Uso de árboles de decisión como herramienta para generar un modelo preventivo de seguridad vial urbana en la ciudad de Trenque Lauquen, Pcia. Buenos Aires

Marcos, Carlos Eduardo*; Martínez Micakoski, Fernanda; Marcos, Candela

Facultad Regional Trenque Lauquen, Universidad Tecnológica Nacional.
Racedo 298. Trenque Lauquen, Bs. As., Argentina. marcoscarloseduardo@gmail.com

RESUMEN.

La seguridad vial se refiere a las medidas adoptadas para reducir el riesgo de lesiones y muertes causadas en el tránsito. Los traumatismos por accidentes de tránsito son un problema de salud pública a nivel mundial. Argentina, a través de la Agencia Nacional de Seguridad Vial, adhirió a los objetivos y finalidades del Decenio de Acción para la Seguridad Vial de la ONU. Entre sus acciones se encuentra la de recabar datos a nivel municipal mediante el uso del Formulario Estadístico Único (actualmente) y sus versiones anteriores, desde finales del año 2011.
El presente trabajo se basa en el uso de árboles de decisión como herramienta para generar un modelo preventivo de seguridad vial urbana, de manera de reducir la proporción de personas hospitalizadas por este tipo de eventos.
Las variables predictoras son una herramienta útil para que los agentes de tránsito puedan realizar acciones preventivas en base a la propia idiosincrasia de accidentalidad vial de la comunidad, poniendo en valor los datos registrados en siniestros desde el año 2012 a 2017.
Se utilizó el software libre R como facilitador de investigación estadística reproducible.
Los resultados reflejan que variables asociadas principalmente al factor humano permiten predecir la existencia de lesiones en los participantes de un siniestro.

Palabras Claves: siniestralidad vial, modelización, factor humano, random forest, machine learning.

ABSTRACT

Road safety refers to measures taken to reduce the risk of injuries and deaths caused in transit. Road traffic injuries are a global public health problem. Argentina, through the National Road Safety Agency, adhered to the goals and purposes of the UN's Decade of Action for road safety. Among its actions is to collect data at the local urban level through the use of the single statistical form (currently) and its previous versions, from the end of the year 2011.
This work is based on the use of decision trees as a tool to generate a preventive model of urban road safety, in order to reduce the proportion of people hospitalized by this type of events.
Predictor variables are a useful tool for transit agents to carry out preventive actions based on the community's own idiosyncrasy of road accidents, valuing the data recorded in accidents from 2012 to 2017.
Free software R was used as a facilitator of reproducible statistical research.
The results show that variables mainly associated with the human factor make it possible to predict the existence of injuries in the participants of a claim.

1. INTRODUCCIÓN

En el marco del proyecto de investigación “Buenas prácticas en la planificación de asignación de recursos de la Dirección de Tránsito en una ciudad de 50.000 habitantes. Caso de aplicación ciudad de Trenque Lauquen”, homologado por la Secretaria de Ciencia, Tecnología y Posgrado de la Universidad Tecnológica Nacional se desarrolló una **modelización**, en base a los registros históricos de siniestralidad vial urbana, que facilita la priorización de las variables a tener en cuenta para **minimizar la proporción de participantes que requieren atención médica** como consecuencia de un colisión en la vía pública.

Los registros de tránsito se basan en el **Formulario Estadístico Único** [1] (actualmente) y sus versiones anteriores, desde finales del año 2010, por lo que las variables intervinientes en un siniestro que no puedan ser obtenidas a partir del mismo son ignoradas en el presente trabajo. Adicionalmente se incorporan una serie de variables con información de “causa aparente” que solicitaba la Provincia de Buenos Aires y que fueron relevadas por los agentes de tránsito.

La modelización de los datos pretende **identificar las variables que mejor predicen la *hospitalización*** de los participantes involucrados en el siniestro de manera que los Agentes de Tránsito puedan realizar tareas preventivas sobre las mismas. Se considera *hospitalización* toda atención médica que no se brinde en el lugar del hecho sino en el Hospital Municipal.

La **Minería de Datos** o **Explotación de Información**, es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, siendo su principal objetivo **encontrar información oculta o implícita**, que no es posible obtener mediante métodos estadísticos convencionales. La entrada al proceso de minería está formada por contenedores de información diversos, esto incluye bases de datos relacionales, almacenes de datos (Datawarehouse), documentos en texto libre, datos de la Web, entre otros [2].

La técnica **SEMMA**, acrónimo de las cinco fases del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Evaluación), surge en el año 2000 y tiene como fin establecer las etapas principales de un proceso de minería de datos [3].

En los últimos años, los investigadores han empleado cada vez más enfoques basados en **Machine Learning** a los registros de datos de siniestros viales buscando ayudar a comprender los factores causales de las colisiones y la gravedad de las lesiones. Esto permitiría a los responsables de la toma de decisiones formular mejores políticas de control de la seguridad vial.

Chong et. al. [4], aplicaron redes neuronales, árboles de decisión y una combinación híbrida de árboles de decisión y redes neuronales para construir modelos que puedan predecir la severidad de las lesiones. La precisión de clasificación obtenida en sus trabajos revelan que, para las lesiones no incapacitantes, las lesiones incapacitantes y las lesiones fatales, el enfoque híbrido funcionó mejor que la red neuronal y los árboles de decisión. Para la ausencia de lesiones y las distintas clases de lesiones, el enfoque híbrido obtuvo un mejor resultado que la red neuronal. Finalmente, la ausencia de lesiones o no de lesiones podrían modelarse mejor directamente mediante árboles de decisión.

Bravo Rocca et. al. [5], desarrollaron una aplicación con el objetivo de reducir el riesgo de integridad física de los peatones mediante la geolocalización, en tiempo real, de lugares más seguros para caminar posibilitando identificar las áreas que tienen la mayor incidencia de diferentes tipos de incidentes. Esta funcionalidad permite a los usuarios elegir rutas más seguras teniendo en cuenta la información proporcionada para cada sector. Los incidentes considerados fueron: "Incendios", “Atropello de peatones por vehículos”, “Colisiones de vehículos”, “Despiste de vehículos”, “Caída de peatones”, “Fuga de gas”, “Vuelco de vehículo” e “Inundación”. El modelo predictivo que utiliza la Regresión Logística Múltiple obtuvo un rendimiento pobre comparado con el algoritmo Random Forest para este tipo de datos.

Shanti y Ramani [6] publicaron un trabajo donde se enfatiza la importancia de los algoritmos de clasificación de Data Mining en la predicción de los patrones de colisión de vehículos a través de datos históricos de accidentes. Los algoritmos de clasificación C4.5, C-RT, CS-MC4, Decision Tree, ID3, Naïve Bayes y Random Forest se aplicaron para predecir los patrones de colisión de vehículos. Los resultados experimentales indicaron que el algoritmo de clasificación Random Forest logró una mayor precisión que otros algoritmos en la clasificación de la forma de colisión que aumenta la tasa de mortalidad en accidentes de tráfico.

Krishnaveni y Hemalatha [7] investigaron sobre los modelos de clasificación para predecir la gravedad de las lesiones que se producen durante los accidentes de tráfico. Compararon las opciones Naive Bayes Bayesian, AdaBoostM1 Meta, PART Rule, J48 Decision Tree y Random Forest para clasificar el tipo de severidad de lesión de 34.500 accidentes de tráfico. El resultado final demostró que el clasificador de Random Forest supera a los otros cuatro algoritmos.

El objetivo principal de este trabajo es obtener las variables predictoras de hospitalización en base a los datos de siniestralidad vial históricos aplicando la técnica de agregación de múltiples árboles de clasificación mediante remuestreo (**random forest**) utilizando software libre [8].

2. METODOLOGÍA

El proceso de análisis se basó en la técnica SEMMA.

Para realizar este trabajo se accedió a la base de datos con registros de siniestros mantenida por investigadores de la Facultad Regional Trenque Lauquen de la Universidad Tecnológica Nacional en el marco de proyectos de seguridad vial en la ciudad desde el año 2011 a la fecha en forma ininterrumpida.
Se conformó un archivo con todas las observaciones y se importó para ser procesado mediante el software libre R.
Durante la exploración de las variables se determinó que la vista minable posee 76 variables y 4.895 observaciones. La primera es del 01 de Diciembre de 2011 y la última del 21 de Abril de 2018. La variable a predecir es "*hospitalizado"*.
Se generaron nuevas variables relacionadas con el tiempo a partir de la variable *fecha*, como hora, día de la semana (lunes a domingo), semana del mes (primera a cuarta), mes y año, procediendo a eliminar la variable original *fecha*.
Se eliminaron variables cuya metodología de registración es arbitraria para modelar, como por ejemplo la calle, la altura, entre calles, etc. El mismo criterio se siguió con aquellas variables en las que había subregistro o el mismo era dudoso.
Se auditaron los datos en base a información adicional como pertenencia o no a la zona (urbana, suburbana, rural, ruta), el momento del día en base al registro de salida y puesta del sol, la presencia o no de rotondas y semáforos en el lugar del evento, cercanía del paso a nivel, etc. Se utilizó la redundancia de información en las variables para corregir errores de registro en las observaciones.
Las variables disponibles se agruparon en base al modelo de la Matriz de Haddon [9] y se eliminaron todas aquellas sobre las que no es posible actuar de manera preventiva, es decir, las que no son anteriores al evento.
Para poder modelizar se identificaron las variables con datos faltantes y se completaron dichos valores con la leyenda '*Na'* para generar los modelos y luego evaluar la coherencia del mismo en base a las variables predictoras seleccionadas.
Para permitir que las variables temporales sean representativas de los datos se tomaron observaciones de años completos, por lo que se limitaron las observaciones al rango de años 2012 a 2017 inclusive.
Para la modelización la tabla minable contenía 43 variables y 3.410 observaciones. A continuación se puede observar el listado de las mismas (ver Tabla 1 y 2):

Tabla 1. Variables de estudio cualitativas.

Variable	Posibles Valores	Variable	Posibles Valores
Lugar Vía Pública	Curva – Recta - Intersección	Causa Aparente Embestir Animal	Si - No
Tipo de Vía	Calle – Otro - Distribuidor - Paso a nivel FF.CC. - Rotonda	Causa Aparente Mal Vehículo	Si - No
Estado Bueno	Sí - No	Causa Aparente Maniobra Riesgosa	Si - No
Estado Baches	Sí - No	Causa Aparente Meteorología Adversa	Si - No
Estado Ahuellamiento	Sí - No	Vehículo	Auto – Bici – Moto – Nc – Otro - Peatón Utilitario
Estado Mojado	Sí - No	Hospitalizado	Si - No
Estado Escarcha	Sí - No	Sexo	F - M
Estado Resbaladizo	Sí - No	Cinturón	Si – No – Nc
Tiempo	Bueno – Lluvia - Granizo – Llovizna - Nublado	Casco	Si – No – Nc
Viento Fuerte	Sí – No	Ubicación en el Vehículo	Conductor - Acompañante delantero - Acompañante trasero - Acompañante v2r - Nc
Niebla	Sí – No	Licencia Vehículo	Si – No – Nc
Luminosidad	Amanecer – Día – Noche - Atardecer	Seguro Vehículo	Si – No – Nc

Luz Artificial	Sí – No
Prioridad Semáforo	Sí – No
Prioridad Señal Ceda el Paso	Sí – No
Prioridad Marca Vial	Sí – No
Prioridad Rotonda	Sí – No
Prioridad Ninguna	Sí – No
Vía Dividida	Cordón o boulevard divisorio –Guardarrail – Línea divisoria – Otro – Ninguna
Causa Aparente Cansancio	Sí – No
Causa Aparente Deficiencia Vía	Sí – No

VTV Vehículo	Si – No – Nc
Cuadrícula geográfica	Matriz A0 a E5
Semana	Primera, Segunda, Tercera, Cuarta
Día de la Semana	Lun, Mar, Mie, Jue, Vie, Sab, Dom
Hora	0 a 23
Causa Aparente Distracción	Sí – No
Causa Aparente Enfermedad	Sí – No
Causa Aparente Exceso Velocidad	Sí – No

Tabla 2. Variables de estudio cuantitativas.

Variable	Posibles Valores
Edad	0 a 99

Variable	Posibles Valores
Mes	1 a 12

Se verificó qué variables predictoras poseían valor único (varianza cero) o cuáles poseían pocos valores únicos en relación al número de observaciones y la mayor tasa entre la frecuencia del más común de los valores respecto a la frecuencia del segundo valor más común para determinar si se los mantenía o excluía del modelo [10].
Se dividió el dataset en dos partes, el 70% para entrenamiento y el 30% restante para testeo. Se verificó que la proporción de la variable respuesta en ambas fuera similar.
Las variables de tipo *carácter* se convirtieron a *factor*.
Se activó el procesamiento en paralelo con tres núcleos en la computadora de análisis, cuyas características son Intel I5-3330 con 8 GB de memoria RAM.
Se aplicó la modelización mediante aprendizaje automático utilizando *random forest*. Inicialmente con los parámetros por defecto para el paquete *randomForest* [11], y luego se optimizaron los parámetros: número de variables como candidatas para cada división y número de árboles para el mínimo error de out-of-bag.
La evaluación de la performance del modelos se basó en la tasa de acierto y el área bajo la curva ROC (Receive Operating Characteristics) [12].
Finalmente se analizó la dependencia parcial de las principales variables predictoras [13] del modelo random forest en función de la variable *hospitalizado*.
En nuestro país el 89,5% de los siniestros se producen por error humano de acuerdo con la información del Centro de Experimentación y Seguridad Vial [14], un 88% de los accidentes tienen como causa principal al factor humano, seguido por cuestiones del medio (10%) y del automóvil (2%). El informe advierte que generalmente participan varios factores al mismo tiempo, por eso se releva cuál es el que tuvo más incidencia en el hecho. Es esperable que las variables más significativas para el modelo estén asociados a la conducta humana, por sobre el entorno y el vehículo.

3. RESULTADOS OBTENIDOS

Visualmente se procedió a conocer la distribución de las variables en función de la variable hospitalizado. Un ejemplo de variable continua se puede observar en la distribución de *edad* (ver Figura 1). En gris se representan las observaciones donde el participante no requirió atención médica en el hospital municipal y en naranja los que sí requirieron algún tipo de atención, ya sea en la guardia, internación en sala común, cuidados intensivos o morgue en caso de fallecimiento. Participantes de todas las edades han resultado ilesos o con lesiones, pero la proporción con mayor hospitalización se produce hasta los 30 años.

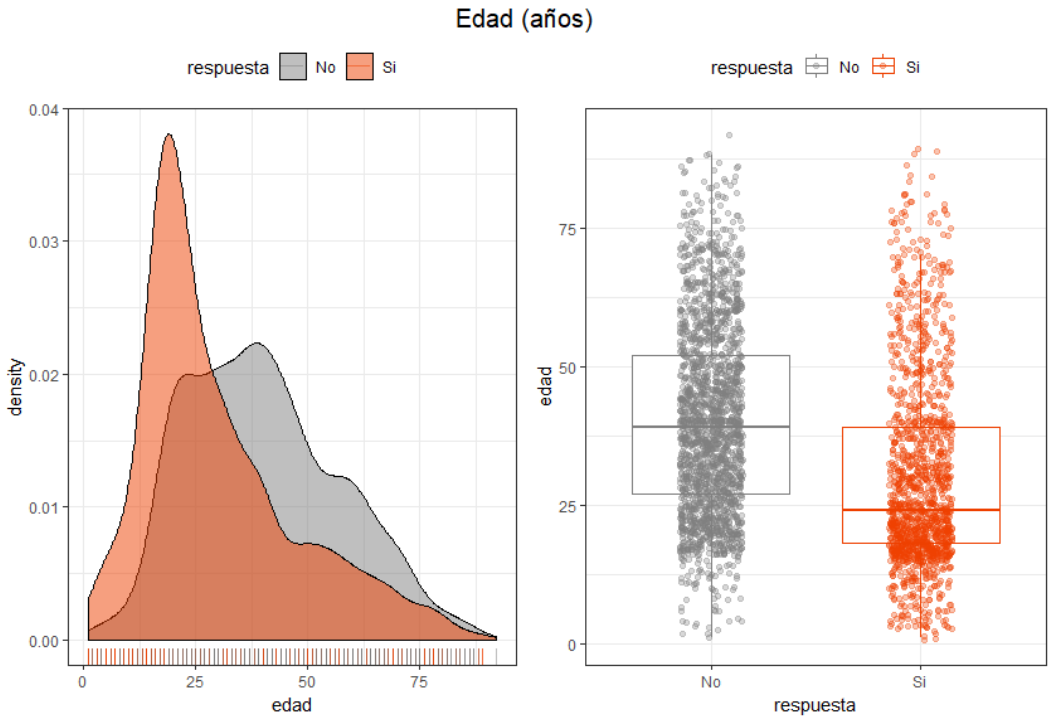

Figura 1 *Distribución de la variable Edad respecto de la variable respuesta Hospitalizado.*

Para el caso de variables categóricas, resulta ilustrativo el ejemplo de *vehículo* (ver Figura 2 y Tabla 1).

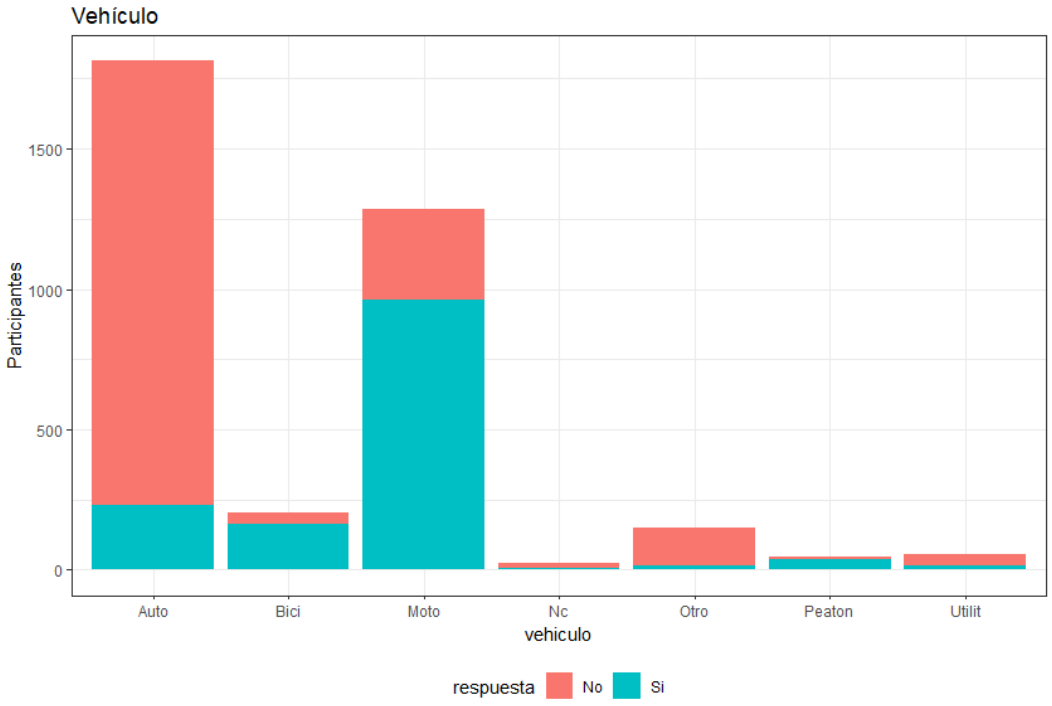

Figura 2 *Distribución de la variable Vehículo respecto de la variable respuesta Hospitalizado*

Tabla 1 *Proporción de participantes hospitalizados por tipo de movilidad*

Vehículo	***No hospitalizado***	***Sí hospitalizado***
Automóvil	0.87	0.13
Bicicleta	0.20	0.80
Motocicleta	0.25	0.75
Sin registrar	0.80	0.20
Otra movilidad	0.90	0.10
Peatón	0.11	0.89
Utilitario	0.75	0.25

Cuando una variable tiene varianza igual o cercana a cero suele añadir al modelo más ruido que información, por lo que debe evaluarse si es conveniente excluirla. Si alguno de los niveles de una

variable cualitativa tiene muy pocas observaciones en comparación a los otros niveles, existe la posibilidad que, durante la validación cruzada o bootstrapping, algunas particiones no contengan ninguna observación de dicha clase, lo que puede dar lugar a errores.
La evaluación de valores únicos y varianzas cercanas a cero arrojó los siguientes resultados para las 20 variables más comprometidas (ver Tabla 2):

Tabla 2 *Valores únicos y varianzas cercanas a cero*

Variable	Ratio de frecuencias	% de Valores únicos	Varianza nula	Varianza cercana a cero
estadoEscarcha	1.691,50	0,0590842	FALSE	TRUE
vientoFuerte	845,25	0,0590842	FALSE	TRUE
prioridadSenalCedaPaso	676,00	0,0590842	FALSE	TRUE
niebla	281,08	0,0590842	FALSE	TRUE
estadoAhuellamiento	210,56	0,0590842	FALSE	TRUE
prioridadRotonda	198,11	0,0590842	FALSE	TRUE
estadoBaches	115,72	0,0590842	FALSE	TRUE
estadoResbaladizo	90,48	0,0590842	FALSE	TRUE
causaCansancio	79,59	0,0590842	FALSE	TRUE
causaEmbestirAnimal	74,22	0,0590842	FALSE	TRUE
causaEnfermedad	71,02	0,0590842	FALSE	TRUE
causaMalEstadoVehiculo	66,70	0,0590842	FALSE	TRUE
causaMeteorologiaAdversa	42,39	0,0590842	FALSE	TRUE
causaDeficienciaVia	28,18	0,0590842	FALSE	TRUE
nroUbicacionVehiculo	14,89	0,1772526	FALSE	FALSE
prioridadSemaforo	14,31	0,0590842	FALSE	FALSE
Tiempo	14,05	0,1772526	FALSE	FALSE
causaExcesoVelocidad	13,10	0,0590842	FALSE	FALSE
estadoMojado	12,12	0,0590842	FALSE	FALSE

Ratio de frecuencias: ratio entre la frecuencia del valor más común y la frecuencia del segundo valor más común.
Porcentaje de valores únicos: número de valores únicos dividido entre el total de muestras (multiplicado por 100).
Dado que ninguna posee varianza cero se las mantuvo en el modelo para luego analizarlas en caso que se encuentren entre las variables predictivas más importantes.
La proporción de la variable respuesta (hospitalizado) positiva fue del 41% de las observaciones. Los grupos de datos de entrenamiento y de testeo obtuvieron similar proporción respecto de la variable hospitalizado (ver Tabla 3):

Tabla 3 Proporción de la variable respuesta en los *datasets de entrenamiento y testeo*

respuesta	***Proporción***
No	0.589
Sí	0.411

El *modelo 1* con los parámetros por defecto de la función *randomforest* obtuvo una tasa de error para las observaciones out-of-bag de 17,27% y la matriz de confusión que se observa a continuación (ver Figura 2).

```
Call:
 randomForest(formula = respuesta ~ ., data = datos_train_rf,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 17.24%
Confusion matrix:
     No  Si class.error
No 1159 252   0.1785967
Si  161 823   0.1636179
```

Figura 2 *Tasa de error OOB y Matriz de confusión del modelo 1*

La optimización del número de variables como candidatas para cada división mediante la estrategia de *random search* con el paquete *caret* [15] fue de 10 (ver Figura 3).
Accuracy es el porcentaje de instancias correctamente clasificadas respecto del total de instancias. Es más útil en una clasificación binaria que en los problemas de clasificación multiclase porque puede ser menos claro cómo se desglosa exactamente la precisión entre esas clases.

Kappa o Kappa de Cohen es como la precisión de clasificación, excepto que excluye de la concordancia observada aquella que es debida al azar en el conjunto de datos. Es una medida más útil para usar en problemas que tienen un desequilibrio en las clases.

```
Random Forest

2395 samples
  42 predictor
   2 classes: 'No', 'Si'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2154, 2156, 2156, 2156, 2156, 2155, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   10   0.8286786  0.6493395
   12   0.8274182  0.6466923
   13   0.8275640  0.6471088
   16   0.8265860  0.6452129
   23   0.8256161  0.6433876
   31   0.8251994  0.6425122
   33   0.8245021  0.6411793
   46   0.8233910  0.6387273
   53   0.8225553  0.6369325
   61   0.8214425  0.6346381
   62   0.8231068  0.6378762
  103   0.8200455  0.6316389
  106   0.8203273  0.6320514
  130   0.8189361  0.6293714

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 10.
```

Figura 3 *Número de variables como candidatas para cada división por el método de random search*

Y su representación gráfica (ver Figura 4).

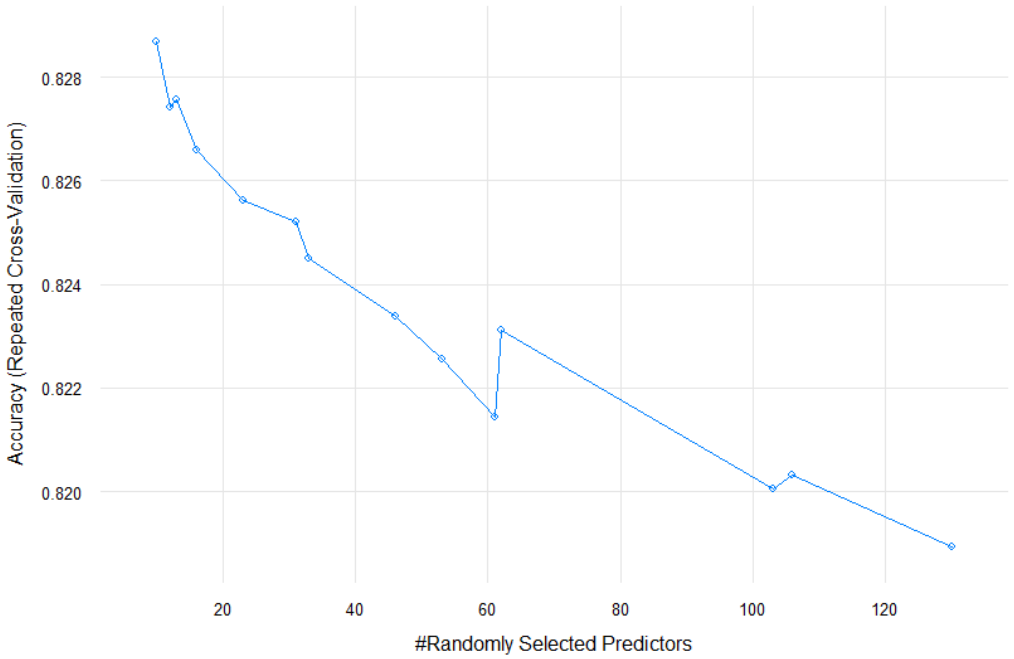

Figura 4 *Representación gráfica del número de variables como candidatas para cada división por el método de random search*

Mediante el método *grid search*, verificamos la mayor exactitud entre un listado de cantidad de variables para cada división que van de 2 a 17, por lo que contiene los valores por defecto (6) y el óptimo para el algoritmo de *random search* (10). A continuación los resultados (ver Figura 5).

```
Random Forest

2395 samples
  42 predictor
   2 classes: 'No', 'Si'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2154, 2156, 2156, 2156, 2156, 2155, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8087785  0.6014173
   3    0.8281184  0.6468970
   4    0.8295097  0.6504543
   5    0.8288140  0.6493877
   6    0.8278401  0.6472985
   7    0.8292348  0.6502786
   8    0.8290953  0.6500099
   9    0.8275646  0.6468463
  10    0.8268696  0.6454859
  11    0.8276989  0.6472208
  12    0.8288158  0.6496613
  13    0.8270056  0.6460214
  14    0.8263082  0.6444298
  15    0.8267249  0.6454231
  16    0.8268650  0.6456946
  17    0.8260281  0.6441167

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

Figura 5 *Número de variables como candidatas para cada división por el método de random search*

Gráficamente, y en diferente escala respecto del gráfico 4, podemos observar las variaciones en la exactitud en función de las variables candidatas para cada división.

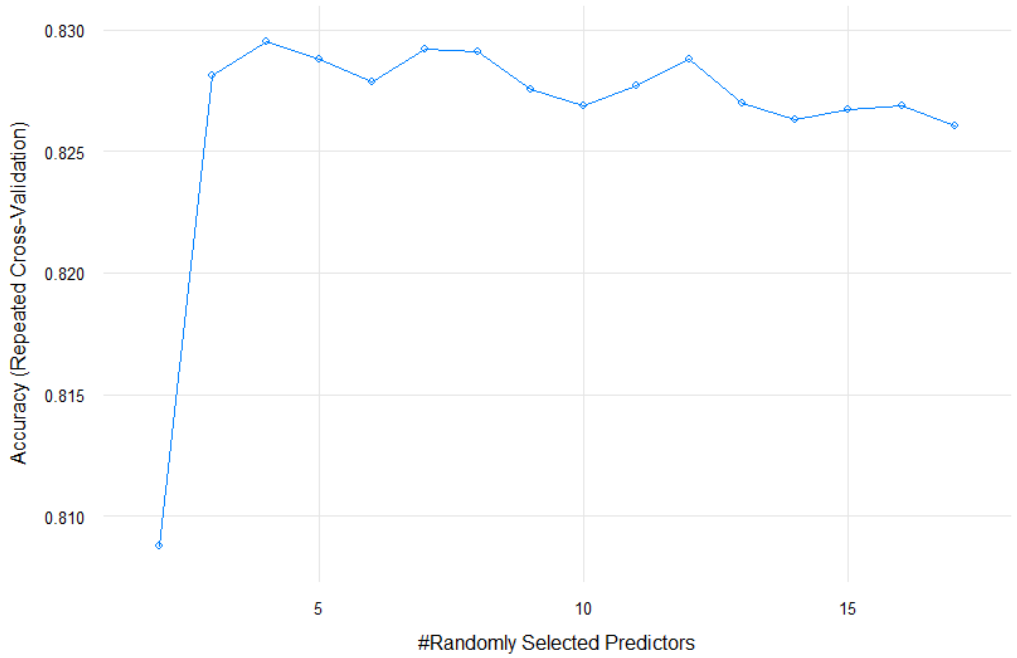

Figura 6 *Representación gráfica del número de variables como candidatas para cada división por el método de grilla search*

Por último, la optimización del parámetro mediante la función 'tuneR' del paquete *caret* generó un valor de 3 (ver Figura 7).

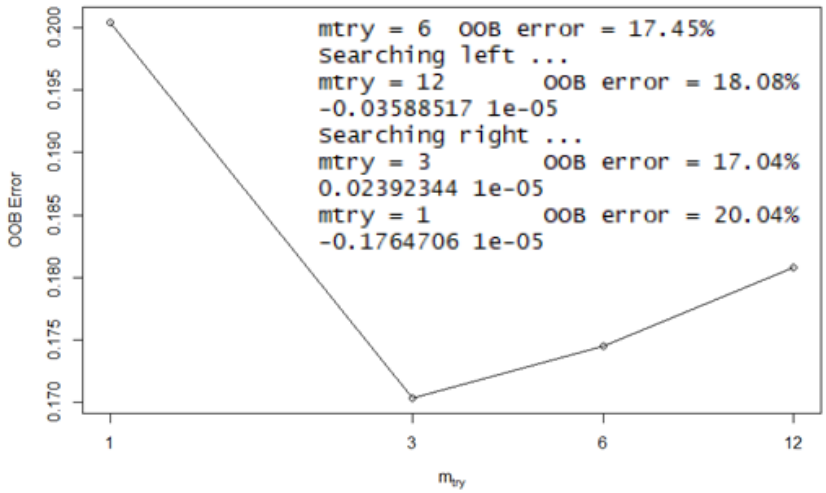

Figura 7 *Representación gráfica del número de variables como candidatas para cada división por el método de tuneRF*

Como el valor de error de out-of-bag obtenido en la última opción es el menor y adicionalmente pertenece al mismo paquete con que se realiza el método de *random forest* se optó por el valor de 3 para la cantidad de variables candidatas para cada división de los árboles.

A partir del nuevo *modelo 2* se observó que en general para una modelización de más de 200 árboles no mejora la tasa de error estimada de out-of-bag (ver Figura 8) por lo que se optará por este número de árboles para modelizar.

```
Call:
 randomForest(formula = respuesta ~ ., data = datos_train_rf,      importance = TRUE, mtry = 3)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 16.83%
Confusion matrix:
     No  Si class.error
No 1167 244   0.1729270
Si  159 825   0.1615854
```

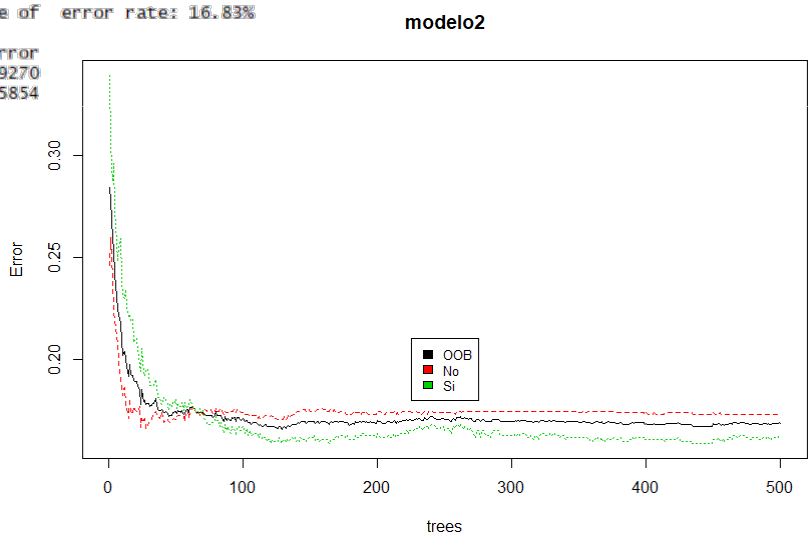

Figura 8 *Tasa de error estimada de out-of-bag para el modelo 2 en función del número de árboles utilizado*

A partir de los parámetros óptimos definidos se calculó el modelo final con los siguientes resultados (ver Figura 9).

```
Call:
 randomForest(formula = respuesta ~ ., data = datos_train_rf,      importance = TRUE, mtry = 3, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 3

        OOB estimate of  error rate: 17.04%
Confusion matrix:
     No  Si class.error
No 1164 247   0.1750532
Si  161 823   0.1636179
```

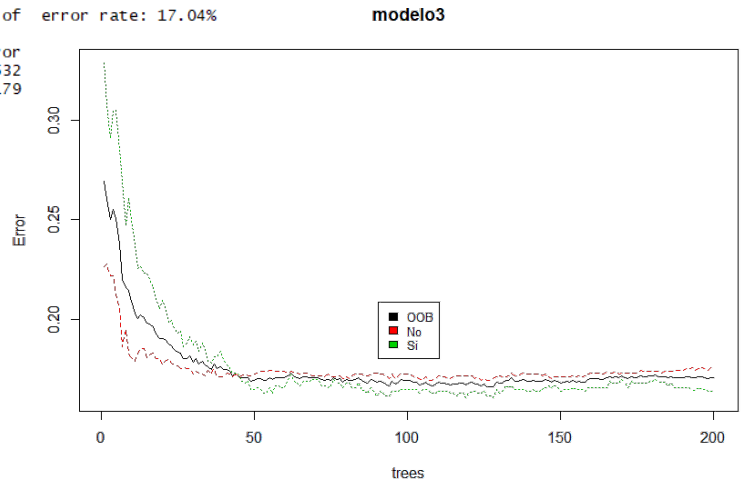

Figura 9 *Modelo 3 'óptimo'*

Para conocer qué variables fueron las más *importantes* en la construcción del modelo graficamos la disminución media de la precisión. Cuanto más disminuye la precisión del Random Forest debido a la exclusión (o permutación) de una sola variable, más importante se considera que esa variable es.
En el ranking de importancia de las variables predictoras del modelo se observa en primer lugar el tipo de movilidad del participante, su edad y sexo, atributos asociados a la forma de utilizar el vehículo y la ubicación geográfica donde ocurrió el siniestro (ver Figura 10).

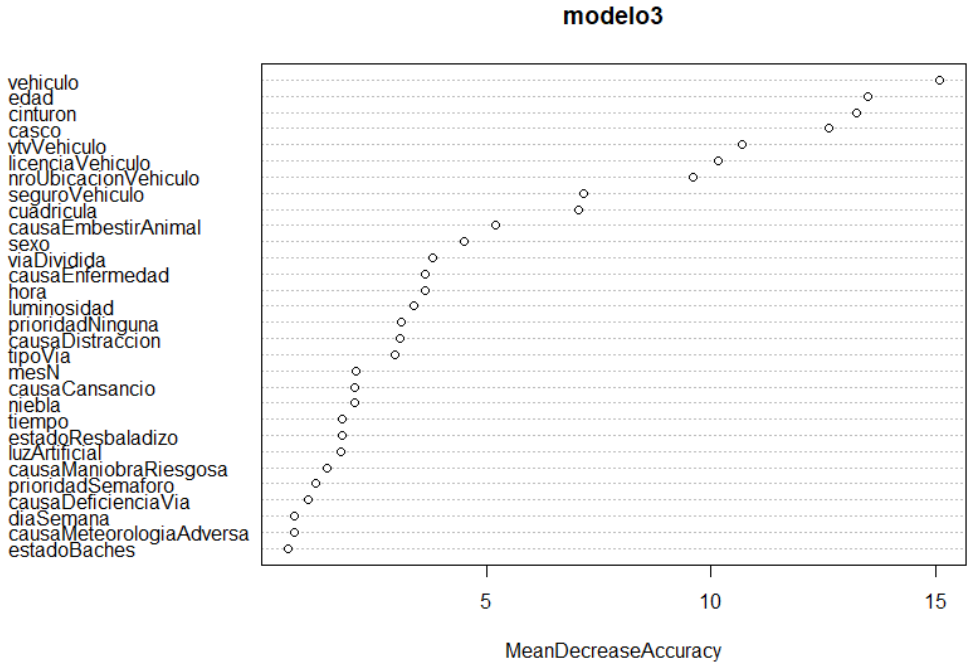

Figura 10 *Importancia de las variables del modelo óptimo*

A partir de la matriz de confusión para la predicción realizada con el modelo óptimo sobre los datos de testeo se obtiene que la exactitud predictiva predictiva fue de 81,17% (ver Tabla 4).

Tabla 4 Matriz de confusión para los datos de testeo con el modelo óptimo

		Real	
		No	***Sí***
Predicción	***No***	485	74
	Sí	119	347

El área bajo la curva ROC para los datos de testeo fue de 0.859 con un intervalo de confianza de 0.836 a 0.883 (ver Figura 11).

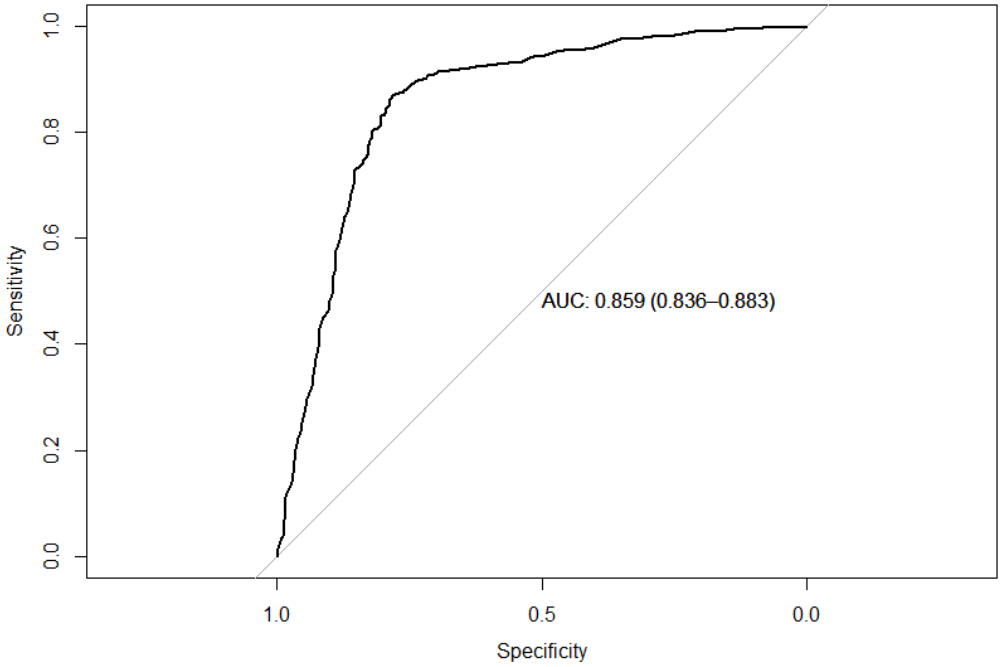

Figura 11 *Área bajo la curva ROC e intervalo de confianza*

La dependencia parcial de las principales variables predictoras se pueden observar en la siguiente gráfica (ver Figura 12).

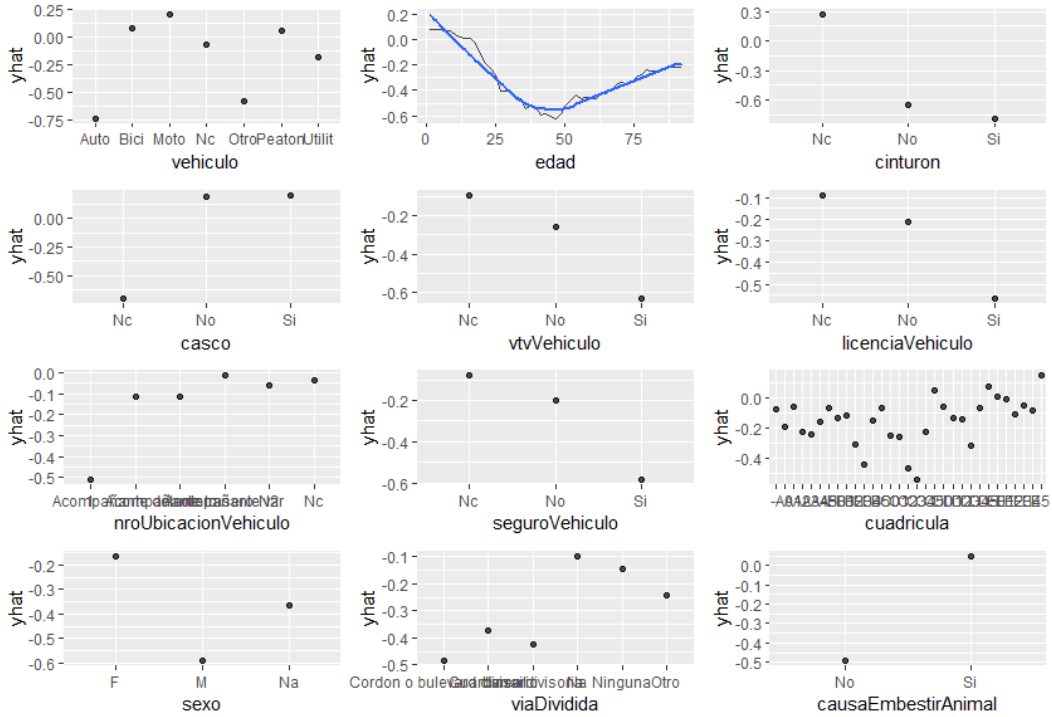

Figura 12 *Dependencia parcial de las 12 variables predictoras con mayor importancia*

4. CONCLUSIONES.

A partir de los datos relevados en el Formulario Estadístico Único es posible predecir la variable *hospitalizado* con una exactitud del 81% en áreas urbanas mediante un modelo de *random forest*. Las variables predictoras más importantes están relacionadas con:

- el tipo de movilidad: vehículo
- caracterización del participante: edad, sexo, ubicación dentro del vehículo
- utilización de los elementos de seguridad vial: cinturón, casco.
- aptitud del conductor y el vehículo: licencia, seguro, VTV.
- zona de la ciudad: cuadrícula

Se verifica que de las cinco variables con mayor peso, 3 de ellas están asociadas directamente al factor humano, en concordancia con los estudios previos mencionados sobre el tema.

También surge la causa aparente de resultar hospitalizado por evitar embestir un animal en la vía pública.
A partir de estas variables es posible generar acciones de prevención a través de los medios de comunicación locales, cursos de capacitación y operativos de los agentes de tránsito.
Los próximos pasos serán generar una aplicación en entorno web que permita a los agentes de tránsito discriminar las variables predictoras por cuadrícula, día de la semana y hora, de manera que los controles de sensibilización sobre seguridad vial sean más eficientes.

5. REFERENCIAS.

[1] Ministerio del Interior y Transporte. (2016). "Disposición 456/2014 AGENCIA NACIONAL DE SEGURIDAD VIAL". Recuperado 9 de septiembre de 2018, de *http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=259461*.

[2] Hernández Orallo, Ramirez Quintana y Ferri Ramirez. (2004). *Introducción a la Minería de Datos*. Editorial Pearson Prentice Hall. España. ISBN 84-205-4091-9.

[3] SAS Institute: Data Mining and the Case for Sampling. (1998). *http://nas.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf*. Recuperado el 17 de agosto de 2018.

[4] Chong M., Abraham A., Paprzycki M. (2004). "Traffic Accident Data Mining Using Machine Learning Paradigms", *Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04)*, Hungary, ISBN 9637154302, pp. 415- 420.

[5] Rocca G., Castillo-Cara M., Levano R., Herrera J. & Orozco-Barbosa L. (2016). *Citizen security using machine learning algorithms through open data*. 2016 8th IEEE Latin-American Conference on Communications (LATINCOM) (pp. 1-6). https://doi.org/10.1109/LATINCOM.2016.7811562

[6] Shanthi, S., Ramani R. (2011). "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms", *International Journal of Computer Applications (0975 – 8887) Volume 35– No.12, December 2011, pp 30-37.*

[7] S. Krishnaveni, M. Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 23– No.7, June 2011.

[8] R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/*.

[9] Haddon, W., Jr. (1968). "The changing approach to the epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based". *Am J Public Health 58: 1431-1438*.

[10] Kuhn, M., and Johnson, K. (2013). "Applied Predictive Modeling". Springer.

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

[12] Fawcett, T. (2006). "An introduction to ROC analysis", *Pattern Recognition Letters, 27, 861-874*.

[13] Brandon M. Greenwell (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R. Journal, 9(1), 421--436*. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

[14] CESVI, "Reconstrucción de Accidentes de Tránsito realizados por CESVI - Histórico 2004 - 2dosem_2016.pdf". (s.f.). Recuperado de https://home.cesvi.com.ar/WebSitesFiles/CesviArgentina/RAT%20-%20Historico%202004%20-%202do%20sem_2016.pdf

[15] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). "caret: Classification and Regression Training". *R package version 6.0-80*. https://github.com/topepo/caret/.

Agradecimientos

Los autores de este trabajo desean agradecer a la Dirección de Protección Ciudadana de la Municipalidad de Trenque Lauquen por su colaboración en el registro de los siniestros viales desde el año 2012 a la fecha.