FACULTAD DE ESTUDIOS ESTADÍSTICOS

MASTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2014/2015

Trabajo de Fin de Master

TITULO: METODOLOGÍA DE MINERÍA DE DATOS PARA EL ESTUDIO DE TABLAS DE SINIESTRALIDAD VÍAL

***Alumno*: Guillermo Villarino Martínez**

***Tutor*: Rosario Cintas del Río/Jose Luis Brita-Paja Segoviano**

Noviembre de 2015

UNIVERSIDAD COMPLUTENSE
MADRID

1. INTRODUCCIÓN

La siniestralidad en las carreteras ha sido, desde la generalización del uso de vehículos a motor, una de las principales causas de muerte en España y por ello es foco de gran preocupación para la sociedad y sus autoridades.

Han pasado muchos años desde el máximo histórico de **fallecidos en accidentes de tráfico en España** de 1989. Aquel aciago año, último de la década de los ochenta y en pleno aumento del parque de vehículos automóviles, **9.344 personas** perdieron su vida en un accidente de tráfico. Entonces no se llegaba a 15 millones de vehículos en total.

Debido a los esfuerzos realizados y a la gran cantidad de recursos destinados, estas elevadas cifras han experimentado, afortunadamente, un **descenso** muy significativo. Entre las causas de esta disminución se encuentran la mayor **concienciación** de la sociedad en materia vial, la **mejora de las infraestructuras** de la red de transportes, los grandes **avances** en materia de **seguridad de los vehículos** automóviles y de **detección de infracciones** mediante cinemómetros y cámaras de seguridad.

España se encuentra entre los 10 países con menor siniestralidad de la Unión Europea, y en concreto ocupa la 7ª posición. La sociedad debe ser consciente de que reducir las cifras actuales no es una tarea sencilla y que para ello es necesario impulsar políticas eficaces de seguridad vial, basadas en la **evidencia científica**; políticas que tengan en consideración los diferentes sectores implicados y que comprometan de manera efectiva a estos sectores, tanto públicos como privados, en la reducción de las cifras de siniestralidad vial. Las **1.903 víctimas mortales y los 10.444 heridos graves**, según los informes policiales, ocasionados en vías urbanas e interurbanas son motivación más que suficiente para mejorar las estrategias que nos lleven a erradicar este grave problema de salud como son las lesiones por accidente de tráfico[1].

[1] Las principales cifras de la siniestralidad vial en España 2012. Dirección General de Tráfico (DGT)

2. DATOS OFICIALES DE SINIESTRALIDAD VIAL EN 2012[2]

Los datos oficiales de siniestralidad se recogen en el informe citado a pie de página que es realizado por la Dirección General de Tráfico cada año y en el que se presenta un estudio pormenorizado, a nivel descriptivo, de los factores en la accidentalidad vial. Se destacan, a continuación las conclusiones más relevantes, no sin antes introducir una serie de definiciones necesarias para la mejor comprensión de los datos a tratar durante el estudio.

Definiciones:

- Se consideran **accidentes de tráfico con víctimas** los que se producen, o tienen su origen en una de las vías o terrenos objeto de la legislación sobre tráfico, circulación de vehículos a motor y seguridad vial y a consecuencia de los mismos una o varias personas resultan muertas y/o heridas.
- Se considera **víctima mortal**, toda persona que, como consecuencia del accidente, fallezca en el acto o dentro de los treinta días siguientes.
- Se consideran **heridos graves**, aquellas personas heridas en un accidente de circulación y cuyo estado precise una hospitalización superior a veinticuatro horas.
- Se consideran **heridos leves,** aquellas personas heridas en un accidente de circulación a los que no puede aplicarse la definición de herido grave.

El cómputo de muertos se realiza a 30 días:

- Desde 1993 a 2010, el cómputo de muertos se realiza a 30 días como resultado de la **aplicación de los factores correctores** deducidos del seguimiento real de una muestra representativa de heridos graves.
- Para los años 2011 y 2012 el cómputo de muertos se realiza a 30 días según la metodología que se especifica en el Anexo de esta publicación[3].

[2] Las principales cifras de la siniestralidad vial en España 2012. Dirección General de Tráfico (DGT)
[3] Series Estadísticas 1993-2012: Accidentes con víctimas, nuestros computados a 30 días, heridos graves, heridos leves, heridos y víctimas. DGT (2012)

La mayoría de los accidentes de tráfico que se producen anualmente en nuestro país ocasionan únicamente daños materiales originando importantes pérdidas económicas. Sin embargo, por su trascendencia para la salud de la población lo que resulta fundamental es conocer el número de accidentes con alguna víctima, las características en relación a la gravedad de las lesiones y los factores que desencadenan el accidente.

Durante el año 2012, las diferentes policías notificaron 83.115 ***accidentes con víctimas***. Según los informes policiales, estos accidentes ocasionaron 1.903 fallecidos en el momento del accidente o hasta 30 días después del mismo, 10.444 personas fueron ingresadas en un centro hospitalario y 105.446 resultaron heridos leves. Estas cifras, aun siendo elevadas, han supuesto una reducción con respecto al año anterior, a pesar de que el número de accidentes ha permanecido estable y el número de heridos leves ha aumentado.

El parque de automóviles ha crecido casi 6 millones en el último decenio en todas las categorías de vehículos, y los turismos representan el 67 % del mismo. No obstante, en 2012 se observa por primera vez en los últimos diez años un descenso en la cifra total del parque respecto del año anterior. La antigüedad media del parque de automóviles con menos de 25 años oscila entre los 8,3 años de los tractores industriales y los 10,7 de los camiones y furgonetas, siendo la antigüedad media para los turismos de 9,5 años. La mitad de los turismos tienen 10 ó más años.

El censo de conductores ha aumentado un 1 % en 2012 comparado con el año anterior, registrando 672 conductores por 1.000 habitantes con edad habilitada para conducir. Se observa un envejecimiento en los últimos años, pasando del 24 % de conductores con una antigüedad del permiso inferior a cinco años en 2008 a un 19 % en 2012.

Respecto a ***los fallecidos***, destacar que el 76 % eran varones, el 51 % tenían 45 años de edad o más, el 46 % estuvieron implicados en un accidente como ocupantes de un turismo, el 76 % tuvo un accidente en vía interurbana y en concreto, un 79 % de estos accidentes se produjo en vías secundarias. El 61 % de los fallecidos eran conductores y el 20 % peatones. El 65 % de los accidentes donde falleció al menos una persona sucedieron en días laborables y, en un 63 % de éstos, el accidente fue entre las 8 de la

mañana y las 8 de la tarde. El 35 % de las víctimas mortales fallecieron en un accidente por una salida de la vía. En 2012, los fallecidos en accidentes de tráfico en España se distribuyen de manera uniforme a lo largo de los días, semanas y meses. El **número medio diario de fallecidos fue de 5,2**, concretamente 3,9 fallecidos en vías interurbanas y 1,2 en urbanas.

El número total de **víctimas mortales en 2012** con respecto al año anterior ha **descendido un 8 %.** En la mayoría de los distintos tipos de vehículos se observa un descenso, salvo en los usuarios de bicicletas que han aumentado un 47 %. También hay más fallecidos en autopistas y vías urbanas. Por edades aumentan los fallecidos de 0 a 14 años y los mayores de 75.

Según ***el lugar del accidente***, el número de fallecidos ha descendido en la mayoría de las comunidades autónomas, salvo en Aragón y el Principado de Asturias donde se mantienen las cifras y en Illes Balears, Canarias, Cataluña y Comunidad Foral de Navarra donde hubo un aumento. Los fallecidos en vías urbanas han aumentado respecto al año 2011, produciéndose 4 fallecimientos más en 2012.

Los ***usuarios de bicicletas*** se vieron implicados en 5.150 accidentes en los que **fallecieron 72 ciclistas** y resultaron heridos graves 572. El 72 % de los accidentes tuvieron lugar en vías urbanas, resultando heridos leves 4.362 ciclistas, el 73 % del total de heridos leves. En vías interurbanas se produjo el mayor número de fallecidos con 52 ciclistas.

3. ESTADO DEL ARTE

En este apartado se pretende hacer una recopilación de información relevante para el estudio que ha sido publicada en los últimos años, con el objetivo de consolidar un conocimiento previo sobre la actualidad en materia de técnicas de minería de datos para la clasificación de clases poco representadas y anteriores estudios sobre siniestralidad vial estableciendo así líneas de investigación.

La **minería de datos** se considera la principal herramienta para la extracción de conclusiones sobre grandes bases de datos, y se han desarrollado multitud de algoritmos de aprendizaje estadístico y computacional como los árboles de clasificación (Breiman, 1984) y métodos de mayor complejidad basados en ellos como bagging (Breiman, 1996) y random forest (Breiman, 2001). Se recogen muchos de estos algoritmos en (Hastie, 2009).

Un serio problema que afecta a los datos de siniestralidad es la baja prevalencia de fallecidos en la población.

En un caso ideal todos los datos pertenecientes a cada clase se encuentran agrupados entre ellos y claramente diferenciables del resto de clases. La realidad es bien distinta y con frecuencia los datos presentan diferentes problemas que dificultan la labor de los clasificadores y disminuyen la calidad de la clasificación realizada.

Uno de los estos problemas es el **desbalanceo de clases**, que ocurre cuando el número de instancias de cada categoría de la variable a clasificar es muy diferente. En estas circunstancias los clasificadores presentan una tendencia de clasificación hacia la clase mayoritaria, minimizando de ésta manera el error de clasificación y clasificando correctamente instancias de clase mayoritaria en detrimento de instancias de la clase minoritaria.

Se han propuesto multitud de soluciones a este problema dado la gran relevancia que los sucesos 'raros' tienen en muchos campos de aplicación, como la utilización de distintos tipos de métricas para la evaluación de los resultados en este tipo de datos (García V., 2010), el rebalanceo de clases por medio de diversos algoritmos como SMOTE (Wang J., 2006) y sus variaciones. Se apunta, así mismo, a las técnicas de remuestreo múltiple como medio de entrenamiento de los algoritmos de aprendizaje (Japkowicz, 2004) que minimiza la tasa de fallos en clasificación, y a métodos de entrenamiento de modelos basados en costes (Domingos P., 1999) que penalizan en mayor medida la incorrecta clasificación de los sucesos de la clase minoritaria de interés haciendo que el algoritmo se 'esfuerce' en reconocer los patrones que subyacen en éstas.

En lo que se refiere a estudios sobre siniestralidad vial publicados, destaca la utilización de modelos de regresión logística (Ali S., 2002) y de métodos no paramétricos como

árboles de clasificación (Li-Yen C. 2006), obteniéndose buenos resultados mediante la combinación de ambas técnicas en (Petra M, 2000 y Serna S, 2009). Se aplica la aproximación mediante redes bayesianas en (Oña J. 2010).

4. OBJETIVOS

La fuente de información del presente estudio la constituye la base de datos de accidentes ocurridos en el año 2012 en España, proporcionada por la Dirección General de Tráfico (DGT). Dicha base de datos está integrada por 83.115 registros que se corresponden con los accidentes con víctimas notificados por los diferentes cuerpos de policía y guardia civil y 202.804 registros referentes a datos específicos de las propias víctimas.

El objetivo principal de este trabajo es la creación de una metodología para el estudio de una base de datos de siniestralidad vial a partir de un procedimiento semi-automático para facilitar el tratamiento de las tablas de datos de siniestralidad vial que, con periodicidad anual, son recogidas por la Dirección General de Tráfico. Esta metodología se puede dividir en las cuatro etapas secuenciales siguientes:

- Preprocesamiento de los datos
  - Proponer un procedimiento para el estudio y recodificación de las variables que históricamente se consideran de influencia en la siniestralidad vial.
  - Determinar las subpoblaciones objetivo mediante estudio de la posible segmentación por tipo de vehículo (bicicletas, motos, camiones, turismos, autobuses..) o por tipo de víctima (peatón, conductor, pasajero..) .
- Determinación de factores de riesgo
  - Identificar y evaluar los factores de riesgo influyentes en los accidentes con víctimas mortales en las distintas subpoblaciones.
  - Determinar perfiles de víctimas y escenarios de accidentalidad como resultado de la integración de distintas técnicas de clasificación.

- ❖ Abordar el problema de clasificación de las clases poco representadas.

- Modelos de clasificación en minería de datos
  - ❖ Crear funciones para facilitar el ajuste de algoritmos de aprendizaje a los datos y valoración de los resultados.
  - ❖ Establecer una comparativa de bondad de ajuste entre distintos métodos de clasificación en minería de datos tales como Random Forest, Gradient Boosting y Redes Neuronales, en la clasificación de los fallecidos en las distintas subpoblaciones de víctimas.

En definitiva se pretende programar un procedimiento mediante el cual sea posible extraer conclusiones de los datos de forma semiautomática y con la máxima fiabilidad posible. Hay que observar en este punto que se han utilizado únicamente los datos del año 2012 por lo que quedaría abierta la línea de investigación para el tratamiento de datos de naturaleza longitudinal.

En cualquier caso, se considera que esta metodología proporciona una base sólida para obtener buenos resultados en los conjuntos de datos de siniestralidad vial recogidos de esta forma.

5. METODOLOGÍA

Para alcanzar los objetivos planteados se recurre a una metodología que incluye muchos aspectos del tratamiento de datos. Por una parte es fundamental establecer un método para el **preprocesamiento** de las variables disponibles en el conjunto de datos y en este sentido se recogen algunas posibles actuaciones para tal fin.

En una segunda fase se utilizan distintas **técnicas de clasificación** que se comentan en este punto a fin de comprender sus posibles fortalezas y debilidades para una correcta monitorización e interpretación.

Finalmente, resulta interesante integrar los resultados obtenidos con las anteriores técnicas, obteniendo nuevos clasificadores mediante distintos métodos de **ensamble de modelos**.

En última instancia y ya que se trata de una metodología construida en **lenguaje R**, es pertinente comentar las herramientas utilizadas y evaluar su capacidad.

5.1 PREPROCESAMIENTO

Como bien es sabido, todo análisis estadístico debe ir precedido de una imprescindible etapa de depuración de los datos, cuyo objetivo es minimizar el 'ruido' que ciertas distribuciones de variables pueden introducir en los modelos, con la consecuente pérdida de precisión e incluso la posible obtención de conclusiones erróneas.

Debido a esto se realiza una cuidadosa labor de examen y depuración de las variables que resultarán de interés a lo largo del estudio y cuyas distribuciones se presentarán en el apartado correspondiente.

El proceso de estudio y depuración de los datos ha sido exhaustivo y pormenorizado, por lo que se explicarán los métodos básicos utilizados para la recategorización de las variables, pues pueden resultar de interés general y se expondrán los casos de mayor relevancia.

En cuanto a los métodos de **recodificación de variables**, por un lado y ya que el objetivo fundamental es preparar el conjunto de datos y sus variables para un análisis de regresión logística o clasificación binaria en general, se han llevado a cabo uniones de niveles de variables de naturaleza categórica mediante el examen de los resultados de **modelos de regresión logística univariante** en los que la variable respuesta es el suceso de interés de este estudio, la variable *muerte a 30 días*[4] de naturaleza dicotómica, y la variable explicativa es la variable a recodificar. La metodología es la que sigue:

- Ajuste de un modelo de regresión logística con la variable a recodificar como variable explicativa (con categoría de referencia adecuada en cada caso) y la variable de interés como respuesta (muerte a 30 días).
- Examen de la significación de los parámetros para todas las categorías en relación a la referencia. Categorías que no resultan significativas son susceptibles de unión.

[4] La variable *muerte a 30 días* refleja el hecho constatado del fallecimiento del siniestrado en accidente de tráfico pasados 30 días del mismo.

- Comprobación del sentido de cambio de la probabilidad estimada en las categorías a unir, es decir, el signo de los coeficientes estimados. Si ambos son del mismo signo unir categorías.
- Si los resultados no son completamente satisfactorios estamos ante dos posibles situaciones, la primera es que la variable no sea significativa en sí misma para explicar el suceso de interés, en cuyo caso la desecharemos del estudio, y la segunda es que la referencia no es correcta, por lo que se probará con otras posibles categorías como referencia.

El otro procedimiento de **recategorización** de variables utilizado han sido los **Arboles de clasificación CHAID**[5] (**CH**i-squared **A**utomatic **I**nteraction **D**etection), método de segmentación o clasificación no paramétrico, jerárquico de tipo descendente y divisivo que tiene por objetivo la división de la población en las categorías de la variable considerada como respuesta por medio de los valores de las variables predictoras o independientes, poniendo de manifiesto de forma rápida las relaciones significativas entre las variables consideradas.

5.2 TÉCNICAS DE CLASIFICACIÓN EMPLEADAS

En lo que se refiere a modelos de clasificación, se presta especial atención a la regresión logística como la técnica clásica debido a su base estadística y a la posibilidad de cuantificar el efecto de las variables sobre la respuesta mediante los odds ratio, frente a la metodología utilizada por otros algoritmos de aprendizaje estadístico y computacional. A continuación se enumeran las distintas técnicas utilizadas:

La **Regresión Logística** es una técnica estadística multivariante de clasificación cuyo objetivo es predecir los valores de una variable respuesta dicotómica o categórica (utilizando el modelo logit o probit multinomial) en función de una serie de variables predictoras que pueden ser de naturaleza categórica o continua. El resultado de esta técnica es la construcción de una función que, a través de una transformación no lineal de tipo exponencial (logit), predice la probabilidad de que la variable respuesta tome el valor considerado como referencia para cada combinación de los valores de las

[5] La metodología CHAID fue desarrollada en el año 1980 en Sudáfrica por Gordon V. Kass en su tesis PhD.

variables independientes del modelo. La estimación de parámetros se realiza por máxima verosimilitud y la distribución de los errores es binomial a diferencia de la regresión lineal. La gran ventaja de ésta técnica es la posibilidad de cuantificar los efectos de los predictores sobre la respuesta a través de los *Odds Ratio*.

Las **Redes Neuronales** aparecieron por primera vez en los años 50 como intento de elaborar una herramienta capaz de resolver de problemas de forma análoga a como lo haría un cerebro animal. Una de las grandes ventajas es que la red neuronal es capaz de elaborar sus propias "reglas" a fin de hallar la mejor respuesta a una entrada determinada.

El cuerpo de la neurona se representa como un sumador lineal de los estímulos externos z_j, seguida de una función no lineal $y_j = f(z)_j$ que se denomina función de activación y es la que utiliza la suma de estímulos para determinar la actividad de salida de la neurona.

A partir de este se puede llegar a elaborar un perceptrón multicapa. En el cual se consideran las capas de datos Input (entradas, estímulos), capa de nodos ocultos (puede ser más de una) y finalmente la capa de salida. Para utilizar una red neuronal con garantías se requieren muchas observaciones (al no haber inferencia propiamente dicha se necesitan datos test para validar el modelo).

El entrenamiento de una red neuronal, tiene por objetivo modificar iterativamente los pesos de la red con el fin de minimizar el error entre la predicción y la respuesta esperada. Dentro de los parámetros que definen una red, la función de red más utilizada es de tipo lineal, y como función de activación más empleada está la función sigmoidea.

Random forest es una técnica mejorada de Bagging que mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual. Esta aleatorización puede introducirse en la partición del espacio (construcción del árbol), así como en la muestra de entrenamiento.

El algoritmo Random Forest, a diferencia de Bagging introduce de forma aleatoria en cada nodo p variables de todas la originales, y de estas selecciona la mejor para realizar la partición. Se presenta a continuación el proceso del algoritmo:

- Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes set de datos.

- Al crear los arboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (sin podar).
- Crea un árbol de decisión con cada set de datos, obteniendo diferentes arboles, ya que cada set contiene diferentes individuos y diferentes variables.
- Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.

Gradient Boosting (Friedman H., 2001) se basa en la idea de entrenar el algoritmo mediante la actualización de los pesos de las observaciones pertenecientes a las clases del suceso de interés a través de la optimización en dirección descendente de una función de pérdida o error determinada, consiguiendo dar mayor relevancia en cada iteración a las observaciones mal clasificadas en pasos anteriores.

En todas las técnicas de minería de datos se ha aplicado el método de **validación cruzada repetida** que consiste en dividir el archivo en n partes construyendo el modelo con n-1 de ellas y reservando la restante para la validación de los resultados obtenidos, con lo que finalmente para cada repetición del algoritmo se ajustarán n modelos validados sobre observaciones 'nuevas'. Este proceso se repite m veces consiguiendo por lo tanto m x n modelos ajustados con conjuntos de entrenamiento distinto y validado sobre observaciones no utilizadas en su construcción. Se ha elegido n = 3 y m = 4 en este estudio.

5.3 ENSAMBLE DE MODELOS

Con el objetivo de mejorar la precisión alcanzada por los modelos de clasificación empleados en el estudio y reducir la varianza de los errores cometidos, se proponen distintos métodos de ensamble de clasificadores mediante la técnica de *stacking*.

Este método consiste en construir clasificadores dados por la combinación, lineal o no, de las probabilidades estimadas por los modelos ajustados, algunos de los cuales son ensambles en sí mismos (Random Forest, Gradient Boosting). Con ello se consiguen las probabilidades estimadas conjuntas y se realiza la clasificación mediante la técnica del punto de corte óptimo de la probabilidad estimada.

Cabe destacar que para obtener mejores resultados es conveniente realizar un estudio de correlaciones entre las predicciones para descartar los ensambles de probabilidades estimadas altamente correladas que usualmente no proporcionan mejora respecto al mejor modelo (Dzeroski, S., 2004).

5.4 LA AUTOMATIZACIÓN. LENGUAJE R

Se puede decir que el lenguaje R es un dialecto libre del lenguaje S, desarrollado por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en 1993. R es un software estadístico de código abierto con gran capacidad de análisis y una amplia comunidad de usuarios y colaboradores que lo ha llevado a ser uno de los entornos estadísticos más utilizados.

Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R y tiene un constante crecimiento gracias a los paquetes desarrollados por su comunidad de usuarios.

El **paquete caret**[6] ("**c**lassification **a**nd **re**gression **t**raining") de R contiene funciones para la construcción y evaluación de modelos predictivos de una forma simple, con un interfaz común que llama a los distintos paquetes de técnicas supervisadas de clasificación y regresión de R con los que realmente se construyen los modelos.

El paquete caret contiene herramientas para la **construcción de los modelos**, la **optimización de los parámetros** usando técnicas de remuestreo (validación cruzada, bootstrapping..) y la **predicción** de la variable respuesta en nuevas observaciones. Además, incorpora funciones para el pre-procesamiento de los datos, **cálculo de la importancia de las variables** en los modelos ajustados y visualizaciones adecuadas a cada técnica.

Por tanto, este paquete de R constituye una base fundamental para la construcción de la metodología propuesta. Se programarán los procedimientos para automatizar el estudio en la medida de lo posible con el fin de facilitar un posible estudio de nuevos datos de esta naturaleza.

[6] https://cran.r-project.org/web/packages/caret/caret.pdf

6. PREPROCESAMIENTO DE LOS DATOS

A partir de los datos proporcionados por la DGT, se ha construido un único fichero que recoge la información que resulta de interés para el estudio y que se ha denominado *Tabla Personas-General-Vehículos.* El archivo se compone de 202804 observaciones y 106 variables, de las cuales se seleccionan 32 por su capacidad informativa sobre el suceso de interés y por ser los factores que históricamente se han considerado de influencia en la siniestralidad vial en los distintos estudios citados y en fuentes oficiales.

El análisis descriptivo se estructura en tres partes, una primera en la que se realiza una recodificación de algunas de las variables, una segunda en la que se presenta el estudio univariante básico de las variables seleccionadas, ya recategorizadas, mediante tablas de frecuencias y gráficos fundamentalmente, y una segunda parte en la que se muestran los cruces de variables que resultan de mayor interés.

6.2 RECODIFICACIÓN DE VARIABLES

Mediante el procedimiento iterativo anterior se ha llevado a cabo la recodificación de la mayoría de variables de naturaleza categórica presentes en el archivo de interés como zona, superficie, tipo de accidente, tipo de vía, infracción del conductor y factores atmosféricos. La mayoría de estas variables contenían categorías muy poco representadas y con escaso valor predictivo para la clasificación el suceso de interés.

En particular y como caso claro de la necesidad de aplicación de este estudio previo de depuración, la variable infracción del conductor constaba originalmente de 23 niveles, lo que hace muy difícil su estudio e interpretación en cualquier modelo de regresión, ya que no puede ser tratada como categorías ordenadas por ausencia de cualquier componente ordinal. Esta misma situación se da en la variable tipo de accidente que, en inicio constaba de 33 niveles y que, tras la aplicación del método de recodificación logístico, quedó con 10 categorías casi todas ellas con influencia en la modelización de la probabilidad logística del suceso de interés.

En el caso de la única variable de naturaleza continua de interés en el archivo, la edad, se decide, por motivos de interpretabilidad y por su falta de linealidad frente a la función logit, categorizarla mediante el procedimiento CHAID comentado pasando a tener cuatro niveles cuyos puntos de corte son 37,49 y 57 años. Se mantiene la variable como continua por si pudiera ser de interés durante el estudio.

```
str(TPGV_red)

## 'data.frame':    202804 obs. of  33 variables:
##  $ muert30  : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ accseg   : Factor w/ 2 levels "No","Si": 2 2 2 2 2 2 2 2 2 2 ...
##  $ alcohol  : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ edad     : num  46 33 46 51 28 52 30 25 25 23 ...
##  $ sexo     : Factor w/ 2 levels "Mujer","Varón": 2 2 2 2 2 1 2 1 1 1 ...
##  $ infrvel  : Factor w/ 5 levels "Velocidad inadecuada para las condiciones existentes",..:
##  $ tipoveh  : Factor w/ 8 levels "Autobus","Bici",..: 8 8 3 3 8 8 8 8 8 8 ...
##  $ infcond  : Factor w/ 8 levels "Ninguna","Distraccion",..: 1 3 3 1 3 3 1 1 1 3 ...
##  $ cansan   : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ velina   : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ infracc  : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ red      : Factor w/ 5 levels "Estatal","Autonómica",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ zona     : Factor w/ 4 levels "Carretera","Zona Urbana",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ lumin    : Factor w/ 4 levels "Crepusculo","Noche",..: 4 4 4 4 4 4 4 4 1 1 ...
##  $ superf   : Factor w/ 4 levels "Seca y limpia",..: 1 1 1 1 1 3 3 3 3 3 ...
##  $ barrera  : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 2 2 2 2 ...
##  $ mediana  : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 2 2 ...
##  $ tipoacc  : Factor w/ 10 levels "Frontal","Frontolateral",..: 5 5 5 5 5 8 3 3 4 4 ...
##  $ facatm   : Factor w/ 5 levels "Buen tiempo",..: 1 1 1 1 1 4 2 2 4 4 ...
##  $ tipovia  : Factor w/ 3 levels "Autopistas","Secundarias",..: 2 2 2 2 2 2 3 3 3 3 ...
##  $ posveh   : Factor w/ 9 levels "Conductor vehículo",..: 2 1 1 1 1 1 1 1 2 1 ...
##  $ lesivid  : Factor w/ 5 levels "Muerto","Herido grave",..: 3 4 4 4 4 4 3 3 3 4 4 ...
##  $ provin   : Factor w/ 52 levels "Álava","Albacete",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hergrav30: Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ distracc : Factor w/ 2 levels "No","Si": 2 2 2 2 2 2 2 2 2 2 ...
##  $ denscir  : Factor w/ 4 levels "0","1","2","3": 2 2 2 2 2 2 2 2 2 2 ...
##  $ hitos    : Factor w/ 2 levels "No","Si": 2 2 2 2 2 2 2 2 2 2 ...
##  $ tipodia  : Factor w/ 4 levels "Anterior a festivo",..: 4 4 4 4 4 3 3 3 3 3 ...
##  $ mes      : Factor w/ 12 levels "Enero","Febrero",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ diasem   : Factor w/ 7 levels "Lunes","Martes",..: 1 1 1 1 1 3 3 3 3 3 ...
##  $ idveh    : Factor w/ 32 levels "0 ","A ","AA",..: 2 2 9 10 11 2 2 9 2 2 ...
##  $ muerograv: Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 1 1 1 1 ...
##  $ edad3    : Factor w/ 4 levels "(0,37]","(37,49]",..: 2 1 2 3 1 3 1 1 1 1 ...
```

Tabla 1: Principales variables de estudio y sus niveles

Se presenta la Tabla 1 que contiene el resumen del conjunto de datos a estudio ya depurado y con las variables re categorizadas que se compone de 202804 registros y 33 variables.

6.3 ESTUDIO DESCRIPTIVO UNIVARIANTE

Todas las variables seleccionadas por su interés son de naturaleza categórica con distinto número de niveles excepto la edad, que posteriormente se recategorizó en los cuatro tramos indicados anteriormente.

De cara a la automatización del proceso de estudio descriptivo se programa un bucle que recorre las variables del archivo creando las tablas de frecuencias y los gráficos para cada una de ellas de tal forma que con la sola ejecución del código que se muestra se genera un informe descriptivo que proporciona un conocimiento del comportamiento de la población a estudio.

```
# Tablas de frecuencias de las variables más relevantes

tab <- function (var) {
  return(as.data.frame(round(prop.table(table(var))*100,2)))
}

# Se crea un bucle para obtener tablas de frecuencia y gráficos de todas las
variables
univar <- function (data) {
  var <- colnames(data)
  for (i in 1:length(var)) {
    cat ("La variable ", i, 'se llama ', var[i], '\n\n')
    if ((length(na.omit(data[,i])))!=0){
    t <- tab(data[,i])
    if (nlevels(t$var[i])<=10){
      g<-ggplot(t,aes(x=var,y=Freq,fill=var))+
        geom_bar(stat='identity',color='darkblue')+
        theme(axis.text.x  =element_text(angle= 45,hjust= 1 ))+
        scale_fill_brewer(palette="Blues")+labs(x=var[i],y="%")+
        ggtitle(paste("Frecuencia relativa de ",var[i]))
      print(g)}
    colnames(t)[1]<-var[i]
    print(t) }
  }
}

univar(TPGV_red)
```

Código 1: Función estudio descriptivo

En primer lugar, la variable objetivo del estudio es muerte a 30 días cuya baja prevalencia en la población, del 0,92%, augura una etapa de clasificación con complicaciones aseguradas.

```
##   muert30  Freq
## 1      No 99.08
## 2      Si  0.92
```

Por su lado la variable tipo de vehículo, Figura 1, fundamental para la creación de las subpoblaciones de interés, presenta una distribución asimétrica con más del 70% de turismos como vehículo implicado en un siniestro, seguido de un 12% de motos. Destaca el hecho de que la proporción de accidentes de conductores de bicis y camiones es similar, cosa que no parece lógica desde el punto de vista probabilístico en relación al

número de desplazamientos de cada vehículo, pero puede explicarse por la mayor lesividad asociada a este tipo de vehículos.

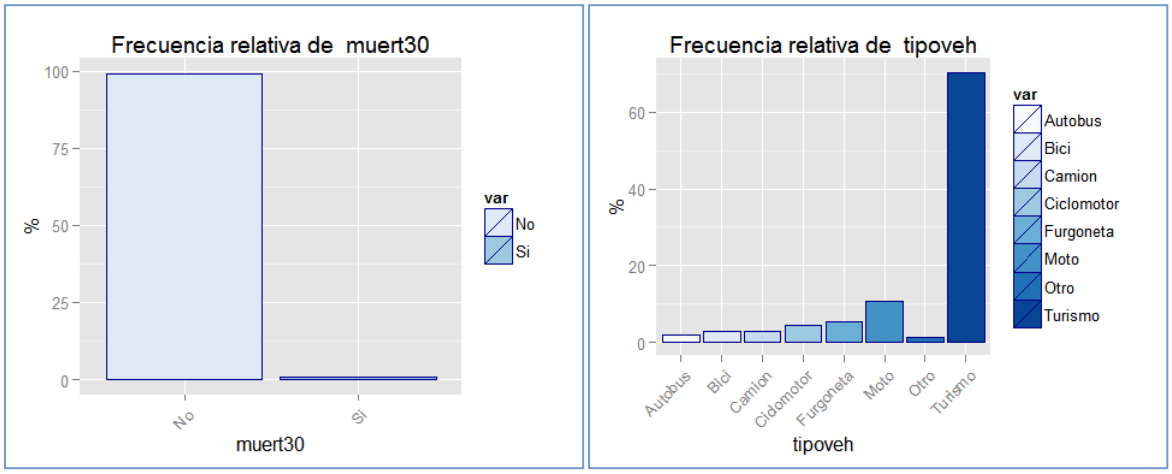


Figura 1: Frecuencia relativa de las variables *muerte a 30 días* y *tipo de vehículo*

En lo referente a los factores propios de la vía se han considerado los elementos presentes en ésta que pueden resultar peligrosos en los siniestros de vehículos de dos ruedas como *Mediana entre calzadas, Barrera de seguridad, Paneles direccionales, Hitos de arista, Captafaros* o estado de la *Superficie*, teniendo también en cuenta variables como *Tipo y Titularidad de la vía, Densidad de circulación y Zona.*

En los gráficos de la Figura 2 se observa que más de la mitad de los accidentes se producen en zona urbana, en torno al 45% en carretera y el 5% restante en travesías o variantes, los accidentes en carreteras secundarias suponen un tercio del total y en autopistas el 15%. Por otra parte la red municipal de carreteras absorbe el 62% de los accidentes con víctimas, seguida de la red estatal con un 14% y la autonómica con un 12%. Más del 90% de los siniestros se producen con niveles de densidad de circulación bajos.

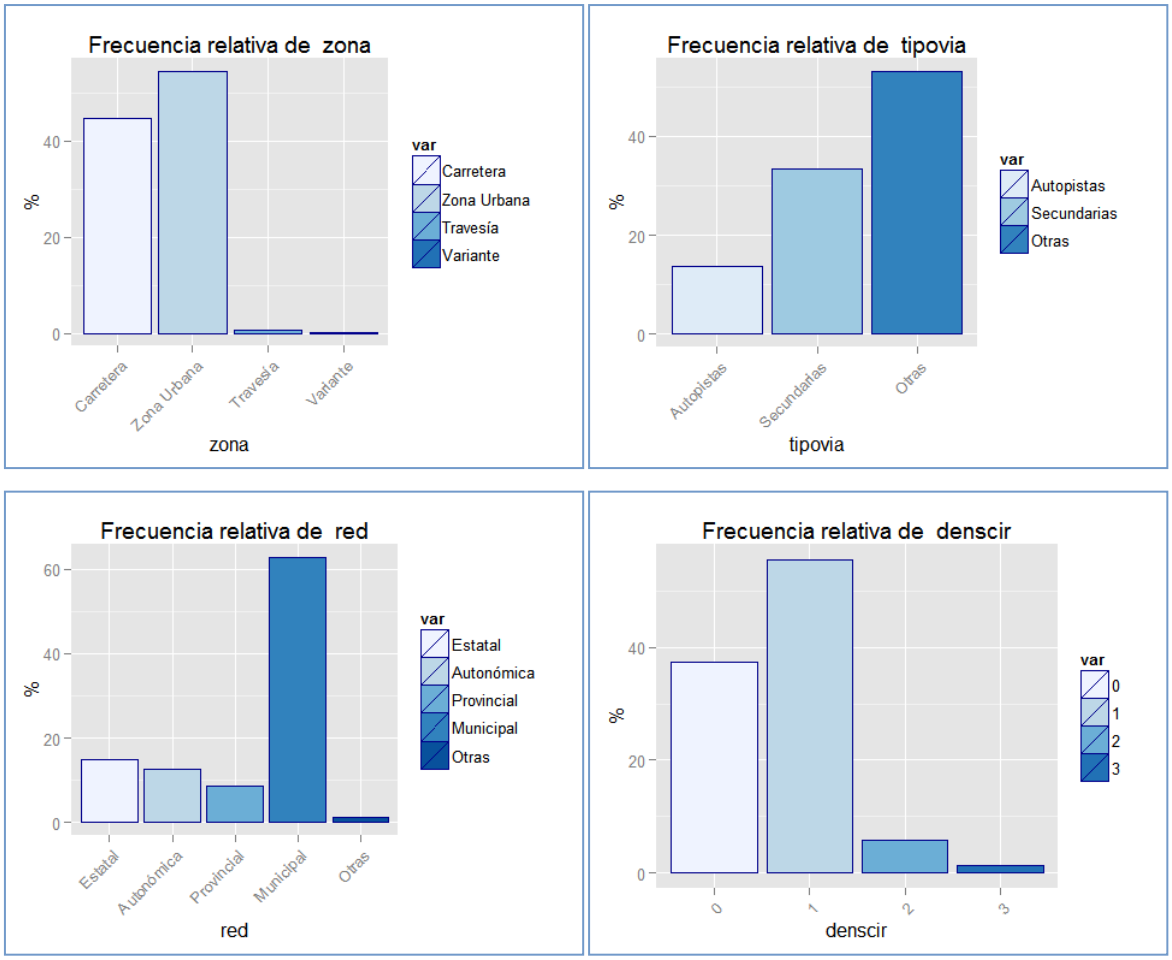


Figura 2: Frecuencia relativa de las variables *zona*, *tipo de vía*, *red* y *densidad de circulación.*

En cuanto a elementos de la vía que pueden influir en la severidad de las lesiones, Figura 3, la barrera de seguridad, la mediana entre calzadas y los hitos de arista están presentes en uno de cada cinco accidentes. Más del 80% de siniestros ocurren con la superficie seca y limpia y el 10% en superficie mojada.

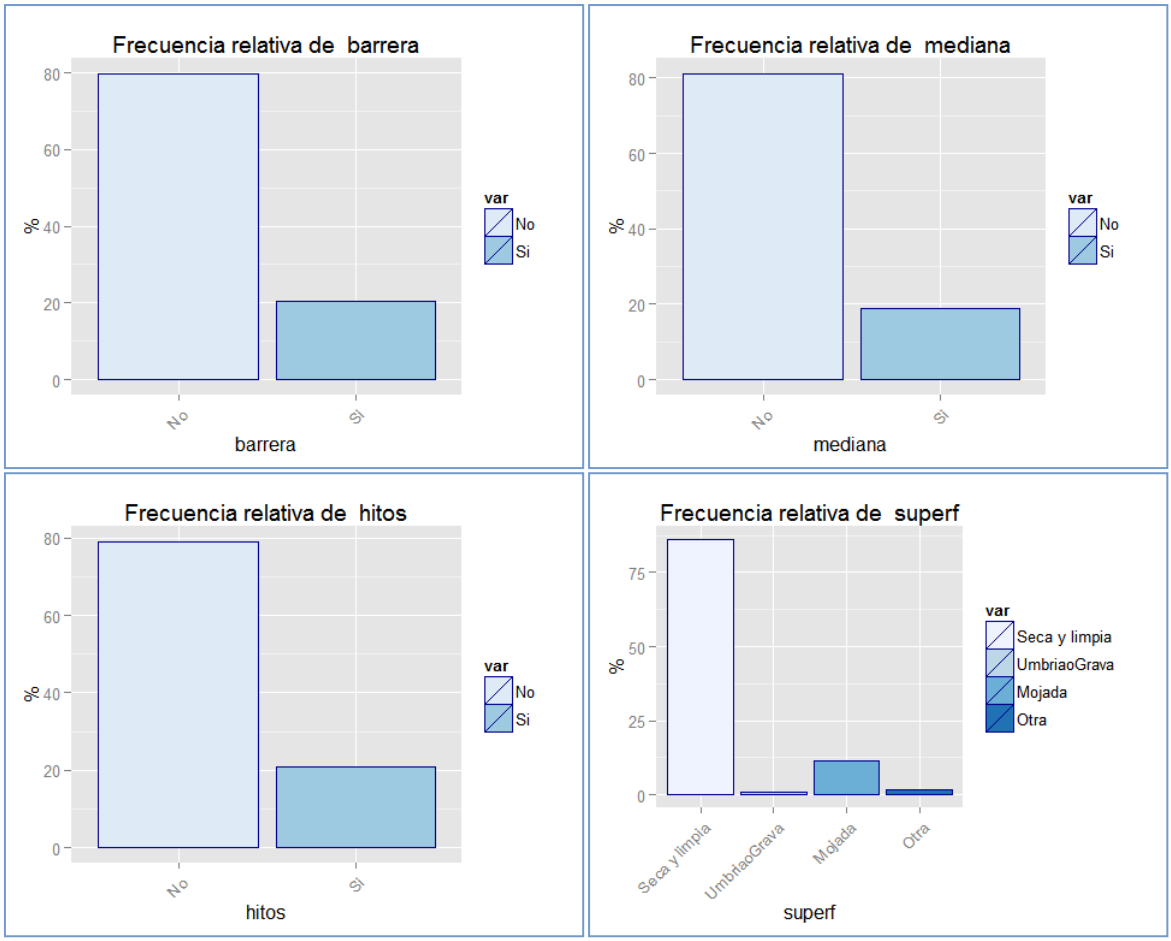


Figura 3: Frecuencia relativa de las variables *barrera, mediana, hitos y superficie.*

A continuación se presentan los factores propios del conductor de vehículo de dos ruedas implicado en siniestro en circulación, entre los que se encuentran variables socio demográficas como *sexo* y *edad,* la *posición en el vehículo* (recordemos que se seleccionaron los siniestros correspondientes a conductores y pasajeros de vehículos de dos ruedas) y conductas que pueden resultar de riesgo en la consecución de un accidente con resultado grave como *distracción al volante, consumo de alcohol y drogas, uso de accesorios de seguridad , distracciones o infracciones* cometidas.

Todas las conductas mencionadas anteriormente representan posibles factores de riesgo que incrementan la probabilidad de sufrir un accidente con resultado grave o mortal y, aunque algunos de ellos representen baja proporción del total, su presencia puede agravar las consecuencias del siniestro.

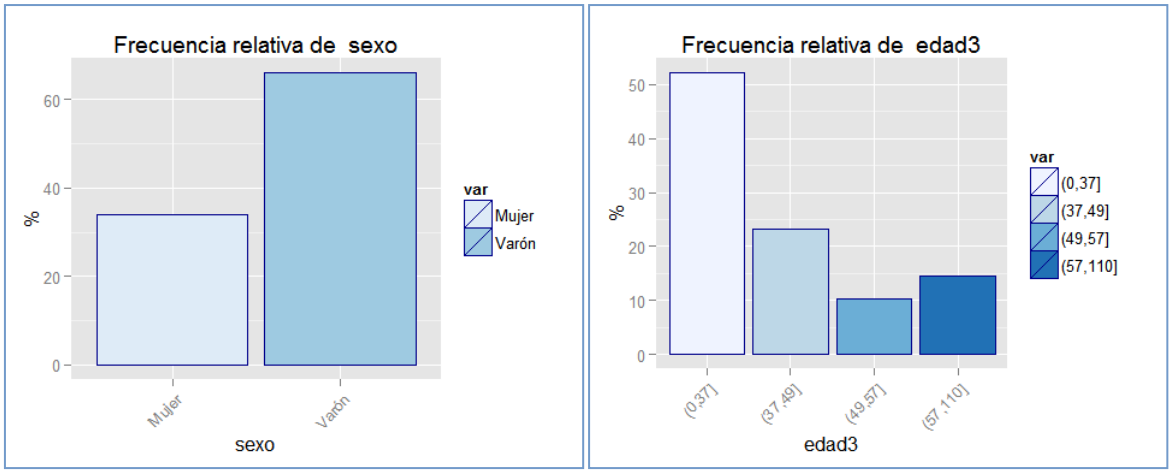


Figura 4: Frecuencia relativa de las variables *sexo y edad recategorizada.*

De los gráficos de la Figura 4 se extrae la información referente al perfil demográfico del accidentado. El 65% de las víctimas en 2012 eran *varones*, frente al 35% de *mujeres*, más de la mitad menores de 37 años, el 23% entre 38 y 49 años y un 15% mayores de 57.

En la Figura 5 se observa, en la distribución de la variable infracción del conductor, que el 70% de los conductores en accidentes con víctimas en 2012 no cometieron ninguna *infracción* y que la distracción se constató en el 10% ellos.

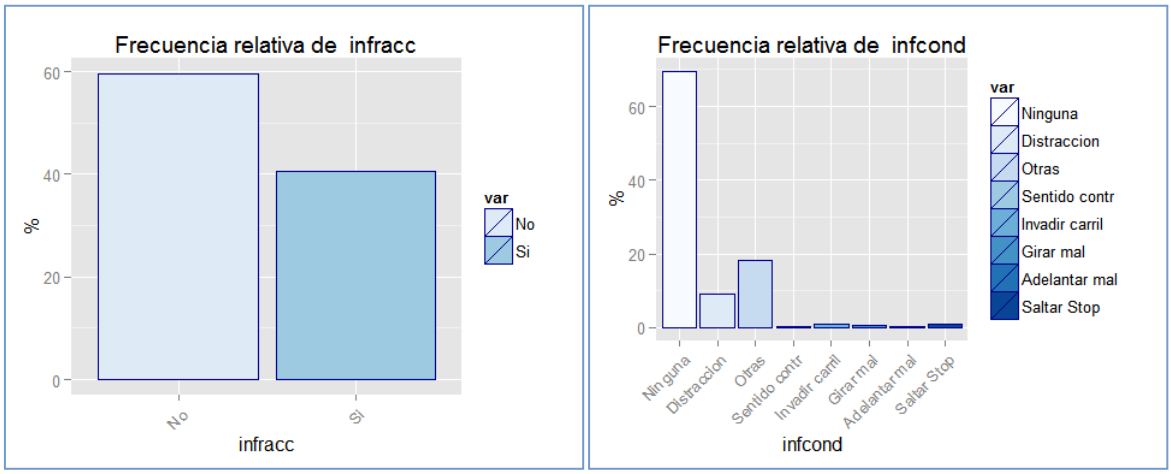


Figura 5: Frecuencia relativa de las variables *infracción* e *infracción del conductor*.

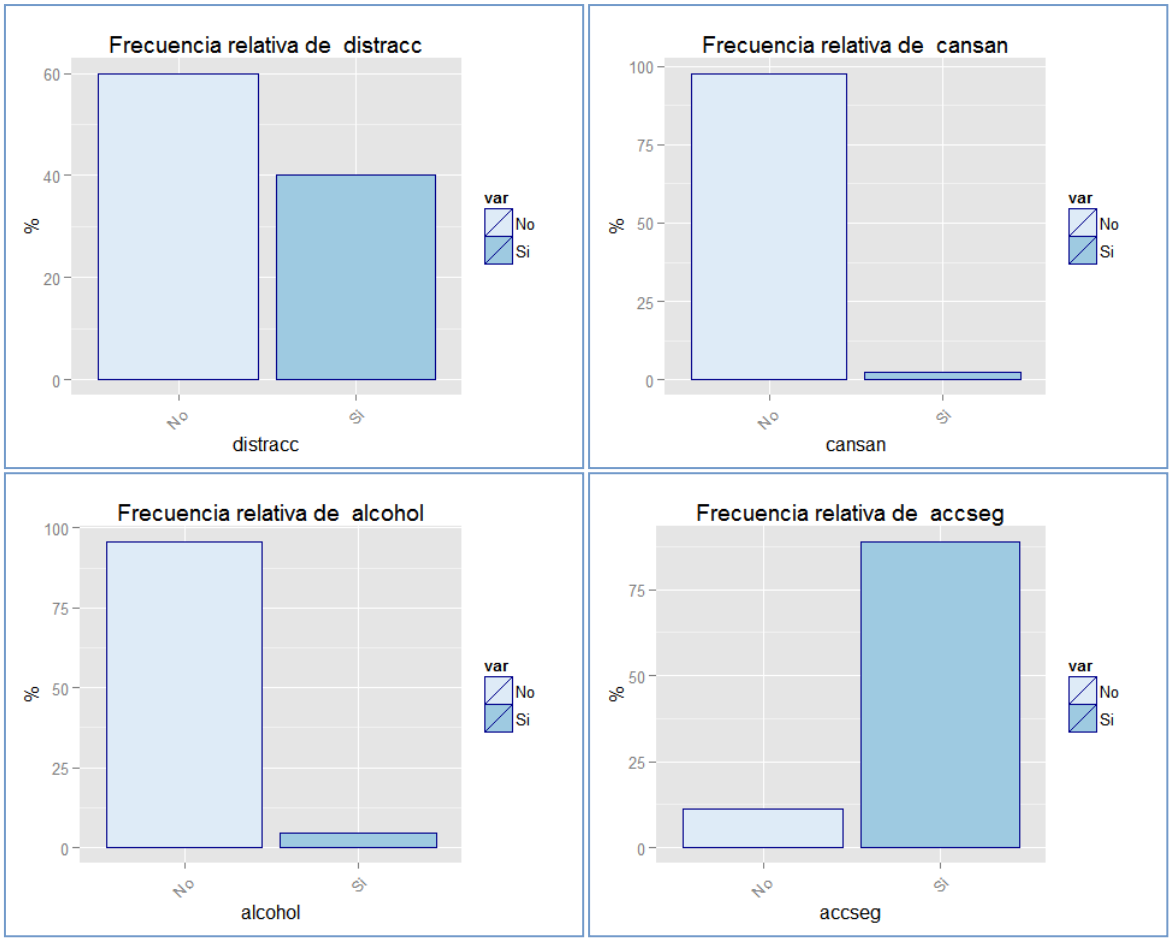


Figura 6: Frecuencia relativa de las variables *distracción, cansancio, alcohol y accesorios de seguridad.*

Se constató *distracción* del conductor el 40% de los casos, *cansancio o sueño* en tan solo el 2,4% de ellos (se achaca este bajo porcentaje a la difícil constatación de este hecho), conducción bajo los efectos de *alcohol o drogas* en un 4,4% de los accidentados y no utilización de *accesorios de seguridad*, en el 11% de estos.

```
##                             posveh  Freq
## 1               Conductor vehículo 56.94
## 2               Pasajero delantero 13.78
## 3       Pasajero trasero izquierdo  3.66
## 4         Pasajero trasero derecho  3.97
## 5          Pasajero trasero cenral  1.50
## 6 Conductor vehículo de dos ruedas 16.61
## 7  Pasajero vehículo de dos ruedas  1.61
## 8         Otros pasajeros sentados  1.80
## 9           otros pasajeros de pie  0.14
```

Tabla2: Frecuencias de la variable *posición en el vehículo*

En cuanto a la distribución en el vehículo, casi el 57% de los accidentados eran *conductores de vehículo* y más del 20% *pasajeros*, el 16% eran conductores de vehículos de dos ruedas.

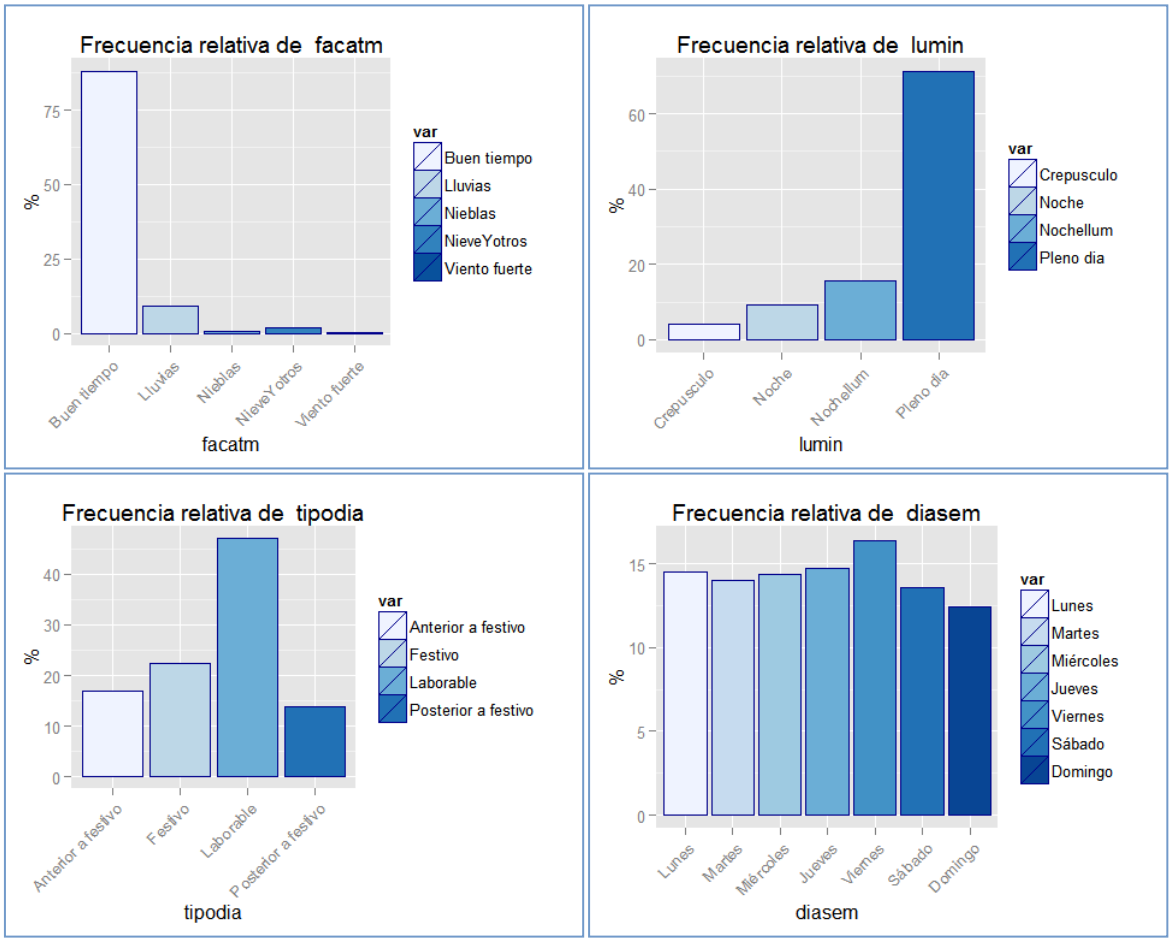


Figura 7: Frecuencia relativa de las variables *factores atmosféricos, luminosidad, tipo de día y día de la semana.*

Los accidentes con buen tiempo suponen más del 80% del total y con lluvia en torno al 10%, ocurren en pleno día más del 70% y el 25% de noche, solo el 5% en el crepúsculo, más de la mitad en día laborable y por encima del 20% en día festivo, los viernes son los días con mayor proporción de accidentes con víctimas. En cuanto al tipo de accidente predominan la colisión frontolateral y los alcances con un porcentaje acumulado de casi el 50%, seguidas de otro tipo de colisión con más del 10%, las salidas de vía por la derecha, 8% y las colisiones laterales con algo más del 7%.

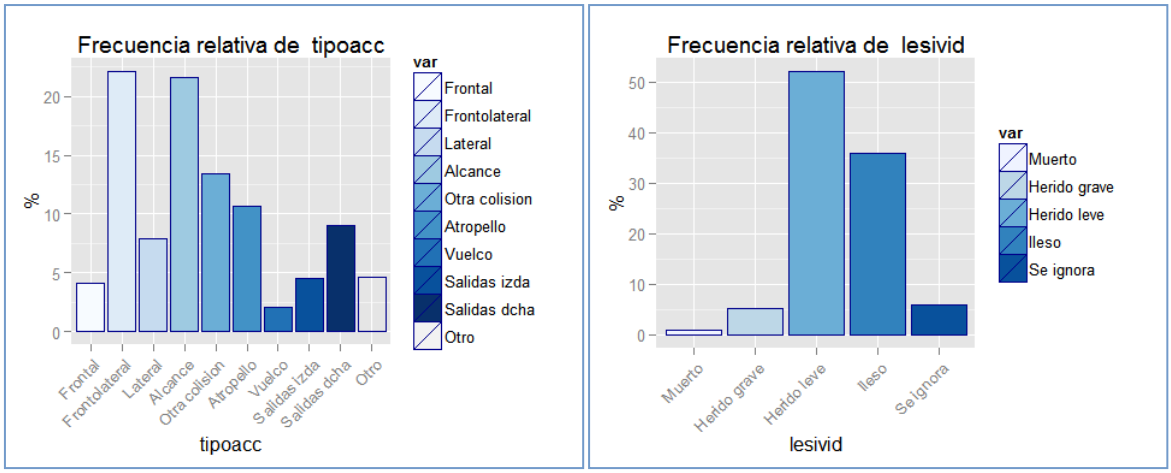


Figura 8: Frecuencia relativa de las variables *tipo de accidente y lesividad.*

La variable lesividad, Figura 8, informa del resultado del siniestro en el momento en el que ocurre y refleja que en 2012 se produjeron accidentes con resultado de muerte en el 0,82% de los casos y que la gran mayoría, un 88%, se saldaron con resultado de ileso o herido leve existiendo un 5% de heridos graves.

6.3 ESTUDIO FRENTE A LA VARIABLE OBJETIVO

Una vez conocidas las distribuciones marginales de las variables seleccionadas, se plantea en este epígrafe un estudio de los cruces de variables que, a priori, presenten mayor interés. Es evidente que la variable de mayor importancia es la denominada *muerte a 30 días* ya que será la que se pretenderá caracterizar mediante los modelos planteados a lo largo del estudio.

Así se presentarán los cruces de esta variable con algunas de las restantes con el objetivo de tener una idea descriptiva de la variación de las frecuencias relativas de ésta en función de los niveles de las otras y, de esta forma, identificar las categorías de las variables independientes con mayor incidencia de fallecidos a 30 días, por tanto, las de mayor riesgo.

Dada la baja presencia de la categoría de interés se decide realizar una comparación de las distribuciones de las variables dentro de las subpoblaciones formadas por la variable muerte a 30 días mediante diagramas de barras apiladas, obteniendo una representación en forma de mosaico sin tener en cuenta la gran diferencia en número de sucesos entre las categorías de la variable objetivo.

Para este fin se construye una función que presenta las tablas de contingencia con la variable muerte a 30 días en filas y cada una de las independientes en columna calculando el porcentaje fila y posteriormente se utiliza en un bucle análogo al del apartado anterior que también presenta los gráficos de este mosaico.

```
# Se crea una fuanción para tablas de contingencia (porcentaje fila)
xtab <- function (var) {
  return(as.data.frame(round(prop.table(table(muert30,var),1)*100,3)))
}
# Se crea un bucle para obtener tablas de frecuencia y gráficos de todas las
variables

cruces <- function (data) {
  attach(data)
var <- colnames(data)
for (i in 2:length(var)) {
  cat ("La variable ", i, 'se llama ', var[i], '\n\n')
  if ((length(na.omit(data[,i])))!=0){
  t <- xtab(data[,i])
  print(t)

  if (nlevels(t$var)<=10){
    g<-ggplot(t,aes(x=muert30,y=Freq,fill=var))+
      geom_bar(stat='identity',color='darkblue')+
      theme(axis.text.x  =element_text(angle= 45,hjust= 1 ))+
      scale_fill_brewer(palette="Blues")+labs(y="%")+
      ggtitle(paste("Mosaico ",var[i]))
    print(g)}
  colnames(t)[2]<-var[i]
   }
}
}
cruces(TPGV_red)
```

Código 2: Función estudio frente a la respuesta

Esta forma de visualización permite identificar de forma rápida las diferencias en las distribuciones de los factores dentro de cada categoría de la variable objetivo para poner de manifiesto cuáles de ellos pueden ejercer influencia sobre ésta.

En primer lugar en lo que se refiere a factores propios de la vía, destaca el cambio en la frecuencia relativa de los accidentes ocurridos en la red estatal y autonómica, siendo casi el doble en la población de fallecidos, por su parte aunque los siniestros en vías de la red municipal suponen más del 60% en la población de accidentados no fallecidos, solo suceden en un 25% de los que presentan resultado mortal.

Se pone de manifiesto la menor severidad de los siniestros en zona urbana, al contrario de os que sucede en zona de carretera. Las carreteras secundarias se erigen como las de

mayor riesgo en el resultado mortal de los accidentes de tráfico y el nivel 1 de densidad de circulación resulta mucho más frecuente en los siniestros mortales.

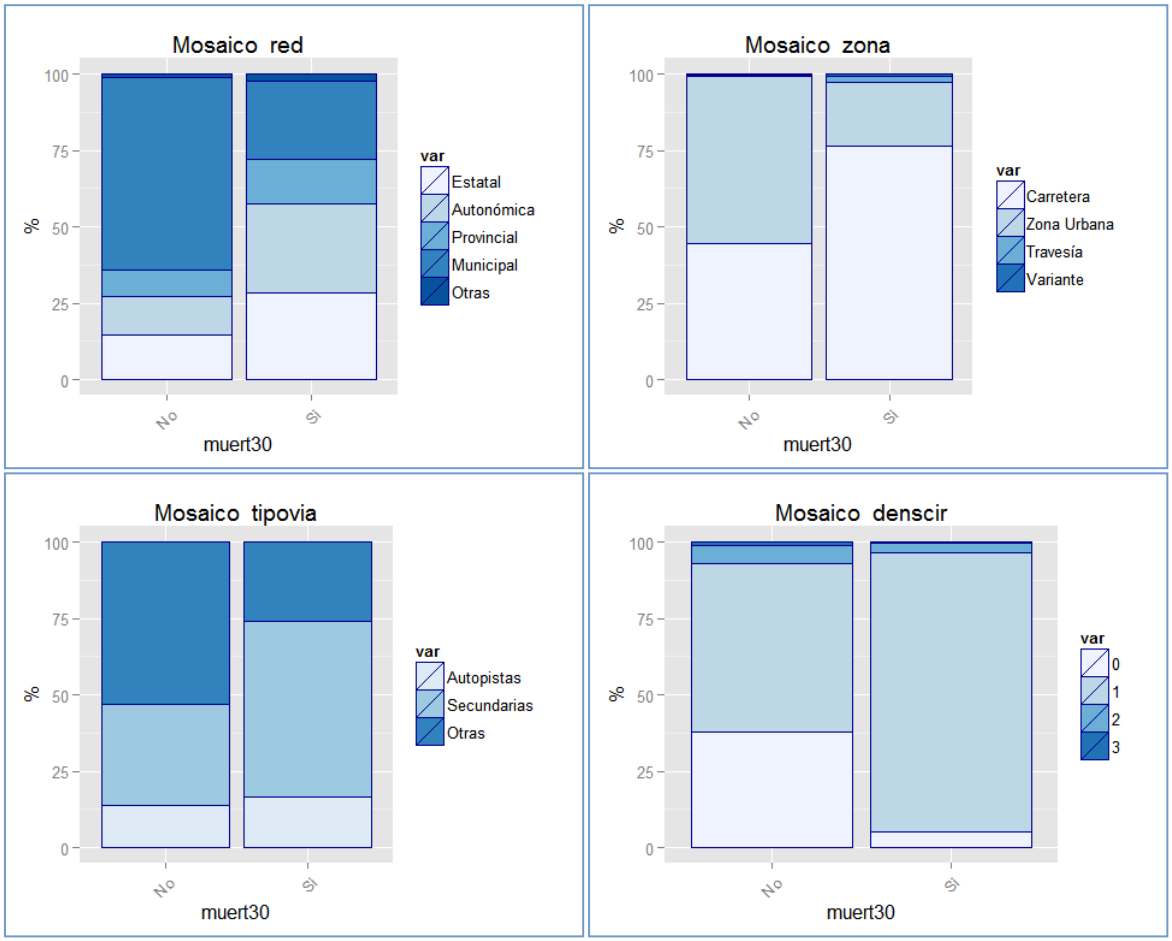


Figura 9: Mosaico de las variables *red, zona, tipo de vía y densidad de circulación.*

Por otra parte los accidentes ocurridos de noche en condiciones de baja iluminación tienen mayor frecuencia en la categoría 'Si' de la variable objetivo, así como la presencia de elementos en la vía como barrera de seguridad, hitos de arista y en menor medida mediana entre calzadas.

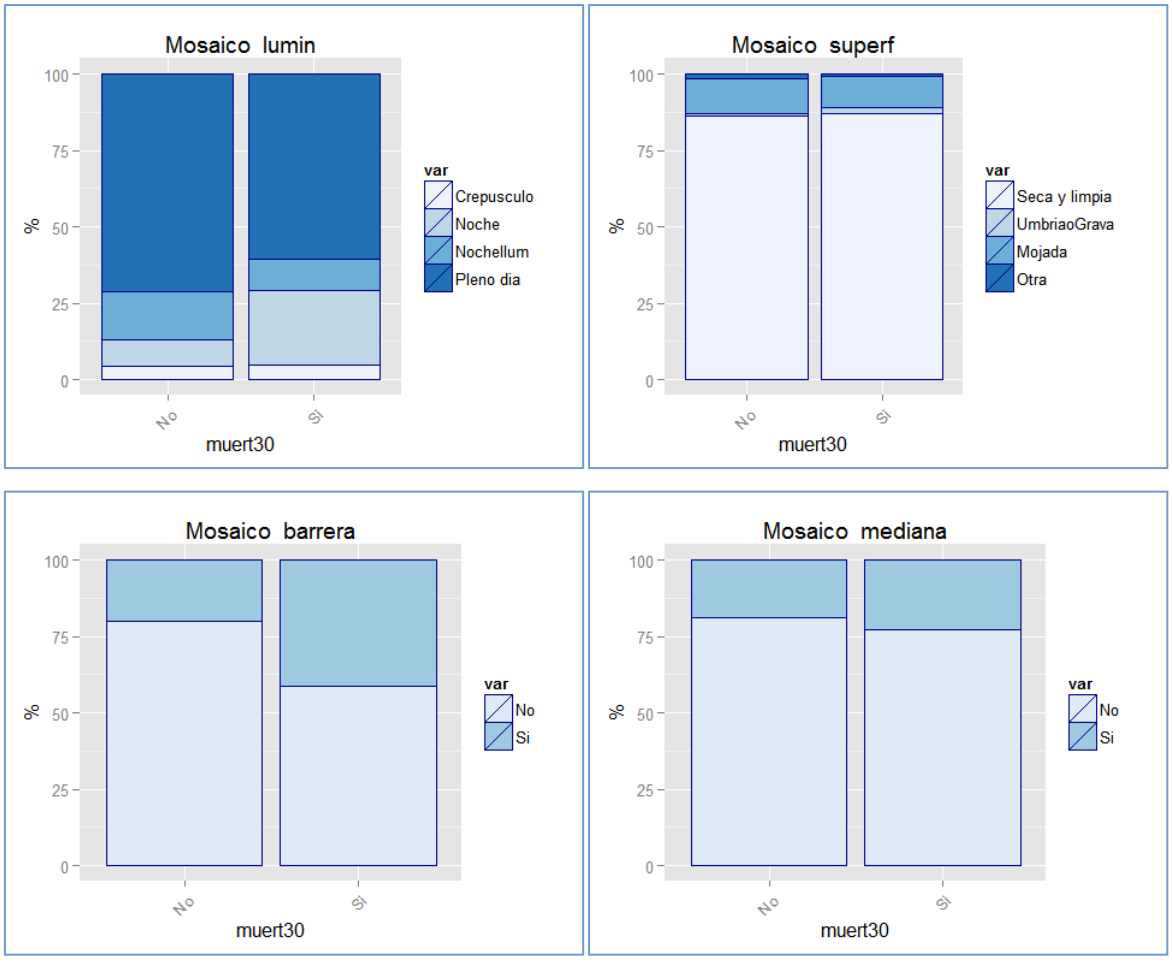


Figura 10: Mosaico de las variables *Luminosidad, superficie, barrera y mediana.*

En lo que se refiere a los factores propios del conductor, la no utilización de accesorios de seguridad y la velocidad inadecuada tienen una prevalencia mucho mayor en los accidentes con resultado mortal, así como el sexo varón y la edad mayor de 57 años

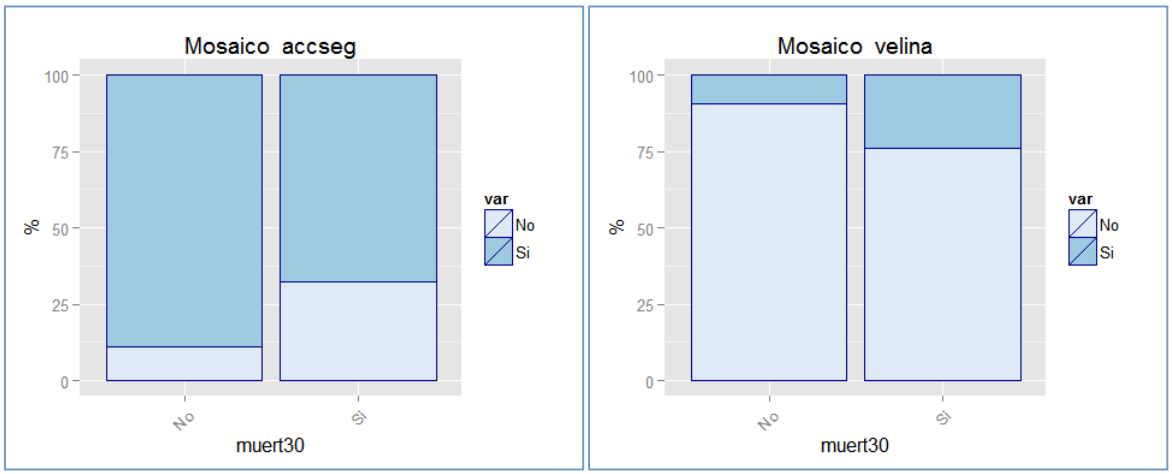


Figura 11: Mosaico de las variables *accesorios de seguridad y velocidad inadecuada*

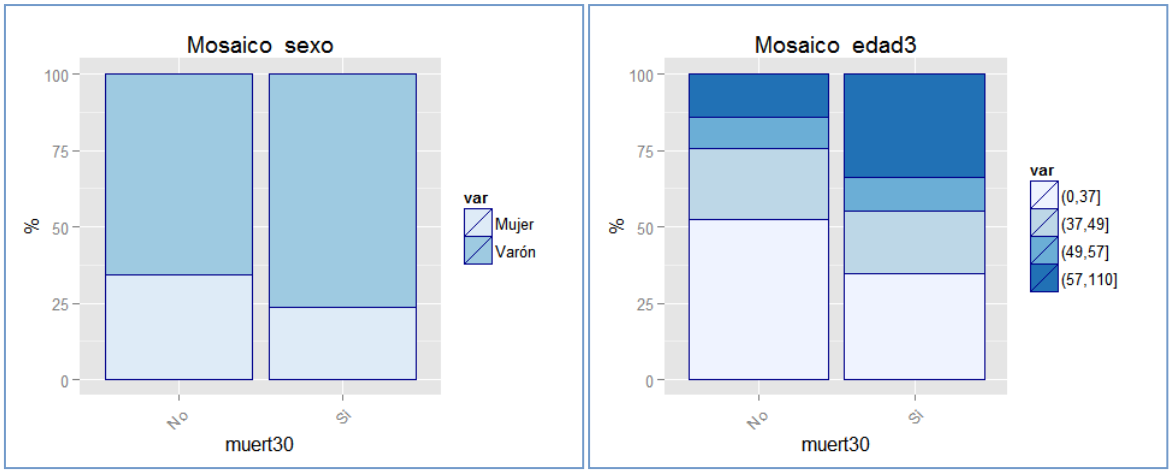


Figura 12: Mosaico de las variables *sexo y edad recategorizada.*

En la Figura 13 no parece existir una diferencia apreciable en las distribuciones de la variable alcohol en ambas categorias de la variable muerte a 30 días. Respecto al tipo de vehículo, es claro que la prevalencia de las motos en la población formada por los accidentes con fallecidos es mucho mayor.

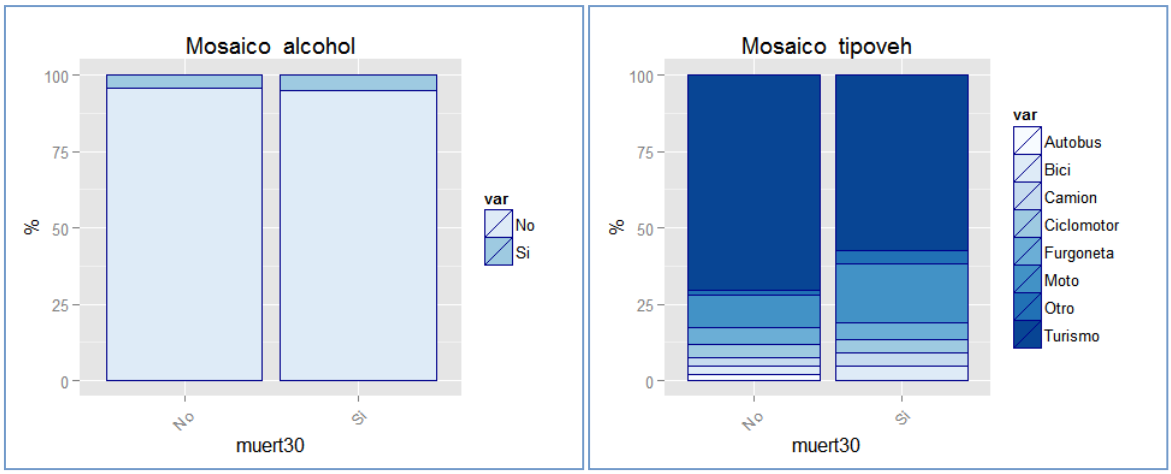


Figura 13: Mosaico de las variables *alcohol y tipo de vehículo.*

En el caso del tipo de accidente, Figura 14, las mayores diferencias se encuentran en salidas de vía por la izquierda, colisión frontal, salidas de vía por la derecha y atropellos, existiendo un porcentaje mayor en la población de fallecidos.

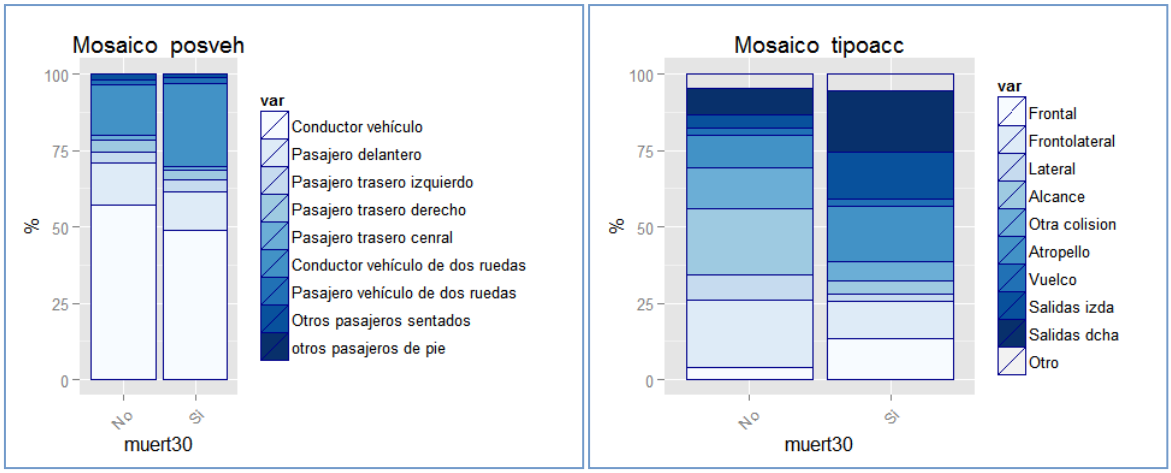


Figura 14: Mosaico de las variables *posición en el vehículo y tipo de accidente.*

El estudio del cruce de las variables posición en el vehículo y muerte a 30 días revela que son los conductores de vehículos de dos ruedas los que presentan mayor frecuencia en accidentes con víctimas mortales por lo que es conveniente estudiar especialmente las subpoblaciones formadas por estos vehículos.

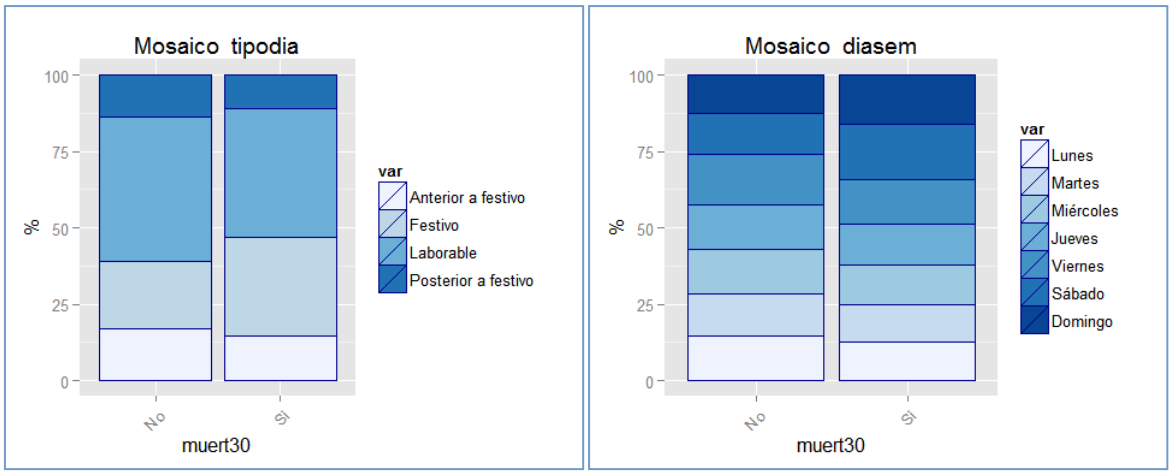


Figura 15: Mosaico de las variables *tipo de día y día de la semana.*

En lo referente al tipo de día existe una diferencia en la distribución de los días festivos con mayor proporción en la población de accidentes con fallecidos. Lo mismo sucede con sábado y domingo de la variable día de la semana.

Una vez que se ha realizado un estudio descriptivo suficientemente informativo sobre la naturaleza de los datos, se está en disposición de plantear los modelos estadísticos necesarios para la caracterización de las consecuencias de los siniestros de conductores o pasajeros de vehículos de dos ruedas en España en el año 2012.

7. SUBPOBLACIONES DE INTERÉS

En este punto y teniendo en cuenta la información extraída del análisis descriptivo de la población de accidentados en siniestros de tráfico en España en el año 2012, así como criterios basados en el conocimiento de la siniestralidad vial, se está en disposición de decidir cuáles serán las subpoblaciones de interés para este estudio.

Como se ha comentado anteriormente, los vehículos de dos ruedas presentan mayor severidad de las lesiones en accidentes que el resto de vehículos, por lo que necesariamente serán objeto de estudio. Dentro de este tipo de vehículos se considera relevante distinguir entre **bicicletas, ciclomotores y motocicletas** ya que debido a sus diferencias, los factores de influencia en la gravedad de las lesiones producidas pueden ser de distinta naturaleza.

Por otra parte, se considerará la subpoblación de **turismos** por ser, con diferencia, la más numerosa en la población aun siendo la que menor proporción de fallecidos a 30 días presenta de entre todas las consideradas, y también la subpoblación de **camiones** por considerarse distinta en cuanto a la capacitación del conductor de este tipo de vehículo (frecuentemente profesional) y a las consecuencias del tipo de accidente.

Por último, parece muy interesante estudiar la subpoblación de **peatones** ya que presenta características muy distintas a las demás.

En resumen, en este estudio se consideran seis de las subpoblaciones más relevantes en la siniestralidad en las carreteras, caracterizando los factores de mayor influencia en cada una de ellas por diversas técnicas, y proponiendo modelos de clasificación para el reconocimiento de patrones que inducen al resultado fatal de los accidentes en cada una de ellas.

Antes de comenzar con el proceso de ajuste de modelos y caracterización de perfiles de víctimas y escenarios de accidentalidad, se señalan en este punto las diferencias más relevantes entre las subpoblaciones consideradas a nivel descriptivo.

Para llevar a cabo el análisis descriptivo tanto univariante como de cruces con la variable de interés muerte a 30 días, se utilizan las funciones presentadas en el epígrafe

de estudio descriptivo con la única modificación del nombre del archivo a considerar, de forma que se generan informes descriptivos completos para cada subpoblación de manera rápida y sencilla.

Estos informes comentados no se presentan por su elevada extensión y falta de interés para los objetivos marcados, poniendo de manifiesto los aspectos que resultan de mayor relevancia en lo que a diferencias entre subpoblaciones se refiere.

Se presentan en la Tabla 3 las proporciones de las categorías de interés de las variables dicotómicas propias de la víctima así como la variable objetivo muerte a 30 días. Se observa la baja incidencia del evento de interés en todas las poblaciones, destacando al alza la subpoblación de peatones con una mortalidad del 3,1%. La no utilización de accesorios de seguridad es más acusada en las poblaciones de bicis y peatones debido a su falta de obligatoriedad, siendo los conductores de ciclomotores y motos los que más responsables en ese aspecto. Este hecho contrasta con las infracciones, que son cometidas en mayor medida por conductores de ciclomotores y motos con un 50,46% y un 47,78%, respectivamente.

		Camiones	Bicis	Motos	Ciclomotores	Peatones	Turismos
muert30	*% Si*	1,23	1,28	1,41	0,74	3,1	0,64
accseg	*% No*	7,92	32,94	6,54	5,37	39,26	8,02
alcohol	*% Si*	2	1,05	1,98	2,67	2,59	5,31
infracc	*% Si*	35,97	46,03	47,78	50,46	41,08	38,59
distracc	*% Si*	46,93	33,15	28,47	30,11	37,27	42,83
velina	*% Si*	10,48	5,58	8,84	5,2	13,53	9,99
Num. Observaciones		5512	5535	20716	8535	11504	134878

Tabla 3: Comparativa de subpoblaciones. *Muerte a 30 días* y factores de la víctima.

Las distracciones son más frecuentes en las subpoblaciones de camiones y turismos (46,93 y 42,83%), siendo estos últimos los que mayor tasa de consumo de alcohol presentan (5,31%), aproximadamente el doble que la siguiente subpoblación en el ranking, los ciclomotores (2,67%).

Respecto a otras variables de interés cuyos gráficos se presentan en el Anexo II, la presencia de víctimas mujeres es superior al 50% en la subpoblación de peatones siendo inferior al 5% en la de camiones. La distribución de la edad es distinta en la subpoblaciones destacando la alta frecuencia de víctimas mayores de 57 años en peatones y de aquellas que tienen entre 38 y 47 años en camiones. Los accidentes ocurridos en zona urbana son clara mayoría en las subpoblaciones de peatones,

ciclomotores, bicis y motos pero se constata en todas la mayor mortalidad de los accidentes en carretera. En lo referente al tipo de accidente, destaca el vuelco como factor de riesgo en camiones, la colisión frontal y las salidas por la izquierda y derecha en turismos, presentando estas últimas mayor mortalidad en motos, ciclomotores y bicis. Destaca la peligrosidad de los alcances en bicis.

8. FACTORES DE INFLUENCIA EN LA MORTALIDAD

El primero de los dos grandes apartados de modelización en este estudio está dedicado a la determinación de los factores de influencia en el resultado fatal de los siniestros de tráfico en España en 2012, poniendo especial énfasis a la capacidad de generalización del proceso a datos de otros años.

Se comienza ajustando modelos clásicos de regresión logística a las subpoblaciones por la gran ventaja que éstos tienen a la hora de cuantificar los efectos de los factores que resultan influyentes a través de los odds ratio. Cabe destacar que no se pretende crear los modelos óptimos pues el objetivo fundamental es construir una metodología que sea generalizable y constituya una base para el análisis y las posteriores modificaciones o mejoras siempre podrán hacerse ad hoc con el mínimo esfuerzo.

Con el objetivo de encontrar los factores que resultan de mayor influencia se recurre al ajuste de modelos con algoritmos propios de la minería de datos mediante procesos automáticos de selección de los mejores ajustes, utilizando la técnica de validación cruzada repetida ya comentada en la metodología cuyo fin es evitar el sobreajuste a los datos, proporcionando así modelos con mayor capacidad de generalización. Los algoritmos utilizados han sido comentados en el apartado de metodología.

Finalmente, y dado que, debido a la elevada extensión de los resultados (los más relevantes se encuentran en el Anexo I), solo se presentarán los correspondientes a la subpoblación de bicicletas, se dedica el último apartado a la comparación de los resultados de las distintas técnicas empleadas en lo referente a la extracción de factores o variables de influencia en el resultado de muerte a 30 días en accidentes de tráfico. La

idea básica es valorar las diferencias existentes y determinar los factores que con mayor frecuencia son elegidos como variables independientes a lo largo de los modelos pues de esta forma su influencia se considerará contrastada.

8.1 MODELO CLÁSICO. REGRESIÓN LOGÍSTICA

En esta parte del estudio se pretende poner de manifiesto cuales son los **factores de riesgo** en la accidentalidad con resultado de muerte en las distintas subpoblaciones de interés. El objetivo fundamental es determinar cuáles de estos factores aumentan en mayor medida la probabilidad de accidente con resultado mortal.

Para este objetivo se ajusta un modelo de regresión **logística binomial multivariante**. En este modelo, la variable respuesta que se pretende caracterizar es la *muerte a 30 días* y, para determinar las variables de influencia de forma automática se lleva a cabo un procedimiento stepwise con el criterio de Akaike, de forma que se seleccionará el modelo final más simple posible que no presente diferencias significativas en su verosimilitud con el siguiente más complejo.

Se programa una función que realiza esta tarea de ajuste para cada población mostrando a su vez el resumen del modelo con la significación estadística de los parámetros estimados, los odds ratio para cada uno de ellos y también la curva ROC con el punto de corte de la probabilidad estimada que se considera óptimo y las matrices de confusión del ajuste tomando como punto de corte de la probabilidad estimada la prevalencia a priori del evento en la población y otro punto de corte de libre elección cuyo valor por defecto es 0.5. De esta forma se puede comparar la capacidad de clasificación en función de dicho punto de corte.

```
logiPredBici<-predicciones(logiBici,bicis)

##
## Call:
## glm(formula = muert30 ~ edad3 + sexo + alcohol + accseg + denscir +
##     facatm + barrera + lumin + velina + infracc + red + tipoacc,
##     family = binomial(link = "logit"), data = data_na)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4592  -0.1298  -0.0742  -0.0481   3.6713
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -6.55334    0.76947  -8.517  < 2e-16 ***
```

```
## edad34                   2.07265    0.33135   6.255 3.97e-10 ***
## alcoholSi                2.76553    0.52489   5.269 1.37e-07 ***
## accsegSi                -0.50683    0.28505  -1.778 0.075400 .
## denscir                  0.69656    0.32157   2.166 0.030302 *
## facatmViento fuerte      2.78984    0.83211   3.353 0.000800 ***
## barreraSi                0.66620    0.31360   2.124 0.033643 *
## luminCrepusculo          1.46495    0.45204   3.241 0.001192 **
## luminNoche               1.60817    0.43725   3.678 0.000235 ***
## infraccSi                0.79533    0.28301   2.810 0.004951 **
## redEstatal               1.35201    0.43653   3.097 0.001954 **
## redOtras                 2.39060    0.64435   3.710 0.000207 ***
## redProvincial            1.04670    0.37474   2.793 0.005220 **
## tipoaccFrontolateral    -1.24742    0.40500  -3.080 0.002070 **
## tipoaccLateral          -1.53947    0.56078  -2.745 0.006047 **
## tipoaccOtra colision    -1.27797    0.68547  -1.864 0.062270 .
## tipoaccSalidas dcha     -1.94532    1.10021  -1.768 0.077038 .
## tipoaccVuelco           -1.21520    0.61846  -1.965 0.049429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 752.70  on 5271  degrees of freedom
## Residual deviance: 547.21  on 5241  degrees of freedom
## AIC: 609.21
##
## Number of Fisher Scoring iterations: 18
```

Tabla 4: Modelo de regresión logística para la subpoblación de bicis. Coeficientes

Se muestra en la Tabla 4 el resultado del ajuste del modelo de regresión logística en la subpoblación de bicis, con 12 variables independientes seleccionadas y una significación de los parámetros que se considera mejorable (solo se presentan los parámetros significativos al 90% de confianza). Cabe destacar que, al tratarse de variables categóricas, la significación de uno de sus niveles puede suponer razón suficiente para mantener la variable completa en el modelo. Por otro lado se observan ciertas anomalías en los errores estándar de estimación que son debidos a la baja frecuencia (en ocasiones la no presencia) de la categoría correspondiente en la población.

Un parámetro con signo positivo se asocia, en regresión logística a un OR (Odds Ratio) mayor que la unidad, es decir a un factor cuya presencia aumenta la probabilidad del evento de interés, en este caso la muerte. Análogamente se considerarán los parámetros con signo negativo como factores atenuantes.

Para cuantificar el efecto de los factores tanto de riesgo como de amortiguación se recurre al cálculo de los OR asociados a cada uno de los parámetros estimados del modelo mediante la función *logistic.display()* disponible en el paquete *epicalc* de R.

```
## Resumen del modelo logístico ajustado
##
##                          OR lower95ci upper95ci Pr(>|Z|)
## edad34                 7.95      4.15     15.21     0.00
## alcoholSi             15.89      5.68     44.45     0.00
## accsegSi               0.60      0.34      1.05     0.08
## denscir                2.01      1.07      3.77     0.03
## facatmViento fuerte   16.28      3.19     83.16     0.00
## barreraSi              1.95      1.05      3.60     0.03
## luminCrepusculo        4.33      1.78     10.50     0.00
## luminNoche             4.99      2.12     11.77     0.00
## infraccSi              2.22      1.27      3.86     0.00
## redEstatal             3.87      1.64      9.09     0.00
## redOtras              10.92      3.09     38.61     0.00
## redProvincial          2.85      1.37      5.94     0.01
## tipoaccFrontolateral   0.29      0.13      0.64     0.00
## tipoaccLateral         0.21      0.07      0.64     0.01
## tipoaccOtra colision   0.28      0.07      1.07     0.06
## tipoaccSalidas dcha    0.14      0.02      1.23     0.08
## tipoaccVuelco          0.30      0.09      1.00     0.05
```

Tabla 5: Modelo de regresión logística subpoblación de bicis. Odds Ratio (OR)

En cuanto a los intervalos de confianza, parecen exageradamente amplios en algunos casos. En este punto es importante distinguir entre los OR asociados a los parámetros que no resultan significativos en el modelo, a los cuales no se les concederá importancia, y aquellos que resultan efectivamente significativos que merecen una investigación sobre las causas de esta anomalía. Este hecho puede deberse a la baja frecuencia de la categoría y a la relativamente alta incidencia del suceso de interés, *muerte a 30 días*, dentro de la misma, de esta forma la estimación resulta imprecisa.

A continuación se resumen las conclusiones que se pueden extraer de este modelo en cuanto a factores de riesgo.

- **Factores atmosféricos**

El ***Viento fuerte*** es un factor de riesgo que **incrementa** la probabilidad de muerte **16,2** veces respecto a *Buen tiempo*. IC[3,2 , 83,1]. Intervalo demasiado amplio.

- **Alcohol**

La ingesta de alcohol aumenta el riesgo de accidente mortal **15,9** veces. IC[5,7 , 44,5]

- **Edad**

El grupo de edad ***Mayores de 57 años*** presentan un riesgo de accidente mortal **7,9** veces **superior** al grupo de referencia (*Menores de 37 años*). IC[4,1 , 15,2]

- **Infracción**

Cometer una ***Infracción*** **incrementa el riesgo** de accidente con resultado mortal **2,2** veces. IC[1,3 , 3,9].

- **Luminosidad**

Los accidentes ***nocturnos*** tienen riesgo de muerte **4,9 veces superior** que en pleno día. IC[2,1 , 11,8]

Los accidentes en el ***crepúsculo*** tienen riesgo de muerte **4,3 veces superior** que en pleno día. IC[1,8 , 10,5]

- **Red**

Otras: incremento del riesgo de muerte de **10,9** veces respecto a la red municipal. IC[3,1 , 38,6]

Estatal: incremento del riesgo de muerte de **3,9** veces respecto a la red municipal. IC[1,6 , 9,1]

Provincial: incremento del riesgo de muerte de **2,8** veces respecto a la red municipal. IC[1,4 , 5,9]

Los factores que disminuyen la probabilidad de muerte en siniestros de vehículos de dos ruedas son los siguientes.

- **Vuelco**

En los accidentes con ***Vuelco*** el riesgo de mortalidad a 30 días en **0,30** veces respecto al *alcance*. IC [0,09, 1]. Cautela en la interpretación debido al excesivo intervalo de confianza. Es posible que existan interacciones significativas con otros factores como el uso de accesorios de seguridad.

- **Colisiones frontolateral y lateral**

Las colisiones frontolaterales y laterales presentan un riesgo de muerte de **0,29 y 0,21** veces el riesgo de los producidos por alcance, respectivamente.

Como resumen visual de los OR para el modelo de regresión logística ajustado a la subpoblación de bicis se presenta el siguiente gráfico que muestra el valor de aquellos correspondientes a los parámetros que resultan significativos en el modelo.

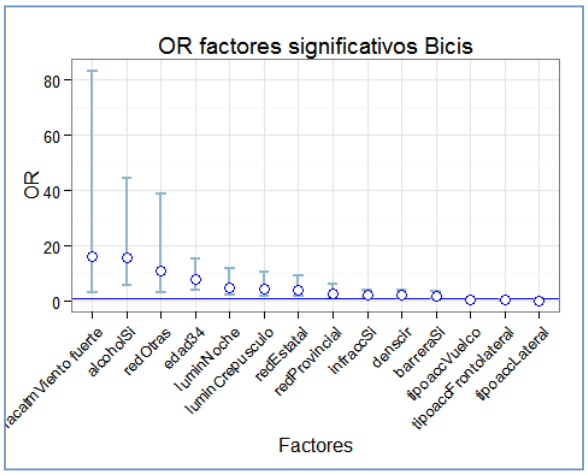


Figura 16: OR y sus intervalos de confianza 95% (bicis).

De esta forma se tiene una idea rápida de los factores que resultan de riesgo y de atenuación en el resultado letal a 30 días de los siniestros de tráfico. Esta información será utilizada en el apartado de extracción de factores con mayor frecuencia de influencia a través de los modelos.

8.2 TÉCNICAS EN MINERÍA DE DATOS

En este apartado se ajustan modelos de clasificación propios de minería de datos a las subpoblaciones de interés con el fin de poner de manifiesto los factores de influencia en el resultado mortal de los siniestros de tráfico a través de las medidas de importancia de las variables de las que dispone el paquete caret con su función *varImp()*.

Con el objetivo de la automatización de los procesos para esta metodología, se crean funciones (Anexo III) que ajustan los modelos considerados variando los parámetros básicos de control de los mismos y validando los resultados mediante la técnica de validación cruzada repetida. De esta forma con una sola ejecución se obtienen, para cada subpoblación, los modelos que presentan mayor valor de la métrica de evaluación escogida que será en este caso el área bajo la curva ROC, de entre todos los estimados.

Conviene señalar que la opción por defecto para la métrica en la función train() de caret es la precisión, pero teniendo en cuenta la naturaleza de los datos y la acusada falta de balanceo, ésta tiende a proporcionar modelos con muy baja sensibilidad, es por ello que se escoge la métrica ROC que tiene en cuanta la relación entre sensibilidad y especificidad.

En cuanto al procedimiento de presentación de resultados, por cada técnica se desarrollarán los pasos fundamentales en la exploración y ajuste del modelo óptimo para la subpoblación de bicicletas, por su especial interés en la actualidad debido al reciente aumento en el uso de este medio de desplazamiento y se presentará una tabla final con los mejores modelos para la técnica considerada en todas las subpoblaciones de interés.

8.2.1 REGRESIÓN LOGISTICA CON BOOSTING

En primer lugar y con el objetivo de evitar el sobreajuste a los datos que el modelo logístico anterior pudiera presentar, se lleva a cabo un proceso de estimación de un modelo de regresión logística utilizando la técnica de aprendizaje boosting explicada en el apartado de metodología, en este caso se utiliza la función de pérdida logística, *deviance*, para actualizar los pesos de las observaciones mal clasificadas.

Al igual que el Gradient Boosting, este método tiene por objetivo aprender de los errores cometidos por el clasificador individual considerado en la etapa anterior para así conceder mayor importancia a la clasificación de las observaciones mal clasificadas.

Todo esto unido a la utilización de la validación cruzada repetida, hace que los resultados sean, a priori, más fiables en cuanto a capacidad de generalización a nuevas observaciones.

En la práctica se utiliza una rejilla para variar el parámetro básico de esta técnica, el número de iteraciones del algoritmo, cuya sintaxis para utilización en la función train() es la siguiente.

```
# Rejilla de parámetros para logitBoost
lbGrid <- expand.grid(.nIter= c(1,10,50,100,500,1000,2000))
```

Con esta especificación se construyen siete modelos de clasificación con este algoritmo para los datos y a su vez se validan en las muestras determinadas por el número

especificado en el control de aprendizaje, en este caso 12. Por tanto se obtienen modelos suficientemente contrastados cuyos resultados son los siguientes.

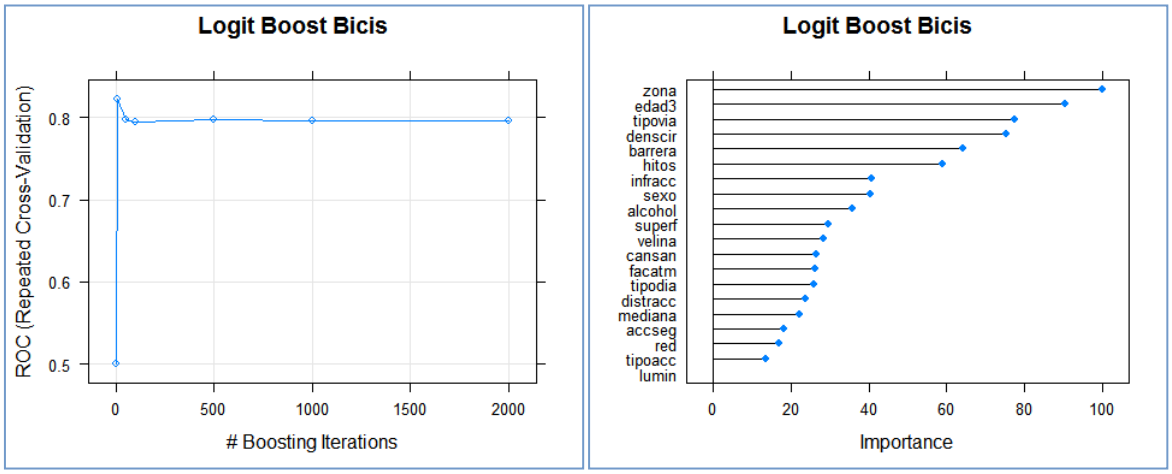


Figura 17: Iteraciones de boosting e importancia de variables Logit Boost (Bicis).

En cuanto a las iteraciones boosting, la métrica ROC lleva a la elección de un modelo logit boost con solamente 10 iteraciones del algoritmo puesto que los resultados son mejores, con un valor del área bajo la curva ROC de 0,82.

Para este modelo se calcula la importancia de las variables como medida de influencia de los factores en el suceso de interés. Así en el gráfico de la derecha se observa que las variables con mayor importancia resultan ser *zona, edad, tipo de vía, densidad de circulación* y *barrera*. Estas serán tenidas en cuenta para la posterior valoración global de los factores con mayor influencia.

De manera análoga y sin más dificultad que la ejecución de la función programada al efecto (Anexo III), se crean los modelos para las subpoblaciones restantes, resumiendo en la siguiente tabla los mejores modelos para cada una de ellas con el parámetro óptimo y el valor de ROC.

```
mejorModelologboost

##          nIter ROC LogiBoost
## Camiones 2000  0.6914192
## Motos    100   0.8260587
## Bicis    10    0.822796
## Peatones 1000  0.7998874
## Ciclos   1000  0.8107651
## Turismos 100   0.8433899
```

Tabla 6: Mejores modelos LogiBoost

Destaca el hecho de que la subpoblación de bicicletas es la que presenta menor valor de número de iteraciones, 10, siendo el siguiente valor nada menos que 100. En general los valores de ROC parecen aceptables, estando por encima del 75% excepto para la subpoblación de camiones.

8.2.2 REDES NEURONALES

A pesar de que las redes neuronales no son, en principio, el mejor método para el estudio de este tipo de datos debido en primer lugar a la existencia de variables categóricas cuyo tratamiento no es óptimo y, en segundo lugar a la falta de interpretabilidad del modelo, en este caso y gracias a las bondades del paquete Caret de R se consigue que las redes neuronales presenten ajustes muy buenos.

El por qué reside en la capacidad de este paquete para crear automáticamente un set de variables indicador para las categorías de las variables nominales permitiendo de esta forma que se consideren solamente aquellas categorías que realmente tienen influencia disminuyendo sustancialmente el número de parámetros (de por sí elevado) a estimar por este algoritmo. Por otro lado mediante la función varImp() se puede obtener una medida de importancia de las variables que viene dada por el cálculo de los valores absolutos de los pesos de la red seleccionada con lo que ambos inconvenientes pueden ser soslayados.

Cabe destacar que resulta cuestionable la medida de importancia de las variables ya que en el momento de entrada a la capa oculta de la red neuronal los pesos pueden modificarse obteniendo en la salida resultados que poco tienen que ver con lo supuesto en la capa de entrada. Aún así se tendrán en cuenta con cautela.

En la práctica los parámetros básicos a monitorizar en este tipo de redes neuronales son el número de nodos en la capa oculta (son redes con una única capa oculta) y el parámetro de actualización de los pesos, *weight decay*.

Tras diversas pruebas se decide tomar la siguiente rejilla de parámetros.

```
# Rejilla de parámetros para logitBoost
lbGrid <- expand.grid(.size= c(1,2,5,10,15,20),.decay= c(0.1,0.4,0.6))
```

De esta forma se ajustan tantas redes neuronales como combinaciones de estos parámetros y se validan las 12 muestras dadas por la validación cruzada repetida definida para todos los modelos. De todas las redes ajustadas se selecciona la que mayor valor del área bajo la curva ROC presenta.

Observando en la Figura 18 la evolución de la métrica seleccionada a través de los parámetros se pueden extraer conclusiones acerca de la combinación más adecuada para los datos, en este caso se selecciona una red neuronal muy simple con solamente 2 nodos en la capa oculta y un parámetro de actualización de los pesos de 0.4 que obtiene un valor de ROC de 0.87.

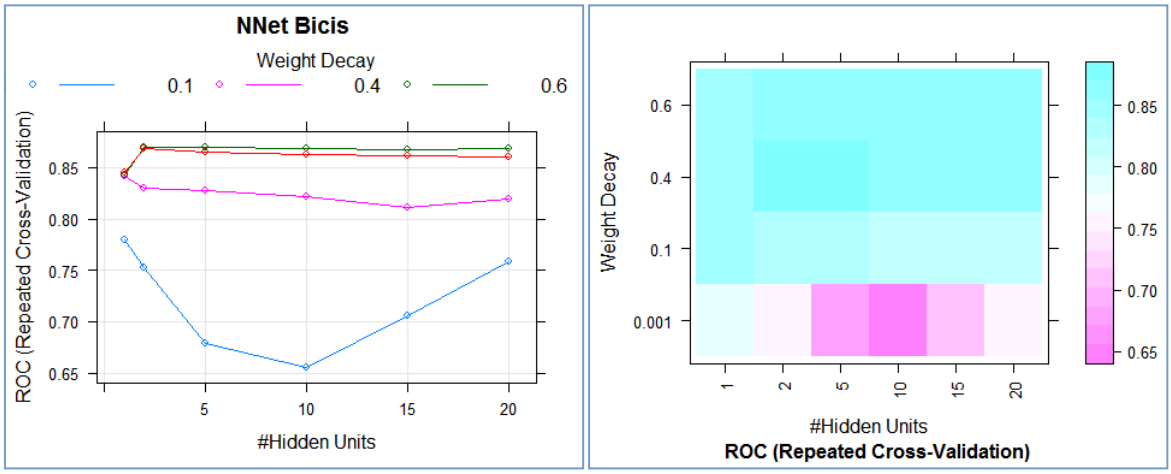
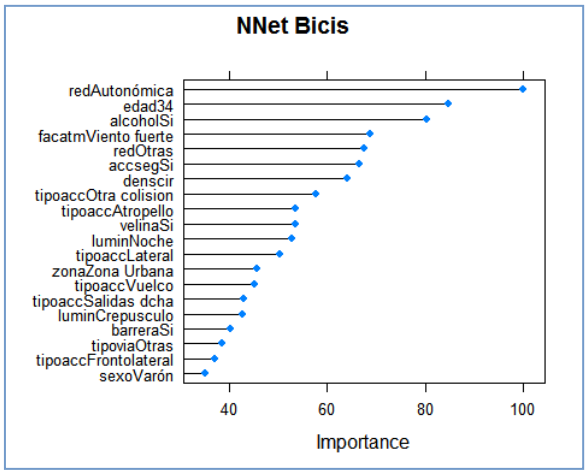


Figura 18: Gráficos de exploración de los parámetros de Red Neuronal (Bicis).



En lo referente a la importancia de variables, las redes neuronales señalan como factores influyentes la ***red autonómica***, la ***edad*** más elevada (personas mayores de 57 años), la ingesta de ***alcohol***, el ***viento fuerte***, otras redes y el uso de ***accesorios de seguridad***.

Figura 19: importancia de variables NNet (Bicis).

De manera análoga se obtiene el estudio para las subpoblaciones restantes resumiendo las características de la red neuronal óptima para cada caso y el valor de ROC.

```
mejorModelonnet

##          size decay ROC NNet
## Camiones 15   0.6   0.7746693
## Motos    10   0.6   0.873362
## Bicis    2    0.4   0.8699829
## Peatones 15   0.6   0.8554402
## Ciclos   10   0.4   0.856608
## Turismos 10   0.6   0.9056526
```

Tabla 7: Mejores modelos NNet

La capacidad de clasificación de las redes neuronales supera a la proporcionada por la técnica de LogitBoost del apartado anterior.

8.2.3 RANDOM FOREST

Continuando con el proceso de ajuste de modelos de minería de datos, se construye un procedimiento para la construcción de modelos de Random Forest para las subpoblaciones consideradas. Este algoritmo introduce dos fuentes de variación, el re muestreo de observaciones y de variables, siendo por tanto un método fiable para el control del sobreajuste. Como se explicó en el apartado de metodología, este algoritmo sortea aleatoriamente un determinado número de variables en cada nodo de cada árbol de decisión ajustado y posteriormente promedia los resultados de todos ellos dando lugar a un clasificador que, en sí, ya constituye un ensamble de modelos.

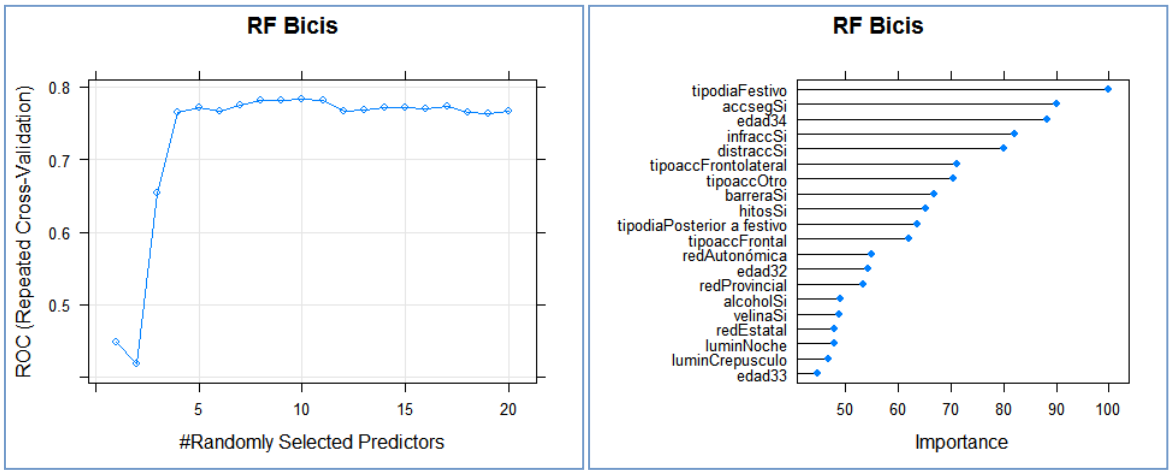


Figura 20: Número de variables a muestrear en cada partición e importancia de variables Random Forest con 100 árboles (Bicis).

En la práctica, el paquete Caret muestra como parámetro básico a monitorizar el número de variables a sortear en cada uno de los nodos, mtry. Así pues se crea una rejilla para

variar este parámetro y se construyen dos modelos de RF con distinto número de árboles, uno con 100 y otro con 500.

En el primer caso los resultados se muestran en la Figura 20 y revelan que el mejor modelo se obtiene cuando se considera 10 variables a muestrear en cada nodo, obteniendo un valor de ROC de 0.78, bastante inferior al proporcionado por las anteriores técnicas.

```
mejorModelorf

##          mtry ROC RF
## Camiones 11   0.7473441
## Motos    12   0.7770139
## Bicis    10   0.7836856
## Peatones 16   0.7880832
## Ciclos   11   0.6864795
## Turismos 13   0.7651845
```

Tabla 8: Mejores modelos RF100

En general, como se observa en la tabla, la precisión del ajuste de estos modelos no es competitiva por lo que se decide construir bosques más poblados.

En cualquier caso las variables más influyentes son **día *festivo***, ***accesorios de seguridad***, ***personas mayores***, ***infracción*** y ***distracción***.

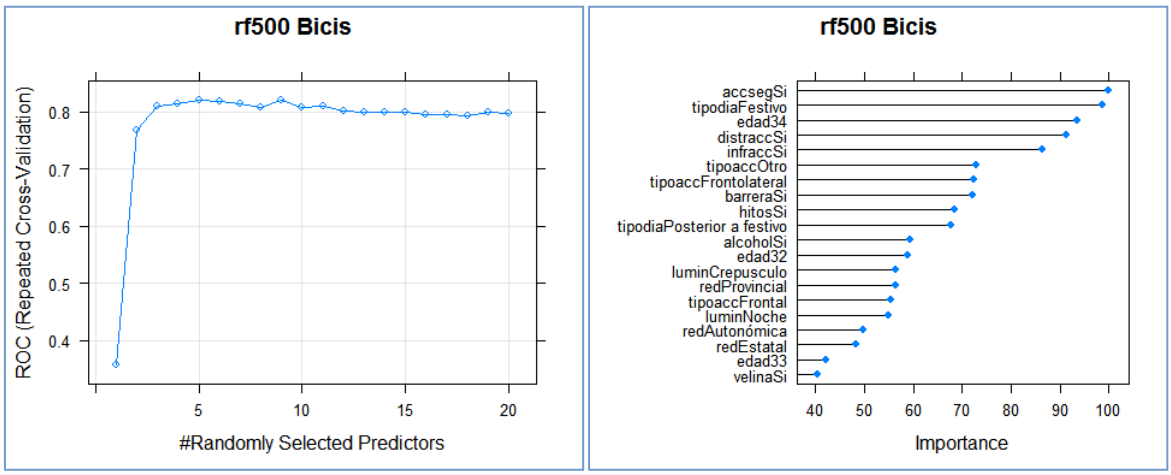


Figura 21: Número de variables a muestrear en cada partición e importancia de variables Random Forest con 500 árboles (Bicis).

Cuando se construyen los RF con 500 árboles, Figura 21, la calidad del ajuste mejora obteniendo un valor de ROC de 0.82 y se seleccionan como variables de mayor relevancia las mismas que para el modelo anterior.

Del mismo modo se ajustan estos modelos para las restantes subpoblaciones obteniendo los resultados que refleja la siguiente tabla.

```
mejorModelorf500

##          mtry ROC rf500
## Camiones 9    0.7683329
## Motos    10   0.8145995
## Bicis    9    0.8216705
## Peatones 19   0.7991191
## Ciclos   10   0.8052082
## Turismos 9    0.8121429
```

Tabla 9: Mejores modelos RF500

El valor de la métrica de ajuste ROC es mejor en general pero no supera a la técnica estrella hasta el momento, la redes neuronales.

Cabe destacar que la monitorización de los parámetros de cada árbol ajustado es importante para la obtención de mejores resultados y no es una opción por defecto de la función train() del paquete Caret.

8.2.4 GRADIENT BOOSTING

El Gradient Boosting es una de las técnicas de clasificación más potentes hoy en día, cuyos resultados suelen superar al Random Forest en muchas ocasiones. El proceso de optimización de la clasificación por medio de la disminución de la función de error o pérdida en el sentido del gradiente descendente proporciona unos resultados muy buenos evitando en gran medida el sobreajuste a los datos.

En el caso del Stocastic GB, se añade la potencia de las técnicas basadas en Bagging de tal forma que se construye cada uno de los clasificadores individuales con un conjunto de entrenamiento diferente, con las ventajas que ello reporta.

Esta es una técnica compleja en cuanto al número de parámetros a monitorizar, cuatro en el caso básico, que son el parámetro de aprendizaje llamado Shrinkage, el mínimo de observaciones en los nodos finales de cada árbol ajustado (esto controla la complejidad de los clasificadores individuales), el número de iteraciones del algoritmo (número de árboles creados) y el máximo de interacciones entre variables permitidas.

Teniendo esto en cuenta se crea una rejilla de parámetros para ser utilizada en la función train() y se programa la función que realiza este proceso de búsqueda del mejor modelo para cada una de las subpoblaciones.

Siguiendo el esquema habitual se presentan los resultados del proceso para la subpoblación de bicicletas y la tabla de resumen de los mejores modelos para cada una de las restantes.

En los gráficos de control del proceso, Figura 22, se tiene los valores de ROC para cada una de las combinaciones de los cuatro parámetros comentados y se observa que el ajuste óptimo se obtiene para un modelo con 400 iteraciones del algoritmo, con parámetro de aprendizaje de 0.01, mínimo de observaciones en los nodos finales de 10 y permitiendo interacciones de orden 3 en las variables. Se obtiene un valor de ROC de 0.85.

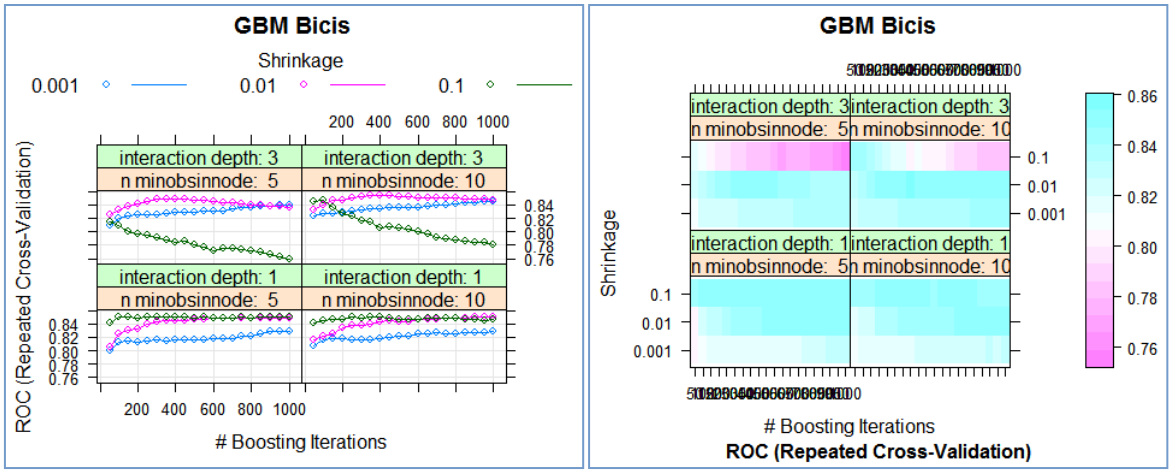
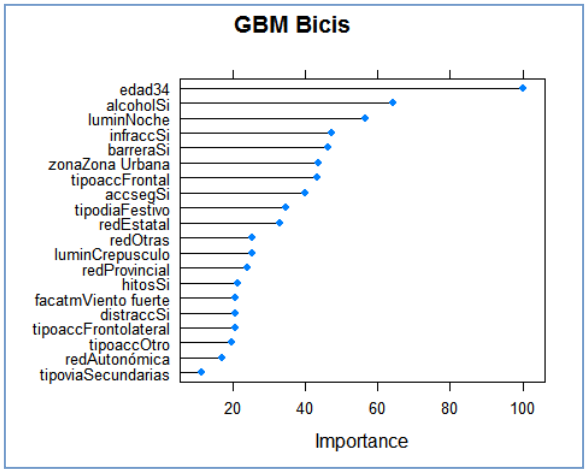


Figura 22: Gráficos de exploración de los parámetros de Gradient Boosting (Bicis).



En cuanto a la importancia de las variables, éste modelo resalta como relevantes la ***edad avanzada***, el ***alcohol***, la ***noche sin iluminación***, la ***infracción*** y la ***barrera***. En menor medida se encuentran variables como ***zona urbana***, ***colisión frontal*** y ***accesorios de seguridad***.

Figura 23: importancia de variables GBM (Bicis).

```
mejorModelogbm
```

```
##           n.trees interaction.depth shrinkage n.minobsinnode ROC GBM
## Camiones 250     1                 0.1       10             0.7812254
## Motos    1000    3                 0.01      10             0.8706486
## Bicis    400     3                 0.01      10             0.8538974
## Peatones 650     3                 0.01      5              0.8552003
## Ciclos   800     3                 0.01      10             0.8537044
## Turismos 750     3                 0.1       10             0.905152
```

Tabla 10: Mejores modelos GBM

En la tabla resumen de los mejores modelos para cada subpoblación se observa un ajuste muy bueno en general, con valores de ROC parecidos a los arrojados por las redes neuronales.

Dada la naturaleza compleja de los datos a clasificar se hace uso de una variante del algoritmo llamado Extreme Gradient Boosting que, por su construcción con un mayor suavizado está indicado para la clasificación de conjuntos de datos con falta de balanceo de las clases de interés. Así mismo este algoritmo tiene mayor velocidad de ejecución.

Como es habitual se crea una rejilla de parámetros a monitorizar dada por el número de iteraciones del algoritmo, la profundidad máxima del los arboles en cada iteración y el parámetro de aprendizaje eta.

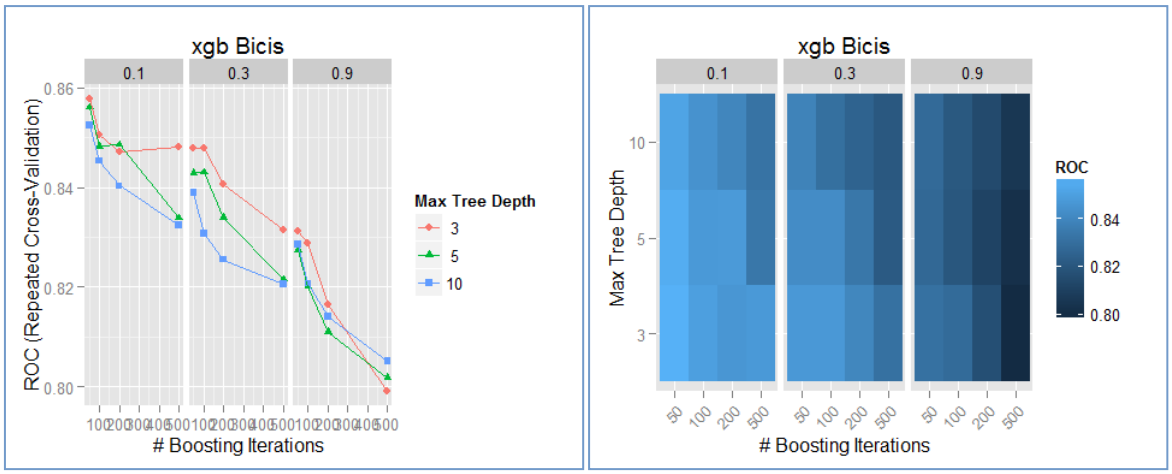
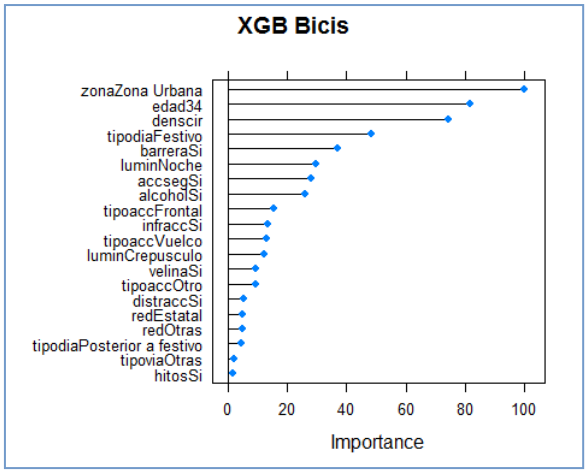


Figura 24: Gráficos de exploración de los parámetros de XGB (Bicis).

Los gráficos de control del proceso de ajuste del modelo óptimo muestran los valores de la métrica ROC para cada combinación de los parámetros comentados, eligiendo para la subpoblación de bicicletas un modelo con 50 iteraciones del algoritmo, profundidad máxima de 3 y tasa de aprendizaje de 0.1.



En cuanto a la importancia de las variables, éste modelo resalta como relevantes la ***zona urbana***, la ***edad avanzada***, el *alcohol*, la ***densidad de circulación***, **el *tipo de día festivo*** y la ***barrera.*** En menor medida se encuentran variables como ***noche sin iluminación*** y los ***accesorios de seguridad.***

Figura 25: Importancia de variables XGB (Bicis).

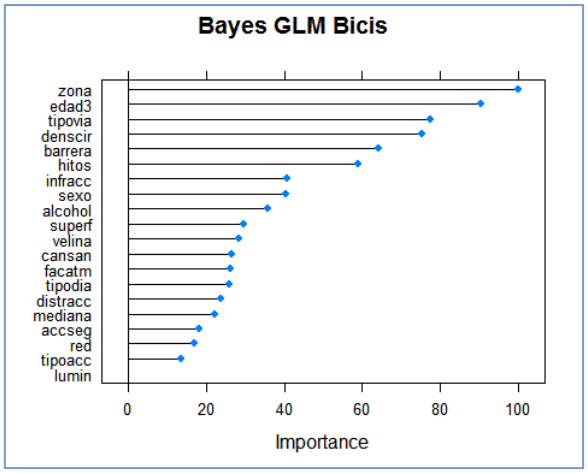
```
mejorModeloxgb

##          nrounds max_depth eta ROC XGB
## Camiones 50      5         0.1 0.7764203
## Motos    100     3         0.1 0.8779354
## Bicis    50      3         0.1 0.8577661
## Peatones 100     3         0.1 0.8607013
## Ciclos   100     3         0.1 0.8813464
## Turismos 200     5         0.1 0.9096459
```

Tabla 11: Mejores modelos XGB

En la tabla resumen de los mejores modelos para cada subpoblación se observa que el valor de la métrica ROC es muy similar a la arrojada por el modelo anterior lográndose como en todos los casos el mejor ajuste en la población de turismos debido con mucha probabilidad al elevado número de observaciones.

8.2.5 MODELO BAYESIANO

Por último y con el fin de mostrar alguna técnica basada en mayor medida en conceptos estadísticos que en aprendizaje computacional, se prueba el ajuste de un enfoque bayesiano para la estimación de modelos lineales generales. La gran diferencia de ésta técnica con las utilizadas hasta este punto es que se infiere una distribución a priori para los coeficientes estimados que serán actualizados en cada etapa del proceso de estimación por mínimos cuadrados ponderados. La distribución a priori es la t de Student, ya que se demuestra que puede solucionar el problema de separación de los datos muy frecuente en regresión logística. (Gelman, A. 2008).



En cuanto a la importancia de las variables, éste modelo resalta como relevantes la ***zona***, la ***edad,*** el ***tipo de vía***, la ***densidad de circulación*** y la ***barrera***. En menor medida se encuentran variables como ***hitos***, ***infracción*** y ***sexo***.

Figura 26: Importancia de variables Bayes GLM (Bicis).

En la práctica y con el paquete caret, este algoritmo no tiene parámetros a monitorizar con lo que proporciona un resultado directo sin la necesidad de elección del mejor modelo a través de una rejilla de parámetros a explorar.

Se presenta la tabla de resumen de los modelos ajustados para cada una de las subpoblaciones de interés en la que se observa un valor del área bajo la curva ROC muy aceptable para todas ellas.

```
mejorModelobayes

##           parameter ROC.Bayes.GLM
## Camiones          1     0.7651676
## Motos             1     0.8604090
## Bicis             1     0.8535707
## Peatones          1     0.8485578
## Ciclos            1     0.8195094
## Turismos          1     0.8901861
```

Tabla 12: Mejores modelos BayesGLM

Es relevante comentar que existen en la función bayesglm() del paquete arm de R multitud de parámetros que pueden ser variados para conseguir mejores resultados.

8.3 ESTUDIO COMPARATIVO

Como resumen de este epígrafe y tras el proceso de análisis de la importancia de las variables en los distintos modelos ajustados se propone una medida de influencia de los distintos factores sobre el resultado fatal en accidentes de tráfico en las subpoblaciones de víctimas de siniestros viales en España en el año 2012.

Se construye, para cada subpoblación del estudio, tablas que contienen las cinco variables más relevantes en cada uno de los modelos ajustados y se realiza un conteo de las frecuencias relativas de aparición de cada variable a lo largo de los ocho modelos. La idea fundamental es que las variables que aparecen como importantes en los distintos modelos con mayor frecuencia han de ser los factores que mayor influencia tienen sobre el suceso de interés a haber sido seleccionados por distintos algoritmos para crear los modelos de clasificación.

Bajo esta perspectiva se tiene en la Figura 27 que, para la **subpoblación de camiones** los factores con mayor frecuencia de aparición en el 'top 5' de la importancia de variables a lo largo de los modelos son accesorios de seguridad, barrera, hitos y salidas de vía por la izquierda. En menor medida con frecuencia de aparición 2 se encuentra factores muy propios de los accidentes de camiones como el vuelco, la nocturnidad, la distracción y la edad entre 38 y 47 años.

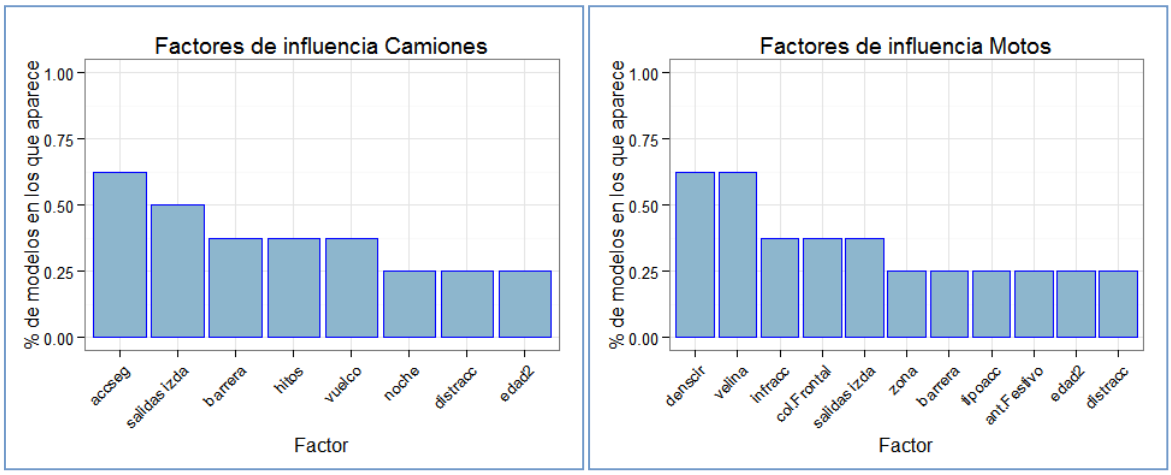


Figura 27: Factores de influencia en las subpoblaciones de Camiones y Motos.

En cuanto a la **subpoblación de motos**, los factores que se consideran de mayor influencia en el resultado de los siniestros de tráfico con esta metodología resultan ser la densidad de circulación y la velocidad inadecuada con elevada presencia, la infracción del conductor y elementos como la barrera, la zona, la edad de 38 a 47 en tipos de accidentes como colisiones frontales y salidas de vía por la izquierda.

La Figura 28 contiene los factores de influencia para la **subpoblación de bicis**, en la que la edad parece jugar un papel fundamental en el resultado del siniestro, apareciendo

su categoría más elevada (a partir de 57 años) como la de mayor influencia y siendo considerada así mismo en otros modelos (que no realizan la descomposición en variables indicador) como el logitBoost y el bayesGLM. Por otro lado se extraen **escenarios de accidentalidad** como la presencia de barrera, la densidad de circulación el día festivo y perfiles de accidentados con la edad comentada, infracción, distracción y presencia de alcohol y accesorios de seguridad.

En el caso de los **peatones**, esta metodología de extracción de factores de influencia, revela que la densidad de circulación, la nocturnidad con falta de iluminación, las personas mayores, las distracciones, los hitos y la utilización de accesorios de seguridad son las variables de mayor influencia.

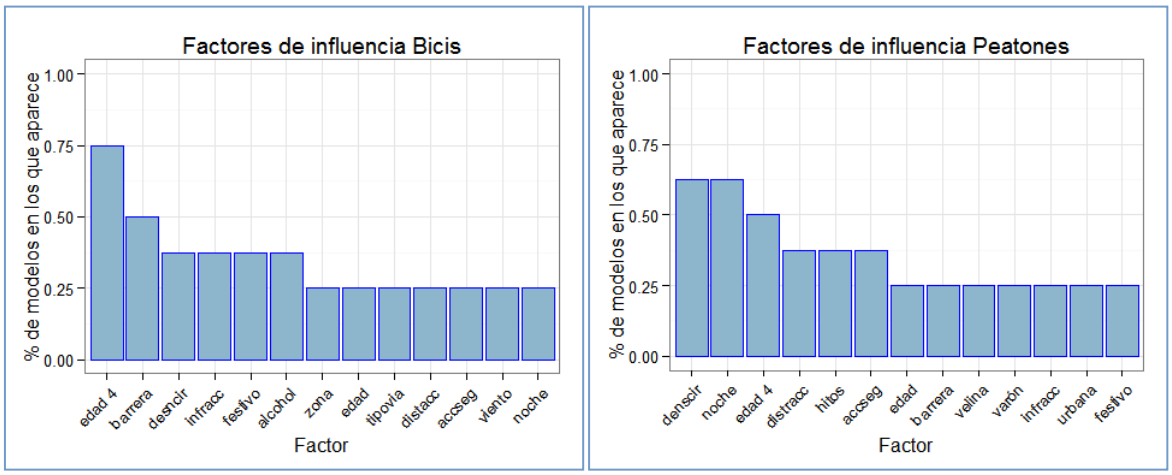


Figura 28: Factores de influencia en la subpoblaciones de Bicis y Peatones.

En la **subpoblación de ciclomotores** (Figura 29), por su parte, tienen especial relevancia factores como la utilización de accesorios de seguridad, el sexo, las distintas edades (con mucha influencia), la densidad de circulación y la infracción. Destacan también escenarios como vías urbanas en días festivos con presencia de cansancio o distracción.

Por último, en la **subpoblación de turismos** es muy relevante la utilización de accesorios de seguridad, siendo las colisiones frontales el tipo de accidente que resulta más influyente en la clasificación de los siniestros. Por otro lado la presencia de barrera o hitos en la calzada, la densidad de circulación o el sexo varón presentan también relativa importancia.

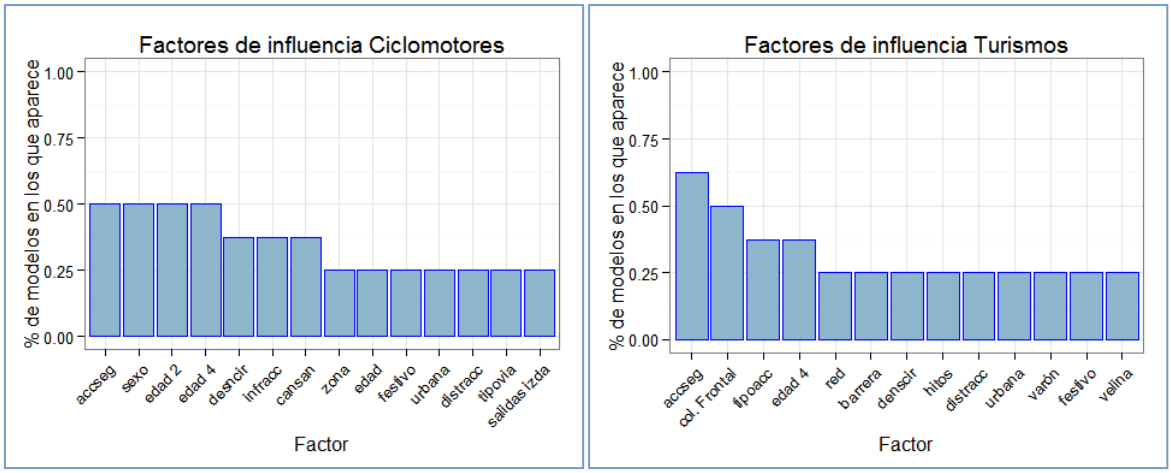


Figura 29: Factores de influencia en la subpoblaciones de Ciclomotores y Turismos.

Se han puesto de manifiesto, mediante esta metodología, los factores tanto propios de la vía como inherentes al conductor que presentan una mayor influencia en la clasificación del evento de interés en el estudio.

Cabe destacar las ventajas e inconvenientes de este método para la selección de factores de influencia ya que por un lado supone un método robusto para esta tarea ya que es un compendio de técnicas, y no un solo algoritmo, el que ha decidido extraer esa variables como influyente, evitando así posibles fallos o sesgos en la selección de variables de cada uno de ellos de manera individual.

Por otra parte, la principal desventaja de este método es la imposibilidad de cuantificar la importancia de estas variables así como el sentido de influencia en la clasificación del evento de interés debido a que se seleccionan para realizar la partición digamos en un nodo (en el caso de los métodos basados en árboles) pero es difícil saber si esa categoría de esa variable desemboca en un aumento o en un detrimento de la probabilidad estimada. Este hecho no tiene especial importancia ya que se dispone de métodos complementarios como la inspección descriptiva de la población y la interpretación de los OR de la regresión logística, con los que se puede decidir ese sentido de influencia.

9. CAPACIDAD DE CLASIFICACIÓN

A lo largo de este apartado se evalúa la capacidad de clasificación de los modelos creados sobre los datos de las subpoblaciones de interés desde dos perspectivas distintas. La clasificación automática y la clasificación mediante probabilidades estimadas.

9.2 CLASIFICACIÓN AUTOMÁTICA

Por un lado se estudia el proceso de ajuste por remuestreo con validación cruzada repetida de los modelos de minería de datos mediante gráficos, midiendo el valor de la métrica seleccionada (ROC) a través de las muestras creadas. Esta es la forma habitual de medir la calidad del ajuste de los modelos a los datos y también evaluar el sobreajuste por medio del estudio de los intervalos de confianza para la métrica ROC en su distribución sobre las muestras de validación (en este caso 12).

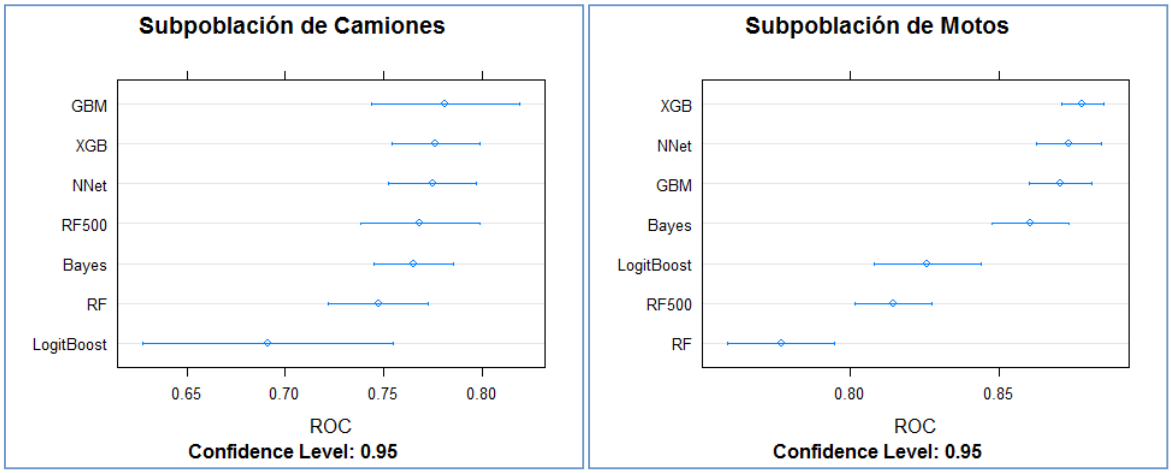


Figura 30: Comparativa de precisión (ROC) en Camiones y Motos.

En los gráficos anteriores se observa que en la subpoblación de camiones el mayor valor de la métrica ROC es alcanzado por el método de Gradient Boosting, sin embargo debido a la elevada dispersión quizá en Extreme Gradient Boosting sea más competitivo a la hora de generalizar. La subpoblación de motos tiene un claro ganador el XGB que presenta un valor de ROC de 0,88. Como se comprobará posteriormente el valor de la sensibilidad de estos modelos es muy bajo.

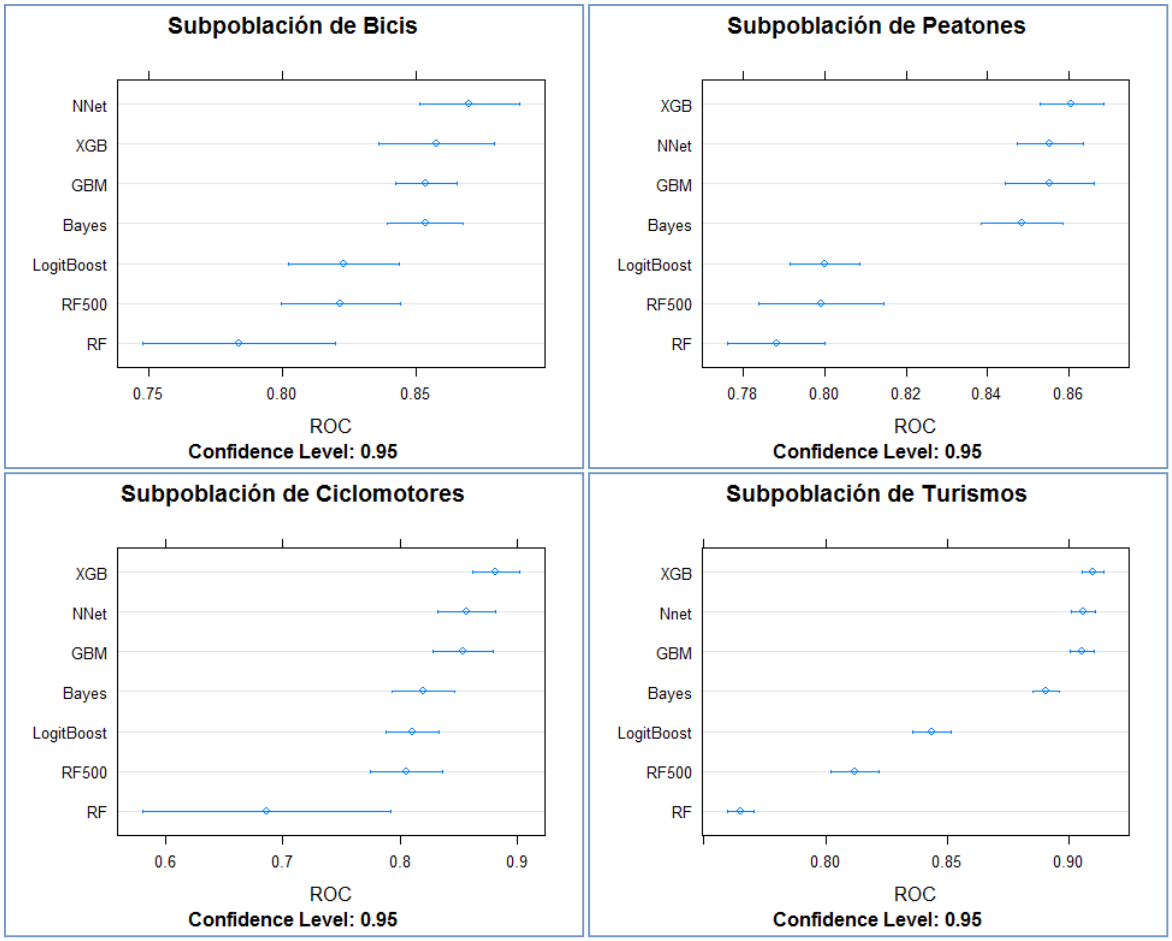


Figura 31: Comparativa de precisión (ROC) en Bicis, Peatones, Ciclos y Turismos.

El resultado para las demás subpoblaciones es el mismo siendo el algoritmo Extreme Gradient Boosting el que mejor ajusta en todas ellas a excepción de las bicis, caso en el que las Redes Neuronales funcionan aparentemente mejor.

Cuando el conjunto de datos a estudio es balanceado, es decir, que existe equilibrio en la distribución de las clases a predecir, éste método de evaluación arroja buenos resultados en cuanto a capacidad de clasificación, no ocurriendo de la misma forma en la clasificación de clases minoritarias, que es precisamente el escenario de este estudio.

9.1 CLASIFICACIÓN MEDIANTE PROBABILIDADES ESTIMADAS.

En este caso, al ser el suceso de interés muy poco probable, las probabilidades estimadas suelen ser bajas y esto hace que el punto de corte que maximiza la relación entre sensibilidad y especificidad de la clasificación no sea el 0.5.

Por ello y para una mejor clasificación, se programan funciones de predicciones (Anexo III) que genera las matrices de confusión de las clasificaciones con dos puntos de corte para la probabilidad estimada distintos, el primero será la prevalencia del evento en la subpoblación y el segundo a elección del usuario (por defecto 0.5), así mismo recurre a la función ROC para dibujar la curva y estimar el punto de corte óptimo para la probabilidad, el que hace máxima la relación entre sensibilidad y especificidad para la probabilidad estimada frente a la clase real.

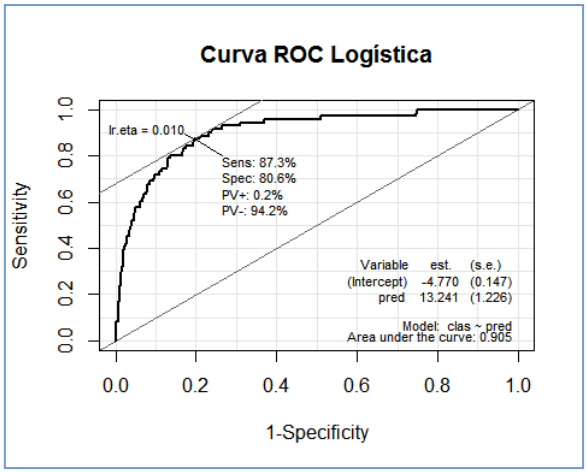
Se puede comprobar que existe poca diferencia entre este valor estimado y la prevalencia a priori del evento que, sin embargo producen grandes cambios en número de mal clasificados. Este hecho es habitual en conjuntos de datos no balanceados debido a las bajas probabilidades estimadas que crean una la frontera de decisión muy dispersa entre las clases a predecir y esta poca consistencia de la estimación del punto de corte es la mayor pega de la utilización de este método.

En cualquier caso se considera muy superior la capacidad de clasificación del evento por medio de este procedimiento, máxime al tratarse un suceso fatal que ha de ser evitado. Es lógico actuar bajo la premisa de que el coste de mala clasificación es muy superior en la clase de interés y por ello ha de relajarse el umbral para la especificidad incurriendo en mayor tasa de falsos positivos.

Para ilustrar el proceso de extracción de predicciones y clasificación óptima por medio del procedimiento programado, se presentan los principales resultados para una de las técnicas, la regresión logística en la subpoblación de bicicletas.

Cabe destacar que la forma de obtener las probabilidades estimadas varía en algunas de las técnicas por lo que es necesario programar más de una función de predicciones. En cualquier caso los resultados obtenidos tienen siempre la misma estructura.

En Primer lugar se presenta la curva ROC para el ajuste de la probabilidad estimada frente a la clase real con el punto de corte óptimo.



En la Figura 32 se observa que el punto de corte de la probabilidad estimada que hace máxima la suma de sensibilidad y especificidad resulta ser 0,010, con valores de 87,3 y 80,6% respectivamente, produciendo un área bajo la curva de 0,905 lo que implica un ajuste muy aceptable del modelo a los datos.

Figura 32: Gráfico curva ROC del modelo logístico para bicicletas

Seguidamente se obtienen las matrices de confusión de las clasificaciones obtenidas con el punto de corte de la probabilidad estimada igual a la prevalencia a priori en la población y con otro punto de corte a elegir, que por defecto será el clásico 0,5.

De esta forma se obtienen tres clasificaciones posibles y se compara la capacidad de cada una de ellas. Se constata el hecho comentado anteriormente de la falta de consistencia en la clasificación ante pequeñas variaciones en el punto de corte seleccionado y la falta de adecuación de la clasificación estándar (con el punto de corte en 0,5) en la clasificación de eventos poco probables.

```
## Confusion Matrix and Statistics
##
## predClas   No   Si
##       No 4269   11
##       Si  932   60
##
##                Accuracy : 0.8211
##                  95% CI : (0.8105, 0.8314)
##     No Information Rate : 0.9865
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.09
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.84507
##             Specificity : 0.82080
##          Pos Pred Value : 0.06048
##          Neg Pred Value : 0.99743
##              Prevalence : 0.01347
##          Detection Rate : 0.01138
##    Detection Prevalence : 0.18816
##       Balanced Accuracy : 0.83294
##
##        'Positive' Class : Si
```

Tabla 13: Matriz de confusión prevalencia a priori. Regresión logística. Bicis

En la anterior salida del procedimiento (Tabla 13) se concluye que la prevalencia del evento en la población es del 1,36%, constituyendo un fenómeno de muy difícil modelización. Con este punto de corte para la probabilidad estimada se obtiene una sensibilidad del 84,5% y una especificidad del 82,1%, valores que difieren de los comentados anteriormente. Se clasifican correctamente 60 de las 71 observaciones de la clase de interés y 4269 de las 5201 observaciones de la clase mayoritaria. Quedará a elección del analista la clasificación a utilizar finalmente, siendo esta última más conservadora en cuanto a falsos positivos.

```
##  Moviendo el punto de corte de la probabilidad estimada a  0.5
##
## Confusion Matrix and Statistics
##
## predClas   No   Si
##       No 5200   68
##       Si    1    3
##
##                Accuracy : 0.9869
##                  95% CI : (0.9835, 0.9898)
##     No Information Rate : 0.9865
##     P-Value [Acc > NIR] : 0.4363
##
##                   Kappa : 0.0787
##  Mcnemar's Test P-Value : 1.935e-15
##
##             Sensitivity : 0.0422535
##             Specificity : 0.9998077
##          Pos Pred Value : 0.7500000
##          Neg Pred Value : 0.9870919
##              Prevalence : 0.0134674
##          Detection Rate : 0.0005690
##    Detection Prevalence : 0.0007587
##       Balanced Accuracy : 0.5210306
##
##        'Positive' Class : Si
```

Tabla 14: Matriz de confusión clasificación automática. Regresión logística. Bicis

En cuanto a la clasificación obtenida por medio de la separación de las clases para el punto de corte de la probabilidad en 0,5 (Tabla 14), se observa que la sensibilidad arrojada es muy baja reconociendo solamente a 3 de los 71 fallecidos por lo que no se considera válido este método para los datos de este estudio.

Realizando este procedimiento para todos los modelos en todas las subpoblaciones se obtienen los valores del estadístico c o área bajo la curva ROC para el punto de corte óptimo de la probabilidad estimada. En total se han realizado 8 predicciones para cada una de las 8 subpoblaciones, con lo que se han ajustado 48 modelos finales en estos datos escogidos de entre cientos probados.

ÁREA BAJO LA CURVA ROC CON EL PUNTO DE CORTE ÓPTIMO DE LA PROBABILIDAD ESTIMADA								
	Logistica	LogiBoost	NNet	RF100	RF500	GBM	XGB	Bayes
Camiones	0,84	0,79	0,85	0,7	0,76	0,85	0,91	0,86
Motos	0,88	0,83	0,93	0,74	0,78	0,87	0,9	0,88
Bicis	0,9	0,84	0,94	0,74	0,82	0,92	0,93	0,91
Ciclomotores	0,86	0,78	0,91	0,76	0,79	0,87	0,89	0,86
Peatones	0,88	0,85	0,96	0,66	0,75	0,93	0,95	0,9
Turismos	0,89	0,86	0,94	0,72	0,77	0,93	0,95	0,9

Tabla 15: Comparativa de precisión (ROC) global. Punto de corte óptimo.

Es evidente que el resultado de la clasificación por esta vía de elección del punto de corte óptimo para la probabilidad estimada es, en general, mucho mejor. Aumenta en todos los casos el valor de AUC, en especial en las Redes Neuronales que consiguen superar a algoritmo XGB que parecía ser el claro vencedor.

Así, NNet es el algoritmo que mejor ajusta a las subpoblaciones de Motos bicis Peatones y ciclomotores y XGB lo hace mejor en Camiones con mucha diferencia, y en Turismos con no tanta.

10. ENSAMBLE DE MODELOS

Como último apartado del estudio, se proponen aquí algunos métodos de ensamble de los modelos anteriormente ajustados con el fin de obtener clasificadores cuya relación entre sensibilidad y especificidad sea mayor que la proporcionada por los modelos individuales. Para ello se construye un conjunto de datos que contiene las probabilidades estimadas por los ocho modelos ajustados para cada subpoblación con el fin de combinar estas probabilidades de distintas formas. Sin ánimo de profundizar mucho en los modelos de ensamble óptimos, se proponen varios de estos posibles clasificadores combinados y se compara su capacidad.

En primer lugar se construye un clasificador dado por la media aritmética de las probabilidades estimadas por cada uno de los modelos individuales, llamado **ensamble medio (EnsMean)**. A continuación se realizará un ajuste de regresión logística por pasos, backward, con las ocho probabilidades estimadas como predictores para la clasificación del evento y se construirá un clasificador que viene dado por la media ponderada por pesos obtenidos por los coeficientes de la regresión obtenidos, de forma relativa. Este clasificador se llamará **ensamble regresión (EnsRegw)**. Así mismo se

considerará la probabilidad estimada de este modelo de regresión logística como otro posible **ensamble logístico (EnsRegPred)**.

Por último se construye un clasificador dado por la media de los dos mejores modelos para cada subpoblación y otro clásico ensamble dado por la media ponderada de Gradient Boosting y Random Forest que se considera interesante debido a las distintas formas de actuación de sendos algoritmos, estando el primero orientado a reducir la varianza de las estimaciones y con la ventaja de la selección por sorteo de variables del segundo. Se consideran los pesos de 0,8 y 0,2 respectivamente (**Ens2080**).

Antes de crear los ensambles se realiza un estudio de correlaciones entre las probabilidades estimadas ya que conviene integrar los resultados de los modelos que presente mayor independencia para contrarrestar errores de clasificación.

Se presenta el gráfico de correlaciones entre las probabilidades estimadas de todos los modelos ajustados para la subpoblación de bicis. En este caso, y sin ánimo de generalizar, se observa correlación moderada del XGB con la mayoría de los modelos por lo que los ensambles resultan adecuados. Existe alta correlaciones entre algunos de los modelos.

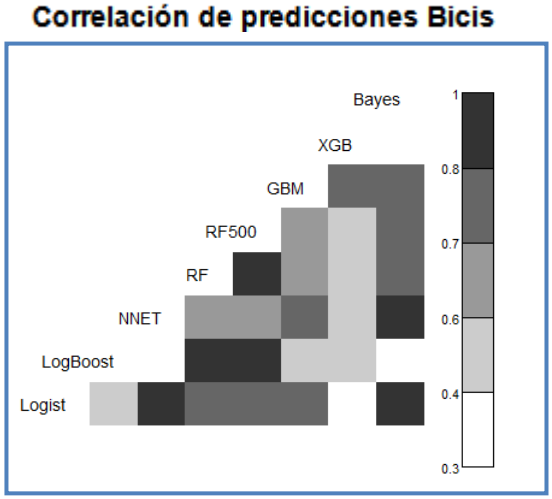


Figura 33: Correlaciones predicciones

Se programa una función (Anexo III) para que realice las tareas de búsqueda de correlaciones y creación de ensambles y se resumen los resultados obtenidos desde la perspectiva de la clasificación mediante probabilidades estimadas en la siguiente tabla.

En la Tabla 15 se observa que existe un empate en frecuencia de primer puesto en cuanto a precisión en la clasificación medida por el área bajo la curva ROC entre los dos ensambles logísticos debido a su similar construcción. Solamente en la subpoblación de ciclomotores se consigue superar al mejor de los algoritmos (NNet), que presentaba un valor de ROC de 0,91. El ensamble mediante la probabilidad estimada del modelo de regresión logística stepwise consigue un valor de 0,96, con una clasificación del evento muy buena.

	Comparativa ensambles			
	EnsRegPred	EnsRegw	EnsMean	Ens2080
Camiones	0,85	0,89	0,87	0,85
Motos	0,93	0,91	0,9	0,89
Bicis	0,92	0,93	0,92	0,92
Ciclomotores	**0,96**	0,94	0,93	0,93
Peatones	0,9	0,89	0,88	0,87
Turismos	0,93	0,95	0,94	0,93

Tabla 16: Tabla Comparativa de ensambles

Se constata la necesidad de crear modelos de ensamble para posible mejora del ajuste obtenido en la calcificación de eventos de baja incidencia.

11. PRINCIPALES CONCLUSIONES

Tras la creación y ejecución de la **metodología de estudio** de tablas de siniestralidad vial propuesta en este trabajo se destacan las conclusiones más relevantes que responden a los objetivos marcados al inicio.

En primer lugar, en lo relativo a la **determinación de factores de riesgo** en la mortalidad en accidentes de tráfico, es preciso resaltar los **perfiles de víctimas** y **escenarios de accidentalidad** que se han extraído para cada subpoblación.

Los camiones presentan un perfil de víctima de entre 37 y 49 años en la que se constata distracción al volante y falta de utilización de accesorios de seguridad y un escenario de accidentalidad caracterizado por la nocturnidad con vuelco y salida de vía por la izquierda como tipos de accidente de mayor riesgo y elementos en la vía como barrera, hitos de arista.

Las motos presentan un perfil de víctima de entre 38 y 47 años en la que se constata alguna infracción al volante y velocidad inadecuada y un escenario de accidentalidad caracterizado por la densidad de circulación con colisión frontal y salida de vía por la izquierda como tipos de accidente de mayor riesgo y elementos en la vía como barrera.

Las bicis presentan un perfil de víctima mayor de 57 años en la que se constata distracción, falta de utilización de accesorios de seguridad, infracción y alcohol y un

escenario de accidentalidad caracterizado por día festivo con densidad de circulación y elementos en la vía como barrera de seguridad.

Los peatones presentan un perfil de víctima varón mayor de 57 años en la que se constata distracción y falta de utilización de accesorios de seguridad y un escenario de accidentalidad caracterizado por la nocturnidad con falta de iluminación en zona urbana en día festivo con densidad de circulación y elementos en la vía como barrera, hitos de arista.

Los ciclomotores presentan un perfil de víctima varón de entre 37 y 49, o mayor de 57 años en la que se constata falta de utilización de accesorios de seguridad, infracción y distracción al volante y un escenario de accidentalidad caracterizado por densidad de circulación en zona urbana en día festivo.

Los turismos presentan un perfil de víctima varón mayor de 57 años en la que se constata falta de utilización de accesorios de seguridad y distracción al volante y un escenario de accidentalidad caracterizado por la densidad de circulación con colisión frontal como tipo de accidente de mayor riesgo y elementos en la vía como barrera, hitos de arista.

En lo que se refiere a **capacidad de clasificación**, en general se considera muy buena utilizando el método de clasificación por probabilidades estimadas, siendo las técnicas que mejores resultados han arrojado el Extreme Gradient Boosting que logra la mejor medida de bondad de ajuste en las subpoblaciones de camiones y turismos y las Redes Neuronales que los hacen en las cuatro restantes.

El **ensamble de modelos** a través de las predicciones de regresión logística con las probabilidades estimadas como variables predictoras en la clasificación del evento de interés mejora sensiblemente el ajuste en la subpoblación de ciclomotores.

Respecto al problema de la clasificación de las **clases poco representadas** se ha constatado la mejora en precisión aportada por las técnicas de minería de datos utilizadas frente al modelo clásico de regresión logística.

Por último, se destaca que el **lenguaje R** supone una potente herramienta para el manejo de grandes volúmenes de datos y la creación de procedimientos para el ajuste de

modelos de minería de datos. En este punto el paquete caret constituye un interfaz común y de fácil manejo con utilidades para el preprocesamiento, ajuste y validación de decenas de modelos predictivos en minería de datos.

12. BIBLIOGRAFÍA Y REFERENCIAS

[1] Ali S. Al-Ghamdi (2002) "Using logistic regression to estimate the influence of accident factors on accident severity". King Saud University, Saudi Arabia. Accident Analysis and Prevention, 34. pp. 729-741.

[2] Beltrán Pascual, M., Muñoz Alamillos, Á. & Muñoz Martínez, A (2012) "Un nuevo clasificador de préstamos bancarios a través de la minería de datos". Departamento de Economía aplicada y Estadística. UNED.

[3] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. G. (1984). "Classification and Regression Trees". Wadsworth International Group, Belmont, California, USA.

[4] Breiman, l. (1996): "Bagging predictors. Machine Learning", Kluwer Academic Publishers, Boston, Manufactured in the Netherlands, vol. 24, 2, pp. 123-140.

[5] Breiman, L. (1996). "Stacked Regression", Machine Learning, 24, pp. 49-64.

[6] Breiman, L. (2001). "Random Forests". Machine Learning, 45, pp. 5-32.

[7] Chang, L.Y. & Wang, H.W. (2006). "Analysis of traffic injury severity: an application of non-parametric classification tree techniques". Accident Analysis and Prevention 38, pp. 1019–1027.

[8] De Oña, J., Mujalli, R. & Calvo, F. (2011). "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks". Accident Analysis and Prevention, 43, pp. 402-411.

[9] Domingos, P. (1999). "MetaCost: a general method for making classifiers cost-sensitive", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.155-164. ISBN:1-58113-143-7. [doi>10.1145/312129.312220].

[10] Dzeroski, S. & Zenko, B. (2004). "Is combining classifiers with stacking better than selecting the best one?". Machine learning, 54, pp. 225-273. Kluwer Academic Publishers.

[11] Estabrooks, A., Jo, T. & Japkowicz, N., (2004), "A Multiple Resampling Method for Learning from Imbalances Data Sets", Computational Intelligence, Volume 20, Number 1, February 2004, pp.18-36.

[12] Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. "Do we need hundreds of classifiers to solve real world classification problems?". Journal of Machine Learning Research, 15. pp. 3133–3181.

[13] García Jiménez V. (2010) "Distribuciones de Clases No Balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje". Tesis doctoral. Universitat Jaume I.

[14] Gelman, A. et all. (2008) "A weakly informative default prior distribution for logistic and other regression models".The Annals of Applied Statistics 2008, Vol. 2, No. 4, 1360–1383.

[15] Hatsie, T., Tibshirani, R. & Friedman J. (2009) "The elements of Statistical Learning. Data mining, inference and prediction". ED. Springer. Second Edition.

[16] Hosmer, D.W. & Lemeshow, S. (2000). "Applied Logistic Regression". John Wiley and Sons.

[17] J. Wang, M. Xu, H. Wang and J. Zhang (2006). "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding". Proc. 8th Int. Conf. Signal Process., vol. 3, pp. 1815 -1818.

[18] Kuhn, M. (2008) "Building Predictive Models in R Using the caret Package" Journal of Statistical Software. Volume 28.

[19] Kuhn, M & Johnson, K. (2013) "Applied Predictive Modeling". Ed. Springer.

[20] Kuhnert, P., Do, K. & McClure, R. (2000) "Combinig non-parametric models with logistic regression: an application to motor vehicle injury data". Journal Computational Statistics & Data Analysis archive Vol. 34 Issue 3, Sept. 28, pp. 371-386. Elsevier Science.

[21] Kurt, I., Ture, M. & Kurum, A. T. (2008). "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications". Expert systems Appl., 34, pp. 366–374.

[22] Santana J.S & Mateos E. (2014) "El arte de programar en R: Un lenguaje para la estadística". ISBN 978-607-9368-15-9.

[23] Serna Pineda, S. C. (2009). "Comparación de árboles de clasificación y regresión y regresión logística". Trabajo presentado para optar a título de Magister en Estadística. Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

[24] Silva Aycaguer, L. C. (1995). "Excursión a la Regresión Logística en Ciencias de la Salud". Ed. Díaz de Santos.

[25] Williams, G. (2011) "Data mining with Rattle and R. The art of excavating data for knowledge discovery". Ed. Springer.

ANEXOS

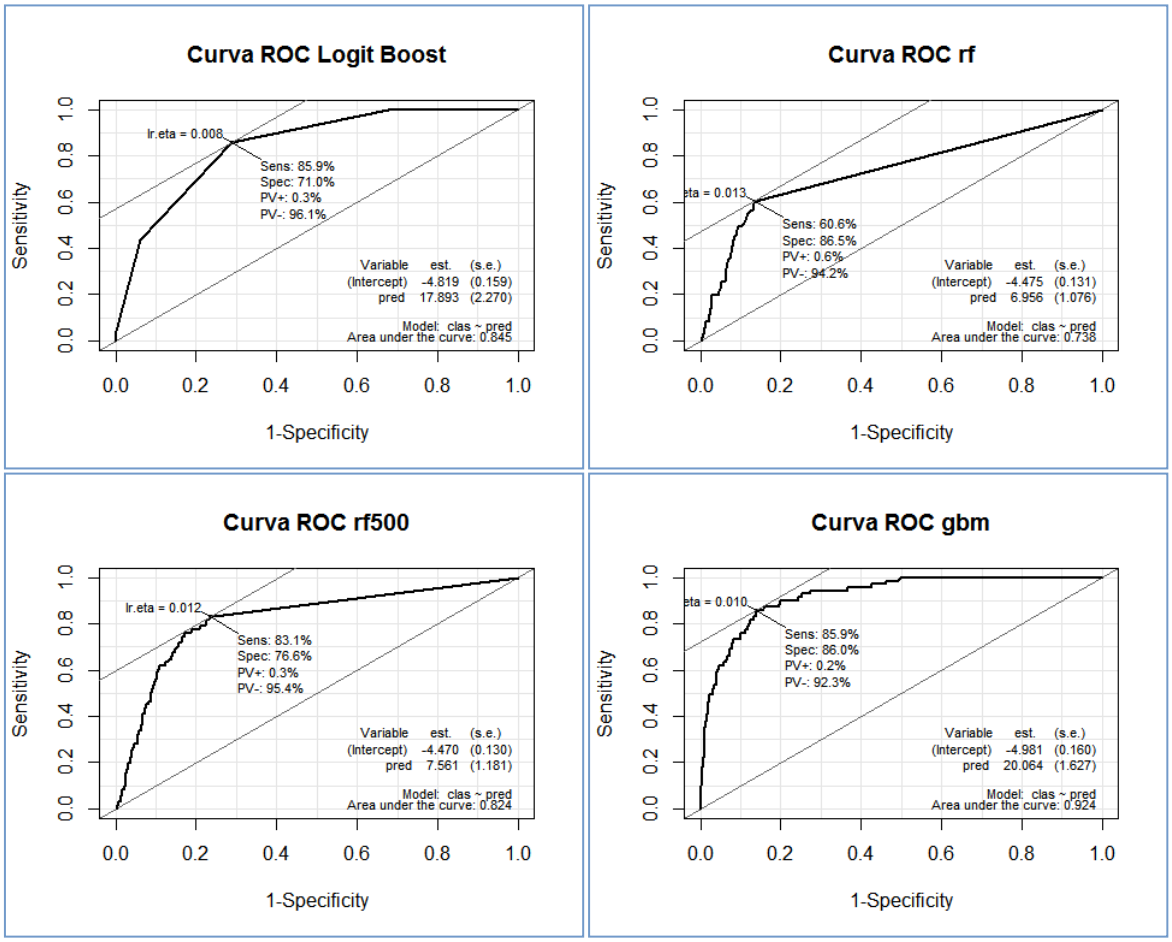
Metodología de minería de datos para el estudio de tablas de siniestralidad vial

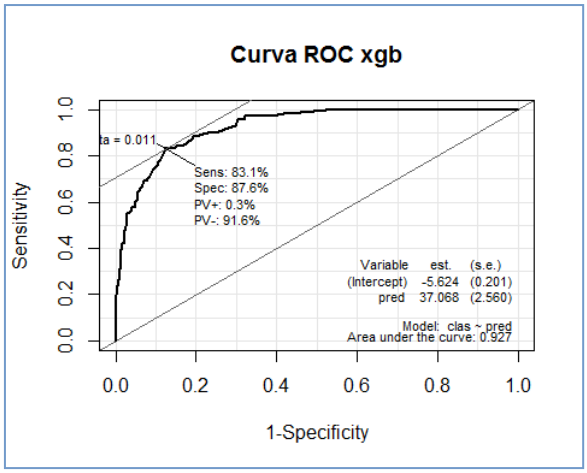
ANEXO I: PRINCIPALES RESULTADOS

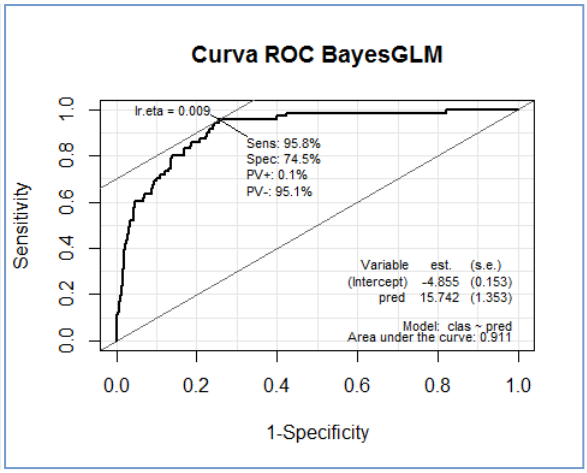
Se ha considerado conveniente incluir en este anexo los principales resultados obtenidos para las restantes subpoblaciones del estudio en lo que se refiere a determinación de los factores de influencia en el resultado de muerte en accidentes de tráfico con víctimas y a la capacidad de clasificación de cada técnica a través del método de probabilidades estimadas presentando la curva ROC con el punto de corte de la probabilidad estimada que se considera óptimo en cuanto a la relación entre sensibilidad y especificidad logradas.

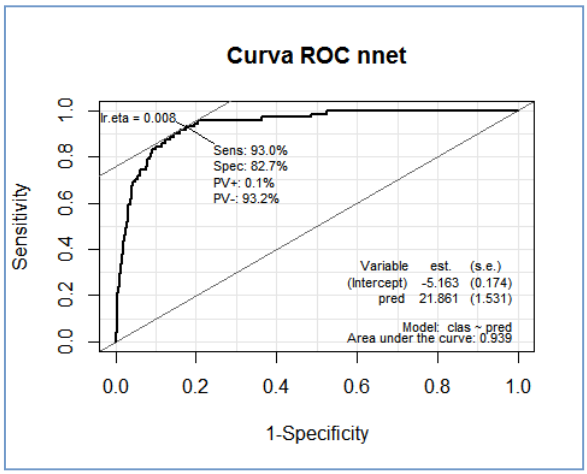
Se incluyen así mismo los resultados para la subpoblación de bicicletas que no han sido mostrados a lo largo del estudio.

BICIS

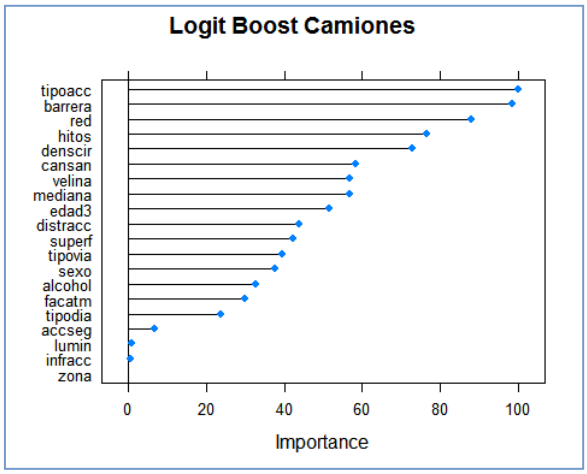


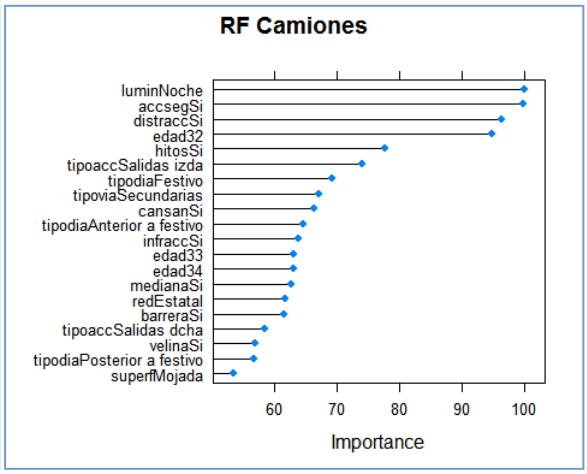

Curva ROC xgb
lr.eta = 0.011
Sens: 83.1%
Spec: 87.6%
PV+: 0.3%
PV-: 91.6%
Variable est. (s.e.)
(Intercept) -5.624 (0.201)
pred 37.068 (2.560)
Model: clas ~ pred
Area under the curve: 0.927
Sensitivity
1-Specificity

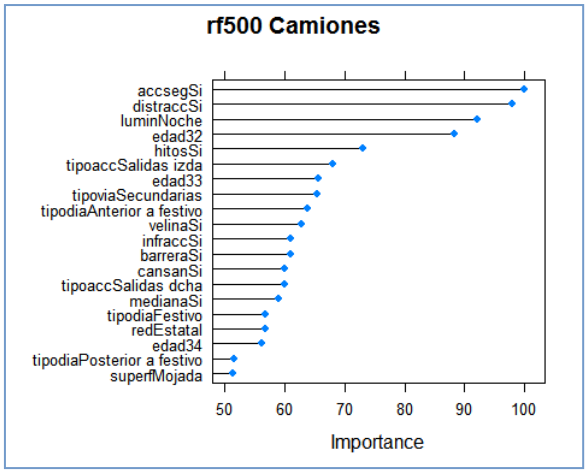

Curva ROC BayesGLM
lr.eta = 0.009
Sens: 95.8%
Spec: 74.5%
PV+: 0.1%
PV-: 95.1%
Variable est. (s.e.)
(Intercept) -4.855 (0.153)
pred 15.742 (1.353)
Model: clas ~ pred
Area under the curve: 0.911
Sensitivity
1-Specificity

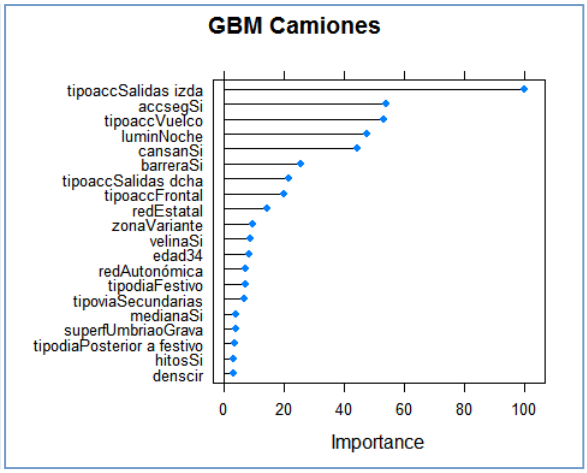

Curva ROC nnet
lr.eta = 0.008
Sens: 93.0%
Spec: 82.7%
PV+: 0.1%
PV-: 93.2%
Variable est. (s.e.)
(Intercept) -5.163 (0.174)
pred 21.861 (1.531)
Model: clas ~ pred
Area under the curve: 0.939
Sensitivity
1-Specificity

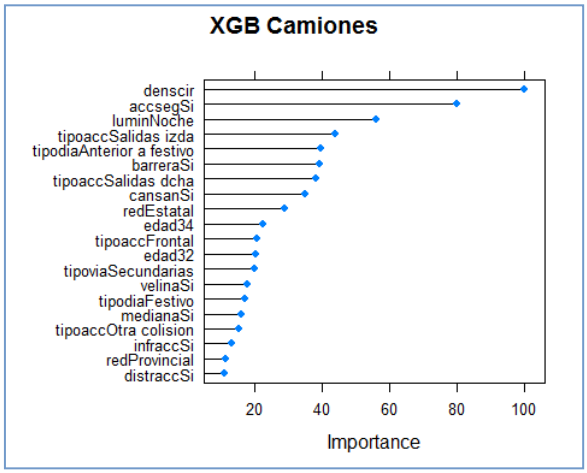
CAMIONES

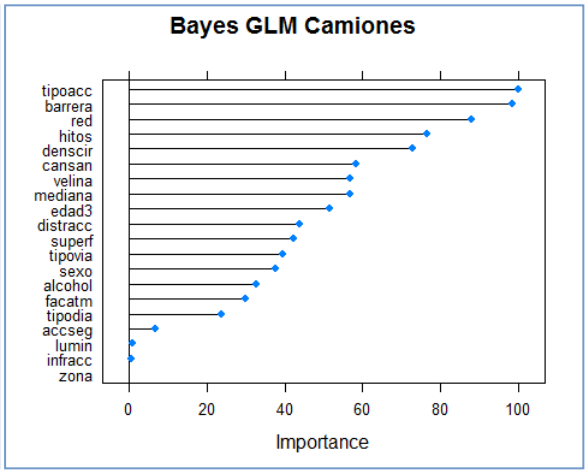

Logit Boost Camiones
tipoacc
barrera
red
hitos
denscir
cansan
velina
mediana
edad3
distracc
superf
tipovia
sexo
alcohol
facatm
tipodia
accseg
lumin
infracc
zona
Importance

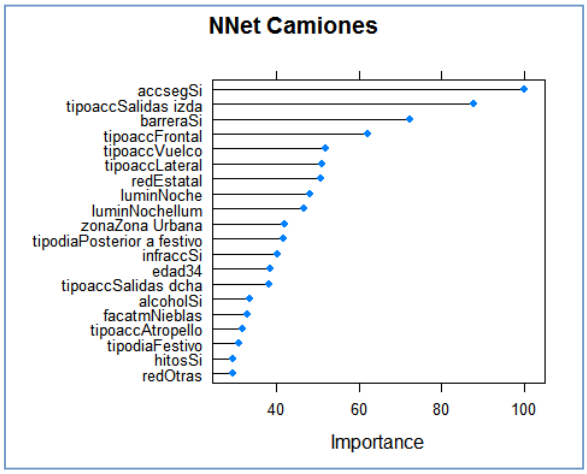

RF Camiones
luminNoche
accsegSi
distraccSi
edad32
hitosSi
tipoaccSalidas izda
tipodiaFestivo
tipoviaSecundarias
cansanSi
tipodiaAnterior a festivo
infraccSi
edad33
edad34
medianaSi
redEstatal
barreraSi
tipoaccSalidas dcha
velinaSi
tipodiaPosterior a festivo
superfMojada
Importance

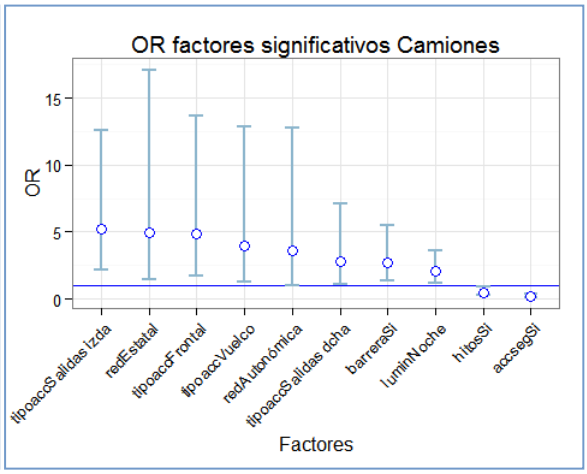

rf500 Camiones
accsegSi
distraccSi
luminNoche
edad32
hitosSi
tipoaccSalidas izda
edad33
tipoviaSecundarias
tipodiaAnterior a festivo
velinaSi
infraccSi
barreraSi
cansanSi
tipoaccSalidas dcha
medianaSi
tipodiaFestivo
redEstatal
edad34
tipodiaPosterior a festivo
superfMojada
50 60 70 80 90 100
Importance

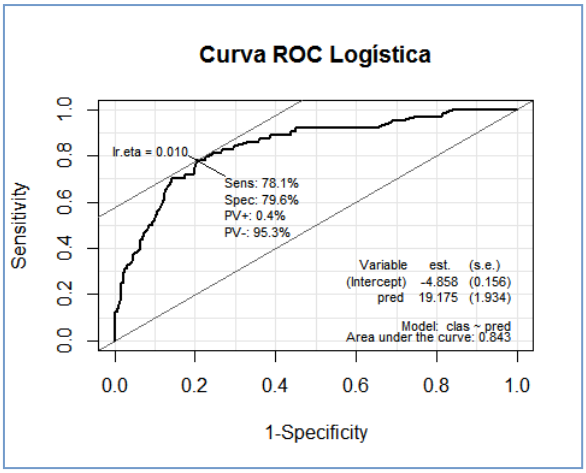

GBM Camiones
tipoaccSalidas izda
accsegSi
tipoaccVuelco
luminNoche
cansanSi
barreraSi
tipoaccSalidas dcha
tipoaccFrontal
redEstatal
zonaVariante
velinaSi
edad34
redAutonómica
tipodiaFestivo
tipoviaSecundarias
medianaSi
superfUmbriaoGrava
tipodiaPosterior a festivo
hitosSi
denscir
0 20 40 60 80 100
Importance

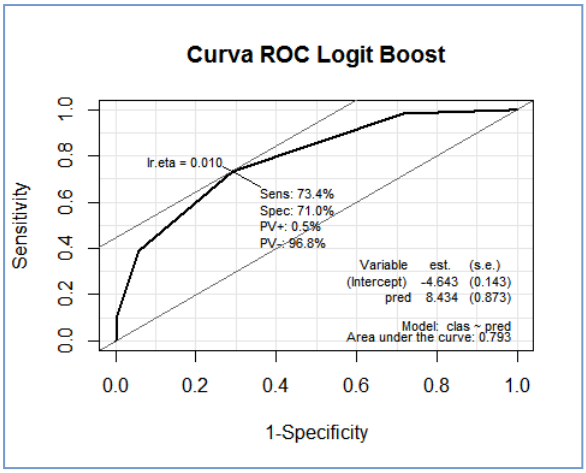

XGB Camiones
denscir
accsegSi
luminNoche
tipoaccSalidas izda
tipodiaAnterior a festivo
barreraSi
tipoaccSalidas dcha
cansanSi
redEstatal
edad34
tipoaccFrontal
edad32
tipoviaSecundarias
velinaSi
tipodiaFestivo
medianaSi
tipoaccOtra colision
infraccSi
redProvincial
distraccSi
20 40 60 80 100
Importance

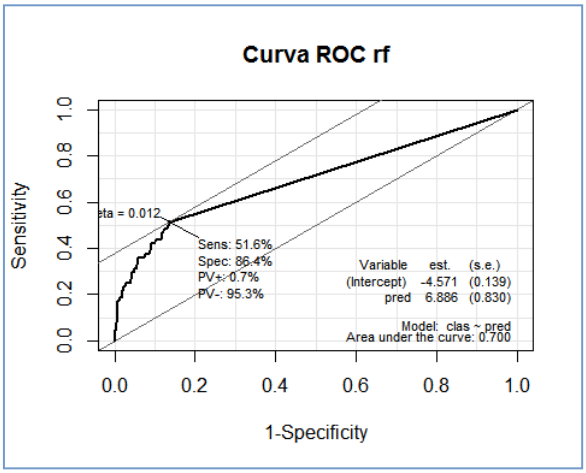

Bayes GLM Camiones
tipoacc
barrera
red
hitos
denscir
cansan
velina
mediana
edad3
distracc
superf
tipovia
sexo
alcohol
facatm
tipodia
accseg
lumin
infracc
zona
0 20 40 60 80 100
Importance

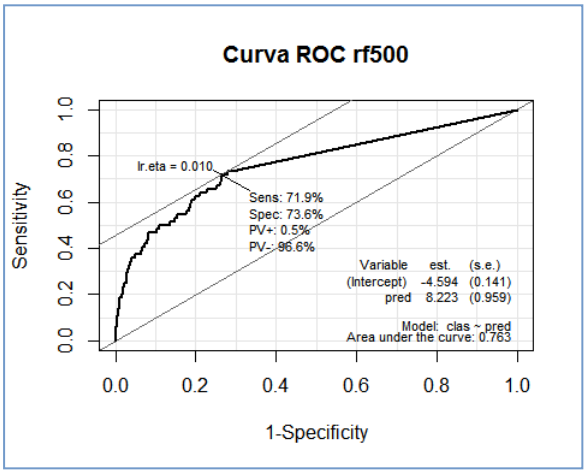

NNet Camiones
accsegSi
tipoaccSalidas izda
barreraSi
tipoaccFrontal
tipoaccVuelco
tipoaccLateral
redEstatal
luminNoche
luminNochellum
zonaZona Urbana
tipodiaPosterior a festivo
infraccSi
edad34
tipoaccSalidas dcha
alcoholSi
facatmNieblas
tipoaccAtropello
tipodiaFestivo
hitosSi
redOtras
40 60 80 100
Importance

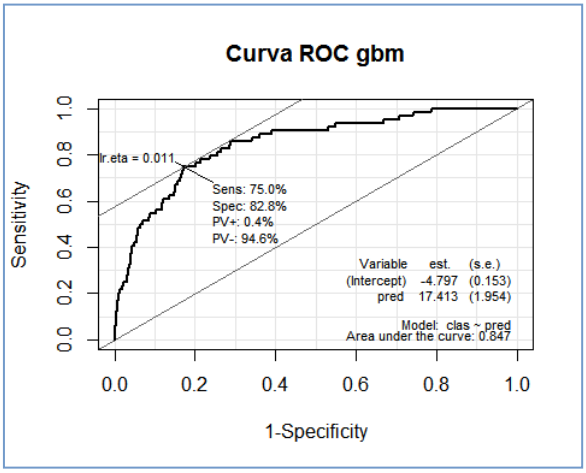

OR factores significativos Camiones
OR
0 5 10 15
tipoaccSalidas izda
redEstatal
tipoaccFrontal
tipoaccVuelco
redAutonómica
tipoaccSalidas dcha
barreraSi
luminNoche
hitosSi
accsegSi
Factores

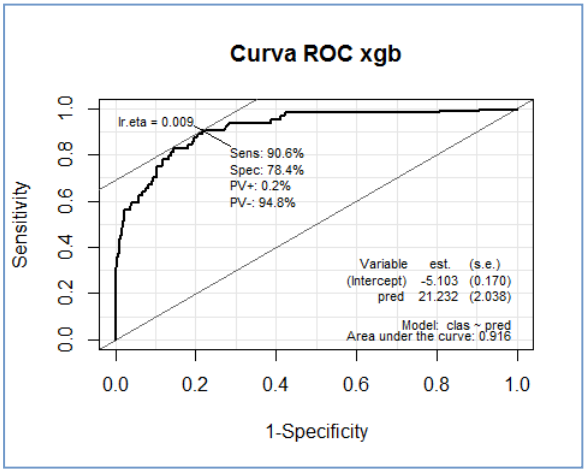

Curva ROC Logística
Sensitivity
1-Specificity
lr.eta = 0.010
Sens: 78.1%
Spec: 79.6%
PV+: 0.4%
PV-: 95.3%
Variable est. (s.e.)
(Intercept) -4.858 (0.156)
pred 19.175 (1.934)
Model: clas ~ pred
Area under the curve: 0.843

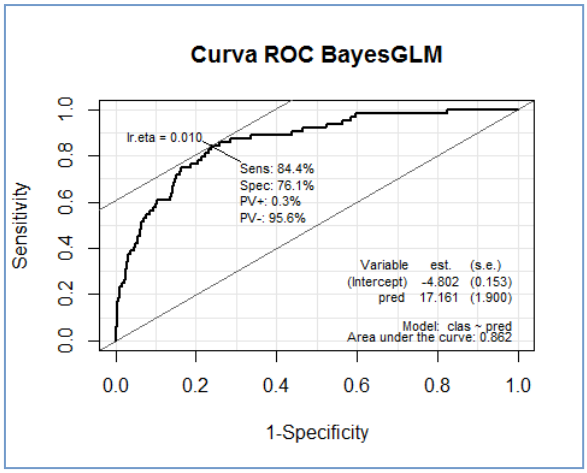

Curva ROC Logit Boost
Sensitivity
1-Specificity
lr.eta = 0.010
Sens: 73.4%
Spec: 71.0%
PV+: 0.5%
PV-: 96.8%
Variable est. (s.e.)
(Intercept) -4.643 (0.143)
pred 8.434 (0.873)
Model: clas ~ pred
Area under the curve: 0.793

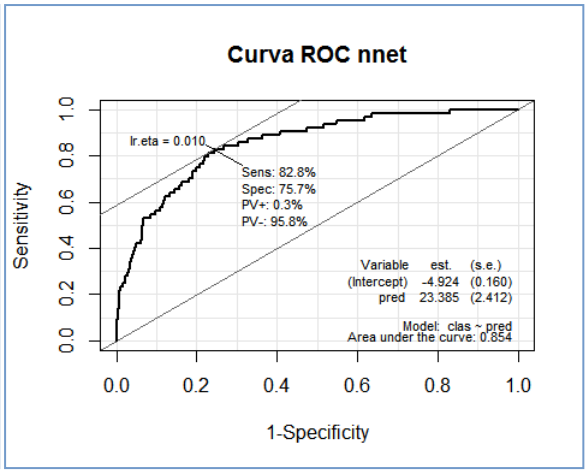

Curva ROC rf
Sensitivity
1-Specificity
Sens: 51.6%
Spec: 86.4%
PV+: 0.7%
PV-: 95.3%
Variable est. (s.e.)
(Intercept) -4.571 (0.139)
pred 6.886 (0.830)
Model: clas ~ pred
Area under the curve: 0.700


Curva ROC rf500
Sensitivity
1-Specificity
lr.eta = 0.010
Sens: 71.9%
Spec: 73.6%
PV+: 0.5%
PV-: 96.6%
Variable est. (s.e.)
(Intercept) -4.594 (0.141)
pred 8.223 (0.959)
Model: clas ~ pred
Area under the curve: 0.763

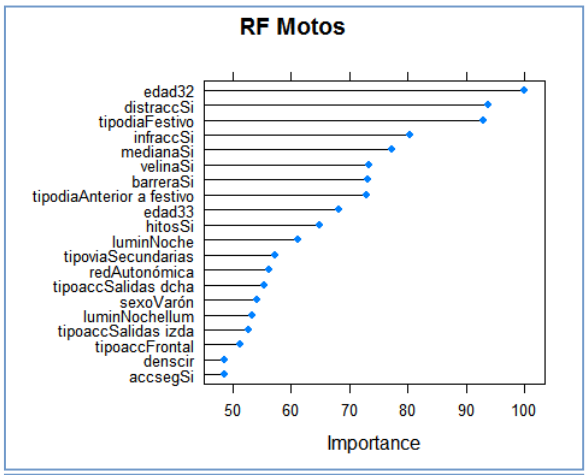

Curva ROC gbm
Sensitivity
1-Specificity
lr.eta = 0.011
Sens: 75.0%
Spec: 82.8%
PV+: 0.4%
PV-: 94.6%
Variable est. (s.e.)
(Intercept) -4.797 (0.153)
pred 17.413 (1.954)
Model: clas ~ pred
Area under the curve: 0.847

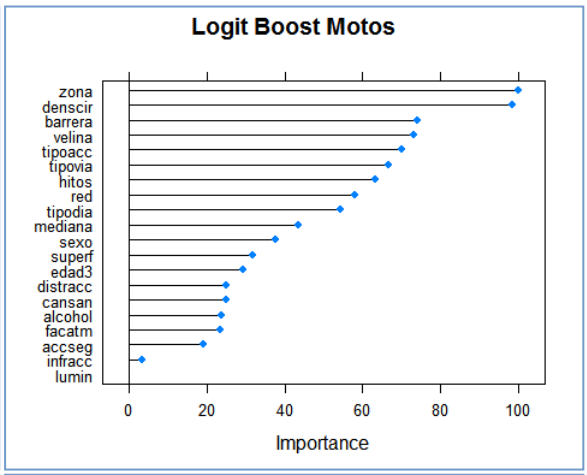

Curva ROC xgb
Sensitivity
1-Specificity
lr.eta = 0.009
Sens: 90.6%
Spec: 78.4%
PV+: 0.2%
PV-: 94.8%
Variable est. (s.e.)
(Intercept) -5.103 (0.170)
pred 21.232 (2.038)
Model: clas ~ pred
Area under the curve: 0.916

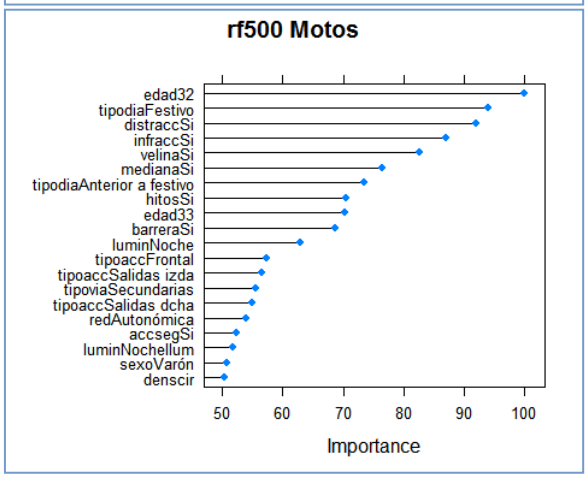

Curva ROC BayesGLM
lr.eta = 0.010
Sens: 84.4%
Spec: 76.1%
PV+: 0.3%
PV-: 95.6%
Variable est. (s.e.)
(Intercept) -4.802 (0.153)
pred 17.161 (1.900)
Model: clas ~ pred
Area under the curve: 0.862
Sensitivity
1-Specificity

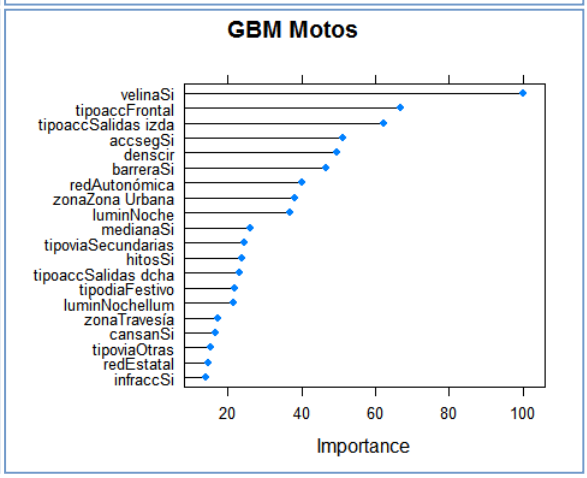

Curva ROC nnet
lr.eta = 0.010
Sens: 82.8%
Spec: 75.7%
PV+: 0.3%
PV-: 95.8%
Variable est. (s.e.)
(Intercept) -4.924 (0.160)
pred 23.385 (2.412)
Model: clas ~ pred
Area under the curve: 0.854
Sensitivity
1-Specificity

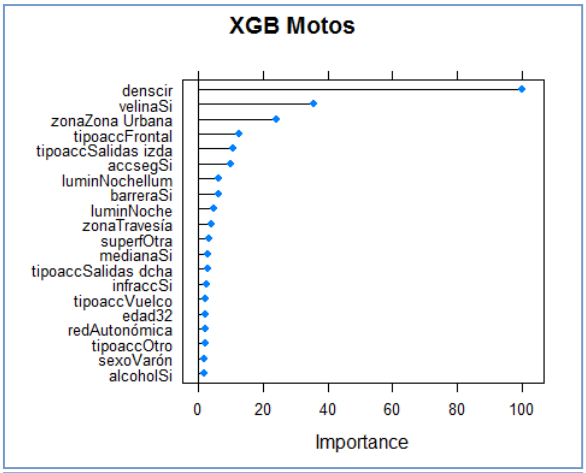
MOTOS

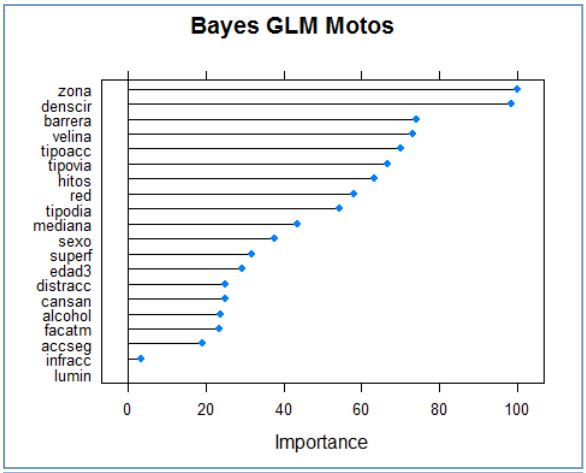

RF Motos
edad32
distraccSi
tipodiaFestivo
infraccSi
medianaSi
velinaSi
barreraSi
tipodiaAnterior a festivo
edad33
hitosSi
luminNoche
tipoviaSecundarias
redAutonómica
tipoaccSalidas dcha
sexoVarón
luminNochellum
tipoaccSalidas izda
tipoaccFrontal
denscir
accsegSi
Importance

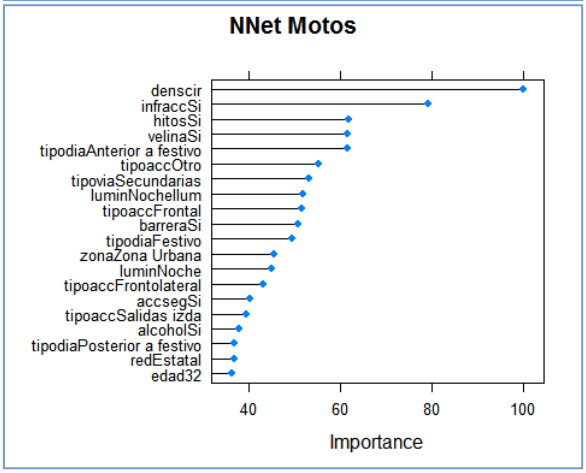

Logit Boost Motos
zona
denscir
barrera
velina
tipoacc
tipovia
hitos
red
tipodia
mediana
sexo
superf
edad3
distracc
cansan
alcohol
facatm
accseg
infracc
lumin
Importance

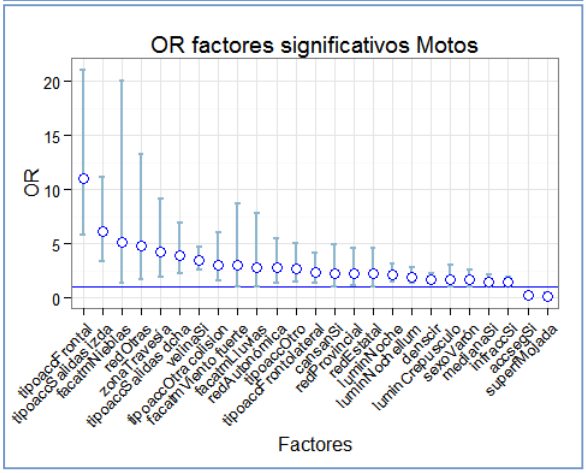

rf500 Motos
edad32
tipodiaFestivo
distraccSi
infraccSi
velinaSi
medianaSi
tipodiaAnterior a festivo
hitosSi
edad33
barreraSi
luminNoche
tipoaccFrontal
tipoaccSalidas izda
tipoviaSecundarias
tipoaccSalidas dcha
redAutonómica
accsegSi
luminNochellum
sexoVarón
denscir
Importance

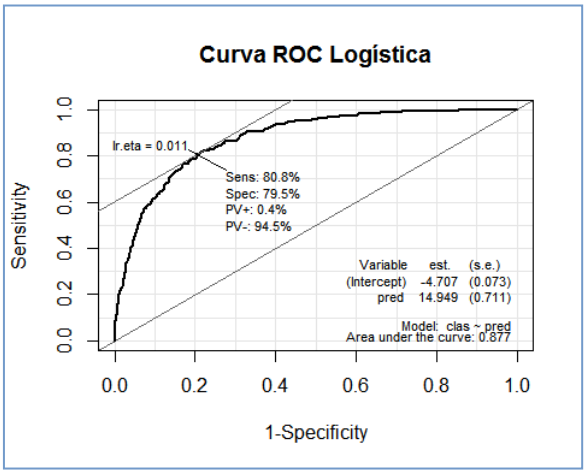

GBM Motos
velinaSi
tipoaccFrontal
tipoaccSalidas izda
accsegSi
denscir
barreraSi
redAutonómica
zonaZona Urbana
luminNoche
medianaSi
tipoviaSecundarias
hitosSi
tipoaccSalidas dcha
tipodiaFestivo
luminNochellum
zonaTravesía
cansanSi
tipoviaOtras
redEstatal
infraccSi
Importance

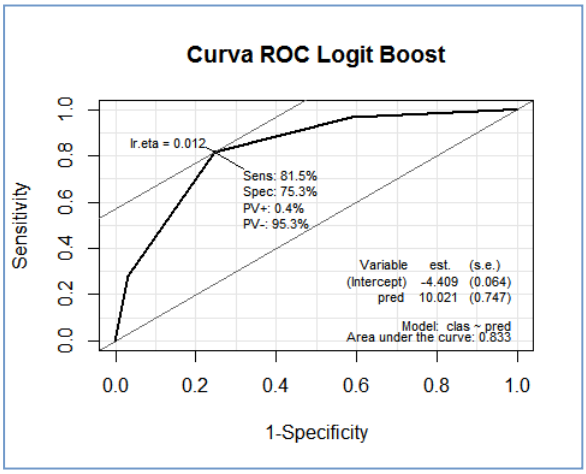

XGB Motos
denscir
velinaSi
zonaZona Urbana
tipoaccFrontal
tipoaccSalidas izda
accsegSi
luminNochellum
barreraSi
luminNoche
zonaTravesía
superfOtra
medianaSi
tipoaccSalidas dcha
infraccSi
tipoaccVuelco
edad32
redAutonómica
tipoaccOtro
sexoVarón
alcoholSi
0 20 40 60 80 100
Importance

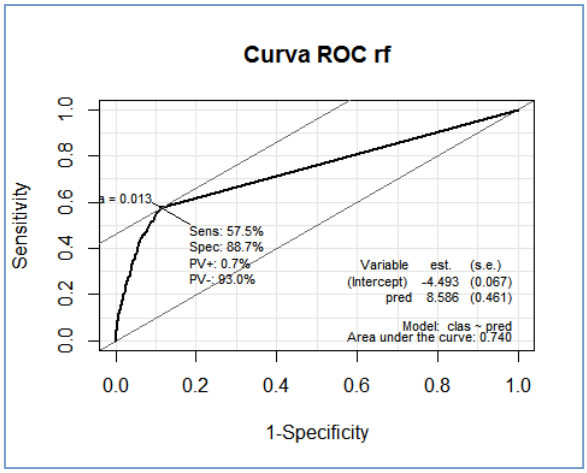

Bayes GLM Motos
zona
denscir
barrera
velina
tipoacc
tipovia
hitos
red
tipodia
mediana
sexo
superf
edad3
distracc
cansan
alcohol
facatm
accseg
infracc
lumin
0 20 40 60 80 100
Importance

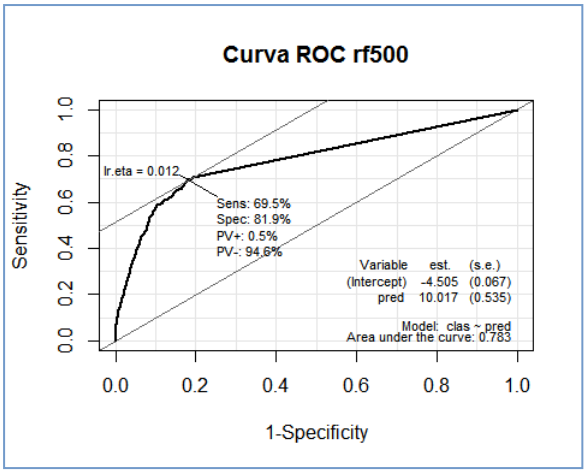

NNet Motos
denscir
infraccSi
hitosSi
velinaSi
tipodiaAnterior a festivo
tipoaccOtro
tipoviaSecundarias
luminNochellum
tipoaccFrontal
barreraSi
tipodiaFestivo
zonaZona Urbana
luminNoche
tipoaccFrontolateral
accsegSi
tipoaccSalidas izda
alcoholSi
tipodiaPosterior a festivo
redEstatal
edad32
40 60 80 100
Importance

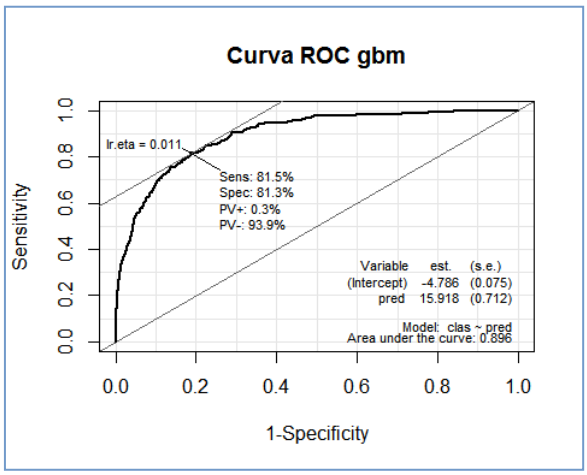

OR factores significativos Motos
OR
20
15
10
5
0
tipoaccFrontal
tipoaccSalidas izda
facatmNieblas
redOtras
zonaTravesía
tipoaccSalidas dcha
tipoaccOtra colision
facatmVuelco fuerte
velinaSi
facatmLluvias
redAutonómica
tipoaccOtro
tipoaccFrontolateral
cansanSi
redProvincial
redEstatal
luminNoche
luminNochellum
denscir
luminCrepusculo
sexoVarón
medianaSi
infraccSi
accsegSi
superfMojada
Factores

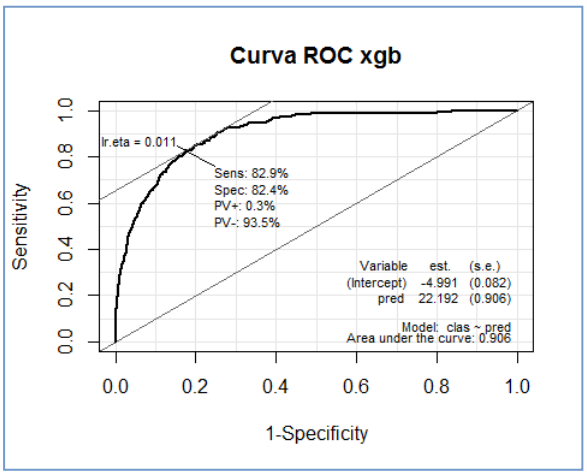

Curva ROC Logística
lr.eta = 0.011
Sens: 80.8%
Spec: 79.5%
PV+: 0.4%
PV-: 94.5%
Variable est. (s.e.)
(Intercept) -4.707 (0.073)
pred 14.949 (0.711)
Model: clas ~ pred
Area under the curve: 0.877
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1-Specificity

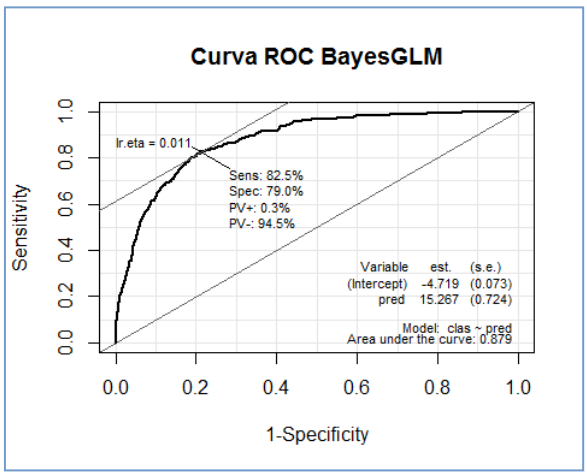

Curva ROC Logit Boost
lr.eta = 0.012
Sens: 81.5%
Spec: 75.3%
PV+: 0.4%
PV-: 95.3%
Variable est. (s.e.)
(Intercept) -4.409 (0.064)
pred 10.021 (0.747)
Model: clas ~ pred
Area under the curve: 0.833
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1-Specificity

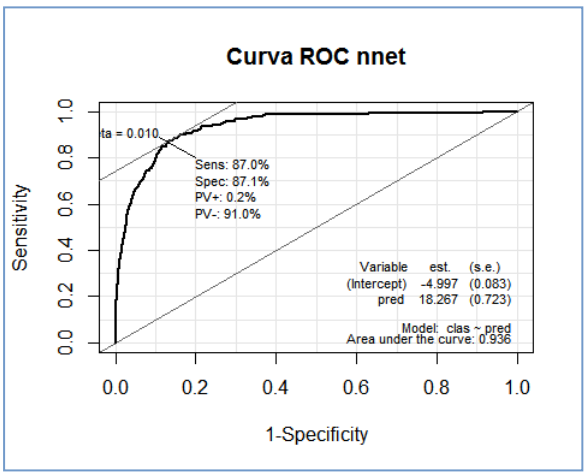

Curva ROC rf
Sensitivity
1-Specificity
= 0.013
Sens: 57.5%
Spec: 88.7%
PV+: 0.7%
PV-: 93.0%
Variable est. (s.e.)
(Intercept) -4.493 (0.067)
pred 8.586 (0.461)
Model: clas ~ pred
Area under the curve: 0.740


Curva ROC rf500
Sensitivity
1-Specificity
lr.eta = 0.012
Sens: 69.5%
Spec: 81.9%
PV+: 0.5%
PV-: 94.6%
Variable est. (s.e.)
(Intercept) -4.505 (0.067)
pred 10.017 (0.535)
Model: clas ~ pred
Area under the curve: 0.783

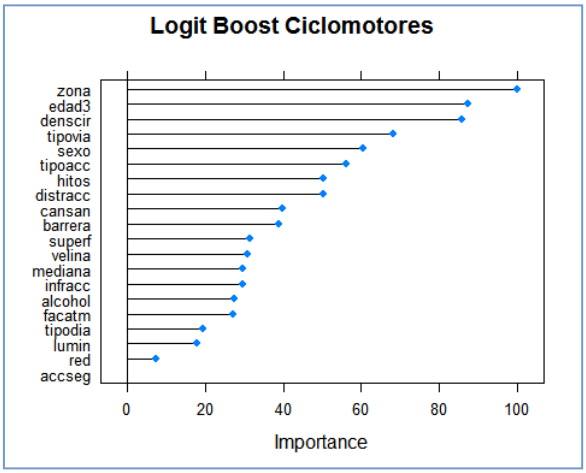

Curva ROC gbm
Sensitivity
1-Specificity
lr.eta = 0.011
Sens: 81.5%
Spec: 81.3%
PV+: 0.3%
PV-: 93.9%
Variable est. (s.e.)
(Intercept) -4.786 (0.075)
pred 15.918 (0.712)
Model: clas ~ pred
Area under the curve: 0.896

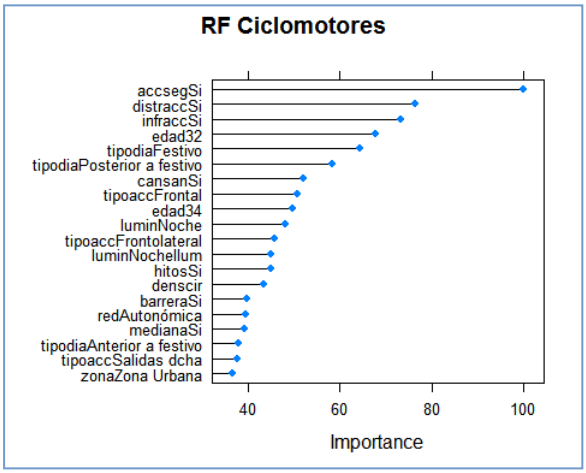

Curva ROC xgb
Sensitivity
1-Specificity
lr.eta = 0.011
Sens: 82.9%
Spec: 82.4%
PV+: 0.3%
PV-: 93.5%
Variable est. (s.e.)
(Intercept) -4.991 (0.082)
pred 22.192 (0.906)
Model: clas ~ pred
Area under the curve: 0.906

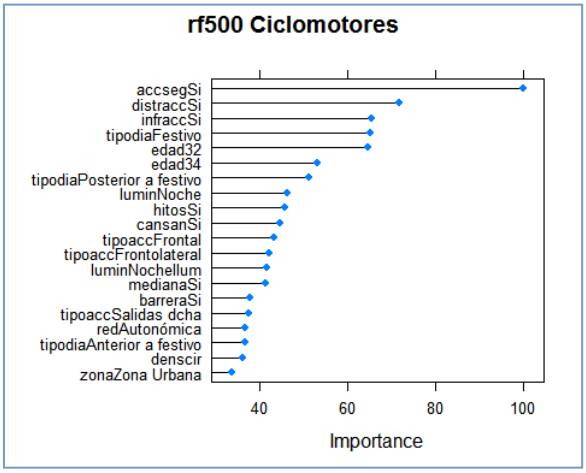

Curva ROC BayesGLM
Sensitivity
1-Specificity
lr.eta = 0.011
Sens: 82.5%
Spec: 79.0%
PV+: 0.3%
PV-: 94.5%
Variable est. (s.e.)
(Intercept) -4.719 (0.073)
pred 15.267 (0.724)
Model: clas ~ pred
Area under the curve: 0.879

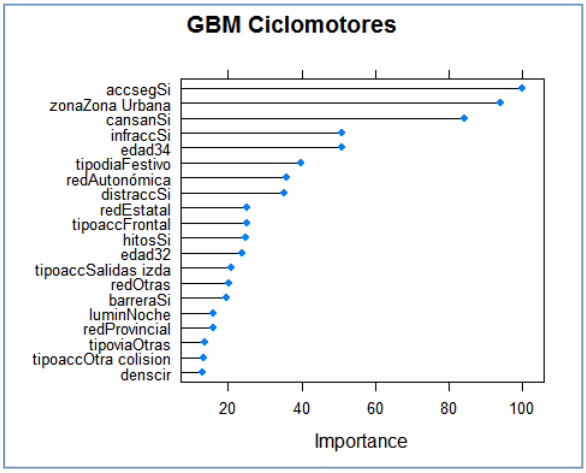

Curva ROC nnet
Sensitivity
1-Specificity
ta = 0.010
Sens: 87.0%
Spec: 87.1%
PV+: 0.2%
PV-: 91.0%
Variable est. (s.e.)
(Intercept) -4.997 (0.083)
pred 18.267 (0.723)
Model: clas ~ pred
Area under the curve: 0.936

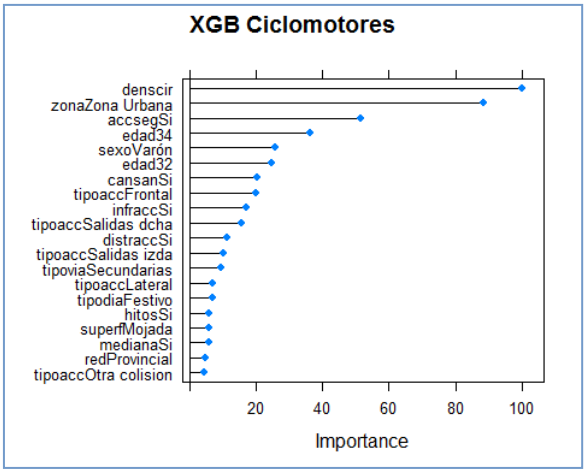
CICLOMOTORES

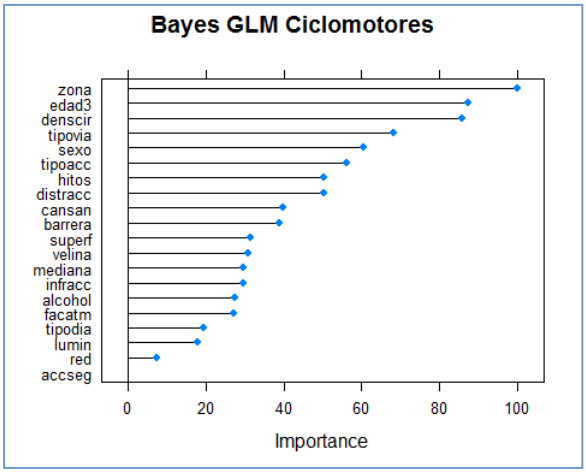


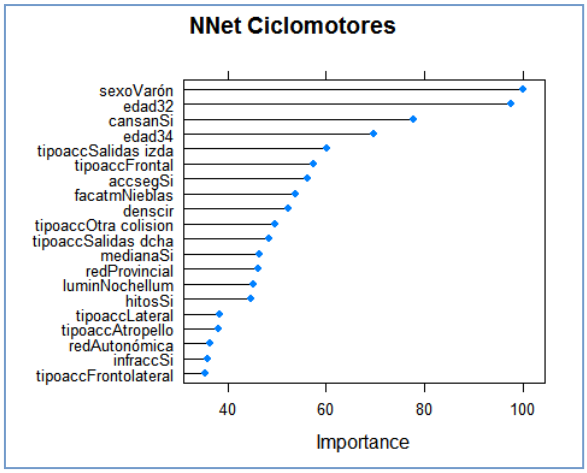


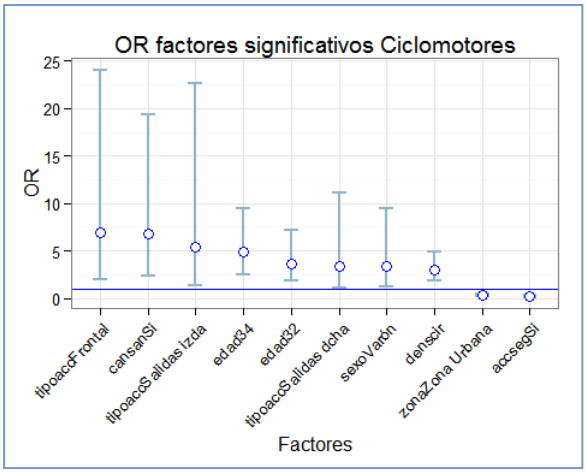


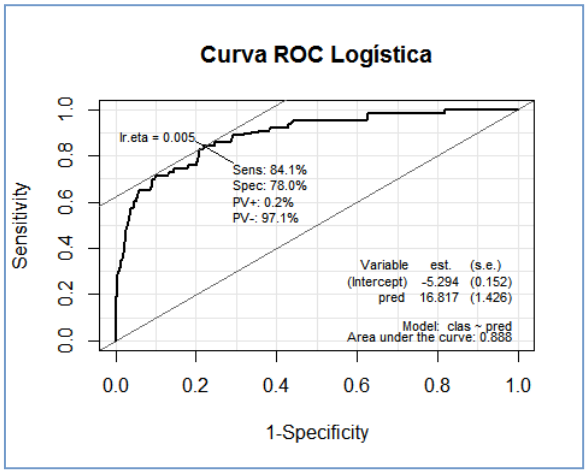


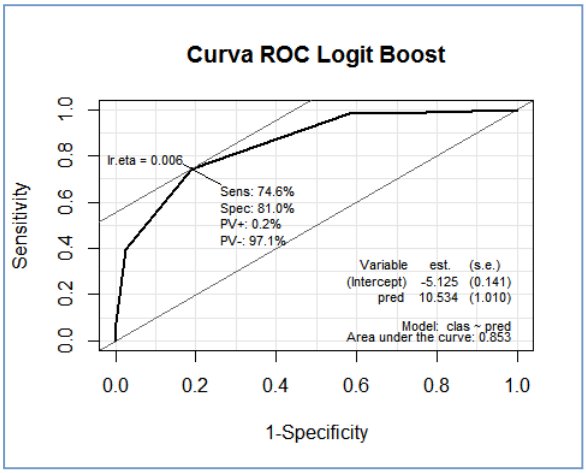


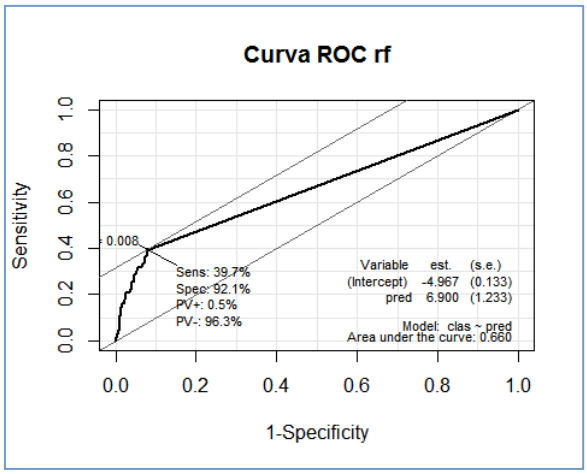


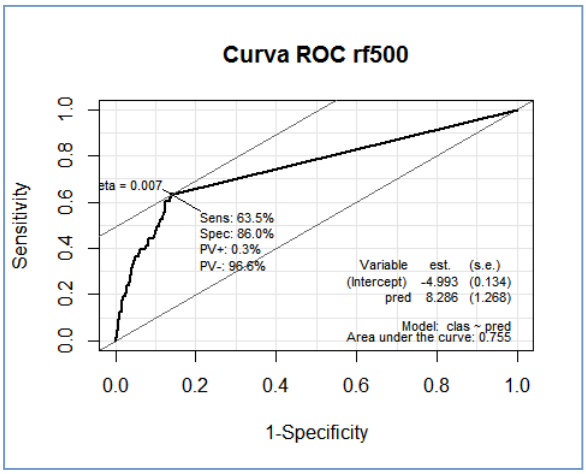

NNet Ciclomotores
sexoVarón
edad32
cansanSi
edad34
tipoaccSalidas izda
tipoaccFrontal
accsegSi
facatmNieblas
denscir
tipoaccOtra colision
tipoaccSalidas dcha
medianaSi
redProvincial
luminNochellum
hitosSi
tipoaccLateral
tipoaccAtropello
redAutonómica
infraccSi
tipoaccFrontolateral
40
60
80
100
Importance

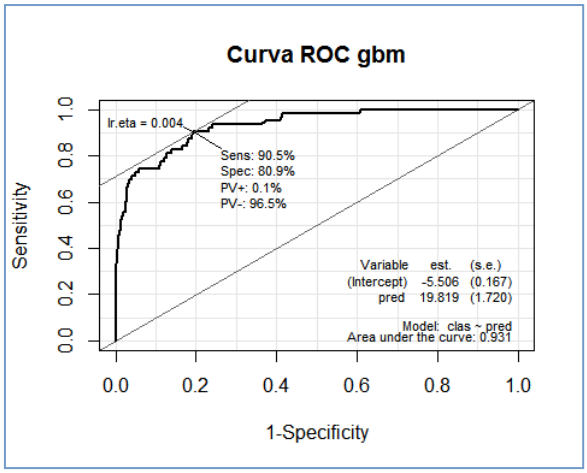

OR factores significativos Ciclomotores
OR
25
20
15
10
5
0
tipoaccFrontal
cansanSi
tipoaccSalidas izda
edad34
edad32
tipoaccSalidas dcha
sexoVarón
denscir
zonaZona Urbana
accsegSi
Factores

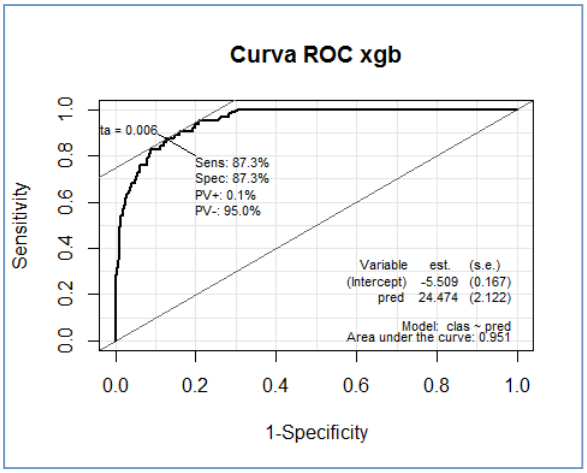

Curva ROC Logística
lr.eta = 0.005
Sens: 84.1%
Spec: 78.0%
PV+: 0.2%
PV-: 97.1%
Variable est. (s.e.)
(Intercept) -5.294 (0.152)
pred 16.817 (1.426)
Model: clas ~ pred
Area under the curve: 0.888
Sensitivity
1-Specificity

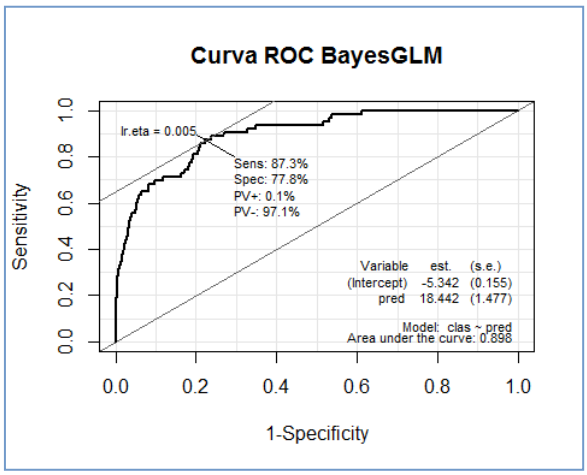

Curva ROC Logit Boost
lr.eta = 0.006
Sens: 74.6%
Spec: 81.0%
PV+: 0.2%
PV-: 97.1%
Variable est. (s.e.)
(Intercept) -5.125 (0.141)
pred 10.534 (1.010)
Model: clas ~ pred
Area under the curve: 0.853
Sensitivity
1-Specificity

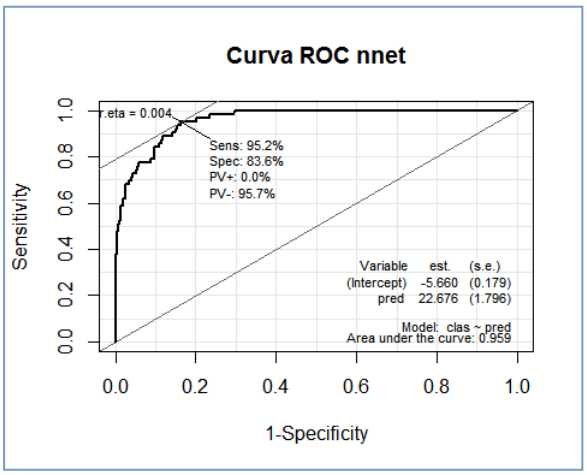

Curva ROC rf
0.008
Sens: 39.7%
Spec: 92.1%
PV+: 0.5%
PV-: 96.3%
Variable est. (s.e.)
(Intercept) -4.967 (0.133)
pred 6.900 (1.233)
Model: clas ~ pred
Area under the curve: 0.660
Sensitivity
1-Specificity


Curva ROC rf500
eta = 0.007
Sens: 63.5%
Spec: 86.0%
PV+: 0.3%
PV-: 96.6%
Variable est. (s.e.)
(Intercept) -4.993 (0.134)
pred 8.286 (1.268)
Model: clas ~ pred
Area under the curve: 0.755
Sensitivity
1-Specificity

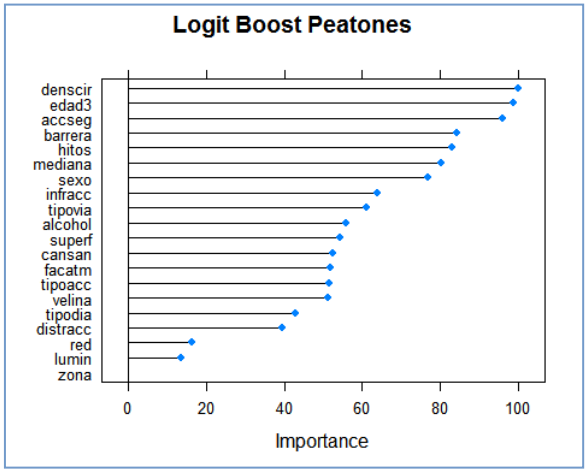

Curva ROC gbm
lr.eta = 0.004
Sens: 90.5%
Spec: 80.9%
PV+: 0.1%
PV-: 96.5%
Variable est. (s.e.)
(Intercept) -5.506 (0.167)
pred 19.819 (1.720)
Model: clas ~ pred
Area under the curve: 0.931
Sensitivity
1-Specificity

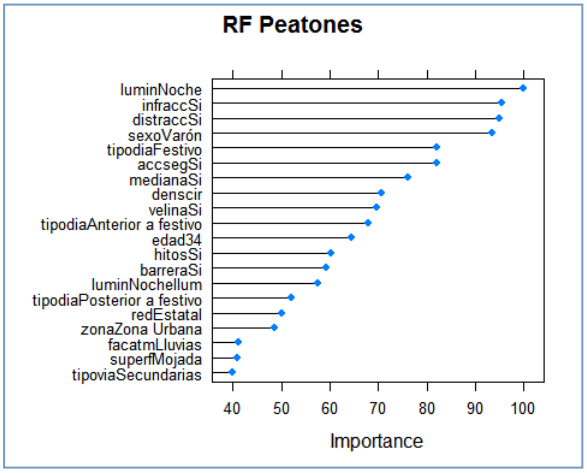

Curva ROC xgb
ta = 0.006
Sens: 87.3%
Spec: 87.3%
PV+: 0.1%
PV-: 95.0%
Variable est. (s.e.)
(Intercept) -5.509 (0.167)
pred 24.474 (2.122)
Model: clas ~ pred
Area under the curve: 0.951
Sensitivity
1-Specificity

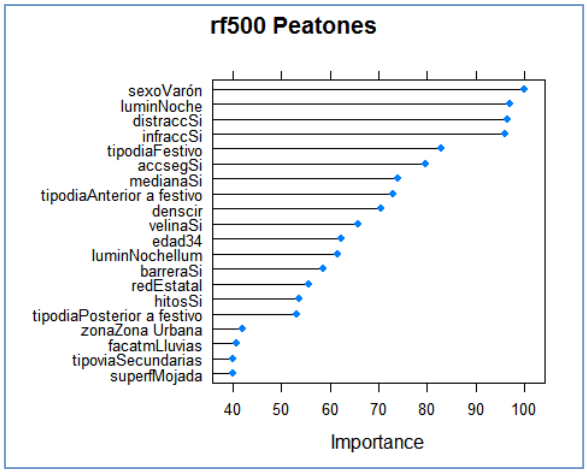

Curva ROC BayesGLM
lr.eta = 0.005
Sens: 87.3%
Spec: 77.8%
PV+: 0.1%
PV-: 97.1%
Variable est. (s.e.)
(Intercept) -5.342 (0.155)
pred 18.442 (1.477)
Model: clas ~ pred
Area under the curve: 0.898
Sensitivity
1-Specificity

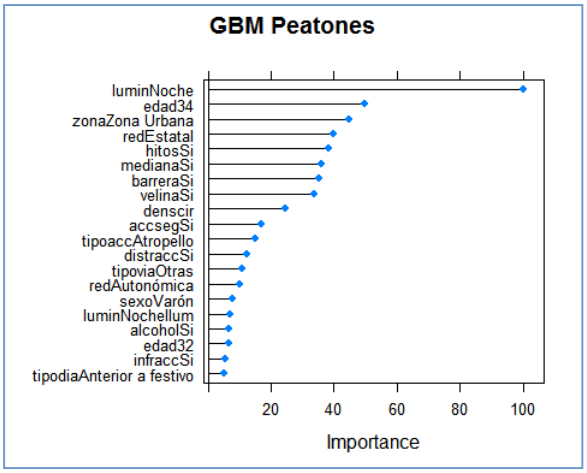

Curva ROC nnet
r.eta = 0.004
Sens: 95.2%
Spec: 83.6%
PV+: 0.0%
PV-: 95.7%
Variable est. (s.e.)
(Intercept) -5.660 (0.179)
pred 22.676 (1.796)
Model: clas ~ pred
Area under the curve: 0.959
Sensitivity
1-Specificity

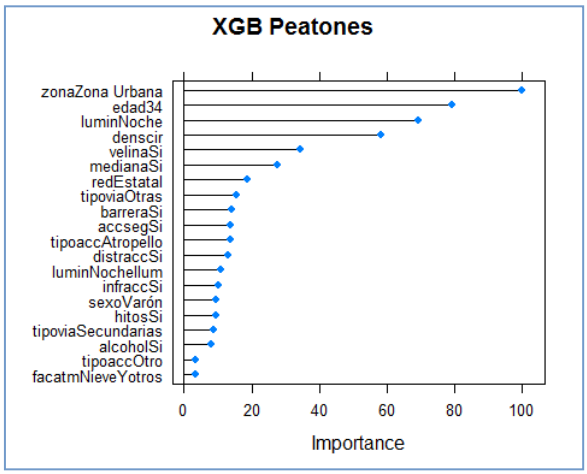
PEATONES

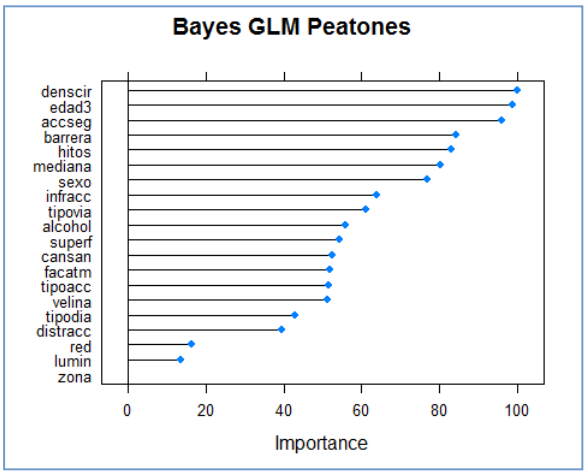

Logit Boost Peatones
denscir
edad3
accseg
barrera
hitos
mediana
sexo
infracc
tipovia
alcohol
superf
cansan
facatm
tipoacc
velina
tipodia
distracc
red
lumin
zona
Importance

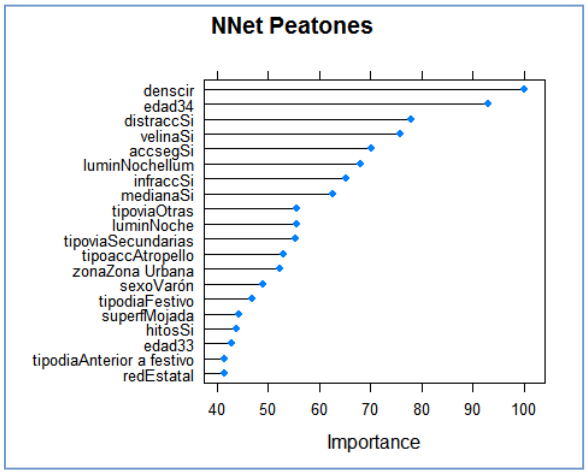

RF Peatones
luminNoche
infraccSi
distraccSi
sexoVarón
tipodiaFestivo
accsegSi
medianaSi
denscir
velinaSi
tipodiaAnterior a festivo
edad34
hitosSi
barreraSi
luminNochellum
tipodiaPosterior a festivo
redEstatal
zonaZona Urbana
facatmLluvias
superfMojada
tipoviaSecundarias
Importance

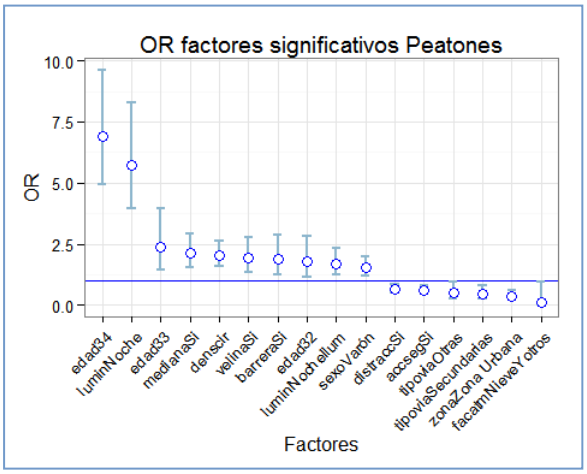

rf500 Peatones
sexoVarón
luminNoche
distraccSi
infraccSi
tipodiaFestivo
accsegSi
medianaSi
tipodiaAnterior a festivo
denscir
velinaSi
edad34
luminNochellum
barreraSi
redEstatal
hitosSi
tipodiaPosterior a festivo
zonaZona Urbana
facatmLluvias
tipoviaSecundarias
superfMojada
40 50 60 70 80 90 100
Importance

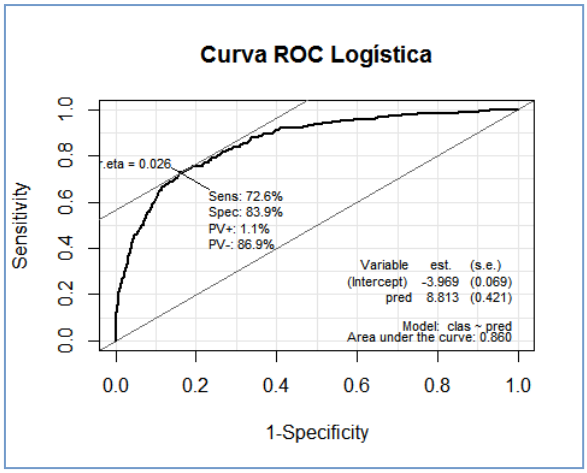

GBM Peatones
luminNoche
edad34
zonaZona Urbana
redEstatal
hitosSi
medianaSi
barreraSi
velinaSi
denscir
accsegSi
tipoaccAtropello
distraccSi
tipoviaOtras
redAutonómica
sexoVarón
luminNochellum
alcoholSi
edad32
infraccSi
tipodiaAnterior a festivo
20 40 60 80 100
Importance

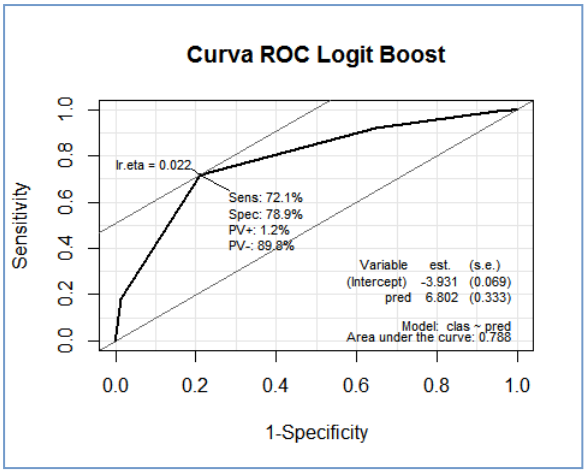

XGB Peatones
zonaZona Urbana
edad34
luminNoche
denscir
velinaSi
medianaSi
redEstatal
tipoviaOtras
barreraSi
accsegSi
tipoaccAtropello
distraccSi
luminNochellum
infraccSi
sexoVarón
hitosSi
tipoviaSecundarias
alcoholSi
tipoaccOtro
facatmNieveYotros
0 20 40 60 80 100
Importance

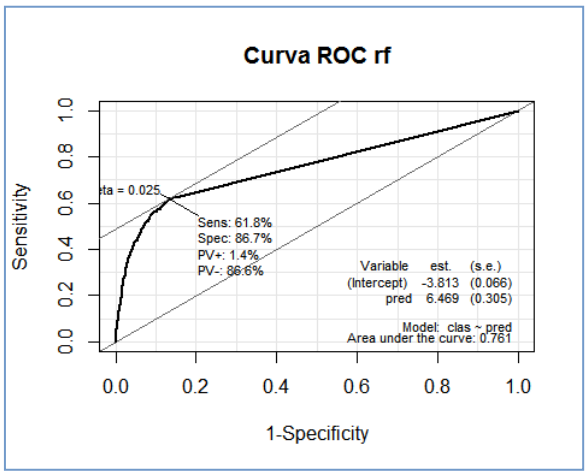

Bayes GLM Peatones
denscir
edad3
accseg
barrera
hitos
mediana
sexo
infracc
tipovia
alcohol
superf
cansan
facatm
tipoacc
velina
tipodia
distracc
red
lumin
zona
0 20 40 60 80 100
Importance

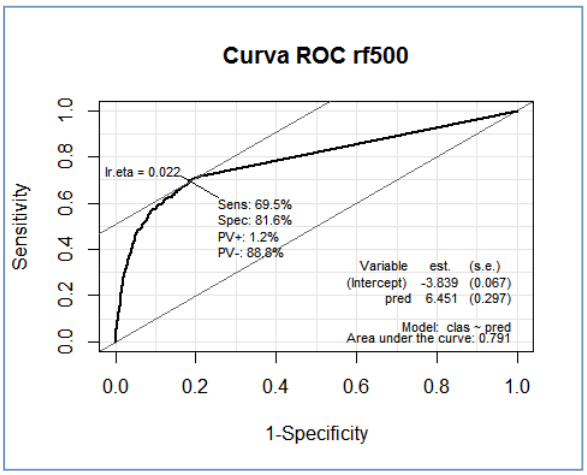

NNet Peatones
denscir
edad34
distraccSi
velinaSi
accsegSi
luminNochellum
infraccSi
medianaSi
tipoviaOtras
luminNoche
tipoviaSecundarias
tipoaccAtropello
zonaZona Urbana
sexoVarón
tipodiaFestivo
superfMojada
hitosSi
edad33
tipodiaAnterior a festivo
redEstatal
40 50 60 70 80 90 100
Importance

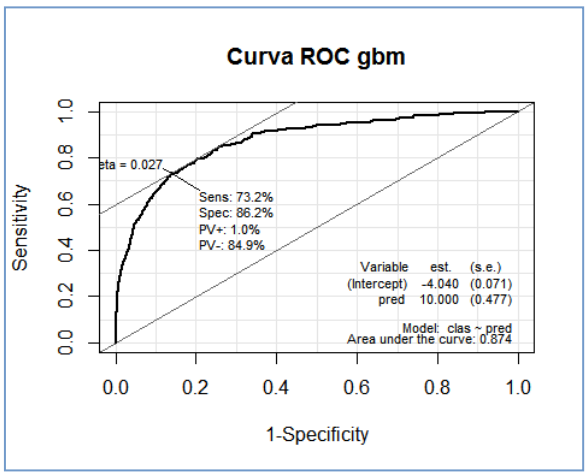

OR factores significativos Peatones
OR
10.0
7.5
5.0
2.5
0.0
edad34
luminNoche
edad33
medianaSi
denscir
velinaSi
barreraSi
edad32
luminNochellum
sexoVarón
distraccSi
accsegSi
tipoviaOtras
tipoviaSecundarias
zonaZona Urbana
facatmNieveYotros
Factores

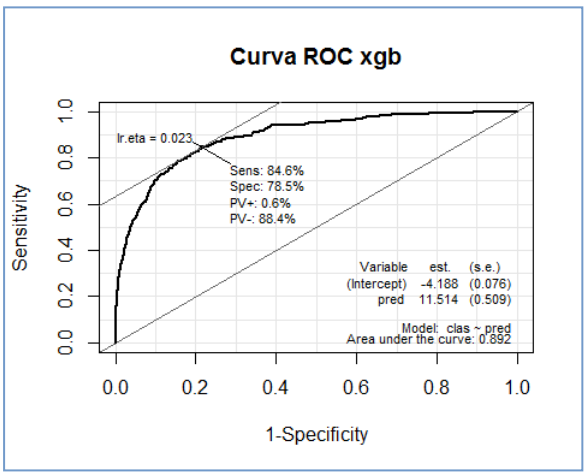

Curva ROC Logística
.eta = 0.026
Sens: 72.6%
Spec: 83.9%
PV+: 1.1%
PV-: 86.9%
Variable est. (s.e.)
(Intercept) -3.969 (0.069)
pred 8.813 (0.421)
Model: clas ~ pred
Area under the curve: 0.860
Sensitivity
1-Specificity

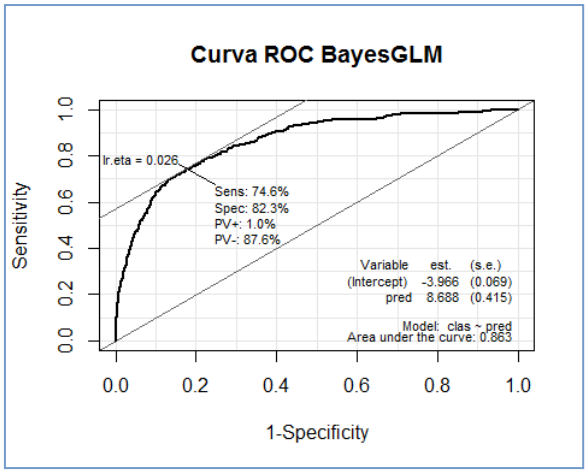

Curva ROC Logit Boost
lr.eta = 0.022
Sens: 72.1%
Spec: 78.9%
PV+: 1.2%
PV-: 89.8%
Variable est. (s.e.)
(Intercept) -3.931 (0.069)
pred 6.802 (0.333)
Model: clas ~ pred
Area under the curve: 0.788
Sensitivity
1-Specificity

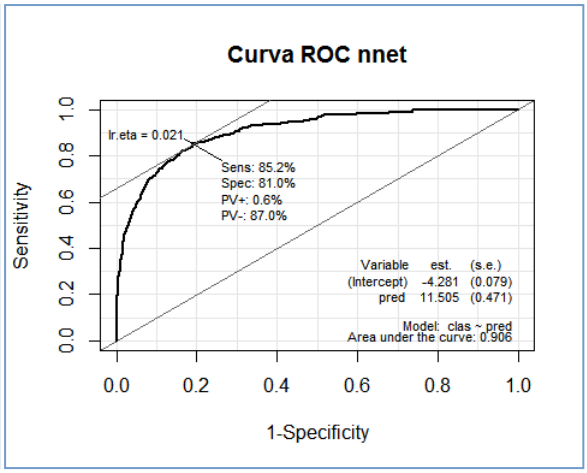

Curva ROC rf
ta = 0.025
Sens: 61.8%
Spec: 86.7%
PV+: 1.4%
PV-: 86.6%
Variable est. (s.e.)
(Intercept) -3.813 (0.066)
pred 6.469 (0.305)
Model: clas ~ pred
Area under the curve: 0.761
Sensitivity
1-Specificity


Curva ROC rf500
lr.eta = 0.022
Sens: 69.5%
Spec: 81.6%
PV+: 1.2%
PV-: 88.8%
Variable est. (s.e.)
(Intercept) -3.839 (0.067)
pred 6.451 (0.297)
Model: clas ~ pred
Area under the curve: 0.791
Sensitivity
1-Specificity

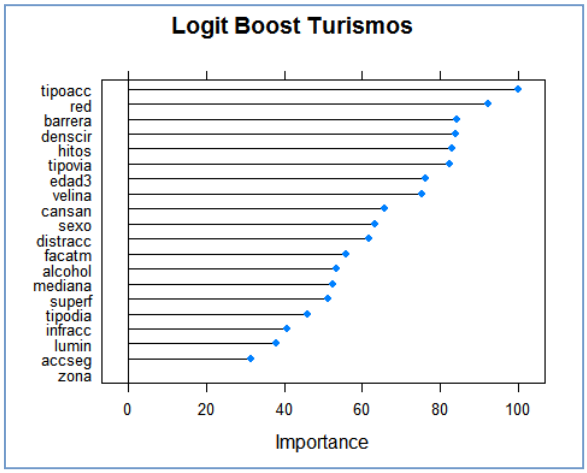

Curva ROC gbm
eta = 0.027
Sens: 73.2%
Spec: 86.2%
PV+: 1.0%
PV-: 84.9%
Variable est. (s.e.)
(Intercept) -4.040 (0.071)
pred 10.000 (0.477)
Model: clas ~ pred
Area under the curve: 0.874
Sensitivity
1-Specificity

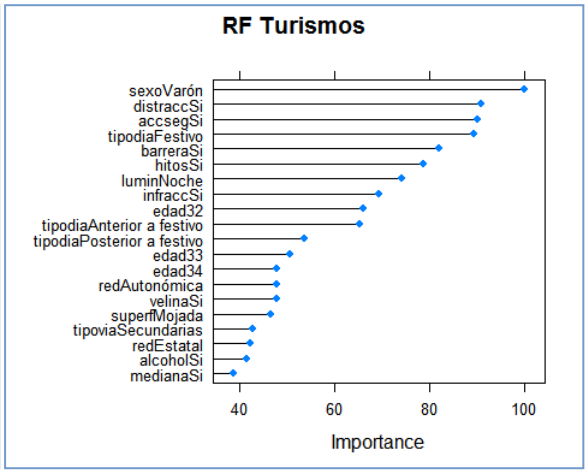

Curva ROC xgb
lr.eta = 0.023
Sens: 84.6%
Spec: 78.5%
PV+: 0.6%
PV-: 88.4%
Variable est. (s.e.)
(Intercept) -4.188 (0.076)
pred 11.514 (0.509)
Model: clas ~ pred
Area under the curve: 0.892
Sensitivity
1-Specificity

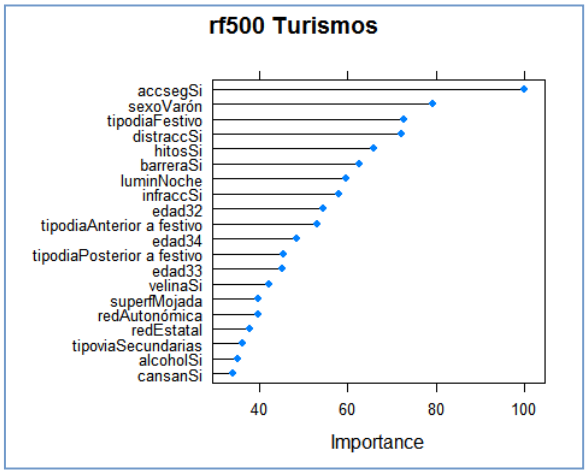

Curva ROC BayesGLM
Sensitivity
1-Specificity
lr.eta = 0.026
Sens: 74.6%
Spec: 82.3%
PV+: 1.0%
PV-: 87.6%
Variable est. (s.e.)
(Intercept) -3.966 (0.069)
pred 8.688 (0.415)
Model: clas ~ pred
Area under the curve: 0.863

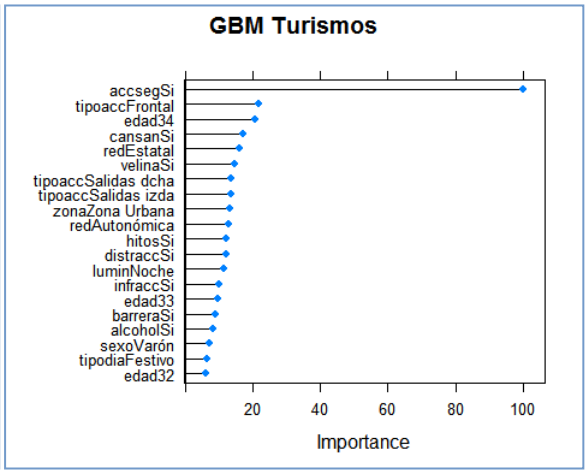

Curva ROC nnet
Sensitivity
1-Specificity
lr.eta = 0.021
Sens: 85.2%
Spec: 81.0%
PV+: 0.6%
PV-: 87.0%
Variable est. (s.e.)
(Intercept) -4.281 (0.079)
pred 11.505 (0.471)
Model: clas ~ pred
Area under the curve: 0.906

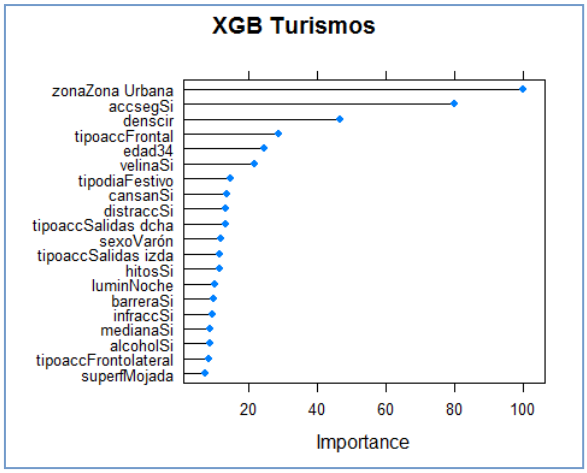
TURISMOS

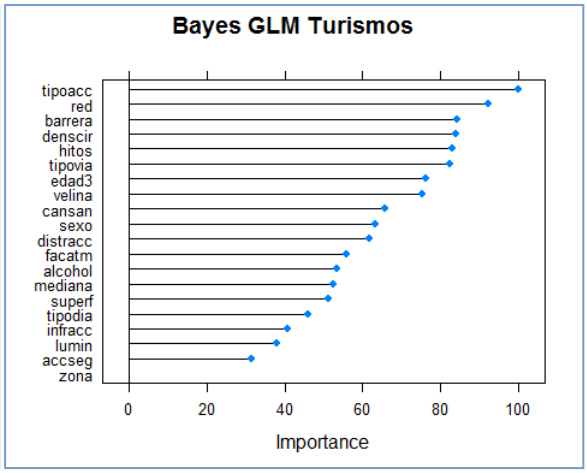

Logit Boost Turismos
tipoacc
red
barrera
denscir
hitos
tipovia
edad3
velina
cansan
sexo
distracc
facatm
alcohol
mediana
superf
tipodia
infracc
lumin
accseg
zona
Importance

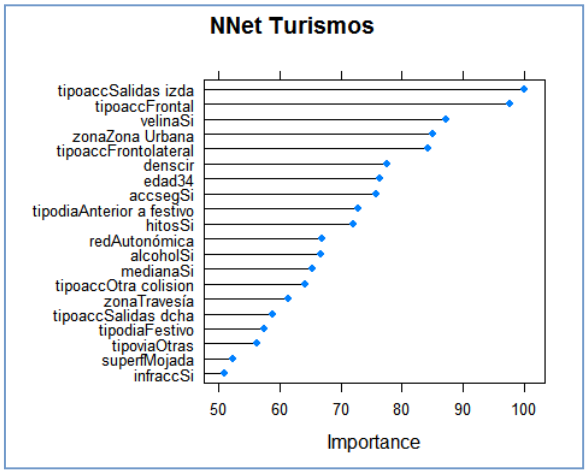

RF Turismos
sexoVarón
distraccSi
accsegSi
tipodiaFestivo
barreraSi
hitosSi
luminNoche
infraccSi
edad32
tipodiaAnterior a festivo
tipodiaPosterior a festivo
edad33
edad34
redAutonómica
velinaSi
superfMojada
tipoviaSecundarias
redEstatal
alcoholSi
medianaSi
Importance

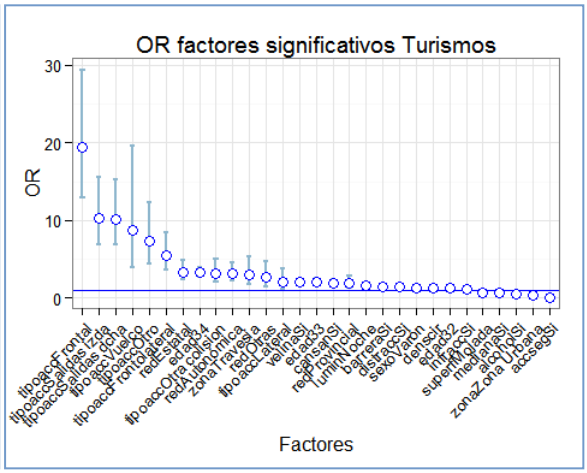

rf500 Turismos
accsegSi
sexoVarón
tipodiaFestivo
distraccSi
hitosSi
barreraSi
luminNoche
infraccSi
edad32
tipodiaAnterior a festivo
edad34
tipodiaPosterior a festivo
edad33
velinaSi
superfMojada
redAutonómica
redEstatal
tipoviaSecundarias
alcoholSi
cansanSi
Importance

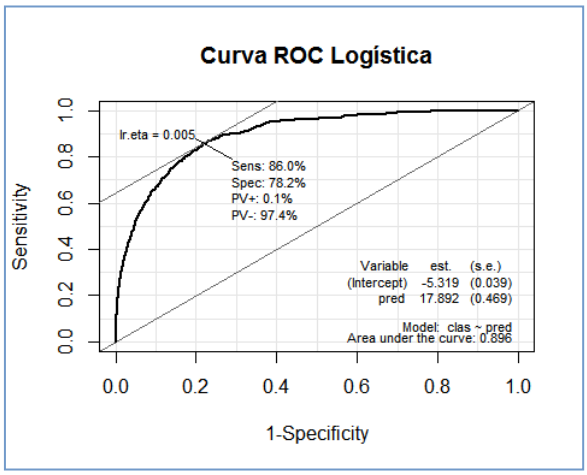

GBM Turismos
accsegSi
tipoaccFrontal
edad34
cansanSi
redEstatal
velinaSi
tipoaccSalidas dcha
tipoaccSalidas izda
zonaZona Urbana
redAutonómica
hitosSi
distraccSi
luminNoche
infraccSi
edad33
barreraSi
alcoholSi
sexoVarón
tipodiaFestivo
edad32
Importance

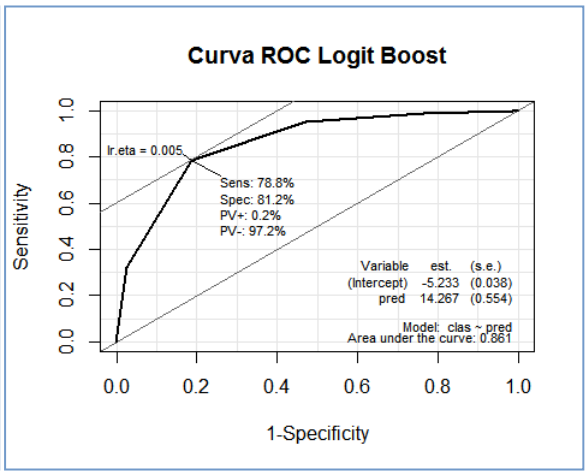

XGB Turismos
zonaZona Urbana
accsegSi
denscir
tipoaccFrontal
edad34
velinaSi
tipodiaFestivo
cansanSi
distraccSi
tipoaccSalidas dcha
sexoVarón
tipoaccSalidas izda
hitosSi
luminNoche
barreraSi
infraccSi
medianaSi
alcoholSi
tipoaccFrontolateral
superfMojada
20
40
60
80
100
Importance

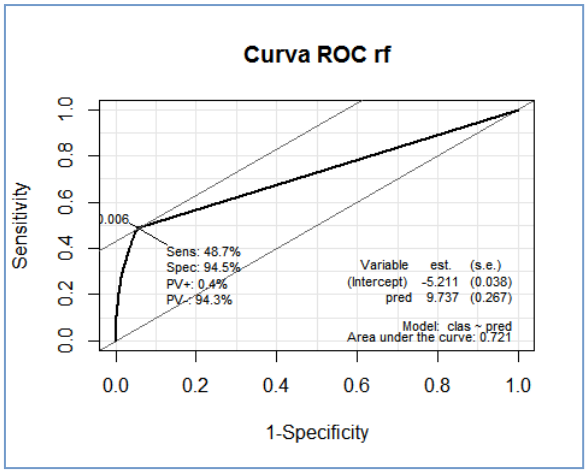

Bayes GLM Turismos
tipoacc
red
barrera
denscir
hitos
tipovia
edad3
velina
cansan
sexo
distracc
facatm
alcohol
mediana
superf
tipodia
infracc
lumin
accseg
zona
0
20
40
60
80
100
Importance

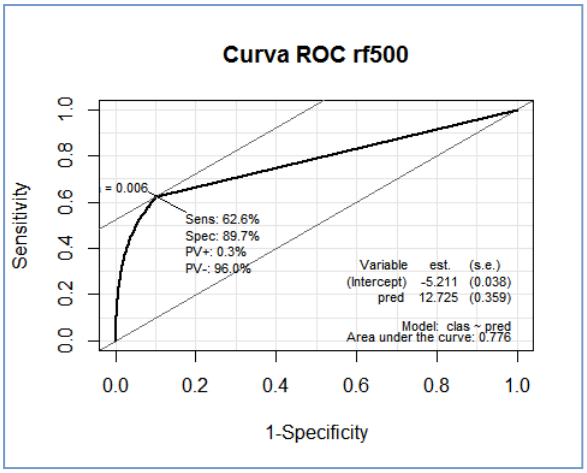

NNet Turismos
tipoaccSalidas izda
tipoaccFrontal
velinaSi
zonaZona Urbana
tipoaccFrontolateral
denscir
edad34
accsegSi
tipodiaAnterior a festivo
hitosSi
redAutonómica
alcoholSi
medianaSi
tipoaccOtra colision
zonaTravesía
tipoaccSalidas dcha
tipodiaFestivo
tipoviaOtras
superfMojada
infraccSi
50
60
70
80
90
100
Importance

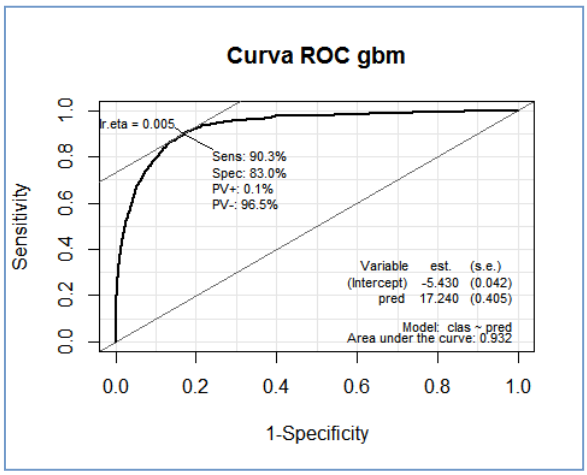

OR factores significativos Turismos
OR
30
20
10
0
Factores

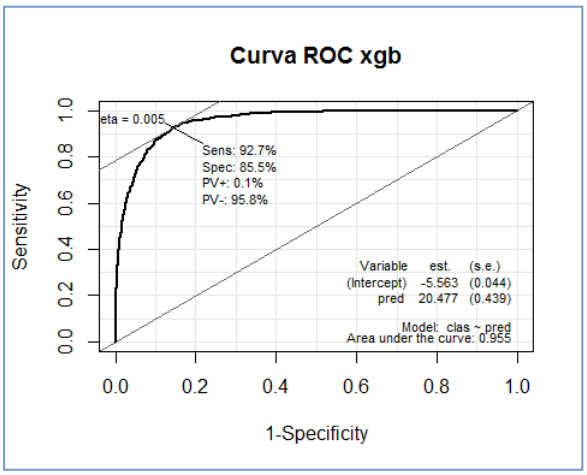

Curva ROC Logística
lr.eta = 0.005
Sens: 86.0%
Spec: 78.2%
PV+: 0.1%
PV-: 97.4%
Variable est. (s.e.)
(Intercept) -5.319 (0.039)
pred 17.892 (0.469)
Model: clas ~ pred
Area under the curve: 0.896
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0

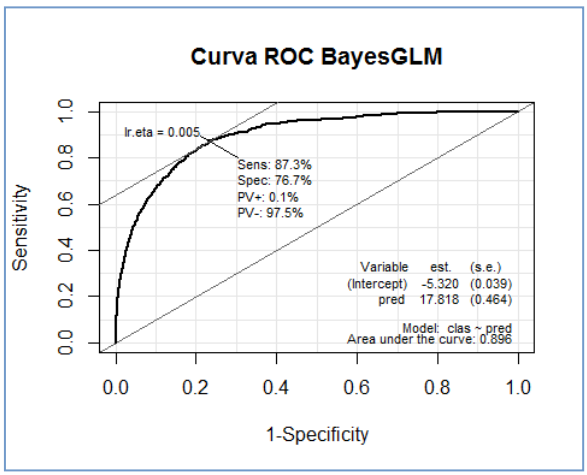

Curva ROC Logit Boost
lr.eta = 0.005
Sens: 78.8%
Spec: 81.2%
PV+: 0.2%
PV-: 97.2%
Variable est. (s.e.)
(Intercept) -5.233 (0.038)
pred 14.267 (0.554)
Model: clas ~ pred
Area under the curve: 0.861
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0

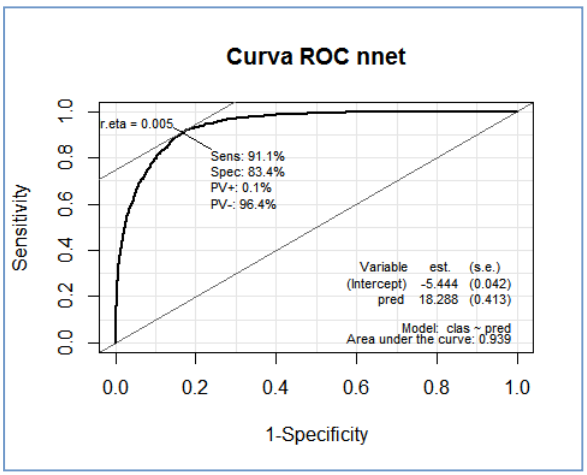

Curva ROC rf
Sensitivity
1-Specificity
0.006
Sens: 48.7%
Spec: 94.5%
PV+: 0.4%
PV-: 94.3%
Variable est. (s.e.)
(Intercept) -5.211 (0.038)
pred 9.737 (0.267)
Model: clas ~ pred
Area under the curve: 0.721


Curva ROC rf500
Sensitivity
1-Specificity
= 0.006
Sens: 62.6%
Spec: 89.7%
PV+: 0.3%
PV-: 96.0%
Variable est. (s.e.)
(Intercept) -5.211 (0.038)
pred 12.725 (0.359)
Model: clas ~ pred
Area under the curve: 0.776

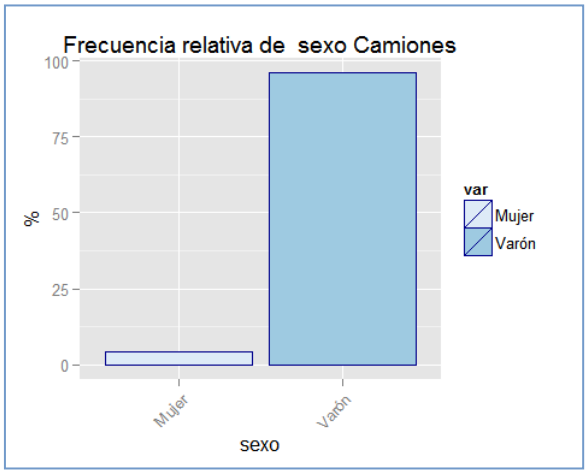

Curva ROC gbm
Sensitivity
1-Specificity
lr.eta = 0.005
Sens: 90.3%
Spec: 83.0%
PV+: 0.1%
PV-: 96.5%
Variable est. (s.e.)
(Intercept) -5.430 (0.042)
pred 17.240 (0.405)
Model: clas ~ pred
Area under the curve: 0.932

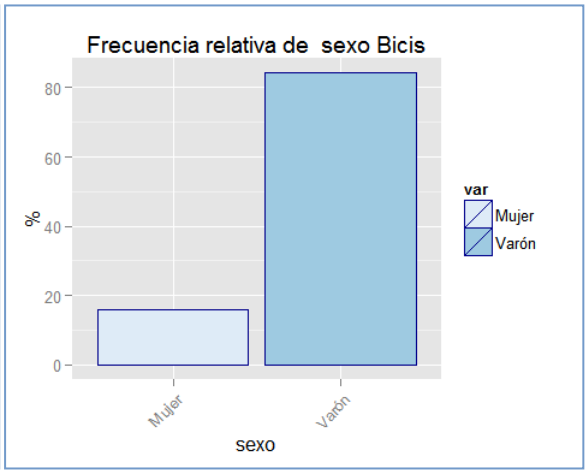

Curva ROC xgb
Sensitivity
1-Specificity
eta = 0.005
Sens: 92.7%
Spec: 85.5%
PV+: 0.1%
PV-: 95.8%
Variable est. (s.e.)
(Intercept) -5.563 (0.044)
pred 20.477 (0.439)
Model: clas ~ pred
Area under the curve: 0.955

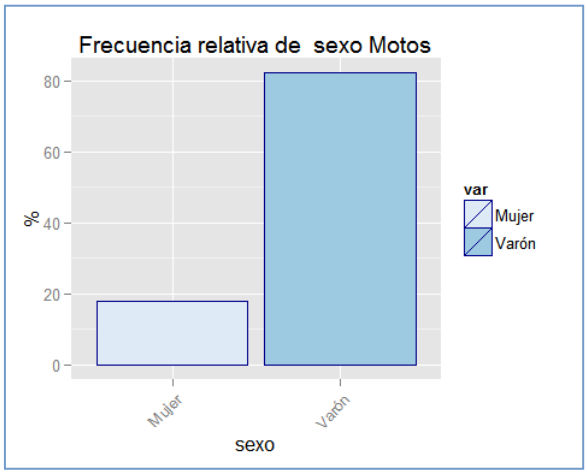

Curva ROC BayesGLM
Sensitivity
1-Specificity
lr.eta = 0.005
Sens: 87.3%
Spec: 76.7%
PV+: 0.1%
PV-: 97.5%
Variable est. (s.e.)
(Intercept) -5.320 (0.039)
pred 17.818 (0.464)
Model: clas ~ pred
Area under the curve: 0.896

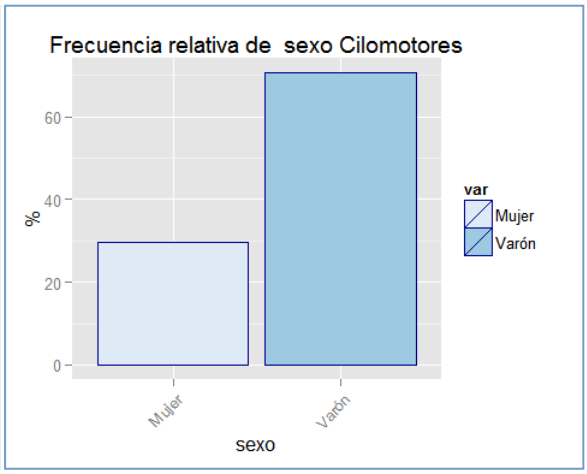

Curva ROC nnet
Sensitivity
1-Specificity
r.eta = 0.005
Sens: 91.1%
Spec: 83.4%
PV+: 0.1%
PV-: 96.4%
Variable est. (s.e.)
(Intercept) -5.444 (0.042)
pred 18.288 (0.413)
Model: clas ~ pred
Area under the curve: 0.939

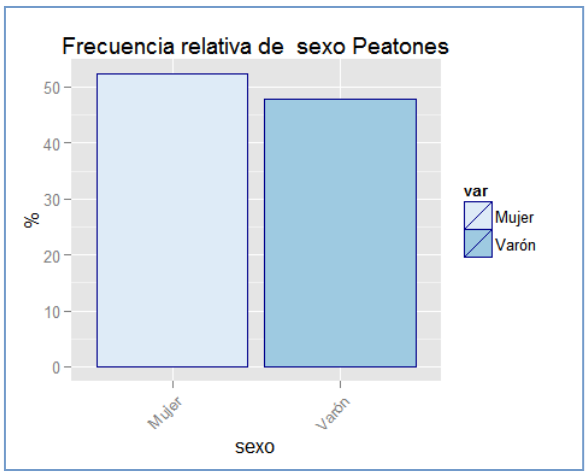
ANEXO II: COMPARACIÓN DESCRIPTIVA

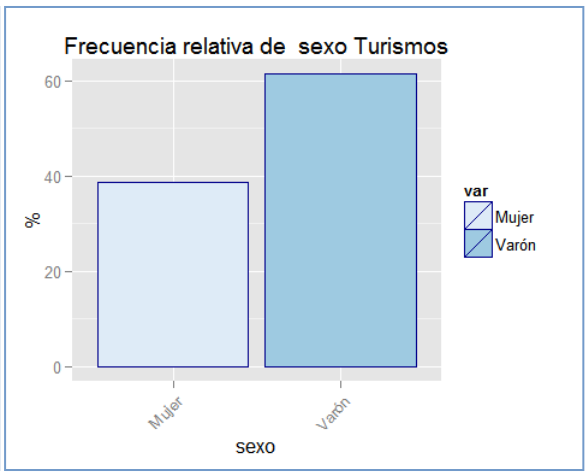


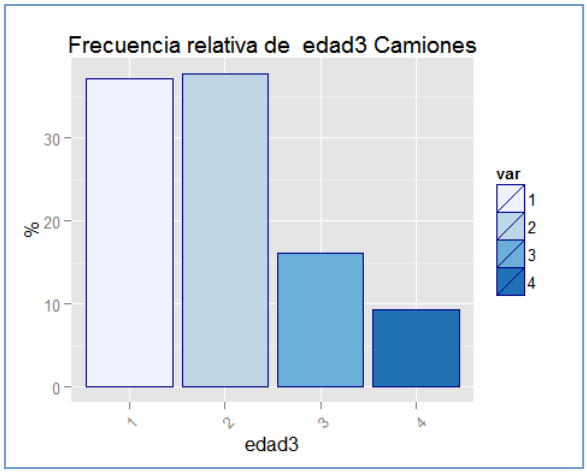


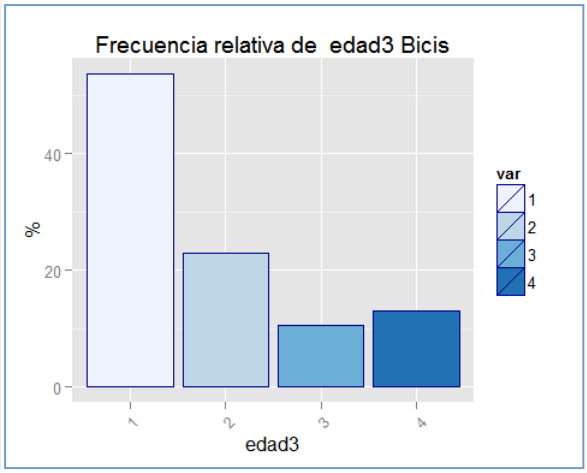


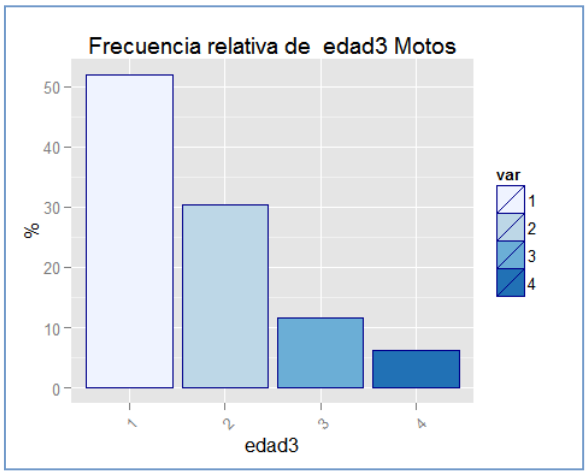


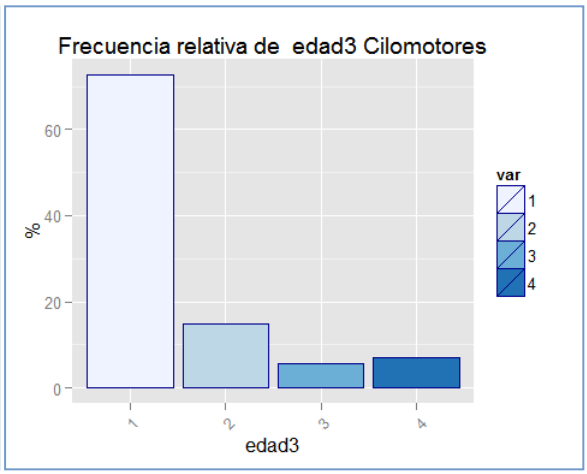


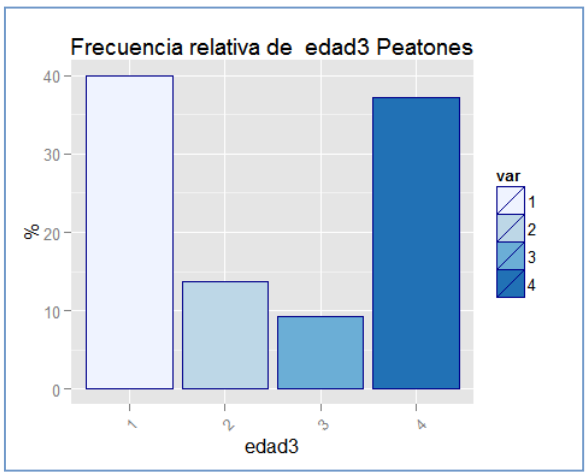


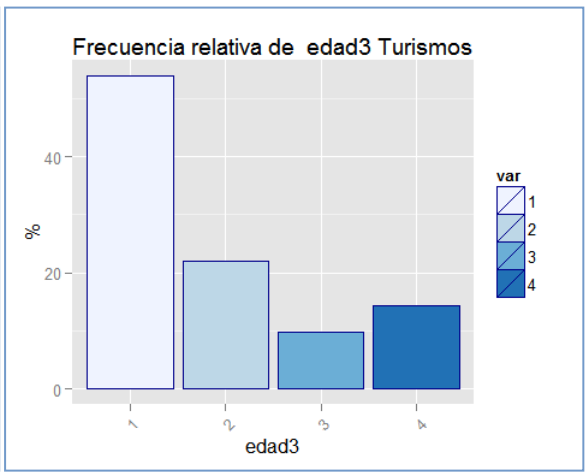

Frecuencia relativa de edad3 Camiones
%
edad3
var
1
2
3
4

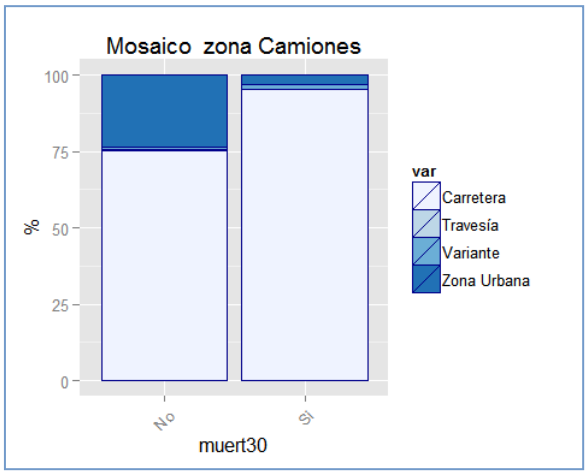

Frecuencia relativa de edad3 Bicis
%
edad3
var
1
2
3
4

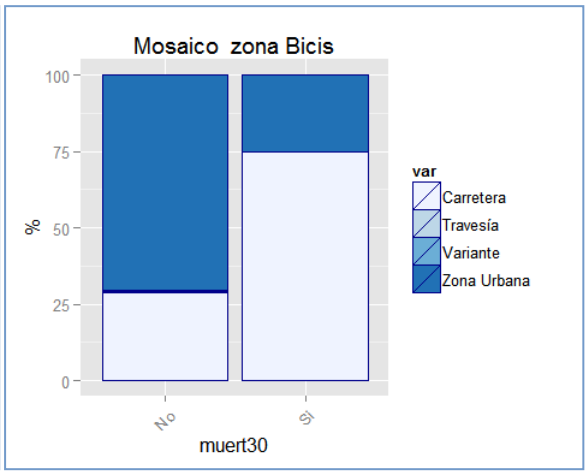

Frecuencia relativa de edad3 Motos
%
edad3
var
1
2
3
4

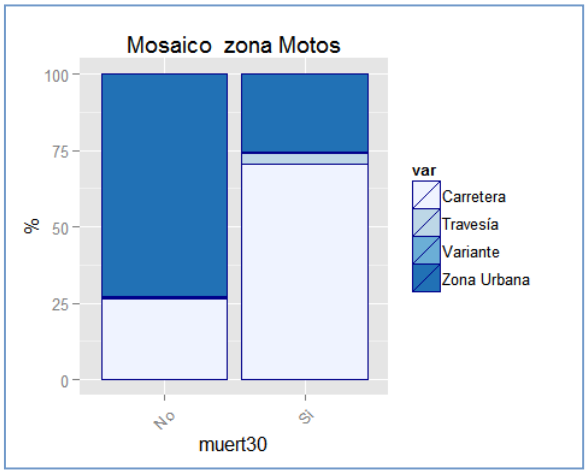

Frecuencia relativa de edad3 Cilomotores
%
edad3
var
1
2
3
4

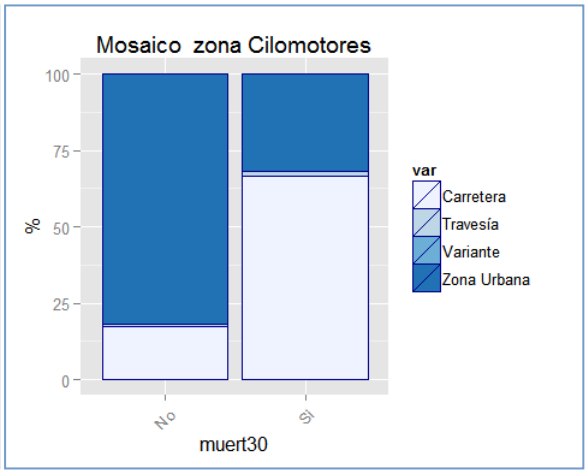

Frecuencia relativa de edad3 Peatones
%
edad3
var
1
2
3
4

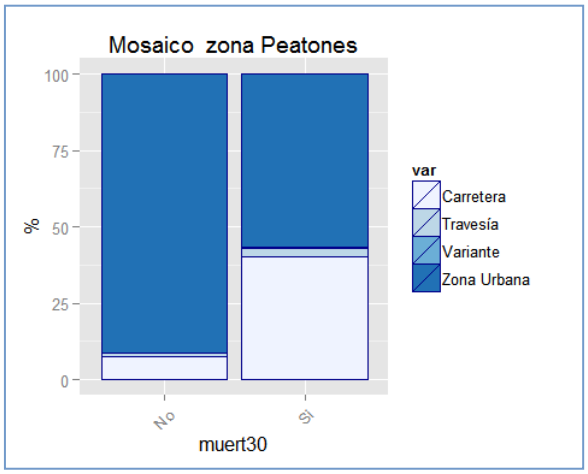

Frecuencia relativa de edad3 Turismos
%
edad3
var
1
2
3
4

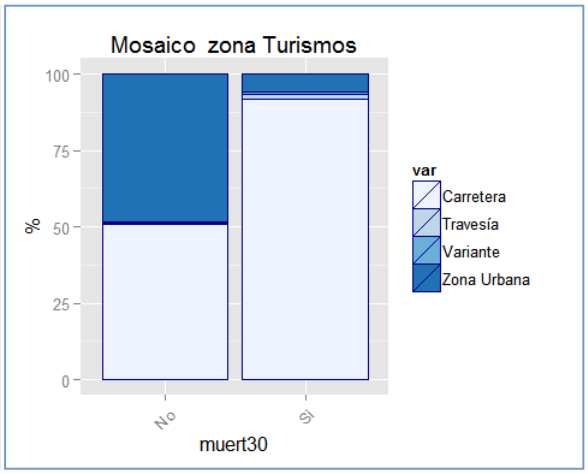

Mosaico zona Camiones
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

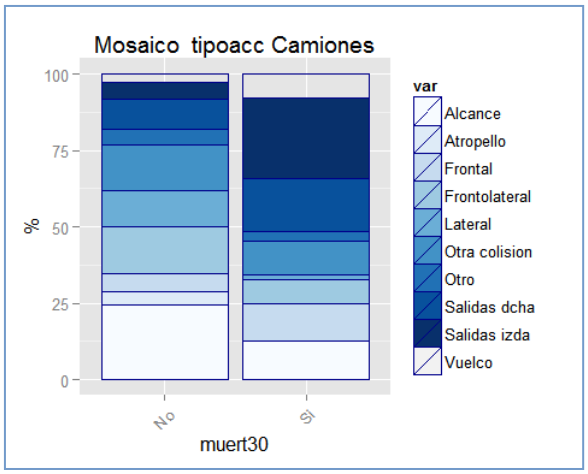

Mosaico zona Bicis
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

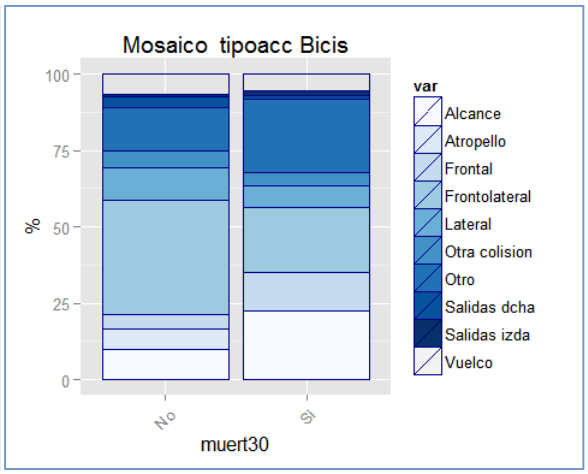

Mosaico zona Motos
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

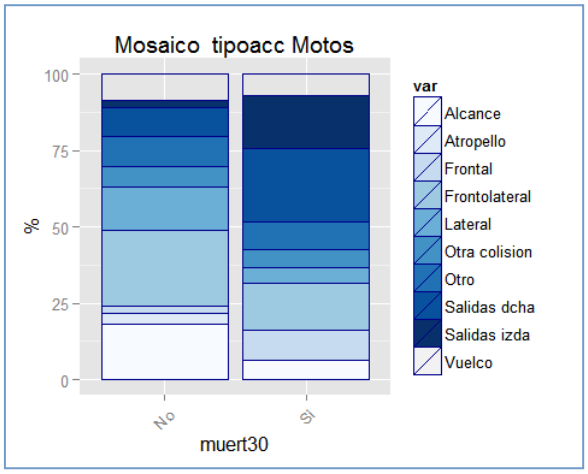

Mosaico zona Cilomotores
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

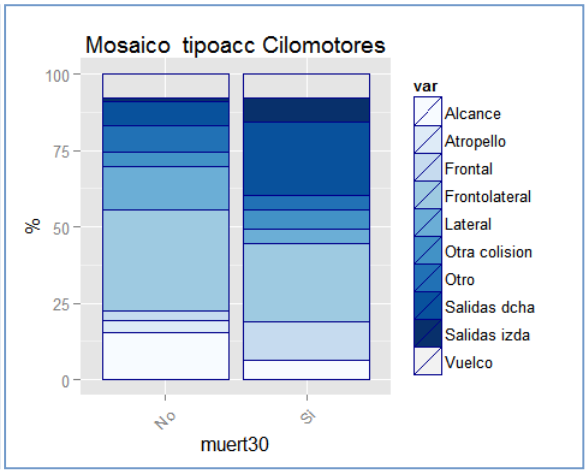

Mosaico zona Peatones
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

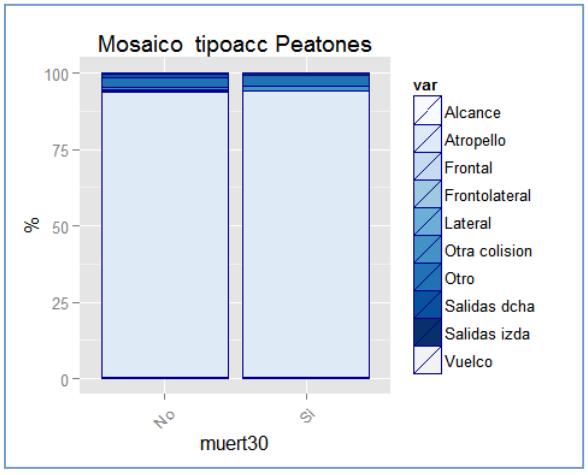

Mosaico zona Turismos
100
75
50
25
0
%
No
Si
muert30
var
Carretera
Travesía
Variante
Zona Urbana

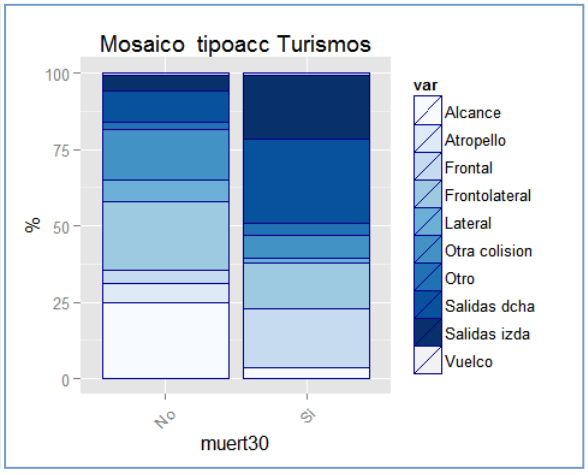

Mosaico tipoacc Camiones
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco


Mosaico tipoacc Bicis
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco


Mosaico tipoacc Motos
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco


Mosaico tipoacc Cilomotores
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco


Mosaico tipoacc Peatones
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco


Mosaico tipoacc Turismos
100
75
50
25
0
%
No
Si
muert30
var
Alcance
Atropello
Frontal
Frontolateral
Lateral
Otra colision
Otro
Salidas dcha
Salidas izda
Vuelco

ANEXO III: CÓDIGO R PRINCIPAL

PREPROCESAMIENTO

```
# Se crea el archivo con las variables de interés para el estudio

TPGV_red<- subset(TPGV,select=c(muert30, accseg , alcohol, edad,
                                sexo , infrvel,tipoveh, infcond,
                                cansan , velina , infracc ,  red,
                                zona, lumin , superf, barrera, mediana ,
                                tipoacc ,  facatm , tipovia,posveh,lesivid,
                                provin,hergrav30,distracc,denscir,hitos,tipodia,
                                mes,diasem,idveh))

# Preprocesamiento: depuración y recategorización de variables
attach(TPGV_red)
require(epicalc) # Contiene la siguiente versión de la función recode (R ver.3.03)
# Se consideran los valores ausentes como categoría 'N' en varias variables
recode(vars=c(alcohol, distracc, infracc,cansan,velina), ' ', 'N',TPGV_red)

# Se establecen los niveles 'Si' y 'No' como criterio unificado
recode(vars=c(alcohol, distracc, infracc,cansan,velina), 'N', 'No',TPGV_red)
recode(vars=c(alcohol, distracc, infracc,cansan,velina), 'S', 'Si',TPGV_red)

# La variable accesorios de seguridad se convierte en dicotómica
TPGV_red$accseg <-as.factor(ifelse(TPGV_red$accseg != 'Ninguno', 'Si', 'No'))

# Se recodifica la variable tipo de vehículo
# para posterior segmentación en subpoblaciones de interés
recode(tipoveh,1,'Bici',TPGV_red)
recode(tipoveh,2,'Ciclomotor',TPGV_red)
recode(tipoveh,11,'Moto',TPGV_red)
recode(tipoveh,c(21,22,23),'Turismo',TPGV_red)
recode(tipoveh,c(41,42,51,52,53,54,55),'Camion',TPGV_red)
recode(tipoveh,43,'Furgoneta',TPGV_red)
recode(tipoveh,c(61,62,63),'Autobus',TPGV_red)
recode(tipoveh,c(10,24,30,31,32,70,80,81,82,90),'Otro',TPGV_red)

TPGV_red$tipoveh <- as.factor(TPGV_red$tipoveh)

# Se recodifican las variables de muerte y herido grave a 30 días
# de tal manera que se consideran casos positivos los realmente constatados
# en ese momento y no se aplican los coeficientes e mortalidad para realizar
# estimaciones con las que se trabaja en las cifras oficiales.
TPGV_red$muert30 <- as.factor(ifelse(TPGV_red$muert30==1,'Si','No'))
TPGV_red$hergrav30 <- as.factor(ifelse(TPGV_red$hergrav30==1,'Si','No'))

# Se crea la variable muerto o herido grave con el fin de estudiar
# esta variable como alternativa a las anteriores, el propósito último es
# el de lograr un mayor balaceo de clases de cara a los modelo de clasificación
# que se ajustarán.
TPGV_red$muerograv <- as.numeric(TPGV_red$hergrav30) + as.numeric(TPGV_red$muert30)

TPGV_red$muerograv <- as.factor(ifelse(TPGV_red$muerograv==3,'Si','No'))


# Se recodifica infracción del conductor de 24 categorías a 8...histótico..
#attach (TPGV_red)
TPGV_red$infcond<-as.factor(TPGV_red$infcond)
```

```
recode(infcond,c("11","25","31","32","41","42","44",
                 "45","46","51","52","53","61","62","71","81"), 'Otras',TPGV_red)
recode(infcond,c('0','91','1','21','22','23','24','43'),
               c('Ninguna','Ninguna','Distraccion','Sentido contr','Invadir carril',
                 'Girar mal', 'Adelantar mal','Saltar Stop'),TPGV_red)
recode.is.na(infcond,'Ninguna',TPGV_red)

# Se recodifica el tipo de accidente de 33 a 10 categorías

recode(tipoacc,c("Atropello a conductor de animales",
                 "Atropello a peaton aislado o en grupo",
                 "Atropello a peaton reparando el vehiculo",
                 "Atropello a peaton sosteniendo bicicleta"),
                 'Atropello',TPGV_red)

recode(tipoacc,c( "Colision con animal conducido o rebaño" ,
                  "Colision con animales sueltos",
                  "Colision con barrera de paso a nivel",
                  "Colision con otro objeto o material",
                  "Colision con valla de defensal",
                  "Colision con vehiculo estacionado o averiado",
                  "Colision multiple o en caravana" ),'Otra colision',TPGV_red)

recode(tipoacc,c(  "Salida de la via por la dcha. con despeñamiento",
            "Salida de la via por la dcha. con vuelco",
            "Salida de la via por la dcha. en llano",
            "Salida de la via por la dcha. otra",
            "Salida de la via por la dcha. y choque con arbol o poste",
            "Salida de la via por la dcha. y choque con cuneta o bordillo",
            "Salida de la via por la dcha. y choque con muro o edificio",
            "Salida de la via por la dcha. y otro choque"  ),'Salidas dcha',TPGV_red)

recode(tipoacc,c(   "Salida de la via por la izda. con despeñamiento",
            "Salida de la via por la izda. con vuelco",
            "Salida de la via por la izda. en llano",
            "Salida de la via por la izda. otra",
            "Salida de la via por la izda. y choque con arbol o poste",
            "Salida de la via por la izda. y choque con cuneta o bordillo",
            "Salida de la via por la izda. y choque con muro o edificio",
            "Salida de la via por la izda. y otro choque"),'Salidas izda',TPGV_red)

recode(tipoacc,c("Colision frontal","Colision frontolateral",
                 "Colision  lateral","Colision por alcance","Vuelco en la calzada"),
               c("Frontal" ,"Frontolateral","Lateral","Alcance",'Vuelco'),TPGV_red)

# Se recodifica superficie mediante logística ##
recode(superf,c('Aceite','Barrillo','Nevada','Helada'),'Otra',TPGV_red)
recode(superf,c('Gravilla sueltal','Umbria'),'UmbriaoGrava',TPGV_red)

# Se recodifica tipovia histórico con ayuda de logística ##
recode(tipovia,c('Autovia','Autopista','Via rapida'),'Autopistas',TPGV_red)
recode(tipovia,c('Via convencional','Via convencional con carril lento'),
'Secundarias',TPGV_red)
recode(tipovia,c('Via de Servicio','Ramal de enlace','Otro tipo','Camino
vecinal'),'Otras',TPGV_red)

# Se elimina la categoría 'Desconcido' de la variable sexo
recode(sexo,'Desconocido',NA,TPGV_red)

# Se recodifica edad mediante arbol chaid. Niveles <=37 (37,49] (49,57] >57
# Se lanza el paquete car que contiene una variante interesante de la función recode
require(car)
```

```
TPGV_red$edad3 <-recode(edad," 0:37=1; 37:49=2; 49:57=3; 57:110=4; 999=NA",
                        TPGV_red,as.factor.result=T)

# Se recodifica la variable luminosidad para evitar nombres muy largos y se unen dos
categorías
TPGV_red$lumin <- recode(lumin, " c('Iluminacion insuficiente noche)',
                         'Sin iluminacion (noche)') = 'Noche';
                        'Iluminacion suficiente (noche)'= NocheIlum'",
                         as.factor.result=T,TPGV_red)

# Se recodifica la variable factores atmosfericos ayuda de logística
TPGV_red$facatm <- recode(facatm, "c('Lloviznando','Lluvia fuerte') = 'Lluvias';
                        c('Granizando', 'Nevando', 'Otro') = 'NieveYotros';
                        c('Niebla ligera','Niebla intensa') = 'Nieblas'",
                          TPGV_red, as.factor.result=T)
```

ESTUDIO DESCRIPTIVO

```
require(ggplot2)

# Función para tablas de frecuencias de las variables más relevantes

tab <- function (var) {
  return(as.data.frame(round(prop.table(table(var))*100,2)))
}

# Función para tablas de contingencia (porcentaje fila)

xtab <- function (var) {
  return(as.data.frame(round(prop.table(table(muert30,var),1)*100,3)))
}

# Función bucle para obtener tablas de frecuencia y gráficos de todas las variables

univar <- function (data, main) {

  var <- colnames(data)
  for (i in 1:length(var)) {

    cat ("La variable ", i, 'se llama ', var[i], '\n\n')
    if ((length(na.omit(data[,i])))!=0){
    t <- tab(data[,i])
    if (nlevels(t$var[i])<=10){
      g<-ggplot(t,aes(x=var,y=Freq,fill=var))+
        geom_bar(stat='identity',color='darkblue')+
        theme(axis.text.x  =element_text(angle= 45,hjust= 1 ))+
        scale_fill_brewer(palette="Blues")+labs(x=var[i],y="%")+
        ggtitle(paste("Frecuencia relativa de ",var[i],main))
      print(g)}
    colnames(t)[1]<-var[i]
    print(t) }
  }
}

univar(TPGV_red, '')

# Función bucle para obtener tablas de contingencia y gráficos de mosaico frente
# a la variable objetivo

cruces <- function (data, main) {
```

```
  attach(data)
  var <- colnames(data)
  for (i in 2:length(var)) {
    cat ("La variable ", i, 'se llama ', var[i], '\n\n')
    if ((length(na.omit(data[,i])))!=0){
    t <- xtab(data[,i])
    print(t)

    if (nlevels(t$var)<=10){
      g<-ggplot(t,aes(x=muert30,y=Freq,fill=var))+
        geom_bar(stat='identity',color='darkblue')+
        theme(axis.text.x  =element_text(angle= 45,hjust= 1 ))+
        scale_fill_brewer(palette="Blues")+labs(y="%")+
        ggtitle(paste("Mosaico ",var[i],main))
    print(g)}
  colnames(t)[2]<-var[i]
   }
}}
# Ejemplo de llamada archivo general
cruces(TPGV_red, '')
```

CREACIÓN DE SUBPOBLACIONES

```
#Subpoblaciones

bicis<-subset(TPGV_red,tipoveh =="Bici")
turismos <- subset(TPGV_red,tipoveh =="Turismo")
motos <- subset(TPGV_red,tipoveh =="Moto")
furgos <- subset(TPGV_red,tipoveh =="Furgoneta")
ciclos <- subset(TPGV_red,tipoveh =="Ciclomotor")
camiones <- subset(TPGV_red,tipoveh == "Camion")
peatones <- subset(TPGV_red,idveh =="P ")
```

GRÁFICOS DE ODDS RATIO, REGRESIÓN LOGÍSTICA

```
require(epicalc)

# Cargar modelos logísticos
logiCamion <- readRDS("logiCamion.RDS")
logiMoto <- readRDS("logiMoto.RDS")
logiBici <- readRDS("logiBici.RDS")
logiPeato <- readRDS("logiPeato.RDS")
logiCiclo <- readRDS("logiCiclo.RDS")
logiTurismo <- readRDS("logiTurismos.RDS")

# Función tabla resumen OR sigifiactivos logistica + graficos

logiOR <- function (model, main){
  resumen<-logistic.display(model, crude.p.value=F,decimal=1,simplified=T)
  tab <- as.data.frame(round(resumen$table[resumen$table[,4]<0.05,],2))
  tab<- tab[order(-tab$OR),]
  tab$x<-factor(rownames(tab),levels=unique(as.character(rownames(tab))))
  g<-ggplot(tab, aes( x=x,y=OR)) +   geom_hline(yintercept=1,colour='blue')+
    geom_errorbar(aes(ymin=lower95ci, ymax=upper95ci),
colour="lightskyblue3",width=.3,size=0.9) +
    geom_line() + xlab("Factores") +theme_bw()+
    theme(axis.text.x  =element_text(angle= 45,hjust= 1 ))+
```

```
    ggtitle(paste("OR factores significativos",main)) +
    geom_point(color='blue',size=3, shape=21, fill="white")
  print(g)

  return(tab)
}

orBici<-logiOR(logiBici,main='Bicis')
```

AJUSTE DE MODELOS CON CARET

```
# Modelos de data mining para las subpoblaciones

fiveStats = function(...) c (twoClassSummary(...), defaultSummary(...))

cv.ctrl = trainControl ( method = "repeatedcv", number = 3 , repeats = 4,
                         classProbs = TRUE,
                         summaryFunction = fiveStats )

# Función para todos los modelos de una subpoblación. (Tiempo de computación muy
elevado!!, se recomienda ejecutar los modelos por separado)

modelosDT <- function (data){

   data_na<-na.omit(subset(data,select=c(muert30,edad3,sexo,alcohol,accseg,
                                         distracc,denscir,facatm,tipodia,superf
                                         ,mediana,barrera,lumin,zona,velina,hitos
                                         ,infracc,cansan,red,tipovia,tipoacc)))
   # Rejilla de parámetros para RF
    rfGrid = expand.grid ( .mtry = c (1:20) )

   # Modelo RF 100 árboles
    rfFit <- train(muert30 ~ ., data = data_na, method = "rf", metric='ROC',
                   trControl = cv.ctrl, ntree=100,tuneGrid=rfGrid)
    print(rfFit)
    print(rfFit$finalModel)

   # Modelo RF 500 árboles
    rf500Fit <- train(muert30 ~ ., data = data_na, method = "rf", metric='ROC',
                   trControl = cv.ctrl, ntree=500,tuneGrid=rfGrid)
    print(rf500Fit)
    print(rf500Fit$finalModel)

   # Rejilla de parámetros para GBM
    gbmGrid = expand.grid (.interaction.depth = c (1,3), .n.trees = (1:20)*50,
                           .shrinkage = c (0.001, 0.01, 0.1),
                           .n.minobsinnode= c(5,10))
   # Modelo GBM
    gbmFit <- train(muert30 ~ ., data = data_na, method = "gbm", metric='ROC',
                   trControl = cv.ctrl, verbose = F, tuneGrid = gbmGrid)
    print(gbmFit)
    print(gbmFit$finalModel)

   # Rejilla de parámetros para XGB
    xgbGrid = expand.grid (.nrounds = c (50,100,200,500), .max_depth = c (3,5,10),
                           .eta = c (0.1,0.3,0.9))
```

```
   # Modelo XGB
    xgbFit <- train(muert30 ~ ., data = data_na, method = "xgbTree", metric='ROC',
                    trControl = cv.ctrl, tuneGrid=xgbGrid)
    print(xgbFit)
    print(xgbFit$finalModel)

   # Rejilla de parámetros para NNet
    nnetGrid = expand.grid (.size = c (1,2,5,10,15,20),
                            .decay = c (0.001, 0.1, 0.4, 0.6))
   # Modelo NNet
    nnetFit <- train(muert30 ~ ., data = data_na, method = "nnet", metric='ROC',
                    trControl = cv.ctrl, trace = F, verbose = F, tuneGrid = nnetGrid)
    print(nnetFit)
    print(nnetFit$finalModel)

   # Rejilla de parámetros para LogitBoost
    lbGrid = expand.grid (.size = c (1,10,50,100,500,1000,2000)
   # Modelo LogitBoost
    LogitBoostFit <- train(muert30 ~ ., data = data_na, method = "LogitBoost",
                     metric='ROC', trControl = cv.ctrl, tuneGrid = lbGrid)
    print(LogitBoostFit)
    print(LogitBoostFit$finalModel)

   # Modelo bayesiano lineal generalizado
    bayesFit <- train(muert30 ~ ., data = data_na, method = "bayesglm",
                     metric='ROC', trControl = cv.ctrl)
    print(bayesFit)
    print(bayesFit$finalModel)

return(list(RF = rfFit, RF500 = rf500Fit, XGB = xgbFit, Bayes =bayesFit,
            LogitBoost = logboostFit, GBM = gbmFit, NNet = nnetFit))
}
# Ejemplo de llamada para camiones
modelosCamion<-modelosDT(camiones)

# Se guarda el objeto creado
saveRDS(modelosCamion, 'modelosCamion.RDS')
```

IMPORTANCIA DE FACTORES, CLASIFICACIÓN Y PREDICCIONES

```
# Factores de influencia (importancia de las variables)

impCamion <- lapply(modelosCamion, varImp)

lapply(impCamion, plot(top=20))

# Ejemplo control del proceso de ajuste por validación cruzada repetida (12 muestras)
resampCamion <- resamples(modelosCamion)

# Ejemplo gráfico ROC clasificación automática (apartado 9.1)
dotplot(resampCamion,metric = "ROC",main='Subpoblación de Camiones')


# Ejemplo de función de predicciones (predict no se comporta de la misma forma en
todos los modelos)
predicciones <- function (model,data,corte=0.5){
  data_na<-na.omit(subset(data,select=c(muert30,edad3,sexo,alcohol,accseg,
                                        distracc,denscir,facatm,tipodia,superf
```

```
                                        ,mediana,barrera,lumin,zona,velina,hitos
                                        ,infracc,cansan,red,tipovia,tipoacc)))
  print(dim(data_na)[1])
  pred <- predict(model,newdata=data_na,type='prob')[,'Si']
  summary(pred)
  print(length(pred))
  clas<-data_na$muert30
  cat('Gráficos de evaluación del modelo xgb')
  r<-ROC(form=clas~pred,stat=T,plot='ROC',PV=T,MX=T,AUC=T,MI=T,data=data_na,
         main=paste('Curva ROC xgb'))
  priorProb <-   as.numeric(table(data_na$muert30)[2])/
                 as.numeric(table(data_na$muert30)[1])

  cat('\n La prevalencia de fallecidos en la subpobalcion es ',priorProb,
      ' con este punto de corte:\n\n')
  predClas <- ifelse(pred>priorProb,'Si','No')
  print(length(predClas))
  print('ok')
  print(confusionMatrix(table(predClas,data_na$muert30),positive='Si'))
  cat('\n Moviendo el punto de corte de la probabilidad estimada a ',corte, '\n\n')

  predClas <- ifelse(pred>corte,'Si','No')
  tabla<-table(predClas,data_na$muert30)
  print(tabla)
  if (dim(tabla)[1]==2){
    print(confusionMatrix(tabla,positive='Si'))} else {
      cat('No se puede calcular la matriz de confusión \n')
    }
  return(pred)
}
# Ejemplo de llamada para camiones (se guardan las probabilidades estimadas)
xgbPredCamion<-predicciones(xgbCamion,camiones)
```

ENSAMBLE DE MODELOS

```
# Estudio de correlaciones entre predicciones(GRÁFICOS DISPERSIÓN)

plot(predCamion, main ='Camiones')
plot(predMoto, main ='Motos')
plot(predBici, main ='Bicis')
plot(predPeato, main ='Peatones')
plot(predCiclo, main ='Ciclos')
plot(predTurismo, main ='Turismos')

# Gráficos de correlacion
corCamion<-cor(predCamion)
corrplot(corCamion)
corMoto<-cor(predMoto)
corrplot(corMoto)
corBici<-cor(predBici)
corrplot(corBici)
corPeato<-cor(predPeato)
corrplot(corPeato)
corCiclo<-cor(predCiclo)
corrplot(corCiclo)
corTurismo<-cor(predTurismo)
corrplot(corTurismo)
```

```
# Función ensambles.

ensambles <- function (pred ,data){
  data_na<-na.omit(subset(data,select=c(muert30,edad3,sexo,alcohol,accseg,
                                        distracc,denscir,facatm,tipodia,superf
                                        ,mediana,barrera,lumin,zona,velina,hitos
                                        ,infracc,cansan,red,tipovia,tipoacc)))
  class <- data_na$muert30

  # Modelo Logístico
     l <- step(glm(class ~ .,data=pred, family=binomial(link='logit')))
     print(summary(l))

  # Enamble Logístico predicciones
       predi <- predict(l,type='response')
       par(mfrow=c(1,1))
       cat('Ensamble Regresión 1: \n')
       print(summary(predi))

       r1<-ROC(form=class~predi,stat=T,plot='ROC',PV=T,MX=T,
               AUC=T,MI=T,data=pred,
               main=paste('Curva ROC EnsRegPred'))

  # Enamble Logístico pesos
    # Coeficientes (valor absoluto)
      v<-abs(l$coeff[-1])
    # Pesos
      w<-v/sum(v)
      w
    # Seleccionar modelos implicados
      dt<-pred[,names(w)]
      ensReg <- as.numeric(as.matrix(dt)%*%w)
      cat('Ensamble Regresión 2: \n')
      print(summary(ensReg))

      r2<-ROC(form=class~ensReg,stat=T,plot='ROC',PV=T,MX=T,
              AUC=T,MI=T,data=pred,
              main=paste('Curva ROC EnsRegw'))

  # Ensamble 80% GBM y 20% RF
     ens2080 <- 0.8*pred$GBM+0.2*pred$RF500
     cat('Ensamble 80% GBM 20% RF: \n')
     print(summary(ens2080))

     r3<-ROC(form=class~ens2080,stat=T,plot='ROC',PV=T,MX=T,
             AUC=T,MI=T,data=pred,
             main=paste('Curva ROC Ens2080'))

  # Ensamble media de los 8 modelos
     ensMean <- apply(pred[1:8],1,mean)
     cat('Ensamble Media aritmética: \n')
     print(summary(ensMean))

     r4<-ROC(form=class~ensMean,stat=T,plot='ROC',PV=T,MX=T,
             AUC=T,MI=T,data=pred,
             main=paste('Curva ROC EnsMean'))
}
# Ejemplo de llamada para camiones (utiliza las predicciones del apartado anterior)
ensambles(predCamion,camiones)
```