uc3m | Universidad Carlos III de Madrid

Máster Universitario en Ciencias Actuariales y Financieras
2017-2018

Trabajo Fin de Máster

"Predicción de la severidad de accidentes de tráfico con víctimas mediante *Random Forest*"

Alejandro Rubén Domingo Gesteiro

Tutor/es

José Miguel Rodriguez-Pardo del Castillo

Jesús Ramón Simón del Potro

Madrid – junio 2018

Resumen:

Los accidentes de tráfico han sido principalmente objeto de estudio actuarial en relación con las carteras de asegurados calibrándose con las características de los individuos. Este trabajo toma otro enfoque, donde desarrolla dos métodos predictivos de *machine learning* para estudiar la severidad de los accidentes de tráfico a partir de variables externas al accidente. En primer lugar, se modeliza un *random forest* para predecir la severidad de las víctimas de accidentes de tráfico no fallecidas; y por otro lado un modelo lineal generalizado binomial para predecir víctimas mortales. Para las dos modelizaciones, las validaciones cruzadas de 5 y 10 iteraciones respectivamente arrojan aproximadamente una precisión del 98% para el *random forest* y un 95% para el modelo lineal generalizado. Este trabajo tiene una limitación interna al estimar modelos simples y por otra parte una limitación externa, debido a que los modelos estudian variables mayoritariamente externas, siendo interesante incluir además variables endógenas al individuo en futuros estudios.

Traffic accidents have been the subject of an actuarial study in relation to insurance portfolios through the characteristics of individuals. This work takes another approach, two predictive methods of machine learning are computed to study the severity of traffic accidents from variables external to the accident. First, a random forest is modeled to predict the severity of victims of non-fatal traffic accidents. On the other hand, a generalized linear binomial model is estimated to predict fatalities. For the two models, the 5 and 10-fold cross validations, respectively, yield approximately 98% accuracy for the random forest and 95% for the generalized linear model. This work has an external limitation of estimating simple models and, on the other hand, an external limitation, due to the fact that the variables are mainly external, which would be interesting adding variables endogenous to the individual in future studies.

Palabras clave: *Random forest*, *machine learning*, accidentes, tráfico, víctimas.

Dedicaciones

A mis padres, Julio y María, a mi novia Fabby, y mi familia.

"*The key to artificial intelligence has always been the representation*"

– Jeff Hawkins

Contenido

1. Introducción

1.1. Motivación del trabajo

Durante los últimos años el *big data* se ha desarrollado en profundidad en diversos ámbitos, lo que ha llevado al nacimiento de nuevas tendencias en modelización como la inteligencia artificial, *Machine Learning* o *Deep Learning*. Gracias a ello, el desarrollo de los modelos estadísticos y predictivos ha evolucionado considerablemente, al igual que los soportes físico-digitales que realizan sus cálculos. Aunque se considere un tema de actualidad, algunos de los algoritmos utilizados hoy día datan de la mitad del siglo veinte, siendo el componente computacional el que ha retrasado su evolución tal y como describe Langley (2011).

Según Delbridge (2017), el sector asegurador también se encuentra en esta transición hacia las nuevas tecnologías. Según el autor, las entidades aseguradoras e instituciones gubernamentales denotan que la modelización por ser tradicional no implica que sea la más precisa, instando por ello a que se produzca esta transición. Sin embargo, este sector, debido a su naturaleza conservadora y altamente regulada, realiza esta transición de manera pausada.

En este punto, Morgan (2017) añade qué, si bien las técnicas actuariales por ser tradicionales no tienen por qué ser inferiores estadísticamente, las nuevas tendencias computacionales añaden una nueva perspectiva de desarrollo en el que gracias al aprendizaje estadístico se podrán mejorar diversas áreas del sector asegurador como el cálculo actuarial, la experiencia del cliente en la contratación o la experiencia de este al recibir sus prestaciones.

El desarrollo tecnológico también ha permitido mejorar la calidad de vida, aumentar la seguridad y la prevención de accidentes con nuevos sistemas como el frenado automático. Sin embargo, aun considerando nuestra evolución digital los accidentes siguen siendo un factor mortal que no se logra erradicar con un nuevo chip o el creciente uso de datos masivos.

Los accidentes, en especial los de tráfico, siguen siendo desde hace muchos años un foco a erradicar por muchos gobiernos, empresas y comunidades. En el caso de España, los accidentes de tráfico generan cientos de miles de víctimas donde un 1,6% del total de accidentes con víctimas sufren la muerte[1]. Las cifras de fallecidos con

[1] Dirección General de Tráfico. (2017). *Anuario Estadístico de Accidentes 2016*. [Documento online]. Recuperado de http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/anuario-estadistico-de-accidentes/Anuario-accidentes-2016.pdf

respecto al de totales no es tan alta como otras defunciones no naturales, pero detrás de este porcentaje subyacen 1.663 fallecidos. Si bien la industrialización ha mejorado la capacidad de movilidad a grandes distancias con vehículos, la tecnología digital podría prevenir estos accidentes y, en todo caso, ayudar a los servicios médicos a salvar vidas atendiendo a las víctimas en menor tiempo y con mejor equipación.

Los seguros de automóviles tienen como objetivo social el compensar la pérdida fortuita de víctimas de accidentes de tráfico mediante la indemnización de los gastos en salud o capitales alzados a los perjudicados o aquellos relativos a la víctima. Esta indemnización por un lado amengua el riesgo que acarrean los conductores a poder hacer frente a eventos azarosos de altos capitales como la responsabilidad civil. Pero, por otro lado, el seguro tiene un objetivo social de ayudar o compensar un perjuicio o pérdida a las víctimas o relativos de un accidente de tráfico (en el caso de los seguros de vehículos a motor). No tienen el mismo objetivo las garantías aseguradas de reparaciones de bajo coste ya que no responden a un fin humano sino a un ámbito de presupuesto personal.

Asimismo, este trabajo puede ser de ayuda en los seguros al estimar la severidad de los accidentes de tráfico en base a los factores externos que puedan ser deterministas en el modelo predictivo. Por ejemplo, por tipo de vehículo, uso de Sistema de Retención Infantil (SRI) o el año de matrícula del vehículo (indicando por consiguiente una temporalidad aproximada de la antigüedad del vehículo). Ciertos factores geográficos como la comunidad autónoma, provincia o la fecha y hora pueden ser potenciales predictores de accidentes comunes que pueden referenciarse en diferentes recargos en los seguros de autos anuales.

1.2. Objetivos

El objetivo de este trabajo trata, en primer lugar, sobre la predicción de la severidad categorizada como heridos no hospitalizados (lesiones leves), heridos con necesidad hospitalaria (lesiones graves) y fallecidos, explicado mediante factores externos de la carretera, condiciones del accidente o factores del vehículo. Gracias a estas predicciones, y como segundo objetivo, se pretende estudiar un modelo predictivo mediante *Machine Learning* para prevenir accidentes de alta intensidad.

Este modelo podría incorporarse en automóviles autónomos o como instrumento a los servicios sanitarios con el fin de obtener información previa a la llegada al lugar del accidente, gracias a los parámetros ofrecidos por el vehículo de manera telemática, de

la gravedad de las lesiones de los integrantes del vehículo accidentado y las necesidades médicas que requerirán.

Por otro lado, el objetivo social de este trabajo se basa en reducir las víctimas que puedan deberse a accidentes de tráfico mediante la prevención de los mismos gracias al análisis de factores externos. En la siguiente ilustración se observa la evolución de fallecidos por accidentes de tráfico en España desde el 1960 hasta el 2015, donde la tendencia de los últimos 26 años es, en promedio, decreciente[1].

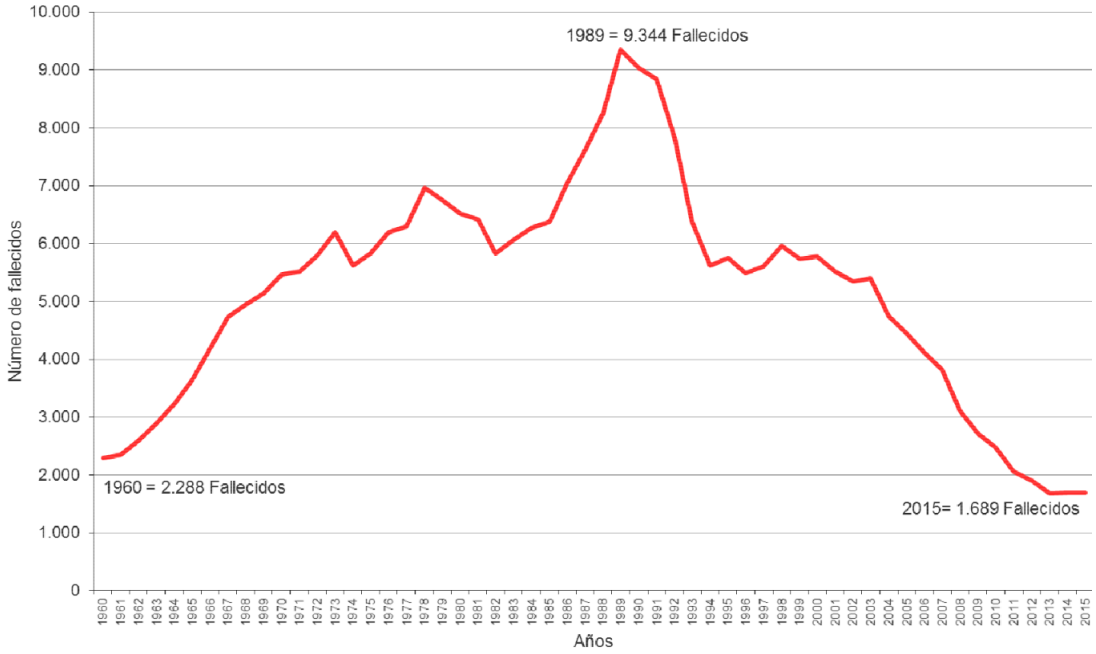


Ilustración 1. Evolución de víctimas mortales en accidentes de tráfico en España

Fuente: Anuario Estadístico de Accidentes 2015. Dirección General de Tráfico

Las mejoras en la seguridad de los vehículos, la conducción automática o el aviso automatizado de una emergencia, sin duda pueden ayudar a reducir el número de fallecidos anuales ilustrados en el gráfico anterior. Sin embargo, el apalancamiento de fallecidos por accidentes de tráfico a partir del 2012 no debe considerarse una tasa permanente, puesto que el incremento en la seguridad vial y el desarrollo en la conducción asistida instan conllevar a decrecimiento concluyente en la reducción de víctimas mortales de accidentes de tráfico.

Las lesiones graves además pueden permitir a las aseguradoras, apreciar la gravedad de los accidentes prontamente, dando paso a poder aplicar tecnologías modernas como la automatización de los partes de accidentes o la aplicación de la metodología *blockhain* en la asignación de sumas aseguradas antes de recibir las reclamaciones. Sin

embargo, los datos de este trabajo son de carácter público, por lo que no dispone de desglose de la lesión corporal y/o mental sufrida, no pudiendo aplicar un modelo concreto en estos aplicativos.

La importancia de esta temática engloba tanto un objetivo social como un objetivo de negocio como puede ser la mejora del servicio o un ajuste en el *pricing* de seguros de automóviles o seguros de responsabilidad civil. La importancia de esta temática es amplia y su aplicación puede determinar mejoras notables en el negocio asegurador en el ramo de no vida, así como una mejora en los servicios sanitarios, o en la seguridad vial española.

2. Revisión de literatura

El alcance metodológico de este trabajo se basa en la predicción de víctimas de accidentes de tráfico en categorías diferenciadas aplicando técnicas de *machine learning* como el árbol de aprendizaje de clasificación o regresión, o el bosque aleatorio.

Estudios similares, donde se enfatiza el uso de técnicas de *machine learning*, se han llevado a cabo como el trabajo de Moreno (2017) sobre la predicción de frecuencias en accidentes de autos en Barcelona atendiendo a factores meteorológicos y a los días festivos en la ciudad, donde oleadas de turismo local pueden determinar un aumento significativo en accidentes de tráfico. Otro ejemplo es el trabajo de Lin *et al.* (2017), donde además de proponer un método de selección de variables basado en árboles de patrones de frecuencia, estudian cuál es el mejor modelo predictivo para los accidentes de tráfico en Estados Unidos, considerados ruidosos y heterogéneos estadísticamente. Por otro lado, Chin y Quddus (2003) analizan también la frecuencia, en este caso, de los accidentes de tráfico en intersecciones señalizadas en Singapur, aplicando una distribución binomial negativa con efectos aleatorios (*Random effect negative binomial* en inglés) debido a la heterogeneidad inobservada y la correlación serial de las observaciones, donde además uno de los factores determinantes fue el volumen de vehículos, uno de los factores incluidos en la muestra para este trabajo.

Sin embargo, ninguno de ellos ha realizado un enfoque sobre la predicción de la severidad de las víctimas de accidentes de tráfico basada en factores externos circunstanciales alrededor del accidente en España en los últimos años. Si bien los datos para este trabajo son de carácter público, implican una depuración y calidad del dato previa que se considera necesaria a la hora de aplicar un modelo y esperar resultados no sesgados. Sobre un tema similar trabajaron Beshah et al. (2012) mediante

el estudio de los factores para poder determinar la precisión predictiva de un modelo que analiza los accidentes de tráfico en general en Etiopía, los datos recabados, la tendencia y los patrones relevantes, con el objetivo de conseguir mejorar la seguridad vial. Por otro lado, Chen y Chen (2017) analiza ciertos factores, alguno externo, para mapear los accidentes de tráfico en la ciudad de Shanghai y alrededores mediante modelos predictivos como árboles de decisión, regresiones lineales y bosques aleatorios. Como puede observarse en estos trabajos el poder predictivo de estas técnicas está siendo utilizado en diversos estudios de accidentes.

Otros estudios con estas técnicas predictivas se han llevado a cabo en el ámbito de los seguros, especialmente mediante el uso del *random forest* como puede observarse en los trabajos de Alshamsi (2014) y de Lin et al. (2017), donde analizan la aplicación de un *random forest* parametrizado en el ámbito del *big data* de los seguros en China.

3. Metodología

El Machine learning se ha convertido en los últimos años en un tema relevante dado el aumento de la capacidad computacional de la que se dispone actualmente. Esto trae consigo la mejora de las técnicas predictivas gracias a la investigación de modelos matemático-estadísticos realizados por de las empresas con el fin de conocer con altas probabilidades la barrera de la incertidumbre.

Este trabajo utiliza un bosque aleatorio (Random Forest en inglés) con el fin de generar un modelo de clasificación, mediante patrones externos al conductor de un vehículo, que pueda predecir con cierto margen de error la gravedad del daño sufrido por las víctimas de un accidente de tráfico.

Para entender la metodología de un random forest, primero se debe entender el cálculo de un árbol de clasificación o regresión desarrollado en primer lugar por Ho (1995). Este posteriormente fue ampliado por Breiman y Cutler (2011) publicado en el paquete R de “randomForest” por Liaw y Wiener (2012). Este paquete es el utilizado en la modelización del presente trabajo.

3.1. Árbol de clasificación o regresión

Un árbol de clasificación o regresión, también conocido como árbol de aprendizaje (por sus siglas ACR) es una técnica comúnmente utilizada en *machine learning* que o bien clasifica en clústeres (árbol de clasificación en castellano), o bien analiza mediante una

regresión lineal (árbol de regresión), observaciones según sus características (variables independientes) para estudiar un elemento concreto (variable dependiente u objetivo) representando los resultados en forma de árbol de decisión. Un árbol de decisión es una representación similar a un árbol invertido, donde partiendo de un "tronco" o nodo de origen, se subdivide la muestra en dos submuestras (ramificaciones) según avanza la estructura, y finalmente se representan las clases finales en "hojas" (terminaciones o ramas finales). Una gran ventaja de estas técnicas es que su representación gráfica es muy intuitiva y sencilla de analizar.

Los árboles de decisión se caracterizan por ser de clasificación o de regresión, según si la variable objetivo a estudiar es una variable categórica o una variable continua, respectivamente. Asimismo, según el tipo de árbol el criterio de selección de cada nodo, es decir, el criterio que decide si una observación se clasifica hacia la izquierda o hacia la derecha del nodo, también varía existiendo varios tipos de árboles según el criterio escogido.

Metodológicamente cada nodo se divide[2] en dos 'ramas' condicionado a cumplir o no un criterio de selección de cada nodo. Si para una observación la condición del criterio es correcta, esta observación se direcciona, por consenso general, hacia la submuestra de la izquierda. Contrariamente si el criterio de selección no se cumple, la observación se clasifica hacia el subconjunto de la derecha de dicho nodo, y así sucesivamente a lo largo del árbol.

Aunque el ACR por normal general clasifica las observaciones en dos ramificaciones por cada nodo, esto no implica que no se analice la interacción de una variable en diferentes niveles o bien las interacciones entre variables, puesto que permite repetir los criterios en nodos posteriores si se consideran criterios explicativos o que aporten información relevante al modelo, creando así interacciones de variables como pudiera ser sexo y edad.

Véase la siguiente figura a modo de ejemplo donde el efecto de la representación de un árbol de clasificación o regresión a la izquierda es igual que una división binaria por criterios consecutivos lineales; esto también puede observarse en el gráfico en tres dimensiones a su derecha:

[2] Se evita el término "cortar" el árbol ya que puede confundirse con "podar" el árbol siendo estas acciones diferentes. En inglés "cut", o cortar en español, se refiere a cuando el nodo se divide ("Split" en inglés) en dos submuestras considerando un criterio de selección. Por otra parte "podar" ("prune" en inglés) se refiere a reducir la profundidad del árbol, es decir, el número de nodos totales. En aquellos casos en donde el árbol tenga muchas variables, existe un punto donde añadir nodos no aumenta significativamente la precisión explicativa del modelo.

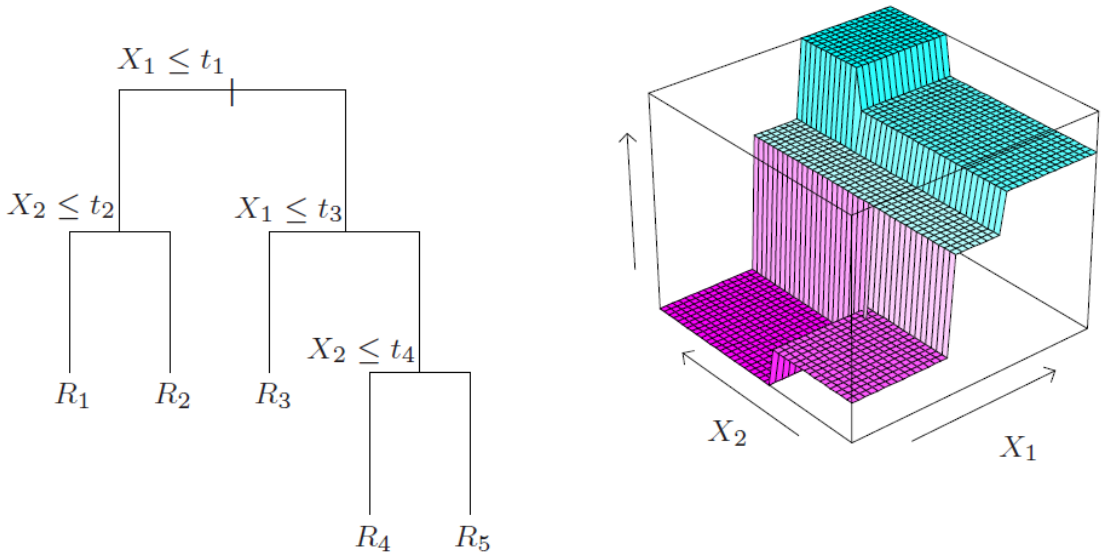


Ilustración 2. A la izquierda una representación jerárquica de un ACR. A la derecha una representación en tres dimensiones del árbol a su izquierda. (Hastie *et al*., 2009)

Puede denotarse de la ilustración derecha, que los árboles tienen la limitación de perder el suavizado esperado de la función subyacente del modelo. Esto implica que la función subyacente se basa en cortes lineales altamente influenciados por los datos de entrenamiento, siendo de cierta manera una función “poco flexible” en el sentido de que crear un árbol con una muestra o con otra (siendo de la misma población) puede cambiar drásticamente el árbol generado.

Por norma general, un árbol de clasificación utiliza como medida de impureza del nodo, el índice de Gini como criterio seleccionador para subdividir la muestra. La impureza nodo medida como el índice de Gini es una medición probabilística de clasificar aleatoria y erróneamente una observación aleatoria en el conjunto de categorías. Esto es expresado formalmente de la siguiente forma en base a la literatura (Hartie, Tibshirani y Friedman, 2009):

Sea un árbol de clasificación donde la variable objetivo es categórica pudiendo tomar números enteros de 1 a *k*, sea a su vez un nodo *m* perteneciente a la región R_m (entiéndase región como el subconjunto debajo de un nodo patriarcal) con N_m observaciones, entonces la proporción de observaciones de clase *k* en el nodo *m* se define como:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum x_i \in R_m I(y_i = k)$$

(Ecuación 1)

Donde entonces la impuridad del nodo mediante el índice de Gini será el peso ponderado del número de observaciones en cada nodo hacia la muestra de la izquierda $N_{m,izquierda}$ y él número de observaciones en la muestra de la derecha del nodo $N_{m,derecha}$, del siguiente índice:

$$\sum k \neq k' \hat{P}_{mk} \hat{P}_{mk'} = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk})$$

(Ecuación 2)

La igualdad es la misma forma de expresar el índice entendiéndose como la ratio de error en el nodo de clasificarse en la categoría *k* con probabilidad $\hat{p}_{mk}$. De la misma forma, se puede concebir binariamente si en cada nodo se cuenta como igual a 1 si se clasifica en *k* o cero en caso negativo, entiéndase por ende la igualdad de la derecha de la ecuación 2.

Por otro lado, en árboles de regresión el criterio seleccionador de los nodos suele ser aquel que minimiza la suma cuadrada de los errores $\sum (y_i - f(x_i))^2$ en cada nodo, por ende, el punto que mejor separa una submuestra es aquel que para una variable *s* y un punto de división *r* resuelva la siguiente ecuación considerando las regiones R_{α_1} y R_{α_2} como los dos subconjuntos creados debajo del nodo:

$$min_{s,r}[min_{\alpha_1} \sum_{x_i \in R_{\alpha_1}(s,r)} (y_i - \alpha_1)^2 + min_{\alpha_2} \sum_{x_i \in R_{\alpha_2}(s,r)} (y_i - \alpha_2)^2]$$

(Ecuación 3)

La ecuación anterior se puede simplificar partiendo de que para minimizar la suma cuadrática de los errores para una región lo más “céntrico” es lo menos separado de todas las observaciones en su conjunto, por lo que para α_1 y α_2 se cumple que el óptimo local es la media de y_i de su región R_α (subconjunto). Formalmente, para *q* =1 o 2, una variable *s* y un punto *r*:

$$\hat{\alpha}_q = \text{promedio}(y_i | x_i \in R_{\alpha q}(s,r))$$

(Ecuación 4)

Nótese que esto no es el mínimo global de la suma cuadrada de los errores, sino una secuencia de búsqueda de mínimos locales por cada nodo, que en conjunto inquieren llegar a un mínimo global. Esto se denomina como algoritmo voraz, conocido también como *greedy algorithm* en inglés. Esta estrategia es una desventaja de los ACR en términos metodológicos ya que no implica que se alcance el óptimo global siguiendo

esté método, no obstante, se utiliza porque es computacionalmente mucho más asequible.

Por último y no menos importante, la capacidad de ajuste a los datos nace de la profundidad que debe alcanzar cada árbol, por ende, es esencial que el tamaño del árbol no sea demasiado grande para no sobre-ajustarse a la muestra de entrenamiento, así como que sea suficientemente extenso para capturar un efecto explicativo. La profundidad del árbol se mide a través el número de nodos. Sería sensato añadir un límite para parar el algoritmo de incluir más nodos, no obstante, en casos particulares puede ser imprudente ya que una variable que no indique un factor con capacidad explicativa superior al límite puede omitirse dejando quizás un subconjunto de observaciones con una clasificación única o un siguiente nodo altamente explicativo.

Para evitar el sobreajuste se intenta elegir un número óptimo de nodos de adelante hacia atrás, es decir, se parte de un árbol en profundidad hasta un máximo fijado como el tamaño mínimo del nodo, al que se le aplica el podado del coste de complejidad (en inglés entendido como *cost complexity prunning*) definido de la siguiente manera (Hastie *et al*, 2009):

Sea un árbol en profundidad T_0 donde $T \subset T_0$, *m* el índice del nodo y T_n el número de nodos terminales en T, sea el "*podado del coste de complejidad"*:

$$CC(T) = \sum_{m=1}^{T_n} N_m Criterio_m(T) + \theta T_n$$

(Ecuación 5)

Donde N_m es el número de observaciones en la ramificación debajo del nodo *m*, y el $Criterio_m(T)$ indica el criterio de selección empleado para dividir los nodos. Véase que $\theta = 0$ juega el papel de un árbol sin podado, mientras que $\theta > 0$ indica un grado de recorte de nodos.

El parámetro θ es el punto clave que minimiza el coste de complejidad, donde para buscar el óptimo se sigue el criterio que para un θ determinado, existe una ramificación que minimiza $CC(T)$, entonces para encontrar dicha ramificación se podan aquellos nodos que incrementen en menor medida el $CC(T)$ reiteradamente condicionado a topar un árbol con un único nodo. Consecuentemente, se puede estimar un $\hat{\theta}$ que

minimice el *coste de complejidad* (Véase Breiman, L., Friedman, J., Stone, C.J. y Olshen, R.A., 1984)[3].

Las ventajas de los árboles de aprendizaje son que la visulazación es muy intuitiva, permitiendoles transmitir con facilidad estructuras complejas de los datos a individuos no expertos en modelos predictivos o con poco conocimiento técnico. Además, una gran ventaja es que no necesitan generar variables *dummy* para cada categoría de una variable factorial como ocurre en otros modelos lineales.

Por otro lado, no todo modelo es perfecto, en el caso de los árboles de aprendizaje si bien son sencillos de interpretar también son mucho menos precisos como modelos predictivos comparado con otras técnicas relativamente sencillas como GLM o MARS (acrónimo para "modelo lineal generalizado" y "*splines* de regresión adaptativa multivariante" respectivamente).

Además otra de las desventajas de los árboles de clasificación regresión es la dificultad de modelar estructuras aditivas, es decir, entiéndase una regresión del tipo $Y = \beta_1 I(X_1 > \sigma_1) + \beta_2 I(X_2 > \sigma_2) + \beta_3 I(X_3 > \sigma_3) + \varepsilon$ donde $I(X_t > \sigma_t)$ es el operador identidad que toma el valor 1 si la condición al interior del paréntesis se cumple o cero en caso contrario; β_t es el coeficiente del punto *t*, ε es ruido blanco y σ_t es el umbral que diferencia patrones. Un ACR puede con sufucientes datos llegar a una estructura similar concatenada, sin embargo su esctructura binomial de generar dos submuestras en cada nodo puede no permitir ver la estructura intrínseca del modelo si se tienen muchos efectos aditivos.

Asimismo, otra de las grandes limitaciones de los ACR es la alta varianza del modelo debido a la estructura jerárquica que poseen. Nótese que cambios en los nodos superiores repercuten directamente en las ramificaciones inferiores, por ende el modelo posee intrínsecamente una alta varianza estructural. Esta no robustez de los árboles puede reducirse mediante la aplicación de técnicas como el *bootstrapping* o *bagging* expuestos más adelante en este trabajo y desarrollados teóricamente en el siguiente apartado.

3.2. Random Forest

Bosque aleatorio (*Random forest* del inglés), es también un algoritmo típicamente utilizado en *machine learning*, donde partiendo de la creación de árboles de clasificación

[3] Breiman, L., Friedman, J., Stone, C.J. y Olshen, R.A., (1984). *Classification and Regression Trees*. Nueva York: Chapman & Hall.

o regresión que analizan una variable numérica o categórica se realiza un *boostrap* (remuestreo) con reemplazo sobre la muestra de entrenamiento, consiguiendo una lotería de árboles de decisión que 'votan' en mayoría o promedian una decisión final para cada observación.

Random forest al igual que los árboles ACR se dividen en bosques de clasificación o de regresión, si el estudio se centra en una variable categórica, donde por tanto el criterio de selección es por normal general el índice de impureza de Gini para los árboles de decisión creados, el modelo predictivo final del *random forest* se centra en seleccionar la clasificación 'más votada', es decir, la clasificación modal. Sea $\hat{C}_j(x)$ la clasificación estimada del árbol *J* de la observación *x*, la predicción del *random forest* se define como:

$$\hat{C}_j(x) = moda\left\{\hat{C}_j(x)\right\}_1^J$$

(Ecuación 6)

Por otra parte, si la variable dependiente es numérica continua, el *random forest* genera árboles de regresión donde generalmente se separa la muestra siguiendo el criterio de selección se minimice la suma cuadrada de los errores. Mismamente, el modelo final resulta de promediar la predicción de la observación *x* en la batería de árboles de regresión generados, es decir:

$$Y(x) = \frac{1}{J}\sum_{j=1}^{J} Árbol_j(x)$$

(Ecuación 7)

Los bosques aleatorios, tienen un componente aleatorio, que les permite ser un algoritmo más fiable que los árboles de decisión o regresión en cuanto a modelo predictivo, esto es el *boostrap aggregating*, o comúnmente conocido como *bagging* (traducido como empaquetado). El objetivo del empaquetado es trazar una media de modelos ruidosos reduciendo su varianza, donde generalmente funciona mejor para modelos con bajo sesgo y alta varianza como los árboles. Esta técnica se basa en promediar una batería de árboles generados aleatoriamente incorrelados entre sí, que en conjunto formen un modelo final.

Para ello, genera *m* muestras de tamaños aleatorios no mayores a los datos de entrenamiento, sobre los que se construye el modelo, en este caso un árbol del bosque aleatorio. Utilizando un bucle que escoja *p* variables aleatoriamente donde de cada *m* y *p*, elige el mejor criterio de selección ("*split criteria*" o "*split-point*" en inglés) para cortar

el árbol en cada nodo. Para que el árbol no genere la profundidad máxima en cada submuestra *m*, se limita el bucle a terminar cuando se alcance un número mínimo de nodos N_{min}. El número mínimo de nodos se optimiza escogiendo aquel que agregando mayor profundidad no produce mejora en el error medio del modelo; más adelante se detalla como evaluar un *ranfom forest* predictivo.

Finalmente, una vez repetido el proceso sobre las *m* submuestras aleatorias, se combinan todas por votación mayoritaria o la moda de cada 'hoja' (en el caso de clasificación), o promediando los *m* árboles (en caso de regresión). Nótese que un *random forest* con un *bagging* de *m* = 1, es decir, de una muestra, es simplemente un árbol de clasificación o regresión con dicha submuestra *m*, donde no hay interés en su aplicación (es una muestra aleatoria más pequeña que generar un ACR sobre la muestra total). Sin embargo, un *random forest* con una selección de *m* submuestras muy elevada puede implicar un ejercicio computacional muy costoso ganando poco ajuste en el modelo predictivo. Por ende, la selección óptima del número de submuestras se estima evaluando el ajuste ganado al introducir número de versus el coste computacional de generar *m* árboles.

Esta técnica de *bagging* propuesta en 1994 por Leo Breiman[4] permite reducir la varianza del modelo predictivo, así como ayudar a evitar el sobreajuste (*overfitting* en inglés) típico de los árboles de decisión a los datos de entrenamiento, nublando la capacidad predictiva del modelo. Debido a que en un bosque aleatorio el *bagging* (o "remuestreo con reemplazo") deja variables y observaciones fuera de la selección aleatoria (al igual que también otras se repiten), permite que se realicen muchos árboles incorrelacionados entre sí que no contemplen todas las opciones en cuanto a variables y observaciones generando un mapeo de resultados alrededor de una media común (modelo final). De esta manera, al obtener el modelo final, que es una media o moda de todos los modelos, no es posible que la predicción se ajuste a un entrenamiento al ser siempre una diversidad de árboles con entrenamientos diferentes, aunque considerados de una misma población, por lo que el modelo predictivo final no percibe sobreajuste a un entrenamiento concreto.

Una de las medidas más relevantes de los bosques aleatorios para entender la estructura del modelo es analizar la importancia final que han tenido las variables sobre el output de todos los árboles. La importancia de las variables se mide generalmente con el criterio de mejora o criterio de selección que cada variable ha obtenido en cada

[4] Breiman, Leo (1994). Bagging predictors. *Technical report 421,* Department of Statistics, University of California at Berkeley.

nodo de aquellos árboles en donde haya tenido un efecto relevante. La suma total del criterio de selección generado para cada variable en cada nodo de cada árbol se acumula para medir en promedio la importancia de una variable en el conjunto de árboles del bosque aleatorio. Asimismo, una ventaja del bosque aleatorio es que todas las variables estiman un criterio que mide su importancia en todos los subconjuntos de nodos y árboles, de manera que es más probable en un bosque aleatorio que se analice e incluyan todas las variables en el modelo que otros algoritmos como el *gradient boosting* o regresiones lineales.

Una vez obtenido el bosque aleatorio, una forma vastamente utilizada para medir el desempeño del modelo predictivo es utilizando la muestra de observaciones *Out-Of-Bag* (del inglés, "fuera de la bolsa" referido a *bagging* o *bootstrap*) para estimar, valga la redundancia, el error *Out-Of.Bag*. La muestra *Out-of-Bag* o comúnmente denominado OOB, es aquella muestra que no se ha seleccionado en la submuestra *m* para el árbol del mismo sub-índice. Por ende, el árbol calculado nunca se ha entrenado con la muestra OOB, siendo entonces un buen estimador del desempeño del árbol *m*.

Esto es comparable a realizar una validación cruzada sobre el árbol *m*-ésimo del bosque con datos de control. Realizando este proceso para los *m* árboles aleatorios del bosque, entonces, el error OOB es la media del error predictivo generado por cada muestra OOB de cada árbol del bosque. En términos generales, es equiparable a realizar una validación cruzada de N-pliegues con los datos de entrenamiento ("*N-fold cross validation*" en inglés) o visto de otro modo, para un *m* grande, es equiparable al método de validación cruzada 'deja-uno-fuera'.

Si bien, el error OOB permite estimar un error de predicción, evitando tener que utilizar una muestra y validación independiente, suele subestimar el desempeño actual por lo que comúnmente se utiliza para escoger el hiperparámetro del número de árboles a introducir en el *random forest*. Es decir, se analiza la evolución del error OOB según se incrementa el número de árboles en el bosque aleatorio, seguidamente, aquel punto donde el error OOB deja de decrecer en unas cuantas interacciones, se recoge el número de árboles óptimo. Véase la siguiente ilustración, donde se muestra la evolución del error OOB frente al error de predicción de una validación cruzada con una muestra de testeo, todo ello sobre el incremento del número de árboles estimados en el *bagging* de un *random forest*.

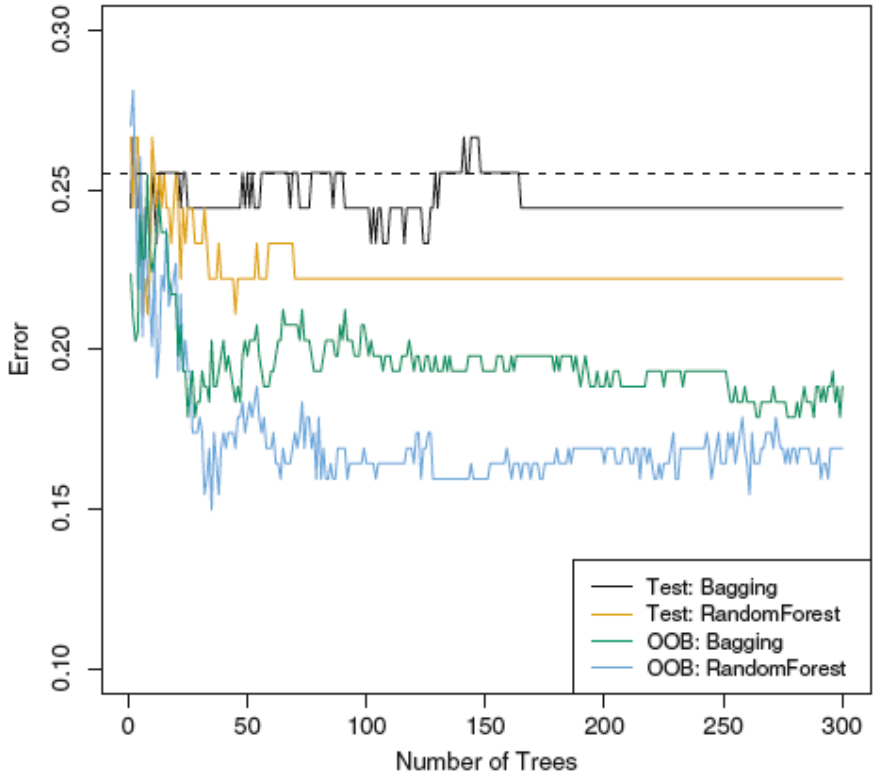


Ilustración 3. Evolución del error OOB en *bagging* y *random forest* versus el error de predicción de una validación cruzada con una muestra de control entorno al número de árboles generados (James, G. *et al*, 2013).

Nótese que el error OOB subestima el desempeño del modelo, por lo que es crucial realizar una validación cruzada porque si no se estaría asignando un ajuste de bondad no real al modelo, pudiendo predecir pobremente nuevas observaciones. Del mismo modo, la validación cruzada permite medir eficientemente el desempeño con una muestra de control no entrenada en el modelo.

4. Modelo económico

A continuación, se detalla la estructuración modelo económico o modelo predictivo del presente trabajo, escrudiñando la fuente de información escogida, el tratamiento de la muestra de datos, la aplicación real del modelo expuesto y, finalmente, la validación de las estimaciones y resultados obtenidos en el mismo para determinar su precisión en las conclusiones extraídas.

Cabe destacar que en este trabajo se ha utilizado RStudio [5] como herramienta estadístico-matemática para el tratamiento y modelización predictiva, y Hojas de Cálculo

[5] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

de Microsoft Excel como soporte de formato y fusión de los ficheros originales de datos públicos.

Los datos recabados para este estudio son microdatos de carácter público provenientes de la web de la Dirección General de Tráfico[6] (en adelante "DGT") referentes a los informes policiales recogidos sobre accidentes de tráfico en el año 2015 en todo el ámbito del territorio español. Los microdatos se publican anualmente en tres hojas de cálculo "*TABLA_ACCVIT_2015.csv*", "*TABLA_PERS_2015.csv*" y "*TABLA_VEHIC_2015.csv*", referentes a datos de víctimas, personas y vehículos respectivamente. Se recoge dicha información de formularios censales o mediante enumeración completa con un nivel mínimo de segregación municipal o inferior. Cada unidad (expresado en filas) se refiere a un accidente de tráfico con víctimas. El objetivo de esta información público-estadística es la de obtener un conocimiento de los accidentes de tráfico con víctimas atendiendo a las circunstancias del evento y, asimismo, a sus respectivas consecuencias, es decir, a las víctimas de los mismos.

Aunque se expresen numéricamente, la mayoría de las variables son de categóricas. La referencia a cada uno de los valores para estas variables se expresa en diccionarios en hojas de cálculo adjuntos a una carpeta comprimida referenciada en "*Diseño de registro año 2011 y anualidades posteriores*" de la misma fuente de la DGT.

A continuación, se explaya la recolección, filtro y depuración de la base de datos para conseguir una muestra en concordancia con los formatos adecuados para la ejecución del *random forest* y los árboles de aprendizaje.

4.1. Tratamiento

El fichero de víctimas recoge datos de todas aquellas personas que sufrieran una lesión leve, grave o el fallecimiento y las condiciones del accidente (tipo de carretera, intersección o curva, prioridad o señalización, visibilidad, tiempo y localización, etc.). El fichero de personas recoge tanto viandantes como conductores involucrados en el accidente, así como terceros ocupantes del vehículo entiéndase copiloto o pasajeros en asientos traseros. Y el fichero vehículos recoge la información relativa a los vehículos involucrados en el accidente como anomalías, tipo de vehículo y año de la matrícula. La tabla 1 a continuación, muestra una breve descripción de la base de datos recogida, donde cada fichero contiene un número diferente de observaciones debido a que cada fichero recoge diferentes datos relativos al mismo accidente.

[6] Dirección General de Tráfico. Microdatos de Accidentes: 2015. [Dataset]. Recuperado de https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces

Tabla 1. Dataset Accidentes 2015

Ficheros	Observaciones	Variables
Víctimas "*TABLA_ACCVIT_2015.csv*"	97.756	39
Personas "TABLA_PERS_2015.csv"	238.476	31
Vehículos "TABLA_VEHIC_2015.csv"	170.749	14

Fuente: Accidentes con víctimas 2015. Dirección General de Tráfico.

Para tratar la base de datos se puede trazar la relación entre ficheros mediante un identificador único "ID_ACCIDENTE" que interrelaciona los tres ficheros. Además, un segundo identificador "ID_VEHICULO" conecta al fichero de Personas con el de Vehículos. Sin embargo, para este trabajo, solamente se ha recabado el "ID_ACCIDENTE" para crear una muestra que a partir de Víctimas recoja una única observación relativa de Personas y de Vehículos. Resultando en una muestra de 97.756 observaciones con 84 variables cada una.

De esta forma se recogen solamente aquellas variables explicativas y referentes al accidente, al vehículo y a la víctima, en donde generalmente es el conductor del vehículo. En aquellos vehículos con accidentes con más de un pasajero por vehículo se han omitido las observaciones e información relativa a las víctimas acompañantes, debido a que se generaría un espacio en blanco referente a los datos del accidente o vehículo. Asimismo, tampoco es prudente agregar a las víctimas acompañantes y repetir los factores del vehículo y de la carretera; ya que entonces el modelo entendería que es otra observación independiente, cuando no lo es, sesgando el modelo predictivo ya que la interdependencia de variables sobreestima valores, en este caso, los criterios de selección y el error OOB.

Una vez los datos se encuentren fusionados por su número identificativo, en un solo fichero en soporte Microsoft Excel, se importan en RStudio utilizando el paquete *readxl* como un fichero de datos (*data frame* en inglés). La organización del fichero de datos en R se basa en el criterio general de establecer las observaciones en filas y las variables en columnas, manteniendo una clara estructura para la visualización de la tabla de datos.

Seguidamente se recortan variables de la muestra de datos siguiendo los siguientes criterios: (a) Si la variable explicativa se repite en otra columna de los datos. (b) Si la variable no tiene un potencial poder explicativo en el modelo predictivo. (c) Si la variable es categórica y además posee más de 53 categorías (limitación de la herramienta y el paquete de cálculo). Variables repetidas como el número de identificación del accidente

se omiten al igual que el año de ocurrencia porque se reitera y no merece de atención en el análisis.

Asimismo, variables como el Id de la persona o del vehículo, o bien el año (reiteradamente que toda la muestra es del 2015), no aportan valor al modelo de predicción por lo que descartarlos es lo más prudente. Una limitación del paquete utilizado con RStudio es que no permite modelar un bosque aleatorio con variables categóricas con más de 53 factores, lo que es sensato puesto que los criterios de selección de dicha variable obligan a dispersar la muestra de datos enormemente.

Ha sido necesario realizar una serie de ajustes de formato antes de construir el fichero de datos final en RStudio para poder conseguir modelizar el bosque aleatorio. En el anexo de este trabajo se incluye la totalidad del código en RStudio utilizado y se adjunta además la base de datos fusionada. Esta base de datos recoge accidentes con víctimas únicamente, incluso si solamente se ha perjudicado una lesión leve a una persona, es decir, sin necesidad hospitalaria. Por otra parte, si no ha habido ningún tipo de lesión o víctima por ninguna de las partes involucradas, ni terceras, no se registra la observación.

La variable por predecir en el modelo es la severidad de las lesiones en los accidentados, para lo que se necesita construir una variable que recoja la totalidad de categorías. Los datos recogen la variable explicativa segregada en tres variantes en formato de contador numérico de la cantidad de lesiones sufridas por cada accidente registrado. Por ejemplo, la variable "*TOT_HERIDOS_GRAVES*" recoge el número total de heridos graves para cada registro con un identificador "".

Sin embargo, como la muestra seleccionada solo recoge información para el primer y principal individuo del accidente (generalmente el conductor), para un accidente con 5 heridos graves, la información relativa a la persona sobre el año del permiso de conducir o la información relativa al vehículo ocupado se encuentra en la misma observación, sin repetirse ningún factor en otra variable u observación.

Tabla 2. Estructura de variables dependientes originales

Víctimas	Heridos Leves	Heridos Graves	Fallecidos
TOT_VICTIMAS	*TOT_HERIDOS_LEVES*	*TOT_HERIDOS_GRAVES*	*TOT_MUERTOS*
TOT_VICTIMAS30D	*TOT_HERIDOS_LEVES30D*	*TOT_HERIDOS_GRAVES30D*	*TOT_MUERTOS30D*
	HERIDO_LEVE_24H	*HERIDO_GRAVE_24H*	*MUERTO_24H*
	HERIDO_LEVE30D	*HERIDO_GRAVE30D*	*MUERTO_30D*

Fuente: Microdatos de Accidentes. Dirección General de Tráfico. Elaboración propia

Para poder modelizar el bosque aleatorio es necesario disponer de la variable dependiente como un único factor explicado del modelo, para lo que se construye una

variable explicada denominada por comodidad “Y”; donde se recoge para cada observación si la severidad sufrida (no necesariamente el número de heridos o fallecidos) es leve o grave.

Para el caso de los fallecidos, como representan solamente el 1,4% del total de la muestra se opta por realizar un estudio independiente sobre la probabilidad de supervivencia con modelo lineales generalizados. Esta división del estudio se debe a que la proporción baja de observaciones afecte el modelo predictivo del *random forest*, además de que los estudios de variables con muy pocas observaciones con la categoría de interés muestran un buen ajuste al ser modelizados por estos algoritmos. Por ende, el enfoque de este trabajo se desarrolla en la creación del *random forest* para estudiar la severidad de los no fallecidos, y por otra parte se estudia un modelo para predecir la categoría de fallecidos en accidentes de tráfico.

La creación de la variable se basa en un bucle que recorre todas las observaciones de la muestra asignando una categoría (por simplicidad “Leve” o “Grave”) según si la existencia de heridos es mayor a cero en sentido decreciente de su intensidad, es decir, del más grave al más leve. De esta manera se les da énfasis a aquellas víctimas con lesiones más graves, antes que, a los leves, en un accidente con múltiples heridos.

La lógica de esto subyace en que se considera que, por ejemplo, en un accidente con 4 víctimas donde dos fallecen, uno es herido grave y el cuarto es herido leve; dado que solamente se asigna una categoría de severidad para una observación, la clasificación más prudente es asignar la categoría de la peor lesión sufrida puesto que ha sido determinante para al menos una víctima los factores externos expuestos, siendo el objetivo de este trabajo.

La siguiente tabla muestra las variables independientes del modelo, utilizadas para estimar la severidad de las lesiones. Mayoritariamente son variables categóricas debido a que se clasifican objetivamente las circunstancias en factores determinísticos como el tipo de vía o el uso obligatorio de casco en vehículos a dos ruedas.

Tabla 3. Variables independientes, tipo y referencia de medida

#	Variable	Tipo	Referencia
1	Severidad (“Y”)	Factor	Leve o Grave
2	Mes	Numérica entera	Enero, febrero, marzo, …
3	Hora	Numérico	…, 12, 13, 14, 15, 16, 17…
4	Dia de la semana	Factor	Lunes, martes, miércoles...
5	Provincia	Factor	Araba, Albacete, Alicante, Almería, …
6	Vehículos implicados	Numérica entera	1, 2, 3, 4 …
7	Zona	Factor	Carretera, zona urbana, travesía …
8	Zona agrupada	Factor	Vías interurbanas o vías urbanas

9	Carretera	Factor	Titularidad estatal, autonómica, …
10	Tipo de vía	Factor	Autopista, autovía, vía para automóviles …
11	Trazado de la vía	Factor	Intersección, recta, curva suave …
12	Tipo de intersección	Factor	En t ó y, en x ó +, enlace de entrada …
13	Prioridad**	Factor	Agente, ceda, semáforo, stop, …
14	Superficie de la calzada	Factor	Seca y limpia, umbría, mojada, helada …
15	Nivel de luminosidad del día	Factor	Pleno día, noche: iluminación suficiente …
16	Factores atmosféricos	Factor	Buen tiempo, niebla intensa, niebla ligera …
17	Visibilidad restringida	Factor	Sin dato, edificios, configuración del terreno …
18	Tipo de accidente	Factor	Colisión de vehículos en marcha (Frontal), colisión en marcha (Frontolateral) …
19	Edad	Numérica	0, 1, 2, 3, 4, 5…
20	Sexo	Factor	Mujer o Hombre
21	Año del permiso de conducir	Numérica entera	…, 73, 74, 75, 76, 77…
22	Posición de la víctima	Factor	Peatones, conductor, pasajero delantero …
23	Seguridad**	Factor	Lleva cinturón, lleva casco, SRI, …
24	Maniobra realizada	Factor	Siguiendo la ruta, adelantando por derecha …
25	Infracción de velocidad	Factor	Sin dato, velocidad inadecuada, sobrepasado …
26	Infracción de la conducción	Factor	Conducción distraída, circular sentido prohibido, …
27	Infracción de apertura	Factor	Apertura de puertas sin precaución, …
28	Infracción de alumbrado	Factor	Incorrecta utilización del alumbrado, …
29	Infracción de carga del vehículo	Factor	Incorrecto estacionamiento, …
30	Infracción peatonal	Factor	No respetar señal de semáforo, no utilizar paso de peatones, …
31	Resumen infracciones	Factor	Infracción cometida, …
32	Año de matrícula	Numérica entera	…, 46, 47, 48, 49, 50…
33	Mes de matrícula	Numérica entera	Enero, febrero, marzo, …
34	Tipo de vehículo	Factor	Bicicleta, Ciclomotor, Coche Minusválido, …
35	Anomalía**	Factor	Dirección, frenos, neumáticos, …
36	Número de ocupantes	Numérica entera	1, 2, 3, 4…
37	Mercancías peligrosas	Factor	Con o sin mercancías peligrosas
38	Vehículo incendiado	Factor	Incendiado o no

Fuente: Microdatos de Accidentes. Dirección General de Tráfico. Elaboración propia.

La variable “Prioridad” es una aglomeración compuesta de siete variables *dummy* originales en el modelo. La razón de transformar estas variables en una sola se debe a que en distintas variables se está repitiendo el mismo efecto. Por ejemplo, en la variable binaria de si en la ubicación del accidente, existía una prioridad de ceda el paso (posible factor de los hechos) cuando esta variable toma el valor uno es porque hay una señal de ceda el paso, pero en caso contrario, un cero indica que no existe dicha señal de ceda el paso. Ahora bien, si por consiguiente la variable “*PRIORIDAD_STOP*” también se incluye en el modelo y también resulta que es cero, es decir que no hay una señalización de “STOP”, entonces se está indicando doblemente lo mismo en dos variables diferente. Comúnmente se denomina a esto multicolinealidad, es decir, las correlaciones entre variables al explicar en el modelo indican el mismo efecto, provocando un sesgo importante en el modelo que se debe evitar.

De igual manera, la variable “SEGURIDAD” es una agrupación de “USO_CINTURON”; “USO_CASCO” y “USO_SRI”, referentes al uso del cinturón de seguridad del vehículo, el uso del casco en ciclomotores y el sistema de retención infantil para menores. Como el cinturón de seguridad es el accesorio más común y el tratamiento incluye mayoritariamente a conductores de vehículos a cuatro ruedas como “Cinturón” o “No_cinturón” este es el factor principal discriminante para la mayoría de las observaciones.

Seguidamente el uso del sistema de retención infantil indica casos muy concretos para mortalidad infantil, donde se clasifican como “SRI” aquellos vehículos que lo dispongan para los ocupantes infantiles. El uso de casco por otra parte contrasta el uso de casco o la falta del mismo, por lo que se señalizan “Casco” para los casos que lleven puesto el protector y “No_Casco” para los casos contrarios. Asimismo, aquellos casos en los que no se disponga de dato, se entabla como “NA/NS”.

En este caso como las variables son categóricas mayoritariamente no se puede contrastar una matriz de correlaciones, por lo que la lógica común de que no se repitan efectos circunstanciales o explicativos es suficiente. El efecto de la multicolinealidad en los *random forest* no implica un hecho determinante en la modelización del modelo como podría ser en una regresión lineal debido a que el árbol en factores binarios como ser hombre o mujer (incluyendo los dos en el modelo) no podría incluir los dos parámetros o una submuestra quedaría en cero. Sin embargo, variables que indiquen los mismos puede dar lugar a sesgar en la clasificación de variables por los mismos parámetros.

La variable “Prioridad” por ende es el resultado de ejecutar un bucle que, al igual que la variable dependiente “Y”, recorre todas y cada una de las observaciones asignando los niveles de “Agente”, “Ceda”, “Marcas”, “Otra”, “Paso”, “Semáforo”, “Stop” o “Ninguna” en los casos respectivos donde exista una prioridad en la vía circulatoria entorno al evento del accidente.

La variable “Anomalía” es también una variable creada artificialmente proveniente de cinco variables dicotómicas originales “anomalía en la dirección”, “anomalía en los neumáticos”, “anomalía en los frenos”, “anomalía de reventón”, o “ninguna anomalía”.

Como se menciona en la sección metodológica la independencia de las variables es esencial para conseguir estimadores insesgados. Si bien, el *bagging* al realizar submuestras aleatorias elimina gran parte de la correlación entre variables y árboles, si una misma variable dicotómica como las anomalías originales describen reiteradamente escenarios idénticos que genera multicolinealidad entre variables y por tanto entre árboles dentro del *random forest*.

4.2. Análisis estadístico-descriptivo

Se pueden conocer ciertos patrones analíticos de los datos con simple estadística descriptiva de los mismos, por ejemplo, el 73% de las víctimas son mujeres, de las cuales el 52% conducían el automóvil. Esto concuerda con que el 67% de los accidentados estaban conduciendo un vehículo a cuatro ruedas en el momento del accidente, en donde el 1,5% de los conductores accidentados de vehículos a cuatro ruedas ha fallecido en el 2015.

En cuanto a los factores externos del accidente, el principal factor es la estructura de la vía en la que ocurre el evento fortuito recogido en "*TIPO_VIA*" y en "*TRAZADO_NO_INTERSECC*". Estas variables describen de forma general la carretera, su clasificación y el tramo en donde ocurre el accidente: ya sea la curva de una autovía o en la recta de una autopista.

Para el tipo de trazado el cual indica la estructura del tramo del accidente, a continuación, se encuentra un desglose por la severidad sufrida por las víctimas en leves y fallecidos, debido a que los grave representan una estructura proporcionalmente equilibrada entre los trazados de la vía circulatoria de leves y fallecidos.

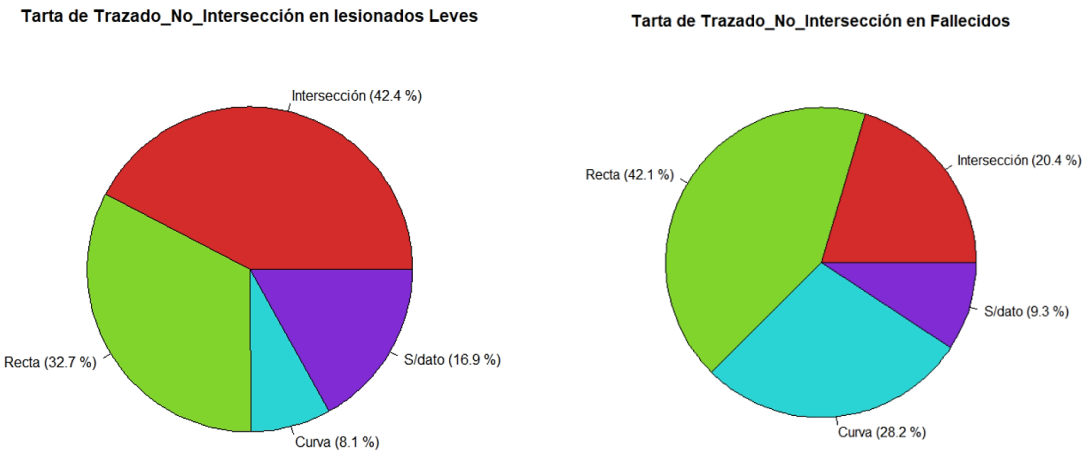


Ilustración 4. Gráfico circular de la proporción de accidentes con víctimas con lesiones leves (izquierda) y fallecidos (derecha) sobre el tipo de trazado de la vía. Elaboración propia.

Se observa que mayoritariamente los accidentes leves ocurren en intersecciones, de donde se deduce que las intersecciones se encuentran mayoritariamente dentro de núcleos urbanos. De ello también se deduce que la velocidad promedio del vehículo dentro de los núcleos urbanos es más reducida y por ende exhibiendo lesiones de menor impacto por colisiones, salidas de vía o atropellos.

Asimismo, la mayoría de los accidentes con fallecidos (42,1%) sucede en rectas, esto es mayoritariamente en carreteras de velocidades elevadas (por ejemplo, autopistas), debido a que la colisión de vehículo-vehículo o vehículo-objeto presenta un impacto de gran escala aumentando considerablemente las probabilidades de lesiones graves o incluso la muerte.

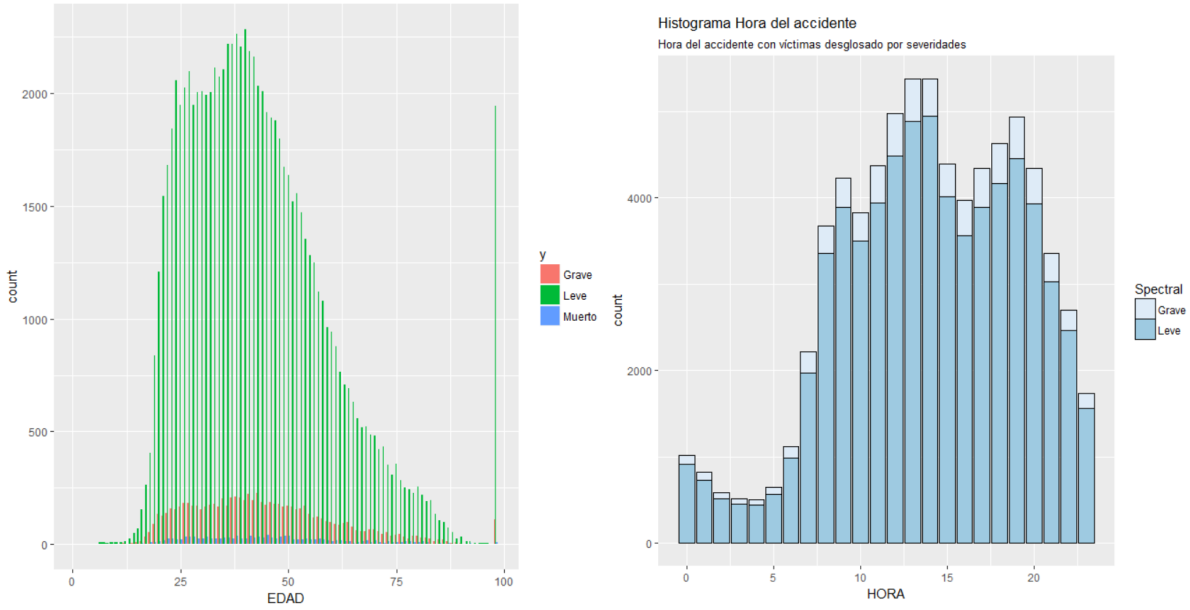


Ilustración 5. Gráfico de barras (izquierda) de la dispersión de edades entre las víctimas desglosado por niveles de severidad. Histograma (derecha) de la distribución de la hora del accidente de tráfico con víctimas desglosado por niveles de severidad.

Se denota que la edad se acumula alrededor de los 25 a los 55, con una media de 40 años aproximadamente para los tres niveles de severidad, no denotando una diferencia significativa entre severidades. Sin embargo, en promedio se describe que la edad promedio de las víctimas también describe la edad promedio del mayor uso del vehículo como transporte principal.

Si bien en edades recientes, el crecimiento de víctimas de tráfico es exponencial se debe al crecimiento en proporción del uso del vehículo, mientras que según se acerca la tercera edad el decrecimiento es suave acorde a ganar experiencia en la conducción y la reducción de la población conductora por el envejecimiento o la falta de necesidad del uso del vehículo.

Por otro lado, para edades inferiores a 18 años para lo que hay pocos casos, significan víctimas menores de edad, que no conductores, pero peatones o pasajeros acompañantes. El pico de edad denotado a los 99 años es un marcaje de aquellas personas que de las que no se dispone dato.

4.3. Random Forest

Esta sección detalla en profundidad la aplicación del modelo predictivo, su optimización y la validación final del algoritmo. Uno de los objetivos del trabajo es utilizar técnicas de *machine learning* para aprender de los datos y determinar un modelo que pueda explicar la severidad en los heridos de accidentes de tráfico. Para ello se opta por aplicar un bosque aleatorio debido a su sencilla interpretabilidad, cálculo y buen ajuste predictivo con bases de datos con gran cantidad de variables categóricas de varios niveles.

El bosque aleatorio como se explica en la sección metodológica se basa en la aplicación recurrente de árboles de clasificación realizando un *bootstrapping* de la muestra de entrenamiento. Si se representa un árbol para una pequeña muestra del modelo obtenemos el siguiente árbol de clasificación:

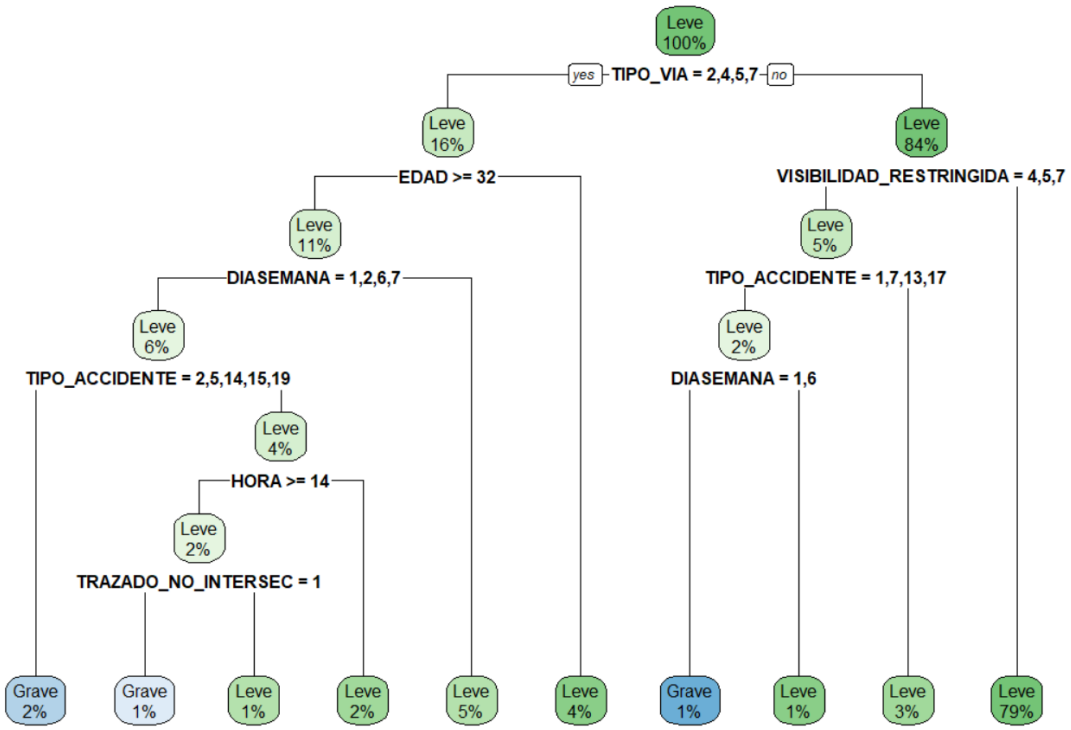


Ilustración 6. Árbol de clasificación para una pequeña muestra de entrenamiento. Véase la tabla de variables para entender el formato y tipo de los nodos expuestos. Elaboración propia.

Se aprecia que el primero nodo selecciona por el tipo de vía en el que ocurre el accidente en este caso por autovía, vía convencional (con y sin carril lento) y vía de servicio, respectivamente. Si el criterio de selección (tipo vía igual a 2, 4, 5 o 7) es verdadero, la ramificación consiguiente se representa la sub-muestra de la izquierda debajo del nodo, es decir, aquella con el nodo de la edad de la víctima.

Si, por el contrario, el tipo de vía no corresponde a los anteriores, el nodo de la derecha es el consiguiente, donde el siguiente criterio de agrupación es el indicador de visibilidad restringida por factores atmosféricos, deslumbramientos y otra causa. Si el criterio del nodo se cumple, le siguen dos criterios más, que si también son verdaderos se clasifican las severidades graves. En otras palabras, esos casos concretos de lesiones graves se categorizan por mayoría en una secuencia de criterios para llegar a esa hoja. Asimismo, si el criterio de selección de la visibilidad restringida no se cumple, ya se está categorizando al 80% de la muestra como leve por ser mayoría con esos criterios.

Cabe destacar que en general las hojas del árbol determinan la causa de los graves donde, en los casos contrarios a la selección del nodo se denotan las lesiones leves. Ello indica que el nivel de clasificación del árbol busca aquellos criterios que expliquen a los graves de los leves por una interacción jerárquica de las variables externas.

Adicionalmente, la representación del árbol permite entender fácilmente los criterios de las clasificaciones y ajustar un modelo a por ejemplo un vehículo con sistema de socorro automatizado. Por ejemplo, una combinación de factores externos condicionantes a sufrir una lesión grave es en un accidente con colisión de vehículos en marcha, un lunes, para una persona de más de 32 años en una autovía (uno de los casos de la primera hoja de graves de la izquierda).

La referencia a cada variable categórica se puede visualizar en el anexo 7.7 Datos de víctimas de accidentes.

Para realizar un seguimiento de la ejecución del modelo se opta por utilizar una validación cruzada de N pliegues de la muestra de datos completa. La validación cruzada de N-pliegues consiste en crear *n* submuestras aleatorias y proporcionales de los datos, en donde en un bucle de *n* periodos independientes se separa la submuestra n-*ésima* de los datos, se entrena el modelo con la proporción $\frac{(n-1)}{n}$ de datos (sin utilizar la submuestra n-*ésima*) y se realiza una validación cruzada con la muestra de control n-*ésima*. Finalmente se puede estimar un modelo final promediando la *n* modelos anteriores. Esta técnica sirve comúnmente para evitar el sobreajuste que puedan sufrir los modelos, al ser una técnica parecida al *bagging*.

Para este modelo, como el bosque aleatorio interiormente realiza un *bootstrap* y validación parecidos, que no es equivalente a hacerlo con el modelo final, se opta por esta técnica para reducir el tiempo computacional de entrenar cada modelo con una cantidad de datos grande. Por lo que, una validación cruzada de 5 particiones

individuales (pliegues) para construir un modelo promedio se considera suficiente en este caso.

Holísticamente, la estructura del modelo se basa entonces en 5 fases donde en cada una se calcula un bosque aleatorio, desarrollado por una lotería de árboles de clasificación; un modelo predictivo resultado de una votación mayoritaria de las clases entre todos los árboles intrínsecos y una evaluación del desempeño y bondad del ajuste como modelo predictivo.

En la siguiente figura se muestra la estructura que se desenvuelve en la totalidad de los cálculos como una visión general estructural. Manteniendo la simbología anterior, *n* indica la validación cruzada de 5 pliegues, donde en cada una se realiza un bosque aleatorio en el que se escoge el hiperparámetro *M* (número de árboles) y se ejecuta un bucle de *m* hasta *M* árboles. Seguidamente se construye el bosque aleatorio y se testean los resultados con la muestra de control n-*ésima*.

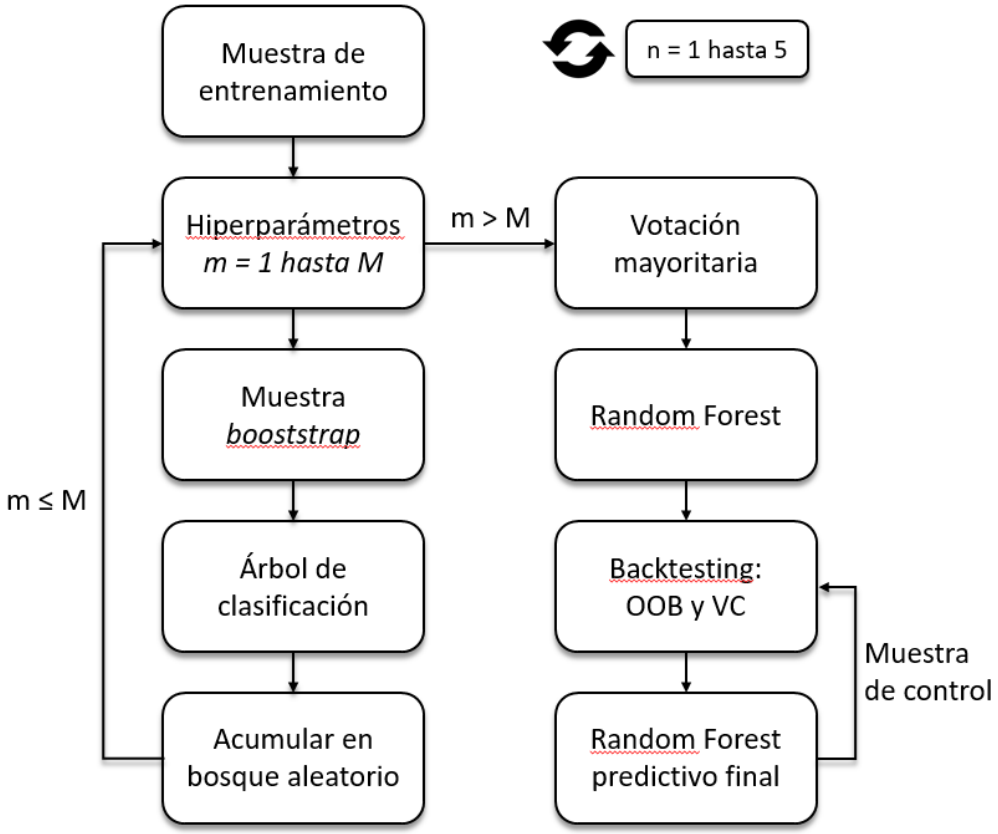


Ilustración 7. Estructura de cálculo del *Random Forest*. De una muestra de datos de entrenamiento, se escogen los hiperparámetros M (número de árboles) y P (número de variables en cada remuestreo booststrap) y se aplica el proceso de cáculo explicado para el bosque aleatorio. Finalmente, se elabora el bosque por la moda de categorías y se testea la precisión del modelo final. Elaboración propia.

4.3.1. El Modelo

El paquete utilizado en la herramienta RStudio para calcular los bosques aleatorios es el desarrollado por Liaw, A. "*randomForest* R package"[7] donde se implementa en código abierto para ejecutar los cálculos en la herramienta estadístico-matemática R.

El modelo consta de una primera tres fases: una primera fase de aplicación estándar donde se ejecuta el modelo con patrones estándar cómo base de modelización del bosque aleatorio. Una segunda etapa donde se parametrizan las variables intrínsecas al bosque aleatorio, entiéndase el criterio de selección, el número de variables utilizadas a la hora de aplicar el empaquetado (*bagging*), el número de árboles óptimo a calcular para cada sub-muestra aleatoria buscando la eficiencia en el cómputo de los datos como en la convergencia estable de los resultados del bosque, entre otros (para todas las parametrizaciones véase el marco teórico del *random forest* en el aparto de Metodologías).

Una última fase ejecuta el modelo con los parámetros óptimos sobre una base de entrenamiento para generar el modelo predictivo final. Se aplica una validación cruzada a la predicción del modelo para entender su veracidad y precisión sobre la muestra real del mismo año, así como una matriz de confusión y la ratio de error *Out-Of-Bag* (véase metodología del *random forest*) sobre todo el modelo estimado y el predicho.

Antes de modelizar el *random forest* se deben escoger los dos hiperparámetros que resultan de optimizar el número de árboles de clasificación que se ejecutan en el bosque aleatorio, y del número de variables predictoras en cada muestra *bagging*. La mejor forma de estimar los parámetros es mediante el análisis de la evolución del error OOB (calculado intrínsecamente en el *random forest*) según se incrementa el número de parámetros.

El número óptimo de árboles a introducir se estimar por prueba y evaluación del error en un bosque aleatorio. Provechosamente el paquete de R de Liaw y Wiener (2002) introduce una función para escanear el seguimiento de *random forest* con un incremento determinado del número de árboles hasta un máximo. La siguiente tabla muestra como utilizando los parámetros predeterminados (número de variables predictores por *bagging* = 6 y número máximo de árboles a testar = 500), la evolución del OOB manifiesta la siguiente estructura para un incremento de 50 en 50 árboles de entrenamiento en el bosque:

[7] A. Liaw and M. Wiener (2002). *Classification and Regression by randomForest*. R News 2(3), 18-22.

Tabla 4. Evolución del número de árboles en el *random forest* y el OOB error

Número de árboles (M)	Error Out-Of-Bag	Lesión Grave	Lesión Leve
50	1,90%	19,53%	0,04%
100	1,88%	19,49%	0,02%
150	1,88%	19,48%	0,02%
200	1,87%	19,48%	0,02%
250	1,87%	19,48%	0,01%
300	1,87%	19,48%	0,01%
350	1,87%	19,49%	0,01%
400	1,87%	19,49%	0,01%
450	1,87%	19,49%	0,01%
500	1,87%	19,49%	0,01%

Fuente: Microdatos Accidentes. DGT. Elaboración propia

Se aprecia en la tabla que un número óptimo de árboles se alcanza con una media de 200 árboles aproximadamente, siendo el punto donde el error OOB general del modelo es más bajo, así como el error condicional de cada una de las categorías de la variable predicha.

Para estimar el número óptimo de variables se evalúa el error OOB en el desempeño general del bosque aleatorio desde 1 variable predictor hasta el máximo número de variables del modelo, en este caso 38 variables predictores. Por simple eficiencia se realizan aumentos cuadráticos del número de variables introducidas en el *bagging*, de manera que la progresión resulta: 1, 2, 4, 8, 16, 32, 38.

Asimismo, si se encuentra un punto en donde el incremento de mejora en el error OOB no supera el 0,1% se opta por parar el cómputo del mismo puesto que el error OOB converge a un punto fijo y no merece la pena calcular los errores posteriores. Asimismo, el número de árboles también debe especificarse y se denota el número de árboles calculado en el paso anterior (m = 200).

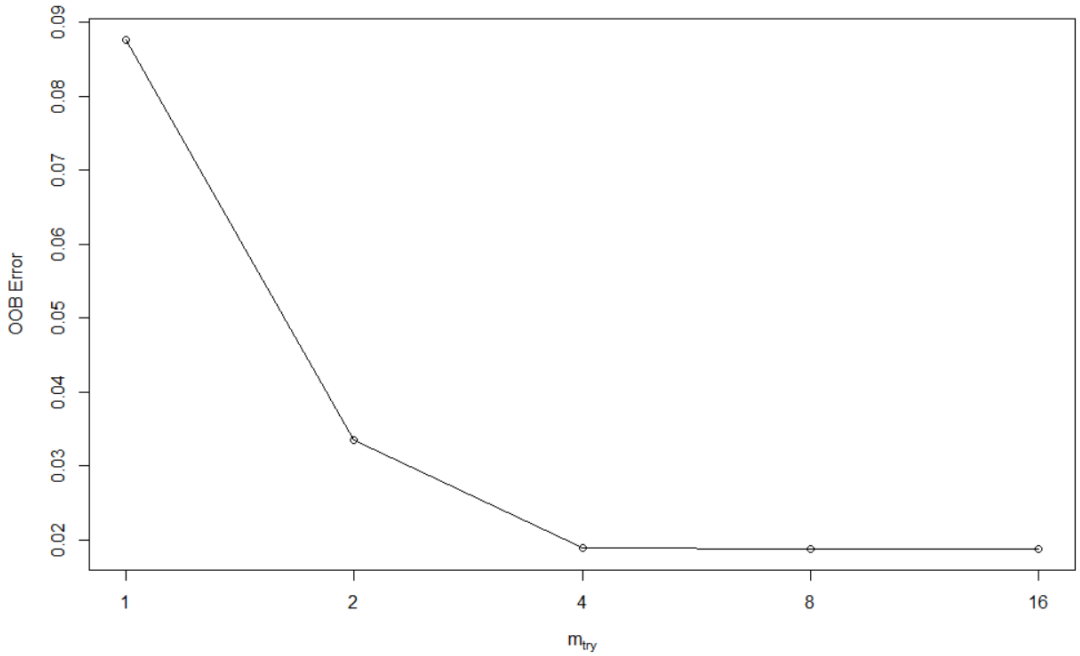


Ilustración 8. Evolución del número de variables introducidas en el *bagging* versus el error OOB medio de la predicción del bosque aleatorio.

De esta ilustración se denota que el óptimo se encuentra en el "codo" de la pendiente, es decir, donde el número de variables predictores es igual a 4. La denotación entre elegir cuatro o dos se debe a que el escoger un número más elevado de variables es más probable de escoger mejores criterios de selección, donde dos es relativamente una baja cantidad. Además, se fija que, para calcular la secuencia de nodos, se establece un parámetro de mejora en el error OOB, por el cual si no se alcanza una mejora mínima del 0,1% el cómputo finaliza en esa observación.

Este hiperparámetro se establece ahora en conjunto con el número de árboles para realizar cada uno de los *random forest* dentro de los cinco pliegues de la validación cruzada. A continuación, se recoge los resultados del entrenamiento del primer *random forest* con la primera partición de la muestra para un *random forest* de 200 árboles e intentar cuatro variables en cada criterio de selección de cada nodo.

4.3.2. Predicción

Con los parámetros ajustados al modelo explicativo de los datos, se procede a estimar el modelo, no obstante, el *random forest* no arroja coeficientes, varianzas, ni p-valores de los que entender el modelo. Por ende, se utiliza la matriz de confusión para contrastar la frecuencia de observaciones correctamente clasificadas y las erróneas según la votación mayoritaria de los doscientos árboles internos del bosque. A continuación, se

recoge el output de computar el *random forest* con los hiperparámetros anteriores y analizar el ajuste del modelo a la muestra de entrenamiento.

Tabla 5. *Random Forest* de clasificación con una partición de entrenamiento.

Tipo	Clasificación
Número de arboles	200
Número de variables en cada división	4
Error OOB de la estimación	1,95 %

Matriz de confusión

	Grave	Leve	Error de clase
Grave	4.489	1.109	19,81 %
Leve	5	53.364	0,009 %

Fuente: Microdatos Accidentes. DGT. Elaboración propia

En la parte superior de la tabla se encuentran los parámetros introducidos para elaborar el *random forest*, resultando en un error *Out-of-Bag* (OOB) medio del modelo de aproximadamente el 2%. Es importante denotar que en general el OOB suele subestimar el poder predictivo del modelo, por lo que se considera más que relevante realizar una validación cruzada por separado debido a este efecto.

La parte media-inferior del recuadro muestra la matriz de confusión. Esta matriz representa visualmente los resultados de la predicción con la muestra OOB del propio grupo de entrenamiento del *random forest* en las diferentes clases que ha predicho el modelo frente a las clases reales. Por norma general las filas denotan la predicción de las clasificaciones o clases, y las columnas las clasificaciones reales que tienen las observaciones reales. Para entenderlo mejor, las clasificaciones correctas se encuadran en la diagonal de la matriz, es decir, donde cada las observaciones tienen la misma clasificación en el índice de la columna que en de la fila.

Para todas aquellas predicciones que se han asignado incorrectamente, véase todas aquellas que no están en la matriz diagonal de la matriz de confusión, son predicciones incorrectas, por lo que se puede calcular el error de predicción condicionado a una clase. Es decir, formalmente para una observación x_h donde su clase real es *h*, y para toda clase *i*, que cumpla $h \neq i$, entonces:

$$Error\ de\ clase = \frac{\sum(x_{h \neq i} \mid clase_i)}{\sum Clase_i}$$

(Ecuación 8)

Se aprecia de la tabla superior que el error por clase es muy bajo, permitiendo que el ajuste sea muy bueno tanto para los Leves como para los heridos Graves. Si bien, los heridos Leves tienen un ajuste casi perfecto del modelo, es también lógico al pensar que las lesiones leves son altamente frecuentes y mucho más probable, que los accidentes graves. Los accidentes graves tienen una incidencia de error de aproximadamente el 20% siendo un muy buen estimador de accidentes más casuales y con mayor impacto en la integridad corporal y la salubridad del individuo. Nótese además que el error medio del modelo OOB se realiza mediante un promedio de los errores de clase proporcionales a la muestra que explican, por ende, que sea muy bajo.

Para entender la evolución del desempeño general del *ranfom forest* en el número de árboles que computa, véase la siguiente ilustración, donde se denota en error computado por el OOB (curva negra), y los errores de clase según el *bagging* crece en tamaño.

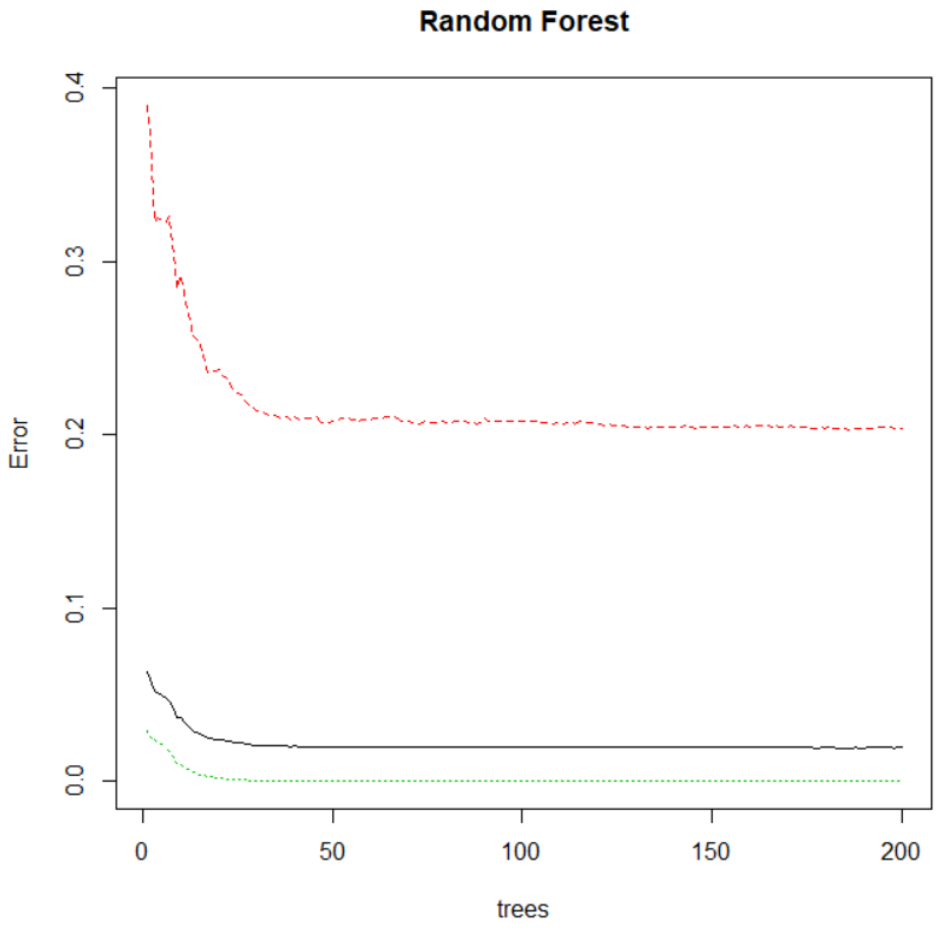


Ilustración 9. Evolución del error OOB (negro), error de clase “Leve” (verde) y error de clase “Grave” (rojo) frente al incremento del número de árboles en el *random forest*.

Desgraciadamente, un *random forest* no se puede graficar estructuralmente en su conjunto por la alta frecuencia del número de árboles que computan; ni sería prudente mostrar la representación de una moda de los nodos ya que no refleja el modelo real puesto que diferentes nodos de diferentes árboles explican hojas distintas o ramificaciones no equivalentes, y por ende no comparables en la votación. Por tanto, para evaluar el desempeño del modelo se analiza la relevancia de las variables en el mismo, esto explica así la estructura del bosque y los juicios que ha seleccionado.

Al igual que en una regresión lineal estimar los intervalos de confianza o un p-valor, en un *random forest*, la relevancia se mide mediante la media decreciente del criterio selector, en este caso de clasificación, la impuridad de Gini, esto no indica si el efecto es suficientemente integro estadísticamente como computar efectos en la media de datos, sino una valoración del uso y desempeño del mismo a la hora de dividir el árbol.

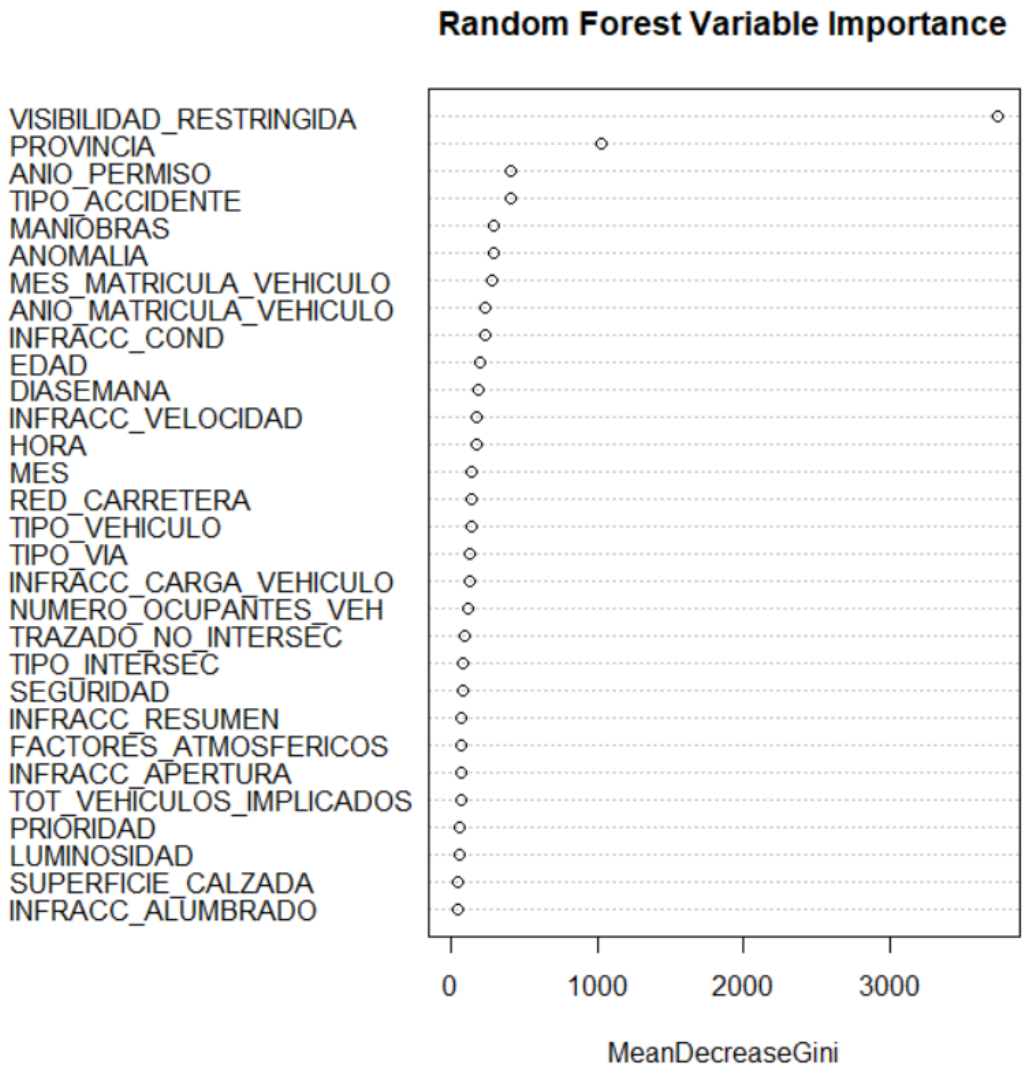


Ilustración 10. Gráfico decreciente de la importancia de las variables como criterios de selección en los árboles del *random forest*, computado a través del índice de Gini.

Se aprecia que la variable más importante con mayor poder predictivo y analítico en los heridos leves y graves es la visibilidad restringida. Esta variable clasifica en nueve categorías si el accidente ocurrió en un punto con falta de campo de visión al volante. La clasificación se divide en, “sin dato”, “edificios”, “configuración del terreno”, “vegetación”, “factores atmosféricos”, “deslumbramiento”, “polvo o humo”, “otra causa” o “sin restricción”.

La siguiente tabla muestra la frecuencia de víctimas (por observación) clasificadas en heridos leves o graves, desglosado por la visibilidad en la restricción. A priori se denota que graves tiene un alto índice de incidencias en factores intermedios a la visibilidad restringida, mientras que las severidades leves presentan la proporción más alta cuando no existe restricción.

Tabla 6. Tabla de frecuencias de visibilidad restringida frente a severidad

(Visibilidad Restringida | Y = y)	Leve	Grave
Edificios	3194	5052
Configuración del terreno	41	135
Vegetación	178	545
Factores atmosféricos	165	149
Deslumbramiento	22	188
Polvo o humo	8	78
Otra causa	812	358
Sin restricción	83.503	1.968

Fuente: Microdatos Accidentes. Dirección General de Tráfico. Elaboración propia

Es lógico intuir que la visibilidad restringida sea un parámetro crítico a la hora de analizar los accidentes. En situaciones adversas donde el conductor sea limitado por un objeto externo su campo de visión deduce que en primer lugar el conductor debe maniobrar el vehículo en base a una información escasa o nula del entorno vial; así como en casos en los que no permita al individuo reaccionar frente a una colisión o atropello por esta causa.

Seguidamente, la segunda variable más importante es, destacadamente, la provincia. Esta dictamina la geolocalización de los accidentes, pudiendo ser la infraestructura o patrones de la seguridad vial locales lo que dictamine mayor frecuencia de accidentes con severidades leves y graves. Otras variables determinantes a la hora de segregar el modelo han sido el año del permiso de conducir del individuo relevante al conductor y

el tipo de accidente ocasionado (salida de carretera, atropello, colisión, etc.). Por una parte, el año del permiso de conducir refleja la experiencia del conductor y el tipo de accidente determina la forma en la que ha ocurrido el mismo.

La geolocalización puede ser además un factor importante no solo a raíz de determinar el patrón de seguridad del lugar, sino que lugares con estructuras viales o mayor incidencia de pérdida de visibilidad puede dar lugar a que la interacción de las dos variables sea un buen predictor de la severidad.

Véase la siguiente figura donde se representa en tres dimensiones la provincia, la visibilidad restringida y; en dos niveles de "altura" se representa la severidad de las lesiones graves (bajo) y leves (alto); y a su derecha un gráfico de dispersión entre la provincia y la visibilidad restringida, donde los círculos rojos representan heridos leves y los triángulos azules representa a la severidad leve de los accidentes. Nótese que algunos triángulos superponen a los círculos, pero la mayoría de los graves se encuentran en dispersiones intermedias.

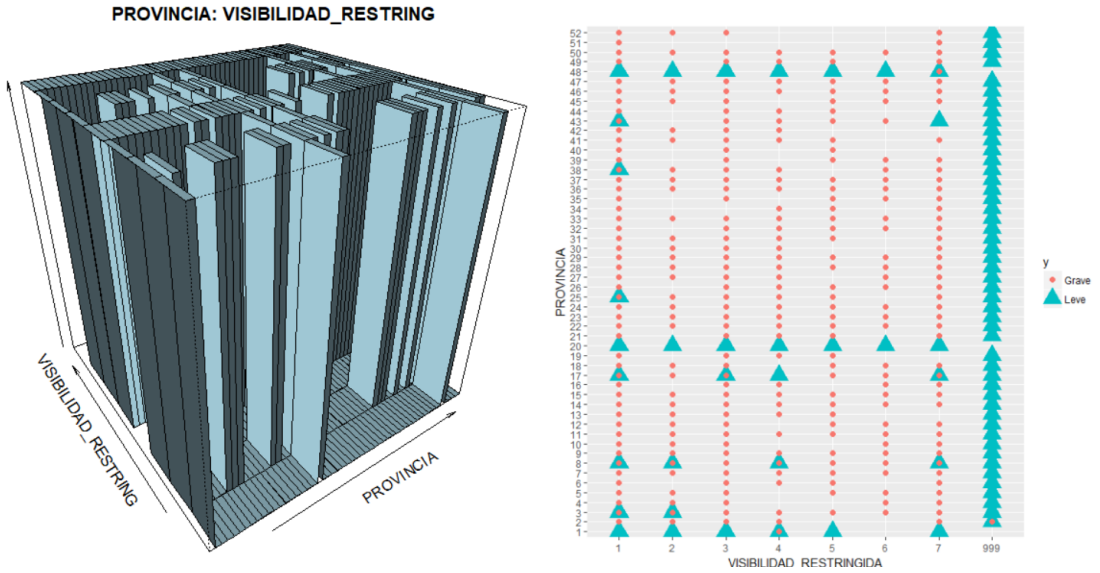


Ilustración 11. A la izquierda: gráfico en tres dimensiones de la variable estimada (severidad "y") en el eje vertical, y un factor de visibilidad restringida y provincia del lugar del accidente, en los ejes horizontales. A la derecha un gráfico de contraste de las mismas variables, denotando colores y forma para los niveles de severidad.

Nótese que esta interacción de las variables es muy importante ya que dictamina en gran medida la dispersión de los accidentes por geolocalización y por la característica de que el accidente se produce en una circunstancia donde el campo de visión de los conductores se perjudica por algún factor como edificios, vegetación o deslumbre de objetos o vehículos. Asimismo, muchos círculos rojos denotan en el gráfico derecho,

que ocurren en variedad del territorio español accidentes con víctimas graves mayoritariamente por falta de visibilidad en la conducción. La visibilidad restringida de 999 indica que no hay restricción en la visibilidad en la conducción. Claramente es un factor muy característico en la sucesión de los accidentes con heridos graves. Del gráfico no se pueden denotar las intensidades o provincias que tienen mayor impacto, o cual es el factor de visibilidad restringida que afecta mayormente en este tipo de accidentes.

Véase en la primera tabla del anexo 7.4 el mapeo de calor en una tabla filtrando la visibilidad restringida por provincias segregando para accidentes con heridos leves y heridos graves por separado, donde se denota claramente que los accidentes leves están en su mayoría en casos en los que la visibilidad no está restringida. Inclusive el alto índice de accidentes en Guipúzcoa y Vizcaya son accidentes leves y no graves, muchos de ellos en relación a falta de visibilidad por edificios.

Por otro lado, la gran cantidad de accidentes con víctimas graves, representado en la segunda tabla del anexo 7.4, en las todas las provincias variando por causa exacta de la falta de visibilidad, pero el factor discriminante sigue siendo el mismo del campo de visión en la conducción. Nótese que el cambio es drástico si se computan las localidades donde más accidentes ocurren, siendo la primera Madrid donde mayoritariamente los accidentes no presentan restricción de visibilidad porque es lógico ya que existe un alto índice de movilidad y transporte de vehículos al ser la capital. Asimismo, sin considerar la no restricción de visibilidad, siguen Valencia y Barcelona, en donde también por ser ciudades más grandes tienen más probabilidad de accidentados debido a tener más volumen de tránsito.

Es destacable que Valencia y Barcelona presentan altos índices en accidentes con heridos graves, y además la mayoría son en falta de visibilidad por causa de edificios. Esto puede indicar patrones en la arquitectura vial de la ciudad, pero también este efecto puede estar inflado por patrones de conducción más temerosos y culpar a la falta de visibilidad de los edificios a velocidades cuestionables. Esto sin embargo es un efecto potencial, no comprobable con la base de datos de accidentes disponible en la DGT.

Otras causas de la restricción de la visibilidad tanto en provincias muy pobladas como en poco pobladas varían en el factor de la visibilidad, pero en general todos representan en conjunto que el factor visibilidad es crucial en la influencia de los accidentes de una intensidad acorde a recurrir en lesiones graves a las víctimas del mismo.

4.3.3. Backtesting

A continuación, se realiza una comprobación del ajuste de bondad del modelo predictivo a una muestra de control con la que no se ha entrenado el modelo. Esta es la mejor manera de medir el desempeño del modelo debido a que el error OOB subestima el error del modelo en general, por lo que no es un estimador confiable sino es para ajustar los hiperparámetros del número de árboles o el número de variables a estudiar en cada criterio de selección de nodos.

La siguiente tabla muestra los resultados de incorporar en el modelo predictivo una muestra de control de 14.742 observaciones. La matriz de confusión que presenta es la mejor manera de analizar la clasificación real que ha elaborado el *random forest* anterior. Para el análisis, solo se ha seleccionado el *random forest* de la primera ejecución de entrenamiento, ya que la variación del modelo frente al resto de los cinco entrenamientos no varía significativamente (véase en el anexo 7.5 los cinco *ranfom forest* con sus respectivos entrenamientos y validaciones cruzadas).

Tabla 7. *Random Forest* de clasificación con una muestra de control.

Tipo	Clasificación
Número de arboles	200
Número de variables en cada división	4
Error de la muestra de control	1,87 %

Matriz de confusión

Predicción \ Real	Grave	Leve	Error de clase
Grave	1.129	275	19,59 %
Leve	0	13.338	0,00 %

Fuente: Microdatos Accidentes. DGT. Elaboración propia

En general el desempeño del *random forest* es muy bueno, sin presentar signos de sobreestimación. Las severidades graves se clasifican con una precisión del 80%, mientras que la precisión de las severidades leves es del 100%. Esto es un efecto típico en los problemas de clasificación, donde en el campo de la medicina se suele diferenciar la probabilidad de clasificación en dos casos concretos: especificidad y sensibilidad. Sensibilidad o probabilidad de detección, es la probabilidad de predecir un evento discriminante como una enfermedad para una población donde el valor real es un factor no discriminante. Especificidad se refiere a la probabilidad de predecir un evento no discriminante donde el valor real es ese evento no discriminante.

Para entender el concepto, en la tabla superior la sensibilidad es la proporción de graves reales que han sido realmente clasificados como graves. Sin embargo, la sensibilidad es la proporción de los leves reales que han sido clasificados como leves.

Una representación adicional del desempeño clasificatorio del modelo es con la representación del margen predictivo del *random forest*. Esta gráfica muy similar a la curva ROC en sistemas clasificatorios similares muestra la ratio de las predicciones correctas e incorrectas con respecto a la proporción máxima de votos. Véase la siguiente ilustración de los márgenes predictivos del *random forest*:

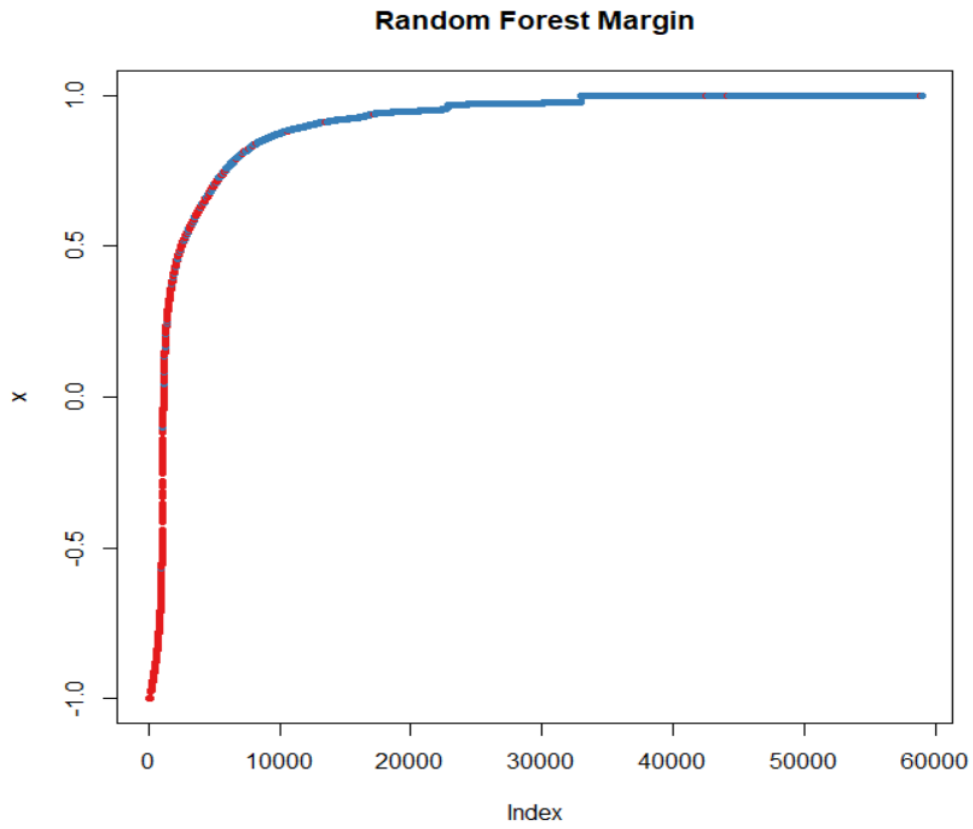


Ilustración 12. Gráfico de los márgenes de predicción del *random forest*.

Los valores del margen varían de -1 a 1 en el eje de ordenadas, donde el valor indica la asignación de las observaciones. Si la mayoría de los valores se encuentra por encima de cero significa que la mayoría de votos dieron una asignación mayoritaria correcta. Del mismo modo, si la mayoría de valores se encuentra por debajo de cero significa que se han realizado asignaciones incorrectas. El color indica la clasificación de las observaciones, es decir, en rojo se representa la severidad grave y en azul la leve. Distinguir que, al graficar, las asignaciones azules superponen a las graves, por ende,

se conoce que, de las tablas anteriores, existe gran cantidad de puntos rojos superpuestos por azules.

Nótese que la mayoría de puntos se encuentran por encima de cero, siendo una cantidad de puntos rojos por debajo, es decir, graves mal asignados y algún que otro punto azul (leve mal asignado). La representación cercana hacia el eje superior y el eje de las ordenadas también indica que rápidamente se supera el cero para un número de observaciones menor a mil. Ello indica que la vasta mayoría se encuentra bien clasificada en su categoría real.

4.4. Análisis de fallecidos

4.4.1. Modelo lineal generalizado

Para estudiar los fallecidos se ha estimado una regresión a través de un modelo linear generalizado para estudiar una variable binomial (fallecidos versus no-fallecidos) con función de enlace *logit* para estimar la probabilidad del suceso debido a un accidente de tráfico con víctimas. Este trabajo se enfoca en definir algoritmos de *machine learning* como principal indicador y en la predicción de la severidad de los accidentes de tráfico, por lo que la teoría metodológica de los modelos lineales generalizados se asumirá como conocida por el lector.

Para realizar esta regresión se necesita transformar la variable dependiente de la severidad, donde se construye una nueva variable “YY” que categoriza el resultado de la severidad del accidente de tráfico en “Vivo” o “Muerto”, denotando que solamente se estudia a una víctima por accidente. El modelo utiliza todas las variables explicativas anteriores sin modificaciones, siendo estas exactamente igual que en el modelo del *random forest*, donde únicamente ha cambiado la variable explicada.

Adicionalmente se resalta que la modelización se realiza también en R, por lo que la multicolinealidad se trata de forma automática. Es decir, al introducir las variables categóricas, el modelo auto reconoce la cantidad de factores, eliminando el primer factor de cada variable categórica al introducirse en el modelo para evitar la multicolinealidad explicada anteriormente.

Para estimar el modelo se utiliza una validación cruzada de 10 iteraciones o “desdoblamientos”. Al igual que el *random forest*, para ganar eficiencia en el modelo, se realizan diez ejecuciones del modelo con subconjuntos de proporción 9/10 y se realiza una validación cruzada con el 1/10 restante, retirando la ejecución con el siguiente

subconjunto no repetido. Asimismo, el modelo final resulta de converger el promedio de los 10 modelos generados en un único modelo predictivo.

De esta manera el modelo linear generalizado binomial con función "*logit*" se puede observar en completitud en el anexo 7.5.b). Donde se denota que las variables más significativas son la provincia, la edad, el sexo, factores lumínicos, la restricción de visibilidad por algún objeto o factor, el tipo de accidente (colisión, salida o atropello), la red de carretera, el uso del casco, entre otras. Una vez estimado el modelo disponible en el anexo b)7.5, se recogen los siguientes estadísticos del mismo:

Tabla 8. Estadística de residuos, criterio de información y desviaciones

Deviance Residuals				
Mínimo	1er Cuartil	Mediana	3er Cuartil	Máximo
-3,924	0,022	0,044	0,078	2,406
Null deviance	14.328,9	de 97.751 grados de libertad		
Residual deviance	7.782,9	de 97.521 grados de libertad		
AIC	8.244,9	No. Fisher scoring iterations		18

Fuente: Modelo binomial lineal generalizado función de enlace *logit*. Elaboración propia

De la desviación nula y la desviación residual se puede construir una medición del ajuste del modelo como un todo para determinar su significatividad como tal. Si bien el criterio de información de akaike (AIC) cuanto menor sea entre modelos, indica mejor ajuste general. Para modelos lineales generalizados se puede computar el ajuste general de bondad del modelo con respecto a un modelo con un solo parámetro, también denominado modelo nulo.

Esto se calcula dado que la diferencia de la desviación (que no variación de la estadística clásica) entre el modelo propuesto y el modelo 'nulo' (con sólo un parámetro) se distribuye como una chi-cuadrado con grados de libertad igual a la diferencia de grados de libertad entre los dos modelos propuestos.

En este modelo lineal generalizado el p-valor que ajusta a una chi-cuadrado con 230 grados de libertad resulta en 0. Esto significa que la hipótesis nula de que el modelo como un todo no ajusta mejor que un modelo nulo, se rechaza. Esta medida es una equivalencia al test estadístico de Fisher para medir la significatividad del modelo como un todo, sin embargo, ello no prueba que el modelo sea bueno o malo, para ello se realiza una predicción y un análisis posterior en el siguiente subapartado.

4.4.2. Predicción y *Backtesting*

A continuación, se realiza un *backtesting* del modelo predictivo contrastado con la validación cruzada de 10 iteraciones. Dicha validación cruzada resulta de juntar la predicción realizada con cada muestra no entrenada en el modelo lineal generalizado. Asimismo, se representan una batería de coeficientes que miden el desempeño del modelo explicados más adelante.

Antes de adentarse al modelo cabe presentar que este modelo predice una probabilidad de ocurrencia de un suceso donde el criterio para determinar en una predicción si recae en una clase o en otra (en un modelo binomial) suele ser escogiendo una banda probabilística que sirva de criterio selector. Al igual que en los nodos de un árbol de clasificación, esta banda indica una probabilidad, en donde si la probabilidad predicha para una observación es mayor a esta banda, esta observación se categoriza como "Vivo", en este modelo. Si, por el contrario, la probabilidad predicha para una observación es menor o igual a la banda probabilística, se categorizará al individuo como "Muerto".

Por norma general en los modelos lineales generalizados con función *logit*, se suele denotar una banda del 50%, sin embargo, para este modelo, esta banda se ha parametrizado para medir un mayor ajuste predictivo y además evitar repercusiones negativas en las predicciones, explicado más adelante.

La muestra completa de datos consta de 97.752 observaciones de donde solamente 1.360 víctimas han fallecido en el 2015, es decir, la base de datos recoge solamente un 1,4% de fallecidos del total de víctimas de tráfico. Por lo que una banda discriminante de supervivencia en proporción a la muestra de datos sería el 98,6%. Si se quisiera conducir a predicciones prudentes sobre el número de fallecidos se deben escoger bandas altas, mientras que por el contrario bandas inferiores.

En este trabajo para no inducir a un parámetro muy ajustado que pueda nublar la visión del modelo, se escoge una banda del 95% como criterio de supervivencia, dado que de esta manera el lector pueda discernir de un modelo menos ajustado hacia los fallecidos a base de medir su poder predictivo con bandas inferiores a la proporción de la muestra.

Tabla 9. Resultados de predicción de fallecidos con GLM binomial (función de enlace *logit*) con una validación cruzada de 10 iteraciones.

GLM binomial con función de enlace "*logit*"			
Precisión	95,32 %	Sensibilidad	81,99 %
Intervalo de confianza al 95%	(95,19 ; 95,45)	Especificidad	95,51 %
Precisión balanceada	88,74 %	Predicción positivos	20,5 %
Ratio de no información	98,61 %	Predicción negativos	99,74 %
P-valor [prec > RNI]	1	Prevalencia	1,39 %
Kappa	0,31	Error de detección	5,57 %
P-valor test de Mcnemar	2e-16	Prevalencia de detección	88,75 %

Matriz de confusión			
	Muerto	Vivo	Total
Muerto	1.115	4.329	5.444
Vivo	245	92.063	92.308
Total	1.360	96.392	97.752

Fuente: GLM binomial con función "*logit*". Elaboración propia.

Se denota en primer lugar que la precisión del modelo es bastante alta, midiendo un desempeño excelente de la ratio de aciertos entre totales en la predicción. Asimismo, una de las preocupaciones más importantes en la predicción es determinar las observaciones que caerán en la categoría de fallecidos y no en la de supervivientes en un accidente por lo que para calcular la precisión de predecir estas clases se denotan en sensibilidad y especificidad, respectivamente.

No obstante lo anterior, el modelo predictivo es recomendable que sea menos preciso en categorizar aquellas observaciones que sobreviven y clasificarlas como fallecidos; y por el contrario ser más preciso en no categorizar como vivos aquellos que fallecen, porque ello indica una pérdida de la eficiencia en un modelo de supervivencia. El primer caso, en donde se clasifica como fallecido una observación mientras que en la realidad ha sobrevivido no tiene repercusión negativa en la aplicación del modelo ni en la aplicación del modelo en la realidad.

Por ende, es preferible un modelo predictivo más prudente en el sentido de categorizar observaciones hacia fallecidos, aunque estas víctimas realmente sobrevivan el accidente; y por el contrario ser muy precavidos de no categorizar como vivo a

observaciones que realmente fallecen, lo que sí es una gran repercusión negativa en la predicción y aplicación del modelo. De la matriz de confusión se puede obtener las siguientes ratios de medición del error de predicción y error de categoría de clasificación:

Tabla 10. Error de predicción, error de clase y error promedio ponderado

	Muerto	Vivo	Promedio ponderado
Error de Clase	18,01 %	4,49 %	4,68 %
Error de Predicción	79,52 %	0,27 %	4,68 %

Fuente: Matriz de confusión de GLM binomial. Elaboración propia

El error principal por minimizar es el error de clase muerto, donde el 18% de las observaciones que realmente han fallecido, se han categorizado como vivos (predicción con repercusiones negativas). Este error se puede mejorar cambiando la banda probabilística a partir de la cual la observación se denota como fallecido o vivo. Este 18% concierne a todas aquellas observaciones que el modelo binomial lineal generalizado arroja con una probabilidad estimada superior al 95% y aun así este individuo ha fallecido por el accidente.

El error medio ponderado se denota como el error de clase o el error de predicción ponderado proporcionalmente por los pesos del número de observaciones en una clase o en una predicción, respectivamente. Cabe destacar que la suma del promedio ponderado corresponde a la equivalencia del error medio del modelo, lo que es idéntico a (1 – ratio de precisión). Asimismo, el desempeño del modelo en general es muy alto, en donde para esta banda probabilística se acepta un error de clase de fallecidos del 18%, pudiendo este parametrizarse con respecto a la aplicación que se quiere sobrellevar.

Por ejemplo, en la implementación en el automóvil de un sistema de alertas de accidentes tráfico automatizado a los equipos de emergencias. Si se incorpora al ordenador del vehículo con un modelo predictivo de severidades como el presente, puede parametrizarse para que categorice con una banda probabilística más alta a fallecidos (aunque sigan vivos) que justamente lo contrario, dado que la incidencia de fallecido también deduce que la severidad sufrida es de intensidad alta-moderada necesitando de una atención médica.

Una de las medidas más utilizadas para medir el ajuste en un modelo de clasificación es mediante el gráfico de ROC (Característica Operativa del Receptor por sus siglas en inglés) y el cómputo del AUC (área debajo de la curva por sus siglas en inglés). La curva

ROC mide la sensibilidad del sistema predictivo clasificador según la ratio de clasificaciones acertadas frente al ratio de clasificaciones incorrectas.

Para entender el siguiente gráfico ROC, las bandas de referencia son que un ROC representado como línea recta con ángulo de 45º significa que el sistema de clasificación no tiene ningún poder predictivo y es igual que una selección aleatoria del 50/50 (clasificaciones binarias). En el caso contrario, un ROC representado como una línea recta de dos tramos que sigue al eje izquierdo de las ordenadas y el eje superior de las abscisas; significa que el modelo predictivo clasificador tiene una precisión perfecta del 100%. Véase a continuación la representación de la curva ROC para el modelo binomial lineal generalizado:

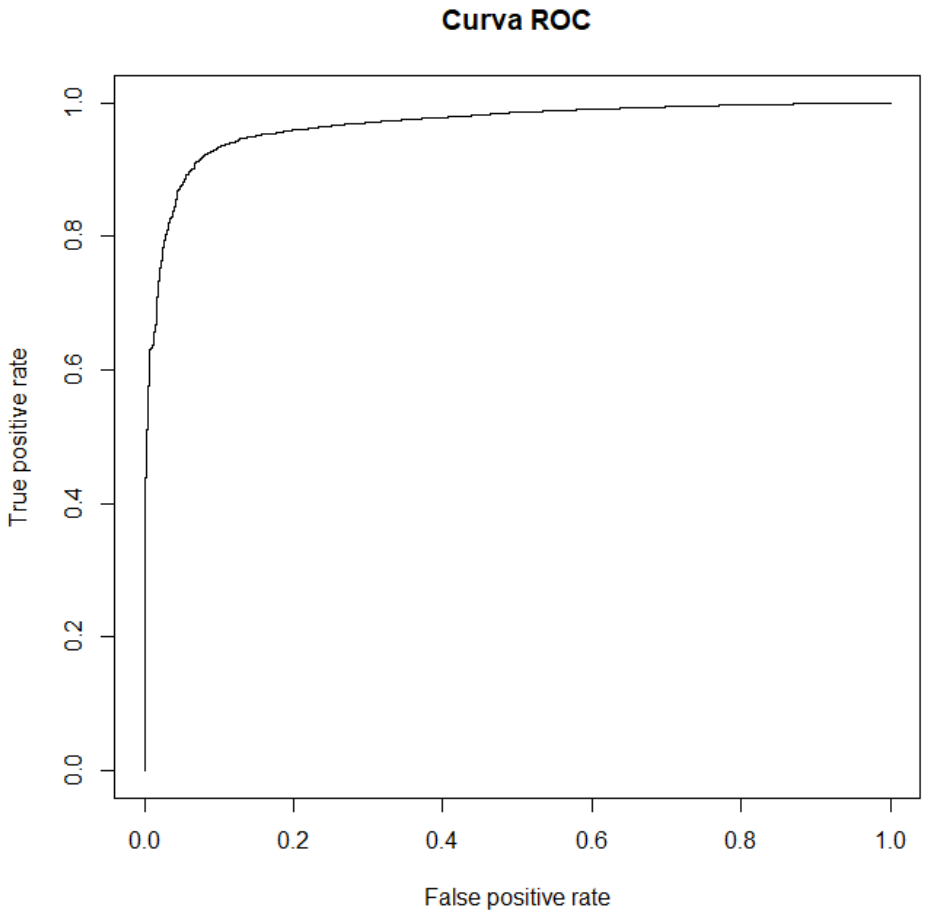


Ilustración 13. Gráfico curva ROC (Característica Operativa del Receptor) sobre el ajuste del modelo lineal generalizado binomial con función de enlace “logit”. Elaboración propia.

Nótese que el ajuste es muy cercano a la predicción perfecta, es decir, se aproxima al eje izquierdo de las ordenadas y el límite superior del recuadro. Asimismo, como la curva ROC es una representación para medir cuantitativamente en valor este ajuste se utiliza

el AUC o área debajo de la curva ROC. Este indicador oscila entre 0,5 y 1, donde para intervalos entre [0,5 ; 0,75) se consideran sistemas malos y regulares, de [0,75 ; 0,9) se considera un modelo bueno, para [0,9 a 0,97) se interpreta como muy bueno y para [0,97 a 1) se considera un resultado excelente.

El área debajo de la curva ROC en esta validación cruzada es de 0,9665; es decir, se encuentra entre los test muy buenos y excelentes. Por otra parte, el AUC puede interpretarse como la probabilidad de clasificar correctamente a una observación en referencia a clasificarla aleatoriamente.

5. Conclusiones y recomendaciones

5.1. Objetivos cumplidos

La aplicación de las metodologías de *Machine Learning* como el *random forest*, ajustan modelos estadísticos con alto poder predictivo en la severidad de los accidentes de tráfico con víctimas, determinado por factores externos. La precisión media del modelo es del 98% aproximadamente. Asimismo, el modelo linear generalizado determina con alta probabilidad posibles accidentes de tráfico que conlleven víctimas mortales, con una precisión promedio del 95%.

Los cuatro principales factores externos para determinar la severidad grave y leve en accidentes de tráfico con víctimas en España son el factor de la visibilidad restringida, la provincia, el año del permiso de conducir y el tipo de accidente. Por otro lado, para predecir aquellas víctimas mortales, el modelo lineal generalizado ha estimado como los más significativos la provincia, la edad, el sexo, factores lumínicos, la restricción de visibilidad por algún objeto o factor, el tipo de accidente (colisión, salida o atropello), la red de carretera y el uso del casco, entre las variables determinantes.

Existen dos limitaciones principales en el alcance del trabajo, que se desglosan en limitaciones endógenas y exógenas. Las limitaciones endógenas se basan en aquellas que intrínsecamente no han permitido desarrollarse en profundidad en el modelo predictivo. Entiéndase modelos estadísticos y predictivos que fueran conjunciones de otros modelos, buscando patrones estadísticos que pudieran crear mejores modelos predictivos. Por ejemplo la utilización de árboles de modelos lineales generalizados u otras metodologías que si bien podrían o no haber obtenido diferentes resultados a los presentes, no han sido consideradas en este estudio. Otro factor endógeno ha sido la utilización de datos de corte transversal, donde una observación se estudia en un único momento del tiempo. Sin embargo la base datos de la DGT dispone de muestras de datos de años anteriores, lo que podría dar algún indicador o patrón temporal sobre los accidentes. No obstante, no se podría realizar estudios de series temporales particulares o datos de panel debido a que las observaciones son distintas en todos los casos.

Por otro lado las limitaciones exógenas son de factores externo al trabajo en sí, siendo estos la limitación computacional y la estructura de datos de carácter público publicada por la DGT. La limitación computacional solamente ha permitido realizar un estudio para la muestra de accidentes de tráfico con víctimas para el periodo del 2015, y además donde se incluye únicamente una observación por accidente, sin analizar terceros individuos que también se han referenciado al mismo identificador del accidente.

Asimismo, esta capacidad ha limitado la utilización de algoritmos de alto coste computacional en los ordenadores y con los materiales disponibles.

La segunda limitación exógena ha sido la incapacidad de obtener datos propios del individuo para realizar un estudio en conjunto que discrimine aquellos patrones intrínsecos en la persona que expliquen ciertos accidentes, de los accidentes ocurridos por factores externos o bien, la iteración de los dos tipos de factores, internos y externos al individuo observado. Si bien, la privacidad de las víctimas y todo aquel tercero relacionado no puede ser de carácter público si consenso explícito, la severidad de las víctimas no se especifica, lo que también restringe los resultados posibles. Por tanto, sería sugestivo realizar un mismo estudio en donde la severidad de los accidentados se desglose en diversos niveles o diversas lesiones por gravedad, localización, costes médicos, entre otros. Esta enfoque sería muy atractivo para el ámbito actuarial en la modelización de sistemas de *pricing* en seguros de accidentes, de autos y de responsabilidad civil.

5.2. Líneas futuras de trabajo

La separación del estudio de severidad en dos secciones, vivos (graves y leves) y fallecidos, no concluye que el *random forest* empeore a la hora de predecir estos casos. Sin embargo, la modelización de una categoría con tan pocas observaciones, como es la de fallecidos, sesga el modelo predictivo en ese ámbito al no ser un criterio discriminatorio fuerte. Por ende, sería interesante la aplicación de *random forest* con implementación de ponderaciones o pesos a muestras con menor proporción de observaciones de una clase para estimar a los graves y fallecidos; quienes justamente son los más perjudicados pero inferiores en robustez estadística. Sin embargo estas metodologías se escapan del alcance de este trabajo limitado.

Asimismo, se anima a la comunidad actuarial a profundizar en otras metodologías, en especial en el ámbito de la inteligencia artificial, que puedan explicar en mejor medida los patrones de datos en un modelo conjunto. A modo de ejemplo, para predecir observaciones muy extremas como los fallecidos, un modelo interesante pudiera ser el *random forest* de percentiles, donde se estima un *random forest* pero para valores poblaciones atípicos, extremos o muy inferiores robustamente.

Asimismo, los datos de carácter público explican variables externas en su mayoría, por lo que la inclusión de variables endógenas al conductor y de los individuos relacionados, en conjunto con estos factores externos puede arrojar resultados concluyentes y de interés actuarial, asegurador y de seguridad vial.

6. Referencias

6.1. Referencias literarias

— Hastie, T., Tibshirani, R. y Friedman, J. (2009). *The Elements of Statistical Learning. Data mining, Inference, and Prediction*. Stanford, California: Springer Series in Statistics.

— James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013). *An introduction to Statistical Learning with Applications in R*. Stanford, Califortnia: Springer Series in Statistics.

— Breiman, Leo (1996). Bagging predictors 24 (2). *Machine Learning*. pp. 123-140

— Breiman, Leo (1994). Bagging predictors. *Technical report 421*, Department of Statistics, University of California at Berkeley.

— Morgan, Blake. FORBES. (25 Julio, 2017). “How Artificial Intelligence Will Impact The Insurance Industry”. [Artículo de revista en web]. Recuperado de https://www.forbes.com/sites/blakemorgan/2017/07/25/how-artificial-intelligence-will-impact-the-insurance-industry/#48e9e1d46531

— Ho, Tin Kam (1995). Random Decision Forests. Comunicación presentada en *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada*. Recuperado de http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf

— Breinman, Leo (2001). “Random Forest”. *Machine Learning*. 45 (1): págs. 5-32

— Liaw, A., Wiener, M. (2012). “Documentation for R package randomForest”. [Documento online]. Recuperado de https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

— Breiman, L., Friedman, J., Stone, C.J. y Olshen, R.A., (1984). *Classification and Regression Trees*. Nueva York: Chapman & Hall.

— A. Liaw and M. Wiener (2002). *Classification and Regression by randomForest*. R News 2(3), 18-22.

— Moreno Hoyo, Alberto. (2017). Universitat Politècnica De Catalunya. Departament D&apos, Estadística I Investigació Operativa, & Langohr, Klaus. (n.d.). Predicting car accidents in Barcelona using a Random Forest model.

— Lin, L., Wang, Q., y Sadek, A. (2017). Real-time Traffic Accident Risk Prediction based on Frequent Pattern Tree.

— Beshah, T., Ejigu, D., Abraham, A., Krömer, P., y Snášel, V. (2012). Knowledge discovery from road traffic accident data in ethiopia: Data quality, ensembling and trend analysis for improving road safety. Neural Network World, 22(3), 215-244.

— Chin, y Quddus. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention, 35(2), 253-259.
— Wuthrich, Mario V. and Buser, Christoph, (2017). Data Analytics for Non-Life Insurance Pricing. Swiss Finance Institute Research Paper No. 16-68. [Documento online] Recuperado de: https://ssrn.com/abstract=2870308
— Lin, W., Wu, Z., Lin, L., Wen, A. y Li, J. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. IEEE Access (vol. 5): 16568-16575.
— Alshamsi, A. (2014). Predicting car insurance policies using random forest. Innovations in Information Technology (INNOVATIONS), 2014 10th International Conference on, 128-132.
— Chen Chen. (2017). Analysis and Forecast of Traffic Accident Big Data. ITM Web of Conferences, 12, 04029.
— Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63.
— Langley, Pat (2011). "The changing science of machine learning". Machine Learning. 82 (3): 275–279.

6.2. Base de datos

— Dirección General de Tráfico. Microdatos de Accidentes: 2015. [base de datos]. Recuperado de https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces

6.3. Herramienta y paquetes

— RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. [Herramienta estadística] Recuperado de http://www.rstudio.com/
— Wickham, H., Bryan, J., RStudio (2018). R package ‘readxl’. [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/readxl/readxl.pdf
— Liaw, A., Wiener, M. (2018). R package ‘randomForest. [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
— Milborrow, S. (2018). R package ‘plotmo’. [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/plotmo/plotmo.pdf
— Wickham, H., Chang, W., RStudio (2016). R package ‘ggplot2’. [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
— Therneau, T., Atkinson, B., Ripley, B. (2018). R package ‘rpart’. [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/rpart/rpart.pdf
— Milborrow, S. (2018). R package ‘rpart.plot’ [paquete de R]. Recuperado de https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf

7. Anexos

7.1. Índice tablas

7.2. Índice Ilustraciones

7.3. Índice Ecuaciones

7.4. Tabla análisis estimadores

Tabla 11. Tabla de víctimas de accidentes de tráfico de lesiones **LEVES**. Mapeo de calor sobre la visibilidad restringida frente a la provincia del accidente.

VISIBILIDAD_RESTRINGIDA		1	2	3	4	5	6	7	999
PROVINCIA		EDIFICIOS	CONFIGURACIÓN DEL TERRENO	VEGETACIÓN	FACTORES ATMOSFÉRICOS	DESLUMBRAMIENTO	POLVO O HUMO	OTRA CAUSA	SIN RESTRICCIÓN
52	Melilla	0	0	0	0	0	0	0	328
51	Ceuta	0	0	0	0	0	0	0	161
50	Zaragoza	0	0	0	0	0	0	0	1255
49	Zamora	0	0	0	0	0	0	0	61
48	Bizkaia	784	13	58	47	6	2	68	0
47	Valladolid	0	0	0	0	0	0	0	666
46	Valencia/València	0	0	0	0	0	0	0	3221
45	Toledo	0	0	0	0	0	0	0	604
44	Teruel	0	0	0	0	0	0	0	91
43	Tarragona	16	0	0	0	0	0	1	1265
42	Soria	0	0	0	0	0	0	0	81
41	Sevilla	0	0	0	0	0	0	0	1351
40	Segovia	0	0	0	0	0	0	0	217
39	Cantabria	0	0	0	0	0	0	0	575
38	S.C.Tenerife	1	0	0	0	0	0	0	1354
37	Salamanca	0	0	0	0	0	0	0	514
36	Pontevedra	0	0	0	0	0	0	0	1448
35	Palmas, Las	0	0	0	0	0	0	0	913
34	Palencia	0	0	0	0	0	0	0	180
33	Asturias	0	0	0	0	0	0	0	1763
32	Ourense	0	0	0	0	0	0	0	289
31	Navarra	0	0	0	0	0	0	0	283
30	Murcia	0	0	0	0	0	0	0	1082
29	Málaga	0	0	0	0	0	0	0	1667
28	Madrid	0	0	0	0	0	0	0	12631
27	Lugo	0	0	0	0	0	0	0	338
26	Rioja, La	0	0	0	0	0	0	0	463
25	Lleida	3	0	0	0	0	0	0	812
24	León	0	0	0	0	0	0	0	499
23	Jaén	0	0	0	0	0	0	0	555
22	Huesca	0	0	0	0	0	0	0	340
21	Huelva	0	0	0	0	0	0	0	542
20	Gipuzkoa	1131	6	79	62	7	4	54	0
19	Guadalajara	0	0	0	0	0	0	0	153
18	Granada	0	0	0	0	0	0	0	885
17	Girona	12	0	1	1	0	0	2	1362
16	Cuenca	0	0	0	0	0	0	0	226
15	Coruña, A	0	0	0	0	0	0	0	958
14	Córdoba	0	0	0	0	0	0	0	1016
13	Ciudad Real	0	0	0	0	0	0	0	470
12	Castellón/Castelló	0	0	0	0	0	0	0	491
11	Cádiz	0	0	0	0	0	0	0	2088
10	Cáceres	0	0	0	0	0	0	0	396
9	Burgos	0	0	0	0	0	0	0	583
8	Barcelona	35	2	0	1	0	0	2	13597
7	Balears, Illes	0	0	0	0	0	0	0	2344
6	Badajoz	0	0	0	0	0	0	0	425
5	Ávila	0	0	0	0	0	0	0	357
4	Almería	0	0	0	0	0	0	0	506
3	Alicante/Alacant	1	1	0	0	0	0	0	1962
2	Albacete	0	0	0	0	0	0	0	321
1	Araba/Álava	509	7	24	27	3	0	6	0

Tabla 12. Tabla de víctimas de accidentes de tráfico de lesiones **GRAVES**. Mapeo de calor sobre la visibilidad restringida frente a la provincia del accidente.

VISIBILIDAD_RESTRINGIDA		1	2	3	4	5	6	7	999
	PROVINCIA	EDIFICIOS	CONFIGURACIÓN DEL TERRENO	VEGETACIÓN	FACTORES ATMOSFÉRICOS	DESLUMBRAMIENTO	POLVO O HUMO	OTRA CAUSA	SIN RESTRICCIÓN
52	Melilla	5	1	1	0	0	0	3	1
51	Ceuta	9	0	0	0	0	0	1	0
50	Zaragoza	171	2	5	3	6	2	5	10
49	Zamora	20	0	2	2	1	0	1	0
48	Bizkaia	60	0	5	2	2	0	7	0
47	Valladolid	63	4	8	4	1	2	2	0
46	Valencia/València	269	3	19	1	8	4	7	10
45	Toledo	86	3	3	0	2	3	4	2
44	Teruel	25	0	5	2	3	0	0	5
43	Tarragona	25	0	18	1	8	1	7	76
42	Soria	15	1	1	2	0	0	0	2
41	Sevilla	127	2	12	4	4	0	5	7
40	Segovia	30	0	2	0	1	0	0	0
39	Cantabria	43	0	4	0	1	1	1	0
38	S.C.Tenerife	127	2	14	2	0	2	4	4
37	Salamanca	54	1	5	2	3	2	5	2
36	Pontevedra	149	6	19	14	4	3	9	4
35	Palmas, Las	94	0	9	0	1	2	7	3
34	Palencia	39	0	0	2	1	0	1	1
33	Asturias	135	3	18	7	2	1	8	2
32	Ourense	59	0	7	1	1	1	4	0
31	Navarra	75	2	3	5	3	0	4	2
30	Murcia	120	6	10	3	0	0	7	8
29	Málaga	118	2	4	2	5	1	8	13
28	Madrid	186	4	10	2	9	5	9	930
27	Lugo	83	1	13	5	0	1	3	1
26	Rioja, La	31	0	1	3	1	1	1	2
25	Lleida	12	2	22	3	2	0	2	103
24	León	97	2	13	5	5	1	4	0
23	Jaén	42	1	9	2	2	1	1	0
22	Huesca	63	2	12	3	3	1	5	3
21	Huelva	49	0	5	2	2	0	1	1
20	Gipuzkoa	119	3	18	4	2	0	13	0
19	Guadalajara	33	0	2	1	1	0	0	0
18	Granada	103	4	17	0	2	2	4	2
17	Girona	30	3	31	0	11	1	4	84
16	Cuenca	27	0	2	1	0	2	2	1
15	Coruña, A	114	4	15	8	2	2	15	5
14	Córdoba	66	2	8	2	2	2	5	2
13	Ciudad Real	37	1	2	1	2	0	0	1
12	Castellón/Castelló	53	2	8	0	4	1	1	3
11	Cádiz	122	7	12	2	4	2	7	5
10	Cáceres	27	1	2	0	0	1	1	0
9	Burgos	83	2	6	2	4	1	2	5
8	Barcelona	257	4	59	7	18	1	35	320
7	Balears, Illes	216	7	26	1	7	3	8	18
6	Badajoz	47	0	7	1	4	0	3	1
5	Ávila	43	3	3	0	2	1	1	1
4	Almería	62	3	10	0	0	2	4	3
3	Alicante/Alacant	177	9	12	1	4	5	13	3
2	Albacete	49	1	6	1	0	0	5	4
1	Araba/Álava	51	1	2	4	1	0	2	0

7.5. Output entrenamiento *random forest y GLM*

a) Output R: modelo *Ranfom Forest*:

```
> ##############################
> # Folds for cross-validation #
> ##############################
> set.seed(1) #because of the sample function ahead
> dgt3 <- dgt2[sample(nrow(dgt2)),]
> dgt3 <- dgt3[, -6] #Drop "COMUNIDAD AUTONOMA"
> dgt3 <- na.omit(dgt3)
> folds <- cut(seq(1, nrow(dgt3)), breaks=5, labels=FALSE)
>
> #################
> # RANDOM FOREST #
> #################
> #i <- 1
> #accuracies <- c()
> inicio <- Sys.time()
> for (i in 1:5){
+   testIndexes <- which(folds == i, arr.ind=TRUE)
+   testData <- dgt3[testIndexes, ]
+   trainData <- dgt3[-testIndexes, ]
+
+   set.seed(2)
+   ranfor <- randomForest(y ~ .,
+                          xtest=testData[,2:ncol(testData)],
+                          ytest=testData$y,
+                          data=trainData,
+                          importance=TRUE,
+                          replace=TRUE,
+                          keep.forest=TRUE,
+                          mtry = 4,    # number of predictors sampled
+                          ntree=200)
+   print(ranfor)
+   print(duracion <- Sys.time()-inicio) # Check computational timing
+
+   assign(paste0("ranfor_", i), ranfor)
+   assign(paste0("dur_", i), duracion)
+
+   #prediction <- predict(ranfor, testData)
+   #testData$acierto <- prediction == testData$y
+   #t <- table(prediction, testData$y)
+   #accuracy <- sum(testData$acierto)/nrow(testData)
+   #accuracies <- c(accuracies, accuracy)
+   #print(accuracy)
+ }

Call:
 randomForest(formula = y ~ ., data = trainData, xtest = testData[,
   2:ncol(testData)], ytest = testData$y, importance = TRUE,      repl
ace = TRUE, keep.forest = TRUE, mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.33%
Confusion matrix:
      Grave  Leve class.error
Grave  5790   980  0.14475628
Leve   3129 67217  0.04448014
```

```
               Test set error rate: 4.73%
Confusion matrix:
      Grave  Leve class.error
Grave  1476   227  0.13329419
Leve    684 16893  0.03891449
Time difference of 2.299714 mins

Call:
 randomForest(formula = y ~ ., data = trainData, xtest = testData[,
   2:ncol(testData)], ytest = testData$y, importance = TRUE,      repl
ace = TRUE, keep.forest = TRUE, mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.88%
Confusion matrix:
      Grave  Leve class.error
Grave  5753   955  0.14236732
Leve   3582 66827  0.05087418
               Test set error rate: 5.59%
Confusion matrix:
      Grave  Leve class.error
Grave  1514   251  0.14220963
Leve    826 16688  0.04716227
Time difference of 4.571544 mins

Call:
 randomForest(formula = y ~ ., data = trainData, xtest = testData[,
   2:ncol(testData)], ytest = testData$y, importance = TRUE,      repl
ace = TRUE, keep.forest = TRUE, mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.82%
Confusion matrix:
      Grave  Leve class.error
Grave  5824   958  0.14125627
Leve   3533 66802  0.05023104
               Test set error rate: 5.65%
Confusion matrix:
      Grave  Leve class.error
Grave  1449   242  0.14311059
Leve    848 16740  0.04821469
Time difference of 6.858694 mins

Call:
 randomForest(formula = y ~ ., data = trainData, xtest = testData[,
   2:ncol(testData)], ytest = testData$y, importance = TRUE,      repl
ace = TRUE, keep.forest = TRUE, mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.11%
Confusion matrix:
      Grave  Leve class.error
Grave  5818   995  0.14604433
Leve   2942 67362  0.04184684
               Test set error rate: 5.05%
Confusion matrix:
```

```
      Grave  Leve class.error
Grave  1439   221  0.13313253
Leve    753 16866  0.04273795
Time difference of 9.122548 mins

Call:
 randomForest(formula = y ~ ., data = trainData, xtest = testData[,
   2:ncol(testData)], ytest = testData$y, importance = TRUE,      repl
ace = TRUE, keep.forest = TRUE, mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.89%
Confusion matrix:
      Grave  Leve class.error
Grave  5893   926  0.13579704
Leve   3613 66685  0.05139549
                Test set error rate: 6.12%
Confusion matrix:
      Grave  Leve class.error
Grave  1399   255  0.15417170
Leve    925 16700  0.05248227
Time difference of 11.31549 mins
```

b) Output R: modelo GLM binomial con función de enlace *logit*:

```
> summary(fit.dead2)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9236   0.0221   0.0438   0.0777   2.4063

Coefficients: (17 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.657e+01  1.520e+03   0.011 0.991303
MES                       -1.264e-02  9.637e-03  -1.312 0.189673
HORA                       1.226e-02  5.575e-03   2.199 0.027910 *
DIASEMANA2                 8.865e-02  1.220e-01   0.726 0.467618
DIASEMANA3                 6.540e-02  1.234e-01   0.530 0.596123
DIASEMANA4                 3.041e-01  1.271e-01   2.392 0.016753 *
DIASEMANA5                 7.967e-02  1.182e-01   0.674 0.500215
DIASEMANA6                 6.174e-02  1.161e-01   0.532 0.594771
DIASEMANA7                -1.545e-02  1.182e-01  -0.131 0.896010
PROVINCIA2                -3.616e+00  5.506e-01  -6.567 5.14e-11 ***
PROVINCIA3                -3.460e+00  4.950e-01  -6.989 2.76e-12 ***
PROVINCIA4                -3.171e+00  5.347e-01  -5.930 3.03e-09 ***
PROVINCIA5                -3.261e+00  5.815e-01  -5.608 2.05e-08 ***
PROVINCIA6                -3.578e+00  5.217e-01  -6.860 6.90e-12 ***
PROVINCIA7                -3.147e+00  4.905e-01  -6.416 1.40e-10 ***
PROVINCIA8                -1.344e+00  4.980e-01  -2.698 0.006984 **
PROVINCIA9                -3.259e+00  5.280e-01  -6.174 6.67e-10 ***
PROVINCIA10               -3.568e+00  5.686e-01  -6.275 3.49e-10 ***
PROVINCIA11               -3.251e+00  5.179e-01  -6.277 3.45e-10 ***
PROVINCIA12               -3.329e+00  5.450e-01  -6.109 1.01e-09 ***
PROVINCIA13               -3.849e+00  5.364e-01  -7.175 7.22e-13 ***
PROVINCIA14               -3.435e+00  5.256e-01  -6.535 6.35e-11 ***
PROVINCIA15               -3.058e+00  5.016e-01  -6.098 1.08e-09 ***
PROVINCIA16               -3.798e+00  5.625e-01  -6.752 1.46e-11 ***
```

```
PROVINCIA17                -1.818e+00  5.182e-01  -3.509 0.000449 ***
PROVINCIA18                -3.569e+00  5.057e-01  -7.058 1.69e-12 ***
PROVINCIA19                -2.914e+00  6.330e-01  -4.604 4.15e-06 ***
PROVINCIA20                -6.449e-01  4.370e-01  -1.476 0.140023
PROVINCIA21                -3.470e+00  5.371e-01  -6.461 1.04e-10 ***
PROVINCIA22                -2.873e+00  5.540e-01  -5.185 2.15e-07 ***
PROVINCIA23                -3.582e+00  5.331e-01  -6.720 1.82e-11 ***
PROVINCIA24                -3.441e+00  5.105e-01  -6.740 1.58e-11 ***
PROVINCIA25                -2.364e+00  5.205e-01  -4.543 5.55e-06 ***
PROVINCIA26                -3.988e+00  5.538e-01  -7.201 5.98e-13 ***
PROVINCIA27                -2.989e+00  5.363e-01  -5.573 2.50e-08 ***
PROVINCIA28                -3.078e+00  4.900e-01  -6.280 3.38e-10 ***
PROVINCIA29                -3.634e+00  4.994e-01  -7.277 3.41e-13 ***
PROVINCIA30                -3.307e+00  5.137e-01  -6.437 1.22e-10 ***
PROVINCIA31                -3.556e+00  5.339e-01  -6.660 2.73e-11 ***
PROVINCIA32                -2.604e+00  6.116e-01  -4.258 2.06e-05 ***
PROVINCIA33                -2.879e+00  5.175e-01  -5.563 2.65e-08 ***
PROVINCIA34                -2.531e+00  6.615e-01  -3.826 0.000130 ***
PROVINCIA35                -2.764e+00  5.388e-01  -5.130 2.89e-07 ***
PROVINCIA36                -2.915e+00  5.044e-01  -5.780 7.45e-09 ***
PROVINCIA37                -2.809e+00  5.861e-01  -4.793 1.65e-06 ***
PROVINCIA38                -3.193e+00  5.016e-01  -6.366 1.94e-10 ***
PROVINCIA39                -3.330e+00  5.449e-01  -6.112 9.85e-10 ***
PROVINCIA40                -3.552e+00  5.703e-01  -6.228 4.73e-10 ***
PROVINCIA41                -3.090e+00  5.037e-01  -6.135 8.49e-10 ***
PROVINCIA42                -3.454e+00  6.874e-01  -5.025 5.04e-07 ***
PROVINCIA43                -2.415e+00  5.056e-01  -4.777 1.78e-06 ***
PROVINCIA44                -3.561e+00  5.990e-01  -5.945 2.77e-09 ***
PROVINCIA45                -3.225e+00  5.242e-01  -6.151 7.69e-10 ***
PROVINCIA46                -3.306e+00  4.894e-01  -6.755 1.42e-11 ***
PROVINCIA47                -3.417e+00  5.443e-01  -6.277 3.44e-10 ***
PROVINCIA48                 5.433e-01  5.057e-01   1.074 0.282654
PROVINCIA49                -4.052e+00  5.803e-01  -6.982 2.91e-12 ***
PROVINCIA50                -3.041e+00  5.207e-01  -5.840 5.23e-09 ***
PROVINCIA51                -3.167e+00  8.085e-01  -3.917 8.97e-05 ***
PROVINCIA52                -2.362e+00  1.162e+00  -2.032 0.042113 *
TOT_VEHICULOS_IMPLICADOS   -1.375e-01  6.078e-02  -2.262 0.023670 *
ZONA2                       1.076e+00  2.699e-01   3.988 6.67e-05 ***
ZONA3                       2.437e-01  2.756e-01   0.884 0.376556
ZONA_AGRUPADA2                     NA         NA      NA       NA
RED_CARRETERA2              1.923e-01  1.054e-01   1.824 0.068202 .
RED_CARRETERA3              3.559e-01  1.373e-01   2.592 0.009544 **
RED_CARRETERA4              1.079e+00  1.975e-01   5.464 4.65e-08 ***
RED_CARRETERA5              4.593e-01  2.147e-01   2.139 0.032419 *
TIPO_VIA2                   4.862e-01  1.882e-01   2.584 0.009772 **
TIPO_VIA3                   3.598e-01  1.800e-01   1.999 0.045614 *
TIPO_VIA4                  -8.745e-02  3.105e-01  -0.282 0.778224
TIPO_VIA5                  -1.627e-01  5.411e-01  -0.301 0.763631
TIPO_VIA6                   2.774e-01  5.419e-01   0.512 0.608677
TIPO_VIA7                  -1.644e-01  3.006e-01  -0.547 0.584431
TIPO_VIA8                          NA         NA      NA       NA
TIPO_VIA9                          NA         NA      NA       NA
TRAZADO_NO_INTERSEC1       -3.981e-01  2.357e-01  -1.689 0.091204 .
TRAZADO_NO_INTERSEC2       -2.476e-01  2.438e-01  -1.016 0.309779
TRAZADO_NO_INTERSEC999     -5.210e-01  2.863e-01  -1.820 0.068788 .
TIPO_INTERSEC1             -9.901e-03  2.499e-01  -0.040 0.968399
TIPO_INTERSEC2              1.366e-02  2.649e-01   0.052 0.958881
TIPO_INTERSEC3              3.048e-01  2.844e-01   1.072 0.283833
TIPO_INTERSEC4                     NA         NA      NA       NA
PRIORIDADCeda              -1.494e+01  1.520e+03  -0.010 0.992162
PRIORIDADMarcas            -1.457e+01  1.520e+03  -0.010 0.992356
PRIORIDADNinguna           -1.474e+01  1.520e+03  -0.010 0.992265
```

```
PRIORIDADOtra              -1.507e+01  1.520e+03  -0.010 0.992090
PRIORIDADPaso              -1.471e+01  1.520e+03  -0.010 0.992282
PRIORIDADSemaforo          -1.510e+01  1.520e+03  -0.010 0.992074
PRIORIDADStop              -1.502e+01  1.520e+03  -0.010 0.992119
SUPERFICIE_CALZADA2        -7.061e-02  3.061e-01  -0.231 0.817568
SUPERFICIE_CALZADA3         1.817e-01  1.610e-01   1.129 0.259070
SUPERFICIE_CALZADA4         1.290e+00  8.065e-01   1.599 0.109834
SUPERFICIE_CALZADA5         1.859e+00  1.363e+00   1.365 0.172405
SUPERFICIE_CALZADA6         1.399e+01  5.047e+02   0.028 0.977888
SUPERFICIE_CALZADA7         7.460e-01  4.070e-01   1.833 0.066793 .
SUPERFICIE_CALZADA999       9.496e-01  6.156e-01   1.542 0.122957
LUMINOSIDAD2               -4.522e-01  1.231e-01  -3.673 0.000239 ***
LUMINOSIDAD3               -3.482e-01  1.063e-01  -3.275 0.001058 **
LUMINOSIDAD4               -5.322e-01  9.082e-02  -5.860 4.63e-09 ***
FACTORES_ATMOSFERICOS2     -3.539e-01  4.180e-01  -0.847 0.397200
FACTORES_ATMOSFERICOS3      2.187e-01  4.540e-01   0.482 0.630067
FACTORES_ATMOSFERICOS4      1.185e-01  2.024e-01   0.586 0.558025
FACTORES_ATMOSFERICOS5      1.514e-01  3.565e-01   0.425 0.670984
FACTORES_ATMOSFERICOS6     -3.191e-01  1.106e+00  -0.288 0.772974
FACTORES_ATMOSFERICOS7     -1.364e+00  9.017e-01  -1.513 0.130346
FACTORES_ATMOSFERICOS8      1.946e-02  3.850e-01   0.051 0.959675
FACTORES_ATMOSFERICOS9     -2.568e-01  1.402e-01  -1.832 0.066886 .
FACTORES_ATMOSFERICOS999   -4.331e-01  2.551e-01  -1.698 0.089512 .
VISIBILIDAD_RESTRINGIDA2    8.898e-01  4.227e-01   2.105 0.035282 *
VISIBILIDAD_RESTRINGIDA3    3.237e-01  1.244e-01   2.602 0.009272 **
VISIBILIDAD_RESTRINGIDA4    2.253e-01  2.856e-01   0.789 0.430324
VISIBILIDAD_RESTRINGIDA5    1.830e+00  3.486e-01   5.250 1.52e-07 ***
VISIBILIDAD_RESTRINGIDA6    7.869e-01  4.123e-01   1.909 0.056303 .
VISIBILIDAD_RESTRINGIDA7    3.369e-01  1.858e-01   1.813 0.069877 .
VISIBILIDAD_RESTRINGIDA999  3.762e+00  9.557e-02  39.367  < 2e-16 ***
TIPO_ACCIDENTE2             4.756e-01  1.569e-01   3.031 0.002435 **
TIPO_ACCIDENTE3             2.465e+00  3.815e-01   6.462 1.03e-10 ***
TIPO_ACCIDENTE4             7.829e-01  1.802e-01   4.344 1.40e-05 ***
TIPO_ACCIDENTE5             4.744e-01  2.349e-01   2.020 0.043366 *
TIPO_ACCIDENTE6             3.271e-01  2.906e-01   1.126 0.260371
TIPO_ACCIDENTE7            -4.343e-01  1.693e-01  -2.564 0.010335 *
TIPO_ACCIDENTE8             1.250e+00  5.103e-01   2.450 0.014302 *
TIPO_ACCIDENTE9             1.506e+00  3.619e-01   4.162 3.15e-05 ***
TIPO_ACCIDENTE10            1.140e+00  2.774e-01   4.110 3.95e-05 ***
TIPO_ACCIDENTE11            1.453e+01  3.359e+03   0.004 0.996548
TIPO_ACCIDENTE12            3.995e-01  1.874e-01   2.132 0.033017 *
TIPO_ACCIDENTE13           -4.508e-02  3.057e-01  -0.147 0.882753
TIPO_ACCIDENTE14            6.968e-01  2.189e-01   3.184 0.001454 **
TIPO_ACCIDENTE15            8.876e-01  2.713e-01   3.272 0.001067 **
TIPO_ACCIDENTE16            3.943e-01  1.787e-01   2.206 0.027358 *
TIPO_ACCIDENTE17            3.841e-01  2.795e-01   1.374 0.169444
TIPO_ACCIDENTE18            9.631e-01  2.113e-01   4.557 5.19e-06 ***
TIPO_ACCIDENTE19            1.293e+00  2.471e-01   5.232 1.68e-07 ***
TIPO_ACCIDENTE20           -2.108e-01  1.793e-01  -1.175 0.239825
EDAD                       -1.030e-02  2.767e-03  -3.722 0.000197 ***
SEXO2                       3.418e-01  9.567e-02   3.573 0.000353 ***
SEXO999                     7.904e-01  6.472e-01   1.221 0.221989
ANIO_PERMISO               -7.125e-04  3.096e-03  -0.230 0.817990
POSICION1                   1.957e+00  1.157e+00   1.692 0.090660 .
POSICION2                   1.452e+01  2.179e+03   0.007 0.994684
POSICION3                   1.849e+01  1.075e+04   0.002 0.998628
POSICION4                   1.504e+01  5.740e+03   0.003 0.997909
POSICION6                   3.491e+00  1.830e+00   1.907 0.056506 .
POSICION7                   1.411e+01  3.283e+03   0.004 0.996571
POSICION99                  1.097e+01  3.675e+03   0.003 0.997618
USO_CINTURON1               2.461e-02  1.048e-01   0.235 0.814339
USO_CINTURON2              -9.421e-01  1.277e-01  -7.380 1.58e-13 ***
```

USO_CINTURON3	NA	NA	NA	NA	
USO_SRI3	NA	NA	NA	NA	
USO_SRI99	NA	NA	NA	NA	
USO_CASCO1	7.490e-01	2.022e-01	3.705	0.000212	***
USO_CASCO2	-7.457e-02	2.244e-01	-0.332	0.739695	
USO_CASCO3	NA	NA	NA	NA	
MANIOBRAS1	-1.353e+00	9.547e-01	-1.417	0.156574	
MANIOBRAS2	-1.201e+00	9.639e-01	-1.246	0.212636	
MANIOBRAS3	-1.219e+00	9.628e-01	-1.266	0.205592	
MANIOBRAS4	-3.561e-01	1.427e+00	-0.250	0.802942	
MANIOBRAS5	-1.555e+00	9.817e-01	-1.584	0.113303	
MANIOBRAS6	5.227e-01	1.411e+00	0.370	0.711113	
MANIOBRAS7	-1.889e+00	1.069e+00	-1.767	0.077193	.
MANIOBRAS8	-1.622e+00	1.036e+00	-1.566	0.117404	
MANIOBRAS9	-1.922e+00	1.071e+00	-1.795	0.072700	.
MANIOBRAS10	1.264e+01	1.381e+03	0.009	0.992697	
MANIOBRAS11	-1.753e+00	1.003e+00	-1.748	0.080404	.
MANIOBRAS12	-1.095e+00	9.965e-01	-1.099	0.271960	
MANIOBRAS13	1.291e+01	4.571e+02	0.028	0.977463	
MANIOBRAS14	3.462e-01	1.402e+00	0.247	0.804950	
MANIOBRAS15	1.243e+01	3.064e+02	0.041	0.967647	
MANIOBRAS16	-1.592e+00	1.016e+00	-1.567	0.117086	
MANIOBRAS17	-4.837e-01	1.140e+00	-0.424	0.671252	
MANIOBRAS18	-1.199e+00	1.015e+00	-1.181	0.237540	
MANIOBRAS19	-1.588e+00	9.866e-01	-1.609	0.107546	
MANIOBRAS20	-1.410e-01	1.202e+00	-0.117	0.906623	
MANIOBRAS21	-1.099e+00	9.842e-01	-1.117	0.263941	
MANIOBRAS22	-1.101e+00	1.054e+00	-1.044	0.296346	
MANIOBRAS23	-7.738e-01	1.175e+00	-0.658	0.510302	
MANIOBRAS24	-1.249e+00	1.004e+00	-1.244	0.213522	
MANIOBRAS25	-1.391e+00	1.038e+00	-1.341	0.179978	
MANIOBRAS26	-2.203e+00	1.205e+00	-1.828	0.067556	.
MANIOBRAS27	1.250e+01	1.407e+03	0.009	0.992911	
MANIOBRAS29	5.687e-01	1.404e+00	0.405	0.685372	
MANIOBRAS30	1.385e+01	6.484e+02	0.021	0.982959	
MANIOBRAS31	-1.704e+00	1.084e+00	-1.572	0.115865	
MANIOBRAS41	1.937e-01	1.445e+00	0.134	0.893356	
MANIOBRAS42	1.350e+01	1.891e+03	0.007	0.994304	
MANIOBRAS43	-1.185e+00	1.401e+00	-0.846	0.397473	
MANIOBRAS51	-2.128e+00	1.036e+00	-2.055	0.039919	*
MANIOBRAS52	-1.708e+00	9.810e-01	-1.741	0.081762	.
MANIOBRAS61	NA	NA	NA	NA	
MANIOBRAS71	6.428e-01	1.415e+00	0.454	0.649552	
MANIOBRAS72	-9.020e-01	1.000e+00	-0.902	0.367054	
MANIOBRAS77	-2.017e+00	9.844e-01	-2.049	0.040426	*
INFRACC_VELOCIDAD1	-2.206e-01	1.234e-01	-1.788	0.073794	.
INFRACC_VELOCIDAD2	1.483e+01	1.385e+03	0.011	0.991459	
INFRACC_VELOCIDAD3	4.893e-01	1.029e-01	4.756	1.98e-06	***
INFRACC_VELOCIDAD4	NA	NA	NA	NA	
INFRACC_COND1	4.149e-01	3.220e-01	1.288	0.197574	
INFRACC_COND2	6.943e-01	2.368e-01	2.932	0.003370	**
INFRACC_COND3	6.327e-01	2.477e-01	2.554	0.010642	*
INFRACC_COND4	-2.057e-01	3.275e-01	-0.628	0.529888	
INFRACC_COND5	-8.710e-02	1.808e-01	-0.482	0.629963	
INFRACC_COND6	-1.341e-01	3.265e-01	-0.411	0.681335	
INFRACC_COND7	7.183e-01	2.673e-01	2.687	0.007211	**
INFRACC_COND8	-7.458e-02	2.037e-01	-0.366	0.714271	
INFRACC_COND9	9.890e-03	1.032e-01	0.096	0.923631	
INFRACC_COND10	NA	NA	NA	NA	
INFRACC_APERTURA1	1.546e+01	8.498e+02	0.018	0.985484	
INFRACC_APERTURA2	9.599e-01	2.944e-01	3.261	0.001110	**
INFRACC_APERTURA3	NA	NA	NA	NA	

```
INFRACC_ALUMBRADO1          -1.006e+00  4.560e-01  -2.206 0.027379 *
INFRACC_ALUMBRADO2                  NA         NA      NA       NA
INFRACC_ALUMBRADO3                  NA         NA      NA       NA
INFRACC_CARGA_VEHICULO1      1.545e+01  9.801e+02   0.016 0.987425
INFRACC_CARGA_VEHICULO2     -7.809e-01  3.005e-01  -2.599 0.009355 **
INFRACC_CARGA_VEHICULO3             NA         NA      NA       NA
INFRACC_RESUMEN1            -1.252e-02  1.497e-01  -0.084 0.933365
INFRACC_RESUMEN2             1.362e-01  1.267e-01   1.075 0.282341
INFRACC_RESUMEN3                    NA         NA      NA       NA
ANIO_MATRICULA_VEHICULO      6.982e-03  5.093e-03   1.371 0.170449
MES_MATRICULA_VEHICULO       2.126e-02  9.230e-03   2.303 0.021269 *
TIPO_VEHICULO3              -1.452e+00  4.005e-01  -3.625 0.000289 ***
TIPO_VEHICULO4              -1.183e+00  3.865e-01  -3.060 0.002210 **
TIPO_VEHICULO5               7.236e-02  1.656e+00   0.044 0.965155
TIPO_VEHICULO6               5.301e-01  1.653e+00   0.321 0.748443
TIPO_VEHICULO7               4.074e-01  2.012e+00   0.203 0.839516
TIPO_VEHICULO8              -2.128e-01  1.674e+00  -0.127 0.898829
TIPO_VEHICULO9               4.206e-01  1.655e+00   0.254 0.799442
TIPO_VEHICULO10              2.253e-01  1.674e+00   0.135 0.892916
TIPO_VEHICULO11                     NA         NA      NA       NA
TIPO_VEHICULO12             -2.942e-01  1.661e+00  -0.177 0.859425
TIPO_VEHICULO13             -2.343e-01  1.883e+00  -0.124 0.900964
TIPO_VEHICULO14             -2.153e-01  1.705e+00  -0.126 0.899559
TIPO_VEHICULO15              5.227e-02  1.663e+00   0.031 0.974930
TIPO_VEHICULO16              2.553e-01  1.677e+00   0.152 0.878999
TIPO_VEHICULO17             -1.009e+00  2.066e+00  -0.488 0.625455
TIPO_VEHICULO18              7.504e-01  1.751e+00   0.429 0.668261
TIPO_VEHICULO19             -1.712e+00  1.743e+00  -0.982 0.325964
TIPO_VEHICULO20              1.341e+00  1.715e+00   0.782 0.434091
TIPO_VEHICULO21              2.167e+00  1.673e+00   1.295 0.195282
ANOMALIAFrenos              -3.803e-01  1.422e+00  -0.267 0.789097
ANOMALIANeumático            8.756e-01  1.359e+00   0.644 0.519503
ANOMALIANinguna              8.234e-01  1.337e+00   0.616 0.537905
ANOMALIAReventón             2.204e+00  1.545e+00   1.427 0.153715
`ANOMALIASin dato`          -1.448e+00  1.377e+00  -1.051 0.293142
NUMERO_OCUPANTES_VEH        -1.740e-02  1.364e-02  -1.276 0.201860
MERCANCIAS_PELIGROSAS1       1.551e+00  9.600e-01   1.616 0.106202
VEHICULO_INCENDIADO1        -9.999e-01  4.256e-01  -2.349 0.018823 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14328.9  on 97751  degrees of freedom
Residual deviance:  7782.9  on 97521  degrees of freedom
AIC: 8244.9

Number of Fisher Scoring iterations: 18
```

7.6. Glosario

Acrónimo	Descripción
ACR	Árbol de Clasificación y/o Regresión
AUC	Area Under the Curve (Área debajo de la curva ROC)
DGT	Dirección General de Tráfico. Ministerio del Interior. Gobierno de España.
OOB	Muestra Out-Of-Bag (observaciones fuera de la selección aleatoria del *bagging*)
ROC	Receiver Operating Characteristic, (Característica Operativa del Receptor)

7.7. Datos de víctimas de accidentes

Tabla 13. Estructura de los datos

Variable	**Tipo**	**Referencia**
Severidad (Y)	Factor	Leve o Grave
Mes	Numérica entera	Enero, febrero, marzo, …
Hora	Numérico	…, 12, 13, 14, 15, 16, 17…
Dia de la semana	Factor	Lunes, martes, miércoles...
Provincia	Factor	1 araba/álava 2 albacete 3 alicante/alacant 4 almería 5 ávila 6 badajoz 7 balears, illes 8 barcelona 9 burgos 10 cáceres 11 cádiz 12 castellón/castelló 13 ciudad real 14 córdoba 15 coruña, a 16 cuenca 17 girona 18 granada 19 guadalajara 20 gipuzkoa 21 huelva 22 huesca 23 jaén 24 león 25 lleida 26 rioja, la 27 lugo 28 madrid 29 málaga 30 murcia 31 navarra 32 ourense 33 asturias 34 palencia 35 palmas, las 36 pontevedra 37 salamanca 38 s.c.tenerife 39 cantabria

		40 segovia 41 sevilla 42 soria 43 tarragona 44 teruel 45 toledo 46 valencia/valència 47 valladolid 48 bizkaia 49 zamora 50 zaragoza 51 ceuta 52 melilla
Comunidad autónoma	Factor	1 andalucía 2 aragón 3 asturias, principado de 4 balears, illes 5 canarias 6 cantabria 7 castilla y león 8 castilla-la mancha 9 cataluña 10 comunitat valenciana 11 extremadura 12 galicia 13 madrid, comunidad de 14 murcia, región de 15 navarra, comunidad foral de 16 rioja, la 17 país vasco 18 ceuta y melilla
Vehículos implicados	Numérica entera	1, 2, 3, 4…
Zona	Factor	1 carretera 2 zona urbana 3 travesía 4 variante
Zona agrupada	Factor	1 vías interurbanas 2 vías urbanas
Carretera	Factor	1 titularidad estatal 2 titularidad autonómica 3 titularidad provincial (diputación, cabildo o consell) 4 titularidad municipal 5 otras titularidades
Tipo de vía	Factor	1 autopista 2 autovía 3 vía para automóviles 4 vía convencional con carril lento 5 vía convencional 6 camino vecinal 7 vía de servicio 8 ramal de enlace 9 otro tipo
Trazado de la vía	Factor	0 intersección 1 recta 2 curva suave 3 curva fuerte sin señalizar 4 curva fuerte con señal y sin velocidad señalizada 5 curva fuerte con señal y velocidad señalizada
Tipo de intersección	Factor	0 no aplica 1 en t ó y 2 en x ó + 3 enlace de entrada 4 enlace de salida 5 giratoria

		6 otros
Prioridad del agente	Factor	0 ninguna 1 agente
Prioridad del semáforo	Factor	0 ninguna 1 semaforo
Prioridad del stop	Factor	0 ninguna 1 señal de “stop”
Prioridad del ceda el paso	Factor	0 ninguna 1 señal de “ceda el paso”
Prioridad de marcas	Factor	0 ninguna 1 solo marcas viales
Prioridad de paso peatonal	Factor	0 ninguna 1 paso para peatones
Otra prioridad	Factor	0 ninguna 1 otra señal
Superficie de la calzada	Factor	1 seca y limpia 2 umbría 3 mojada 4 helada 5 nevada 6 barrillo 7 gravilla suelta 8 aceite 9 otro tipo
Nivel de luminosidad del día	Factor	1 pleno día 2 crepúsculo 3 noche: iluminación suficiente 4 noche: iluminación insuficiente 5 noche: sin iluminación
Factores atmosféricos	Factor	1 buen tiempo 2 niebla intensa 3 niebla ligera 4 lloviznando 5 lluvia fuerte 6 granizando 7 nevando 8 viento fuerte 9 otro
Visibilidad restringida	Factor	0 sin dato 1 edificios 2 configuración del terreno 3 vegetación 4 factores atmosféricos 5 deslumbramiento 6 polvo o humo 7 otra_causa 8 sin restricción
Tipo de accidente	Factor	11 Colisión de vehículos en marcha (Frontal) 12 Colisión de vehículos en marcha (Frontolateral) 13 Colisión de vehículos en marcha (Lateral) 14 Colisión de vehículos en marcha (Alcance) 15 Colisión de vehículos en marcha (Múltiple o en caravana) 21 Colisión de vehículo con obstáculo en calzada (Vehículo estacionado o averiado) 22 Colisión de vehículo con obstáculo en calzada (Valla de defensa) 23 Colisión de vehículo con obstáculo en calzada (Barrera de paso a nivel) 24 Colisión de vehículo con obstáculo en calzada (Otro objeto o material) 31 Atropello a peatón sosteniendo bicicleta 32 Atropello a peatón reparando vehículo 33 Atropello a peatón aislado o en grupo 34 Atropello a conductor de animales 35 Atropello a animal conducido o en rebaño

		36 Atropello a animales sueltos 41 Vuelco en la calzada 51 Salida de la vía por la izquierda con colisión (Choque con árbol o poste) 52 Salida de la vía por la izquierda con colisión (Choque con muro o edifício) 53 Salida de la vía por la izquierda con colisión (Choque con cuneta o bordillo) 54 Salida de la vía por la izquierda con colisión (Otro tipo de choque) 55 Salida de la vía por la izquierda sin colisión (Con despeñamiento) 56 Salida de la vía por la izquierda sin colisión (Con vuelco) 57 Salida de la vía por la izquierda sin colisión (En llano) 58 Salida de la vía por la izquierda sin colisión (Otra) 61 Salida de la vía por la derecha con colisión (Choque con árbol o poste) 62 Salida de la vía por la derecha con colisión (Choque con muro o edifício) 63 Salida de la vía por la derecha con colisión (Choque con cuneta o bordillo) 64 Salida de la vía por la derecha con colisión (Otro tipo de choque) 65 Salida de la vía por la derecha sin colisión (Con despeñamiento) 66 Salida de la vía por la derecha sin colisión (Con vuelco) 67 Salida de la vía por la derecha sin colisión (En llano) 68 Salida de la vía por la derecha sin colisión (Otra) 71 Otro tipo de accidente
Edad	Numérica	0, 1, 2, 3, 4, 5…
Sexo	Factor	1 mujer 2 hombre
Año del permiso de conducir	Numérica entera	…, 73, 74, 75, 76, 77…
Posición de la víctima con respecto al vehículo	Factor	0 peatones 1 conductor vehículo 2 pasajero delantero 3 pasajero trasero izquierdo 4 pasajero trasero derecho 5 pasajero trasero central 6 conductor vehículo de dos ruedas 7 pasajero vehículo de dos ruedas 8 otros pasajeros sentados 9 otros pasajeros de pie
Uso del cinturón	Factor	0 Sin dato 1 Lleva cinturón 2 No lleva cinturón 3 No aplica
Uso del Sistema de Retención Infantil (SRI)	Factor	0 Sin dato 3 Lleva cinturón 99 No lleva cinturón
Uso de casco	Factor	0 Sin dato 1 Lleva casco 2 No lleva casco 3 No aplica
Maniobra realizada	Factor	0 sin dato 1 siguiendo la ruta 2 adelantando por la derecha 3 adelantando por la izquierda 11 girando o saliendo hacia otra vía o acceso por la derecha

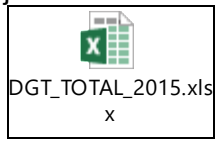
		12 girando o saliendo hacia otra vía o acceso por la izquierda 13 girando en "u" 21 incorporándose desde otra vía o acceso 22 cruzando intersección 23 estacionando o saliendo del estacionamiento 31 circulando hacia atrás 41 maniobra súbita para salvar obstáculo o vehículo 42 maniobra súbita para salvar peatón aislado o en grupo 43 brusca reducción de velocidad 51 retención por imperativo de la circulación 52 parado o estacionado 61 fugado 71 otra 72 se ignora
Infracción de velocidad	Factor	0 sin dato 1 velocidad inadecuada en condición existente 2 sobrepasar la velocidad establecida 3 marcha lenta entorpeciendo la circulación 4 ninguna
Infracción de la conducción	Factor	0 Sin dato 1 Conducción distraída o desatenta 2 Circular o invadir el sentid contrario o prohibido 3 No mantener distancia o frenar sin causa 4 No respetar prioridades, semáforos o señales 5 No indicar maniobra o sin precaución 6 Ciclista o ciclomotor en paralelo o fuera de pista 7 Incorrecta apertura 8 Otra infraccion 9 Ninguna infraccion 10 No aplica
Infracción de apertura	Factor	0 Sin dato 1 Apertura de puertas sin precaución 2 Apertura de puertas con precaución 3 No aplica
Infracción de alumbrado	Factor	0 Sin dato 1 Incorrecta utilización del alumbrado 2 Correcta utilización del alumbrado 3 No aplica
Infracción de carga del vehículo	Factor	0 Sin dato 1 Parado o estacionamiento prohibido o peligroso 2 No hay infracción de estacionamiento o parado 3 No aplica
Infracción peatonal	Factor	0 No aplica 1 No respetar las señalizaciones 2 Irrumpir la vía o estar antireglamentariamente 3 Ninguna infracción
Resumen infracciones	Factor	0 Sin dato 1 Presenta alguna infracción 2 Ninguna infracción 3 NA
Año de matrícula del vehículo	Numérica entera	…, 46, 47, 48, 49, 50…
Mes de matrícula del vehículo	Numérica entera	1 enero 2 febrero 3 marzo 4 abril 5 mayo 6 junio 7 julio 8 agosto 9 septiembre 10 octubre 11 noviembre 12 diciembre

Tipo de vehículo	Factor	1 Bicicleta o triciclo sin motor 2 Ciclomotor 10 Coche de Minusválido 11 Motocicleta 21 Turismo de SP hasta 9 plazas 22 Turismo sin remolque 23 Turismo con remolque 24 Ambulancia 30 Maquinaria de obras y agrícola 31 Tractor agrícola sin remolque 32 Tractor agrícola con remolque 41 Camión (PM <= 3500 K) sin remolque 42 Camión (PM <= 3500 K) con remolque 43 Furgoneta 51 Camión (PM > 3500 K) sin remolque 52 Camión (PM > 3500 K) con remolque 53 Camión cisterna sin remolque 54 Camión cisterna con remolque 55 Vehículo articulado 61 Autobús de línea regular 62 Autobús escolar 63 Otro autobús 70 Tren 80 Carro 81 Otros Vehículos 82 Cuadriciclo 90 Desconocido
Ninguna anomalía	Factor	1 Ninguna anomalia 2 Alguna anomalía 3 Sin dato / No aplica
Anomalía en los neumáticos	Factor	1 Anomalía en neumáticos 2 No anomalía en neumáticos 3 Sin dato / No aplica
Anomalía de reventón	Factor	1 Anomalía de reventón 2 No anomalía de reventón 3 Sin dato / No aplica
Anomalía de dirección	Factor	1 Anomalía en la dirección del vehículo 2 No anomalía en dirección 3 Sin dato / No aplica
Anomalía en los frenos	Factor	1 Anomalía en frenos 2 No anomalía en frenos 3 Sin dato / No aplica
Número de ocupantes del vehículo	Numérica entera	1, 2, 3, 4…
Transporte de mercancías peligrosas	Factor	0 No mercancías peligrosas 1 Sí mercancías peligrosas
Vehículo incendiado	Factor	0 incendiado 1 sí incendiado

Fuente: Microdatos Accidentes de la Dirección General de Tráfico. Tabla de elaboración propia

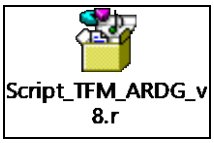
7.8. Datos de Accidentes de Tráfico con Víctimas

Hoja de Microsoft Excel:



Base de datos R (formato .Rda):

7.9. Modelización en R



Script completo ejecutado en R:

```
#   Financial & Actuarial Science Master Thesis
#   Produced by:  Alejandro Rubén Domingo Gesteiro
#   Created:      08/05/2018

# Purge R before any calculus
#cat("\014") #clean screen
#remove(list= ls()) #erase any data

# Define path to Working Directory (unique path to import all exce
getwd()
setwd("C:/Users/alejandrorubendoming/Desktop/Trafico Accidentes")
setwd("D:/adomi/Dropbox/Dropbox/Master - Finance & Actuarial Science/2º Año/2º Cuatrimestre/TFM")

# Install Packages
#install.packages("readxl")
#install.packages("randomForest")
#install.packages("rpart")
#install.packages("rpart.plot")
#install.packages("plotmo")
#install.packages("ggthemes")
#install.packages("ggExtra")
#install.packages("ggplot2")
#install.packages("caret")
#install.packages("e1071")
#install.packages("ROCR")

# Load packages
library(readxl)
library(randomForest)
library(rpart)
library(rpart.plot)
library(plotmo)
library(ggthemes)
library(ggExtra)
library(ggplot2)
library(caret)
library(e1071)
library(ROCR)

#######################################
## Import, Create & Clean dataset ##
#######################################

# Traffic victims (injury, death, etc)
dgt <- read_excel("DGT_TOTAL_2015.xlsx", col_names = TRUE)
dgt <- as.data.frame(dgt)
str(dgt)

dgt$TRAZADO_NO_INTERSEC[is.na(dgt$TRAZADO_NO_INTERSEC)] <- 0
dgt$TIPO_INTERSEC[is.na(dgt$TIPO_INTERSEC)] <- 0
dgt$PRIORIDAD_AGENTE[is.na(dgt$PRIORIDAD_AGENTE)] <- 0
dgt$PRIORIDAD_CEDA[is.na(dgt$PRIORIDAD_CEDA)] <- 0
dgt$PRIORIDAD_MARCAS[is.na(dgt$PRIORIDAD_MARCAS)] <- 0
dgt$PRIORIDAD_OTRA[is.na(dgt$PRIORIDAD_OTRA)] <- 0
dgt$PRIORIDAD_SEMAFORO[is.na(dgt$PRIORIDAD_SEMAFORO)] <- 0
dgt$PRIORIDAD_PASO[is.na(dgt$PRIORIDAD_PASO)] <- 0
dgt$PRIORIDAD_STOP[is.na(dgt$PRIORIDAD_STOP)] <- 0
```

```
# Change Anomalia-dummies into a single "Anomalia" factor variable
for (i in 1:nrow(dgt)){
  if(dgt$ANOMALIA_DIRECCION[i]==1){
    dgt$ANOMALIA[i] <- "Direccion"
  } else if(dgt$ANOMALIA_FRENOS[i]==1){
    dgt$ANOMALIA[i] <- "Frenos"
  } else if(dgt$ANOMALIA_NEUMATICO[i]==1){
    dgt$ANOMALIA[i] <- "Neumático"
  } else if(dgt$ANOMALIA_REVENTON[i]==1){
    dgt$ANOMALIA[i] <- "Reventón"
  } else if(dgt$ANOMALIA_NINGUNA[i]==3){
    dgt$ANOMALIA[i] <- "Sin dato"
  } else {
    dgt$ANOMALIA[i] <- "Ninguna"
  }
}
dgt$ANOMALIA <- as.factor(dgt$ANOMALIA)
sort(unique(dgt$ANOMALIA))

# Change Prioridad-dummies into a single "Prioridad" factor variable
for(i in 1:nrow(dgt)){
  if(dgt$PRIORIDAD_AGENTE[i]==1){
    dgt$PRIORIDAD[i] <- "Agente"
  }else if(dgt$PRIORIDAD_CEDA[i]==1){
    dgt$PRIORIDAD[i] <- "Ceda"
  } else if(dgt$PRIORIDAD_MARCAS[i]==1){
    dgt$PRIORIDAD[i] <- "Marcas"
  } else if(dgt$PRIORIDAD_OTRA[i]==1){
    dgt$PRIORIDAD[i] <- "Otra"
  } else if(dgt$PRIORIDAD_PASO[i]==1){
    dgt$PRIORIDAD[i] <- "Paso"
  } else if(dgt$PRIORIDAD_SEMAFORO[i]==1){
    dgt$PRIORIDAD[i] <- "Semaforo"
  } else if(dgt$PRIORIDAD_STOP[i]==1){
    dgt$PRIORIDAD[i] <- "Stop"
  } else {
    dgt$PRIORIDAD[i] <- "Ninguna"
  }
}
dgt$PRIORIDAD <- as.factor(dgt$PRIORIDAD)
sort(unique(dgt$PRIORIDAD))

for (i in 1:nrow(dgt)){
  if(dgt$USO_CINTURON[i]==1){
    dgt$SEGURIDAD[i] <- "Cinturon"
  } else if(dgt$USO_CINTURON[i]==2){
    dgt$SEGURIDAD[i] <- "NO_Cinturon"
  } else if(dgt$USO_SRI[i]==3){
    dgt$SEGURIDAD[i] <- "SRI"
  } else if(dgt$USO_CASCO[i]==1){
    dgt$SEGURIDAD[i] <- "Casco"
  } else if(dgt$USO_CASCO[i]==2){
    dgt$SEGURIDAD[i] <- "No_Casco"
  } else {
    dgt$SEGURIDAD[i] <- "NA/NS"
  }
}
dgt$SEGURIDAD <- as.factor(dgt$SEGURIDAD)
sort(unique(dgt$SEGURIDAD))
#_________Structure_Check_______________
#any(is.na(dgt$INFRACC_PEATON))
#str(dgt$ACCION_PEATON)
#summary(dgt$ACCION_PEATON)
#sort(unique(dgt$NUMERO_OCUPANTES_VEH))
#_____________________________________
```

```
################################################
## Specify external and objective variables ##
################################################

attach(dgt)
external <- data.frame(
  #                 ID_ACCIDENTE,
  #ANIO,
  MES,
  HORA,
  DIASEMANA,
  PROVINCIA,
  COMUNIDAD_AUTONOMA,
  #                 ISLA,
  #                 COD_MUNICIPIO,
  #                 MUNICIPIO,
  TOT_VEHICULOS_IMPLICADOS,
  ZONA,
  ZONA_AGRUPADA,
  #                 CARRETERA,
  RED_CARRETERA,
  TIPO_VIA,
  TRAZADO_NO_INTERSEC,
  TIPO_INTERSEC,
  # PRIORIDAD_AGENTE,
  # PRIORIDAD_SEMAFORO,
  # PRIORIDAD_STOP,
  # PRIORIDAD_CEDA,
  # PRIORIDAD_MARCAS,
  # PRIORIDAD_PASO,
  # PRIORIDAD_OTRA,
  PRIORIDAD,  # Artificial variable
  SUPERFICIE_CALZADA,
  LUMINOSIDAD,
  FACTORES_ATMOSFERICOS,
  VISIBILIDAD_RESTRINGIDA,
  #   ACERAS,
  TIPO_ACCIDENTE,
  #                 ID_VEHICULO,
  #                 ID_PERSONA,
  #                 ID_CONDUCTOR,
  #                 ID_PASAJERO,
  #                 ID_PEATON,
  EDAD,
  SEXO,
  ANIO_PERMISO,
  POSICION,
  #USO_CINTURON,
  #USO_SRI,
  #USO_CASCO,
  SEGURIDAD,
  #                 DICCIONARIO_MANIOBRAS,
  MANIOBRAS,
  INFRACC_VELOCIDAD,
  INFRACC_COND,
  INFRACC_APERTURA,
  INFRACC_ALUMBRADO,
  INFRACC_CARGA_VEHICULO,
  INFRACC_RESUMEN,
  INFRACC_PEATON,
  #                 DICCIONARIO_ACCION_PEATON,
  #                 ACCION_PEATON,
  #                 ANIO__1,
  #                 ID_VEHICULO.y,
  ANIO_MATRICULA_VEHICULO,
  MES_MATRICULA_VEHICULO,
  TIPO_VEHICULO,
```

```
  # ANOMALIA_NINGUNA,
  # ANOMALIA_NEUMATICO,
  # ANOMALIA_REVENTON,
  # ANOMALIA_DIRECCION,
  # ANOMALIA_FRENOS,
  ANOMALIA,   # Artificial variable
  NUMERO_OCUPANTES_VEH,
  MERCANCIAS_PELIGROSAS,
  VEHICULO_INCENDIADO
  #   ANIO__2
)
external[is.na(external)] <- 0
for(j in 3:ncol(external)){
  external[, j] <- as.factor(external[, j])
}
external$EDAD <- as.numeric(external$EDAD)
external$ANIO_PERMISO <- as.integer(external$ANIO_PERMISO)
external$MES <- as.integer(external$MES)
external$ANIO_MATRICULA_VEHICULO <- as.integer(external$ANIO_MATRICULA_VEHICULO)
external$MES_MATRICULA_VEHICULO <- as.integer(external$MES_MATRICULA_VEHICULO)
external$NUMERO_OCUPANTES_VEH <- as.integer(external$NUMERO_OCUPANTES_VEH)
external$TOT_VEHICULOS_IMPLICADOS <- as.integer(external$TOT_VEHICULOS_IMPLICADOS)
str(external)
summary(external)

#_________Check which variables have more than 53 categories (max for random forest package)_____
#num_col <- c()
#for(j in 1:ncol(external)){
#  num_col <- c(num_col, length(unique(external[, j])))
#}
#names(num_col) <- names(external)
#num_col
#_____________________________________________________________________________________________
_____________

# Objective recovers all the dependent variables to explain
objective <- data.frame()
objective <- data.frame(
  TOT_VICTIMAS,
  TOT_VICTIMAS30D,
  TOT_MUERTOS,
  TOT_MUERTOS30D,
  TOT_HERIDOS_GRAVES,
  TOT_HERIDOS_GRAVES30D,
  TOT_HERIDOS_LEVES,
  TOT_HERIDOS_LEVES30D,
  MUERTO_24H,
  MUERTO_30D,
  HERIDO_GRAVE_24H,
  HERIDO_GRAVE30D,
  HERIDO_LEVE_24H,
  HERIDO_LEVE30D
)
any(is.na(objective))
str(objective)
summary(objective)
detach(dgt)

#____________________Structure_Check____________________
# See how from 32 people involved in an accident, 27 were lightly injured, and 5 had a severe injury
# but estimate freq is not the objective in the paper at hand
#View(objective[objective$TOT_VICTIMAS==32, ])

# *** SELECT OBJECTIVE VARIABLE FOR ANALYSIS ***
# 'y' assigns  "Muerto", "Grave" or "Leve" to the injury category suffered in the accident
y <- c()
for(i in 1:nrow(objective)){
```

```
  if (objective$TOT_MUERTOS[i] > 0){
    y[i] <- "Muerto"
    next
  } else if(objective$TOT_HERIDOS_GRAVES[i] > 0){
    y[i] <- "Grave"
    next
  } else if(objective$TOT_HERIDOS_LEVES[i] > 0){
    y[i] <- "Leve"
    next
  } else {
    y[i] <- "Ileso"
    next
  }
  y <- as.factor(y)
  print(table(y)) # Check how many obs. are in each category
}

# Check structure, summary and unique variables
str(y)
table(y)
summary(y)
sort(unique(y))

# Percentage of Deads from total: Most people didn't die!
#y.test <- dgt$TOT_MUERTOS
#(percent.zeros <- length(y.test[y.test==0])/(length(y.test)))
#(percent.no.zero <- (1 - percent.zeros))

#########################
# Save & Load Dataset #
#########################
dgt2 <- cbind(y, external)
summary(dgt2)
str(dgt2)

# --- SAVE ---
#save(dgt2, file="dgt2.Rda")

# --- Environment ---
setwd("C:/Users/alejandrorubendoming/Desktop/Trafico Accidentes")
setwd("D:/adomi/Dropbox/Dropbox/Master - Finance & Actuarial Science/2º Año/2º Cuatrimestre/TFM")

# --- LOAD ---
load("dgt2.Rda")

# Separate Deads from others
dead <- dgt2
yy <- c()
for (j in 1:nrow(dead)){
  if(dead$y[j]=="Muerto"){
    yy[j] <- "Muerto"
  } else {
    yy[j] <- "Vivo"
  }

}
yy <- as.factor(yy)
dead <- cbind(yy, dead)

sort(unique(dead$yy))
table(dead$y, dead$yy)

dead <- dead[,-2]
str(dead)

dgt2 <- dgt2[dgt2$y!="Muerto", ]
dgt2$y <- factor(dgt2$y)
```

```
dgt2 <- na.omit(dgt2)
str(dgt2)

############################
# Descriptive Statistics #
############################
summary(dgt2)
str(dgt2)
#_____________________________________________
# TIPO_VIA
table(dgt2$TIPO_VIA, dgt2$y)
table(dgt2$TIPO_VIA, dgt2$y)*100/nrow(dgt2)
table(dgt2$TIPO_VIA[dgt2$y=="Grave"], dgt2$y[dgt2$y=="Grave"])*100/sum(dgt2$y=="Grave")
table(dgt2$TIPO_VIA[dgt2$y=="Leve"], dgt2$y[dgt2$y=="Leve"])*100/sum(dgt2$y=="Leve")
table(dgt2$TIPO_VIA[dgt2$y=="Muerto"], dgt2$y[dgt2$y=="Muerto"])*100/sum(dgt2$y=="Muerto")
#_____________________________________________
# TRAZADO_NO_INTERSECC y TIPO_VIA

#_____________________________________________
# EDAD
table(dgt2$EDAD, y)
qplot(y, EDAD, data=dgt2, geom=c("boxplot"),
            fill=y,
            main="Boxplot Lesiones versus Edad víctimas",
            xlab="",
            ylab="Edades víctimas"
            )
#_____________________________________________
table(dgt2$SEXO, y) #1=MUJER, 2=HOMBRE, 999=NA
sum(dgt2$SEXO=="1")/nrow(dgt2) # Women proportion
sum(dgt2$SEXO=="2")/nrow(dgt2) # Men proportion
#_____________________________________________
table(dgt2$POSICION, dgt2$SEXO) # 0=Peaton, 1=Conductor, 6=Motorista
sum(dgt2$SEXO[dgt2$POSICION=="1"]=="1")/nrow(dgt2)
#_____________________________________________
table(dgt2$POSICION, dgt2$y)*100/nrow(dgt2)
sum(dgt2$POSICION[dgt2$y=="Muerto"]=="1")*100/sum(dgt2$POSICION=="1")
#_____________________________________________
table(dgt2$MES, dgt2$y)
summary(dgt2$MES)
windows()
qplot(y, MES, data=dgt2, geom=c("boxplot"),
     fill=y,
     main="Boxplot Severidad versus Edad",
     xlab="",
     ylab="Edades víctimas")
#_____________________________________________
ggplot(data=dgt2) +
  geom_bar(mapping=aes(x=EDAD, fill=y),
  position = "dodge")
#_____________________________________________
windows()
bar.month <- ggplot(data=dgt2) +
  geom_bar(
    mapping = aes(x=PROVINCIA, fill=y),
    show.legend = TRUE,
    width = 1
  ) +
  theme(aspect.ratio = 1) +
  labs(x = NULL, y = NULL)

bar.month + coord_flip() # Sideways plot
#bar.month + coord_polar() # Pizza plot
#_____________________________________________
windows()
ggplot(data=dgt2, mapping=aes(x=DIASEMANA, y=HORA)) +
  geom_boxplot()
```

```
ggplot(data=dgt2, mapping=aes(x=EDAD, y=NUMERO_OCUPANTES_VEH)) +
  geom_boxplot() +
  coord_flip()
#_____________________________________________
table(dgt2$VISIBILIDAD_RESTRINGIDA, y)
table(dgt2$VISIBILIDAD_RESTRINGIDA, y)*100/nrow(dgt2)
#_____________________________________________
windows()
par(mfrow=c(2,2))
boxplot(dgt2$MES, dgt2$DIASEMANA)
boxplot(dgt2$MES, dgt2$HORA)
hist(table(dgt2$TIPO_ACCIDENTE))
hist(dgt2$EDAD)

windows()
g <- ggplot((data=dgt2), aes(x=HORA, stat="count")) +
  scale_fill_brewer((palette = "Spectral"))
g + geom_histogram(aes(fill=y),
                   stat="count",
                   binwidth = .1,
                   col = "black",
                   size=.1) +
  labs(title ="Histograma Hora del accidente",
       subtitle = "Hora del accidente con víctimas desglosado por severidades")

windows()
g <- ggplot((data=dgt2), aes(x=TRAZADO_NO_INTERSEC, stat="count")) +
  scale_fill_brewer((palette = "Spectral"))
g + geom_histogram(aes(fill=y),
                   stat="count",
                   binwidth = .1,
                   col = "black",
                   size=.1) +
  labs(title ="Histograma Factores Atmosféricos",
       subtitle = "Factores atmosféricos frente a la severidad de accidentes de tráfico")

# TRAZADO NO INTERSECCION
table(dgt2$TRAZADO_NO_INTERSEC, dgt2$y)
table(dgt2$TRAZADO_NO_INTERSEC, dgt2$y)*100/nrow(dgt2)
traz1                              <-                              table(dgt2$TRAZADO_NO_INTERSEC[dgt2$y=="Grave"],
dgt2$y[dgt2$y=="Grave"])*100/sum(dgt2$y=="Grave")
traz2                              <-                                table(dgt2$TRAZADO_NO_INTERSEC[dgt2$y=="Leve"],
dgt2$y[dgt2$y=="Leve"])*100/sum(dgt2$y=="Leve")
traz3                              <-                              table(dgt2$TRAZADO_NO_INTERSEC[dgt2$y=="Muerto"],
dgt2$y[dgt2$y=="Muerto"])*100/sum(dgt2$y=="Muerto")

traz1 <- as.table(traz1[,1])
names(traz1) <- c("Intersección", "Recta", "Curva", "S/dato")
windows()
pie(traz1, labels=paste0(names(traz1), " (", round(traz1,1), " %)"),
    col=rainbow(nrow(table(dgt2$TRAZADO_NO_INTERSEC))),
    main="Tarta de Trazado_No_Intersección en lesionados Graves")

traz2 <- as.table(traz2[,2])
names(traz2) <- c("Intersección", "Recta", "Curva", "S/dato")
windows()
pie(traz2, labels=paste0(names(traz2), " (", round(traz2,1), " %)"),
    col=rainbow(nrow(table(dgt2$TRAZADO_NO_INTERSEC))),
    main="Tarta de Trazado_No_Intersección en lesionados Leves")

traz3 <- as.table(traz3[,3])
names(traz3) <- c("Intersección", "Recta", "Curva", "S/dato")
windows()
pie(traz3, labels=paste0(names(traz3), " (", round(traz3,1), " %)"),
    col=rainbow(nrow(table(dgt2$TRAZADO_NO_INTERSEC))),
    main="Tarta de Trazado_No_Intersección en Fallecidos")
```

```
windows()
theme_set(theme_bw())
g2 <- ggplot(dgt2, aes(ANIO_MATRICULA_VEHICULO, ANIO_PERMISO)) +
  geom_count() +
  geom_smooth(method="lm", se=F)
ggMarginal(g2, type="histogram", fill="transparent")

########################
# Classification Tree #
########################
#formulas <- as.formula(paste("y ~", paste(colnames(dgt3)[2:20], collapse="+")))
train2 <- dgt2[1:1000,]
ct <- rpart(y ~., method="class", data = train2, cp=0.01)
windows()
rpart.plot(ct, extra=100)

summary(ct)

windows()
plotmo(ct, type="prob", nresponse="Leve", ngrid2=200)

windows()
plotmo(ct, type="prob", nresponse="Leve",
       type2="image", ngrid2=200,
       pt.col=ifelse(train2$y=="Leve", "red", "lightblue"))

#install.packages("rattle")
#library(rattle)
#fancyRpartPlot(ct)

############################
# Optimal hyperparameters #
############################

# Test number of trees inside the RF
inicio <- Sys.time()
ranfor.z <- randomForest(y ~ .,
                         data=dgt2,
                         importance=TRUE,
                         replace=TRUE,
                         keep.forest=FALSE,
                         #na.action=na.omit,
                         mtry = 6,     # number of predictors sampled randomly
                         do.trace=50,  # Best way to trace testing RF parameters
                         ntree=500)
ranfor.z
(duracion <- Sys.time()-inicio) # Check computational timing
# Best ntree = 200

# Test how many predictors should be sampled for splitting at each node
# ***WARNING: TAKES A REALLY LONG OF TIME TO COMPUTE!!***
tuneRF(x = dgt2[,2:ncol(dgt2)],   # Training dataset
       y = dgt2$y,      # Objective variable to predict
       mtryStart = 1,   # Initial variables quantity
       stepFactor = 2,  # Variables increment
       ntreeTry = 200,   # Number of tree each trial
       improve = 0.001   # Minimum OBB gain to continue to next trial
)
# Best mtry = 4

##############################
# Folds for cross-validation #
##############################
set.seed(1) #because of the sample function ahead
dgt3 <- dgt2[sample(nrow(dgt2)),]
dgt3 <- dgt3[, -6] #Drop "COMUNIDAD AUTONOMA" (hyperparameters doesn't change)
```

```
dgt3 <- na.omit(dgt3)
folds <- cut(seq(1, nrow(dgt3)), breaks=5, labels=FALSE)

##################
# RANDOM FOREST #
##################
#i <- 1
#accuracies <- c()
inicio <- Sys.time()
for (i in 1:5){
  testIndexes <- which(folds == i, arr.ind=TRUE)
  testData <- dgt3[testIndexes, ]
  trainData <- dgt3[-testIndexes, ]

  set.seed(2)
  ranfor <- randomForest(y ~ .,
                         xtest=testData[,2:ncol(testData)],
                         ytest=testData$y,
                         data=trainData,
                         importance=TRUE,
                         replace=TRUE,
                         keep.forest=TRUE,
                         mtry = 4,    # number of predictors sampled randomly
                         ntree=200)
  print(ranfor)
  print(duracion <- Sys.time()-inicio) # Check computational timing

  assign(paste0("ranfor_", i), ranfor)
  assign(paste0("dur_", i), duracion)

  #prediction <- predict(ranfor, testData)
  #testData$acierto <- prediction == testData$y
  #t <- table(prediction, testData$y)
  #accuracy <- sum(testData$acierto)/nrow(testData)
  #accuracies <- c(accuracies, accuracy)
  #print(accuracy)
}

summary(ranfor_1)
str(ranfor_1)

# Conditional inference Trees
#install.packages("party")
#library(party)
#cforest(y ~., data=dgt3, controls=cforest_unbiased())

windows()
plotmo(ranfor_1, ngrid1=52, ngrid2=52)
windows()
plotmo(ranfor_1, degree1=NA, degree2=c("VISIBILIDAD_RESTRINGIDA","PROVINCIA"),
       ngrid1=52, ngrid2=52, persp.theta=-35)

table(dgt2$PROVINCIA, dgt2$VISIBILIDAD_RESTRINGIDA)

# PASAR ESTO AL EXCEL Y HACER "HEAT MAPPING"
table(dgt2$PROVINCIA[dgt2$y=="Leve"], dgt2$VISIBILIDAD_RESTRINGIDA[dgt2$y=="Leve"])
table(dgt2$PROVINCIA[dgt2$y=="Grave"], dgt2$VISIBILIDAD_RESTRINGIDA[dgt2$y=="Grave"])

windows()
plotmo(ranfor_1, degree1=NA, degree2=c("TIPO_ACCIDENTE","EDAD"),
       ngrid1=52, ngrid2=52, persp.theta=-35)

windows()
plot(ranfor_1, main="Random Forest")
importance(ranfor_1)
windows()
varImpPlot(ranfor_1, type=2,  main="Random Forest Variable Importance")
```

```
windows()
plot(randomForest::margin(ranfor_1), main="Random Forest Margin")


windows()
ggplot(dgt2, aes(x=ANIO_PERMISO, y=EDAD, shape=y, color=y)) +
       geom_point()

################
# GLM - Death #
################
dead <- dead[, -6] #Drop "COMUNIDAD AUTONOMA" (hyperparameters doesn't change)
dead1 <- dead
any(is.na(dead1))
dead1 <- na.omit(dead1)
any(is.na(dead1))

str(dead1)
summary(dead1)
sort(unique(dead1$yy))
table(dead1$yy)

# Check if any factor with levels == 1 needs to be dropped out
(hello <- sapply(dead1, function(x) is.factor(x)))
world <- dead1[,hello]
ifelse(basura <- sapply(world, function(x) length(levels(x))) == 1, "DROP", "NODROP")

dat <- na.omit(dead1)
fctores <- lapply(dat[sapply(dat, is.factor)], droplevels)
sapply(fctores, nlevels)
which(sapply(dat, function(x) { length(unique(x)) }) == 1)
# INFRACC_PEATON IS A LIAR ! DROP HIM!
grep("INFRACC_PEATON", colnames(dead1))
dead1 <- dead1[,-(grep("INFRACC_PEATON", colnames(dead1)))]
is.null(dead1$INFRACC_PEATON)

# The row with the dead$MANIOBRAS == 61 has to be erased because theres only 1 obs.
table(dead1$MANIOBRAS)
dead1 <- dead1[!(dead1$MANIOBRAS==61), ]
table(dead1$MANIOBRAS)
# Same for TIPO_VEHICULO==11
table(dead1$TIPO_VEHICULO)
dead1 <- dead1[!(dead1$TIPO_VEHICULO==11), ]
table(dead1$TIPO_VEHICULO)
str(dead1)

#############################################################
# GLM with "caret" package and 10-fold cross-validation #
#############################################################
dead2 <- dead1 # BACKUP dataset
dead.value <- ifelse(dead2$yy == "Vivo", 1, 0)
tc <- trainControl("cv", 10, savePredictions = T)
inicio <- Sys.time()
fit.dead2 <- train(yy ~., data=dead2, method = "glm", family=binomial, trControl=tc)
(duracion <- Sys.time()-inicio)

# Deviance test: values comes from null and residual deviance
summary(fit.dead2)
1-pchisq(14328.9-7782.9, df=(97751-97521))
1-pchisq(fit.dead2$finalModel$null.deviance-fit.dead2$finalModel$deviance,
df=(fit.dead2$finalModel$df.null-fit.dead2$finalModel$df.residual))

#head(fit.dead2)
pred.dead2 <- fit.dead2$finalModel$fitted.values
pred.dead2t <- function(t) ifelse(pred.dead2 > t, 1, 0)
confusionMatrix(pred.dead2t(0.95), dead.value)
```

```
addmargins(table(pred.dead2t(0.95), dead.value, dnn=c("Predicción", "Actual")))

# Cross-validation confusion matrix and errors
cTab <- table(pred.dead2t(0.95), dead.value, dnn=c("Predicción", "Actual"))
addmargins(cTab)
err.dead2 <- (cTab[2] / (cTab[1] + cTab[2]))
err.alive2 <- (cTab[3] / (cTab[3] + cTab[4]))
err.dead <- (cTab[3] / (cTab[1] + cTab[3]))
err.alive <- (cTab[2] / (cTab[2] + cTab[4]))
err.pread2 <- err.dead2 * (((cTab[1] + cTab[2]))/sum(cTab)) + err.alive2 * (((cTab[3] + cTab[4]))/sum(cTab))
err.pread <- err.dead * (((cTab[1] + cTab[3]))/sum(cTab)) + err.alive * (((cTab[2] + cTab[4]))/sum(cTab))
print(paste0("Error Clase  Muerto: ", round(err.dead2*100, 2), "%")) # THE IMPORTANT ONE!
print(paste0("Error Clase  Vivo: ", round(err.alive2*100, 2), "%")) # THE IMPORTANT ONE!
print(paste0("Error Predicción Muerto: ", round(err.dead*100, 2), "%"))
print(paste0("Error Predicción Vivo: ", round(err.alive*100, 2), "%"))
print(paste0("Error Predicción promedio ponderado: ", round(err.pread*100, 2), "%"))
print(paste0("Error Clase promedio ponderado: ", round(err.pread2*100, 2), "%"))

# ROC Curve
pr1 <- prediction(pred.dead2, dead2$yy)
prf1 <- performance(pr1, measure = "tpr", x.measure = "fpr")
windows()
plot(prf1, main="Curva ROC")
# AUC from ROC Curve
auc <- performance(pr1, measure="auc")
(auc <- auc@y.values[[1]])

#####################################
# GLM with a 20% Cross validation # --> (10-fold CV is a better estimate)
#####################################
dead2 <- dead1 # BACKUP dataset
set.seed(1) #because of the sample function ahead
dead2 <- dead2[sample(nrow(dead2)), ]
trainIndex <- round(0.8 * nrow(dead2))
trainData2 <- dead2[1:trainIndex, ]
testData2 <- dead2[(trainIndex+1):nrow(dead2), ]

# GLM
inicio <- Sys.time()
fit.dead <- glm(yy~., family=binomial(link="logit"),
                data=trainData2, na.action=na.omit)
print(duracion <- Sys.time()-inicio)
fit.dead
summary(fit.dead)

### Prediction ###
pred.dead <- predict.glm(fit.dead, testData2, type="response")
str(pred.dead)
summary(pred.dead)

pred.dead.results <- ifelse(pred.dead > 0.5, "Vivo", "Muerto")
misclass.error <- mean(pred.dead.results != testData2$yy)
print(paste0("Accuracy: ", round((1-misclass.error)*100, digits=2), "%"))

#anova(fit.dead, test="Chisq")
pr <- prediction(pred.dead, testData2$yy)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
windows()
plot(prf, main="Curva ROC")

auc <- performance(pr, measure="auc")
(auc <- auc@y.values[[1]])

# Backtesting with prediction error
(prop <- 1 - sum(dead2$yy=="Muerto")/nrow(dead2))
thresh <- prop
predFac <- cut(pred.dead, breaks=c(-Inf, thresh, Inf), labels=c("Muerto", "Vivo"))
```

```
cTab <- table(predFac, testData2$yy, dnn=c("Predicción", "Actual"))
addmargins(cTab)
err.dead <- (cTab[3] / (cTab[1] + cTab[3]))
err.alive <- (cTab[2] / (cTab[2] + cTab[4]))
err.pread <- err.dead * (((cTab[1] + cTab[3]))/sum(cTab)) + err.alive * (((cTab[2] + cTab[4]))/sum(cTab))
print(paste0("Error clase Muerto: ", round(err.dead*100, 2), "%"))
print(paste0("Error clase Vivo: ", round(err.alive*100, 2), "%"))
print(paste0("Error predicción promedio ponderado: ", round(err.pread*100, 2), "%"))

# McFadden pseudo R-squared to asses model fitness value
#install.packages("pscl")
#library(pscl)
#pR2(fit.dead)
#install.packages("DescTools")
#library(DescTools)
#PseudoR2(fit.dead, which="all")
# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
########################
# Step-wise logit GLM #   WAY TOO MUCH TO COMPUTE...
########################
#inicio <- Sys.time()
#fit.dead <- glm(yy~., family=binomial(link="logit"),
#               data=dead2, na.action=na.omit)
#duracion <- Sys.time()-inicio

#backwards <- step(fit.dead)
#formula(backwards)
#summary(backwards)

##############################
# MODELOS DE SUPERVIVENCIA #
##############################
#library(survival)
#survobj <- with(dead2, Surv(MES, yy))
#test1 <- survfit(survobj~1, data=dead)
#summary(test1)
#plot(test1, xlab="Supervivencia en meses",
#     ylab="% Supervivientes", main="Distribución de supervivencia")

# Estimate permutation p-values for importance metrics
#install.packages("rfPermute")
#library(rfPermute)
#inicio <- Sys.time()
#rfP <- rfPermute(y ~ .,
#            data=dgt3[1:20000, ],
#            importance=TRUE,
#            replace=TRUE,
#            mtry = 6,     # number of predictors sampled randomly
#            ntree=200,
#                        nrep=100)
#rfP
#(duracion <- Sys.time()-inicio) # Check computational timing
#rp.importance(rfP)
#rfP$pval
#plot(rfP, imp.type=1)
#str(rfP)
#plotmo(rfP, ngrid1=52, ngrid2=52)

#table(external$VISIBILIDAD_RESTRINGIDA)
#table(external$VISIBILIDAD_RESTRINGIDA,y)
#table(external$COMUNIDAD_AUTONOMA)
#table(external$COMUNIDAD_AUTONOMA,y)
#table(dgt2$VISIBILIDAD_RESTRINGIDA, dgt2$COMUNIDAD_AUTONOMA)

#########
# RF v2 # Random Forest using 'CARET' package
#########
```

```
#install.packages("caret")
#library(caret)
#install.packages("e1071")
#library(e1071)

#set.seed(1)
#Folds <- createMultiFolds(y=dgt2$y, k=5, times=5)

#rfControl <- trainControl(method = "repeatedcv",
#           number = 10, repeats = 10, index = Folds,
#           classProbs = TRUE,
#           allowParallel = TRUE,
#           selectionFunction = "oneSE",
#           returnResamp = "final")

#rfGrid <- expand.grid(mtry = seq(2, 16, by=2))

#rfFit <- train(y ~., data = dgt2,
#           method = "rf",
#           importance = TRUE, ntree = 200,
#           trControl = rfControl, tuneGrid = rfGrid,
#           metric = "Kappa", maximize = TRUE)

#rfPred <- predict.train(rfFit, data, type="raw")

# As TOT_MUERTOS is mostly zeros, let's check the prediction with other than zero values
#testData.nozero <- testData[testData$y != 0, ]
#prediction2 <- predict(ranfor, testData.nozero)
#testData.nozero$acierto <- prediction2 == testData.nozero$y
#t2 <- table(prediction2, testData.nozero$y)
#(accuracy2 <- sum(testData.nozero$acierto)/nrow(testData.nozero))

#MDSplot(ranfor, testData$y)

################################
# Conditional inference trees #
################################
# Install package
#install.packages("party")
#library(party)
# Run Conditional inference tree
#set.seed(1)
#fit <- cforest(y ~.,
#             data = trainData,
#             controls = cforest_unbiased(ntree = 200, mtry = 4))
#predict_cf <- predict(fit, testData, OOB = TRUE, type = "response")

################################
# Quantile Regression Forest #
################################
# Tree-based ensemble method for estimation of conditional quantiles
#https://cran.r-project.org/web/packages/quantregForest/quantregForest.pdf

#-----------------------------------------------------------------------
######################
# Parallel Computing #
######################
#inicio <- Sys.time()
#################################
# Parallel backend regist
#cl <- makeCluster(detectCores())
#registerDoParallel(cl)
#getDoParWorkers()
#################################
#rf <- foreach(ntree=150,
#            # .combine = combine,
#             .multicombine = TRUE,
```

```
#              .packages = 'randomForest') %dopar%
#  randomForest(y ~., data=trainData, r=T, importance=T, mtry=20)
###################
# Stop Cluster
#stopCluster(cl)
#print(duracion <- Sys.time()-inicio)
###################
#-----------------------------------------------------------------
```