UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CCSS.

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA

Identification of Main Factors and Variables Describing the Quantity and Distribution of Fatal Vehicular Accidents in Metropolitan City of Lima Using data Mining Techniques: Random Forest, Boosting, Decision Trees.

TESIS

Para obtener el título profesional de Ingeniero Estadístico

Elaborado Por:

Sindia Sherly Tarazona Tocto

LIMA – PERÚ

2016

Dedicatoria

A mis padres Emerson A. Tarazona Isidro y Ruth Tocto Miraval, por su apoyo y confianza en mí a lo largo de mi vida; a mi hermano Miguel A. Tarazona Tocto por ser mi motivo de superación y a mi tía Gloria Aponte por su apoyo incondicional en mi preparación para ingresar a la universidad.

Agradecimientos

A Dios, por permitirme estudiar esta hermosa carrera

Al profesor Richard Fernández por su empeño y exigencia en el curso de Taller de Tesis

A mis profesores de Pregrado por los conocimientos brindados a lo largo de mi vida universitaria

A mis amigos que colaboraron con sugerencias para el desarrollo de la presente investigación

Al comité revisor, por sus valiosas aportaciones y colaboración incondicional en este trabajo.

Y un agradecimiento especial a mis compañeros del código 2012 I, 2011 con quienes compartí en su mayoría clases y sin su colaboración no estaría actualmente terminado la carrera

ÍNDICE

RESUMEN

INTRODUCCIÓN

RESUMEN

Todos los años, más de 1,2 millones de personas fallecen en el mundo en accidentes de tránsito según el Informe sobre la situación mundial de la seguridad vial 2015 de la Organización Mundial de la Salud; así mismo, un informe de la Comunidad Andina de Naciones (CAN) publicada en el año 2013, ubicaba al Perú como el país con la más alta tasa de accidentes de tránsito por cada 100 mil vehículos, seguido por Bolivia, Colombia y Ecuador.

En el 2014 se realizó uno de los últimos Censos Nacionales de Comisarías, el cual contiene una sección sobre Accidentes de Tránsito. Como resultado del censo se pudo conocer que en Lima Metropolitana ocurren el 45% del Total de accidentes de tránsito registrados en el país.

Abordar el problema de los accidentes de tránsito fatales es un estudio complejo porque intervienen múltiples factores, pero a la vez interesante. La presente investigación es una aplicación en el campo del estudio sobre la siniestralidad vial. El estudio partió del problema ¿Cuáles son los principales factores que influyen en los accidentes de tránsito fatales en Lima Metropolitana, mediante los resultados de la importancia de variables de los modelos de Minería de Datos: Random Forest, Boosting, y Árbol de Decisiones CART?

Para responder el problema de investigación se recurrió al uso de Modelos de minería de datos como Boosting, Random Forest y Árbol de decisiones CART, cuyos resultados arrojaron una tabla de importancia de variables a partir de los cuales se logró identificar que el Tipo de vehículo usado (Camioneta rural(combi), mototaxi), Tipo de vía de tránsito(carretera o avenida), Invasión del carril y Desacato a la señal de tránsito por parte del conductor, son los principales factores que influyen en el resultado fatal de los accidentes de tránsito en Lima Metropolitana.

Palabras Clave

Accidente de tránsito fatal, sensibilidad, Gini, Validación cruzada, Boosting, Random forest, Árbol de Decisiones CART, Smooth

ABSTRACT

Every year, more than 1.2 million people die in the world in traffic accidents according to the World Health Report 2015; Likewise, a report from the Andean Community of Nations (CAN) published in 2013 placed Peru as the country with the highest traffic accident rate per 100,000 vehicles, followed by Bolivia, Colombia and Ecuador.

In 2014, one of the last National Census of Commissaries was held, which contains a section on Traffic Accidents. As a result of the census it was possible to know that 45% of the total traffic accidents in the country occur in Lima Metropolitan.

Addressing the problem of fatal traffic accidents is a complex study because there are many factors involved, but at the same time interesting. The present investigation is an application in the field of the study on road accidents. The study started from the problem What are the main factors that influence the fatal traffic accidents in Metropolitan Lima, through the importance of the variables of the Data Mining models: Random Forest, Boosting, and CART Decision Tree?

To answer the research problem we used the use of Data Mining Models such as Boosting, Random Forest and CART Decision Tree, whose results showed a table of importance of variables from which it was possible to identify that the type of vehicle used (Road or avenue), Invasion of the lane and disrespect to the sign of transit by the driver, are the main factors that influence in the fatal result of the accidents of Transit in Metropolitan Lima.

Keywords

Fatal Traffic Accident, Sensitivity, Gini, Cross Validation, Boosting, Random Forest, Decision Tree CART, Smooth

INTRODUCCIÓN

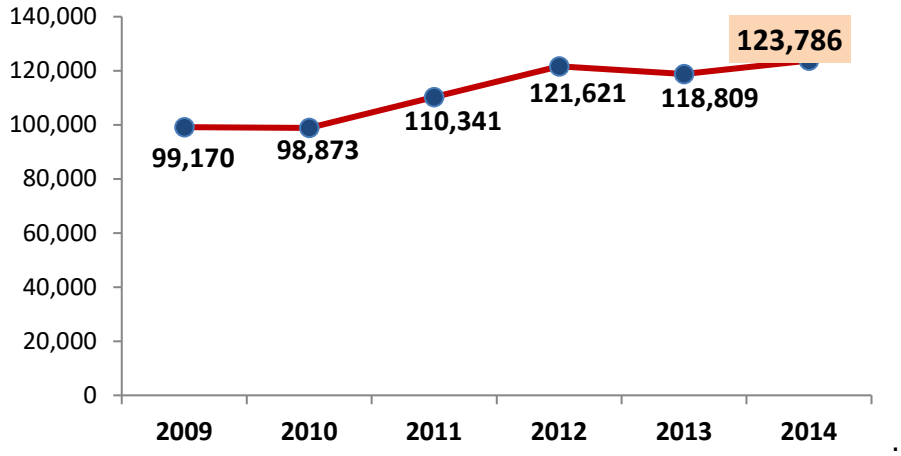
La siniestralidad en las carreteras ha sido, desde la generalización del uso de vehículos a motor, una de las principales causas de muerte en el mundo y por ello es foco de gran preocupación para la sociedad y sus autoridades.

De acuerdo al portal Ojo Científico, las estadísticas y las investigaciones de la Organización Mundial de la Salud (OMS), desarrolladas entre los años 2000 y 2011, y publicadas en julio de 2013, indicaron que los accidentes automovilísticos son la segunda causa principal de muerte entre la humanidad, ocasionando la muerte de aproximadamente, 1.15 millones de personas al año, lo cual quiere decir que, al día, entre 3 mil 500 y 3 mil 600 personas mueren en un accidente de tránsito.

El Ministerio de Salud (MINSA) de nuestro país en el 2014, mostró estadísticas de las principales casusas de mortalidad; los accidentes de transporte se ubican en el puerto 13, de un total de 25 mencionadas.

Sin duda los accidentes de tránsito son un tema importante para el Estado Peruano, motivo por el cual desde el 2010 se han venido desarrollando anualmente Censos Nacionales a Comisarías, de cuyos resultados se observó el siguiente gráfico que muestra la evolución del total de accidentes de tránsito registrado a nivel nacional desde el 2009 hasta el 2014.

Número total de accidentes de tránsito registrados, periodo 2013-2014



FUENTE: Informe sobre la situación mundial de La Seguridad Vial 2015 - OMS
ELABORACIÓN: Propia

Del gráfico mostrado se observa una tendencia ascendente del total de accidentes de tránsito registrados en el Perú, pasando de 99 mil 170 registros en el 2009 a 123 mil 786 en el 2014, aumentando en un 25% (24,616 casos de accidentes de tránsito más).

Por otro lado, como resultado de los censos realizados, también se lograron conocer el número de accidentes de tránsito fatales, que en su mayoría representaron entre el 1% y 2% del total.

Según el informe del Censo Nacional de Comisarías 2015, con registros de accidentes del 2014; en Lima Metropolitana ocurren el 45% del total de accidentes de tránsito y de estos, el 1% son accidentes de tránsito fatales, lo que significa que como resultado del accidente hubo por lo menos una víctima mortal.

El principal interés de la presente investigación es identificar los factores que influyen en los accidentes de tránsito fatales a nivel de Lima Metropolitana, desde un enfoque estadístico con la aplicación de técnicas de Minería de datos. Para lograr el objetivo, inicialmente se partió del problema de desbalanceo en la base de datos, y el cómo abordar este tipo de datos fue el principal interés a un inicio.

La presente investigación se divide en V capítulos. En el Capítulo I se muestra los antecedentes que sirvieron como base para abordar con éxito

los problemas de investigación y proporcionaron distintas alternativas de enfoque de solución.

En el Capítulo II, se describen los problemas de investigación tanto general como específicos, los Objetivos de investigación, las Hipótesis planteadas y la Justificación del estudio.

En el Capítulo III, se muestra de manera detallada las definiciones teóricas de los modelos desarrollados: Random Forest, Boosting y Árbol de decisiones CART; así como, definiciones previas y términos básicos que sirvieron para su entendimiento profundidad.

En el Capítulo IV, se muestra la Metodología seguida para el logro de los objetivos, donde se describe la población en estudio, las variables y el diseño de muestreo.

Más adelante, en el Capítulo V, se muestran los resultados, a nivel descriptivo, bivariado, y los resultados de los modelos desarrollados, así como, la comparación de sus principales indicadores de desempeño y sus gráficos de importancia de las variables.

Finalmente, se presentan las conclusiones y recomendaciones, donde se demuestra las hipótesis estadísticas planteadas previamente, por ende cumpliendo con los objetivos de la presente investigación.

CAPÍTULO I

ANTECEDENTES

1.1 Investigaciones

Se buscaron en primera instancia antecedentes nacionales, no se encontró, por lo que solo se mencionan antecedentes internacionales:

a) Randa, M., López G y Garach L. en el 2015, en su investigación denominada “Bayes classifiers for imbalanced traffic accidents data sets”, realizaron una investigación con el objetivo de Realizar un modelo de predicción de gravedad de lesiones en un accidente de tránsito mediante clasificadores de Bayesianos; las variables que usó fueron: número de vehículos implicados, condiciones de la superficie, velocidad, patrón de accidente, número de direcciones y obtuvo como resultado que el uso de los conjuntos de datos balanceados, usando la técnica de balanceo: oversampling con redes bayesianas mejoraron la clasificación de la gravedad de lesiones en un accidente de tráfico.

b) Klaus M. en el 2013, publicó su investigación titulada “*PREDICTION OF ROAD ACCIDENTS: COMPARISON OF TWO BAYESIAN METHODS”,* con el objetivo de Establecer una metodología que permita desarrollar un modelo de predicción de la ocurrencia de un accidente, la metodología que usaron fue comparar dos Método Bayesianos: Modelos de Redes bayesianas probabilísticas y Método Empírico Bayesiano.

c) Bahar D., Arenas, B., Mira J. en el 2015, publicaron un estudio titulado “Metodología desarrollada para la selección de predictores significativos que explican fatales accidentes de carretera en España”, con el objetivo de proponer una metodología para la selección de los predictores significativos que explican los accidentes mortales en carretera, la metodología utilizada fue proponer un método basado en una red neuronal, esta metodología sigue un razonamiento similar al de Lasso, en el sentido de que los predictores son seleccionados teniendo en cuenta su importancia individual. La idea es construir modelos NN con dos variables independientes (TIM). Todos los modelos se estiman utilizando la cadena de Markov método de Monte Carlos.

d) Villarino, M (2015) en su trabajo de fin de maestría en Minería de datos en la Universidad Complutense de Madrid, cuyo título de su investigación fue *Metodología de minería de datos para el estudio de tablas de siniestralidad vial.* Se propuso el siguiente objetivo general: 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. El investigador realizó una serie de modelos de minería de datos entre ellos Random Forest, Gradient Boosting y Redes Neuronales, en la clasificación de los fallecidos en las distintas subpoblaciones de víctimas. El investigador partió de una base de datos desbalanceada 1%-99%, el cual balanceó usando la metodología Smooth y posteriormente construyó lo modelos obteniendo buenos indicadores de desempeño.

CAPÍTULO II

REALIDAD PROBLEMÁTICA

2.1 Descripción del contexto del problema

Según el Informe sobre la situación mundial de la seguridad vial 2015 publicada por la Organización Mundial de la Salud (OMS); todos los años, más de 1,2 millones de personas fallecen en accidentes de tránsito y entre 20 y 50 millones padecen traumatismos no mortales. Dicho de otra forma, cada 25 segundos en el mundo fallece una persona y otras 40 sufren traumatismo, como consecuencias de los accidentes de tránsito.

A nivel mundial los accidentes de tránsito son uno de las tres principales causas de mortalidad en personas cuya edad se encuentra en el rango de 5 a 44 años, compitiendo en este ranking nada menos que con enfermedades como el VIH/SIDA, y con la tuberculosis.

CUADRO Nº 2.1

PRINCIPALES CAUSAS DE MUERTE EN TRES GRUPOS ETARIOS EN EL MUNDO

N°	5-14 años	15-29 años	30-44 años
1	Infecciones de las vías respiratorias	Traumatismos causados por el tránsito	Infección por el VIH/SIDA

2	Traumatismos causados por el tránsito	Infección por el VIH/SIDA	Tuberculosis
3	Malaria	Tuberculosis	Traumatismos causados por el tránsito

FUENTE: Informe sobre la situación mundial de La Seguridad Vial 2015 - OMS
ELABORACIÓN: Propia

Esta realidad no es ajena al Perú, cada día observamos en la televisión múltiples noticias sobre accidentes de tránsito, muchos de ellos fatales. Un informe de la Comunidad Andina de Naciones (CAN) publicada en el año 2013, ubicaba al Perú como el país con la más alta tasa de accidentes de tránsito por cada 100 mil vehículos, seguido por Bolivia, Colombia y Ecuador.

Para muchos investigadores este problema es interesante, pero a la vez complejo porque es ocasionada a causa de múltiples factores. Para conocer estos factores hace necesario primero realizar una medición y recopilación de información sobre los accidentes de tránsito.

En el Perú, cada vez cobra mayor importancia el hacer frente a esta problemática, motivo por el cual varias instituciones del Estado vienen trabajando conjuntamente en la generación de estadísticas que permitan mostrar datos sustanciales relacionados a los accidentes de tránsito. Un ejemplo de ello es la ejecución del IV Censo Nacional de Comisarías 2015 a cargo del Instituto Nacional de Estadística e Informática, en coordinación con el Ministerio de Economía y Finanzas y el Ministerio del Interior, el cual se ejecutó en el marco de los programas estratégicos: Seguridad Ciudadana y Accidentes de Tránsito, incorporados por el Ministerio de Economía y Finanzas dentro de la política de aplicación del Presupuesto por Resultados (PpR).

Los datos generados a partir del IV Censo Nacional de Comisarías 2015, contienen información de los accidentes de tránsito producidos a nivel nacional, los cuales fueron recopilados a partir de los registros en los

libros de ocurrencias o Sistema de Denuncias Policiales (SIDPOL) de las comisarías censadas.

En base a los resultados del Censo en mención, se conoció que Lima es la Región con mayor cantidad de víctimas de accidentes de tránsito fatales registrados (13% del total); motivo por el cual el presente estudio se basó en estudiar los datos asociados a los accidentes de tránsito producidos en Lima Metropolitana durante el 2014.

2.2 Formulación del problema de investigación

Los accidentes de tránsito son hechos de nuestro día a día, y un gran porcentaje de ellas tiene como resultado víctimas mortales; es por ello que la presente investigación buscará estudiar los factores asociados a los accidentes de tránsito fatales, para lo cual se partirá de la formulación de las siguientes interrogantes:

2.2.1 Problema general

¿Cuáles son los principales factores que influyen en los accidentes de tránsito fatales en Lima Metropolitana, mediante los resultados de importancia de variables de los modelos de Minería de Datos: Random Forest, Boosting, y Árbol de Decisiones CART?

2.2.2 Problemas específicos

- ¿Los modelos de minería de datos: Random Forest, Boosting, y Árbol de decisiones, presentan buenos indicadores de desempeño?

- ¿Los modelos de minería de datos: Random Forest, Boosting y Árbol de Decisiones presentan estabilidad en las tasas de error, mediante la metodología validación cruzada k-fold?

- ¿Cuál es la importancia de las variables de los modelos de Minería de datos: Random Forest, Boosting, y Árbol de decisiones?

2.3 Formulación de los objetivos de la Investigación

2.3.1 Objetivo general

Identificar los principales factores que influyen en los accidentes de tránsito fatales en Lima Metropolitana, mediante los resultados de importancia de variables de los modelos de Minería de Datos: Random Forest, Boosting, y Árbol de decisiones CART

2.3.2 Objetivos específicos

- Determinar los indicadores de desempeño: Sensibilidad, Especificidad, Índice de Gini y Distancia Euclídea, de los modelos de minería de datos: Random Forest, Boosting, y Árbol de decisiones.
- Determinar la estabilidad de las tasas de error de los modelos de minería de datos: Random Forest, Boosting y Árbol de Decisiones, mediante la metodología validación cruzada k-fold, con k=10.
- Determinar la importancia de las variables de los modelos de Minería de datos: Random Forest, Boosting, y Árbol de decisiones.

2.4 Hipótesis de la investigación

2.4.1 Hipótesis general

Los principales factores que influyen en los accidentes de tránsito fatales en Lima Metropolitana, mediante la importancia de las variables de los modelos de clasificación usados son: Tipo de vehículo usado (Camioneta rural(combi), mototaxi), Tipo de vía de tránsito (carretera o avenida) y Desacato a la señal de tránsito por parte del conductor.

2.4.2 Hipótesis específicos

- Los valores de los indicadores de desempeño de los modelos usados son: Gini mayor a 50% para todos los modelos y mayor a 70% para el modelo Boosting, Sensibilidad mayor a 55% para todos los modelos, y Especificidad mayor a 60% para todos los modelos.
- Las tasas de error de los modelos construídos, obtenidos mediante la metodología validación cruzada k-fold, con k=10, son menores a 9% para el Modelo Boosting, menores al 15% para el Modelo Random Forest y menores a 25% para el Modelo Árbol de decisiones CART.
- Las variables(top 3) que figuran en la tabla de importancia de variables del modelo boosting son: Tipo de vía de ocurrencia de del accidente(carretera), tipo de vehículo mayor involucrado(camión-combi) y desacato a la señal de tránsito por el conductor; para el modelo Random forest son: Tipo de vehículo mayor involucrado(camión-combi), Tipo de transporte(público) y Accidente de tránsito por atropello; finalmente para Árbol de decisiones son: Tipo de vía de ocurrencia de del accidente(carretera), tipo de vehículo menor involucrado(mototaxi) y por invasión en el carril contrario.

2.5 Justificación

La presente investigación es de vital importancia ya que mostrará las principales variables que perfilan a los accidentes de tránsito fatales de Lima Metropolitana ocurridos en el 2015.

Los beneficiarios de esta investigación serán las autoridades del Ministerio de Transporte y Comunicaciones, y La Municipalidad Metropolitana de Lima quienes son los responsables de plantear políticas efectivas de seguridad vial; así mismo, en el largo plazo tendrá un impacto positivo en la población de Lima Metropolitana.

Por otro lado, la presente investigación también beneficiará a la comunidad estadística y a todo amante de la estadística, ya que se mostrará diversas variantes del Modelo de Regresión Logística (MRL): MRL con punto de corte óptimo y MRL a partir de muestras balanceadas.

Cabe señalar que en el Perú no existen investigaciones con el mismo fin, recién se está empezando a recolectar datos mediantes censos que a lo mucho son analizadas mediante análisis descriptivo de datos; sin embargo, sí existen en países cercanos como Chile y en países Europeos.

2.6 MATRIZ DE CONSISTENCIA

Formulación del Problema	**Formulación de los objetivos**	**Formulación de la Hipótesis**	**Variables**
Problema General ¿Qué modelo es más adecuado comparando el Modelo Logit con PCO y Modelo Logit con datos balanceados para identificar las principales variables que perfilan a los accidentes de tránsito fatales de Lima	**Objetivo General** Elegir el modelo más adecuado comparando el Modelo Logit con PCO y Modelo Logit con datos balanceados para identificar las principales variables que perfilan a los accidentes de tránsito fatales de Lima Metropolitana	**Hipótesis General** El modelo más adecuado comparando el Modelo Logit con PCO y Modelo Logit con datos balanceados para identificar las principales variables que perfilan a los accidentes de tránsito fatales de Lima Metropolitana según el Censo 2015, resultó el	**Variables Dependiente** Y= Accidente de tránsito fatal Donde: Y=1: Fatal Y=0: No fatal

Metropolitana según el Censo 2015?	según el Censo 2015	Modelo Logit con Punto de Corte 0.4, esto en base a los siguientes indicadores: índice de youden, sensibilidad, especificidad, verdadero positivo, falso negativo e índice de Gini	
Problemas Específicos -¿Qué Modelo comparando el Modelo Logit con PCO y Modelo Logit con datos balanceados presenta los menores porcentajes de error según la técnica de validación cruzada? -¿Cuáles son las principales variables que influyen en la fatalidad de los accidentes de tránsito de Lima Metropolitana, según el Censo 2015? -¿Cuál es el perfil de los accidentes de tránsito de acuerdo a su propensión de fatalidad alta y	**Objetivos Específicos** -Determinar el modelo comparando el Modelo Logit con PCO y Modelo Logit con datos balanceados que presente los menores porcentajes de error según la técnica de validación cruzada. -Determinar las principales variables que influyen en la fatalidad de los accidentes de tránsito de Lima Metropolitana, según el Censo 2015 -Determinar el perfil de los accidentes de tránsito de acuerdo a su propensión de	**Hipótesis Específicas** -El Modelo Logit con PCO es el que presenta los menores porcentajes de error según la técnica de validación cruzada, esto debido al menor valor de error de predicción en comparación con el otro modelo. -Las principales variables que influyen en la fatalidad de los accidentes de tránsito de Lima Metropolitana, según el Censo 2015 son las características del conductor y las características de la carretera. -Los accidentes de tránsito con alta propensión de fatalidad son	**Variables Independientes** - Fecha de ocurrencia del accidente de tránsito - Tiempo de ocurrencia del accidente de tránsito - Lugar de ocurrencia - Tipo de vehículo mayor involucrado. - Tipo de transporte - Número de fallecidos - Número de heridos - Número de ilesos - Factores vinculados al accidente de tránsito

media?	fatalidad alta y media.	aquellos que……. Los accidentes de tránsito con propensión media de fatalidad son aquellos que……. (dependerá de los resultados)	

CAPÍTULO III

MARCO TEÓRICO

3.1 Técnicas Previas

3.1.1 Correlación Chi cuadrado

Una medida muy extendida para medir la dependencia e independencia, es el estadístico Chi-cuadrado, que da una medida de la diferencia entre las frecuencias observadas en la tabla y las "*frecuencias esperadas en caso de independencia*". Recordamos el cálculo de dichas frecuencias esperadas e_{ij}:

$$e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n}$$

Con el estadístico Chi-cuadrado se obtiene una medida de diferencia entre las frecuencias esperadas y las frecuencias observadas. El estadístico se calcula en la forma siguiente:

$$\chi^2_{\exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

Observamos las siguientes propiedades de este estadístico:

- Si todas las frecuencias observadas son iguales a la correspondiente frecuencia esperada, $f_{i,j} = e_{i,j}$ entonces $\chi^2_{\exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$ = $\sum_i \sum_j \frac{(f_{ij} - f_{ij})^2}{e_{ij}} = 0$.

- Esto ocurre sólo cuando las dos variables de la tabla son independientes; Por tanto, si hay independencia entre las dos variables de la tabla, $\chi^2_{\exp} = 0$

- Cuanto mayor sea la diferencia entre las frecuencias observadas y esperadas en la tabla, el valor de Chi cuadrado será mayor. Es decir, a mayor intensidad de la asociación entre las variables, Chi-cuadrado será mayor.

- El valor de Chi-cuadrado siempre es positivo o cero (pues es suma de números positivos, ya que los denominadores de la suma son todos positivos al ser suma de números elevados al cuadrado.

- En general, a mayor número de sumandos, se obtendrá un valor mayor.

Los *grados de libertad* de un estadístico calculado sobre un conjunto datos se refieren al número de cantidades independientes que se necesitan en su cálculo, menos el número de restricciones que ligan a las observaciones y el estadístico. El número de grados de libertad del estadístico Chi-cuadrado se calcula de la siguiente forma:

- Se calcula, en primer lugar el número de sumandos, es decir *m* x *n*, siendo *n* y *m* el número de filas y número de columnas en la tabla.

- A esta cantidad se debe restar el número de restricciones impuestas a las frecuencias observadas. Observamos que podemos cambiar todas las frecuencias de la tabla sin cambiar los totales por filas y columnas, excepto los datos en la última fila y la última columna de la tabla, pues una vez que fijemos todos los valores excepto estos, quedan

automáticamente fijados. Por tanto, si la tabla tiene *m* filas y *n* columnas, *el número de grados de libertad* es *(m-1) x (n-1).* Expresamos esta dependencia en la siguiente forma:

$$\chi^2_{\exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \rightarrow \chi^2_{(n-1)(m-1)}$$

3.1.2 Prueba de homogeneidad de varianzas

Uno de los supuestos que más se requieren en aplicaciones estadísticas populares, tales como el análisis de varianza, el análisis de regresión, etc., es el de la homogeneidad de varianzas. Este supuesto es crucial para garantizar la calidad de los procedimientos estadísticos utilizados tanto en pruebas de hipótesis como en la construcción de intervalos de confianza. (Correa. J [1])

Existen muchas pruebas para verificar si el supuesto de homogeneidad es plausible o no, pero, dada la complejidad del problema, no es posible realizar estudios comparativos entre ellas que sean exhaustivos, ni de su comportamiento para muestras pequeñas, ya que muchas de ellas son de carácterasintótico. En este trabajo estudiamos el nivel real de significancia, el cual es la verdadera probabilidad de rechazar la hipótesis nula cuando es cierta y que en pruebas no exactas es diferente del nivel nominal, o teórico, de significancia, determinado por el usuario, usualmente a niveles del 5 % u otros valores pequeños. Además, se estudia la potencia de las pruebas bajo algunas alternativas abajo enunciadas. En esta simulación se quiere comparar la prueba de Bartlett, la prueba de Levene (Brown & Forsythe 1974), la prueba de Hartley (1950), la prueba de Cochran (1941), la prueba de Fligner & Killeen (1976), la prueba basada en la teoría de la información, la prueba de Layard y algunas de sus variaciones, por medio de la potencia que cada prueba tenga con respecto a diferentes hipótesis alternas. La idea es

[1]Estudio de potencia de pruebas de homogeneidad de varianza- Revista Colombiana de Estadística

saber cuál es la mejor prueba y bajo qué condiciones de número de muestras y tamaños se puede utilizar.

3.1.2.1 Notación

La notación utilizada en el presente artículo será la siguiente:
k = Número de muestras
ni = Tamaño de la i-ésima muestra
sigma2= Varianza estimada para la i-ésima población a partir de una muestra de tamaño ni, N = n1 + n2 + · · · + nk
s 2 = Varianza total estimada
La hipótesis que se quiere probar es:
Ho : σ 2 1 = σ 2 2 = · · · = σ 2 k
Ha : σ 2 i 6= σ 2 j para por lo menos un par (i, j)

3.1.2.2 Prueba de Bartlett

Introducida por Bartlett en 1937, es una modificación del test de Neyman y Pearson para “corregir el sesgo” ; esta prueba es la que se utiliza con más frecuencia para probar la homogeneidad de las varianzas (Conover et al. 1981). En esta prueba los ni en cada tratamiento no necesitan ser iguales; sin embargo, se recomienda que los ni no sean menores que 3 y muchos de los ni deben ser mayores que 5.

El estadístico de prueba se define como:

$$U = \frac{1}{C}\left[(N-k)\ln(s^2) - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2\right]$$

Donde

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)$$

Cuando la hipótesis nula es cierta, el estadístico tiene distribución aproximadamente χ 2 con k − 1 grados de libertad; cuando el muestreo se realiza en poblaciones normales, la aproximación es buena para muestras bastante pequeñas (Layard 1973). No requiere que los tamaños de las

muestras sean iguales. Es muy sensible a alejamientos del supuesto de normalidad (Montgomery 2002, pág. 82). Si tenemos evidencia fuerte de que los datos vienen de hecho de una distribución normal, o casi normal, entonces la prueba de Bartlett tiene un buen desempeño.

3.1.2.2 Prueba de Levene

El estadístico de prueba de Levene se define como:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\overline{Z}_{i.} - \overline{Z}_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_{i.})^2}$$

Donde Zij puede tener una de las siguientes tres definiciones:

1. $Z_{ij} = |X_{ij} - \overline{X}_{i.}|$ donde $\overline{X}_{i.}$ es la media del i-ésimo subgrupo.
2. $Z_{ij} = |X_{ij} - \widetilde{X}_{i.}|$ donde $\widetilde{X}_{i.}$ es la mediana del i-ésimo subgrupo.
3. $Z_{ij} = |X_{ij} - \overline{X'}_{i.}|$ donde $\overline{X'}_{i.}$ la media recortada al 10 % del i-ésimo subgrupo.

$\overline{Z}_{..}$ es la media global de Z_{ij} y $\overline{Z}_{i.}$ es la media del i-ésimo subgrupo de los Z_{ij}.

La prueba de Levene rechaza la hipótesis de que las varianzas son iguales con un nivel de significancia α si W > Fα,k−1,N−k donde Fα,k−1,N−k es el valor crítico superior de la distribución F con k −1 grados de libertad en el numerador y N −k grados de libertad en el denominador a un nivel de significancia α. La prueba de Levene ofrece una alternativa más robusta que el procedimiento de Bartlett, ya que es poco sensible a la desviación de la normalidad. Eso significa que será menos probable que rechace una verdadera hipótesis de igualdad de varianzas sólo porque las distribuciones de las poblaciones muestreadas no son normales.

3.2 Técnicas a usar

3.2.1 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 árboles 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 árboles, 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.

3.2.1.1 Estimación del error con Random Forest

Se define la tasa de error out of bag (OOBi) de una observación xi como el error obtenido al ser clasificada por los árboles del bosque construidos sin su intervención.

- La estimación OOB del error es el promedio de todos los OOBi para todas las observaciones del conjunto de datos

- Es mejor estimador que el error aparente. Parecida a la estimación por validación cruzada

- La medida se puede extrapolar al problema de regresión describiéndola en términos del ECM

3.2.2 Boosting

Es Procedimiento que combina el output de muchos predictores para crear un “comité” de predictores de gran precisión

Incorpora de bagging la idea de la agregación. Sin embargo, el procedimiento de agregación es distinto ya que no todos los predictores del comité van a tener el mismo peso

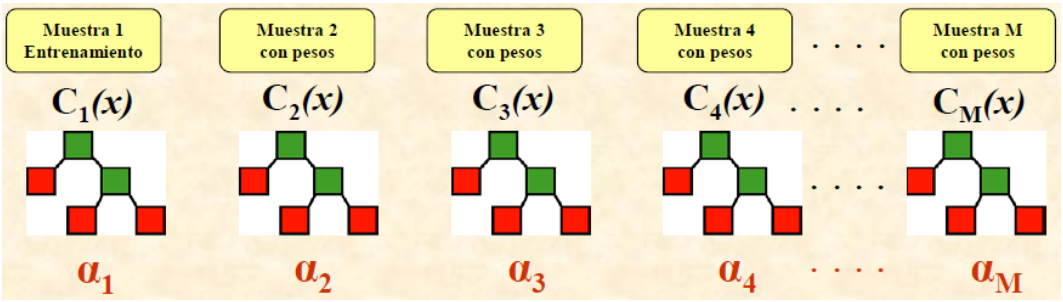
Por tanto, cada uno de los modelos de predicción del comité influyen de manera diferente en la predicción final.

Asimismo, el mecanismo de aprendizaje del Bossting asigna pesos a las observaciones enfocándose en las que son más difíciles de clasificar.

Por otro lado, AdaBoost.M1 es el algoritmo más popular, desarrollado por Freund (California) y Shapire (Princeton) en los años 90 para el problema de clasificación binaria

3.2.2.1 Algoritmo AdaBoost.M1

AdaBoost.M1 ajusta los pesos para ponderar la contribución de cada miembro del comité. La idea es simple. Tenemos las opiniones de un comité de expertos. Debemos considerarlas todas; sin embargo, tenemos que conceder mayor peso a los más “sabios” del comité:



El Predictor final será: sign(C(x)) con C(x) = α1× C1(x) + …+ αM × CM(x)

Algoritmo AdaBoost.M1

Paso 1. Inicializar pesos $w_i^{(1)} = 1/n : i = 1, \ldots, n$

Paso 2. Bucle: repetir para $m = 1$ hasta M

- 2.1 Ajustar un clasificador $C_m(x)$ a la muestra con los pesos $w_i^{(m)}$

- 2.2 Calcular el error con pesos $e_m = \frac{\sum_{i=1}^{n} w_i^{(m)} I(y_i \neq C_m(x_i))}{\sum_{i=1}^{n} w_i^{(m)}}$

- 2.3 Calcular los coeficientes $\alpha_m = \log \left| \frac{1 - e_m}{e_m} \right|$

- 2.4 Actualizar pesos $w_i^{(m+1)} = w_i^{(m)} \exp[\alpha_m \cdot I(y_i \neq C_m(x_i))]$

Paso 3. Clasificador AdaBoost: $C(x) = \sum_{m=1}^{M} \alpha_m C_m(x)$

Paso 4. Predicción AdaBoost: $signo(C(x))$

Donde M es un parámetro prefijado.

Tener en cuenta que los casos erróneamente clasificados en cada etapa tienen mayor peso en la siguiente etapa. Cada Cm del comité “se enfoca” en los fallos de los anteriores.

3.2.3 Árbol de decisiones

Los Árboles de Decisión se pueden utilizar para modelizar problemas de Clasificación Binaria (Fraude vs no fraude) o Multiclase (niveles de satisfacción: completamente, bastante, poco satisfecho, totalmente insatisfecho), así como problemas de Regresión: Pagos que realiza una compañía de seguros, Gasto mensual de los clientes de una cadena de supermercados, etc.

3.2.3.1 Segmentos con modelos de Árbol

Los segmentos de árbol se caracterizan por lo siguientes:

- La partición en cada nodo describe dos conjuntos disjuntos de la base de datos.

- El corte viene dado por una o varias condiciones en una de las variables explicativas
- El particionamiento es recursivo. Se detiene en los nodos terminales
- A cada nodo terminal se le asigna uno de los estados de la variable criterio Y
- Para cada nueva observación, el estado de la variable respuesta se predice por el estado del nodo terminal al que pertenece dicha observación

3.2.3.2 Algoritmos de segmentación

El criterio que determina el corte en cada nodo, el mecanismo de segmentación y el criterio de parada (que permite decidir si un nodo es terminal o no) han dado lugar a distintos algoritmos de segmentación. Dos algoritmos bastante extendidos son CHAID (Es el acrónimo de Chi-Squared Automatic Interaction Detector) y CART (Es el acrónimo de Classification and Regression Trees)

3.2.3.2.1 Algoritmo CHAID

El Algoritmo CHAID posee las siguientes características:

- Procede del ámbito de la Inteligencia artificial. Desarrollado por Kass a principios de los años 80
- Asume que las variables explicativas son categóricas u ordinales. Cuando no lo son, se discretizan
- Inicialmente se diseñó para el caso de variable respuesta Y categórica. Posteriormente se extendió a variables continuas
- Utiliza contrastes de la $\chi 2$ de Pearson y la F de Snedecor
- El corte en cada nodo es multi-vía

El criterio de corte de CHAID se caracteriza por los siguientes:

- CHAID considera todos los cortes posibles en todas las variables. Selecciona el corte que da el menor p-valor asociado a una medida de contraste estadístico

- Si la variable criterio es categórica la medida es la χ2 de Pearson. Si es continua la medida es la del test de la F
- La búsqueda de la variable y el corte óptimo se lleva a cabo en dos fases: merge (fusión de categorías) y split (selección de la variable de corte)

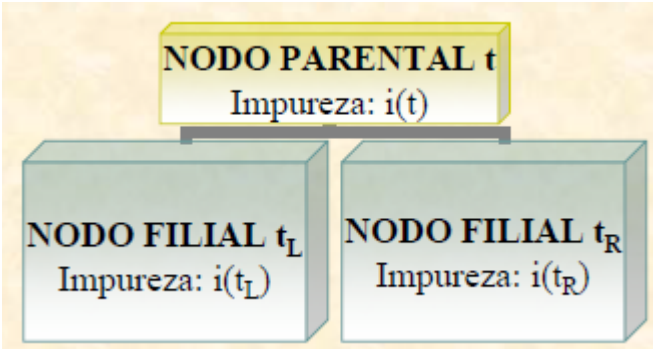
3.2.3.2.2 Algoritmo CART

El Algoritmo CART posee las siguientes características:

- Procede del ámbito de la Estadística. Desarrollado por matemáticos de la universidad de Berkeley y Stanford (Breiman, Friedman, Olshen y Stone) a mediados de los 80
- Trabaja con variables de todo tipo. No necesita discretizar las variables explicativas continuas
- El corte en cada nodo viene dado por reglas de tipo binario. Se pueden formular como preguntas: ¿Es Xk < a? ¿Pertenece Xk a un subconjunto E de estados?
- Da lugar a estructuras de árbol de mayor profundidad

A continuación se detallan puntos importantes a considerar al construir un modelo de árbol de decisiones CART

a) El criterio de corte de CART se caracteriza por los siguientes:
   - Se basa en la idea de impureza. CART selecciona el corte que conduce al mayor decrecimiento de la impureza. Así se consiguen descendientes homogéneos en la variable respuesta Y



Algunas medidas de impureza para el problema de Clasificación:

- La entropía
- El Índice de Gini
- El criterio de Twoing

*La medida de impureza para el problema de regresión es La agregación de las varianzas de todos los nodos temrinales.

b) El criterio de parada

La alternativa de CART es no parar; es decir:

- CART propone segmentar la base de datos hasta obtener una estructura de árbol lo más compleja posible
- Un nodo se declara como terminal sólo si su tamaño es inferior a un umbral preestablecido (normalmente muy pequeño)
- La complejidad de un árbol se mide por el número de nodos terminales
- A continuación, se poda la estructura de árbol maximal que se ha obtenido.

c) La poda de una rama

Una rama del nodo t de un árbol T está formada por él y todos sus descendientes. Podar la rama en t consiste en eliminar todos los descendientes del nodo t

El proceso de poda se apoya en la siguiente medida:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

- Combina el riesgo o coste de predicción y la complejidad.
- El primer sumando mide el riesgo de T (tasa de error si el problema es de clasificación o la suma de las varianzas residuales si es de regresión)
- El segundo sumando penaliza las estructuras de árbol complejas. El parámetro $\alpha \geq 0$ se denomina parámetro de complejidad

3.2.3.2.3 CHAID vs CART

Ambos métodos se diferencian en los siguientes aspectos:

- El particionamiento CHAID es mulitivía; el de CART es binario
- Por tanto, las estructuras de árbol CHAID suelen ser más simples que las dadas por CART
- CART no requiere discretizar variables. CHAID sí
- Análisis experimentales han demostrado que CHAID es más vulnerable a generar falsos positivos
- CART suele tener mayor capacidad predictiva

3.2.4 Métodos para hacer frente a conjuntos de datos no balanceados

Estos métodos tienen como objetivo transformar una data no balanceada en una balanceada aplicando un mecanismo o algoritmo. La modificación se da alterando el tamaño de la data original y estableciendo una proporción similar de balance. Los métodos más utilizados son los siguientes:

3.2.4.1Undersampling: Esta metodología reduce aleatoriamente el número de observaciones de la clase mayoritaria de tal manera que coincida con el número de observaciones de la clase minoritaria y de esta manera obtener un conjunto de datos balanceados.

3.2.4.2 Oversampling: Esta metodología replica las observaciones de la clase minoritaria con el fin de que se tenga una proporción similar a la otra clase y así obtener un conjunto de datos balanceados. La principal ventaja de este método es la no pérdida de información original.

3.2.4.3 Smote

Es un algoritmo de sobre-muestreo de ejemplos utilizado para la clase minoritaria:

- Crea ejemplos sintéticos en lugar de hacer un sobre-muestreo con reemplazo.
- Opera en el espacio de atributos feature space, en lugar del espacio de datos *data space*.
- Crea un ejemplo sintético a lo largo de los segmentos de línea que unen alguno o todos los k vecinos más cercanos de la clase minoritaria.
- Se eligen algunos de los k vecinos más cercanos de manera aleatoria (no seutilizan todos).
- SMOTE utiliza típicamente k = 5.

El algoritmo de SMOTE realiza los siguientes pasos:

- Recibe como parámetro el porcentaje de ejemplos a sobre-muestrear.
- Calcula el número de ejemplos que tiene que generar.
- Calcula los k vecinos más cercanos de los ejemplos de la clase minoritaria.
- Genera los ejemplos siguiendo este proceso:
  - ✓ Para cada ejemplo de la clase minoritaria, elige aleatoriamente el vecino a utilizar para crear el nuevo ejemplo.
  - ✓ Para cada atributo del ejemplo a sobre-muestrear, calcula ladiferencia entre el vector de atributos muestra y el vecino elegido.
  - ✓ Multiplica esta diferencia por un número aleatorio entre 0 y 1.
  - ✓ Suma este último valor al valor original de la muestra.
  - ✓ Devuelve el conjunto de ejemplos sintéticos.

SMOTE es el algoritmo herramienta más utilizada para realizar el sobre muestreo pero presenta los siguientes inconvenientes:

(i) puede generar muchos ejemplos artificiales cuyas semillas son ejemplos con ruido; (ii) al generar un nuevo ejemplo, interpola entre dos ejemplos de la clase minoritaria, sin embargo, pueden existir muchos

ejemplos cercanos o inclusive entre ellos de la clase mayoritaria, generando modelos incorrectos; (iii) sólo funciona con variables continuas y (iv) no tiene una forma clara de decidir cuántos ejemplos generar.

3.2.5 Indicadores usados para la comparación de modelos

3.2.5.1 Sensibilidad: Representa la proporción de 1´s que fueron correctamente pronosticados como tal. $\mathrm{Se} = \mathrm{P}(\hat{\mathrm{Y}}_\mathrm{i} = 1|\mathrm{Y}_\mathrm{i} =$

$$\mathrm{Es} = \mathrm{P}(\hat{\mathrm{Y}}_\mathrm{i} = 0|\mathrm{Y}_\mathrm{i} = 0)$$

3.2.5.3 Clasificación global**:** Representa la proporción de predicciones que fueron correctas.

3.2.5.4 Índice de Youden**:**

$$\mathrm{I} = (\mathrm{Se} + \mathrm{Es}) - 1$$

3.2.5.5 Ratio de verdaderos positivos (TP)**:** $\mathrm{TP} = \mathrm{P}(\mathrm{Y}_\mathrm{i} = 1|\hat{\mathrm{Y}}_\mathrm{i} = 1)$Ratio de verdaderos negativos (FP**):**

$$\mathrm{FP} = \mathrm{P}(\mathrm{Y}_\mathrm{i} = 0|\hat{\mathrm{Y}}_\mathrm{i} = 0)$$

3.2.5.7 Distancia Euclidiana (**δ**)**:**

$$\delta = \sqrt{\left[\mathrm{P}\left(\mathrm{Y}_\mathrm{i} = 1\middle|\hat{\mathrm{Y}}_\mathrm{i} = 1\right) - 1\right]^2 + \left[\mathrm{P}\left(\mathrm{Y}_\mathrm{i} = 0\middle|\hat{\mathrm{Y}}_\mathrm{i} = 0\right) - 1\right]^2}$$

Se selecciona aquel clasificador que presente una menor distancia euclidiana al punto (1,1), el cual representa los valores óptimos para cada medida respectivamente.

3.2.5.8 Curva de ROC**:** Una curva ROC es una representación gráfica de la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte. Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0,5 (prueba inútil).

3.2.5.9 Índice de Gini**:** $\mathrm{Gini} = 2 * (\mathrm{ROC} - 0.5)$ Si el valor del Gini se encuentra entre 0 y 0.25, decimos que el modelo predictivo

tiene una clasificación “Baja”; si el valor del Gini se encuentra entre 0.25 y 0.45, tiene una clasificación “Aceptable”; si el valor del Gini se encuentra entre 0.45 y 0.6, tiene una clasificación “Buena”, y finalmente, si el valor del Gini es mayor a 0.5, el modelo tiene una clasificación de “Muy buena”.

3.2.5Ajuste por balanceo de la distribución de la variable target

Una de las principales tareas cuando se plantea construir un modelo predictivo o de clasificación es establecer la variable objetivo o también denominada target. Si no se encuentra disponible de manera directa en una base de datos, es posible crearla utilizando los patrones de las variables que sí se encuentran disponibles. Luego, de ello lo que queda es determinar la distribución de dicha variable objetivo, en caso de ser una variable categórica, el tener una desproporción marcada entre sus categorías y construir un modelo predictivo puede arrojar conclusiones erróneas.

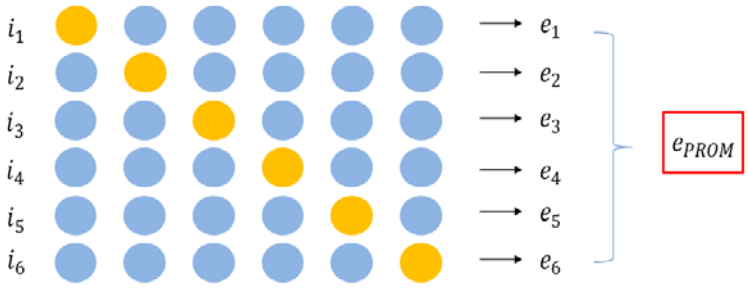
$$þ_i = \frac{p_i * w_o * \pi_1}{(1 - p_i) * w_1 * \pi_0 + p_i * w_o * \pi_1}$$

Donde:

- $þ_i$: Probabilidad transformada
- p_i : Probabilidad obtenida de la muestra balanceada.
- w_o: Pobabilidad de la categoría 0 de la variable Y en la data balanceada.
- w_1: Pobabilidad de la categoría 1 de la variable Y en la data balanceada.
- π_0: Pobabilidad de la categoría 0 de la variable Y en la data inicial.
- π_1: Pobabilidad de la categoría 1 de la variable Y en la data inicial.

3.2.6. K-Fold Cross Validation

Consiste en separar el conjunto de datos en K grupos de igual tamaño. Se realizarán K iteraciones. En la i-ésima iteración, el i-ésimo grupo formará parte de la muestra de validación y los grupos restantes conformarán la muestra de entrenamiento. Para cada iteración se obtendrá una tasa de error (1-precisión global) y validaremos que los errores a lo largo de las iteraciones no muestren variaciones significativas.



3.3 Terminología Básica

- ✓ **Accidentes de tránsito:** Suceso que ocurre sobre la vía y se presenta súbita e inesperadamente, determinado por condiciones y actos irresponsables potencialmente previsibles, atribuidos a factores humanos, vehículos preponderantemente automotores, condiciones climatológicas, señalización y caminos, etc

- ✓ **Accidente de tránsito fatal:** Se considera un accidente de tránsito fatal, cuando el resultado del accidente es la muerte de una persona ya sea dentro del auto o fuera del auto.

CAPÍTULO IV

METODOLOGÍA

4.1 Población en estudio

Son todos los accidentes de tránsito ocurrido en Lima Metropolitana durante el año 2014 que fueron registrados por el Sistema de Denuncias Policiales (SIDPOL) de las comisarías censadas.

4.2 Fuentes de información

Son los libros de ocurrencias o Sistema de Denuncias Policiales (SIDPOL) de las comisarías censadas.

4.3 Tipo de investigación

La presente estudio es una investigación de tipo causal de corte transversal pues se estudian hechos recopilados en un instante de tiempo.

4.4 Definición de variables

A continuación se mostrará las variables que se encuentran en el Cuestionario del Censo Nacional a Comisarías – 2015. A partir del cuestionario se determinará la variable dependiente.

4.4.1 Variable Dependiente:

Fatalidad del accidente (y): y=1 (accidente de tránsito fatal, y=0: Accidente de tránsito no fatal.
La variable dependiente se obtiene a partir de la variable: Consecuencia del accidente, con categorías: “Fatal”, “No fatal” y “Solo daños materiales”. Se considerará a las 2 últimas categorías como “No fatal”

4.4.2 Variables independientes:

Cuadro Nº 4.1

CUADRO DE VARIABLES INDEPENDIENTES DEL CUESTIONARIO DEL CENSO NACIONAL DE COMISARÍAS-ACCIDENTES DE TRÁNSITO 2015

VARIABLE	TIPO DE VARIABLE	ESCALA DE MEDICIÓN	Descripción
Fecha de ocurrencia del accidente de tránsito (Día)	Cuantitativa	Intervalo	
Fecha de ocurrencia del accidente de tránsito (Mes)	Cuantitativa	Intervalo	
Fecha de ocurrencia del accidente de tránsito (Año)	Cuantitativa	Intervalo	
Tiempo de ocurrencia del accidente de tránsito (hora)	Cuantitativa	Intervalo	
Lugar de ocurrencia	Cualitativa	Nominal	1: Autopista 2: Carretera 3: Vía expresa 4: Avenida 5: Calle o jirón 6: Trocha 7: No identificado 8: Otro
Lugar de ocurrencia-tramo de la vía			1: Intersección 2: Recta 3: Curva 4: Rotonda (óvalo) 5: Bifurcación 6: No identificado 7: Otro
Nombre de la vía 1 del lugar de ocurrencia	Cualitativa	Nominal	
Referencia 1 del lugar de ocurrencia	Cualitativa	Nominal	

Nombre de la vía 2 del lugar de ocurrencia	Cualitativa	Nominal	
Referencia 2 del lugar de ocurrencia	Cualitativa	Nominal	
Tipo de accidente de tránsito	Cualitativa	Nominal	1: Atropello 2: Atropello y fuga 3: Caída de pasajero 4: Colisión 5: Colisión y fuga 6: Choque 7: Choque y atropello 8: Choque y fuga 9: Despiste 10: Despiste y volcadura 11: Volcadura 12: Otro
Tipo de vehículo mayor involucrado en el accidente de tránsito	Cualitativa	Nominal	1: Automóvil 2: StationWagon 3: Camioneta pick up 4: Camioneta rural 5: Camioneta panel o furgoneta 6: Ómnibus urbano 7: Camión 8: Remolcador 9: Trayler 10: Otro
Tipo de transporte	Cualitativa	Nominal	1. Público 2. Privado
Número de fallecidos	Cuantitativa	Discreta	
Número de heridos	Cuantitativa	Discreta	
Número de ilesos	Cuantitativa	Discreta	
Factores vinculados al accidente de tránsito	Cualitativa	Nominal	1: Exceso de velocidad 2: Desacato a la señal de tránsito por el conductor 3: Falta de

			iluminación en la vía 4: Exceso de carga 5: Ebriedad del conductor 6: Falla mecánica 7: Vía en mal estado 8: Desacato a la señal de tránsito por el peatón 9: Cansancio o fatiga por el conductor 10: Otro

FUENTE: Cuestionario-Censo Nacional de Accidentes de Tránsito registrado por las Comisarías a nivel nacional en el 2015.
ELABRACIÓN: Propia.

4.5 Diseño de muestras y preparación de datos

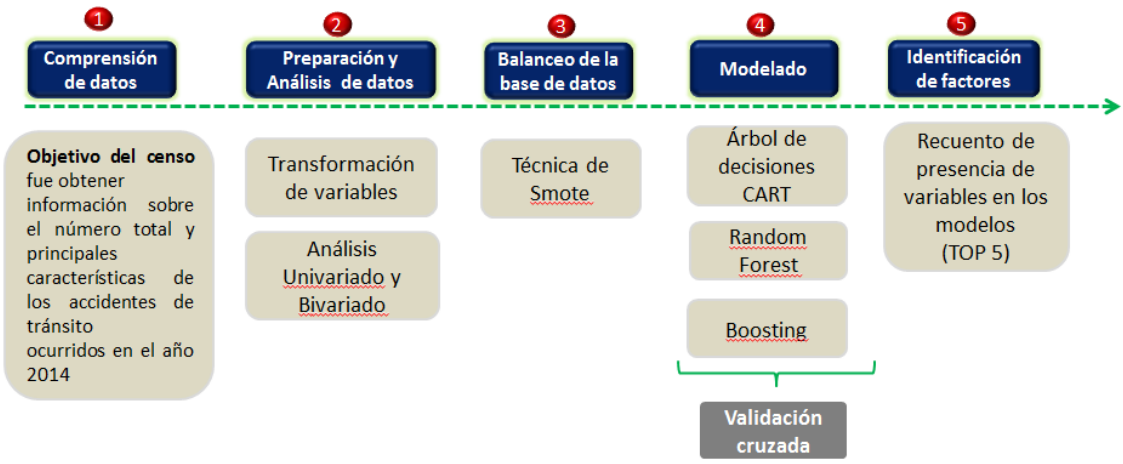
El estudio se centrará a nivel de Lima Metropolitana, como se dispone de datos del censo, se trabajará con toda la población.

4.6 Procedimiento estadístico

Para cumplir los objetivos se siguieron los siguientes pasos:

GRÁFICO Nº 4.1

DIAGRAMA DE PROCEDIMEINTO ESTADÍSTICO



ELABORACIÓN: Propia

4.6.1 Comprensión de los datos

En la etapa de Comprensión de los datos se procedió a identificar las variables y entender su significado, relacionado al contexto en estudio. Implicó la revisión del tipo de variable.

4.6.2 Preparación y análisis de los datos

Involucró la realización de los siguientes:

a) Transformación de variables

- Dicotomización de variables

- Reagrupaación de categorías

b) Análisis univariado

c) Análisis Bivariado

4.6.3 Balanceo de la base de datos

Se dispuso de una base de datos desbalanceada, por lo que la técnica de balanceo más apropiado que se usó fue el Smote.

4.6.4 Modelado

- ✓ Una vez particionado la base de datos en muestra de construcción y validación, y balanceado la muestra de construcción, se procedió

a la construcción de los modelos de minería de datos: Boosting, Random Forest y Árbol de decisiones CART.

- ✓ Se verificó la robustez de cada uno de os modelos construídos a fin de asegurarnos el uso posterior de sus resultados para el logro de los objetivos.
- ✓ Lo anterior se procedió a verificar mediante el uso de la técnica de Validación Cruzada con k fold = 10, con el fin de asegurar que los modelos no presenten sobreajuste y sean estables.

4.6.5 Identificación de variables principales factores

- ✓ A partir del resultado de Importancia de variables de cada uno de los modelos construídos, la metodología de identificación de los principales factores será identificar aquellos con mayor frecuencia de presencia en los tres modelos; siendo los más importantes aquellos que más veces se repitan en los modelos. De esa manera no solo concluimos con los resultados de un modelo, sino que nos aseguramos que la elección sea más confiable.

CAPÍTULO V

RESULTADOS

5.1 Análisis descriptivo de las variables

5.1.1 Análisis univariado de las variables independientes

a) Localización de la dependencia policial

Se analizó de forma descriptiva los accidentes de tránsito según ubicación geográfica, se comparó los distritos de Lima Metropolitana (conformado por 43 distritos) con el resto de departamentos.

Cuadro Nº 5.1

Accidentes de Tránsito registrado según ubicación geográfica

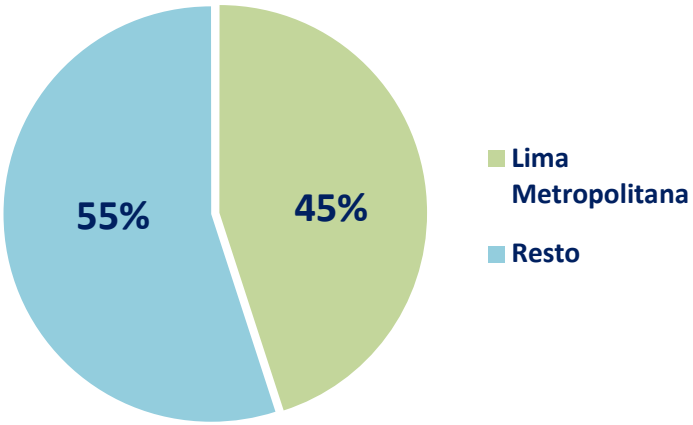
Ubicación Geográfica	Número de accidentes	% Total de Accidentes
Lima Metropolitana	55699	45%
Resto	68087	55%
Total	123786	100%

FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

Gráfico Nº 5.1

Accidentes de Tránsito registrado según ubicación geográfica



FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

Del total de accidentes de accidentes de tránsito registrados en el país, en el año 2014; el 45% fueron registrados en comisarías de Lima metropolitana; es decir casi la mitad de los accidentes de tránsito ocurren en la capital del país

Cuadro Nº 5.2

Frecuencia de accidentes de tránsito registrado en el 2014 en los distritos de Lima Metropolitana

Nº	DISTRITO	Número de Accidentes de Tránsito
1	ANCON	160
2	ATE	2573
3	BARRANCO	487
4	BREÑA	871
5	CARABAYLLO	876
6	CHACLACAYO	261
7	CHORRILLOS	1609
8	CIENEGUILLA	119
9	COMAS	1862
10	EL AGUSTINO	848
11	INDEPENDENCIA	522
12	JESUS MARIA	989
13	LA MOLINA	1838
14	LA VICTORIA	1513
15	LIMA	3908
16	LINCE	1148
17	LOS OLIVOS	2579
18	LURIGANCHO	696
19	LURIN	535
20	MAGDALENA DEL MAR	680
21	MIRAFLORES	3260
22	PACHACAMAC	231
23	PUCUSANA	53
24	PUEBLO LIBRE	923
25	PUENTE PIEDRA	1503
26	PUNTA HERMOSA	101
27	PUNTA NEGRA	31
28	RIMAC	1058
29	SAN BARTOLO	10
30	SAN BORJA	2458
31	SAN ISIDRO	2649
32	SAN JUAN DE LURIGANCHO	3511
33	SAN JUAN DE MIRAFLORES	1760
34	SAN LUIS	660
35	SAN MARTIN DE PORRES	1920
36	SAN MIGUEL	1636
37	SANTA ANITA	949
38	SANTA MARIA DEL MAR	54
39	SANTA ROSA	27
40	SANTIAGO DE SURCO	4435
41	SURQUILLO	1215
42	VILLA EL SALVADOR	1764
43	VILLA MARIA DEL TRIUNFO	1417
	TOTAL	**55699**

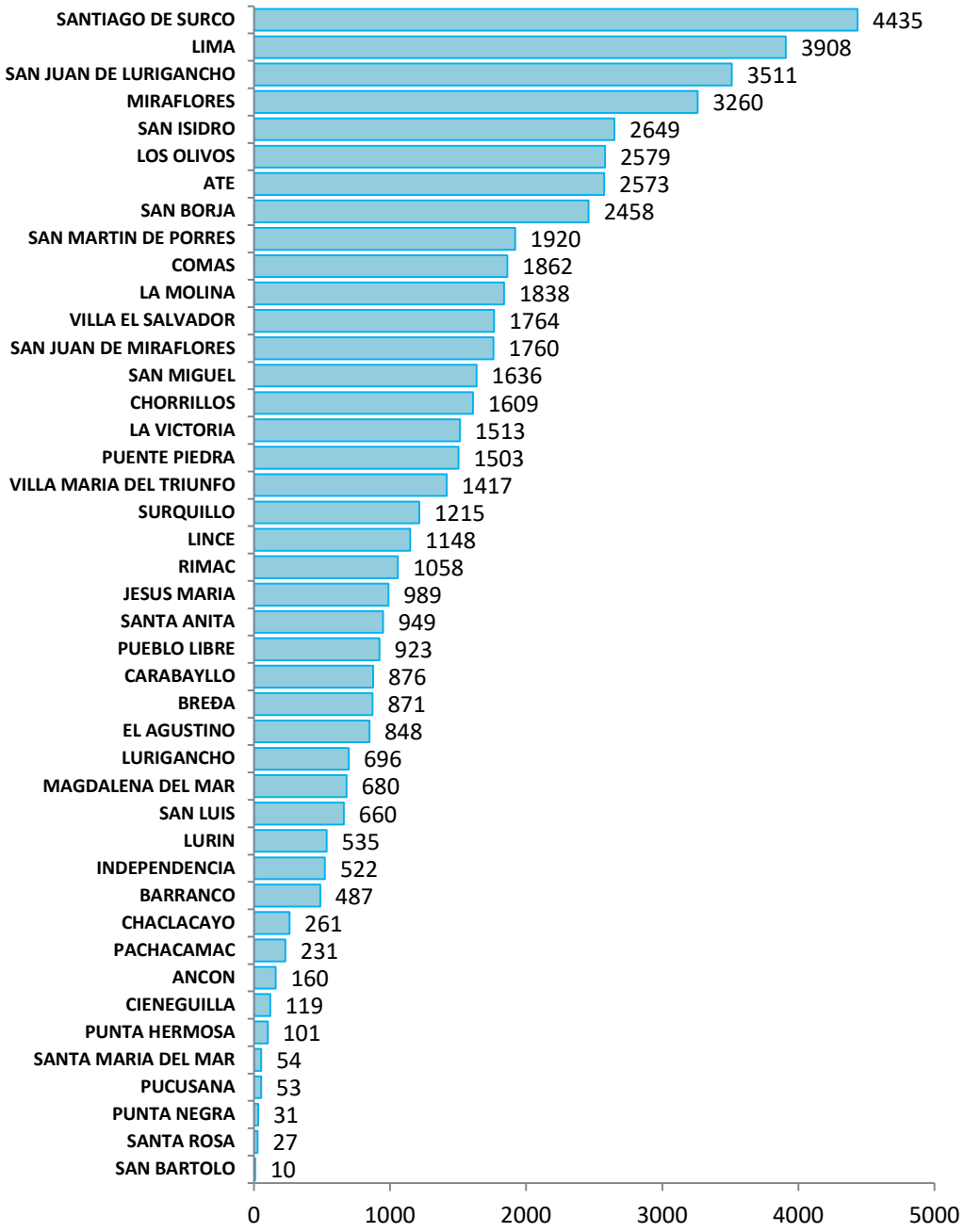
FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

Se parte de un tamaño de población de 55 699 registros de accidentes de tránsito ocurridos en el 2014 en los distritos de Lima Metropolitana (43 distritos)

Gráfico Nº 5.2

Número de Accidentes de Tránsito registrados en el 2014 por las comisarías de los distritos de Lima Metropolitana



FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

Del gráfico Nº 5.1 se observa que Santiago de Surco es el distrito con mayor número de accidentes registrados (4435), seguido de Lima (3908), San Juan de Lurigancho (3511), Miraflores (3260), San Isidro (2649), Los Olivos (2579), Ate (2573), etc.

b) Tipo de accidente

Cuadro Nº 5.3

Tipos de accidentes de tránsito registrados en Lima Metropolitana-2014

Descripción	Frecuencia	Porcentaje
Colisión	14977	27%
Choque	12111	22%
Atropello	9401	17%
Choque y fuga	5551	10%
Despiste	3929	7%
Colisión y fuga	3562	6%
Caída de pasajero	3070	6%
Atropello y fuga	1756	3%
Volcadura	419	1%
Despiste y volcadura	407	1%
Otro	309	1%
Choque y atropello	207	0%
Total	**55699**	**100%**

FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

El mayor tipo de accidente de tránsito en Lima Metropolitana es por colisión (27%), seguido de Choque (22%) y Atropello (17%).

c) Lugar de ocurrencia - Tipo de vía

Cuadro Nº 5.4

Lugar de Ocurrencia de los accidentes de tránsito registrados en Lima Metropolitana-2014

Lugar	Frecuencia	Porcentaje
Avenida	42237	76%
Calle o jirón	7564	14%
Carretera	4149	7%
Autopista	553	1%
Vía expresa	543	1%
No identificado	513	1%
Otro	135	0%
Trocha	5	0%
Total	**55699**	**1**

FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016
ELABORACIÓN: Propia

El 90% de accidentes de tránsito ocurren en Avenidas y Calles o Jirones. En avenidas cerca al 76% de accidentes de tránsito registrados y en Calles o jirones un 14%.

d) Fecha de ocurrencia-mes

Cuadro Nº 5.5

Mes de Ocurrencia de los accidentes de tránsito registrados en Lima Metropolitana - 2014

Mes	Frecuencia	Porcentaje
Marzo	5026	9%
Abril	4781	9%
Junio	4755	9%
Diciembre	4752	9%
Octubre	4746	9%
Septiembre	4746	9%
Julio	4527	8%
Noviembre	4524	8%
Enero	4481	8%
Febrero	4462	8%
Agosto	4455	8%
Mayo	4445	8%
Total	**55699**	**100%**

FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016
ELABORACIÓN: Propia

Del cuadro Nº 5.5, se observa que no influye en gran medida el mes de ocurrencia del accidente de tránsito, los porcentajes están cercanos por mes.

e) Factores vinculados al accidente

Cuadro Nº 5.6

Factores de Ocurrencia de los accidentes de tránsito registrados en Lima Metropolitana - 2014

FACTORES	Número de accidentes de tránsito	Porcentaje
EXCESO DE VELOCIDAD	21417	32%
INVASION DEL CARRIL / MANIOBRAS NO PERMITIDAS	20914	32%
DESACATO A LA SEÑAL DE TRANSITO POR EL CONDUCTOR	9717	15%
DESACATO A LA SEÑAL DE TRANSITO POR EL PEATO	2965	4%
EBRIEDAD DEL CONDUCTOR	2562	4%
FALLA MECÁNICA	1461	2%
VÍA EN MAL ESTADO	731	1%
ESTADO DE EBRIEDAD DEL PEATON	491	1%
EXCESO DE CARGA	406	1%
SEÑALIZACIÓN DEFECTUOSA	399	1%
CANSANCIO O FATIGA DEL CONDUCTOR	367	1%
FACTOR CLIMATICO	333	1%
FALTA DE ILUMINACIÓN EN LA VÍA	114	0%
USO DEL CELULAR O DISPOSITIVOS ELECTRÓNICOS	109	0%
OTRO	4237	6%
TOTAL	**66222**	**100%**

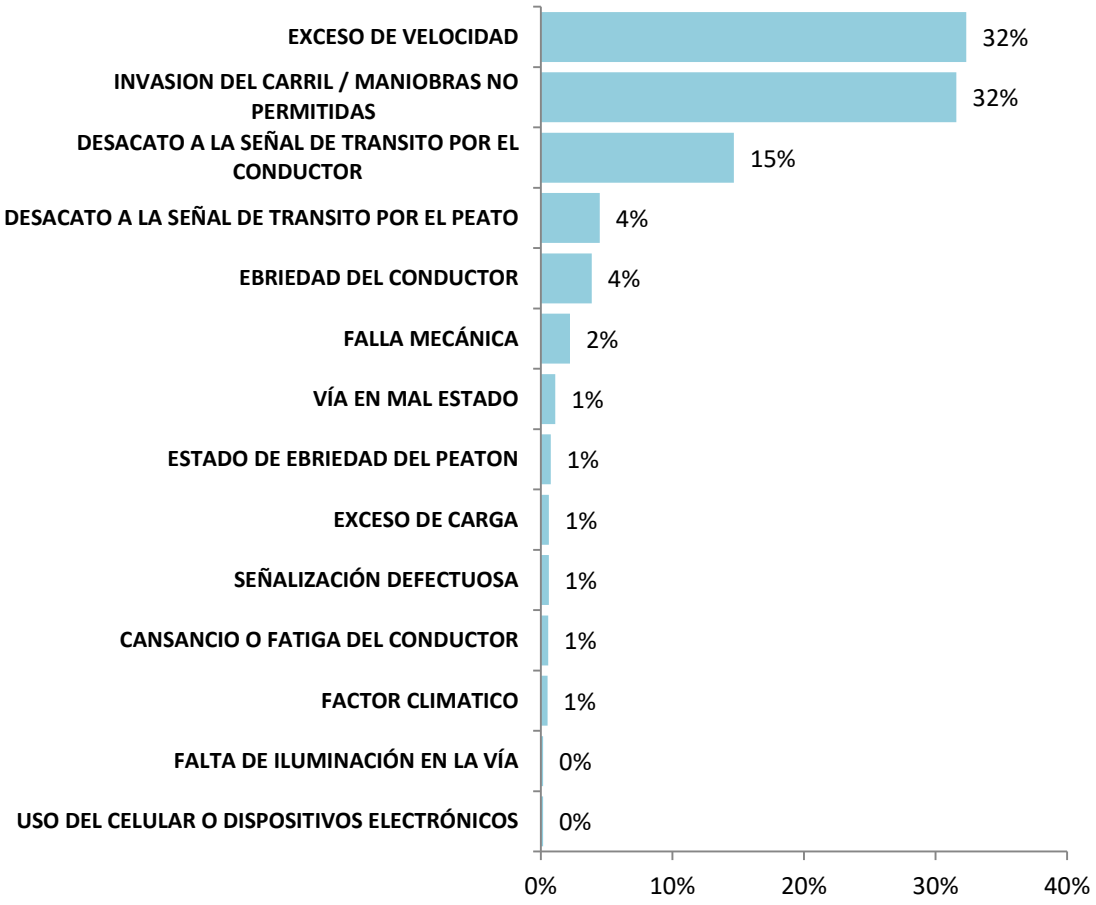
FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016

ELABORACIÓN: Propia

En el cuadro N° 5.6, el total de accidentes de tránsito registrado por factor, no coincide con

Gráfico Nº 5.3

Factores de Ocurrencia de los accidentes de tránsito registrados en Lima Metropolitana - 2014



FUENTE: IV Censo Nacional de Comisarías 2015 31/10/2016
ELABORACIÓN: Propia

Del gráfico Nº 5.3 se observa que los tres principales factores de accidente de tránsito son por Exceso de velocidad en un 32%, seguido de Invasión del carril / maniobras no permitidas en un 32% y desacato a la señal de tránsito por el conductor en un 15.

f) Consecuencia del accidente-fatalidad

Cuadro Nº 5.7

Consecuencia de los accidentes de tránsito registrados en Lima Metropolitana - 2014

Consecuencia	Número de accidentes	Porcentaje
No Fatal	55379	99%
Fatal	320	1%
Total	**55699**	**100%**

FUENTE: IV Censo Nacional de Comisarías 2015 15/11/2016
ELABORACIÓN: Propia

Del total de accidentes de tránsito registrados solo el 1% representa el porcentaje de accidentes fatales.

5.2 Modelado

5.2.1 Partición de la población

Se particionó la población en muestras de construcción y validación. Se consideró la partición 70%(construcción)-30%(validación)
Se observó en el análisis descriptivo el desbalanceo de la base de datos, por lo que seguidamente se procedió a balancear mediante el algoritmo Smote

5.2.2 Balanceo de la muestra de construcción

Para tener resultados más concisos y evitar así la tendencia de clasificación hacia la clase mayoritaria, que podría minimizar el error de clasificación y clasificar correctamente instancias de la clase mayoritaria en detrimento de instancias de la clase minoritaria.

5.2.3 Selección de variables

La técnica de selección de variables usada fue el Criterio de Valor de información, el cual se obtuvo con los paquetes “woe”, “stringr “ e “Information”. Las variables que resultaron con un valor de información alto fueron: "ocurrencia_carretera", "ocurrencia_avenida", "may_automovil", "tramo_interseccion", "transp_privado", "tipo_at_atropelloyfuga", "tramo_recta", "transp_publico", "men_motocar", "may_omnibus_urbano", "f_invasion_carril", "f_exceso_velocidaad", "tipo_at_atropello", "may_camioneta_rural", "f_desacato_señal_dtrans_cond", "f_desacato_señal_dtrans_peaton", "may_camion”

5.2.4 Modelo Boosting

A continuación se muestral los valores de los indicadores en las 10 iteraciones en el modelo:

Cuadro Nº 5.8

MODELO BOOSTING (k fold croos- validation)

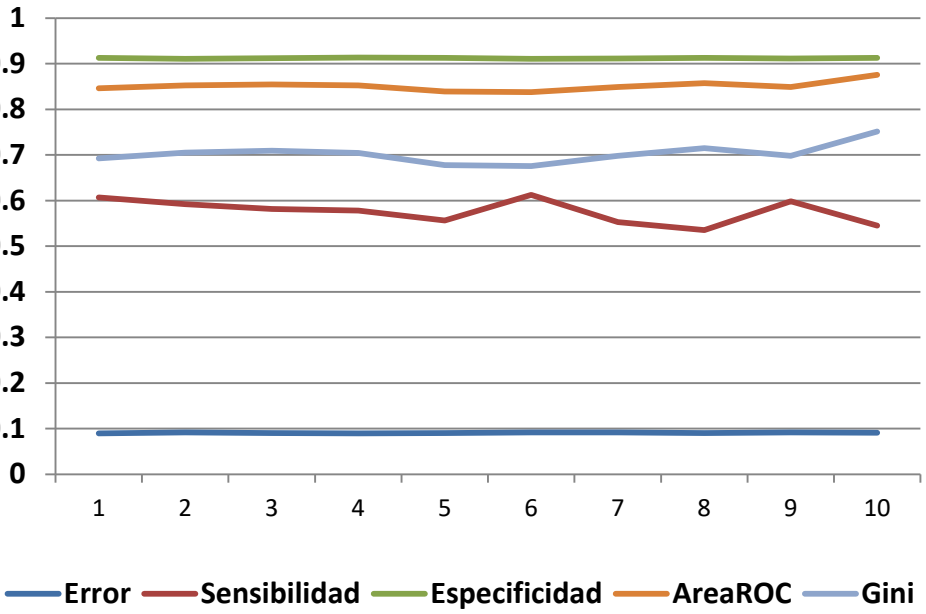
	Error	Sensibilidad	Especificidad	VPP	VPN	AreaROC	Gini	Dist.Eucli
1	0,0895447	60,69%	91,30%	5,68%	99,64%	84,62%	69,24%	0,9431663
2	0,0917881	59,19%	91,09%	5,39%	99,62%	85,26%	70,53%	0,9460844
3	0,0902528	58,16%	91,25%	5,44%	99,61%	85,45%	70,91%	0,9456341
4	0,0894280	57,82%	91,34%	5,51%	99,61%	85,22%	70,43%	0,9449543
5	0,0903700	55,62%	91,27%	5,20%	99,58%	83,90%	67,80%	0,9479850
6	0,0914359	61,22%	91,11%	5,53%	99,63%	83,77%	67,54%	0,9447278
7	0,0913168	55,30%	91,15%	5,30%	99,61%	84,92%	69,84%	0,9469811
8	0,0901351	53,52%	91,30%	5,19%	99,57%	85,73%	71,46%	0,9481146
9	0,0913169	59,82%	91,14%	5,47%	99,62%	84,90%	69,81%	0,9452731
10	0,0906069	54,49%	91,27%	4,91%	99,56%	87,56%	75,12%	0,9508870

FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Gráfico Nº 5.4

IMODELO BOOSTING (k fold croos- validation)



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico se observa estabilidad en los indicadores de desempeño, especialmente en el error de predicción. Si bien, se observa algunas variaciones en otros indicadores, esto solo se presenta con variaciones pequeñas.

Cuadro Nº 5.9

Cuadro de importancia de variables (Modelo boosting)

variables	importancia
ocurrencia_carretera	8,8542255
may_camion	5,4758791
f_desacato_señal_dtrans_cond	5,3404963
may_automovil	3,0284087
ocurrencia_avenida	2,4823858
may_camioneta_rural	2,4239474
tipo_at_atropello	2,3118105
f_invasion_carril	2,0733699
transp_publico	1,8141429
tramo_interseccion	1,6869175

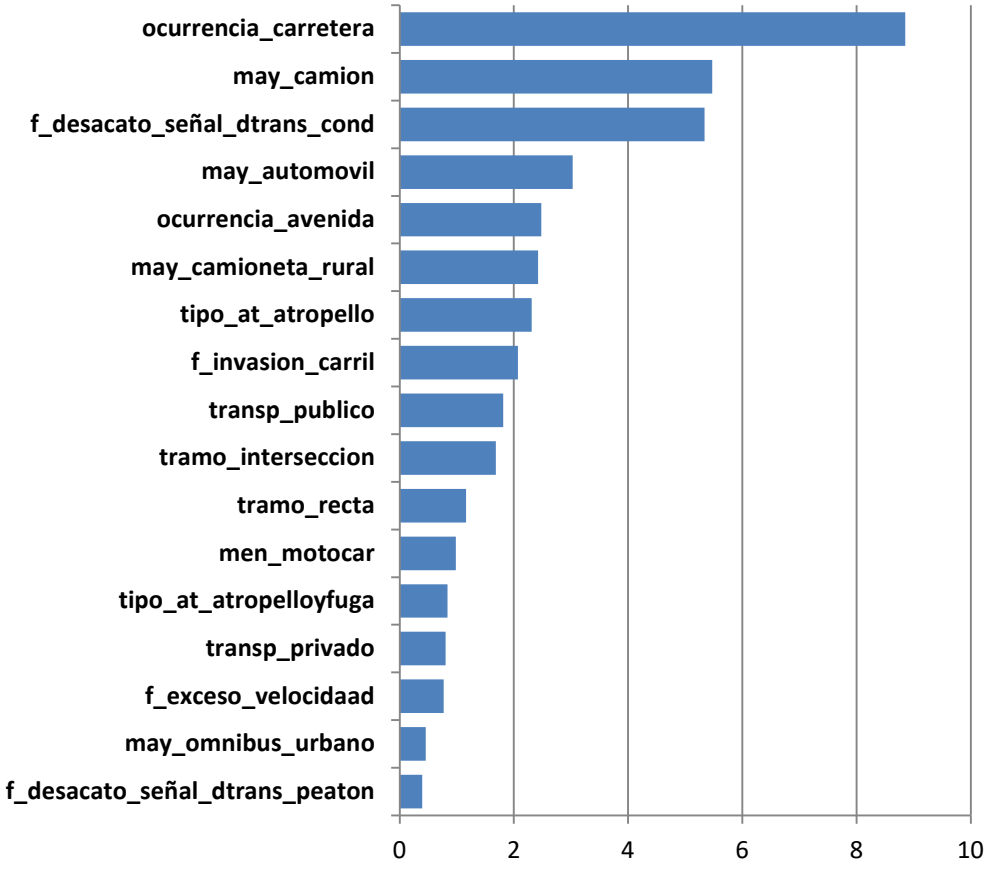
tramo_recta	1,1621104
men_motocar	0,9863531
tipo_at_atropelloyfuga	0,8379125
transp_privado	0,8025781
f_exceso_velocidaad	0,7679292
may_omnibus_urbano	0,4542276
f_desacato_señal_dtrans_peaton	0,3928472

FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Gráfico Nº 5.5

Gráfico de importancia de variables (Modelo boosting)



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico anterior se observa que las variables más importantes (Top 3) para el modelo Boosting son: Tipo de vía(Ocurrencia en carretera), Tipo

de vehículo involucrado(camión) y Ocurrencia por desacato a la señal de tránsito por parte del conductor.

5.2.5 Random Forest

Cuadro Nº 5.10

MODELO RANDOM FOREST (k fold croos- validation)

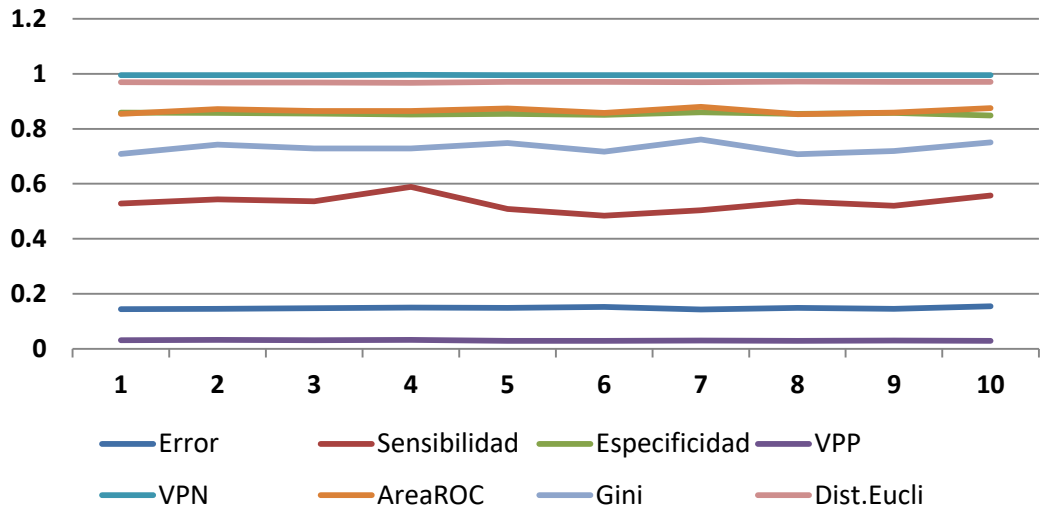
	Error	Sensibilidad	Especificidad	VPP	VPN	AreaROC	Gini	Dist.Eucli
1	0,1436481	52,85%	85,92%	3,08%	99,53%	85,45%	70,90%	0,9692534
2	0,1448347	54,29%	85,79%	3,22%	99,54%	87,14%	74,28%	0,9678534
3	0,1474315	53,67%	85,52%	3,10%	99,54%	86,44%	72,89%	0,9689855
4	0,1501478	58,90%	85,21%	3,27%	99,58%	86,44%	72,88%	0,9673534
5	0,1483778	50,82%	85,47%	2,92%	99,50%	87,39%	74,79%	0,970818
6	0,1519195	48,38%	85,08%	2,89%	99,53%	85,83%	71,66%	0,9711058
7	0,1430609	50,36%	85,99%	3,01%	99,52%	88,05%	76,11%	0,969864
8	0,1483696	53,48%	85,46%	2,83%	99,50%	85,36%	70,73%	0,9717062
9	0,1455431	51,98%	85,75%	2,95%	99,50%	85,94%	71,87%	0,9704882
10	0,1545165	55,68%	84,82%	2,90%	99,53%	87,53%	75,05%	0,9710261

FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Gráfico Nº 5.6

MODELO RANDOM FOREST (k fold croos- validation)



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico se observa estabilidad en los indicadores de desempeño, especialmente en el error de predicción. Si bien, se observa algunas variaciones en otros indicadores, esto solo se presenta con variaciones pequeñas.

Cuadro Nº 5.11

Cuadro de importancia de variables (Modelo Random Forest)

Variables	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
f_desacato_señal_dtrans_cond	0,041021853	0,04241962	0,041733173	270,0524819
may_automovil	0,014677762	0,02962795	0,022248126	193,0967683
may_camion	0,021189228	0,02067301	0,020968156	127,3602163
tipo_at_atropello	0,003519769	0,005931994	0,004749049	29,61812971
f_invasion_carril	0,001491975	0,004295461	0,002899425	28,42461583
transp_publico	0,005771171	0,002923631	0,004333879	26,96966503
ocurrencia_avenida	0,003157334	0,001841494	0,002467092	14,65438297
transp_privado	0,002791618	0,00065477	0,001730768	11,00365083
tipo_at_atropelloyfuga	0,001463878	0,000392327	0,000923154	5,57171037
men_motocar	0,001091352	0,000624562	0,00085816	5,26091014
may_camioneta_rural	0,001022569	0,000492079	0,000761093	3,27666498
may_omnibus_urbano	0,000304997	-3,0047E-05	0,00013816	2,97388086
f_desacato_señal_dtrans_peaton	0,000400309	3,12891E-05	0,000211244	0,75855354
f_exceso_velocidaad	0	0	0	0,38453528
tramo_recta	0	0	0	0,08423473
tramo_interseccion	0	0	0	0

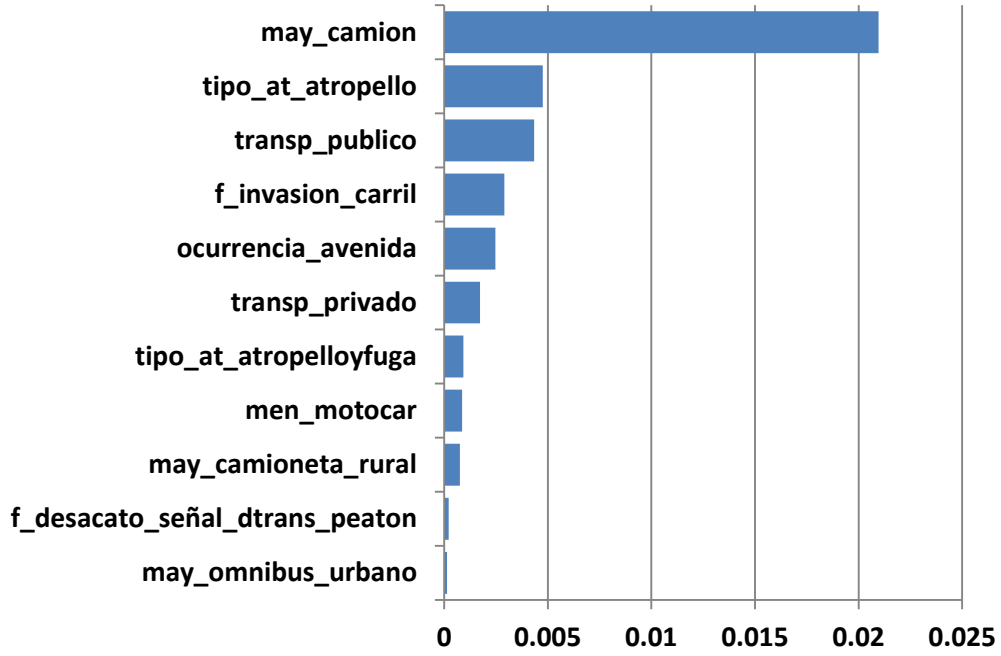
FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

En el cuadro anterior se muestral los valores del aporte de cada variable a la predicción (MeanDecreaseAccuracy) y a la construcción del modelo (MeanDecreaseGini)

Gráfico Nº 5.7

Gráfico de importancia de variables (mean decrease accuracy)



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico anterior se observa que las variables más importantes (Top 3) para el modelo Random Forest son: Tipo de vehículo involucrado (camión-combi), Accidente de tránsito por atropello y Tipo de transporte-público.

5.3.5 Modelo Árbol de decisiones CART

Cuadro Nº 5.12

MODELO ARBOL DE DECISIONES CART (k fold croos- validation)

	Error	Sensibilidad	Especificidad	VPP	VPN	AreaROC	Gini	Dist.Eucli
1	0,2131133	66,13%	78,82%	2,52%	99,61%	85,39%	70,77%	0,9747819

2	0,2131106	63,75%	78,82%	2,53%	99,61%	85,77%	71,53%	0,9746758
3	0,2131115	64,31%	78,82%	2,53%	99,61%	84,36%	68,71%	0,9747478
4	0,2131143	62,57%	78,82%	2,52%	99,61%	84,92%	69,84%	0,9748034
5	0,2131138	62,96%	78,82%	2,52%	99,61%	87,77%	75,54%	0,9747726
6	0,2131147	63,91%	78,82%	2,52%	99,61%	84,77%	69,55%	0,9747766
7	0,2131124	65,15%	78,82%	2,52%	99,61%	85,67%	71,34%	0,9748048
8	0,2131103	63,85%	78,82%	2,54%	99,61%	84,62%	69,24%	0,9746529
9	0,2131114	64,46%	78,82%	2,51%	99,61%	84,47%	68,93%	0,9749064
10	0,2131128	66,39%	78,82%	2,51%	99,61%	86,84%	73,67%	0,9748968

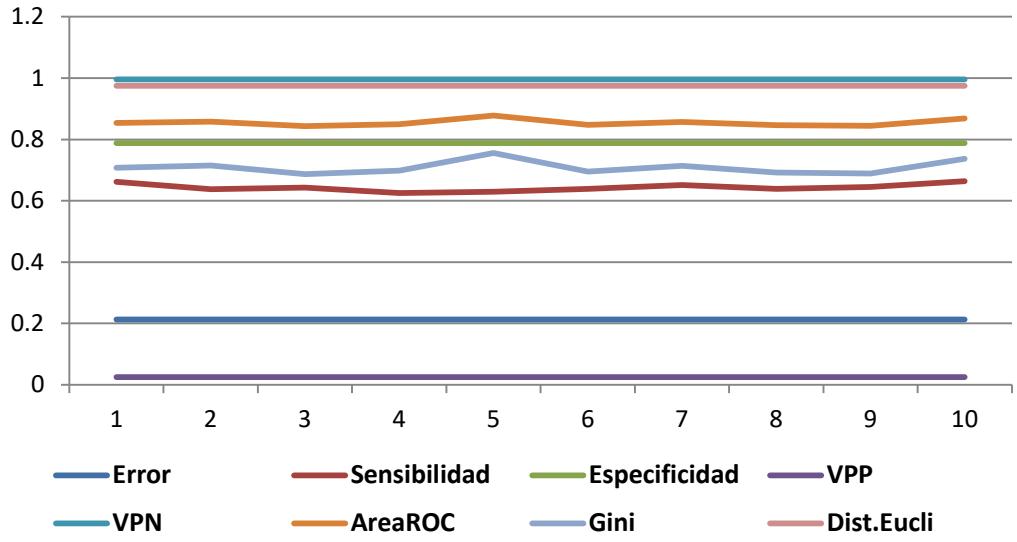
FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico se observa estabilidad en los indicadores de desempeño, especialmente en el error de predicción. Si bien, se observa algunas variaciones en otros indicadores, esto solo se presenta con variaciones pequeñas.

Gráfico Nº 5.8

MODELO RANDOM FOREST (k fold croos- validation)



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Cuadro Nº 5.13

Cuadro de importancia de variables (Modelo Random Forest)

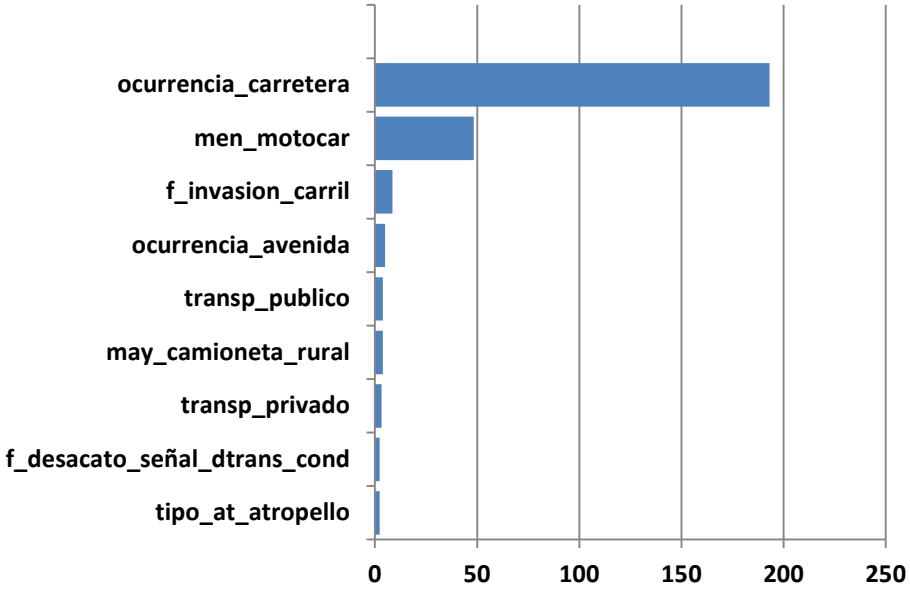
VARIABLES	IMPORTANCIA
ocurrencia_carretera	193,079408
men_motocar	48,273339
f_invasion_carril	8,486862
ocurrencia_avenida	4,982075
transp_publico	3,84083
may_camioneta_rural	3,788215
transp_privado	3,20946
f_desacato_señal_dtrans_cond	2,329727
tipo_at_atropello	2,262406

FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Gráfico Nº 5.9

Gráfico de importancia de variables



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

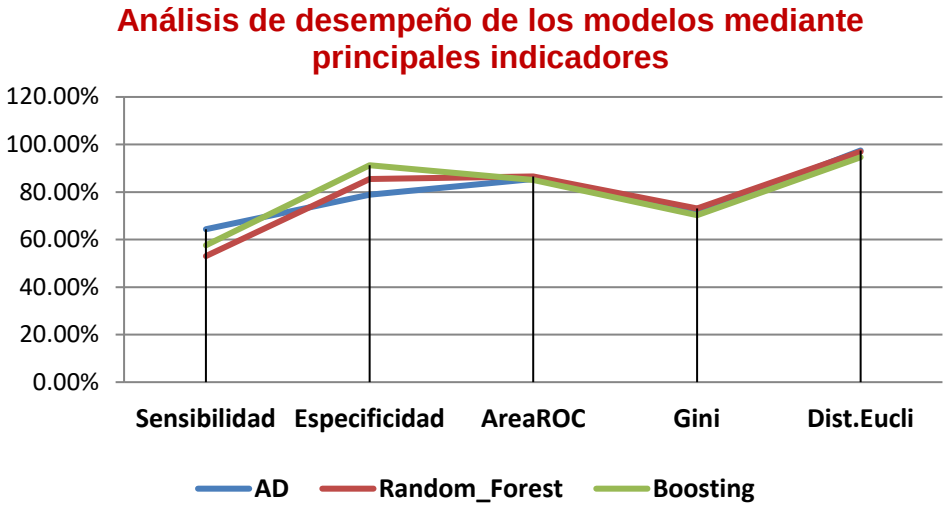
ELABORACIÓN: Propia

Del gráfico anterior se observa que las variables más importantes (Top 3) para el modelo Random Forest son: Tipo de vía de ocurrencia del accidente-caretera, Tipo de vehículo involucrado (motocar y/o mototaxi),

Accidente de tránsito por invasión al carril contrario y Tipo de transporte-público.

5.3.6 Comparación de indicadores de desempeño de los modelos

Gráfico Nº 5.10

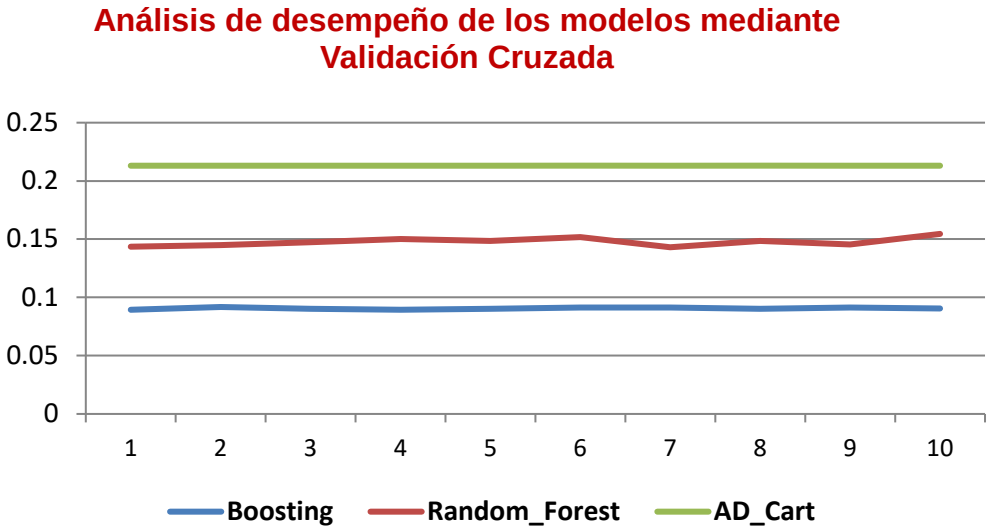


FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Se observa que el modelo que presenta un valor más alto de sensibilidad es el modelo árbol de decisiones CART, seguido de Booting y luego Random Forest; sin embargo ello no es determinante. Asimismo, se observa que para los indicadores Area ROC, GINI y Distancia Euclídea, los modelos presentan similares valores, mostrando así su robustez, pues sus valores superan el 70%

Gráfico Nº 5.11



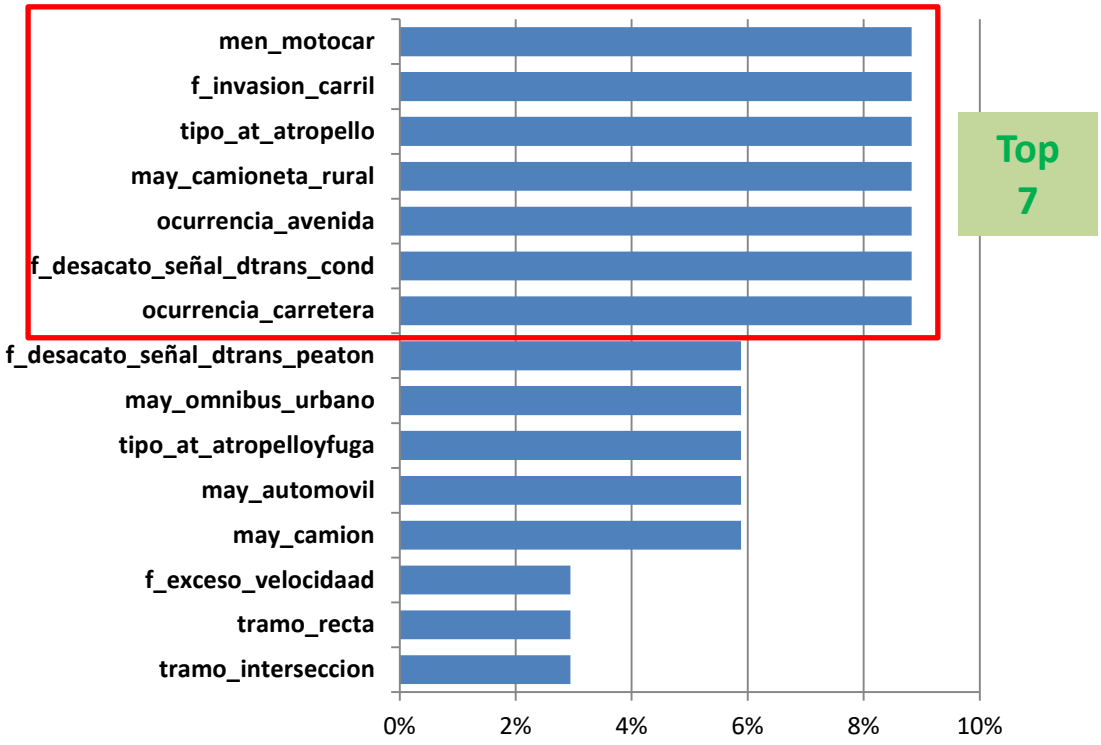
FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Del gráfico anterior se observan los valores del error de predicción para uno de las 10 iteraciones de los tres modelos. Se observa que el modelo con menores tasas de error es el modelo Boosting, seguid por Random Forest y finalmente Árbol de decisiones CART. Asimismo se observa estabilidad de los tres modelos.

Gráfico Nº 5.12

Gráficode Importancia de variables a partir de un recuento de la presencia en los tres modelos realizados



FUENTE: IV Censo Nacional de Comisarías 2015 28/11/2016

ELABORACIÓN: Propia

Para determinar los factores de influencia en los accidentes de tránsito fatales, que se muestran en el gráfico anterior, se realizó un recuento de las variables que más influyen por separado en cada modelo desarrollado, teniendo en consideración que las variables que tuvieron mayor presencia en los tres modelos fueron consideradas como las más importantes.

CONCLUSIONES

- ✓ Los principales factores que influyen en la fatalidad de los accidentes de tránsito en Lima Metropolitana, los cuales se obtuvieron a partir de los resultados de técnicas de minería de datos: Random Forest, Boosting y Árbol de Decisiones, fueron: Que el accidente ocurrió en el tipo de vía carretera o avenida, fue a causa del desacato a la señal de tránsito por parte del conductor, el tipo de vehículo fue una camioneta rural(combi), que sea causado por invadir un carril, que el tipo de vehículo fue mototaxi y/o motocar

- ✓ Los valores de los indicadores de desempeño de los modelos usados son: Gini mayor a 50% para todos los modelos y mayor a 70% para el modelo Boosting, Sensibilidad mayor a 55% para todos los modelos, y Especificidad mayor a 60% para todos los modelos

- ✓ Las tasas de error de los modelos construídos, obtenidos mediante la metodología validación cruzada k-fold, con k=10, son menores a 9% para el Modelo Boosting, menores al 15% para el Modelo Random Forest y menores a 25% para el Modelo Árbol de decisiones CART

- ✓ Las variables(top 3) que figuran en la tabla de importancia de variables del modelo boosting son: Tipo de vía de ocurrencia de del accidente(carretera), tipo de vehículo mayor involucrado(camión-combi) y desacato a la señal de tránsito por el conductor; para el modelo Random forest son: Tipo de vehículo mayor involucrado(camión-combi), Tipo de transporte(público) y Accidente de tránsito por atropello; finalmente para Árbol de decisiones son: Tipo de vía de ocurrencia de del accidente(carretera), tipo de vehículo menor involucrado(mototaxi) y por invasión en el carril contrario.

RECOMENDACIONES

- ✓ Modelar usando la técnica de redes bayesianas o Ecuaciones estructurales, los cuales son usados en estudios de tipo sociales; también permitiría ser interpretado visualmente, lo cual facilitaría su transmisión.

- ✓ Construir modelos logísticos pero con enlaces asimétricos, como Cloglog, Skew probit, Scobit, etc. Los cuales muchas veces presentan mejores indicadores que los modelos Logit tradicionales con enlace estándar 0.5.

- ✓ Construir modelos por subpoblaciones, con el objetivo de obtener resultados más finos y que permitirá tomar decisiones más acertadas.

REFERENCIAS BIBLIOGRÁFICAS

[1] Hernández V. (2012) “Análisis exploratorio espacial de los accidentes de tránsito en Ciudad Juárez”; México.

[2] Klaus M. (2013) “Roadway accident risk predictionbased on bayesian probabilistic networks”;Alemania.

[3] Bahar D., Arenas, B., Mira J. (2015) “Metodología desarrollada para la selección de predictores significativos que explican fatales accidentes de carretera”;España.

[4] Randa, M., López G y Garach L. (2015) “Bayes classifiers for imbalanced traffic accidents datasets”

[6] Iglesias T (2013) “Métodos de Bondad de Ajuste en Regresión Logística”;Universidad de Granada.

[7] Fuentes L., Contreras J., Alonso E., Martínez L. (2014) “Punto de corte óptimo para el diagnóstico confirmatorio de hiperfenilalaninemias por HPLC”.Centro Nacional de Genética Médica. La Habana, Cuba.

[8] Manish S. (2016)”Practical Guide to deal with Imbalanced Classification Problems in R”, Blog: Analytics Vidhya.

[9] Martínez P. (2008) “Métodos Estadísticos para Diagnósticos Médicos”; Barcelona: Universidad de Barcelona.