UNIVERSIDAD DE ESPECIALIDADES ESPÍRITU SANTO

Facultad de Ingeniería

Escuela de Computación y Telecomunicaciones

MODELO PREDICTIVO EN LOS ACCIDENTES DE TRANSITO CON BASE EN DATA SCIENCE.

Trabajo de Titulación que se presenta como requisito para el título de Ingeniero en Sistemas

Autor: Jean Jordy Segarra Arias

Tutor: Ing. Iván Silva Feraud

Samborondón, Febrero de 2020

APROBACIÓN DEL TUTOR

En mi calidad de Tutor del estudiante Jean Jordy Segara Arias, que cursa estudios en el programa de TERCER nivel: Ingeniería en Sistemas, dictado en la Facultad de Sistemas, Telecomunicaciones y Electrónica de la UEES, en modalidad presencial.

CERTIFICO

Que he revisado el Trabajo de Titulación denominado: “Modelo predicitivo en los accidentes de tránsito con base en Data Science”, presentado por el estudiante Jean Jordy Segarra Arias, como requisito previo para optar por el Grado Académico de Ingeniero en Sistemas CERTIFICO que el Trabajo de Titulación ha sido analizado y reúne todos los requisitos para ser presentado y sometido a los procesos de revisión estipulados por la Facultad.

Atte.

_____
Iván Silva Feraud
0915856371

DEDICATORIA

Este trabajo está dedicado a mi Madre querida porque a pesar de los buenos y malos momentos siempre me supo apoyar sin importar las circunstancias, siempre pensé en que el mejor regalo que se le puede dar a una madre es ver a su hijo convertirse en un profesional.

También está dedicado a mi Abuelita quien ha sabido apoyarme siempre que lo he necesitado, y aunque mi Abuelo se fue de este mundo hace mucho siento que cada logro va dedicado para él.

RECONOCIMIENTO

Agradezco mucho el tiempo y la sabiduría de mi tutor, Iván Silva Feraud por guiarme hasta este punto el cual considero un paso muy grande hacia mi vida profesional.

Es importante para mi la perspectiva que obtuve también al trabajar con el profesor Francisco Bolaños quien me supo compartir sus conocimientos.

INDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

La razón de accidentes en las vías se ha convertido en una de las principales causas que mayor número de víctimas anualmente. Este problema se concentra especialmente en países de ingresos medios y bajos donde no se cuenta con un plan de movilidad segura ni medidas de prevención efectivas referentes a accidentes de tránsito para evitar muertes y lesiones en la sociedad.

En el Ecuador y en el cantón Guayaquil específicamente, no es ajeno este problema de accidentalidad vial, por lo que en este estudio se realizó un análisis estadístico de los principales sectores de la ciudad, tomando en cuenta factores como la hora y el día de la semana de estos incidentes en la ciudad, tomando como referencia la base de accidentes de la ciudad, y con base en los resultados, se pueda identificar propuestas y soluciones con el fin de reducir las cifras negativas de esta problemática social, con la información proporcionada por las entidades que controlan el tránsito en el cantón, con ayuda de software estadísticos.

Se conceptualiza términos utilizados en el estudio teórico del tema analizado y los métodos estadísticos que se utilizan para este proyecto, posteriormente, mediante la base de datos de accidentes de tránsito 2016-2018, se realizó un análisis estadístico para conocer el estado actual de la proporcionalidad y ocurrencia de accidentes en el cantón, luego se desarrolló el análisis de la probabilidad de accidentes mediante el modelo de regresión logística binaria para obtener una estimación sobre que si existe o no un accidente de tránsito para el año 2019 en la ciudad.

Palabras Clave: Accidentes de tránsito, Logit, prevención, estadística, leyes.

ABSTRACT

The reason for road accidents has become one of the main causes of the greatest number of victims existing annually; This problem is especially concentrated in middle and low-income countries where there is no safe mobility plan or effective prevention measures related to traffic accidents to prevent deaths and injuries in society.

In Ecuador and in the canton of Guayaquil specifically, this road accident problem is no stranger, so in this study a statistical analysis was carried out where the main sectors of the city were analyzed, taking into account factors such as time and day of the week of these incidents in the city, taking as a reference the base of accidents of the city, and based on the results, proposals and solutions can be identified in order to reduce the negative figures of this social problem, with the information provided by the entities that control the traffic in the canton, with the help of statistical software.

Terms used in the theoretical study of the analyzed topic and the statistical methods used for this project are conceptualized, subsequently, through the 2016-2018 traffic accident database, a statistical analysis was carried out to know the current status of proportionality and occurrence of accidents in the canton, then the analysis of the probability of accidents was developed using the binary logistic regression model to obtain an estimate of whether or not there is a traffic accident for the year 2019 in the city.

Keywords: Traffic accidents, Logit, prevention, statistics, laws.

1. INTRODUCCIÓN

1.1 Antecedentes

En América Latina los accidentes de tránsito han sido los responsables para que muchas avenidas se pinten de sangre en varios de los casos por culpa de ciertos conductores que infringen las leyes de transito lo que provoca no solo la pérdida de un familiar o ser querido sino también en muchos de los casos las mutilaciones de varios de sus miembros que los dejan no solo con una discapacidad temporal sino permanente, y a su vez de los peatones quienes de igual forma poco o nada de caso hacen a las señales de tránsito que en el transcurso de las vías se encuentran.

En materia de seguridad vial, América Latina sigue ocupando el primer lugar en el triste ranking mundial de las regiones con las tasas de mortalidad más altas por accidentes de tránsito. Es por esta razón que la seguridad vial se ha convertido en un tema de conversación obligado entre los gobiernos de los países de América Latina y el Caribe (Salud, 2018).

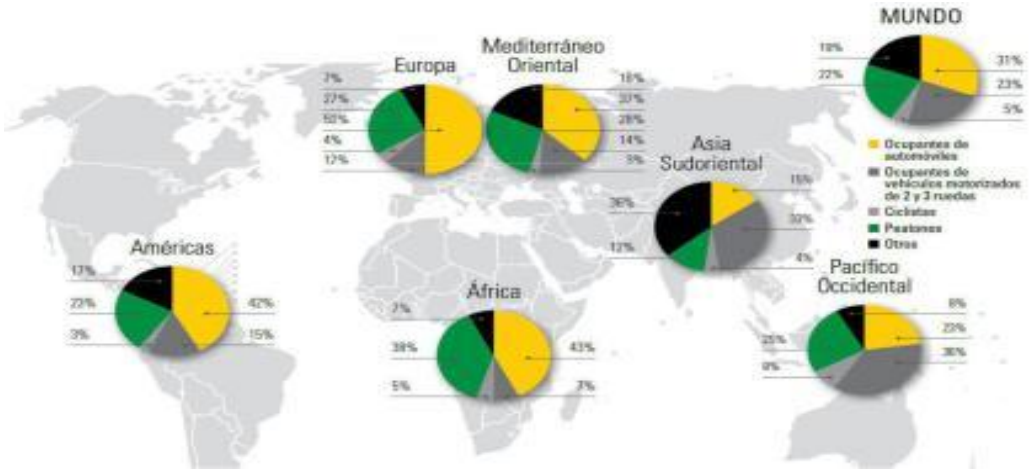


Figura Nro. 1.- Enfoques de Seguridad Vial
Elaborado por.- OMS
Fuente.- Organización Mundial de la Salud (OMS)

La Organización Mundial de la Salud (OMS) es uno de los organismos que inicialmente tomo cartas en el asunto, junto con otras organizaciones de primera instancia como la Organización de Naciones Unidas (ONU) quienes dieron a conocer el problema de seguridad vial y las catastróficas consecuencias a nivel mundial. A partir de ello ha ido adquiriendo la debida importancia aunque la OMS y la ONU reconocen que aún hay mucho trabajo por adelante (Salud, 2018).

En Ecuador esta problemática surge a diario y aunque exista una ley especifica que sancione a los conductores que no cumplen con lo que la ley expresa, muchos de los conductores poco o nada hacen por cambiar esta situación, con tasas de 20,4% por cada 100,000 habitantes, el Ecuador se encuentra en el puesto 70 en el ranking de muertes por causa de los accidentes de tránsito, la ciudad de Guayaquil en el segundo lugar del país por tener el mayor número de accidentes de tránsito en el año 2018 (Transito, 2018).

La Agencia Nacional de Tránsito registra que en la mayoría de los casos los accidentes se deben a la embriaguez, que es la tercera causa, las distracciones al momento de manejar y el exceso de velocidad, que se encuentra en cuarto puesto en la actualidad, así como también la Ley Orgánica de Transporte Terrestre Tránsito y Seguridad Vial dispone que los peatones deban ser sujetos de sanción pecuniaria por infringir la Ley cuando cometan una infracción de tránsito; y serán privados de la libertad cuando sean responsables de una contravención de tránsito (Transito, 2018).

1.2 Descripción del Problema

La ciudad de Guayaquil al ser considerada como una de las más grandes ciudades que tiene el Ecuador, posee varios sectores considerados como peligrosos por generarse a diario los accidentes de tránsito, debido a que uno de los principales factores para que se produzcan los accidentes de tránsito es la imprudencia y negligencia por parte de conductores y a su vez de peatones, dentro de esta problemática se puede establecer que una de las causas para generarse este suceso es el exceso de velocidad lo que según datos estadísticos provoca que cada 20 minutos surja un accidente de tránsito y cada cuatro horas alguien muera por esta causa (Transito, 2018). Así también parte de esta problemática es el mal estado de las vías, la falta de señalización, el consumo de alcohol por parte tanto de conductores como de peatones, y por no usar el cinturón de seguridad o el casco (en el caso de motocicletas), lo que provoca que se dé un gran número de accidentes de tránsito y muertes.

Gran parte de esta problemática la tienen los Agentes Civiles de Transito, ya que es responsabilidad de ellos controlar a diario el tránsito, lo que genera que se vulnere la seguridad jurídica, adicional a ello, la falta de un instructivo jurídico que sirva de guía para los Agentes Civiles de Tránsito en el control vehicular es fundamental para disminuir del accidentes de tránsito especialmente los producidos por imprudencia y negligencia de conductores y peatones.

A su vez deberá realizarse una descripción total sobre el nivel de accidentes de tránsito que a diario se vive en la ciudad de Guayaquil, los motivos por la que los conductores no respetan las normas de seguridad vial todo esto en base a datos estadísticos proporcionados por la Agencia Nacional de Tránsito.

1.3 Alcance y Delimitación del Objeto

El alcance del proyecto de titulación radica en establecer un modelo de predicción para definir una probabilidad de accidentes de tránsito, dadas diversas condiciones, dentro de la ciudad de Guayaquil.

Para implementar este proceso se utiliza software estadístico especializado para generar los análisis, para que anualmente se realice la calibración de los pesos de las variables utilizadas.

1.4 Justificación

La impericia por parte de los conductores se ha vuelto común, estos problemas se viven a diario en Ecuador, especialmente en Guayaquil, al ser una de las ciudades más grandes del país es necesario mejorar el control de estos siniestros.

Es importante recalcar que para mejorar dicho control sobre los accidentes ocurridos es necesario mirar hacia atrás, analizando la información existente creando una regresión para de esta manera aprender sobre lo ocurrido.

La Agencia Nacional de Tránsito ha podido registrar que la mayor cantidad de casos de accidentes de tránsito han ocurrido por el estado de embriaguez por parte de los conductores, seguido de las distracciones al momento de manejar y el exceso de velocidad, gracias a este registro se ha podido concluir con esta información. Sin embargo, es necesario saber identificar las variables de la información obtenida por parte de la Agencia Nacional de Tránsito para un mejor análisis y un modelo más apegado a la realidad.

Es importante establecer algún tipo de predictibilidad generando probabilidades de ocurrencias de accidentes de tránsito, con la información existente hasta el día de hoy. La falta de un control adecuado en la circulación vehicular por parte de la Agencia Nacional de Tránsito se debe a que no poseen algún tipo de instructivo jurídico específico en el que se mencione el/los mecanismos necesarios para disminuir los accidentes de tránsito.

1.5 Objetivos Generales y Específicos

1.5.1 Objetivo General

Establecer una probabilidad de ocurrencia de un accidente de tránsito mediante la identificación de variables que sirva de guía para los Agentes Civiles de tránsito en el control vehicular, para tomar medidas de mitigación de estos eventos en la ciudad de Guayaquil.

1.5.2 Objetivos Específicos

- Realizar un estudio de datos sobre el comportamiento de los accidentes de tránsito producidos por las diversas causas establecidas en la data histórica.
- Analizar cuáles son las causas fundamentales más relevantes que provocan los accidentes de tránsito en la ciudad de Guayaquil.
- Generar una modelo predictivo según el comportamiento de los accidentes de tránsito por días y horas para establecer un patrón.
- Recomendar los días y las horas en los que los Agentes Civiles de Transito puedan mitigar los accidentes en la ciudad con base en más controles.

2. MARCO REFERENCIAL

2.1 MARCO TEÓRICO

2.1.1 Data science para la toma de decisiones

Para definir una solución a un problema cotidiano se poseen diversas formas de resolver los acontecimientos, en el pasado se evaluaba más con la intuición del investigador y la experiencia del consultor o de la dirección del proyecto, en la actualidad con la información que se tiene y el procesamiento de la data, nacen los conceptos de data science para que, por medio de herramientas estadísticas y software capaces de procesar diversa información se puedan generar diversos escenarios para la toma de decisiones.

Según un artículo publicado en CHEST Journal en el año 2018, se define a data science como el conjunto de principios fundamentales que apoyan y guían la extracción basada en principios de información y conocimiento de los datos (Churpek, 2018).

El Data science nace de la necesidad de generar estos “escenarios” como la información obtenida mediante el histórico de accidentes de una ciudad, donde el investigador evalúa diversas formas de generar resultados con base el tratamiento de los datos y la definición de hipótesis para poder llegar al objetivo.

Si bien es cierto, el investigador debe tener un conocimiento matemático-estadístico para poder manipular estas herramientas, el analista debe estar en la capacidad de poder gestionar, analizar, construir y definir que data es la que nos va a ayudar a resolver el problema y en la que la matemática definida en las herramientas informáticas puede hacer el trabajo.

En base a las definiciones anteriores, podemos establecer diversas etapas o fases en las que se efectúa el análisis de datos hasta su interpretación en un modelo conceptual, la cual radica en lo siguiente (Churpek, 2018).

- Recopilamiento de datos.
- Tratamiento y pre procesamiento.
- Entrenamiento del modelo.
- Testeo.
- Visualización e interpretación de los resultados

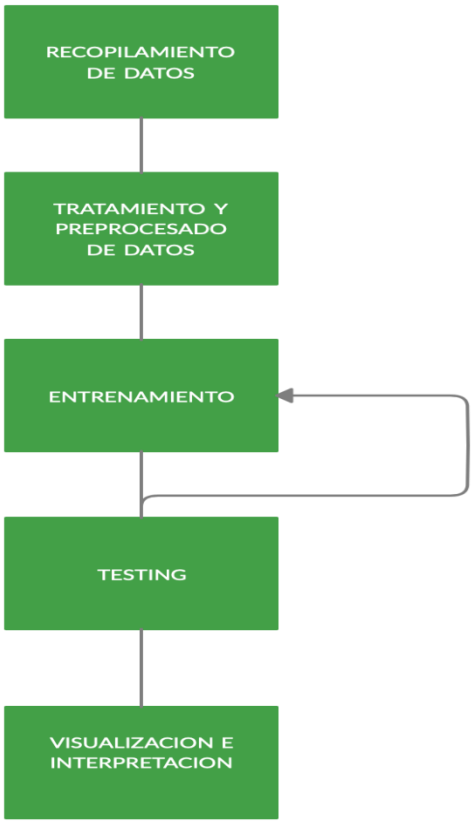


Figura Nro. 2.- Fases de Data science
Elaborado por. Autor
Fuente. Elaboración propia

Recopilamiento de Datos

La información, no siempre se encuentra a disposición de los investigadores, en la actualidad existe diversos orígenes de datos de la misma las cuales pueden alojarse desde una base de datos hasta en hojas de cálculo, archivos textos, imágenes, etc. La recopilación de la información puede realizarse mediante

diversas técnicas tales como la colocación de sensores en las aplicaciones o bases de datos cuando se ejecuta alguna transacción, scrapeo de la web para extraer datos de comportamiento de clics de trafico de páginas, peticiones a APIs o robots para extraer información, diversos formularios que nos permita la tecnología del negocio y poder aplicarla.

Esto es obtener una idea de los datos actuales y comprender lo que significa cada parte de los datos. Esto puede implicar averiguar qué datos serían los más necesarios y las mejores formas de adquirirlos. Esto también significa descubrir qué significa cada uno de los puntos de datos en términos del proyecto. Por ejemplo, si recibe un conjunto de datos de un cliente (Agencia Nacional de Tránsito), debe saber qué representan cada columna y fila. ¿Las filas representan a un solo cliente? ¿Esta columna con un encabezado de lo que parece ser un acrónimo tiene una gran relación con los datos? Realmente no podemos saber esto sin entender exactamente significa (Said, 2019).

.

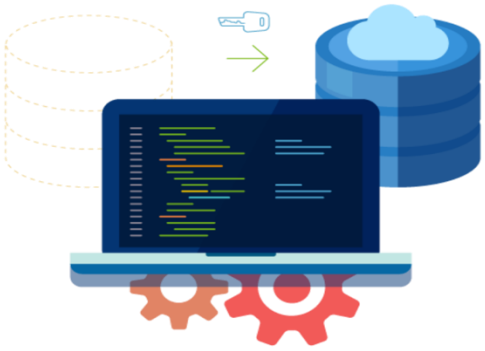


Figura Nro. 3.- Recopilación de Datos
Elaborado por. Autor
Fuente.- Elaboración propia

Tratamiento y preprocesado de Datos

Cuando el proceso de extracción de datos se ha realizado se debe de reprocesar la data que se considera fundamental para resolver el problema propuesto con base en un modelo conceptual, para ello se debe realizar diversos análisis a los datos, análisis de variables, consistencia y persistencia de la base de datos, es decir, categorizar la información que se obtuvo en la extracción para conocer qué datos se deben de preparar para la siguiente fase del estudio.

Hay diversas técnicas y métodos que pueden aplicarse en el reprocesamiento y análisis, entre los que destacamos los siguientes:

- Reducción de la dimensionalidad.
- Elección de Variables discretas y continuas.
- Normalización de datos.
- Cuantificación de Filas.
- Elección de ventanas de tiempo.

La parte de preparación de datos del proceso es donde se invierte la mayor parte del tiempo. La limpieza de los datos puede ser más una forma de arte que una ciencia, ya que es necesario darse cuenta si los datos correctos están disponibles para proceder a un buen modelo y saber cómo limpiarlo correctamente para que no corrompa su modelo. También se considera que tener datos confiables es parte de esto. Hay un viejo dicho, "basura adentro, basura afuera". El modelo no será muy efectivo si los datos ingresados son incorrectos (Gardner, 2019).

.

El pre proceso de los datos es una fase vital para el éxito de la implementación de la solución ya que condiciona a todo el proceso analítico y crear distorsión en los resultados esperados del modelo conceptual implementado.

Figura Nro. 4.- Pre Proceso de Datos
Elaborado por. Autor
Fuente.- Elaboración propia

Entrenamiento

Entrenar un modelo conceptual quiere decir que al modelo implementado como solución se procede a alimentarlo con datos para evaluar su efectividad, para ello existen diversas técnicas para realizar lo mencionado, pero la técnica que está a la vanguardia de las demás es el proceso de entrenamiento de algoritmos de machine learning.

Los algoritmos de machine learning tiene la función específica de pronosticar información nueva, a raíz de haber sido “entrenados” con data histórica que defina un comportamiento de diversas variables, por ello es de suma importancia establecer un modelo matemático, econométrico o macroeconómico acorde a lo que se debe resolver.

Otras técnicas de definir algoritmos de predicción son los siguientes:

- Los árboles de decisión.
- Las redes neuronales.
- Los algoritmos de clusterización.

Los algoritmos de predicción se pueden clasificar como de aprendizaje según resultados anteriores o como de aprendizaje pronosticado, la diferencia radica en que al momento de realizar la evaluación de la predicción en el primer caso tenemos información de contraste, mientras que en el segundo no se posee data para comparar.

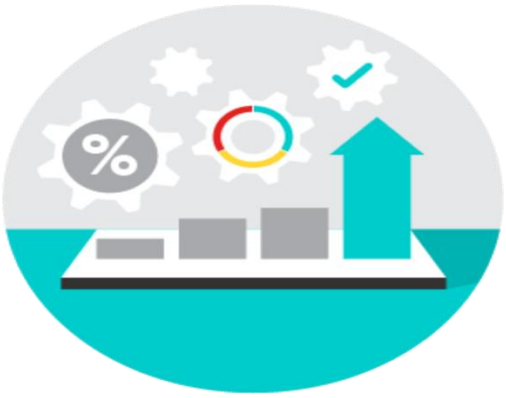


Figura Nro. 5.- Entrenamiento
Elaborado por. Autor
Fuente.- Elaboración propia

Testing

Esta es la parte donde se prueba para ver si se obtiene un buen modelo o no antes de implementarlo o presentarlo. Como indica el diagrama, esta es también la parte en la que se asegura de que el modelo responda las preguntas comerciales que tenía al comienzo de este proceso. Quizás incluso pueda descubrir más preguntas que son más importantes.

De una forma conceptual no se puede establecer que algoritmo dará el mejor resultado, por ello, la manera de definir la mejor solución o ajuste es realizando un proceso de testing, con diferentes configuraciones, hasta validar el mejor ajuste con el mínimo de desviación sugerida en los indicadores estadísticos.

Para ello, se establece una metodología experimental llamada, prueba-error dado que el modelo “adivina” el resultado y con base en el modelo implementado

es muy difícil ajustar la configuración de las variables para que directamente se obtenga los mejores resultados.

Existen diferentes técnicas para validar los resultados de un algoritmo de machine learning o de predicción de datos, entre ellos son las pruebas de bondad y ajuste y los llamados coeficientes de determinación para definir la probabilidad válida para la implementación, de cualquier forma en el análisis de los modelos se recomienda que al realizar el testing se tome como referencia resultados anteriores. (Pedrycz, 2019).

.

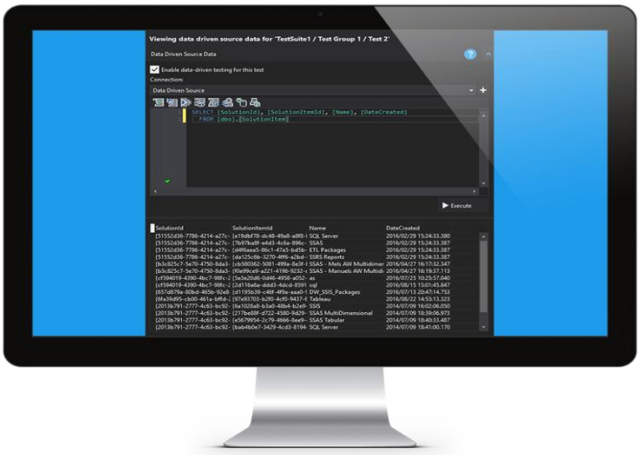


Figura Nro. 6.- Testing
Elaborado por. Autor
Fuente.- Elaboración propia

Visualización e Interpretación

Para finalizar el proceso, y poder interpretar los resultados para terceros, se tiene que representar de una forma clara y especifica del estudio realizado, por ello diversos software tiene en sus API's reportería y visualizaciones que ayudan a crear soluciones que muestren al usuario el contenido del estudio de una forma clara y precisa para la toma de decisiones.

Aquí tambien es donde se comparte los hallazgos de los datos. Esto no se limita a tener una API para llamar que use el modelo. Simplemente se podría documentar los hallazgos en un correo electrónico, un documento compartido o presentando a un grupo de personas u organización. Si bien es fácil hablar de manera técnica, la clave de este paso es transmitir lo que encuentre en los datos a un equipo de ventas o a los ejecutivos para que puedan tomar medidas con ellos, en el caso de este proyecto nuestro objetivo son los agentes de tránsito para que tomen las medidas necesarias según lo consideren. (Said, 2019).

MACHINE LEARNING

Es el campo de estudio que se centra en cómo las computadoras aprenden de los datos y el desarrollo de algoritmos que hacen posible este aprendizaje y poseen las siguientes características:

Los elementos de datos, también conocidos como variables independientes, utilizados para entrenar un modelo. Las características pueden ser simples transformaciones de los datos sin procesar (p. Ej., Frecuencia cardíaca promedio en las últimas 24 h) o transformaciones complejas, como las realizadas por redes neuronales.

Los elementos de datos, también conocidos como variables dependientes, representan el objetivo para la capacitación en un modelo de aprendizaje supervisado. Los resultados pueden ser categóricos (p. Ej., Sí / no) o continuos (p. Ej., Duración de la hospitalización de un paciente). Por otro lado, los resultados binarios categóricos son los más comunes en medicina (p. Ej., Muertos o vivos a los “n” días). Tambien, los resultados binarios generalmente se representan como una lógica booleana (es decir, verdadero / falso o 1/0). (Churpek, 2018).

MODELO DE ENTRENAMIENTO

El proceso mediante el cual los algoritmos de aprendizaje automático desarrollan un modelo de datos al aprender las relaciones entre características y,

en el aprendizaje supervisado, entre características y resultados. Esto también se conoce como derivación del modelo o ajuste de datos (Butler, 2016).

VALIDACION DEL MODELO

El proceso de medir qué tan bien un modelo se ajusta a datos nuevos e independientes. Por ejemplo, evaluar el rendimiento de un modelo supervisado para predecir un resultado en nuevos datos. Este enfoque también se conoce como prueba modelo (Churpek, 2018).

MODELO PREDICTIVO

Un modelo generalmente entrenado para predecir la probabilidad de una condición, evento o respuesta. La Administración de Alimentos y Medicamentos de EE. UU. Considera específicamente las estrategias predictivas como aquellas orientadas a identificar grupos de pacientes con más probabilidades de responder a una intervención (Churpek, 2018).

TIPOS DE ALGORITMOS EN DATA SCIENCE

Los algoritmos de aprendizaje automático generalmente se dividen en dos categorías: supervisados y no supervisados. Los algoritmos semi supervisados representan un híbrido de los dos. Finalmente, los algoritmos de aprendizaje profundo desafían esta clasificación, a pesar de que derivan de algoritmos de redes neuronales artificiales, que generalmente se clasifican como algoritmos supervisados. La característica más definitoria del aprendizaje profundo es su enfoque en el aprendizaje de representaciones de datos (o características) que luego pueden usarse en problemas supervisados, no supervisados o semi-supervisados (Kellerher, 2018).

ALGORITMOS DE APRENDIZAJE SUPERVISADOS

Los algoritmos de aprendizaje supervisados se utilizan para descubrir la relación entre variables de interés y uno o más resultados objetivo. Para problemas supervisados, se deben conocer los resultados objetivo. Por ejemplo, si los investigadores quieren saber si un conjunto de características clínicas (p. Ej., signos vitales, pruebas de laboratorio) pueden predecir la mortalidad, podrían aplicar un algoritmo de aprendizaje supervisado a un conjunto de datos en el que cada registro del paciente contiene el conjunto de características clínicas de interés y una etiqueta que especifique su resultado ("sobrevivió" o "no sobrevivió" en este caso). Los ejemplos de algoritmos de aprendizaje supervisado incluyen métodos basados en regresión (p. Ej., Regresión lineal y logística, lazo, red elástica), métodos basados en árboles (p. Ej., Árboles de clasificación y regresión, bosque aleatorio, árboles potenciados por gradiente), k-vecino más cercano, artificial redes neuronales y máquinas de vectores de soporte (Kellerher, 2018).

ALGORITMOS DE APRENDIZAJE NO SUPERVISADOS

Los algoritmos de aprendizaje no supervisados se utilizan para descubrir patrones o agrupaciones naturales en los datos, sin apuntar a un resultado específico. El caso de uso más convincente del aprendizaje no supervisado en la atención médica es en medicina de precisión, en la cual el objetivo es descubrir subconjuntos de pacientes quienes comparten características clínicas o moleculares similares y son, en teoría, más propensos a responder a terapias dirigidas a su patobiología subyacente compartida. Por ejemplo, un algoritmo de aprendizaje no supervisado puede usarse para descubrir subgrupos de pacientes con sepsis que tienen distintas características moleculares y clínicas y responderán de manera diferente a medicamentos específicos, como los corticosteroides. Algunos ejemplos de algoritmos de aprendizaje no supervisados incluyen algoritmos de agrupamiento (por ejemplo, agrupamiento jerárquico, agrupamiento de k-

medias), análisis de clase latente y análisis de componentes principales (Kellerher, 2018).

Ejemplos de uso de algoritmos en ciencia de datos

Regresión clásica, ejemplo: regresión lineal, regresión logística

Descripción: la regresión lineal es un algoritmo de aprendizaje supervisado que modela la relación entre una o más características y un resultado continuo al ajustar una línea de regresión que minimiza la suma de todos los residuos, que son las distancias entre cada característica en los datos de entrenamiento y línea que se ajusta para modelarlos. La regresión logística es una generalización del modelo lineal que utiliza la función logística para estimar la probabilidad de un resultado binario. Para hacer esto, la curva ajustada en forma de sigmoide de la función logística mapea los valores de las características en una probabilidad entre 0 y 1 (Kellerher, 2018)**.**

Regresión regularizada, ejemplo: lazo, regresión de cresta, red elástica

Descripción: una extensión de los algoritmos de regresión clásicos en los que se impone una penalización al modelo ajustado para reducir su complejidad y disminuir el riesgo de sobreajuste (Zheng P., 2018).

Tree-Based, Ejemplo: árboles de clasificación y regresión, bosque aleatorio, árboles impulsados por gradiente.

Descripción: una clase de algoritmo de aprendizaje supervisado basado en árboles de decisión. Los árboles de decisión son una secuencia de divisiones "if-then-else" que se derivan separando iterativamente los datos en grupos según la relación de las características con el resultado. El bosque aleatorio y los árboles impulsados por gradientes son ejemplos de modelos de árboles de conjunto. Los modelos de conjunto combinan la salida de muchos modelos entrenados para

estimar un resultado, Ejemplo: árboles de clasificación y regresión, bosque aleatorio, árboles impulsados por gradiente (C.K.S., 2019).

K-Nearest neighbor, Ejemplo: K-Nearest neighbor.

Descripción: un tipo de algoritmo de aprendizaje supervisado que representa datos en un espacio de características multidimensionales y utiliza información local sobre las observaciones más cercanas a un nuevo ejemplo para predecir el resultado de ese ejemplo (Kellerher, 2018).

Red neuronal, Ejemplo: red neuronal artificial, red neuronal profunda.

Descripción: una clase de algoritmos no lineales creados con capas de nodos que extraen características de los datos y realizan combinaciones que mejor representan la estructura subyacente, generalmente para predecir un resultado. Las redes neuronales pueden ser poco profundas (p. Ej., Un perceptor con dos capas) o profundas (varias capas), que forman la base del campo del aprendizaje profundo (Kellerher, 2018).

2.1.2 Estado del Arte

En este apartado se realiza la verificación de la investigación bibliográfica con diversos autores de textos y estudios para establecer un símil y obtener mejores prácticas de diversas implementaciones de este trabajo en particular.

Según el artículo Traffic Accidents Severity Prediction publicado en el año 2019 en la universidad de Yunnan en China, los métodos de predicción de gravedad de los accidentes de tránsito existentes utilizan principalmente modelos de predicción de gravedad superficial y modelos estadísticos. El modelo TASP-CNN propuesto en este estudio se implementó en Python utilizando el marco de aprendizaje profundo de código abierto de Google TensorFlow.

Para ilustrar la efectividad del modelo TASP-CNN, este experimento comparó el modelo con seis modelos estadísticos y tres modelos de aprendizaje automático. Los seis modelos estadísticos fueron: algoritmo vecino más cercano a K (K-NN), DT. Clasificadores de Naive Bayes (NBC), Regresión logística (LR), Incremento de gradiente (GB) y Máquinas de vectores de soporte (SVM, también conocidas como redes de vectores de soporte). En consecuencia, los tres algoritmos de aprendizaje automático son: redes neuronales (NN) o sistemas conexionistas, red neuronal recurrente de memoria a corto plazo (LSTM-RNN) - y convolución 1D.

El rendimiento del modelo TASP-CNN propuesto se evaluó utilizando los datos de accidentes de tránsito durante un período de 8 años (2009-2016) del Ayuntamiento de Leeds, y se comparó con el NBC, el KNN, el LR, el DT, el GB, SVC, Conv1D, NN y LSTM-RNN. Los resultados muestran que el modelo TASP-CNN propuesto es mejor que los modelos competitivos.

Cabe recalcar que los datos propuestos en el artículo de la universidad de Yunnan tienen un rango de tiempo más amplio comparando con los datos obtenidos para el modelo predictivo de Guayaquil, considerando como ventaja el tener un rango más amplio en cuanto a datos, el modelo ajustará mejor. También cabe recalcar que el enfoque es parecido, sin embargo el artículo de la universidad de Yunnan resuelve un problema más complejo como la severidad del accidente, lo cual lo han logrado gracias a información extra y un amplio equipo de laboratorio, se llega a la conclusión que un modelo propuesto en comparación a modelos existentes ajusta mejor según el propósito de cada individuo ya que en ambos casos al ser comparados con modelos existentes se pudo demostrar el mejor ajuste en el modelo propuesto en cada artículo (ZHENG, 2019).

Otro artículo realizado por la Universidad Politécnica de Henan en China en el año 2011, comprende desde factores de personas, vehículos, carreteras y medio ambiente hasta accidentes de tránsito, construyendo un modelo de red neuronal BP de tres capas basado en los factores de influencia de los accidentes de tránsito, entrena y predice los accidentes de tránsito de los años 1998-2009 en China. Los

resultados muestran que el modelo de redes neuronales de BP utilizado para predecir los accidentes de tránsito es preciso y factible.

El modelo de predicción se basa en la información sobre accidentes de tránsito y datos relacionados del año 1998 al año 2009 de China, usando los datos del año 1998 al año 2006 como muestra de capacitación, los datos del año 2007 al año 2009 como muestra de predicción, los resultados de la predicción son número de accidentes, número de muertes y pérdidas económicas directas (Aixia Zhang, 2017).

En este caso después de los análisis necesarios se pudo llegar a la conclusión de que el uso de una red neuronal para predecir los accidentes de tránsito es factible debido al histórico de datos que poseen los investigadores, el cual es un punto a favor para este caso en particular en la Politécnica de Henan y un punto en contra para el caso de los accidentes de tránsito en Guayaquil ya que no se cuenta con los mismos datos que se menciona en el documento analizado razón por la cual se decidió proponer nuestro propio modelo el cual generó un mejor ajuste que una red neuronal.

En otro artículo realizado por la Universidad de Ciencia y Tecnología de China, la predicción detallada de accidentes de tránsito en toda la ciudad es de gran importancia para la gestión del tráfico urbano. Según Zhengyan Zhou los enfoques existentes aplican principalmente métodos clásicos de aprendizaje automático basados en registros históricos de accidentes. Por lo tanto, no pudieron involucrar los datos de dominio cruzado, que contienen dependencia espacial y temporal. Recientemente, con más datos urbanos entre dominios disponibles, aprovechando los datos entre dominios mediante algoritmos de aprendizaje profundo para predecir accidentes, proponen un marco ResNet basado en la atención para modelar la correlación sofisticada entre datos urbanos.

En el artículo publicado por Zhou se recopiló los datos de dominio cruzado relacionados con el tráfico en la ciudad de Nueva York (excepto Staten Island) en 2017. El artículo propone un nuevo marco ResNet basado en la atención para

predecir los accidentes de tránsito de gran escala en toda la ciudad. El modelo de inferencia de velocidad que considera la región adyacente y distante se aplica para completar los valores faltantes. Además, ResNet y el mecanismo de atención se introducen en la tarea de predicción de accidentes, lo que hace que el modelo supere a otros métodos de aprendizaje profundo (Zhou, 2019).

En este caso el documento analizado tiene un enfoque en la predicción detallada de accidentes de tránsito, lo cual obliga a los autores a una recopilación de datos exhaustiva desde la estructura de red de carreteras hasta datos meteorológicos, sin embargo la finalidad de Zhou y este documento mantienen el enfoque en la predicción, se puede observar que en ambos casos la mejor opción es un modelo propuesto siempre y cuando se compare con los modelos ya existentes para conseguir el ajuste con mayor porcentaje.

Así mismo, la Facultad de Ingeniería de Comunicación y la Información de la Universidad de Correos y Telecomunicaciones de Nanjing publicó el artículo Vehicle Accident Risk Prediction Based On AdaBoost-SO in VANETs, afirmando que con la rápida expansión del tráfico por carretera y la escala de los vehículos, los ciudadanos deben enfrentar graves riesgos de seguridad de la vida causados por los accidentes de vehículos. Afortunadamente, las redes ad hoc vehiculares brindan información importante sobre big data de vehículos, lo que hace tener nuevos métodos para analizar los accidentes de tránsito de vehículos.

El artículo, aborda el problema de predecir el riesgo de accidentes de vehículos y propone la tricotomía Adaboost con algoritmo SMOTE y codificación One-Hot (AdaBoost-SO) para lograr un modelo de predicción de riesgo de accidentes de vehículos. En el artículo, según Haitao Zhao la predicción del riesgo de accidentes se basa principalmente en el uso de minería de datos grandes y el análisis de datos de accidentes de la vida real (Haitao Zhao, 2019).

A diferencia del modelo propuesto para la predicción de accidentes de tránsito en Guayaquil, el documento de Zhao obtiene los datos en tiempo real, los mismos que analizan que tan peligroso es un vehículo para producir un accidente,

al igual que en el caso de la ciudad de guayaquil el mejor modelo con el mayor ajuste fue un modelo propuesto, es por esto que se llega a la conclusión que un modelo predictivo con datos obtenidos en casos específicos ajustará en el mayor de los casos en comparación con los modelos existentes.

2.1.3 Minería de Datos e Inteligencia Artificial

La minería de datos puede manejar diversos conceptos de inteligencia artificial en algunas fases vistas en el apartado anterior, ya que algunos de los algoritmos que implementan el uso de machine learning forman parte de técnicas de grafos y de inteligencia artificial.

La inteligencia artificial busca definir un proceso de simulación y establece el razonamiento lógico con base en “hipótesis” o reglas definidas por el investigador en las que se incluye el análisis de patrones, clasificación y predicción de eventos con base en un histórico, también para generar datos en concordancia y a un comportamiento dirigido a una distribución estadística de una variable ya sea esta continua o discreta, ejemplos de esta aplicación, es el cálculo de rutas, preferencias de géneros o de artículos de compra o venta, asistentes de atención al cliente, etc.

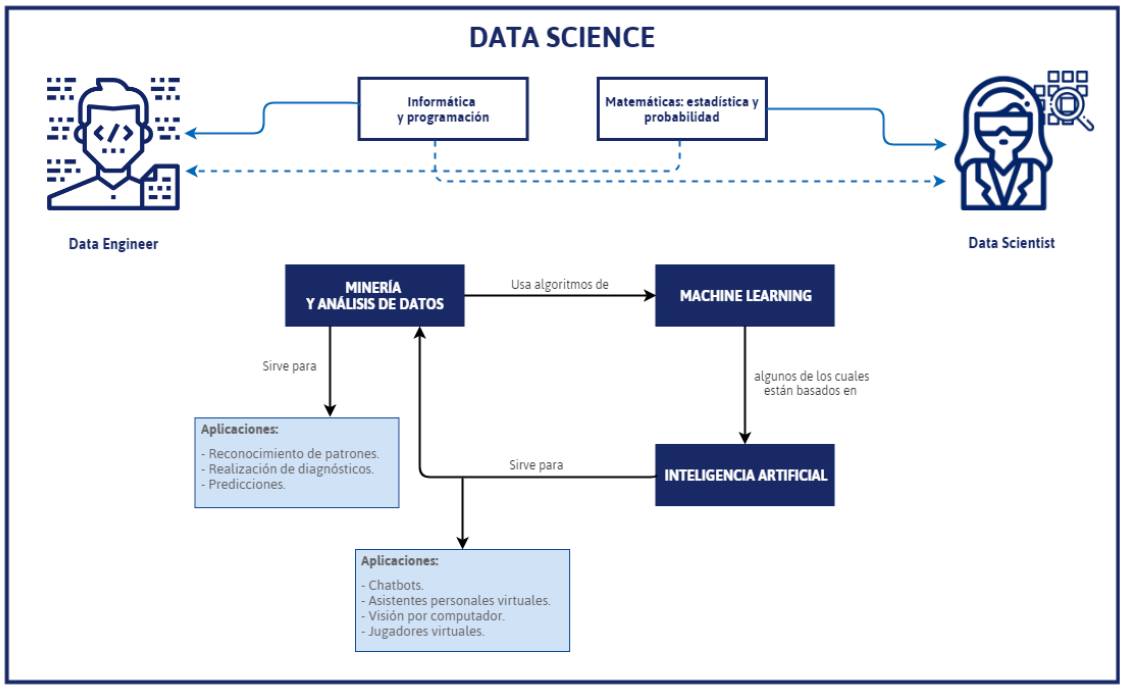


Figura Nro. 7.- Proceso de Data science
Elaborado por. Dpto. Data Science Aukera
Fuente.- Aukera.es

La inteligencia artificial tiene como objetivo resolver problemas basados en la representación del conocimiento y sistemas basados en el conocimiento, dado que el rendimiento de un programa puede incrementarse si el programa aprende de la actividad realizada así como también se han desarrollado herramientas que permiten extraer conocimiento a partir de bases de datos.

2.1.4 Software estadístico R

R es una implementación de software libre del lenguaje S pero con soporte de alcance estático. Se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y gráficas, R es parte del sistema GNU y

se distribuye bajo la licencia GNU GPL. Está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux (Micheaux, 2017).

2.1.5 Regresión Logística

El método de regresión logística es el recomendado cuando se trabaja con una variable cualitativa con dos niveles, tanto con uno (regresión logística simple) como con múltiples predictores (regresión logística múltiple).

Los modelos de Regresión Logística son modelos de regresión que permiten estudiar la dependencia lineal de una variable respuesta con dos categorías (dicotómica) respecto a otras variables explicativas (categóricas o cuantitativas). Representaremos las dos posibles respuestas de la variable respuesta como 0 y 1 (Cáceres, 2017).

Los métodos de clasificación permiten predecir variables cualitativas o categóricas. Tres de los clasificadores más usados son:

- Regresión logística
- Análisis discriminante lineal y cuadrático
- K-vecinos más cercanos (K-nearest neighbours)

Al igual que en el caso de regresión, en los problemas de clasificación contamos con un set de observaciones de entrenamiento (x1,y2),...,(xn,yn) que usamos para generar el clasificador. El objetivo es que el modelo funcione bien con las observaciones de entrenamiento y con nuevas observaciones.

¿Por qué no regresión lineal?

Para una variable respuesta binaria (dos niveles) podríamos crear dos variables dummy (0/1) y predecir la variable codificada como 1 si Y > 0,5 (o usando un límite mayor o menor dependiendo del interés). En este caso no importaría que nivel fuera codificado como 0 o 1, el modelo de regresión lineal diera el mismo resultado. Si la variable cualitativa contara con más de dos niveles, el orden en que

se establecieran las variables dummy (no habiendo un criterio que indique en qué forma se tienen que ordenar) influiría en el modelo resultante, por lo que este enfoque no se considera apto para estas situaciones. Por otro lado, algunos de los valores estimados mediante una recta de mínimos cuadrados pueden ser < 0 o > 1, lo que entra en conflicto con el hecho de que toda probabilidad está comprendida entre [0, 1] (DAVILA, 2015).

Llamaremos a la variable dependiente, que refleja la ocurrencia o no del suceso. Puesto que es dicotómica, admitamos que puede asumir los dos valores siguientes:

$$h = 1 \qquad h = 0$$

Un proceso binomial está caracterizado por la probabilidad de éxito (p) y la probabilidad de fracaso (1-p) ya que, evidentemente, las probabilidades deben sumar 1.

Concepto de ODDS o razón de probabilidad, ratio de ODDS y logaritmo de ODDS

Los ODDs o razón de probabilidad de verdadero se definen como el ratio entre la probabilidad de evento verdadero y la probabilidad de evento falso p/q (Lorenzo, 2017).

Supóngase que la probabilidad de que un evento sea verdadero es de 0.8, por lo que la probabilidad de evento falso es de 1 - 0.8 = 0.2.

En este caso los ODDs de verdadero son 0.8 / 0.2 = 4, lo que equivale a decir que se esperan 4 eventos verdaderos por cada evento falso.

En la regresión lineal simple, se modela el valor de la variable dependiente Y en función del valor de la variable independiente X. Sin embargo, en la regresión logística, tal como se ha descrito en la sección anterior, se modela la probabilidad de que la variable respuesta y pertenezca al nivel de referencia 1 en función del

valor que adquieran los predictores, mediante el uso de LOG of ODDs (Cáceres, 2017).

La trasformación de probabilidades a ODDs es monotónica, si la probabilidad aumenta también lo hacen los ODDs, y viceversa. El rango de valores que pueden tomar los ODDs es de [0,∞]. Dado que el valor de una probabilidad está acotado entre [0,1] se recurre a una trasformación logit (existen otras) que consiste en el logaritmo natural de los ODDs. Esto permite convertir el rango de probabilidad previamente limitado a [0,1] a [−∞,+∞] (Lorenzo, 2017).

Los ODDs y el logaritmo de ODDs cumplen que:

Si p(verdadero) = p(falso), entonces odds(verdadero) = 1

Si p(verdadero) < p(falso), entonces odds(verdadero) < 1

Si p(verdadero) > p(falso), entonces odds(verdadero) > 1

A diferencia de la probabilidad que no puede exceder el 1, los ODDs no tienen límite superior.

Si odds(verdadero) = 1, entonces logit(p) = 0

Si odds(verdadero) < 1, entonces logit(p) < 0

Si odds(verdadero) > 1, entonces logit(p) > 0

La transformación logit no existe para p = 0

Los denominados odds, indican cuánto más probable es el éxito que el fracaso. Obviamente entre la probabilidad del suceso y el odds correspondiente hay una clara relación directa. Si p=0, entonces el odds también es nulo; pero en la medida que tiende a la unidad, el odds tiende a infinito. La siguiente figura refleja gráficamente la relación existente entre ambas magnitudes:

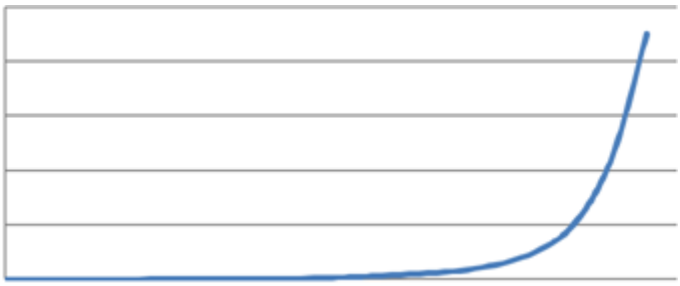


Figura Nro. 8.-Valores de "p"
Elaborado por. Autor
Fuente.- Elaboración propia

2.1.6 Regresión Logística Simple

En los modelos de Regresión Logística se pretende estudiar si la probabilidad de éxito de una variable de este tipo depende, o no, de otra u otras variables (DAVILA, 2015).

Dada una variable respuesta categórica con dos niveles según (Lorenzo, 2017), la regresión logística modela la probabilidad de que Y pertenezca a una categoría o nivel particular, dados los valores de un único predictor X. La clasificación depende del límite o threshold que se establezca.

$$\Pr(Y = k \mid X = x)$$

En regresión logística utilizamos la función logística:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

que siempre producirá una curva en forma de S, comprendiéndose los valores de Y entre [0, 1]. La ecuación anterior puede reestructurarse como

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

donde p(X) / [1 – p(X)] corresponde a los odds, pudiendo tomar cualquier valor entre 0 (muy baja probabilidad de éxito) y ∞ (muy alta probabilidad de éxito). Este ratio, pues, indica cuanto más probable es el éxito que el fracaso.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Introduciendo el logaritmo en ambos lados de la ecuación, obtenemos una función lineal

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

La parte izquierda de la ecuación es lo que se conoce como logaritmo de odds (log-odds) o logit.

La transformación de probabilidad a odds es monotónica, lo que significa que los odds aumentan conforme aumenta la probabilidad, y viceversa. Todas estas transformaciones se implementan para evitar la restricción del rango de probabilidad [0, 1] en la variable respuesta, ya que transformación logística (logaritmo de odds) permite mapear desde menos infinito hasta más infinito.

Interpretación de los coeficientes de regresión

Mientras que en regresión lineal β1 se corresponde con el cambio promedio en Y asociado a un incremento de una unidad en X, en regresión logística β1 es el valor que indica cuanto cambia el logaritmo de odds cuando X se incrementa en una unidad (Lorenzo, 2017), o equivalentemente, multiplica los odds por eβ1. La cantidad con la que p(X) cambia debido a un cambio en X dependerá del valor actual de X, pero independientemente de ello, si β1 es positivo, entonces aumentar X provocará un aumento de p(X). La intersección β0 corresponde con el resultado predicho para el nivel de referencia.

Estimación de los coeficientes de regresión

Los coeficientes β0 y β1 de la ecuación logística son desconocidos, y han de estimarse a partir de los datos de entrenamiento. Mientras que en regresión lineal los coeficientes del modelo se estiman por mínimos cuadrados, en regresión logística se utiliza el método de máxima verosimilitud (máximum likelihood): se buscan coeficientes tales que la probabilidad prevista p^(xi) de éxito se aproxime lo máximo posible a las observaciones reales (Cáceres, 2017).

"Los coeficientes estimados por el modelo para las variables se corresponden al valor del logaritmo de odds, o lo que es lo mismo, multiplica los odds por eβ1 (Cáceres, 2017)***".***

Podemos medir la precisión de los coeficientes estimados a partir de sus errores estándar. Además, en este modelo se emplea el estadístico Z para obtener el nivel de significancia del predictor (p-value), a diferencia del estadístico t en regresión lineal, aunque juegan el mismo papel. Por ejemplo, el estadístico z asociado a β1 sería igual a

$$\hat{\beta}_1 / SE(\hat{\beta}_1)$$

Un valor alto (absoluto) de Z indica la evidencia en contra de la hipótesis nula

$$H_0 : \beta_1 = 0$$

La cual implica que la probabilidad de éxito no depende de la variable independiente X

$$p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

Si el p-value es menor que el nivel de significancia establecido, podemos deducir que hay una relación entre el predictor X y la probabilidad de éxito. La ordenada en el origen β1 estimada en el modelo no suele ser de interés.

2.1.7 Regresión Logística Múltiple

La regresión logística múltiple es una extensión del modelo de regresión logística simple en el que se predice una respuesta binaria en función de múltiples predictores (Cáceres, 2017), que pueden ser tanto continuos como categóricos. La ecuación con la que podemos obtener las predicciones en este caso es

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

donde X = (X1,...,Xp) son los p predictores.

De nuevo usamos el método de máxima verosimilitud para estimar los coeficientes β0,β1,...,βp. Cada coeficiente se interpreta manteniendo fijos al resto.

Al igual que en el caso de la regresión lineal, los resultados obtenidos usando solo un predictor pueden diferir respecto a aquellos obtenidos usando múltiples predictores, especialmente cuando existe correlación entre ellos. Este fenómeno se conoce como confusión (confounding).

Condiciones Del Modelo Logístico

La regresión logística no requiere de ciertas condiciones como linealidad, normalidad y homocedasticidad de los residuos que sí lo son para la regresión lineal. Las principales condiciones que este modelo requiere son:

- Respuesta binaria: La variable dependiente ha de ser binaria.
- Independencia: las observaciones han de ser independientes.
- Multicolinealidad: se requiere de muy poca a ninguna multicolinealidad entre los predictores (para regresión logística múltiple).
- Linealidad entre la variable independiente y el logaritmo natural de odds.
- Tamaño muestral: como regla general, se requiere un mínimo de 10 casos con el resultado menos frecuente para cada variable independiente del modelo (DAVILA, 2015).

La multicolinealidad en regresión es una condición que ocurre cuando algunas variables predictoras incluidas en el modelo están correlacionadas con otras variables predictoras. La multicolinealidad severa es problemática, porque puede incrementar la varianza de los coeficientes de regresión, haciéndolos inestables. Las siguientes son algunas de las consecuencias de los coeficientes inestables:

Los coeficientes pueden parecer insignificantes incluso cuando exista una relación significativa entre el predictor y la respuesta, los coeficientes de los predictores muy correlacionados variarán ampliamente de una muestra a otra, la eliminación de cualquier término muy correlacionado del modelo afectará considerablemente los coeficientes estimados de los demás términos muy correlacionados. Los coeficientes de los términos muy correlacionados incluso pueden tener el signo equivocado (Cáceres, 2017).

Para medir la multicolinealidad, se puede examinar la estructura de correlación de las variables predictoras. También se puede examinar los factores de inflación de la varianza (FIV). Los FIV miden qué tanto aumenta la varianza de un coeficiente de regresión estimado aumenta si los predictores están correlacionados.

Si todos los FIV son 1, no hay multicolinealidad, pero si algunos FIV son mayores que 1, los predictores están correlacionados. Cuando un FIV es > 5, el coeficiente de regresión para ese término no se estima adecuadamente. Si la correlación de un predictor con otros predictores es casi perfecta, Minitab muestra un mensaje indicando que el término no se puede estimar. Los valores de FIV para los términos que no se pueden estimar por lo general superan un mil millones.

La multicolinealidad no afecta la bondad de ajuste ni la bondad de predicción. Los coeficientes (función discriminante lineal) no pueden interpretarse de forma fiable, pero los valores (clasificados) ajustados no se ven afectados (Little, 2016).

Multicolinealidad exacta

Afirmamos que hay colinealidad exacta, cuando una o más variables, son una combinación lineal de otra, es decir, existe un coeficiente de correlación entre estas dos variables de 1. Esto provoca que la Matriz X'X tenga determinante 0, y sea singular (no invertible) (Little, 2016).

Multicolinealidad aproximada

Afirmamos que hay colinealidad aproximada, cuando una o más variables, no son exactamente una combinación lineal de la otra, pero existe un coeficiente de determinación entre estas variables muy cercano al uno y por lo tanto:

$$Det\ x'x \sim 0$$

La Multicolinealidad en modelos de Regresión Lineal Múltiple

El tratamiento de la Multicolinealidad ha sido abordado ampliamente en la literatura estadística desde que introdujeron el estimador de Regresión Ridge (Cáceres, 2017), de modo que en la actualidad es ya un tópico tratado por muchos autores en distintas publicaciones.

Para resolver el problema de la presencia de multicolinealidad entre las variables independientes en el modelo lineal de regresión, yi = ß0 + ß1x1i +... + ßpxpi + ei, i =1, 2,...,n. se han planteado distintos métodos. Por ello, existen dos métodos para detectar y tratar de paliar su efecto, uno de ellos es la Regresión Ridge, el cual ha resuelto el problema de encontrar mejores estimadores de los parámetros del modelo y el otro es la Regresión sobre Componentes Principales, técnica multivariada de amplio uso en la actualidad.

La **homocedasticidad**, es una característica de un modelo de regresión lineal que implica que la varianza de los errores es constante a lo largo del tiempo (Lorenzo, 2017).

Este término, que es lo contrario de heterocedasticidad, se emplea para nombrar la propiedad de algunos modelos de regresión lineal en los que los errores de estimación son constantes a lo largo de las observaciones. Una varianza constante nos permite disponer de modelos más fiables. Además, si una varianza, aparte de ser constante es también más pequeña, nos dará como resultado una predicción del modelo más fiable.

La palabra homocedasticidad se puede desglosar en dos partes, homo (igual) y cedasticidad (dispersión). De tal manera que, si unimos estas dos palabras adaptadas del griego, obtendríamos algo así como misma dispersión o igual dispersión.

La homocedasticidad en un modelo de regresión lineal

La homocedasticidad es una propiedad deseable de los errores de un modelo de regresión simple. La homocedasticidad, como hemos dicho anteriormente, nos permite realizar modelos más fiables. Y esa fiabilidad se ve reflejada en que sea mucho más fácil para los profesionales trabajar con el modelo (Cáceres, 2017).

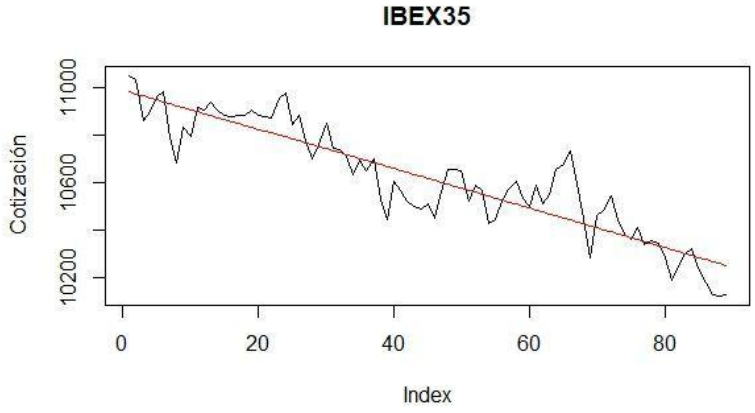


Figura Nro. 9.-Modelo de Regresión Lineal Simple
Elaborado por. IBEX
Fuente.- Yahoo Finance

En la imagen anterior podemos ver un gráfico que representa la cotización del IBEX35. La cotización hace referencia a un periodo escogido al azar de 89

periodos. La línea roja representa la estimación del IBEX35 indiciado por el tiempo. EL IBEX35 fluctúa abajo y arriba sobre esa línea de forma más o menos homogénea (DAVILA, 2015).

Para ver si nuestro modelo tiene la propiedad de homecedasticidad, es decir, para ver si la varianza de sus errores es constante, calcularemos los errores y los representaremos en un gráfico (DAVILA, 2015).

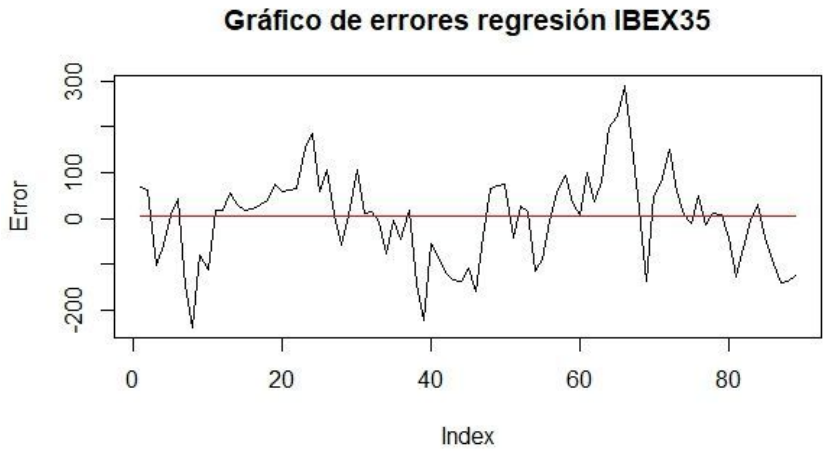


Figura Nro. 10.- Errores de Regresión Lineal Simple
Elaborado por. IBEX
Fuente.- Yahoo Finance

No podemos afirmar con seguridad que el modelo tenga la propiedad de homocedasticidad. Para ello deberíamos realizar los test correspondientes. Sin embargo, la forma del gráfico indica que sí. Un ejemplo perfecto de proceso homocedástico realizado a propósito con un programa informático está reflejado en el siguiente gráfico (DAVILA, 2015).

Tal como se indica, existen ciertas consecuencias de que un modelo no cumpla la hipótesis de homocedasticidad. Recordemos, que si un modelo no cumple el supuesto de homocedasticidad, entonces sus errores tienen heterocedasticidad.

Diferencias entre homocedasticidad y heterocedasticidad

La heterocedasticidad se diferencia de la homocedasticidad en que en ésta la varianza de los errores de las variables explicativas es constante a lo largo de todas las observaciones. A diferencia de la heterocedasticidad, en los modelos estadísticos homocedástico el valor de una variable puede predecir otra (si el modelo es insesgado), y por tanto los errores son comunes y constantes durante el estudio (Cáceres, 2017).

Las situaciones principales en las que aparecen perturbaciones heterocedásticas son los análisis con datos de corte transversal donde los elementos seleccionados, ya sean empresas, individuos o elementos económicos no tienen un comportamiento homogéneo entre ellos.

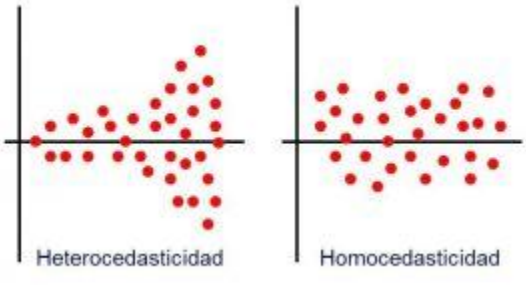


Figura Nro. 11.- homocedasticidad vs heterocedasticidad
Elaborado por. Tomas del Barrio
Fuente.- (Castro, 2018)

2.1.8 Comparación Entre Regresión Logística, Lda, Qda Y Knn

La regresión logística y el LDA son métodos muy próximos: ambos casos son funciones lineales de x, por lo que ambos producen límites de decisión lineales, aunque LDA se aplica para casos en los que la variable respuesta cuenta con predictores con más de dos clases. La única diferencia es que en regresión logística, β0 y β1 se estiman mediante el método de máxima verosimilitud, mientras que en

el caso del LDA los estimadores se corresponden a la media y varianza de una distribución normal. Así, la regresión logística puede dar mejores resultados si la condición de normalidad no se cumple (Cáceres, 2017).

LDA es un método mucho menos flexible que QDA y sufre de menos varianza. Ello puede suponer una mejora en la predicción, pero hay un inconveniente: si la asunción del LDA de que todas las clases comparten la misma matriz de covarianza no es correcta en realidad, el LDA puede sufrir un sesgo. Visto de otra manera, LDA suele ser mejor que QDA si contamos con relativamente pocas observaciones de entrenamiento y reducir la varianza es importante. Por el contrario, se recomienda QDA si el set de observaciones de entrenamiento es muy grande y la varianza del clasificador no supone un problema, o si el supuesto de una matriz de covarianza común entre las clases claramente no se cumple (Cáceres, 2017).

Si el verdadero límite de Bayes es lineal, LDA será una aproximación más precisa que QDA. Si por el contrario no es lineal, QDA será una mejor opción, cabe recalcar que el método KNN adquiere un enfoque distinto a la hora de clasificar: identifica las K observaciones más cercanas a x para clasificar dicha observación. La observación es clasificada en la clase mayoritaria a la que pertenecen las K observaciones vecinas más cercanas. Además, se trata de un método no paramétrico, ya que no asume ninguna forma sobre el límite de decisión.

Por lo tanto, cuando los límites de decisión son altamente no lineales, KNN puede superar el QDA si el número de observaciones no es limitado. Una desventaja del KNN cuando p aumenta es que hay muy pocas observaciones “cerca” de cualquier observación de test, por lo que es importante que si aumentan el número de predictores, lo haga también el número de observaciones (este fenómeno se conoce como curse of dimensionality) (Cáceres, 2017).

2.1.9 Interpretación del modelo

A diferencia de la regresión lineal, en la que β1se corresponde con el cambio promedio en la variable dependiente Y debido al incremento en una unidad del predictor X, en regresión logística, β1 indica el cambio en el logaritmo de ODDs debido al incremento de una unidad de X, o lo que es lo mismo, multiplica los ODDs por eβ1. Dado que la relación entre p(Y) y X no es lineal, β1 no se corresponde con el cambio en la probabilidad de Y asociada con el incremento de una unidad de X.

Cuánto se incremente la probabilidad de Y por unidad de X depende del valor de X, es decir, de la posición en la curva logística en la que se encuentre (Little, 2016).

Condiciones

- **Independencia**: las observaciones tienen que ser independientes unas de otras.
- **Relación lineal** entre el logaritmo natural de odds y la variable continua: patrones en forma de U son una clara violación de esta condición.
- **La regresión logística no precisa de una distribución normal** de la variable continua independiente.
- **Número de observaciones:** no existe una norma establecida al respecto, pero se recomienda entre 50 a 100 observaciones.

Predicciones

Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$p(y = 1|x) = e\beta 0^{+\beta 1^{x}} + e\beta^{+\beta 1x} \ (2)$$

Más adelante en el momento de implementar el desarrollo en el software, se emplea la función glm() con family="binomial" para ajustar modelos de regresión logística. Esta función predice por defecto el log(ODDs) de la variable respuesta. Para obtener las probabilidades P(y=1) hay que aplicar la ecuación (2), donde el valor eβ0^+β1^X es el log(ODDs) devuelto por el modelo. Otra opción es indicar el argumento type="response" en la función predict() (Little, 2016).

3.- METODOLOGÍA

3.1 DEFINICIÓN DE TECNICA DE DATA SCIENCE PARA IMPLEMENTACIÓN DE SOLUCIÓN DEL PROBLEMA

Debido a que debemos establecer un modelo matemático para pronosticar la probabilidad de accidentes en la ciudad de Guayaquil, y se debe procesar gran cantidad de información, el analista puede implementar técnicas de data science para construir un modelo de alta complejidad y de precisión por los cálculos matemáticos, por ello, el analista tomará la decisión del software estadístico que se usará basandose en el software que mayor se ajuste a este escenario, ya que el mismo será usado para implementar la solución, así como también se debe comprobar conceptualmente que el modelo funciona de manera correcta, tal como se muestra en el siguiente diagrama conceptual de la metodología.

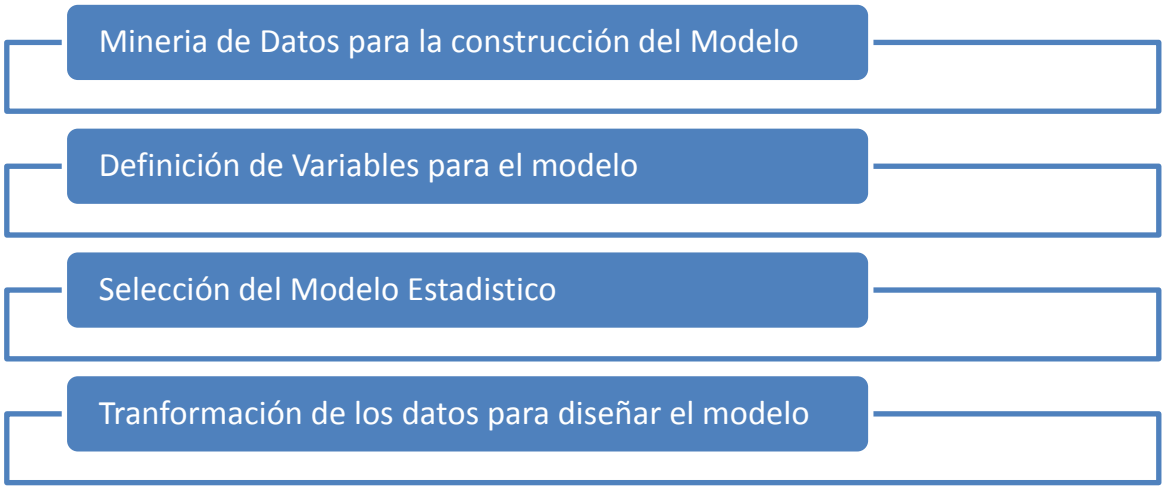


Figura Nro. 12.- Diagrama de Metodología de Elaboración de Modelo Matemático
Elaborado por. Autor
Fuente.- Elaboración propia

Para definir la técnica dando paso a la implementación de la solución al problema, se debe tomar en consideración el tipo de modelo que se va a implementar, para el proyecto se definen varias opciones las cuales son las siguientes:

- ✓ Modelos de Regresión Lineal / Múltiple
- ✓ Modelos Binarios Logit / Probit
- ✓ Redes Neuronales
- ✓ Modelos Autoregresivos

Debido a que el resultado debe ser una probabilidad, y las variables definidas para el análisis debe de ser una variable que defina si existió o no el evento en “x” o “y” lugar, hora, día, etc. Se establece que los modelos que más se ajustan a lo que se necesita para el proyecto son los Modelos Binarios Logísticos.

Cabe recalcar que estos modelos en ningún software definen las variables y como realizar los análisis, por ello se debe de construir un modelo conceptual para luego implementarlo en las herramientas estadísticas que nos ayudaran a realizar

los análisis matemáticos sin la necesidad de conocer la técnica que usan o el ajuste, en algunos casos solo se debe de ingresar las variables para que las aplicaciones generen los resultados, aquí es donde trabaja el analista de data science.

El analista debe de conocer que herramienta puede ayudarnos a resolver el modelo conceptual porque el éxito de implementar estas soluciones radica en 4 directrices:

- La velocidad de procesamiento
- La cantidad de Información procesada
- Las herramientas estadísticas
- La facilidad de implementar la solución conceptual

En este ambiente R Studio tiene diversas áreas, una de las más importantes es la consola de ejecución donde se introducen los comandos para generar los escenarios solicitados en el proyecto.

Una de las partes claves de esta herramienta es que es de software libre, es decir, que no necesita licenciamiento, adicional a ello, la comunidad de desarrolladores tiene implementadas diversas soluciones para algunos problemas, y con la ayuda de los “CRAN”, podemos importar diversas funciones o librerías para incrementar las funcionalidades de la herramienta, lo cual lo hace escalable para resolver diversos problemas planteados.

En el mercado existen herramientas tales como R, Python, SPSS, @Risk, Crystal Ball entre otras, al momento de definir el tipo de herramienta a usarse se consideró las ventajas y desventajas de cada una, tomando en cuenta que Python y R son de software libre fueron los dos favoritos para implementar cualquier modelo que se plantee de forma conceptual, para el proyecto se elige R por motivo de sus abundantes librerías y modulos existentes por el hecho de ser de software libre, con su plataforma grafica R Studio, también cabe recalcar que el analista posee los conocimientos para implementar el modelo conceptual y generar los resultados para analizar las probabilidades en esta plataforma.

3.2 PREPARACIÓN DE LA BASE, ELECCIÓN DEL MODELO Y VARIABLES

3.2.1 Inclusión de las Variables en el Modelo

Antes de iniciar el proceso de modelado, las variables han de definirse de la forma adecuada, es decir teniendo en cuenta el tipo de variable que se posee para realizar la elección del modelo: variables continuas (cuantitativas) y los factores (variables discretas nominales y ordinales).

Para esta investigación, se analizó la base de datos de la Agencia Nacional de Tránsito/INEC, el metadata que maneja la institución para definir los tipos de campo con los que se trabajara el proyecto.

INEC instituto nacional de estadística y censos		**Tabla de Metadatos No. 01**	**Institución:** Instituto Nacional de Estadística y Censos. **Versión (Actualización):** 01 **Código Documento:** TM-01-2018-V01
Código Documento Referencia (Hoja de Ruta):			
No.	**Nombre Metadato**	**Descripción**	
1	*Título o Nombre del Conjunto de Datos:*	2017_Siniestros_de_Tránsito_BDD	
2	*Descripción*	Contiene información de siniestros de tránsito; número de siniestros de tránsito por clase, causa y número de víctimas. La información disponible corresponde a los eventos registrados por las autoridades a nivel nacional, según sus competencias, los datos que son procesados, consolidados y validados por la Agencia Nacional de Tránsito (ANT).	
3	*Institución*	Instituto Nacional de Estadística y Censos.	
4	*Fecha de Publicación*	14/12/2018	
5	*Fecha de Modificación*	11/12/2018	
6	*Frecuencia de actualización*	Anual	
7	*Versión*	V01	
8	*Formato*	csv	
9	*Idioma*	ES	
10	*Fuente*	Dirección de Estadísticas Económicas	
11	*Licencia*	Referencia GPP-DA-v01, Tabla 3 Licencias de Datos Abiertos (ODbL-1.0)	
12	*Etiquetas*	Estadísticas de Transporte 2017.	
13	*Email de contacto de la Fuente*	roberto_chaves@inec.gob.ec	
14	*URI*	http://www.ecuadorencifras.gob.ec/transporte/	

Figura Nro. 13.- Tabla de Metadatos del INEC
Elaborado por. ANT/INEC
Fuente.- Agencia Nacional de Transito

Este estudio se ha realizado sobre 9858 personas, de accidentes de tránsito entre el 1 de enero 2016 y el 31 de diciembre de 2018. Así como también se revisan los tipos de datos de la base proporcionada por el INEC en función de las estadísticas recogida por la Agencia Nacional de tránsito.

instituto nacional de estadística y censos	**Diccionario de Datos No. 01**	**Institución:** Instituto Nacional de Estadística y Censos (INEC) **Versión (Actualización):** 01 **Código Documento:** DD-01-2018-V01
Código Documento Referencia (Hoja de Ruta):		
Nombre del Conjunto de Datos:	2017_Siniestros_de_Tránsito_BDD	
Nombre del Recurso:	Participación a nivel nacional, provincia y cantón.	
Descripción del Recurso:	Contiene información referente a: siniestros de tránsito, en cuanto a causas del siniestro, mes, parroquia, número de fallecidos, lesionados; proporcionada por la Agencia Nacional de Tránsito (ANT), referente al 2017.	
URI del Recurso:	http://www.ecuadorencifras.gob.ec/transporte/	

No.	Nombre del Campo (Encabezado Columna)	Descripción del Campo	Categoría
1	PARROQUIA	PARROQUIA	1 Parroquia Ayacucho 2 Parroquia Bolívar(Sagrario) 3 Parroquia Carbo(Concepción) 4 Parroquia de Chongón 5 Parroquia Febres Cordero 6 Parroquia García Moreno 7 Parroquia Letamendi 8 Parroquia Nueve de Octubre 9 Parroquia Olmedo (SanAlejo) 10 Parroquia Pascuales 11 Parroquia Roca 12 Parroquia Rocafuerte 13 Parroquia Sucre 14 Parroquia Tarqui 15 Parroquia Urdaneta 16 Parroquia Ximena

3	MES	Mes	1=ENERO 2=FEBRERO 3=MARZO 4=ABRIL 5=MAYO 6=JUNIO 7=JULIO 8=AGOSTO 9=SEPTIEMBRE 10=OCTUBRE 11=NOVIEMBRE 12=DICIEMBRE
4	DIA	Día de ocurrencia	1=LUNES 2=MARTES 3=MIÉRCOLES 4=JUEVES 5=VIERNES 6=SÁBADO 7=DOMINGO
5	HORA	Hora del suceso	0=00:00 - 00:59 1=01:00 - 01:59 2=02:00 - 02:59 3=03:00 - 03:59 4=04:00 - 04:59 5=05:00 - 05:59 6=06:00 - 06:59 7=07:00 - 07:59 8=08:00 - 08:59 9=09:00 - 09:59 10=10:00 - 10:59 11=11:00 - 11:59 12=12:00 - 12:59 13=13:00 - 13:59 14=14:00 - 14:59 15=15:00 - 15:59 16=16:00 - 16:59 17=17:00 - 17:59 18=18:00 - 18:59 19=19:00 - 19:59 20=20:00 - 20:59 21=21:00 - 21:59 22=22:00 - 22:59 23=23:00 - 23:59

6	CLASE	Clase del siniestro	1=ATROPELLOS 2=CAÍDA DE PASAJEROS 3=CHOQUES 4=ESTRELLAMIENTOS 5=ROZAMIENTOS 6=VOLCAMIENTOS 7=PÉRDIDA DE PISTA 8=OTROS
7	CAUSA	Causa probable del siniestro	1=EMBRIAGUEZ O DROGA 2=MAL REBASAMIENTO INVADIR CARRIL 3=EXCESO VELOCIDAD 4=IMPERICIA E IMPRUDENCIA DEL CONDUCTOR 5=IMPRUDENCIA DEL PEATÓN 6=DAÑOS MECÁNICOS 7=NO RESPETA LAS SEÑALES DE TRÁNSITO 8=FACTORES CLIMÁTICOS 9=MAL ESTADO DE LA VÍA 10=OTRAS CAUSAS
8	ZONA	Zona	1=URBANA 2=RURAL
9	NUM_LESIONADO	Número de lesionados	Numérico
10	NUM_FALLECIDO	Número de fallecidos	Numérico
11	TOTAL_VICTIMAS	Total de víctimas	Numérico

Figura Nro. 14.-Tabla de Metadatos del INEC
Elaborado por. ANT
Fuente.- Agencia Nacional de Transito

Las variables parroquia, mes, día, Hora, clase y causa, son explicativas, es decir, el contenido de las variables “explican” con detalle el suceso que ocurrió, por ello, lo que nos puede decir la lectura de estas variables es la característica del accidente de tránsito, y no se puede establecer un comportamiento con base en una cifra, dado el objetivo del estudio.

En los modelos de Regresión Logística, al igual en Regresión Lineal, en la modelización se asume que las variables son “cuantitativas”. Si una variable categórica se introduce en el análisis con las categorías {0,1,2,3}, la interpretación del coeficiente que acompañará a la misma indicará que para la categoría “2” el

efecto es el doble que para la categoría codificada como “1”, cuando esto puede no tener ningún sentido.

La solución es crear tantas variables como categorías menos 1, denominadas variables indicadoras o dummy. Las variables dummy toman únicamente el valor cero y uno, de forma que su combinación lineal expresa todos los niveles de la variable original. Si la variable a introducir tiene a niveles se necesitarán a-1 variables indicadoras X2, X3,…, Xa, que toman los valores, siendo así el primer nivel de la variable el nivel de referencia. Las a-1 variables explicativas se introducirán en la parte lineal de modelo con los correspondientes coeficientes.

1	0	0	…	0
2	1	0	…	0
3	0	1	…	0
…	…	…	…	…
	0	0	…	1

Figura Nro. 15.- Variables dummy o indicadoras
Elaborado por. Autor
Fuente.- Elaboración propia

Una variable continua puede ser introducida como tal en el modelo, sin necesidad de recodificación. Aparecerá en la parte lineal del modelo con su correspondiente coeficiente β.

3.2.2 Estrategia de Selección del Modelo

Cualquier modelo de regresión puede tener dos objetivos:

- **Predictivo**, en el que el interés del investigador es predecir lo mejor posible la variable dependiente, usando un conjunto de variables explicativas.

- **Estimativo**, en el que el interés se centra en estimar la relación de una o más variables independientes con la variable dependiente. El segundo objetivo es el más frecuente cuando se trata de encontrar factores determinantes de una enfermedad o un proceso.

El objetivo de elegir una estrategia de selección de modelos es la de definir diversos procedimientos de estimaciones o pronósticos de datos en una o varias series temporales a partir de información histórica. Estas técnicas no tratan de normalizan los datos para establecer el comportamiento de una o varias variables, sino que realizan un análisis para la construcción de un modelo conceptual en la que se generen resultados a través de la serie. Por ello, para este estudio se consideran los siguientes modelos conceptuales: modelo estacionario, con tendencia lineal y con estacionalidad (González, 2017).

Las hipótesis que definen la validez del modelo conceptual son, la estabilidad en el indicador de validación (> al 90%) y que valor tome la variable en cualquier período t con una perturbación aleatoria definida en el modelo (5% de variación) (González, 2017).

El modelo desarrollado para la investigación es el predictivo debido a que se establece como dependientes diversas variables explicativas tales como la hora de los accidentes de tránsito, la parroquia, el día y el mes del suceso, y en el capítulo de desarrollo de la investigación se determina como se estable la significancia de las variables para el modelo final.

Mediante este proceso seleccionaremos qué variables explicativas entran en el modelo a través de un proceso sistemático de introducción y eliminación de variables comparándolos en cada paso con los dos modelos anidados obtenidos.

La estrategia seguida en la selección del modelo que sugiere Collet (Collett, 2013) ya que las rutinas automáticas en los software estadísticos no siempre producen el modelo más adecuado, es el uso de la regresión por pasos, para ello

primero se realizó un análisis del comportamiento de las variables para establecer si se definía el modelo por distribuciones.

Dadas las variables la única que se puede proyectar y establecer una distribución es la de resultado, es decir la de “nro. de accidentes” por lo que las demás son de carácter explicitas, es decir que no varía el resultado de que si ocurrió o no un accidente puesto que el evento se generó, es por ello que técnicas como una red neuronal, o modelos en función del comportamiento de las variables no tienen un ajuste especifico.

Por ello, el modelo a usar debe de tener las características de “leer” las variables explicativas y en función de ellas establece un patrón binario de respuesta, es decir, una probabilidad de que exista o no el evento de accidente de tránsito, de allí que los modelos binarios son la respuesta para este problema.

Una de las funciones básicas para empezar a plantear el modelo es la función de verosimilitud, que resume la información que los datos contienen acerca de los parámetros desconocidos en un modelo, por tanto un estadístico resumen es el valor de la función de verosimilitud (Lorenzo, 2017).

Función de Verosimilitud

La función de verosimilitud (o, simplemente, verosimilitud) es una función de los parámetros de un modelo estadístico que permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones.

Es decir, Dada una muestra observada $X1, X2, \dots \dots Xn$ y una ley de probabilidad p0*,* la verosimilitud cuantifica la probabilidad de que las observaciones provengan efectivamente, de una muestra (teórica) de la p0 (Lorenzo, 2017).

Tomemos el ejemplo de lanzar 10 veces una moneda. La muestra binaria observada es, por ejemplo:

0	1	1	0	1	1	1	0	0	1

Para una muestra de tamaño 10, la ley de Bernoulli de parámetro p, la probabilidad de una tal realización es de $p^6(1-p)^4$.

3.2.3 Capacidad predictiva del Modelo

Para valorar la capacidad predictiva de los modelos utilizamos un procedimiento que evalúa la capacidad de discriminación. Esta herramienta será utilizada en las dos bases de datos analizadas (base del estudio y base posterior que ya posea la probabilidad de ocurrencia del evento) y con ello obtendremos una validación externa del modelo (INEGI, 2010).

La discriminación consiste en la habilidad del modelo para ordenar a las personas según su odds (El odds ratio (OR) expresa la probabilidad de ocurrencia de un evento), de modo que los de mayor riesgo de accidente obtienen una mayor probabilidad de sufrir un siniestro. Esta discriminación es fácilmente cuantificable utilizando el índice de concordancia, que es la versión no paramétrica del área bajo la curva de ROC (Receiver Operating Characteristic), cuyo rango oscila entre 0,5 (discriminación aleatoria) a 1 (perfecta discriminación). Esta área representa la probabilidad de que cuando dos personas son aleatoriamente seleccionadas, el individuo con el mejor pronóstico sufrirá un accidente antes que el otro (INEGI, 2010).

Luego de ello se realiza lo que se denomina “Matriz de confusión” es decir, se realiza el análisis partiendo del cálculo de las probabilidades de una muestra de la base de eventos para luego establecer un límite o punto de corte entre la ocurrencia o no del evento.

	SI	NO
SI	K1	K2
NO	N1	N2

Figura Nro. 16.- Matriz de Confusión
Elaborado por. Autor
Fuente.- Elaboración propia

Lo que se realiza en esta matriz es establecer mediante la base el número de elementos que ocurren y que no ocurren, para luego definir mediante esa muestra cuales realmente fueron evento y cuales no ocurrieron mediante un análisis de probabilidades de un modelo establecido para el efecto.

El resultado de esta comparación es la de la significancia de los resultados de la matriz y del error esperado, con ello se establece que si los resultados de la matriz concuerdan más allá del 90% de datos quiere decir que el modelo predice de forma correcta, sin embargo el error debe establecerse en un parámetro menor al de 5% para que el modelo funcione, si existe un incremento en el error establecido del modelo se deberá calibrar en función de ventanas de tiempo o de pesos de las variables (INEGI, 2010).

3.2.4 Descripción de los Datos

Este estudio se ha realizado sobre 9858 personas, de accidentes de tránsito entre el 1 de Enero 2016 y el 31 de Diciembre del 2018.

Con la presente investigación se analiza la probabilidad de que exista un suceso (accidente), también es de vital interés para este estudio determinar qué variables influyen significativamente en la ocurrencia del suceso.

Las variables consideradas apropiadas para el estudio se han seleccionado teniendo en cuenta los factores que pueden influir en los accidentes ya sea por el factor humano como por factores externos al mismo, estos datos registrados para cada caso forman las llamadas variables explicativas, variables independientes o posibles factores pronóstico. En el modelo de regresión logística la variable dependiente será “accidentes en entre los años 2016-2018”.

A continuación se detallan las variables definidas que constituyen la base de datos:

Factor de Tiempo

Hora: Define la hora del evento (accidente)

CODIGO	RANGO	
1	00:00	00:59
2	01:00	01:59
3	02:00	02:59
4	03:00	03:59
5	04:00	04:59
6	05:00	05:59
7	06:00	06:59
8	07:00	07:59
9	08:00	08:59
10	09:00	09:59
11	10:00	10:59
12	11:00	11:59
13	12:00	12:59
14	13:00	13:59
15	14:00	14:59
16	15:00	15:59
17	16:00	16:59
18	17:00	17:59
19	18:00	18:59
20	19:00	19:59
21	20:00	20:59
22	21:00	21:59
23	22:00	22:59
24	23:00	23:59

Figura Nro. 17.- Rangos de Hora de Accidentes
Elaborado por. Autor
Fuente.- Elaboración propia

Día: Establece el día del evento (accidente)

CODIGO	DESCRIPCION
1	LUNES
2	MARTES
3	MIÉRCOLES
4	JUEVES
5	VIERNES
6	SÁBADO
7	DOMINGO

Figura Nro. 18.- Días de accidentes
Elaborado por. Autor
Fuente.- Elaboración propia

Mes: Se asigna el mes del evento (accidente)

CODIGO	DESCRIPCION
1	ENERO
2	FEBRERO
3	MARZO
4	ABRIL
5	MAYO
6	JUNIO
7	JULIO
8	AGOSTO
9	SEPTIEMBRE
10	OCTUBRE
11	NOVIEMBRE
12	DICIEMBRE

Figura Nro. 19.- Mes de Accidentes
Elaborado por. Autor
Fuente.- Elaboración propia

Factor de Espacio

Parroquia: Establece el lugar donde ocurrió el evento (accidente) en la ciudad de Guayaquil.

CODIGO	PARROQUIA
1	Parroquia Ayacucho
2	Parroquia Bolívar(Sagrario)
3	Parroquia Carbo(Concepción)
4	Parroquia de Chongón
5	Parroquia Febres Cordero
6	Parroquia García Moreno
7	Parroquia Letamendi
8	Parroquia Nueve de Octubre
9	Parroquia Olmedo (SanAlejo)
10	Parroquia Pascuales
11	Parroquia Roca
12	Parroquia Rocafuerte
13	Parroquia Sucre
14	Parroquia Tarqui
15	Parroquia Urdaneta
16	Parroquia Ximena

Figura Nro. 20.- Parroquias de Guayaquil
Elaborado por. Autor
Fuente.- Elaboración propia

Factor Causalidad

Clase de Accidente: Define el tipo de accidente que ocurrió en la ciudad

CODIGO	CLASE
1	ATROPELLOS
2	CAÍDA DE PASAJEROS
3	CHOQUES
4	ESTRELLAMIENTOS
5	ROZAMIENTOS
6	VOLCAMIENTOS
7	PÉRDIDA DE PISTA
8	OTROS

Figura Nro. 21.- Clases de Accidentes
Elaborado por. Autor
Fuente.- Elaboración propia

Causa de Accidente: Define la causa del accidente tipificada en categorías.

CODIGO	CAUSA
1	EMBRIAGUEZ O DROGA
2	MAL REBASAMIENTO INVADIR CARRIL
3	EXCESO VELOCIDAD
4	IMPERICIA E IMPRUDENCIA DEL CONDUCTOR
5	IMPRUDENCIA DEL PEATÓN
6	DAÑOS MECÁNICOS
7	NO RESPETA LAS SEÑALES DE TRÁNSITO
8	FACTORES CLIMÁTICOS
9	MAL ESTADO DE LA VÍA
10	OTRAS CAUSAS

Figura Nro. 22.- Causas de Accidentes
Elaborado por. Autor
Fuente.- Elaboración propia

Nro. de Accidentes: Define el número de accidentes ocurridos.

3.2.5 Agrupación de Variables y Definición de Variables Dummy

Las variables de los modelos quedarían definidas de la siguiente forma:

- La variable nro. de accidentes se introducen como variable continua.
- Para introducir las variables categóricas con tres o más niveles es necesario definir las variables dummy o indicadoras, tal y como se explicó en el apartado anterior de establecer variables dummy. En estos casos los niveles de referencia vendrán determinados por aquellas variables indicadoras que toman el valor cero.
- Las variables categóricas con dos o menos niveles se definen de la siguiente forma: Mes (0= mes donde no ocurrió el evento, 1=mes de la ocurrencia del evento), Día (0= día donde no ocurrió el evento, 1=día de la ocurrencia del evento), Clase de Accidente (0= clase invalida del evento, 1=clase de accidente que suscitó del evento), Sector (0= sector donde no ocurrió el evento, 1=sector donde ocurrió el evento), Hora (0= hora donde no ocurrió el

evento, 1=hora de la ocurrencia del evento), Causa del Accidente (0= causa no valida del evento, 1=causa valida del evento).

La población objeto de estudio está representada por 9858 registros de accidentes de tránsito en la ciudad de Guayaquil entre los años 2016 a 2018, proporcionada por la Agencia Nacional de Tránsito y contrastada por los datos muestrales del INEC (Transito, 2018).

4.- DESARROLLO DE LA PROPUESTA Y RESULTADOS

Con el objetivo de establecer la construcción de un modelo de comportamiento estadístico que nos permita generar una proyección de la ocurrencia o no de un evento de tránsito en una determinada parroquia de la ciudad de Guayaquil, se toma como input la descripción de las variables del capítulo anterior, así como su agrupación.

4.1 Análisis Descriptivo de las Variables

Para analizar los resultados de la base referencial de accidentes de tránsito, en primer lugar se ha realizado un análisis descriptivo expresando los resultados como media (± desviación estándar) para variables cuantitativas y con proporciones para variables cualitativas.

Como ejemplo, una de las variables para el análisis es la **Variable Hora**, en la que se establecen grupos de horas para reducir la dispersión de los datos, luego se realizó una contabilización del número de eventos de transito producidos en los grupos tal como se muestra a continuación:

GRUPO DE HORA	OCURRENCIA	
0-3	986	
4-7	1196	1.212981744
8-11	1822	1.523411371
12-15	1929	1.058726674
16-19	2007	1.040435459
20-24	1917	0.955156951

Figura Nro. 23.- Datos Agrupados de variable "HORA"
Elaborado por. Autor
Fuente. Base de datos de la Agencia Nacional de Transito

Luego, se elabora el análisis descriptivo de la variable en función de los estadísticos desviación estándar, media aritmética y coeficiente de correlación, tal como se muestra en la siguiente tabla:

DESVEST	0.2244
MEAN	1,643
COEF. CORR	(0.29061)

Figura Nro. 24.- Tabla de Estadísticos de Grupo de Horas
Elaborado por. Autor
Fuente. Base de datos de la Agencia Nacional de Transito

4.2 Calculo de Coeficientes de las Variables en R Studio

En este proceso se seleccionara, qué variables explicativas aplican en el modelo a través de un proceso sistemático de significancia y eliminación de variables comparándolos en cada paso los estadísticos resultantes del análisis del modelo de regresión logística.

En la comparación de los dos modelos se ha utilizado el estadístico -2logL ( ) (menos 2 veces el Logaritmo Neperiano de la función de verosimilitud L).

Una vez obtenida la relación lineal entre el logaritmo de los ODDs y la variable predictora X, se tienen que estimar los parámetros β0 y β1; la combinación óptima de valores será aquella que tenga la máxima verosimilitud (maximum likelihood ML),

es decir el valor de los parámetros β0 y β1con los que se maximiza la probabilidad de obtener los datos observados (DAVILA, 2015).

Otra forma para ajustar un modelo de regresión logística es empleando descenso de gradiente. Si bien este no es el método de optimización más adecuado para resolver la regresión logística, está muy extendido en el ámbito del machine learning para ajustar otros modelos, para este análisis no se toma en consideración esta prueba de ajuste.

Existen diferentes técnicas estadísticas para calcular la significancia de un modelo logístico en su conjunto (p-value del modelo). Todos ellos consideran que el modelo es útil si es capaz de mostrar una mejora respecto a lo que se conoce como modelo nulo, el modelo sin predictores, solo con β0, dos de los más empleados son:

- Wald chi-square: está muy expandido pero pierde precisión con tamaños muestrales pequeños.
- Likelihood ratio: usa la diferencia entre la probabilidad de obtener los valores observados con el modelo logístico creado y las probabilidades de hacerlo con un modelo sin relación entre las variables. Para ello, calcula la significancia de la diferencia de residuos entre el modelo con predictores y el modelo nulo (modelo sin predictores) (DAVILA, 2015).

Para determinar la significancia individual de cada uno de los predictores introducidos en un modelo de regresión logística se emplea el estadístico Z y el test Wald chi-test. En R, este es el método utilizado para calcular los p-values que se muestran al hacer summary() del modelo, adicional a ello se usa las siguientes instrucciones:

```
data <- subset(Base.Accidentes,select=c(1,2,3,4,5,6,7,8,9,10))
View (data) “Cargo el data set de la base de datos”
train <- data[1:1000,] “partición de la base para entrenar al modelo”
```

test <- data[1001:9857,] “partición de la base para calcular al modelo”

model <- glm(Grupo.6 ~.,family=binomial(link='logit'),data=train) “instrucción para ejecutar el modelo Logit”

summary(model) “Output de resultados”

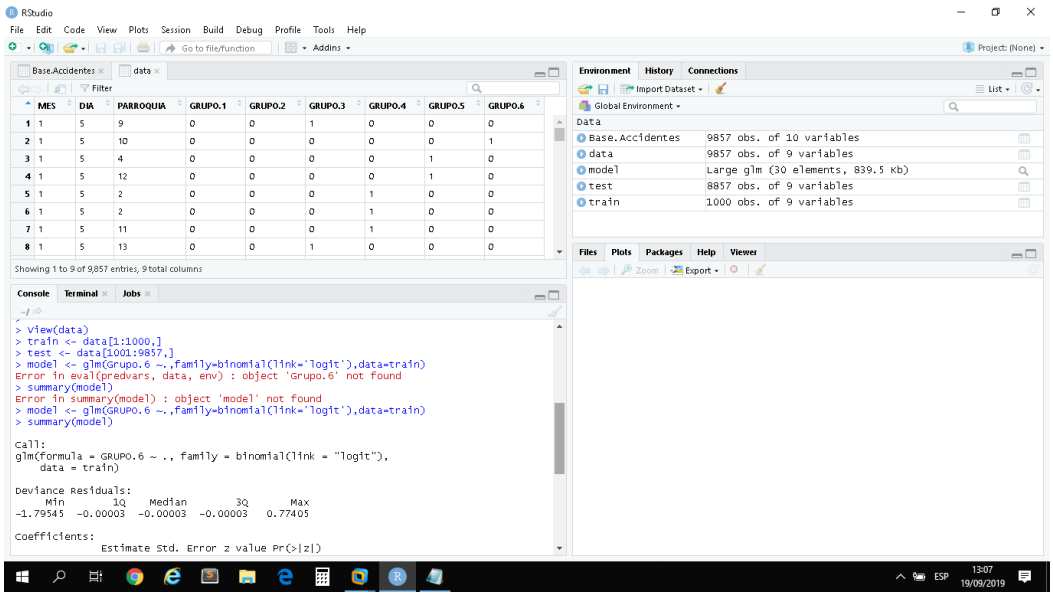


Figura Nro. 25.- Calculo de Coeficientes en R
Elaborado por. Autor
Fuente.- R Studio

Una vez realizada esta modelización se obtiene el modelo de Cox donde están recogidos los coeficientes estimados correspondientes a las variables seleccionadas, los correspondientes errores estándar, el estadístico de Wald con su p-valor asociado, el valor estimado de exp( ) y un intervalo de confianza al 95% de esta cantidad.

Para obtener los resultados del modelo, es necesario dividir el proceso en cuatro pasos:

Primero, se ajusta todos los modelos que contienen una única variable. Los valores de -2log para todos estos modelos se comparan con el correspondiente al modelo nulo (modelo sin variables) para conocer qué variables reducen significativamente el valor de este estadístico. En el gráfico adjunto se recogen las

variables que al ser introducidas por separado en el modelo nulo han reducido el valor de -2log con un p-valor<0,05, dando como resultado las variables Hora, Mes, Día y Parroquia.

	coeff b	s.e.	Wald	p-value	exp(b)	lower	upper	
Intercept	2.53605428	0.31158442	66.2469115	0.00000	12.6297391			
MES	-0.04166738	0.02443523	2.90776637	0.08815389	0.95918877	0.91433381	1.00624421	**VARIABLE SIGNIFICATIVA**
DIA	-0.0824106	0.04351239	3.5870731	0.05823072	0.92089376	0.84561313	1.00287625	**VARIABLE SIGNIFICATIVA**
PARROQUIA	0.03872141	0.01981932	3.81702195	0.05073434	1.03948085	0.99987626	1.08065417	**VARIABLE SIGNIFICATIVA**
2do GRUPO	-23.3629204	1223.60148	0.00036456	0.98476646	7.1386E-11	0	0	
3ro GRUPO	-23.4411603	1061.25676	0.00048788	0.98237767	6.6014E-11	0	0	
4to GRUPO	-23.4720458	950.661409	0.00060961	0.98030205	6.4006E-11	0	0	
5to GRUPO	-23.4741597	893.089373	0.00069086	0.97903064	6.3871E-11	0	0	
6to GRUPO	-23.4544225	891.096906	0.00069279	0.97900143	6.5144E-11	0	0	

Figura Nro. 26.- Calculo de p-value en R
Elaborado por. Autor
Fuente.- R Studio

En R Studio podemos ejecutar estos odds con la siguiente función para poder visualizar por pantalla:

anova(model, test="Chisq")

Donde model es el modelo que se ejecutó con la cláusula glm() y se define el test de pruebas, la más usual es usar Chi-cuadrado.

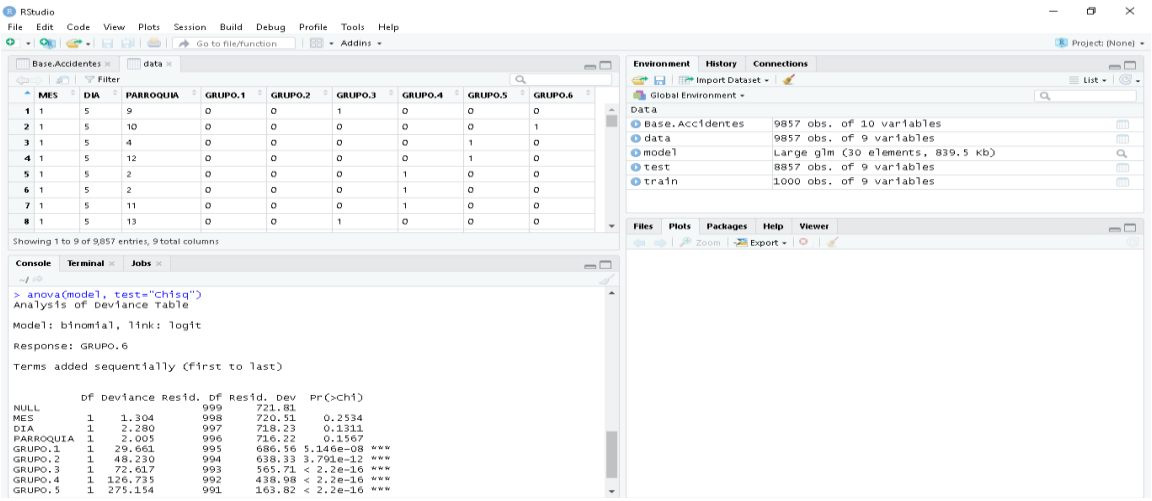


Figura Nro. 27.- Análisis ANOVA en R
Elaborado por. Autor
Fuente.- R Studio

Como **Segundo** paso, se puede observar en la corrida de las variables se puede identificar que las variables mes, día y parroquia son las que define la teoría como significativas, cabe recalcar que para realizar el análisis se definió realizar 10000 iteraciones de cada generación del modelo por cada variable, es decir, se corrieron 180 escenarios con diversos valores aleatorios, pero si se desea generar más de 20 escenarios, depende del procesamiento del equipo para hacerlo.

Luego, como **tercer** paso, se toma los coeficientes de las variables que han resultado relevantes se ajustan al modelo juntas. En ocasiones ciertas variables pueden dejar de ser significativas cuando se ajusta junto a otras. Por lo tanto, aquellas variables que no incrementen significativamente el valor de -2log cuando son omitidas del modelo y serán descartadas, para este modelo se descartaron las variables de clase y causa de accidentes.

	Coeff
Intercept	2.53605428
MES	-0.04166738
DIA	-0.0824106
PARROQUIA	0.03872141
2do GRUPO	-23.3629204
3ro GRUPO	-23.4411603
4to GRUPO	-23.4720458
5to GRUPO	-23.4741597
6to GRUPO	-23.4544225

Figura Nro. 28.- Coeficientes de variables del modelo propuesto
Elaborado por. Autor
Fuente.- R Studio

LL0	-4034.93276
LL1	-511.247927

Figura Nro. 29.- Coeficiente de determinación
Elaborado por. Autor
Fuente.- R Studio

Chi-Sq	7047.36967
df	8
p-value	0
alpha	0.05
sig	yes
R-Sq (L)	0.87329456
R-Sq (CS)	0.51082425
R-Sq (N)	0.91377002
Hosmer	926.772437
df	5342
p-value	1
alpha	0.05
sig	no

Figura Nro. 30.- Estadísticos de validación de significancia
Elaborado por. Autor
Fuente.- R Studio

Adicional a ello, se tienen en cuenta aquellas variables que no resultaron significativas por ellas solas frente al modelo nulo (primer paso) debido a que ahora pueden ser significativas con las resultantes del modelo del segundo paso, en el modelo se categorizo a la variable hora produciendo mejor ajuste.

Covariance Matrix

0.09708485	-0.00413662	-0.00846786	-0.00271697	-0.00721364	-0.01383377	-0.01594794	-0.01687511	-0.01459921
-0.00413662	0.00059708	-1.511E-05	-1.5455E-05	0.00053887	0.00047358	0.00057237	0.00055334	0.00049855
-0.00846786	-1.511E-05	0.00189333	-3.3046E-05	-0.00017913	0.00153659	0.001921	0.00208274	0.00169412
-0.00271697	-1.5455E-05	-3.3046E-05	0.00039281	-0.00037685	-0.0004963	-0.00054815	-0.00050722	-0.00051604
-0.00721364	0.00053887	-0.00017913	-0.00037685	1497200.59	0.00772632	0.00782131	0.00774961	0.00775025
-0.01383377	0.00047358	0.00153659	-0.0004963	0.00772632	1126265.91	0.00962705	0.00969343	0.00935087
-0.01594794	0.00057237	0.001921	-0.00054815	0.00782131	0.00962705	903757.114	0.01026295	0.00983347
-0.01687511	0.00055334	0.00208274	-0.00050722	0.00774961	0.00969343	0.01026295	797608.629	0.00991072
-0.01459921	0.00049855	0.00169412	-0.00051604	0.00775025	0.00935087	0.00983347	0.00991072	794053.696

Figura Nro. 31.- Matriz de covarianzas de las variables del modelo
Elaborado por. Autor
Fuente.- R Studio

En R Studio se usa el análisis de correlación para establecer las matrices que comparan las variables que se maneja en el modelo de accidentes, el código para ejecutar la instrucción es la siguiente:

cor(Base.Accidentes)

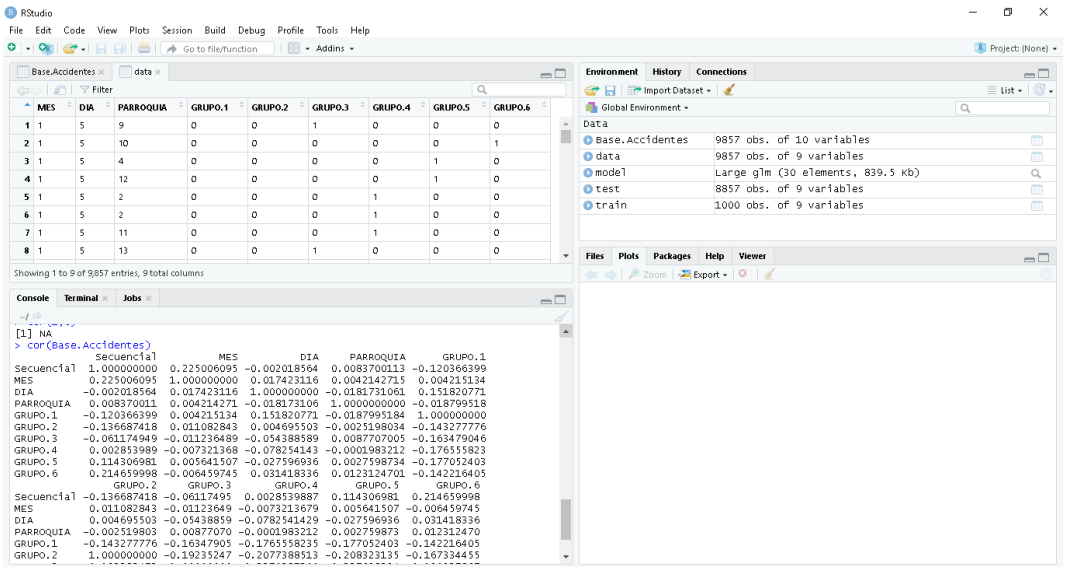


Figura Nro. 32.- Coeficiente de Correlación de variables del modelo
Elaborado por. Autor
Fuente.- R Studio

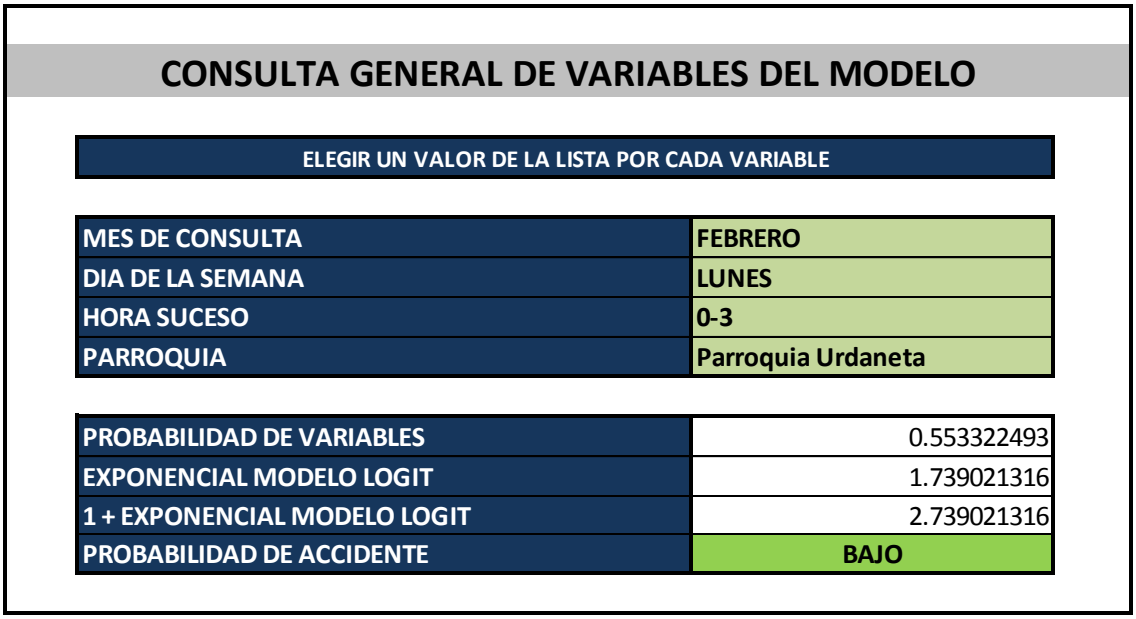
Como **cuarto** paso, se realiza la multiplicación de los coeficientes con cada variable dummy para generar la probabilidad, dependiendo de cálculo de cada coeficiente en función de la variable se obtiene la siguiente ecuación que representa al modelo de predicción de la investigación:

$$probabilidad\ p = k_0(rango\_hora) + k_1(mes) + k_2(dia) + k_3(parroquia)$$

Donde k_0, k_1, k_2, k_3, son los coeficientes obtenidos del análisis de regresión logística, los valores de las variables son la codificación asignada en el estudio, y p es la probabilidad sumada de todas las variables.

Para finalizar, y poder establecer la probabilidad del evento de riesgo de accidente se define la función exponencial ROC de la regresión, que será el resultado de la probabilidad del evento y se lo calcula mediante la siguiente formula.

$$probabilidad\ final = \frac{e^p}{e^p + 1}$$



CONSULTA GENERAL DE VARIABLES DEL MODELO

ELEGIR UN VALOR DE LA LISTA POR CADA VARIABLE	
MES DE CONSULTA	FEBRERO
DIA DE LA SEMANA	LUNES
HORA SUCESO	0-3
PARROQUIA	Parroquia Urdaneta

PROBABILIDAD DE VARIABLES	0.553322493
EXPONENCIAL MODELO LOGIT	1.739021316
1 + EXPONENCIAL MODELO LOGIT	2.739021316
PROBABILIDAD DE ACCIDENTE	BAJO

Figura Nro. 33.- Modelo de Proyección de Accidentes con 4 Variables
Elaborado por. Autor
Fuente.- Elaboración Propia

Si se desea establecer una probabilidad de accidentes de tránsito por variable disgregada, es decir, por cada una por separado, el modelo no tiene una significancia alta para establecer la efectividad en función de 1 sola variable, sin embargo, como objetivo académico, se estimó la probabilidad co n los resultados obtenidos del modelo logístico múltiple obteniendo el modelo de la siguiente forma:

$$probabilidad\ p = k_{variable}(variable)$$

Donde k es el coeficiente obtenido de la regresión logística, variable es la codificación definida en el modelo y p es la probabilidad obtenida por la variable definida para consulta.

Luego se define la exponencial tal como se realiza para este tipo de modelos, pero con una variante que es la de adicionar al final la desviación estándar o el coeficiente de correlación de la variable analizada según lo establezca la siguiente definición:

"Si la desviación estándar de la variable es "mayor que 1" se toma el coeficiente de correlación de la variable explicativa, caso contrario se adiciona la desviación estándar."

$$probabilidad\ de\ accidente = \frac{e^{p(variable)}}{e^{p(variable)} + 1} + (desv.estandar|coef.correlación)\ variable$$

Dando como resultado la probabilidad del accidente, en función de cada variable explicativa, recalcando que se debería realizar para cada variable un estudio por separado para establecer la probabilidad de ocurrencia del evento.

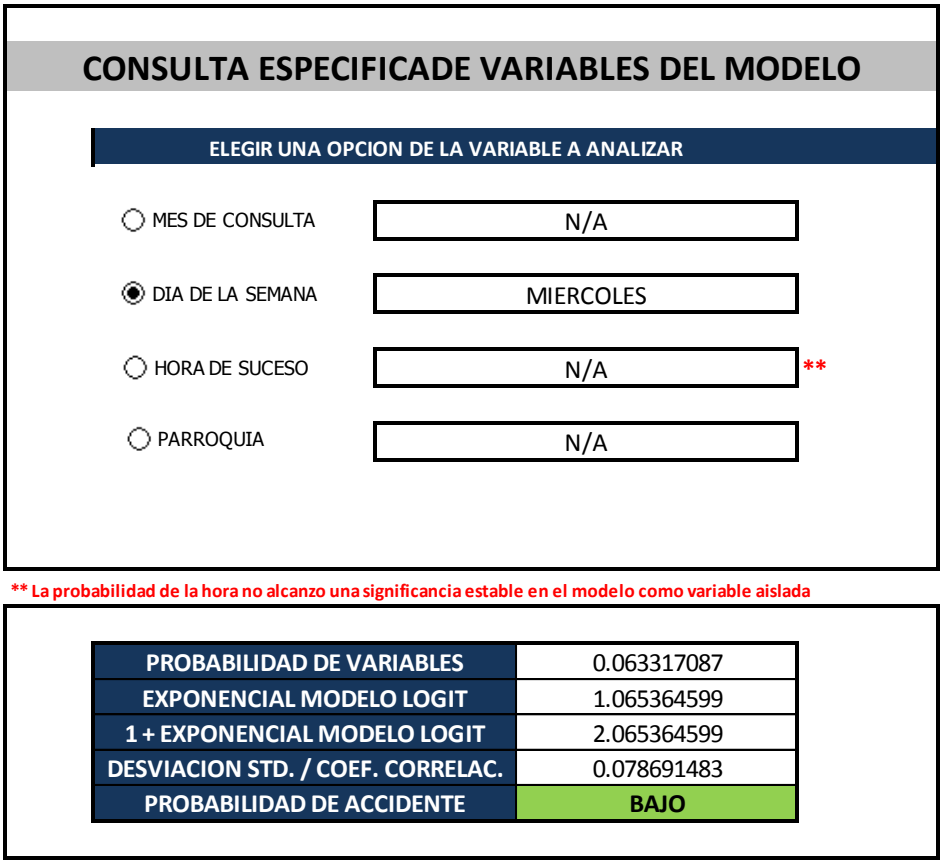


Figura Nro. 34.- Probabilidad de Accidente en función de cada Variable Explicativa
Elaborado por. Autor
Fuente.- Elaboración Propia

4.3 Semaforización

Para realizar la lectura de las probabilidades se estableció un proceso de semaforización para tener una visibilidad más clara acerca de la ejecución de planes para la mitigación de accidentes según la zona, hora, día o mes del evento.

La semaforización se la realizó en función de los niveles de riesgo, alto, medio y bajo el cual estadísticamente es representado por 3 percentiles, es decir, que dependiendo del percentil se le asignará el valor del nivel de riesgo, es por ellos que se definió según el análisis exploratorio un límite o threshold definiendo el sector del evento según como sigue:

PROBABILIDAD	NIVEL
0.00 a 70.00	BAJO
70.00 a 80.00	MEDIO
80.00 a 100.00	ALTO

Figura Nro. 35.- Semaforización de Probabilidades con 4 variables
Elaborado por. Autor
Fuente.- Elaboración Propia

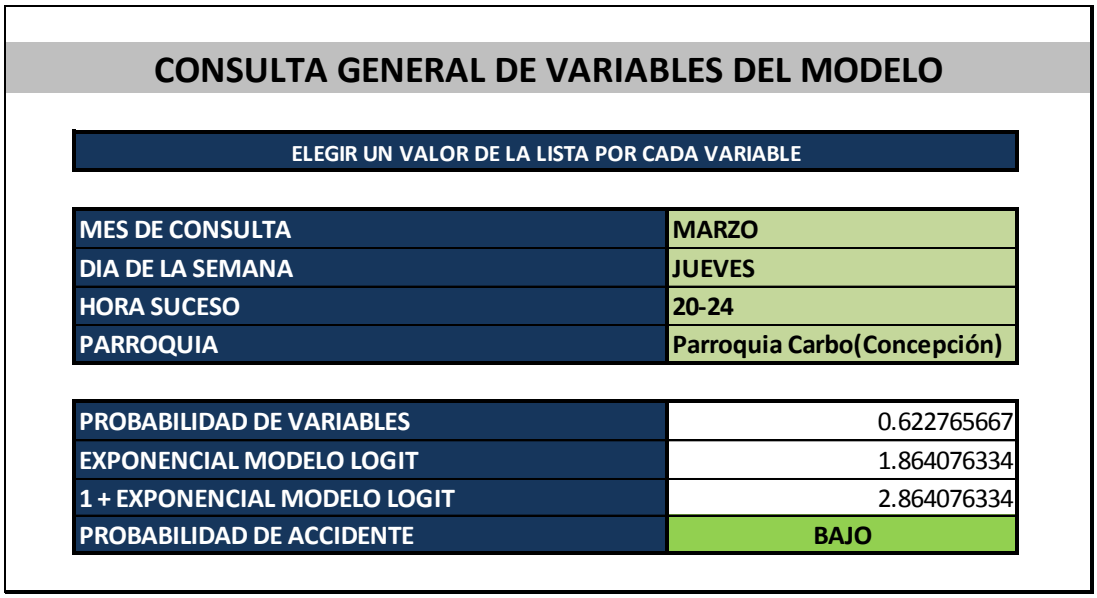
Esto quiere decir que la probabilidad baja es menor o igual de 70%, la probabilidad media se define entre 70% y 80% y la probabilidad alta se define en más del 80%.

CONSULTA GENERAL DE VARIABLES DEL MODELO

ELEGIR UN VALOR DE LA LISTA POR CADA VARIABLE	
MES DE CONSULTA	NOVIEMBRE
DIA DE LA SEMANA	JUEVES
HORA SUCESO	20-24
PARROQUIA	Parroquia Carbo(Concepción)

PROBABILIDAD DE VARIABLES	0.929901505
EXPONENCIAL MODELO LOGIT	2.534259553
1 + EXPONENCIAL MODELO LOGIT	3.534259553
PROBABILIDAD DE ACCIDENTE	MEDIO

Figura Nro. 36.- Probabilidad Media de 4 Variables
Elaborado por. Autor
Fuente. Modelo Matemático de Accidentes



CONSULTA GENERAL DE VARIABLES DEL MODELO

ELEGIR UN VALOR DE LA LISTA POR CADA VARIABLE	
MES DE CONSULTA	MARZO
DIA DE LA SEMANA	JUEVES
HORA SUCESO	20-24
PARROQUIA	Parroquia Carbo(Concepción)

PROBABILIDAD DE VARIABLES	0.622765667
EXPONENCIAL MODELO LOGIT	1.864076334
1 + EXPONENCIAL MODELO LOGIT	2.864076334
PROBABILIDAD DE ACCIDENTE	BAJO

Figura Nro. 37.- Probabilidad Baja de 4 Variables
Elaborado por. Autor
Fuente. Modelo Matemático de Accidentes

Nótese, en este ejemplo que en la figura 37 solo el mes de consulta “Marzo”, la probabilidad de ocurrencia es Baja, a diferencia de la figura 36 donde la probabilidad al mes de “Noviembre” es media.

Para el modelo de 1 variable, se establece un límite o threshold diferente, debido a que al ser de 1 sola variable no se define los rangos de Alto, medio o bajo sino solo una probabilidad binaria de ocurrencia, es decir “Alto” o “Bajo”, tal como sigue.

PROBABILIDAD	NIVEL
0.00 a 72.50	BAJO
72.51 a 100.00	ALTO

Figura Nro. 38.- Semaforización de Probabilidades con 1 variable
Elaborado por. Autor
Fuente.- Modelo Matemático de Accidentes

Esto quiere decir que la probabilidad baja es menor o igual de 72.50%, la y la probabilidad alta se define en más del 72.51%.

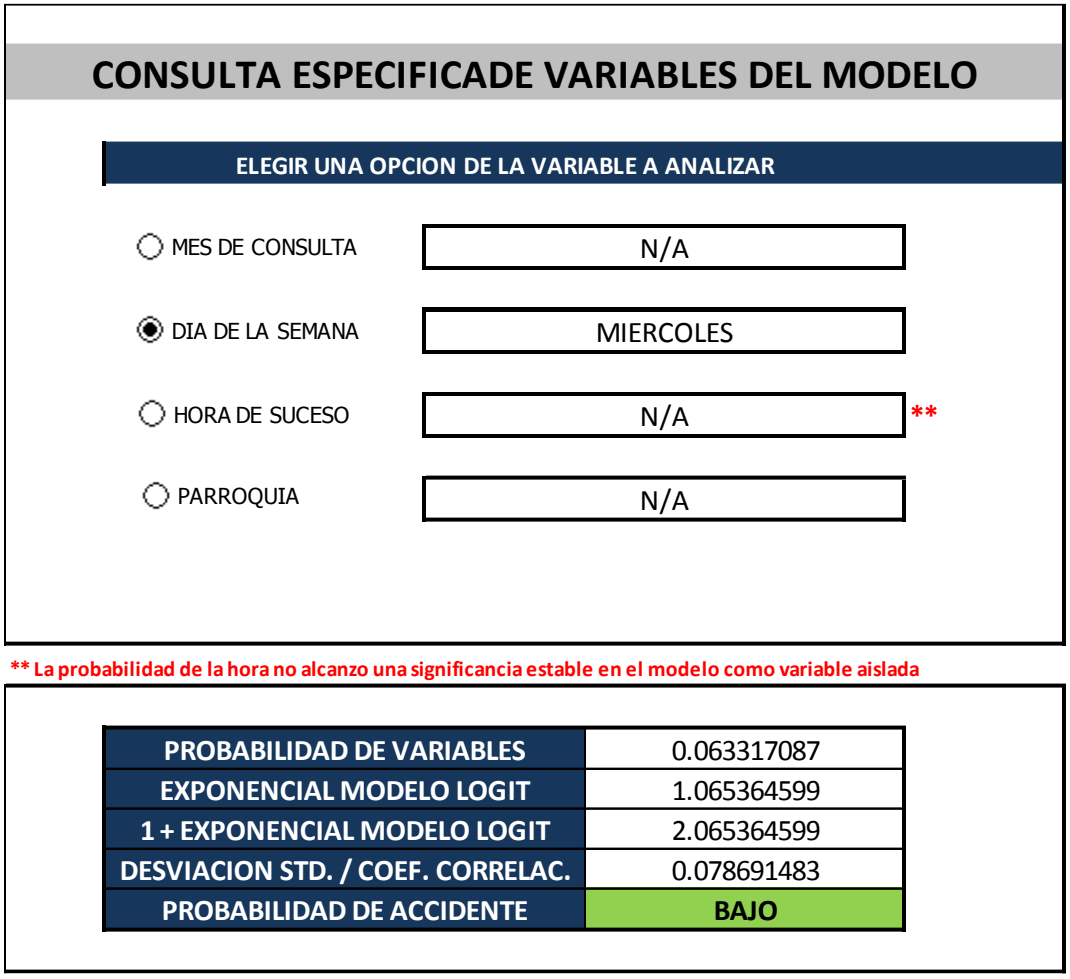


Figura Nro. 39.- Probabilidad Baja con 1 Variable
Elaborado por. Autor
Fuente. Modelo Matemático de Accidentes

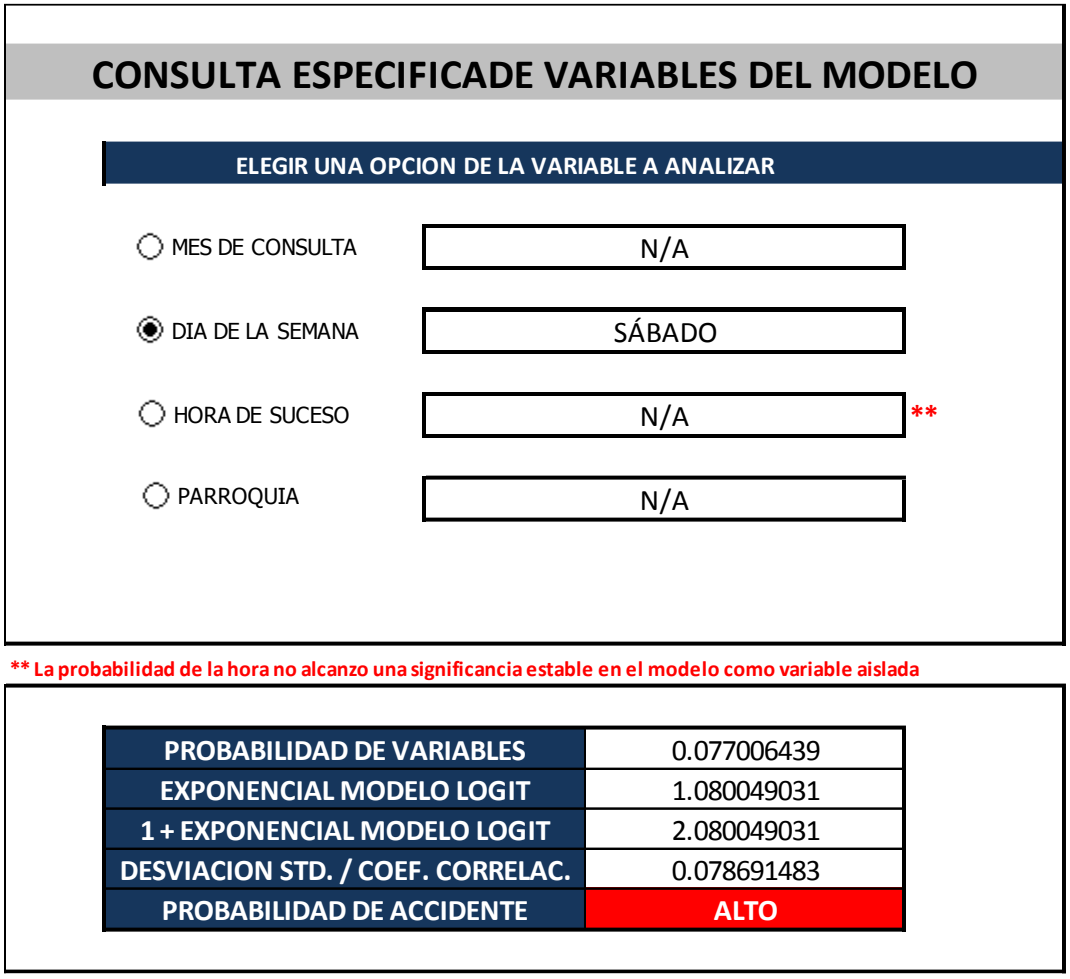


Figura Nro. 40.- Probabilidad Alta con 1 Variable
Elaborado por. Autor
Fuente. Modelo Matemático de Accidentes

4.4 Validación del Modelo

Para empezar con la implementación de los análisis se establecen los diversos procesos que nos ayudan a realizar el análisis exploratorio de los datos

Análisis Exploratorios de las Variables

Para definir el análisis exploratorio se debe establecer los criterios para su selección, según la característica del universo; puede ser de manera aleatoria

(dando la oportunidad a cualquier registro de ser elegido); de manera sistemática (dividiendo la población entre el tamaño de la muestra, obteniendo un valor que servirá para establecer un intervalo para recoger la muestra); por bloques (seleccionando cierta cantidad de registros por meses y aplicando la metodología sistemática en cada bloque); y por juicio del revisor tomando en cuenta los registros materiales u otro criterio.

Para el modelo de accidentes se define un análisis de muestra aleatoria, la cual, puede realizarse por diversos métodos tales como:

- Generación de números aleatorios para seleccionar los registros
- Sentencias aleatorias de la base de datos que estemos manejando los datos del modelo definido por el investigador.
- Funciones de aleatoriedad de Microsoft Excel.

Luego de definir con que método se generaran los valores elegidos se debe de calcular la muestra para definir el número de aleatorios que necesitamos para correr el método elegido sobre la base de datos.

Esto se lo realiza con la fórmula de la muestra para población finita, debido a que tenemos un límite de datos recolectados y es como sigue:

$$\frac{N * (\alpha_C * 0{,}5)^2}{1 + (e^2 * (N - 1))}$$

Dónde:

α_c = Valor del nivel de confianza (varianza)

· **Nivel de confianza**, es el riesgo que aceptamos de equivocarnos al presentar nuestros resultados (también se puede denominar grado o nivel de seguridad), el nivel habitual de confianza es del 95%.

e = Margen de error

· **Margen de error**, es el error que estamos dispuestos a aceptar de equivocarnos al seleccionar nuestra muestra; este margen de error suele ponerse en torno a un 3%.

N = Tamaño Población (universo)

El valor de la muestra para 9858 datos es de 370 registros, para realizar la evaluación del modelo.

Evaluación del modelo

Para evaluar si el modelo es válido, se analiza tanto el modelo en su conjunto como los predictores que lo forman. El modelo se considerará útil si es capaz de mejorar la predicción de las observaciones respecto al modelo nulo sin predictores. Para ello se analiza la significancia de la diferencia ("Deviance") de residuos entre ambos modelos ("Null deviance" y "Residual deviance"), con un estadístico que sigue la distribución chi-cuadrado con grados de libertad correspondientes a la diferencia de los grados de libertad de ambos modelos (Micheaux, 2017).

Es importante también analizar el porcentaje de predicciones correctas además de los falsos positivos y falsos negativos que hace nuestro modelo para evaluar su potencial. Para este ejemplo utilizaremos un threshold de 0,7. Si la probabilidad predicha de que el accidente de tránsito sea positivo es mayor de 0.7, la observación se asignará como "Positivo", caso contrario se asignará como "Negativo". Además de evaluar el test-error global, es conveniente identificar como se reparte este error entre falsos positivos y falsos negativos, ya que puede ocurrir que un modelo sea mucho mejor prediciendo en una dirección que en otra. Esto se ve directamente influenciado por límite de clasificación o threshold establecido, por ello, para evaluar el modelo se establecieron los siguientes parámetros:

- Punto de corte: 0.7
- Calculo de la muestra aleatoria: 370
- Nro. de Registros que cumplen hipótesis de >= 0.7: 218
- Nro. de Registros que cumplen hipótesis de < 0.7 : 152

Luego se realiza un análisis con el modelo propuesto de cada registro validando que el resultado sea la probabilidad "Alta" de ocurrencia de accidente con

el modelo de 4 variables, caso contrario si es “Media”, o “Baja”, será considerado como no ocurrencia de accidente, tal como se muestra a continuación:

	SI	NO
SI	218	12
NO	5	152

Figura Nro. 41.- Matriz de Dirección
Elaborado por. Autor
Fuente. Modelo Matemático de Accidentes

Esto quiere decir que de los 218 registros con probabilidad “Alta” 12 registros fueron en realidad probabilidades con niveles “Medio o Bajo”, y de los 152 registros con probabilidad “Medio o Bajo”, 5 registros fueron en realidad probabilidades con nivel “Alta”.

Luego se realiza una relación entre los falsos positivos y los positivos dando como resultado 94.78% de ajuste del modelo en función de la probabilidad “Alta” de que ocurra un evento de tránsito, el error del modelo se calcula con la relación entre el falso negativo y el negativo, dando como resultado 3.18% que es la distorsión que puede ocurrir en el modelo de predicción de eventos de tránsito.

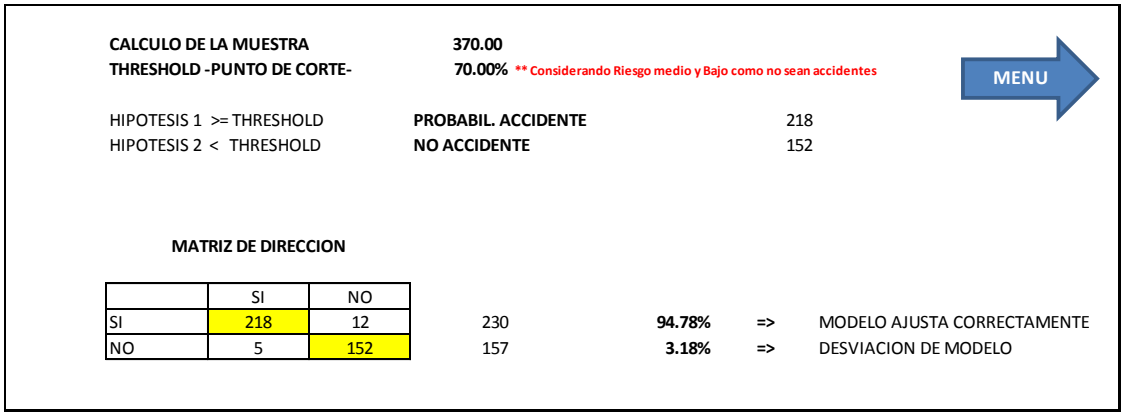


Figura Nro. 42.- Validación del modelo de accidentes
Elaborado por. Autor
Fuente.- Modelo de Validación

5.- CONCLUSIONES Y RECOMENDACIONES

El modelo se utilizó para identificar los factores predictores, es decir las variables que pueden influir en la ocurrencia de un accidente de tráfico. Además, estudia si la relación entre las variables predictoras y la probabilidad de sufrir un accidente es significativa. En el presente proyecto, han resultado significativas las variables referentes al mes, el día, la hora y la parroquia donde se suscitó el evento.

Una vez desarrollados los distintos modelos que nos han permitido definir las variables predictoras del riesgo, se han utilizado dos herramientas para visualizar fácilmente la información obtenida.

El modelo de regresión logística nos ayudó a establecer niveles de riesgos para que sirvan como un predictor de probabilidades de un posible evento de tránsito, esta herramienta se puede automatizar para que sea utilizado por las autoridades de tránsito para tomar acciones preventivas como correctivas y definir el riesgo inherente trasladándolo a proceso que busquen la educación al conductor de la ciudad.

Se recomienda realizar la calibración de los coeficientes del modelo de forma anual para que se cuente con las probabilidades por sector/lugar más actualizadas según lo que se encuentre en los históricos, e inclusive establecer distribuciones a las variables significativas y más adelante generar una red neuronal para definir modelos de contraste y ajuste del presente modelo.

Se recomienda automatizar los procesos de generación de registros de la base de datos en un App, para que se tenga libre acceso a la información y realizar pruebas de stress en el modelo que se implementó para automatizar los pesos de las variables de forma automática y en línea.

BIBLIOGRAFIA

A. Klar, D. Kuhne, R. Wegener, Mathematical Models for Vehicular Traffic.

Aixia Zhang, X. L. (2017). Study on traffic accidents prediction based on bp neural network. *Ieee access*, 19.

Abbas, M., Bullock, D., Head, L., & Trb. (2001). Real-time offset transitioning algorithm for coordinating traffic signals. Advanced Traffic Management Systems and Vehicle-Highway Automation 2001: Highway Operations, Capacity, and Traffic Control, 26-39.

Baena, G. M. (2014). *Metodología de la investigación.* Mexico D.F.: Grupo Editorial Patria.

Butler, D. C. (2016). Machine learning for molecular and materials science. *Nature*, 559.

C.K.S., L. (2019). *Big data analysis and mining. In Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics .* USA: IGI Global.

Cáceres, R. Á. (2017). *Probabilidad y estadística para ingenieros.* Chicago: Prentice-Hall.

Castro, T. d. (2018). *Modelo de Regresion Lineal Multiple.* Madrid: UOC.

Cerrolaza, M. (2012). *Modelos matemáticos en ingeniería moderna.* Merida: Universidad Central de Venezuela.

Churpek, M. M. (2018). BIG DATA AND DATA SCIENCE IN CRITICAL CARE. *researchgate*, 25.

Davila, C. D. (2015). *Ejercicios Resueltos de Econometría. .* Las Palmas - España: Delta Publicaciones.

Gardner, A. K., & Dunkin, B. J. (2019). Pursuing excellence: the power of selection science to provide meaningful data and enhance efficiency in selecting surgical trainees. Annals of surgery, 270(1), 188-192.

González, C. G. (2017). *Tratamiento de datos con R, SPSS y Estadistica.* Vigo-España: Diaz de Santos.

H., L. E. (2017). *Ingeniería del Tránsito.* Santiago: Universidad de Chile. Escuela de Ingeniería.

Haitao Zhao, H. Y. (2019). VEHICLE ACCIDENT RISK PREDICTION BASED ON AdaBoost-SO in VANETs. *IEEE Access*, 21.

INEGI. (2010). *Memorias del XIX Foro Nacional de Estadistica.* Mexico DF: INEGI.

J. Delgado, P. Saavedra y R. M. Velasco; “Taller de Modelación Matemática II”, (2011)

Kellerher, J. (2018). Data Science. *MIT Press*, 25.

K. Nagel, Particle Hopping Models and Traffic Flow Theory. Physical Review E. Vol. 53, Num. 5, (1996).

Landeta, J. M. (2017). *Modelos matemáticos para la toma de decisiones.* Mexico D.F.: Instituto Mexicano de Contadores Publicos.

Little, T. D. (2016). *An Asymmetric Logit Model to explain .* Chicago: The Oxford Handbook.
Lorenzo, J. M. (2017). *Introducción al análisis estadístico .* Madrid: Diaz de Santos.
Medrano, S. L. (2000). *Modelos matemáticos.* Trillas.
McNeil, D. R. (1968). A solution to the fixed-cycle traffic light problem for compound Poisson arrivals. Journal of Applied Probability, 624-635.
Micheaux, P. L. (2017). *Introducción al análisis de datos con R y R Commander.* Illinois: Springer.
Pedrycz, W., & Chen, S. M. (2019). Deep Learning: Concepts and Architectures
Rocha, C. M. (2018). *Metodología de la investigación.* Mexico: Oxford University.
Ross, S. M. (2014). *Introducción a la estadística.* Editorial Revelte.
Said, A., & Torra, V. (Eds.). (2019). Data Science in Practice. Springer
Smith, Hal, And Introduction to Delay Diferential Equations with Applications to the Life Sciences, Springer-Verlag, New York, 2011.
Salud, O. m. (2018). Boletin de Accidentes de Transito en el Mundo. *Principales causas de muerte de personas en el mundo*, 46.
Transito, A. N. (2018). Accidentes de transito: a) factores determinantes, b) recomendaciones, c) estadísticas. *Boletin Informativo*, 22.
Varsrvsarky, O. (2002). *Modelos matemáticos y experimentación numérica.* IICA.
Walpole, R. E. (2016). *Probabilidad y estadística para ingenieros .* Illinois: Pearson.
Zegarra, J. D. (2016). Tipos de traumatismo en fallecidos por accidentes de transito. *Universidad Nacional de San Agustin*, 38.
Zheng P., A. T. (2018). A unified framework for sparse relaxed regularized regression. *IEEE Access*, 1404.
ZHENG, M. (2019). TRAFFIC ACCIDENTS SEVERITY PREDICTION. *IEEE Access*, 14.
Zhou, Z. (2019). ATTENTION BASED STACK RESNET FOR CITYWIDE TRAFFIC ACCIDENT PREDICTION. *IEEE Access*, 25.

ANEXOS

ANEXO 1.- Carga de Datos en R Studio

Para realizar la carga de datos lo podemos realizar por consola de comandos o con el asistente para importar datos, en el proyecto se utiliza el asistente tal como sigue:

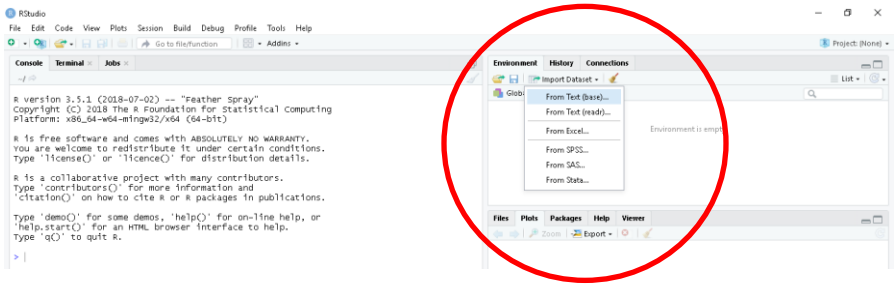


Figura Nro. 43.-Carga de Datos en RStudio
Elaborado por. Autor
Fuente.- R Studio

Se elige la opción “from Text (Based)” debido a que la información de la base de datos del modelo se encuentra en formato texto, y se elige los separadores y si el archivo tiene cabeceras o no tal como se muestra en la figura siguiente:

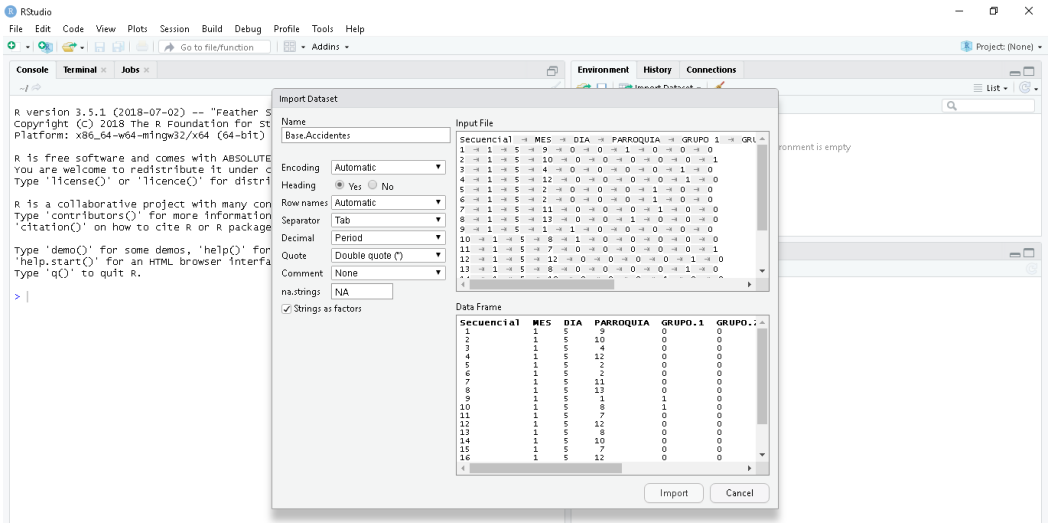


Figura Nro. 44.- Configuración de Carga de Datos
Elaborado por. Autor
Fuente.- R Studio

Al realizar el clic en “Import” se genera dentro de R studio la importación de los datos, y el código fuente para establecer en consola y la pre visualización del contenido del archivo en el entorno de la aplicación.

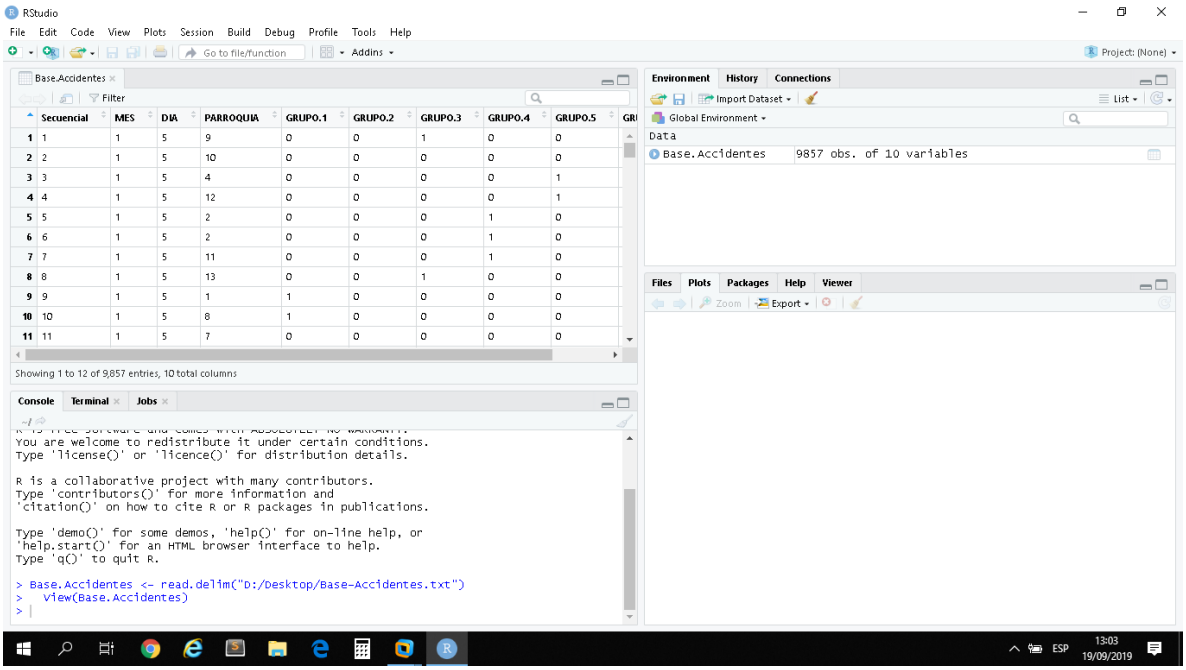


Figura Nro. 45.- Pre visualización de Carga de Datos
Elaborado por. Autor
Fuente.- R Studio

ANEXO 2.- Instalación de R Studio

Básicamente, para instalar R es necesario visitar su página web ubicada en el Uniform Resource Locator (URL) o dirección web www.r-project.org (González, 2017), y posteriormente bajar e instalar el paquete. En el margen izquierdo la página contiene una liga con la leyenda, “Download, Packages and CRAN”

Una vez que se le da click al link, se nos conduce a otra página con los principales servidores o mirrows de R en el mundo. Desde luego, hay que escoger aquí a un servidor que esté lo suficientemente cerca de nuestra ubicación geográfica, Acorde a lo anterior, se elige en enlace correspondiente al servidor, a través página web de la misma dirección URL y damos click en el enlace ”**The Comprehensive R Archive Network”** (González, 2017), en la sección titulada Download and Install R encontramos tres enlaces enumeradas como sigue:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

Dependiendo de nuestro sistema operativo, bajamos el paquete correspondiente y lo instalamos.

Instalación para Sistema Operativo Linux

Para Linux habrá dos paquetes de R que corresponden a sus correspondientes versiones de distribución ya sea Debian/deb, redhat/rpm, suse/rpm, ubuntu/deb. Para instalarlo, por ejemplo, en un sistema tipo debían, se abre una terminal en el sistema (González, 2017) y se le da el comando dpkg siguiente:

```
$ sudo dpkg -i r-base-core_2.11.0-1~etchcran.0_i386.deb
```

Después de darle la contraseña de superusuario para el comando que permite que la instalación se efectúe con los privilegios de un administrador de Linux, el programa R se instalará en nuestro sistema.

Para un sistema Linux tipo Red Hat [13] se instala como en el siguiente ejemplo en modo de superusuario

```
# rpm -i R-core-2.10.0-2.fc11.i586.rpm
```

Instalación para Sistema Operativo Windows

En un sistema Windows se le debe de dar doble click al archivo .exe, en modo de administrador. El archivo al tiempo de este escrito se llama como sigue:

R-3-0-1-win.exe

Luego, se le da doble click, después de lo cual se siguen las instrucciones y después de pasar por varias opciones de configuración, el programa R queda instalado

Instalación para Sistema Operativo Mac

En un sistema o Mac OS X se baja el paquete llamado

R-3-0-1-pkg

y se le da doble click, después de lo cual se siguen las instrucciones y después de pasar por varias opciones de configuración, el programa R queda instalado.

El paquete NCSTATS

Dentro del lenguaje R se pueden cargar paquetes, que añaden funcionalidad a R. Uno de estos paquetes es NCStats, que contiene las funciones necesarias de soporte para poder graficar, en nuestro caso, las distribuciones aplicables a las variables experimentales de eventos de accidentes (Micheaux, 2017).

El paquete contiene funciones y simulaciones para soportar la Estadística introductoria, para instalar NCStats dentro de R, es conveniente ejecutar un comando que va a explorar, los sitios de respaldo de R, para descargarlo e instalarlo.