Modelo predictivo de accidentes de tránsito

John Fredy Redondo Morelo

Informe de práctica presentado para optar al título de Ingeniero de Sistemas

Asesora Interna Diana Margot López, Ingeniera de Sistemas UdeA
Asesor Externo Omar Alejandro Hurtado Rincón

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería de Sistemas
Medellín, Antioquia, Colombia
2023

Cita	(Redondo Morelo, 2023)
Referencia **Estilo IEEE (2020)**	[1] Redondo Morelo, J. F. (2023). *Modelo predictivo de accidentes de tránsito,2023*[Informe de práctica]. Universidad de Antioquia, Medellín, Colombia.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Rector: John Jairo Arboleda Céspedes.
Decano/Director: Jesús Francisco Vargas Bonilla.
Jefe departamento: Diego José Luis Botia Valderrama

Tabla de contenido

Lista de tablas

Lista de figuras

Dedicatoria

En honor a todas aquellas personas que a causa de un accidente de tránsito han perdido algún ser querido o han tenido una lesión de por vida.

Agradecimientos

Agradezco a Dios, mi madre y mi abuela por el apoyo brindado durante el transcurso de toda la carrera, a los distintos profesores que tuve durante mi pregrado y a mi asesora Diana Margot un agradecimiento especial quien me aportó mucho con sus conocimientos al desarrollo de este informe.

Resumen

Los accidentes de tránsito son un problema para todos los países del mundo y es que además de ser una de las causas principales de muerte a nivel mundial, también generan gastos a todos los gobiernos y sin dejar a un lado las secuelas graves que ocasionan a los involucrados desde el punto de vista físico, psicológico y económico.

Durante el desarrollo del presente trabajo se realizó un análisis de las variables cualitativas asociadas al conductor y de las variables cuantitativas asociadas al vehículo, con la finalidad de implementar un modelo de predicción de accidentes de tránsito teniendo en cuenta las anteriores variables.

Para comenzar lograr plantear un modelo predictivo de accidentes de tránsito primero se realiza una minuciosa investigación de los anteriores modelos planteados y de las variables de entrada que tienen en cuenta, luego de esto se plantea el posible modelo y por último se selecciona los métodos de aprendizaje de máquina que se utilizaran en el entrenamiento del modelo.

Luego de realizar una limpieza de la data utilizada, de plantear el modelo y luego de entrenarlos, encontramos que dicho modelo no converge a predicción exacta de un accidente de tránsito por lo cual se deja abierta la posibilidad para trabajos futuros la validación de dicho modelo teniendo en cuenta otras variables de entrada a este.

***Palabras clave*: accidentes de tránsito, modelo predictivo, aprendizaje de máquinas**

Abstract

Traffic accidents are a problem for all countries in the world and it is that in addition to being one of the main causes of death worldwide, they also generate expenses for all governments and without leaving aside the serious consequences that they cause to be involved from the physical, psychological and economic point of view.

During the development of this work, an analysis of the qualitative variables associated with the driver and the quantitative variables associated with the vehicle was carried out, in order to implement a traffic accident prediction model taking into account the previous variables.

To begin to achieve a predictive model of traffic accidents, first a thorough investigation of the previous models proposed and the input variables that they take into account is carried out, after this the possible model is proposed and finally the learning methods are selected. machine that will be used in training the model.

After cleaning the data used, proposing the model and after training them, we found that said model does not converge to the exact prediction of a traffic accident, for which the possibility of validation of said model is left open for future work. account other input variables to it.

Keywords: traffic accidents, predictive model, machine learning

I. Introducción

La masificación de los automóviles en los últimos años ha producido una serie de transformaciones sociales, que han dado lugar a un profundo cambio en la vida contemporánea de cada individuo, creando en los países una verdadera convulsión física, psíquica, ecológica, económica y cultural. Los vehículos han servido para estimular la creación de múltiples infraestructuras e incluso para transformar los procesos sociales y las comunicaciones, sin embargo, paralelamente se han desatado una serie de efectos adversos como la contaminación ambiental, un deterioro de los medios urbanísticos, el ruido y el aumento de accidentes de tránsito. Este último quizás, el mayor efecto debido a sus consecuencias en términos de víctimas, incapacidades, invalidez, costos asistenciales y económicos. La edad de las personas involucradas en los accidentes de tránsito, son en su gran mayoría jóvenes o personas en edad productiva, afectando la economía de los países y la calidad de vida de las personas. Al evaluar el impacto a largo plazo de los accidentes de tránsito y tomando en cuenta el correspondiente número de heridos, se pronostica que para el 2020 los choques en la vía alcanzarán el tercer lugar en la tabla de muerte e incapacidad a nivel mundial, apenas detrás de las depresiones clínicas y las enfermedades cardíacas, pero por encima de las infecciones respiratorias, la tuberculosis, la guerra y el VIH **[1]**.

Estos altos índices de muertes resultan impactantes al conocer que el 45% de los muertos en tránsito en Colombia, tienen edades entre los 15 y los 34 años; además que el promedio de años perdidos por esta causa es de 41 años. Desde los presupuestos para salud, pasando por los daños de los vehículos y el tiempo del personal involucrado en atender estos siniestros, los accidentes de tránsito cuestan millones de dólares a cualquier nación. Según el análisis realizado por la Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja sugiere que los accidentes de carretera cuestan un mínimo del 1% del producto interno bruto- PIB de cualquier país **[2]** En Colombia, según estadísticas del Ministerio de Transportes y Tránsito de 2016, esta es la segunda causa de muerte después de la violencia. De cada tres heridos en accidentes de tránsito, dos son hombres y siete de cada 10 muertos, son peatones.

Hace algunos años, se consideraba que solo la prevención de los accidentes viales era responsabilidad de las entidades de tránsito y transporte. Sin embargo, ese punto de vista ha cambiado a tal punto que las aseguradoras, están preocupadas por la implementación de un modelo de “mejoramiento vial” en el que se pueda intervenir un gran número de factores y causas de los accidentes de tránsito. La identificación de los peligros, la evaluación de los riesgos y el cálculo aproximado de los costos de las pérdidas que se presentan en la compañía por causa de los incidentes y accidentes de tránsito, les permitirá a las directivas hacer un cálculo del costo versus beneficio para implementar con éxito este modelo y alcanzar el objetivo de disminuir la accidentalidad vial.

II. Objetivos

A. Objetivo general

Crear un modelo predictivo de accidentes de tránsito a partir de la clasificación cualitativa y cuantitativa del conductor, vehículo y entorno de movilización.

B. Objetivos específicos

- Diseñar una clasificación de conductores de acuerdo a los hábitos de conducción, factores fisiológicos, cumplimiento de la ley y jornadas de trabajo para conocer su grado de riesgo al volante.

- Implementar un modelo de alertas proactivas al cliente, sobre los riesgos intrínsecos de los vehículos.

- Identificar información sobre los conductores más propensos a sufrir accidentes, a partir del modelo de clasificación diseñado. Esta información será el insumo básico para la construcción de una cultura de autocuidado y cultura vial.

- Construir un modelo de información personalizada de hábitos de conducción, compra y cuidado del vehículo a fin de generar venta cruzada de productos y servicios ofrecidos en las aseguradoras y facilitar los procesos de entrenamiento sistemático.

III. Planteamiento del problema

Según el código nacional de tránsito, un accidente de tránsito es un evento generalmente involuntario generado al menos por un vehículo en movimiento, que causa daños a personas y/o bienes e igual afecta la circulación de los vehículos que se movilizan por la vía.

Generalmente está determinado por condiciones y actos irresponsables potencialmente previsibles, atribuidos a factores humanos, vehículos, condiciones climatológicas, señalización y caminos, los cuales ocasionan pérdidas prematuras de vidas humanas y/o lesiones, así como secuelas físicas o psicológicas **[3]**

Sin embargo, en el contexto de los seguros, un accidente de tránsito, es un acontecimiento que materializa el riesgo de conducir un vehículo de transporte (Automóvil, motocicleta, bicicleta, entre otros), el cual origina daños materiales y humanos que están garantizados en una póliza hasta determinada cuantía, obligando a la aseguradora a restituir, total o parcialmente, al asegurado o a sus beneficiarios, el capital garantizado con el contrato de seguros **[4]**.

Actualmente para una gran parte de las aseguradoras del país, el conductor es el principal responsable de las tragedias automovilísticas con una participación del 83% de los casos; con un 8% de las causas están relacionadas con la vía y el entorno y el 9% con causas asociadas al automóvil **[4]**, identificándose como principales causas de los accidentes:

- La impericia del conductor.
- El exceso de velocidad.
- Exceso de confianza.
- Distracciones.
- Fatiga.
- Uso de medicamentos contraindicados.
- Fallas mecánicas.
- Infracciones de las señales de tránsito.
- Consumir licor antes y durante se conduce, entre otros.

Factores en los cuales hoy en día se trabaja en algunas de las compañías aseguradoras, desde la parte de fomento de la cultura para prevenirlos a través de capacitaciones impartidas. Sin embargo, actualmente no se lleva ningún registro de estas variables asociadas al conductor. En el caso del vehículo se cuenta con información calificativa a través del historial de revisión técnicas que se le han realizado al vehículo en los centros de servicio asociados a algunas aseguradoras.

IV. Marco Teórico

Con el objetivo de entender la evolución de la teoría respecto a la siniestralidad y a la vez de tener un marco de referencia a seguir en el desarrollo de un modelo predictivo el cual se abarca en el desarrollo del proyecto; se mostrarán de manera cronológica modelos teóricos que son usados por la mayoría de los investigadores que estudian los accidentes de tránsito.

A. Modelo de Smeed

Smeed identificó basándose en el análisis de una muestra de accidentes de tránsito de 20 países; una relación exponencial inversa entre las defunciones por vehículo y el número de vehículos per cápita. Posteriormente, en 1949, a las variables inicialmente tenidas en cuenta por Smeed en su estudio inicial al modelo se le introducen nuevas relaciones tales como: el nivel de riqueza nacional, la densidad vehicular, el número de habitantes por cama de hospital y la cantidad de habitantes por médico. La formulación general del modelo de Smeed es la siguiente:

$$D = .0003(np^2)^{\frac{1}{3}}$$

o, expresada per cápita:

$$\frac{D}{p} = .0003 \times \sqrt[3]{\frac{n}{p}}$$

Donde ***D*** es la cifra anual de muertes en el tránsito, ***n*** el número de vehículos registrados y ***p*** la población total. La ecuación permite predecir que, al duplicarse la tasa de motorización de un país (el número de vehículos per cápita), se producirá un incremento de 26% en la tasa de mortalidad por habitante y una disminución de 37% en la tasa de mortalidad por vehículo. El modelo fue reformulado varias veces evolucionando paulatinamente hacia una teoría general de la siniestralidad vial que Smeed resumió en su célebre afirmación: “El número de accidentes mortales de un país corresponde al número que el país está dispuesto a tolerar” **[5]**.

B. Modelo DRAG

Desarrollado en 1984, por M. Gaudry el cual dio a conocer un sofisticado modelo predictivo de la siniestralidad vial denominado ***DRAG***, abreviatura de "***Demande Routière des Accidents et de leur Gravité***", el cual emplea 40 variables clasificadas en 7 categorías que influyen en la producción de siniestros y víctimas. Entre las variables de mayor influencia se incluyen el consumo de combustible, los límites de velocidad legalmente establecidos, las tarifas del transporte público, la obligatoriedad del casco en los motociclistas y del cinturón de seguridad y la normativa legal sobre el alcohol y otras drogas con la conducción de vehículos.

Una característica fundamental del DRAG es el desarrollo de una estructura multi-capa que integra las tres dimensiones principales de la inseguridad: exposición, frecuencia y severidad, cada una de las cuales es objeto de una ecuación propia que considera la influencia de las variables mencionadas anteriormente **[6]**. Para estimar ciertas propiedades y relaciones de las variables independientes que resultan muy difíciles de capturar, el ***DRAG*** se caracteriza por el empleo de la llamada Transformación de ***Box-Cox***.

Este recurso matemático hace que como herramienta sea muy útil para el estudio de la siniestralidad vial, ya que muchas veces el modelizador no tiene modo de saber a priori como deben especificarse ciertas funciones de comportamiento o de riesgo. El DRAG se ha ramificado constituyendo en la actualidad una verdadera subfamilia modélica derivada del original, cuyos componentes son adaptaciones especiales para ciertos países o regiones donde se les da un nombre distinto, p. ej. SAAQ (Quebec), SNUS (Alemania), DRAG-Stockholm (Suecia), TAG (Francia), TRAVAL (California), TRULS (Noruega) y DRAG-España (España).

C. Modelo de cambios de velocidad

En este modelo se consideran las tres dimensiones del problema de la seguridad vial: *exposición, riesgo y consecuencias*, fue desarrollado por G. Nilsson, en su tesis doctoral presentada en la universidad sueca de Lund en el año de 2004, Nilsson articuló un modelo predictivo conocido como "*Power Model*". El modelo se centra en el factor velocidad promedio del tránsito, el cual le permite describir la situación de seguridad de acuerdo con esta variable a fin de servir de herramienta para estimar predictivamente los efectos de los cambios aislados, así como otras importantes relaciones concernientes a los cambios en el riesgo de siniestro y sus consecuencias destructivas. Dicho de otro modo: el cambio de la energía cinética es usado por el Power Model para explicar el cambio en el riesgo y las consecuencias, o sea, en el número de fallecidos y de lesionados dado que ambas dimensiones tienen una fuerte relación con el cuadrado de los cambios de velocidad relativa.

Este modelo permite predecir los efectos de los cambios de velocidad promedio de la circulación sobre la seguridad y puede ser empleado para aislar el efecto de los cambios de la misma en relación con otras medidas o cambios. Asimismo, tiene una fuerte relación con la energía cinética pues las fuerzas del tránsito estrictamente no dependen de la velocidad en sí misma sino de sus variaciones, es decir, de las aceleraciones y deceleraciones de los vehículos debido a que, de hecho, las colisiones y atropellos no constituyen sino un subconjunto de inesperadas deceleraciones tan importantes como la velocidad o sea que, la probabilidad de lesiones o muertes entre los usuarios involucrados es tan alta como las deceleraciones.

El modelo fue evaluado y refinado por Elvik, estableciendo que los cambios en el número de lesiones causadas por siniestros resultantes del cambio de la velocidad promedio puede ser descritos con la siguiente ecuación:

$$\left(\frac{\text{Siniestros (antes)}}{\text{Siniestros (después)}} = \frac{\text{Velocidad (antes)}}{\text{Velocidad (después)}} \right)^{2.0}$$

Para obtener el número esperado de usuarios con lesiones graves es necesario emplear el exponente 3.0 y para los fallecidos 4.5. Una prueba práctica de la validez y la utilidad del modelo es

que actualmente se emplea en varios países desarrollados como herramienta estándar de análisis y planificación, v. gr., España, Dinamarca, Noruega y Suecia.

D. Modelo RIPCORD-ISEREST

Desarrollado por la Unión Europea, donde la formulación del modelo es:

$$E(\lambda) = \alpha Q_{MA}^{\beta} Q_{MI}^{\beta} e^{\sum \gamma_i x_i}$$

Donde el número estimado esperado de siniestros ***E (λ),*** es una función del volumen de tráfico ***Q*** y de un conjunto de factores de riesgo, ***xi (i = 1, 2, 3, …, n).*** El efecto del volumen sobre los siniestros es modelado en términos de una elasticidad que es una potencia ***β***, a la cual el volumen de tráfico es incrementado. Los efectos de varios factores de riesgo que influencian la probabilidad de siniestros, a una cierta exposición dada, es modelada como una función exponencial, que es como es (la base de los logaritmos naturales) incrementada a la suma del producto de coeficientes ***γi***, y los valores de variables ***xi*** que denotan factores de riesgo. Como puede verse, las variables consideradas son el volumen de tránsito y los factores de riesgo.

E. Aplicación del modelo de predicción de accidentes viales (CPM) del HSM para evaluación de seguridad en segmentos de carreteras de dos carriles

La entidad federal estadounidense *American Association of State Highway and Transportation Officials*, AASHTO, en el 2010 publicó la primera edición del manual de seguridad Vial **[7]** en el cual propuso un modelo que es considerado como el nuevo paradigma de la predicción de siniestros viales. Los paradigmas anteriores de modelos predictivos de accidentes viales eran de naturaleza puramente descriptiva basándose en los datos históricos de la frecuencia de las colisiones, de la tasa de las mismas y de los daños a la propiedad. En cambio, el paradigma propuesto es cuantitativo, pues predice el número esperado de colisiones en función de las características geométricas y operativas de las vías, las condiciones existentes y las futuras condiciones proyectadas, así como los diseños alternativos que se proyectan aplicar.

Mediante ecuaciones de regresión el modelo predice el número promedio de colisiones por año para un sector vial determinado, una intersección o un segmento de vía como una función del volumen de tráfico, la cual típicamente no es una relación lineal. Incluso permite cuantificar el cambio esperado de las colisiones producidas en un sitio dado por la implementación de un tratamiento de ingeniería particular de contramedidas puntuales de reducción de los factores de colisión (CRF). Como por ejemplo la instalación de señales de tránsito en una intersección controlada por señal de “Pare”, el HSM ha dado lugar a muchos estudios de accidentalidad por su adaptabilidad a los entornos locales, p. ej. Aplicación del HSM en camino rural de dos carriles en Brasil **[8]** uso del módulo de predicción de accidentes (CPM) del HSM para la evaluación de seguridad en segmentos de carreteras de dos carriles en Colombia **[9]**.

V. Metodología

Para el desarrollo del modelo de clasificación de gestión de riesgos para la prevención de incidentes y accidentes de tránsito se seguirá la metodología planteada por Fayyad, Piatetsky-Shapiro y Smyth nombrada como *Knowledge Discovery in Databases* (KDD) **[10]**, cuyas etapas permiten la aplicación apropiada de técnicas de minería de datos en cualquier proyecto o investigación relacionada a esta disciplina, aumentando la probabilidad de éxito de los proyectos, como también el desempeño y confiabilidad de los modelos generados. El KDD consta principalmente de cinco etapas las cuales son selección, preprocesamiento, transformación, minería de datos, interpretación y evaluación. A continuación, se describen los cuatro pasos y algunas consideraciones que deben tenerse en cuenta al aplicar esta metodología al proyecto actual.

A. Selección

Se seleccionan las variables y registros con los que se trabajará. Teniendo en cuenta que estas deben de ser:

- Potencialmente explicativas del fenómeno en estudio.
- Contar con nulo o poco error de registro
- Estar disponibles en el futuro si se vuelve a realizar el análisis.
- Medibles antes que ocurra el evento en estudio.

En este estudio se realizará inicialmente una exploración de los datos, de esta manera se podrá detectar inicialmente las posibles variables que sean predictores de la accidentalidad vial, se obtendrá información respecto a las escalas y tipos de datos, detección de valores extremos, sesgos en los datos y qué tan dispersos están.

B. Preprocesamiento

Una vez obtenido el conjunto de variables y registros seleccionados deben estar libres de ruidos para no generar sesgo durante su procesamiento. Para esto se tienen en cuenta los siguientes criterios:

Completitud: No existencia de datos faltantes (missing values) en los atributos.

Consistencia: El formato y codificación en un atributo en particular deben ser idénticos para todos los registros.

Coherencia: Los datos deben responder a reglas lógicas básicas según el contexto de la base. Se evalúan las observaciones que distan mucho de otras.

Validez: Los registros deben ser coherentes con la organización y/o actividad que se documenta.

Con el objetivo de que los datos cumplan con los 4 criterios, se evaluará los siguientes tratamientos a los datos vacíos:

Reemplazo Missing Values por Moda y Media: Los valores vacíos pueden ser reemplazados por la moda (variables categóricas) o media (variables numéricas).

Reemplazar Missing Values por Modelo Predictivo: Los datos vacíos o outliers pueden ser reemplazados a través del uso de modelos predictivos tales como máquinas de aprendizaje que permitan la predicción del valor más posible para el dato, considerando los otros datos que sí tienen valor.

Reducir Registros: En el caso que la cantidad de registros con valores vacíos sea muy pequeña en comparación con la base, los registros pueden ser no considerados para la base de entrenamiento. Así, se asegura que los datos usados para el entrenamiento y validación reflejan el comportamiento real del fenómeno en estudio.

Para este trabajo se decide utilizar la reducción de la base de datos, seleccionando los registros en donde ningún atributo es nulo o missing value.

C. Transformación

En la etapa *Minería de Datos* se aplican muchos modelos cuyos requisitos son que todas las variables sean numéricas. Sin embargo, la realidad es otra, existen datos que por naturaleza no cumplen esta condición, tales como la variable de texto categórica que almacena la ciudad de residencia de un propietario de vehículo. Se realiza el siguiente tratamiento a cada variable:

1. Variables Binomiales: Para este tipo de datos los valores posibles son solamente dos. De esta manera, la solución de transformación es asignar un valor numérico a cada categoría *n* de las variables evaluadas.

2. Variables Polinomiales: En este caso los valores posibles son generalmente más de dos categorías. La forma en que se deben transformar a números es generando $n - 1$ atributos binomiales, donde *n* es la cantidad de categorías.

Además de que los datos deben ser numéricos, algunos algoritmos de minería requieren que estos estén normalizados para obtener resultados eficientes. Además, la normalización soluciona el problema de las diferencias generadas en los rangos y medidas que ocupan las variables, como por ejemplo la edad y los hábitos de conducción.

Las bases de datos utilizadas en este estudio contienen datos del tipo binomial y polinomial, por lo que las transformaciones anteriormente explicadas serán aplicadas según corresponda. Así mismo, estos se normalizan para que los algoritmos sean implementados eficientemente.

D. Minería de Datos

Con los datos ya capturados, se procede a sus estudios y procesamiento (variables dummy, escalado de valores, entre otros) para adaptarlos y poder calcular regresiones y redes neuronales a partir de estos. Para este proceso se utiliza Python (Python Software Foundation, 2017) ya que se trata de un lenguaje de programación de alto nivel, lo cual permite desarrollar rápidamente los scripts necesarios. Como editor de código se utiliza Kate. Para este trabajo se ha decidido aplicar las siguientes máquinas:

1. Artifical Neural Net (ANN),
2. Logística Regresión (LR).

El primer modelo que se construyó se calcula aplicando regresión lineal sobre los datos, utilizando la función lm () de Para la edición de código se utiliza el IDE RStudio.

Luego, se procede a calcular redes neuronales artificiales, aplicando validación cruzada (cross validation) para validar los modelos, dividiendo los datos disponibles en dos conjuntos: conjunto de entrenamiento y conjunto de test. Se utiliza la raíz cuadrada del error cuadrático medio (RMSE) para comparar los modelos. En conclusión, los desempeños de los modelos serán evaluados principalmente en dos criterios: Exactitud (accuracy) y Costo de Clasificación. Para medir ambas métricas se hará uso de:

Validación Cruzada: Se utiliza para medir el desempeño de los modelos de minería de datos aplicados en cualquier proyecto. El objetivo principal es utilizar la misma base de datos para generar el modelo de predicción y posteriormente evaluarlo y así obtener una proyección del desempeño de predicción. La forma de validar el modelo se realiza a través de la división de la Base de Datos en dos, siendo una de ellas la de entrenamiento y la otra de testeo. Normalmente, el 70% del total de los registros es utilizado para el entrenamiento, mientras que el 30% restante es para la medición del desempeño.

E. Interpretación y Evaluación

En esta etapa se evalúa el desempeño de los modelos aplicados en la etapa anterior. Además, se visualiza e interpreta los patrones que el (los) mejor(es) modelo (s) entregan. En esta etapa el juicio de experto juega un rol fundamental, ya que se deberá evaluar si los patrones extraídos tienen sentido en el contexto que fueron aplicados. Cabe destacar que, en esta etapa, al igual que las anteriores, existe la posibilidad que se decida volver al primer paso o a una etapa previa según corresponda. Los desempeños de los modelos pueden variar por muchos factores, tales como las variables escogidas y el manejo que se hizo a los datos. Por lo tanto, se hace obligatorio utilizar mecanismos para evaluar el desempeño de cada uno de ellos. De esta manera el patrón identificado tendrá un sustento más objetivo.

Finalmente se obtienen las conclusiones y se establece un plan de trabajo a futuro con el que mejorar los resultados obtenidos.

.

VI. Desarrollo del Producto

Para llegar al desarrollo del modelo predictivo se ha de seguir los pasos según la metodología propuesta en el modelo KDD. Durante los siguientes apartados se explicará de donde se obtuvieron los datos, la descripción de la tecnología utilizada en el desarrollo del modelo. Luego, se estudiará y se transformarán los datos para luego aplicar regresiones lineales y redes neuronales sobre los mismo:

A. Tecnología.

Para el desarrollo del proyecto se utilizaron dos lenguajes de programación: Python y R. Para trabajar con Python se utilizó SublimeText, mientras que para trabajar con R se ha utilizado RStudio. A continuación, se da una mejor explicación de cada tecnología.

Lenguaje Python

Es un lenguaje de programación interpretado, multiplataforma y de alto nivel. Es libre y gratuito, funciona en todas las plataformas. Se ha elegido Python para este trabajo por ser gratuito, multiplataforma y de alto nivel. Soporta multitud de lenguajes y, en el caso de Python, permite ejecutar el código directamente desde el editor

Kate como editor de Python

Sublime Text es un editor de código avanzado. Permite ejecutar el código directamente desde el editor.

Lenguaje R

R es un lenguaje y entorno orientados al análisis estadístico y visualización de datos. R ofrece una amplia variedad de técnicas estadísticas y de visualización de datos y está disponible como software libre bajo los términos de Free Software Foundation’s GNU General Public License (GNU, 2017) y para la mayoría de sistemas UNIX, Windows y Mac OS

RStudio

R Studio es un entorno de desarrollo integrado (IDE por sus siglas en inglés) para R. RStudio puede ser utilizado como aplicación de escritorio en los tres sistemas operativos principales, así como en un navegador web para accesos remotos. Posee una licencia Open Source para fines no comerciales.

B. Selección de Datos

Como primer paso se realizó un preprocesamiento de los datos, eliminando redundancias e inconsistencias para luego normalizar los datos para facilidad de las tareas posteriores.

En el desarrollo del modelo se utilizaron tres bases de datos, una simulada (Conductor), donde se simula la información de las personas que presentan aseguramiento por parte de alguna aseguradora y que hayan o no presentado siniestralidad automovilística, otra de las bases de datos utilizada (Vehículo) es la cual se encuentra información técnica de las revisiones de los vehículos que son asegurados y por último los datos públicos de accidentalidad de la secretaría de tránsito y transporte del municipio de Medellín (Entorno de movilidad). Estos datos se encuentran almacenados en planillas de Excel. Para el estudio se analizaron dos años de siniestralidad.

Con el propósito de buscar un primer modelo se separaron en tres tipos de entidades causantes de siniestros de tránsito, cada una con sus respectivos atributos tales como:

- **Conductor:** Identificador del accidente, sexo, edad, experiencia conduciendo, consumo de drogas, medicamentos, factores psicológicos, horas de trabajo, consumo de alcohol, horas de sueño, agresividad, comportamiento vial.
- **Vehículo:** Identificador del accidente, servicio, tipo de vehículo, neumáticos, dirección, suspensión, frenos, alumbrado, cinturón de seguridad, airbag, casco.
- **Entorno de Movilidad:** Identificador del accidente, fecha, hora de accidente, lugar, causas, tipo de accidente, estado atmosférico, condición de la vía, tipo de vía, urbano/rural.

Se debe de tener en cuenta que solamente se utilizaron aquellos registros de las bases de datos que estuvieron "completos" para el estudio, ya que el eventual reemplazo de valores vacíos puede generar ruido en el desempeño del modelo. Por tanto, se tuvieron en cuenta aquellos datos que reflejen las posibles causas de siniestralidad vial. En otras palabras, se eliminaron de las bases de datos los registros que contienen al menos una variable nula y fueron identificados como outliers a través del análisis estadístico de rangos.

C. Preprocesamiento y Transformación

Las variables categóricas tales: Identificador del accidente, servicio, tipo de vehículo, neumáticos, dirección, suspensión, frenos, alumbrado, cinturón de seguridad, airbag, casco, entre otras fueron transformados a números para ser trabajados en las técnicas de minería de datos. Tales como sexo Hombre 0 Mujer 1, Presenta Siniestralidad 0 No presenta Siniestralidad 1, entre otras.

De esta manera, para cada atributo se generaron *en nuevas* columnas, donde *n* era la cantidad de distintas categorías únicas del atributo. De esta nueva columna se seleccionaron $n-1$, con el objetivo de eliminar problemas de multicolinealidad ya que las variables del tipo numérico tienen distintos rangos y esta diferencia puede generar problemas de desempeño y ruido al momento de aplicar los algoritmos de las técnicas de minería de datos. Por lo tanto, todas las variables numéricas fueron escaladas en un rango de 0 a 1 utilizando la siguiente fórmula:

$$x'_{ij} = \frac{x_{ij} - \min(\mathbf{x}_j)}{\max(\mathbf{x}_j) - \min(\mathbf{x}_j)}$$

Donde x'_{ij} es la nueva escala para el registro, la variable j − ésima del registro i − ésimo de la base de datos y xj el vector de datos de la variable j − ésima. Esta fórmula es denominada técnicamente como normalización de rango 0-1 **[11].**

Una vez analizados y tratados los datos como se deseaban, se cruzan para obtener un único "data frame" con información de los siniestros de tránsito según la información contenida en las bases de datos del conductor, vehículo y entorno de movilidad. Tarea que se desarrolla en Python. Como resultado de esta tarea se ha obtenido campos con información útil para la construcción del modelo y otros que son meramente informativos. Pero se debe tener en cuenta que se tiene una fila por cada accidente producido, con toda la información enriquecida, pero es necesario realizar algunas modificaciones para poder aplicar regresiones y redes neuronales.

Como primer paso se agrupó los registros del fichero por los campos que se van a utilizar para determinar si se produce un accidente o no. Estos campos son los siguientes: sexo, edad, NormHour, Quarter, Prec, TMed, VelMedia. Este "data frame" contiene los datos de los siniestros de conductores en accidentes de tránsito enriquecidos con datos del vehículo y entorno de movilidad.

D. Minería De Datos

Para el estudio se aplicaron dos máquinas de aprendizaje: Regresión Logística y Redes Neuronales Artificiales. Para cada una se optimizó (Tabla No 1) la configuración paramétrica a través de la comparación del desempeño. Por último, se procede a dividir el fichero en dos, uno de ellos contiene lo datos que se van a utilizar para realizar el entrenamiento de la red neuronal y el otro se utilizará para realizar el test y comprobar el resultado obtenido

Máquina de Aprendizaje	Identificador	Grilla de Parámetros
Redes Neuronales	NN	Tamaño Capa Oculta: 1, 5, 10, Automático. Ciclos de Entrenamiento: 500, 1000, 1500.
Regresión Logística	LR	C: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

Tabla No 1

Configuración parámetros regresión logística y redes neuronales artificiales

Regresiones Lineales.

Como se ha detallado anteriormente, el proceso que se ha llevado hasta el momento ha sido la captura de datos, estudiarlos, manipularlos y transformarlos para que cumplan los requerimientos del estudio.

Teniendo ya los datos disponibles de la forma deseada, se aplicaron regresiones lineales con el objetivo de poder predecir la probabilidad de ocurrencia de un accidente de tránsito con las características de un conductor, vehículo y entorno de movilidad combinados.

Al aplicar una regresión lineal podemos conocer la ecuación de regresión la cual permite calcular predicciones de una variable a partir de los valores que tomen otra variable (s). El método de los mínimos cuadrados (MMC) es el procedimiento matemático que permite calcular la ecuación de la regresión y sus componentes.

En R se utiliza la función *lm* para realizar regresiones lineales. Para nuestro caso se carga en R el "data frame" con las probabilidades de que ocurran accidentes. Luego con la función lm, se realiza la regresión lineal utilizando la variable Probability como variable dependiente y las variables Sexo, Edad, NormHour, Quarter, Prec, TMed, VelMedia como predictoras. Para esto se realiza un script en Python.

Redes neuronales

En la sección anterior se ha comprobado que utilizar regresiones lineales con los datos disponibles no genera un modelo predictivo fiable. Ahora se va a crear una red neuronal para comparar el resultado. Para llevar a cabo esta parte del trabajo se ha utilizado R, concretamente el paquete neuralnet **[12]**. Y para validar este modelo se utiliza la técnica de validación cruzada la cual se expone en apartados anteriores de este documento. Con la función neuralnet se entrena la red neuronal estableciendo los siguientes parámetros:

- ***Fórmula***: se indica qué variables se quieren predecir y cuáles son las predictoras.
- ***Data:*** los datos que contienen las variables especificadas en la fórmula.
- ***Hidden:*** indica el número de capas ocultas o intermedias de la red neuronal y el número de neuronas de cada una de ellas.
- ***Threshold:*** valor numérico con el que, si el error obtenido es menor que dicho valor, el algoritmo se para y da por bueno el resultado.
- ***Stepmax:*** número máximo de iteraciones del algoritmo.
- ***Rep:*** número de entrenamientos de la red neuronal.
- ***Algorithm:*** algoritmo utilizado para calcular la red neuronal.
- ***Err.fct:*** función utilizada para calcular el error.

La red neuronal se entrena con la función *neuralnet*. Luego, con la función *compute* se aplica dicha red al set de datos que poseemos "data frame" de test. Además, se utiliza la raíz cuadrada del error cuadrático medio (RMSE por sus siglas en inglés) para poder comparar los modelos y determinar cuál de ellos es más preciso. Además, se realizan varias parametrizaciones moviendo los valores Hidden, threshold y stepmax, buscando la que sea la raíz cuadrada del error cuadrático medio sea cercano a cero.

VII. Resultados

Al analizar los datos obtenidos al aplicar la regresión lineal se puede apreciar que el valor de R2 es muy bajo 0.002646 (Figura No 1), lo cual indica que la variable dependiente Probability no sea muy bien explicada por las variables predictoras en el caso de la probabilidad de ocurrencia de un accidente de tránsito con ellas

Probability ~ Sexo, Edad, NormHour, Quarter, Prec, TMed, VelMedia

Ahora, si observamos en la *Figura No 1* podemos apreciar que la mayoría de los parámetros correspondientes a las variables predictoras no son significativos, como lo prueba el bajo valor que toma el estadístico t de Student para estos (columna t value).

```
Call:
lm(formula = Probability ~ Sexo + Edad + NormHour +
       Quarter + Prec + TMed + VelMedia, data = csvTest)

Residuals:
    Min      1Q  Mediam      3Q     Max
-0.02552 -0.01311 -0.01071 -0.00819 0.99280

Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 0.0064092   0.0028970   2.212    0.0270 *
Sexo        0.0022877   0.0024774   0.923    0.3558
Edad       -0.0001962   0.0030176  -0.065    0.9482
NormHour    0.0065520   0.0024414   2.684    0.0073 **
Quarter     0.0054484   0.0026451   2.060    0.0394 *
Prec        0.0090929   0.0063826   1.425    0.1543
TMed       -0.0072370   0.0029819  -2.427    0.0152 *
VelMedia   -0.0036080   0.0036290  -0.994    0.3201
---
Signif. codes: 0 '***' 0.001 '***' 0.01 '*' 0.05 '.'0.1''1

Residual standard error: 0.06446 on 8032 degress of freedom
Multiple R-squared:  0.002646, Adjusted R-squared: 0.001776
F-statistic: 3.044 on 7 and 8032 DF, p-value: 0.003372
```

Figura 1
Resultados regresiones lineales

Para confirmar la teoría de que el modelo no es confiable, se procede a analizar los gráficos obtenidos a partir de la aplicación de la regresión lineal, Figura No 2. donde se compara los valores observados (puntos negros) de accidentalidad y los valores de probabilidad de accidentalidad obtenidos al realizar la regresión lineal (puntos rojos). Por lo tanto, se reafirma que el modelo no predice correctamente debido al amplio margen de incompatibilidad de los resultados.

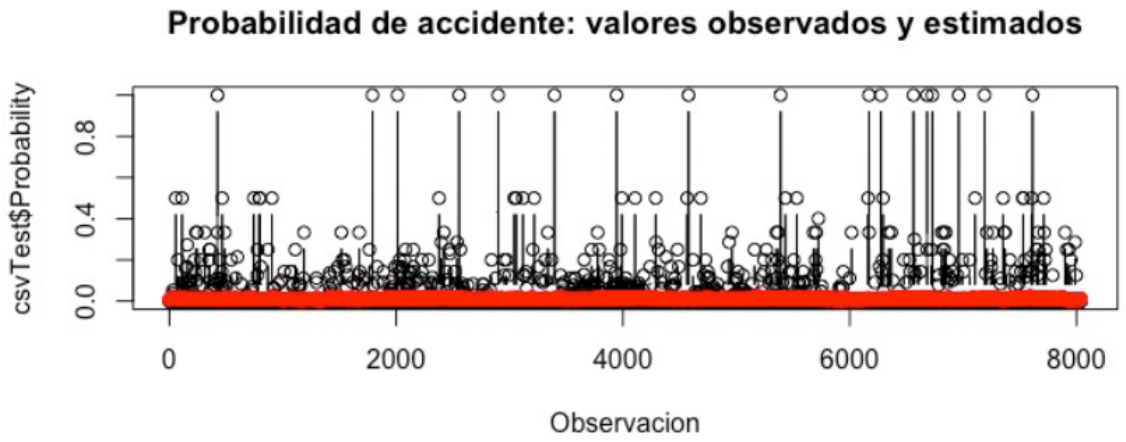

Figura 2
Probabilidad de ocurrencia de un accidente de transito

Ahora podemos analizar la Figura No 3, en donde se observan los residuos, es decir, las diferencias entre los valores observados de accidentalidad y los estimados para la variable (Probability) de accidentalidad. Para esto se debe tener claro que la hipótesis de tener media nula se debe de cumplir. Pero como se puede observar en el gráfico esto no se cumple, por lo que se tiene otro motivo más para confirmar la teoría de que el modelo no es bueno.

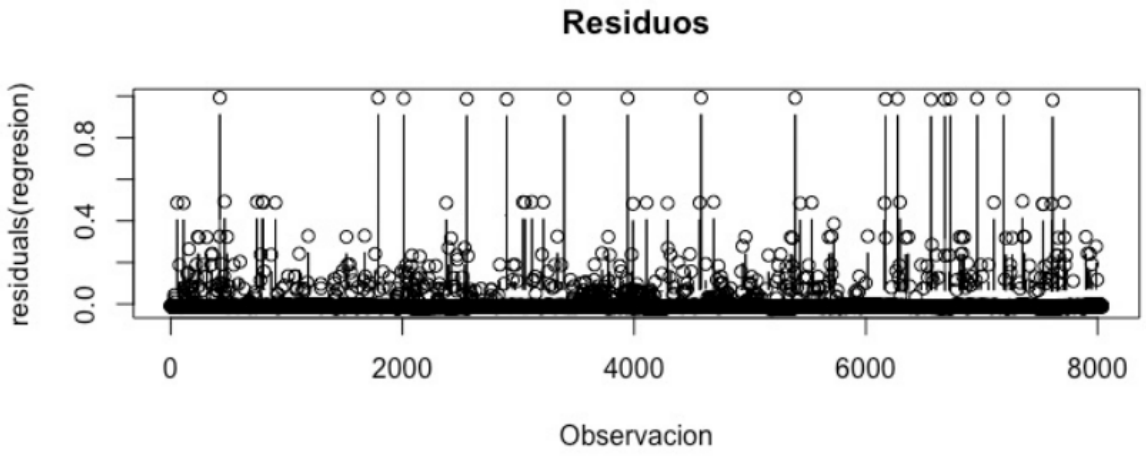

Figura 3

Diferencias entre la probabilidad de ocurrencia de un accidente de tránsito y valores estimados

Para obtener los resultados de aplicar la red neuronal se realizó las siguientes parametrizaciones obteniendo los resultados que se describen a continuación:

Parametrización Red Neuronal uno:

La fórmula utilizada en todas las parametrizaciones es: "Probability ~ Sexo, Edad, NormHour, Quarter, Prec, TMed, VelMedia". Sin embargo, se va a probar distintas configuraciones de la red neuronal. En primera instancia, los valores utilizados son los siguientes:

Hidden = 2 ; threshold = 0.01 ; stepmax = 1e5

El algoritmo utilizado que utiliza la función *neuralne*t es *resilient backpropagation* (rprop). Para esta parametrización el valor del RMSE de la red neuronal es de 0.05795262207. Lo cual es cercano a 0. Pero, no es confiable debido a que todos los valores estimados son cercanos al 0 y la mayoría de las probabilidades también lo son. Por tal motivo, el error en la mayoría de las observaciones es muy pequeño. Sin embargo, si se analizan únicamente las observaciones con probabilidad mayor que 0 para comprobar el error, se observa que el valor del RMSE obtenido es 0.2240609515, lo cual significa que el valor predicho es muy lejano al valor observado del modelo.

Parametrización Red Neuronal dos:

Para este caso se va a incluir una capa oculta más de tres neuronas para comprobar el efecto que produce en el resultado. La configuración es la siguiente:

Hidden=c(2,3) ; threshold=0.01 ; stepmax=1e5

En este caso el valor obtenido de RMSE para esta configuración es 0.1073068407. Al añadir una capa de neuronas al modelo se esperaba un resultado mejor que en la parametrización anterior. Pero el error obtenido es mayor, por lo que el modelo pierde precisión con respecto al obtenido con la parametrización 1.

Parametrización Red Neuronal tres:

Para esta parametrización y con el objetivo de obtener modelos más precisos que los anteriores, se procede a probar el algoritmo *backpropagation* con un learning rate de 0.0001 (Se fija este valor ya que para valores mayores de este parámetro el algoritmo no converge). Así, las cosas la parametrización en este caso es la siguiente:

Hidden = 0 ; threshold = 0.01 ; stepmax = 1e5 ;

Para la anterior configuración el RMSE obtenido fue de 0.05771381917, si bien es mucho mejor que en la parametrización dos, es similar al de la parametrización uno. Lo que nos induce a que el modelo que se obtuvo con la actual configuración no mejora notablemente los resultados.

Parametrización Red Neuronal Cuatro:

A fin de calibrar el modelo se procede a realizar de nuevo otra configuración. Se espera obtener un mejor resultado ya que las capas ocultas no son 0, lo que debería dar una precisión mayor comparado con las parametrizaciones anteriores. La configuración es:

Hidden = 2 ; threshold = 0.01 ; stepmax = 1e5 ;

Para la anterior configuración el RMSE obtenido fue de 0.05774828713, error que es similar al de la parametrización uno y cuatro. Lo que nos induce a que el modelo que se obtuvo con la actual configuración no mejora notablemente los resultados teniendo en cuenta que se utilizó una capa oculta de neuronas.

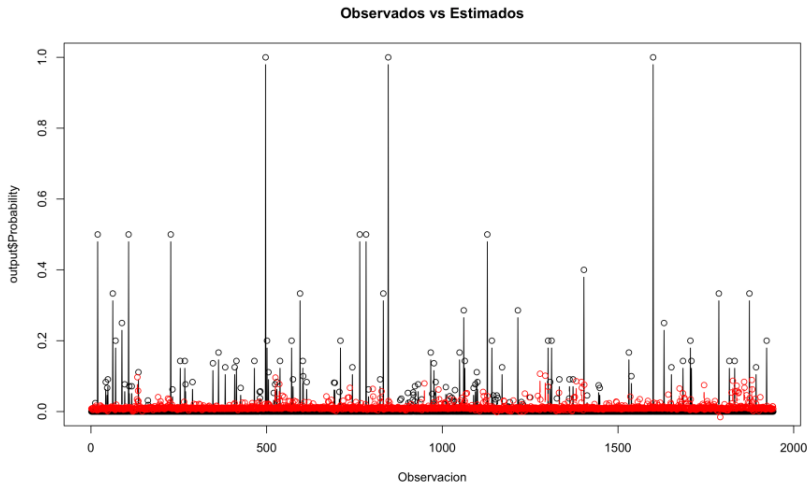

Figura 4
Comparación de valores observados vs estimados

Ahora analizando la Figura 4. podemos observar que tanto los valores de predicción de siniestralidad (puntos rojos) y observados (puntos negros) se acumulan principalmente cerca del valor 0, a lo cual el RMSE obtenido para el conjunto de todos los datos es bajo. Pero, si se centra la atención en las observaciones que toman un valor mayor que 0, se ve que los valores correspondientes de las estimaciones de siniestralidad (puntos rojos) no se acercan al valor de la observación. Por tal motivo, el RMSE para este conjunto de datos es mucho mayor. Lo cual nos indica que, para las variables en las que se puede producir un accidente, el modelo no predice con exactitud la probabilidad de ocurrencia del mismo.

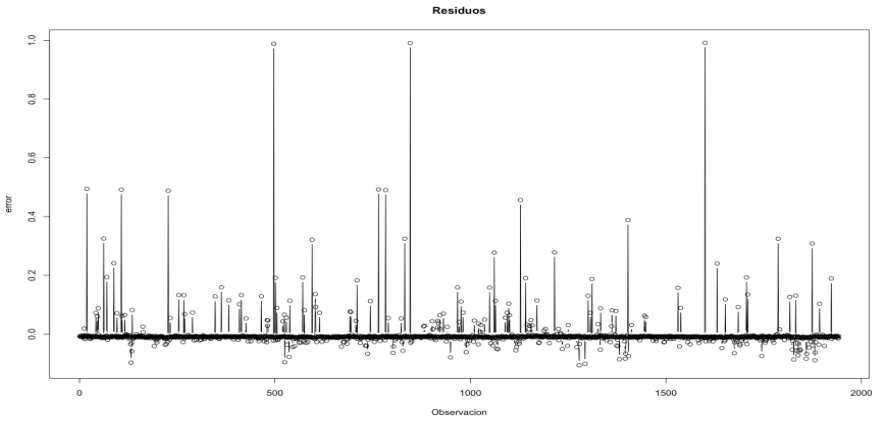

Figura 5
Comparación de valores observados vs estimados

Otras parametrizaciones de la Red Neuronal

Debido a que, con las parametrizaciones configuradas al modelo, se ajustaron el número de capas ocultas, bajando el threshold y aumentando el número máximo de iteraciones, se esperaba mejorar los resultados, debido a que al bajar el threshold, el error obtenido debería ser menor. Lo que se obtuvo fue que en muchos casos el algoritmo no convergerá a un resultado en el que el error sea menor o igual que el threshold configurado se pueden concluir las siguientes premisas:

Productos próximos

Se puede inferir que uno de los motivos por los que los resultados obtenidos no son excesivamente positivos puede deberse a que las variables elegidas para la predicción probablemente no sean las más adecuadas, ya que la correlación lineal existente entre las variables predictoras y la variable dependiente no es alta

VIII. Conclusiones

Muchas de las parametrizaciones que se realizaron al modelo utilizando SVM no han llegado a converger a un error inferior al threshold elegido en el máximo de iteraciones del algoritmo; lo cual nos dice que el entrenamiento no ha sido satisfactorio. Así las cosas, el resultado obtenido tanto en la regresión lineal como en las parametrizaciones es muy similar lo que nos lleva a pensar que se debe de hacer un mejor análisis en la selección de las variables que expliquen el comportamiento de los accidentes de tránsito.

El resultado obtenido tanto en la regresión lineal como en las parametrizaciones es muy similar, el RMSE obtenido con las predicciones es aproximadamente 0.22, una cifra considerablemente peor que el RMSE obtenido con el total de los datos. Esto indica que, para los casos en los que se puede producir un accidente, el modelo no predice con exactitud la probabilidad de ocurrencia del mismo.

Se debe tener en cuenta que la mayoría de analistas consideran que los modelos lineales generales en sus alternativas de Poisson y de binomial negativa son los más adecuados para la construcción de modelos de predicción de frecuencia de accidentes de tráfico (Blanca Arenas, 2008). Por lo cual se debe de tener en cuenta esta técnica para la construcción de modelos predictivos con el objetivo de obtener una mayor precisión en estos junto con el ajuste de dichos modelos mediante el tratamiento de outliers.

Se desarrollaron por parte del equipo habilidades de investigación para la recopilación del material necesario para el desarrollo del proyecto, tales como organización, reflexión, resumir, entre otras.

Referencias

[1] Cañas Zea Oscar E. Análisis Estadístico de la accidentalidad vial en la ciudad de Medellín. Universidad Nacional de Colombia. Tesis especialización en Estadística. Medellín, 2000.

[2] Organización de las Naciones Unidas, ONU. Plan mundial para el decenio de acción para la seguridad vial 2011-2020-www.who.int/roadsafety/decade_of_action/.

[3] Ministerio de Tránsito y Transporte de Colombia, 2002. Código Nacional de Tránsito y Transporte. URL: http://www.colombia.com/actualidad/codigos-leyes/codigo-de-transito.

[4] Grupo Nacional Provisional, 2012. Reflexiones sobre paradigmas de la seguridad vial y su visión futuro, 2012. URL: https://www.gnp.com.mx/wps/portal/portalesgnp/personas.

[5] U.S. National Highway Traffic Safety Administration, NHTSA; Department of Transportation – Tri-level accident investigation study (Indiana Tri-Level Study) – Final Report – Washington, D.C., 1973 – DOT HS-800 912

[6] Izquierdo F.A. Análisis de la seguridad vial española: Un modelo integrado para la evaluación de los principales factores de influencia. URL: www.institutoivia.com/cisev.../analisis...aa/Francisco_Aparicio.pdf

[7] Hakkert A.S, Gitelman V. & Vis M.A. (Eds.) - Road safety performance indicators; Theory - EU Project FP6 SafetyNet - Deliverable D3.6 – 2007. URL: http://www.ripcord-iserest.com.

[8] Berardo Maria Graciela. Aplicación del modelo de predicción de accidentes viales del HSM (2010) en camino rural de dos carriles en Brasil, Córdoba, Argentina, 2015.

[9] Perez Rojas Jazmín. Uso del módulo de predicción de accidentes (CPM) del IHSDM para evaluación de seguridad en segmentos de carreteras de dos carriles. Respuestas. 2013; 18(2): 87-95.

[10] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

[11] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier.

[12] Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). *Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley and Sons.