = POSICION, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "POSICIÓN DEL INDIVIDUO", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable INFRACC_VELOCIDAD:
ctable(data_subset$INFRACC_VELOCIDAD, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(INFRACC_VELOCIDAD) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = INFRACC_VELOCIDAD, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "INFRACCIÓN DE VELOCIDAD COMETIDA", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = INFRACC_VELOCIDAD, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "INFRACCIÓN DE VELOCIDAD", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
```

```
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()

# Variable INFRACC_PEATON:
ctable(data_subset$INFRACC_PEATON, data_subset$ILESO_30D)
#
percentData <- data_subset %>% group_by(INFRACC_PEATON) %>% count(ILESO_30D) %>%
  mutate(ratio = scales::percent(n/sum(n)))
ggplot(data_subset, aes(x = INFRACC_PEATON, fill = ILESO_30D)) +
  geom_bar(position = 'fill') +
  geom_text(data = percentData, aes(y = n, label = ratio),
            position = position_fill(vjust = 0.5), size = 3) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "INFRACCIÓN DE PEATÓN", y = "% INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()
#
ggplot(data_subset, aes(x = INFRACC_PEATON, fill = ILESO_30D)) +
  geom_bar(width = 0.7) +
  theme_bw() +
  scale_y_continuous(expand = c(0, NA), labels = scales::comma_format(big.mark = "")) +
  labs(x = "INFRACCIÓN DE PEATÓN COMETIDA", y = "INDIVIDUOS ACCIDENTADOS") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold")) +
  coord_flip()


## ANÁLISIS CON DOS PREDICTORES

data_subset %>%
  group_by(ILESO_30D, ID_PERSONA) %>%
  summarise(mean = mean(EDAD),
            sd = sd(EDAD))

ggplot(data_subset, aes(x = ILESO_30D, y = EDAD, fill = ID_PERSONA)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "", y = "EDAD DEL INDIVIDUO") +
  scale_fill_discrete(name = "POSICIÓN") +
  theme(axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10),
        legend.title = element_text(size = 10, face = "bold"))


## DIVISIÓN DE LOS DATOS EN TRAINING Y TEST

library(caret)
index <- createDataPartition(y = data_subset$ILESO_30D, p = 0.75, list = FALSE)
# 75% training, 25% test
training_data <- data_subset[index, ]
test_data <- data_subset[-index, ]
dim(test_data)
dim(training_data)


## AJUSTE DEL MODELO DE REGRESIÓN LOGÍSTICA

# Ajuste:
fit1 <- glm(ILESO_30D ~ HORA + DIASEMANA + COMUNIDAD_AUTONOMA + TOT_VEHICULOS_IMPLICADOS +
              TOT_VICTIMAS30D + TIPO_VIA + TIPO_INTERSEC + LUMINOSIDAD + SUPERFICIE_CALZADA +
              EDAD + POSICION, data = training_data, family = binomial)
summary(fit1)

# Coeficientes:
round(exp(coef(fit1)), 2)
library(sjPlot)
plot_model(fit1) + theme_bw()


## IMPORTANCIA DE LAS VARIABLES PREDICTORAS

fit.caret <- train(ILESO_30D ~ HORA + DIASEMANA + COMUNIDAD_AUTONOMA +
                     TOT_VEHICULOS_IMPLICADOS + TOT_VICTIMAS30D + TIPO_VIA + TIPO_INTERSEC +
                     LUMINOSIDAD + SUPERFICIE_CALZADA + EDAD + POSICION,
                   data = training_data, method = "glm")
(imp <- varImp(fit.caret, scale = FALSE))
```

```
plot(imp, top = 10, xlab="IMPORTANCIA DE LA VARIABLE")


## DIAGNÓSTICO DEL MODELO

# R-Cuadrado:
library(performance)
r2_tjur(fit1)

# Prueba de multicolinealidad (factor de inflación de la varianza):
library(car)
vif(fit1)

# Contraste de Hosmer & Lemeshow de la calidad del ajuste:
library(ResourceSelection)
hoslem.test(fit1$y, fitted(fit1))

# ANOVA:
anova(fit1, test = "Chisq")
Anova(fit1)

# Panel de diagnósticos de residuos:
library(ggResidpanel)
resid_panel(fit1)


## EVALUACIÓN DE LA CAPACIDAD PREDICTIVA DEL MODELO

# Cálculo y transformación de la predicción a partir de los datos de prueba
pdata <- predict(fit1, newdata = test_data, type = "response")
prediction <- c()
z = 1
pdata <- as.numeric(pdata > 0.5)
for(i in pdata) {
  if (i == 0) {
    prediction[z] <- 1
  }
  else {
    prediction[z] <- 2
  }
  z <- z+1
}

# Matriz de confusión:
prediction <- factor(as.numeric(prediction), labels = c("No ileso", "Ileso"))
confusionMatrix(data = prediction, reference = test_data$ILESO_30D)

# Curva ROC:
library(Epi)
ROC(prediction, test_data$ILESO_30D, plot = "ROC")
```