DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

TEMA: MODELO DE PREDICCIÓN DE PUNTOS DE EXCESO DE VELOCIDAD EN LA AUTOPISTA GENERAL RUMIÑAHUI DE QUITO, A TRAVÉS DE TÉCNICAS DE GESTIÓN DE DATOS PARA EL CONTROL DE OCURRENCIA DE ACCIDENTES

AUTOR: ARELLANO AGUILAR, LUIS EDUARDO

DIRECTOR: ING. ALMACHE CUEVA, MARIO GIOVANNY

SANGOLQUÍ

2019

CERTIFICADO DEL DIRECTOR

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**MODELO DE PREDICCIÓN DE PUNTOS DE EXCESO DE VELOCIDAD EN LA AUTOPISTA GENERAL RUMIÑAHUI DE QUITO, A TRAVÉS DE TÉCNICAS DE GESTIÓN DE DATOS PARA EL CONTROL DE OCURRENCIA DE ACCIDENTES**" fue realizado por el señor **ARELLANO AGUILAR LUIS EDUARDO** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido, por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 28 de enero de 2019.

Atentamente,

Ing. Mario Almache

CC: 1708718950

DIRECTOR

AUTORÍA DE RESPONSABILIDAD

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

AUTORÍA DE RESPONSABILIDAD

Yo, **ARELLANO AGUILAR LUIS EDUARDO**, con cédula de identidad Nro. 172328441-8, declaro que el contenido, ideas y criterios del trabajo de titulación: "**MODELO DE PREDICCIÓN DE PUNTOS DE EXCESO DE VELOCIDAD EN LA AUTOPISTA GENERAL RUMIÑAHUI DE QUITO, A TRAVÉS DE TÉCNICAS DE GESTIÓN DE DATOS PARA EL CONTROL DE OCURRENCIA DE ACCIDENTES**" es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 28 de enero de 2019.

Arellano Aguilar Luis Eduardo

CC: 1723284418

AUTORIZACIÓN

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

AUTORIZACIÓN

Yo, **ARELLANO AGUILAR LUIS EDUARDO**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación "**MODELO DE PREDICCIÓN DE PUNTOS DE EXCESO DE VELOCIDAD EN LA AUTOPISTA GENERAL RUMIÑAHUI DE QUITO, A TRAVÉS DE TÉCNICAS DE GESTIÓN DE DATOS PARA EL CONTROL DE OCURRENCIA DE ACCIDENTES**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 28 de enero de 2019.

Arellano Aguilar Luis Eduardo

CC: 1723284418

DEDICATORIA

A mi madre Jexenia Aguilar, por ser mi confidente y amiga en todo este tiempo, sin su cariño y paciencia, nunca podría haber llegado tan lejos.

A mi padre Luis Arellano, por sus consejos, sus enseñanzas y por su vasto conocimiento, siempre será mi inspiración y ejemplo.

A mi abuela Elvira Vera, que desde el cielo me cubre con su bendición, llevo presente en mi corazón sus detalles, besos, abrazos y su amor.

A mi abuelo Marcelo Aguilar, que desde el cielo me guía por el camino del bien, recordaré siempre su alegría, sabiduría y afecto.

A mi ñaño Mario Arellano, que con su experiencia impartió grandes consejos en los momentos más idóneos.

A mi tío Rogelio Aguilar, por su eterno ejemplo de superación y esfuerzo, siempre le tendré un gran respeto y admiración.

A esa persona que estuvo a mi lado desde el inicio de este viaje, por su cooperación y cariño incondicional.

AGRADECIMIENTO

Agradezco a Dios por ser el guía de mi vida. A mis padres que han sido mi soporte todos estos años y que gracias a su eterno compromiso han logrado la realización de este proyecto.

A mis ñaños Edgar Arellano, José Vera y Mercedes Arellano, a mis primos Andrés Vera y Ernesto Montalvo, que a la distancia me acompañaron siempre con sus buenos deseos.

A mi primo Josué Aguilar, por ser un gran amigo y compañero, por brindarme su ayuda y motivación hasta en los momentos más difíciles.

A mi familia, por su tolerancia e impulso, los cuales fueron fundamentales para el cumplimiento y culminación de esta larga travesía.

A la Ingeniera Graciela Guerrero por su fiel amistad y gran conocimiento siendo un pilar para la cristalización de este objetivo.

¡Gracias!

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El crecimiento continuo de países, ciudades y comunidades han generado una saturación de los transportes públicos, obligando a las personas a la adquisición de vehículos con el fin de movilizarse de manera rápida y cómoda. Sin embargo esto acarrea una cantidad innumerable de automóviles en las autopistas, produciendo un sin número de accidentes de tránsito que afectan el día a día de las ciudades. Estos accidentes son causados por varios factores, entre ellos humano, ambiental, mecánico y vial. Este proyecto se enfoca en el desarrollo de un modelo de predicción de zonas de exceso de velocidad, con el fin de brindar un aporte para el control de accidentes. Se obtuvo información mediante el uso de una aplicación móvil diseñada para el sistema operativo Android, esta permitió obtener los datos requeridos de diferentes vehículos dentro de la Autopista General Rumiñahui. Para el análisis de la información se usó la herramienta RapidMiner, aquí se hizo un pre procesado de los datos previamente obtenidos para evitar aquellos valores irrelevantes o incorrectos para el caso de estudio, en la última etapa se evaluaron varias técnicas de predicción con el fin de obtener un resultado con el mejor índice de precisión.

Palabras clave:

- **ACCIDENTES DE TRÁNSITO**
- **TRÁGICOS**
- **EXCESO DE VELOCIDAD**
- **MINERÍA DE DATOS**
- **MODELOS DE PREDICCIÓN**
- **RAPIDMINER**

ABSTRACT

The continued growth of countries, cities and communities has generated a saturation of the public transportation service, this forced people to buy vehicles in order to move quickly and comfortably. However, this cause a lot of cars on highways and produce a lot of traffic accidents that affects the cities day by day. This accidents are inflicted by many factors, like human, environmental, mechanic and vial. The project focus on develop of a prediction model of speeding zones, as a contribution to accident control. Information was obtained through a mobile application designed for Android system, this allowed to obtain the required data of different vehicles on the General Rumiñahui Highway. For the data analysis was used the RapidMiner tool, here a preprocessing of the previously obtained data was done to avoid irrelevant or incorrect values for the study case, in the last stage several prediction techniques were evaluated in order to obtain a result with the best precision.

Keywords:

- **TRAFFIC ACCIDENTS**
- **TRAGIC**
- **EXCESO DE VELOCIDAD**
- **DATA MINING**
- **PREDICTION MODELS**
- **RAPIDMINER**

CAPÍTULO I

INTRODUCCIÓN

1.1.ANTECEDENTES

Las técnicas de predicción sirven para obtener estimaciones o pronósticos de valores futuros a partir de la información histórica contenida en un fenómeno observado para establecer un modelo de comportamiento de los distintos parámetros a ser tratados (Taamneh, Alkheder, & Taamneh, 2017). Diferentes artículos científicos plantean modelos que permiten simular y predecir siniestros en carreteras, dentro de los cuales se consideran diversos factores como posibles causantes de los diversos siniestros de tránsito.

En el Ecuador, el Ministerio de Transporte y Obras Públicas (MTOP), a través de sus entidades adscritas como la Agencia Nacional de Tránsito (ANT) y la Comisión de Tránsito del Ecuador (CTE), reportan alrededor de 25 mil accidentes de tránsito anualmente; con un aproximado de 1.000 personas fallecidas (ANT, 2018). Para reducir estos siniestros, el MTOP dispuso la implementación de radares y controles en las principales vías del país, así como la realización de campañas de educación vial. Ayudando a la gestión de los límites legales de tránsito vehicular los cuales son segmentados dependiendo del tipo de vía y vehículo (ANT, 2018), considerando que el exceso de velocidad es una de las causas de accidentes en el Ecuador (Ana Benitez, 2018).

Al aplicar un sistema de monitoreo de tránsito en el país, los datos obtenidos se pueden utilizar como una herramienta de análisis, como por ejemplo (Ana Benitez, 2018) expone que en

el año 2014 se registró un total de 15.572 multas, mientras que con la mejora de la monitorización para el año 2017 esté número incrementó a 260.063, aproximadamente un 1500 % en un periodo de tan solo 3 años.

1.2.PLANTEAMIENTO DEL PROBLEMA

El exceso de velocidad dentro de las vías del Ecuador, es considerado una de los principales causas de accidentes (ANT, 2018), como vuelcos, salidas de automóviles de las carreteras y derrapes, los cuales dejan trágicas secuelas, como daños materiales, lesiones o pérdida de la vida. En el año 2017 la Agencia Nacional de Tránsito, contabilizó 28.967 siniestros viales, los casos más alarmantes son causados por el transporte interprovincial, en los cuales un total de 1.011 personas fallecieron (ANT, 2018).

Tomando en cuenta que las autopistas y carreteras son vías en las cuales el rango de velocidad es de 90 km/h y 100 km/h respectivamente (ANT, 2018) y, en las cuales el riesgo de que se produzcan accidentes de tránsito es mayor, es necesario conocer puntos en los cuales exista un exceso de estos límites de velocidad, ya que estos ayudarán a la obtención de comportamientos que permitan la creación de un modelo de predicción, como aporte a la prevención de accidentes en el Ecuador.

1.3.JUSTIFICACIÓN

El exceso de velocidad en las diferentes vías del Ecuador, es un problema grave que desencadena en varios accidentes fatales, fruto de la falta de concientización por parte de los conductores (Protección de Tránsito, 2018), la carencia de controles en las vías por Agentes de Tránsito, o la limitada o mala ubicación de radares en las carreteras (INEC, 2018).

Es necesario establecer modelos o herramientas que, permitan un correcto análisis predictivo de lugares en los cuales los rangos de velocidad sean sobrepasados, ayudando a la correcta distribución de controles por las diferentes carreteras del país, beneficiando a una reducción considerable de los accidentes de tránsito, así evitando pérdidas materiales, lesiones e incluso muertes.

La presente investigación proporcionará un modelo de predicción que permitirá conocer los puntos de excesos de velocidad dentro de la autopista General Rumiñahui, mediante la determinación y selección de variables que intervengan o influyan en un accidente de tránsito, esta información será validada, procesada y tratada, con la finalidad de evitar datos incongruentes, asegurando la precisión del prototipo, que servirá como aporte a la prevención de accidentes.

1.4 OBJETIVOS

1.4.1 Objetivo General

Estructurar un modelo de predicción de zonas de exceso de velocidad, en la autopista general Rumiñahui de la Ciudad de Quito, utilizando técnicas de gestión de datos, para disminuir los accidentes de tránsito.

1.4.2 Objetivos Específicos

- Investigar sobre la ocurrencia de accidentes de tránsito dentro de las autopistas del Ecuador, desde el año 2014, mediante la recolección de información disponible en la ANT, así como también obtener características de trabajos relacionados de ciudades con alta circulación vehicular, que permitan conocer los principales factores que intervienen en los mismos.
- Obtener datos a través de una aplicación móvil para el sistema operativo Android, que permita almacenar información de diferentes vehículos en la autopista General Rumiñahui.
- Realizar un análisis de los datos obtenidos mediante técnicas de gestión, que permitan obtener un modelo de predicción que determine zonas de exceso de velocidad en la autopista General Rumiñahui.
- Validar el modelo predictivo para la determinación de las zonas de exceso de velocidad, mediante el testeo de la información obtenida en la investigación, para verificar su precisión.

1.5 ALCANCE

La presente investigación comprende un análisis de las principales causales de accidentes de tránsito desde el año 2014, información contenida en la ANT en estadísticas dispuestas de forma mensual. Estos archivos permitirán determinar factores que intervengan en la ocurrencia de accidentes viales.

Para la selección y almacenamiento de factores o variables que intervienen dentro de un accidente de tránsito, se utilizará una aplicación móvil para el sistema operativo Android, que guardará esta información en una Base de Datos en la nube.

Previo al filtrado de datos se realizará una validación, con esto se planea evitar información errónea dentro del tratamiento de la información. Al tratar la información se podrá obtener patrones de comportamiento que permitirán la creación del modelo predictivo que determinará los puntos en las cuales los conductores exceden la velocidad permitido, dentro de la autopista General Rumiñahui.

Con la finalidad de validar dicho modelo, se llevarán a cabo pruebas que permitan verificar la precisión del modelo, mediante la información previamente obtenida, asegurando la exactitud predictiva.

Para delinear de forma adecuada el alcance de la investigación planteada, se proponen varias preguntas de investigación asociadas a los objetivos específicos, tal como se muestra en la Tabla 1.

Tabla 1.
Preguntas de Investigación

Objetivo específico	Pregunta de investigación
i. **Investigar sobre la ocurrencia de accidentes de tránsito dentro de las autopistas del Ecuador, desde el año 2014, mediante la recolección de información disponible en la ANT, así como también obtener características de trabajos relacionados de ciudades con alta circulación vehicular, que permitan conocer los principales factores que intervienen en los mismos.**	a. ¿Qué factores influyen en la ocurrencia de accidentes automovilísticos? b. ¿Cuáles son las principales causales de accidentes de tránsito?
ii. **Obtener datos a través de una aplicación móvil para el sistema operativo Android, que permita almacenar información de diferentes vehículos en la autopista General Rumiñahui.**	a. ¿Qué factores son de vital importancia para la determinación del modelo de predicción? b. ¿Qué porcentaje de confianza nos brinda la aplicación al recabar la información en la autopista general Rumiñahui?
iii. **Realizar un análisis de los datos obtenidos mediante técnicas de gestión que permitan obtener un modelo de predicción que determine zonas de exceso de velocidad en la autopista General Rumiñahui.**	a. ¿Qué técnica de gestión de datos es la más adecuada y cuál es la técnica seleccionada? b. ¿Qué tan preciso es el modelo predictivo obtenido con la técnica seleccionada?
iv. **Validar el modelo predictivo para la determinación de las zonas de exceso de velocidad, mediante el testeo de la información obtenida en la investigación, para verificar su precisión.**	a. ¿Cuántas pruebas se deben llevar a cabo para validar el modelo? b. ¿Qué porcentaje de error existe al testear el modelo?

Para la obtención de datos dentro de la autopista general Rumiñahui se utilizará una aplicación para dispositivos móviles con sistema operativo Android, representada por un modelo de tres niveles, los cuales se describen en la Figura 1.

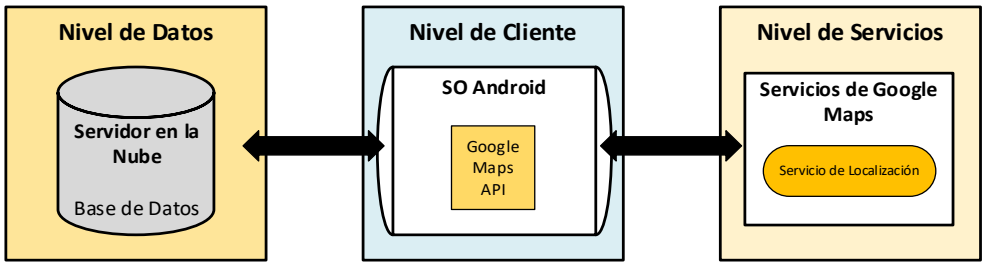


Figura 1. Arquitectura de la aplicación

En la figura podemos observar el nivel de Datos, un servidor en la nube almacenará una base de datos con los datos requeridos para el presente caso de estudio. En el nivel de Cliente, se visualizará la velocidad actual del vehículo y su ubicación, por medio del dispositivo móvil. Finalmente, el nivel de Servicios se utilizará todos aquellos servicios que brinda la aplicación de localización a través de las prestaciones obtenidas de Google Maps.

CAPÍTULO II

METODOLOGÍA Y MARCO TEÓRICO

2.1 METODOLOGÍA

2.1.1 Design Science

Design Science es una metodología aplicada comúnmente a investigaciones en el área de Ciencias de la Computación y, consiste en el diseño y la investigación de artefactos que pertenezcan al contexto. Esta metodología consta de un proceso, el cual abarca áreas relacionadas al proyecto, como la resolución de un problema importante para la comunidad, seguido de una construcción y evaluación rigurosa del modelo a ser implementado, de manera que asegure dar una solución real a la problemática central (Cataldo, 2015).

El objetivo de Design Science es desarrollar conocimiento, de modo que el profesional de una disciplina específica lo pueda utilizar para diseñar soluciones en su campo de estudio.

2.2 MARCO TEÓRICO

2.2.1 Siniestros de Tránsito

Un siniestro se considera un evento de tránsito en el que intervienen como mínimo un automotor, y el cual sea producido o provocado dentro de una vía, y en el que una persona resulte en un estado trágico. El termino siniestro se acuña a partir del año 2014 en el Ecuador ya que su uso es recomendado por organismos internacionales (OISEVI, 2018).

Cada año suceden alrededor de 26.000 siniestros de tránsito, esta información es almacenada de manera mensual y provincial dentro del Ecuador, figura 7 y 8.

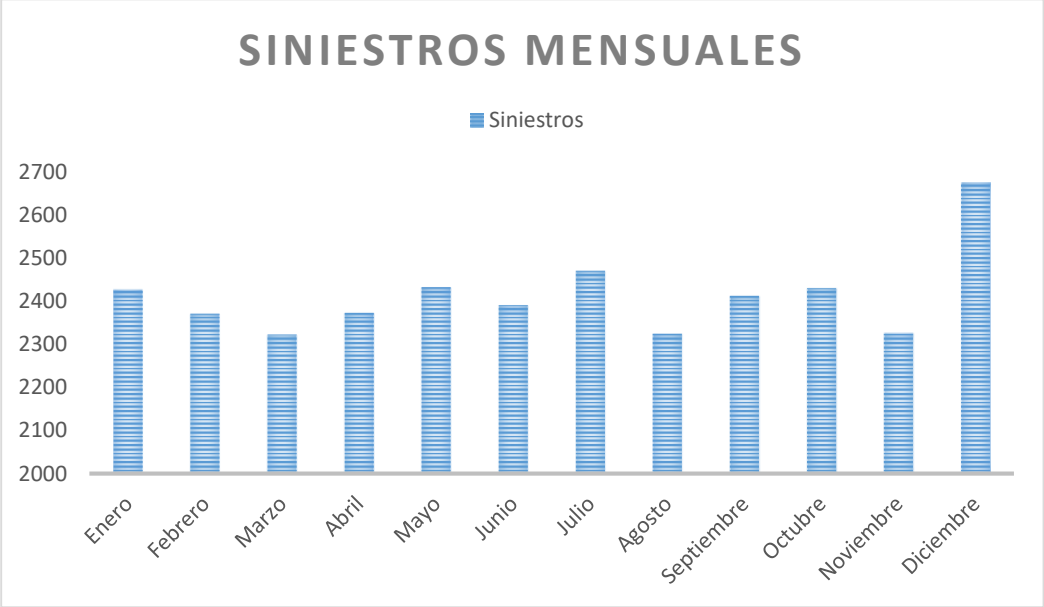


Figura 2. Siniestros Por Mes a Nivel Nacional – 2017
Fuente: (ANT, 2017)

En la figura podemos notar que el mes con mayor incidencia de siniestros a nivel provincial son los meses de Diciembre, Julio y Enero, debido a su alto índice de movilidad vehicular (ANT, n.d.).

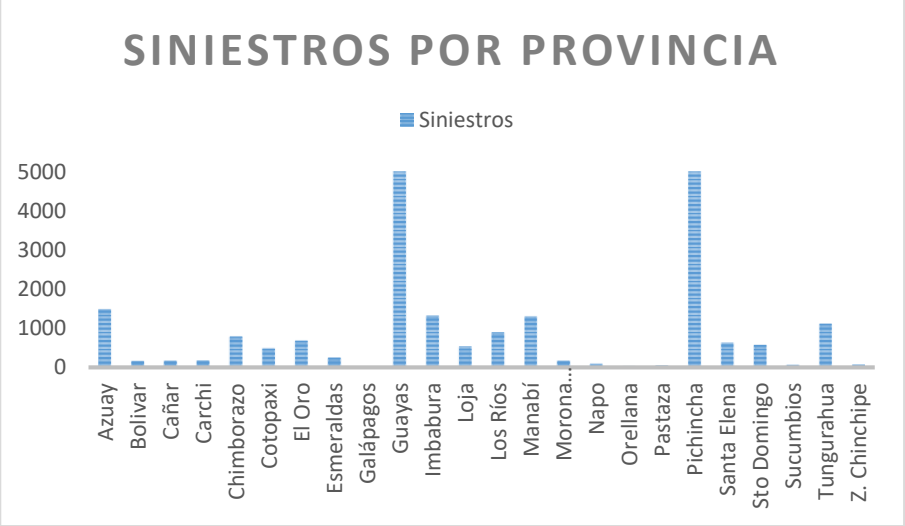


Figura 3. Siniestros Por Provincia a Nivel Nacional – 2017
Fuente: (ANT, 2017)

En esta figura se puede observar que las provincias con mayor porcentaje de siniestros son las provincias con mayor cantidad de población, como Guayas, Pichincha y Azuay.

Es así que estos accidentes resultan en un aproximado de 2000 personas fallecidas cada año, como se muestra en la figura 9 y 10.

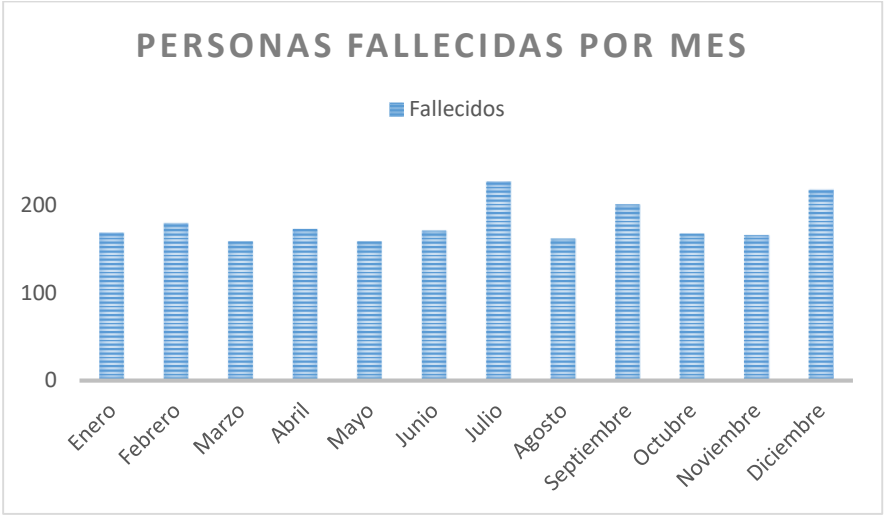


Figura 4. Fallecidos Por Mes a Nivel Nacional – 2017
Fuente: (ANT, 2017)

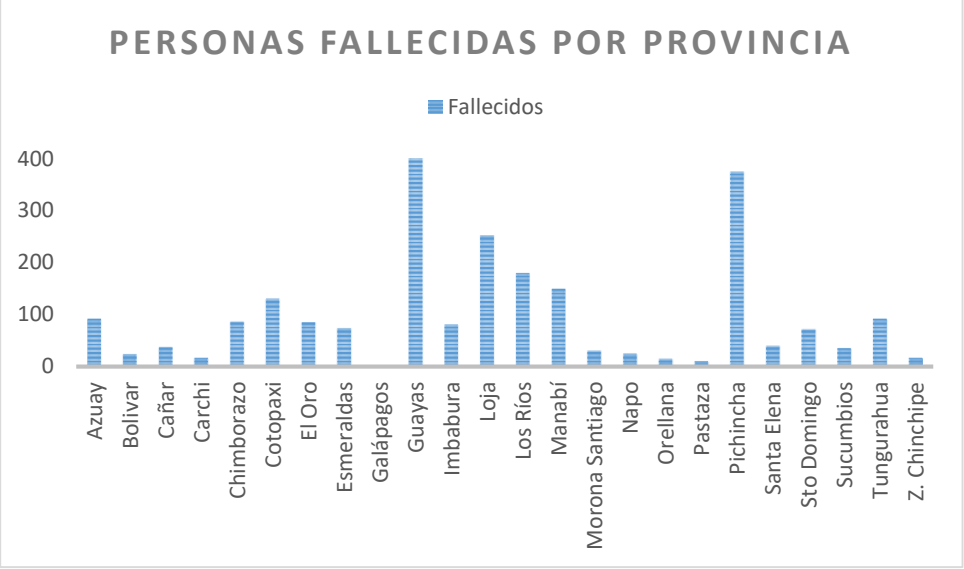


Figura 5. Fallecidos Por Provincia a Nivel Nacional – 2017
Fuente: (ANT, 2017)

En estas dos figuras podemos notar que la mayor cantidad de fallecidos la podemos encontrar en las provincias de Pichincha y Guayas, así como una mayor ocurrencia en los meses de Diciembre y Julio.

a) Factores que Intervienen en un Siniestro de Tránsito

Se considera varias causas según la ANT (2018).

i. Factores humanos (conductor):

- Edad avanzada.
- Cansancio del conductor.
- Ingesta de alcohol.

- Sobreestimar la experiencia o la capacidad del vehículo.
- Conducir un vehículo en mal estado.
- Desconocer las leyes de tránsito.
- Factores psicológicos.
- Sueño.
- Usar el celular al conducir.
- Distraerse al conducir.

ii. Factores humanos externos (peatones, otros conductores):

- Otros conductores en condiciones incorrectas.
- Conducción agresiva de otros conductores.
- Vehículos con velocidad por debajo del promedio.
- Imprevisión del peatón.

iii. Factores humanos externos (mecánicos y logísticos):

- Incorrecto mantenimiento del vehículo.
- Bajo conocimiento del mecánico.
- Uso de repuestos inapropiados.
- Mala visibilidad por luces, espejos, etc. defectuosos o en mal estado.

b) Principales Causas de los Siniestros de Tránsito

i. Humanas: causas en las que participa el ser humano según Mayou (1993)**.**

- **Exceso de Velocidad:** la velocidad de un vehículo recae en manos de un conductor, el exceso de velocidad trae consigo varios efectos como menor adherencia de las llantas y un periodo de reacción menor. Las normas del país definen límites de velocidad que se muestran en la figura 11.

VEHICULOS	TIPO DE VIA	LIMITE MÁXIMO	RANGO MODERADO	FUERA DE RANGO MODERADO
	Urbana	50 Km/h	50 Km/h a 60 Km/h	Más de 60 Km/h
	Perimetral	90 Km/h	90 Km/h a 120 Km/h	Más de 120 Km/h
	Rectas en Carretera	100 Km/h	100 Km/h a 135 Km/h	Más de 135 Km/h
	Curvas en Carretera	60 Km/h	60 Km/h a 75 Km/h	Más de 75 Km/h
	Urbana	40 Km/h	40 Km/h a 50 Km/h	Más de 50 Km/h
	Perimetral	70 Km/h	70 Km/h a 100 Km/h	Más de 100 Km/h
	Rectas en Carretera	90 Km/h	90 Km/h a 115 Km/h	Más de 115 Km/h
	Curvas en Carretera	50 Km/h	50 Km/h a 65 Km/h	Más de 65 Km/h
	Urbana	40 Km/h	40 Km/h a 50 Km/h	Más de 50 Km/h
	Perimetral	70 Km/h	70 Km/h a 95 Km/h	Más de 95 Km/h
	Rectas en Carretera	70 Km/h	70 Km/h a 100 Km/h	Más de 100 Km/h
	Curvas en Carretera	40 Km/h	40 Km/h a 60 Km/h	Más de 60 Km/h

Figura 6. Rango de Velocidades Respecto al Tipo de Vehículo
Fuente: (ANT, n.d.)

La velocidad de un vehículo es controlada por el velocímetro, este permite conocer la rapidez a la que circula el vehículo de manera precisa, este es un tacómetro calibrado en Km/h que se basa en la medición de la rotación del árbol de la caja de velocidades o el giro de las ruedas (EDIMEC, 2014).

- **Malas Maniobras:** los accidentes son usualmente causados por malas maniobras al rebasar o adelantar a otro vehículo, así como al ser imprudentes o negligentes al conducir un vehículo.
- **Embriaguez:** ingesta de bebidas alcohólicas que ocasiona pérdida temporal de las facultades mentales y físicas.

ii. **Mecánicos:** fallas de los principales elementos del vehículo: frenos, dirección, suspensión, etc. Estos accidentes pueden ser evitados con un correcto mantenimiento, aunque pueden existir fallas imprevisibles (Abdel-Aty & Radwan, 2000).

iii. **Viales**: la vía no es una causa directa de los accidentes, pero si puede aportar a la ocurrencia de una tragedia. La falta de señalización en la vía puede ser una causa. También a lo largo de la vía podemos encontrar varios peligros según (Choquehuanca-Vilca, Cárdenas-García, Collazos-Carhuay, & Mendoza-Valladolid, 2010).

- **Derrumbes:** producidos por laderas que se arrastran hacia la vía convirtiéndose en un obstáculo, impidiendo la circulación normal de los vehículos.
- **Hundimientos:** producidos por la circulación de vehículos que exceden el peso que puede soportar una vía.
- **Grietas:** ruptura de la vía de menor tamaño que un bache, si un vehículo cae dentro de una grieta puede provocar que el capot se alce violentamente obstruyendo la visibilidad del conductor.
- **Bache:** ruptura de la vía en forma de hueco, este puede causar desequilibrio.

iv. **Ambientales:** fenómenos naturales que afectan la visibilidad del conductor (Híjar-Medina, Carrillo-Ordaz, Flores-Aldana, Anaya, & López-López, 1999).

- **Lluvia:** afecta la captación de objetos que están frente, a los lados y detrás del vehículo, esta puede ocasionar un hidroplaneo por la presencia de una cantidad considerable de agua en la vía.
- **Niebla:** afecta a la percepción de las señaléticas de la vía, en este caso es indispensable la reducción de la velocidad del vehículo.
- **Neblina:** es una nube de menor densidad que la niebla.
- **Tormenta:** afecta en la estabilidad del vehículo, ya que es una combinación de lluvia y viento.
- **Granizo:** agua congelada en forma de granos, sólidos y gruesos, afectan generalmente la adherencia de las llantas.

En la figura 12, se muestran las circunstancias que alteran la visibilidad del conductor.

Fenómeno	Característica	Visibilidad
Calima	Partículas secas trasladadas por el viento	2 Km
Neblina	Partículas húmedas en suspensión	
	○ Débil	> 4 Km
	○ Medio	2 a 4 Km
	○ Fuerte	1 a 2 Km
Niebla	Partículas húmedas agrupadas en suspensión	
	○ Débil	200 a 500 m
	○ Media	200 a 500 m
	○ Fuerte	50 a 200 m
	○ Muy fuerte	< 50 m
Smog	Niebla más humo	Disminución variable de la visibilidad según la intensidad
Lluvia	○ ligera ○ fuerte	Dificulta la visual en forma variable según la intensidad
Nieve	Agua congelada que desciende como copos	Dificulta visibilidad Impide la visual
	○ ligera ○ ventisca	
Granizo	Agua congelada que desciende como grumos, en forma violenta	Dificulta la visibilidad, varía según intensidad puede impedir la visual
Viento	Depende de dirección del viento	

Figura 7. Factores Ambientales que Afectan la Visibilidad del Conductor
Fuente: (EDIMEC, 2014)

2.2.2 Conocimiento en Bases de Datos (KDD)

El KDD es un término usado para referirse al procesamiento de la información para descubrir conocimiento (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

En un proceso de KDD según Beltrán (2014) se pueden definir al menos 6 fases:

1. **Recolección de Datos:** esta fase del KDD es la más importante ya que se define de que fuentes se extraerá la información. Los siguientes pasos serán mucho más sencillos si la fuente de la información es accesible. (Fayyad, Piatetsky-Shapiro, et al., 1996).

2. **Selección, Filtrado y Transformación de Datos:** se procede a eliminar la mayor cantidad de datos incorrectos o inconsistentes e irrelevantes, para esto se usan:
    - Histogramas
    - Muestreo
    - Redefinición de Atributos

3. **Minería de Datos:** una vez que se han recolectados una cantidad considerable de datos, se decide el tipo de patrón que se quiere encontrar y el tipo de conocimiento a extraer.

4. **Evaluación y Validación:** la fase 3 da como resultado uno o varios modelos, para escoger el modelo es necesario usar criterios para su evaluación (Fayyad, Haussler, & Stolorz, 1996).

5. **Interpretación y Difusión:** en ocasiones el modelo suele ser sencillo, pero en otras es necesario un proceso de interpretación.

6. **Actualización y Monitorización:** este es el proceso de mantenimiento y monitoreo del modelo.

Minería de Datos

Con el avance tecnológico a pasos agigantados, las empresas de la actualidad poseen software y hardware cada vez más sofisticado, que permite y facilita el almacenamiento de enormes cantidades de información (Hand, 2007).

La Minería de Datos o Data Mining se encarga de recabar patrones de conocimiento o interés dentro de millares de datos. Esta descubre tendencias, comportamientos y patrones con el fin de ayudar en el proceso de toma de decisiones (Hand, 2007).

El proceso de Data Mining comienza con la definición de los datos, para esto debemos determinar qué datos se requieren, dónde y cómo encontrarlos. Al obtener los datos estos deben ser almacenados en una base de datos. Antes de realizar el análisis de los datos, debemos tener claro lo que se desea obtener (predicciones, modelos, comportamientos, acciones, etc) (Han, Kamber, & Pei, 2011). En la figura 2 podemos ver de manera general el proceso de minería de datos.

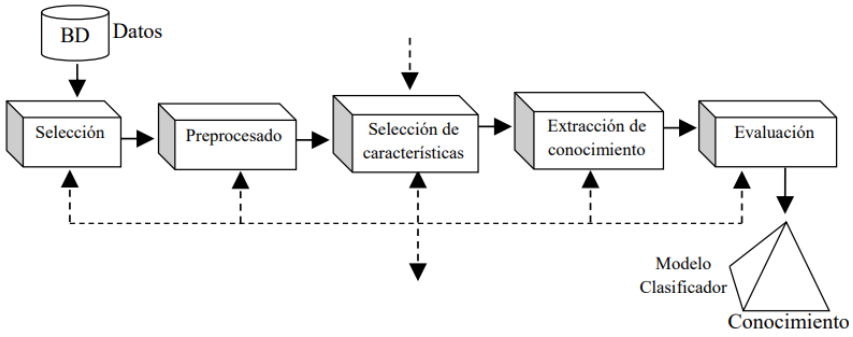


Figura 8. Pasos de la Minería de Datos
Fuente: (Beltrán Martínez, 2014)

En la figura podemos observar las fases constitutivas del proceso de minería de datos, desde que se selecciona la Base de Datos a ser tratada, filtrado, selección de características, extracción de

conocimiento y su evaluación. Estos pasos son descritos por Han et al (2011) a mayor detalle, a continuación.

1. **Procesado de Datos:** generalmente los datos obtenidos siempre cuentan con basura, que impide su tratamiento, en este estado se filtran los datos, de manera que se extraen valores indeseados.

2. **Selección de Características:** aquí se deben seleccionar categorías para reducir el tamaño de los datos, aquí se eligen la variable o variarles más predominantes del problema.

3. **Extracción de Conocimiento:** usando una o varias técnicas de minería de datos se obtendrá un modelo.

4. **Evaluación y Validación:** el modelo obtenido debe ser validado, verificando que los resultados son válidos y satisfactorios.

Clasificación de las Técnicas de Aprendizaje

Según Riquelme, Ruiz, & Gilbert (2006), tenemos:

1. **Técnicas No Supervisadas y Descriptivas**

   Métodos Descriptivos

   - Correlación y Asociaciones
   - Reglas de Asociación

- Algoritmos de Búsqueda de Asociaciones.
- Métodos Representativos.

2. Técnicas Supervisadas o Predictivas

Métodos Predictivos

- Interpolación y Predicción Secuencial
- Aprendizaje supervisado
  - **K-NN (Nearest Neighbour)**

Conocido como el vecino más cercano, se basan en la búsqueda dentro de un grupo de prototipos de los k prototipos más cercanos al patrón a clasificar. Las predicciones se basan en ejemplos que tengan el mayor grado de similitud al de la predicción (García & Gómez, 2006).

Este algoritmo está formado por dos fases la de entrenamiento y la de clasificación. Su funcionamiento radica en comparar los vectores de entrada con un conjunto de vectores qué presentan distinto comportamiento, de esta manera se apunta a la clase más cercana del conjunto de vectores(Saini, 2013).

. Las principales ventajas de este método es que no tiene costo para su aprendizaje y puede predecir valores continuos. Las desventajas radican en determinar k ya que no hay un mecanismo que permita definirlo (García & Gómez, 2006).

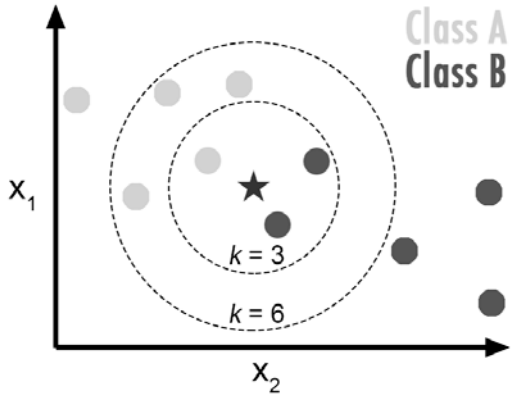


Figura 9. Algoritmo K-NN
Fuente: (Venkat, 2017)

En la figura podemos observar el comportamiento del algoritmo del K-NN, en el cual tenemos 2 clases A y B, representadas por puntos de color gris y negro respectivamente. Estas clases se hallan dispersas de manera irregular por todo el espacio, el valor de K representa la medida que abarca un grupo de vectores, este valor de K puede disminuir o aumentar según la dispersión de los datos y la precisión de la predicción. En esta figura podemos notar que a mayor K existe una mayor cantidad de vectores contenidos para la Clase A, mientras que para la Clase B el valor no varía (Saini, 2013).

- **K-means clustering**

El algoritmo de K-means usado por MacQueen en 1967, es uno de los algoritmos mayormente usados, este se encarga de encontrar K conjuntos de datos con una característica en particular, su mayor limitante es que los datos con los cuales trabaja deben ser solo valores numéricos (Zhexue, 2011).

Este algoritmo divide los objetos en un número pre definido, permite tener una comprensión cualitativa y cuantitativa de los datos, es muy conocido por ser eficiente al manipular grandes cantidades de datos (Cáceres, 2016).

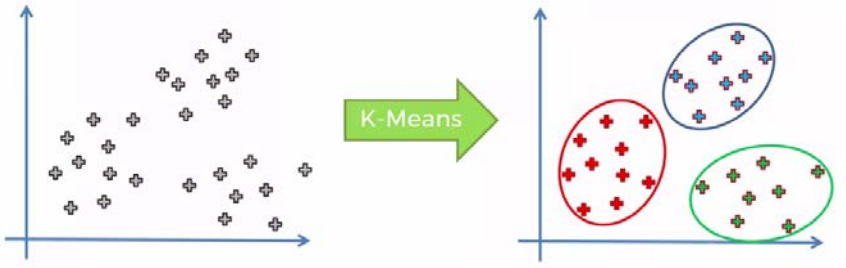


Figura 10. Algoritmo K-means
Fuente: (Prasad Patil, 2018)

En la figura se puede notar el comportamiento del algoritmo, con lo cual se comienza con un conjunto de varios datos dispersos, para lo cual procede a formar grupos o particiones con datos en los cuales el comportamiento sea semejante, consiguiendo conjuntos ordenados de datos con características similares, esto se logra mediante varias observaciones (Prasad Patil, 2018).

- **Perceptron Learning**

Es un tipo de neurona, el cual calcula mediante entradas que son afectadas por una función para producir una salida deseada (Acuna, 2010). Para esto en la tabla 2 podemos ver los tipos de funciones.

Tabla 2.
Funciones de Activación

Nombre	Función
Lineal	A(x)=x
Logística	$A(x)=(1+e^{-x})^{-1}$
Gaussiana	$A(x)=\exp - x^2/2$
Threshold	A(x)=0 si x<0, A(x)=1 en otro caso

La tabla de Funciones de Activación consta de 2 columnas, en la primera podemos observar el nombre de la función, mientras que en la segunda el tipo de función que describe dicha función.

El perceptrón consta de varias entradas cada una de ellas con sus respectivos pesos, una función de activación y su salida (Rosenblatt, 1958), como se puede ver en la figura 5.

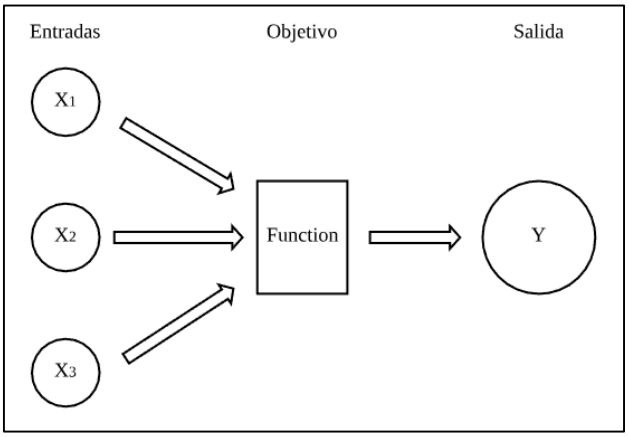


Figura 11. Perceptron Simple

La figura muestra la estructura básica de un perceptrón, la cual se define por varias entradas (X), las cuales se dirigen hacia su objetivo (función), para así obtener un comportamiento deseado o salida (Y).

- **Redes Neuronales**

Forman parte la Inteligencia Artificial (IA), estas son un conjunto de unidades básicas con un número definido de conexiones, en donde se pueden encontrar 3 tipos de unidades (González, 2014).

a. **Unidades de entrada:** aquellas que reciben señales del entorno.
b. **Unidades de salida:** envían las señales (dan respuesta).
c. **Unidades ocultas:** estas están conectadas siempre con otra unidad (dentro de la red).

El tipo de red neuronal más usado es el **Perceptron Multicapa**, esto se debe a que es capaz de aprender de cualquier tipo de función (continua), dependiendo principalmente de la complejidad de la red. El tipo de aprendizaje es supervisado ya que el usuario define la salida deseada (Lippman & Lippman, 1987). La figura 6 muestra el esquema de un perceptron multicapa.

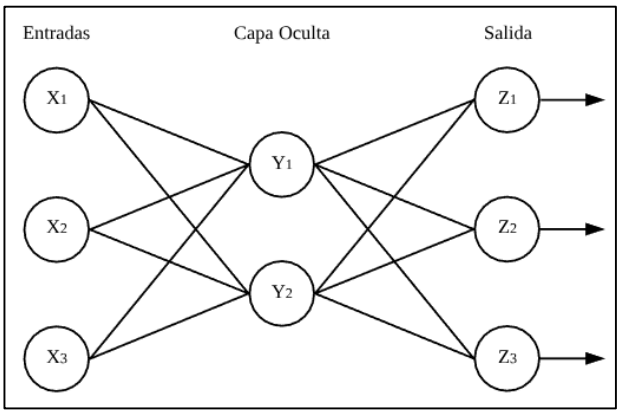


Figura 12. Perceptron Multicapa Simple

En la figura podemos observar la estructura de un perceptron multicapa o un conjunto de perceptrones, que se constituye por varias entradas (X), dirigidas a su capa oculta u objetiva (Y) para producir varias salidas (Z).

2.2.3 Herramientas para Minería de Datos

A continuación se enumeran alguna de las herramientas más usadas para el proceso de Data Mining, en los diferentes ámbitos.

1. RapidMiner

Es una herramienta conocida anteriormente como YALE, fue desarrollada en conjunto por varios colaboradores Klinkengerg, Mierswa y Fischer. Esta permite un modelado estadístico, pre procesado de datos, análisis de negocio, optimizaciones y predicciones. Es mayormente usada por grandes compañías como Ford, Honda, IBM, entre otras (Yadav, Malik, & Chandel, 2015).

2. R Studio:

Este software es independiente ya que tiene incluido editores para la escritura de documentos con LaTeX, además tiene una integración con lenguajes de marcado y permite el uso de herramientas como C++, CSS y otros lenguajes de programación. R Studio permite realizar todas las actividades que incluye R, pero simplifica mucha de ellas, brindando agilidad y rapidez dentro de su interfaz y su editor de texto (Gandrud, 2015).

3. Clementine:

Es un sistema muy intuitivo y con un alto grado de usabilidad, se basa principalmente en la unión de la minería de datos con otras actividades y sistemas de negocio que permitan obtener predicciones inteligentes con tiempos eficientes en sus distintas operaciones. En sus versiones más recientes esta incluye reglas de scoring y permiten el modelamiento de árboles de decisión (Rodriguez & Díaz, 2009).

4. Weka:

Conocido como proyecto WEKA, fue fundada en 1993, los primeros años de la misma se focalizaron en el desarrollo de su interfaz, generalmente desarrollada en el lenguaje C. Weka es un software que se actualiza continuamente y en cada una de sus mejoras incluye nuevos algoritmos de aprendizaje, filtros para pre procesamiento, así como mejoras en su usabilidad. Este software es generalmente usado en varios proyectos relacionados a descubrir conocimiento en biología, procesamiento de lenguaje, minería de datos paralela y distribuida, entre muchos más (Hall Mark, Eibe Frank, 1997).

CAPÍTULO III

ESTADO DEL ARTE

3.1 PLANTEAMIENTO DE LA REVISIÓN DE LITERATURA

En esta fase se abordó la problemática de investigación para delimitar un contexto para la búsqueda de estudios científicos; posteriormente se procedió a definir un objetivo de búsqueda con lo cual se plantearon diferentes preguntas de investigación que permitan alinear la búsqueda con el problema de investigación y finalmente, se definieron los criterios de inclusión y exclusión.

3.2 CONFORMACIÓN DEL GRUPO DE CONTROL (GC)

Fue necesaria la participación de 2 investigadores. Los cuales propusieron estudios que podrían ser parte del grupo de control. Posteriormente se realizó un focus group donde se establecieron los artículos científicos seleccionados para el grupo de control, el cual se indica en la Tabla 3.

Tabla 3.
Grupo de Control

Título	Cita	Palabras clave
Accidentes automovilísticos fatales en la Zona Metropolitana de la Ciudad de México: una perspectiva en el espacio y en el tiempo	(RAMOS, ARAM; SILVA, ELID;AGUIRRE, 2015)	fatal, car, auto, accidents, mortality.
Study on accident prediction models in urban railway casualty accidents using logistic regression analysis model	(Jin & Lee, 2017)	accident investigation, accident prediction, falling accident, logistic regression, railway casualty accident.
Traffic speed prediction and congestion source exploration: A deep learning method	(Wang, Gu, Wu, Liu, & Xiong, 2016)	convolutional neural network, time series prediction, data mining, traffic information, speed.

Tras un análisis de los estudios del GC, se seleccionaron las palabras más relevantes respecto al objetivo de la búsqueda, en este caso fueron: fatal, car, auto, accidents, crash, prediction, predict, data, data managment, data mining.

3.3 CONSTRUCCIÓN Y AFINACIÓN DE LA CADENA DE BÚSQUEDA

Con las palabras clave que fueron obtenidas de los artículos científicos del grupo de control se conformó la cadena de búsqueda: ((“ACCIDENT” OR “CRASH”) AND (“CAR” OR “AUTO”) AND (“DATA MANAGMENT” OR “DATA MINING”) AND (“PREDICTION” OR “PREDICT”)), misma que se utilizó en la base digital SCOPUS.

Sin embargo, la cadena antes mencionada no obtuvo artículos científicos. Después de realizar varias iteraciones, combinando y cambiando palabras, se logró obtener la cadena definitiva,

la cual contenía palabras de interés y de importancia para la temática a tratar; como resultado se planteó la cadena: “((“ACCIDENT” OR “CRASH”) AND (“DATA MINING”) AND (“PREDICTION”))”.

3.4 SELECCIÓN DE ESTUDIOS

Al aplicar la cadena de búsqueda en la base digital SCOPUS, se obtuvo alrededor de 161 artículos relacionados con el tema, número de artículos que se consideró manejable y a los que se denominó estudios candidatos.

Los 161 estudios candidatos pasaron por un proceso de filtrado, el cual es descrito a continuación:

1. **Vigencia:** Se filtraron únicamente estudios realizados a partir del año 2016 hasta la presente. Se eligió este año debido a que la tecnología avanza rápidamente, por lo que es necesario tener estudios con una relativa actualidad.

En base al filtro antes mencionado, se obtuvieron los estudios primarios, que en este caso fueron 6, los cuales constituyen la base para encontrar las características del estado del arte en torno a la problemática identificada, los cuales se muestran en la Tabla 4.

Tabla 4.
Estudios Seleccionados

Código	Título	Cita
EP1	An Accident Prediction in Military Barracks Using Data Mining	(Shin, Yoo, & Nasridinov, 2016)
EP2	Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques	(Kaur & Kaur, 2017)
EP3	Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates	(Taamneh, Alkheder, & Taamneh, 2017)
EP4	Real-time crash prediction on freeways using data mining and emerging techniques	(You, Wang, & Guo, 2017)
EP5	The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents	(Shiau, Tsai, Hung, & Kuo, 2016)
EP6	A review on road accident data analysis using data mining techniques	(Science, Sakhare, & Science, 2017)

3.5 ELABORAR EL ESTADO DEL ARTE

EP1 (Yoo, Shin, Nasridinov, 2016): An Accident Prediction in Military Barracks Using Data Mining

En el trabajo de investigación titulado “An Accident Prediction in Military Barracks Using Data Mining” sus autores exponen ideas acerca del uso de técnicas de minería de datos para prevenir que hayan asesinatos y suicidios dentro de los cuarteles militares en Corea Del Sur, para esto se recopilaron datos de varios soldados como redes sociales, datos personales y médicos, luego se aplican técnicas de data mining para mejorar y manejar dichos datos y por ultimo realizan una evaluación de rendimiento de cada técnica para verificar la aplicabilidad de cada una a dicho caso.

En este artículo los autores recalcan la importancia de una limpieza e integración de datos antes de realizar un análisis de data mining, para esto usaron ranking, clustering, calificación y minería de texto.

EP2 (Gagandeep, Harpreet, 2017): Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques

"Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques", es un trabajo de investigación donde se analizan datos de accidentes de tránsito de las carreteras estatales y las carreteras ordinarias, con técnicas de minerías de datos. El trabajo se centra en un análisis paramétrico usando diversas herramientas, técnicas y metodologías, usadas en otros artículos científicos. Es así que recopilan un total de 11 artículos, los cuales usan técnicas de data mining, en las cuales los autores generan sus inputs, simulaciones y posibles outputs (resultados) al aplicar dichas técnicas.

El problema se planea solventar mediante una investigación paramétrica de los atributos que se deben considerar dentro de un accidente, como tipo, lugar, intersección, etc. La limitación de este artículo es que aún se necesita una explicación adicional para las tendencias, patrones, instancias y definición de nuevas instancias, las cuales se pueden obtener a través de varios algoritmos, los cuales serán aplicados en un futuro.

Para la simulación usan la herramienta RStudio, el cual es un entorno de desarrollo integrado para la herramienta R.

EP3 (Madhar, Sharaf, Salah, 2016): Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates

En el paper “Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates”, explica el uso de información histórica sobre accidentes de tránsito, para la obtención de factores más influyentes en estos.

Para el pre procesado de los datos se realizaron cambios como: eliminación de atributos invariantes, eliminación de atributos descriptivos, eliminación de atributos irrelevantes, eliminación de los registros con valores desconocidos, categorización de atributos con exceso de valores y, eliminación de información redundante.

Para los modelos de clasificación se usó el software WEKA en el cual se emplearon 4 algoritmos: árbol de decisión, inducción de reglas, bayes y perceptron multicapa. De esta manera se procedió a la evaluación de la efectividad de cada método para predecir la gravedad del accidente: 1) El conjunto de datos se usó como un conjunto de entrenamiento para el algoritmo y la precisión de la clasificación se determinó en función de que tan bien predijo la clase de accidente. 2) Se evaluó la precisión mediante validación cruzada con 10 pliegues. 3) Se re muestreo el conjunto de datos para sesgar la distribución de la gravedad del accidente hacia una distribución uniforme y se reutilizo la validación cruzada con 10 pliegues para evaluar su desempeño.

Este documento ofrece una doble contribución: establecer modelos para predecir la severidad de la lesión de cualquier accidente con una precisión razonable. Establece un conjunto de reglas que los analistas de seguridad pueden utilizar para identificar los principales factores que contribuyen a la gravedad de las lesiones

EP4 (You, Wang, Guo, 2017): Real-time crash prediction on freeways using data mining and emerging techniques

"Real-time crash prediction on freeways using data mining and emerging techniques" usa como área de prueba la autpista G60 en Shangai,China. Esta posee una longitud de 48,7 Km con 3 a 5 carriles por cada sentido. Para la recopilación de información se considera los datos del tráfico, clima y choques. La técnica de Random Forest se usó para seleccionar los factores contribuyentes antes del primer proceso de modelado, en dicho proceso la muestra se dividió en dos partes al azar: los datos de entrenamiento y los de prueba. Para revelar los efectos variables se empleó el método del valor de importancia media usado en redes neuronales para evaluar los efectos relativos de las variables.

EP5 (Shiau, Tsai, Hung, Kuo, 2016): The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents

En el paper "The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents" la investigación se divide en 3 partes, primero se realiza la recolección de información, en las cuales se consideran 17 variables de entrada, incluyendo el clima, rayos solares, sitios de accidentes y patrones de caminos, en este caso particular también se consideran las variables de salida, para el pre-procesado se usa la técnica RFE que permite ordenar los datos de acuerdo a su importancia, para posteriormente ser sustituidas en BPNN que usa redes neuronales para tratar los datos obtenidos, de manera que se pueda eliminar y limitar los falsos

positivos y negativos, cada experimento es repetido 5 veces, de esta manera se verifican los cambios al procesar los datos.

Para finalizar la investigación se lleva a cabo un análisis de Regresión Logística importando los datos en el mismo para la predicción de clasificación de acuerdo al conjunto de datos de prueba, seguido de una combinación con el análisis FRPCA. Los resultados de este experimento muestran una mejora considerable al utilizar la regresión logística combinada con la FRPCA.

EP6 (Sakhare, Kasbe, 2017): A review on road accident data analysis using data mining techniques

"A review on road accident data analysis using data mining techniques" se centra en el análisis de información obtenida del sitio us.gov la cual contiene 15 atributos y 2634 datos de entradas relacionados a accidentes de tránsito del año 2015, con esto se planea realizar un análisis de los principales factores y razones para accidentes, de esta manera se podrá obtener una predicción anual, mensual y diaria para mejorar la gestión de las vías.

Lo primero a realizar es el clustering de manera correcta para lo cual se usara K-mean, seguido de técnicas de programación neurolingüística, después se planea organizar el mapa de la información, y para finalizar se llevara a cabo la implementación de la técnica propuesta y una prueba para comparar los resultados con el cluster de K-mean.

3.5.1 Características del estado del arte

En conclusión, de acuerdo al estudio del estado del arte realizado, se obtuvo que las técnicas más usadas para una gestión de datos son K-mean, Bayes y redes neuronales; estas son aplicadas para obtener un modelo de predicción, en estos artículos se llevan a cabo un análisis para la obtención de variables requeridas y necesarias en cada problemática, estas variables permiten determinar los factores que influyen en la ocurrencia de accidentes, como: el clima, el tipo de vehículo, el tiempo, entre otras.

Estas investigaciones sirven de apoyo para definir técnicas predictivas que ayuden a la generación del modelo requerido para el problema de zonas de exceso de velocidad, así como la determinación de la que provea un alto índice de confiabilidad y la cual pueda ser validada.

CAPÍTULO IV

DISEÑO DE LA INVESTIGACIÓN Y RECOLECCIÓN DE LA INFORMACIÓN

4.1 DESIGN SCIENCE

Esta metodología se centra principalmente en el diseño, como tal este es un proceso y un producto, ya que el resultado de un conjunto de pasos (proceso), da como resultado un artefacto (producto) (Wieringa, 2010).

A continuación se definen sus pasos, aplicados al objeto de estudio de la investigación.

1. **Relevancia del Problema:** La actual problemática es de sumo interés para la comunidad, debido a que el exceso de velocidad es causado por una sobreestimación de experiencia por parte de los conductores dentro de las autopistas, trayendo consigo, accidentes que provocan daños materiales, lesiones e incluso la muerte.
2. **Diseño:** La actual investigación requiere de un software para adquirir toda la información primordial para su posterior análisis, obteniendo con esto un modelo que permita predecir puntos en los cuales el exceso de velocidad sea recurrente.
3. **Rigor:** En este punto se considera que para la creación del modelo se requiere de una cantidad elevada de datos, que posteriormente serán validados, a la par que permitirán obtener un modelo de predicción correctamente validado y testeado.
4. **Proceso:** Para la recopilación de datos, se usará una aplicación para dispositivos Android, esto por su alto índice de portabilidad, con lo cual se obtendrán los datos

requeridos para la creación del modelo de predicción dentro de la herramienta de RapidMiner

5. **Evaluación:** Dentro del proceso de evaluación se evaluará en primera instancia los datos obtenidos a través de la aplicación con el fin de evitar incongruencias, seguido a esto el modelo será validado y testeado con pruebas reales y simuladas, que aseguran su índice de precisión.
6. **Contribuciones:** Como contribución se focaliza en Fundamentos, ya que se desarrollara un modelo predictivo.
7. **Comunicación:** La principal audiencia serán entes encargados de la administración vial, ya que el problema tiene un alto grado de importancia y la forma en la cual se define la solución es poco convencional.

4.2 PROCESO DE INVESTIGACIÓN

4.2.1 Obtención de la Información

En esta etapa del proyecto se obtuvo información proveniente de la A.N.T., la cual funciona como ente centralizador de datos de las entidades encargadas del control vial, en la cual incluso se ve involucrado el Servicio Integrado de Seguridad (ECU 911).

Para adquirir la información contenida por la A.N.T., fue necesario aplicar una solicitud dentro de la página web https://www.ant.gob.ec/ , se accede al apartado Formulario de Información Estadística, aquí se procedió a llenar el formulario en el cual se debió especificar los campos contenidos en la Figura 13.

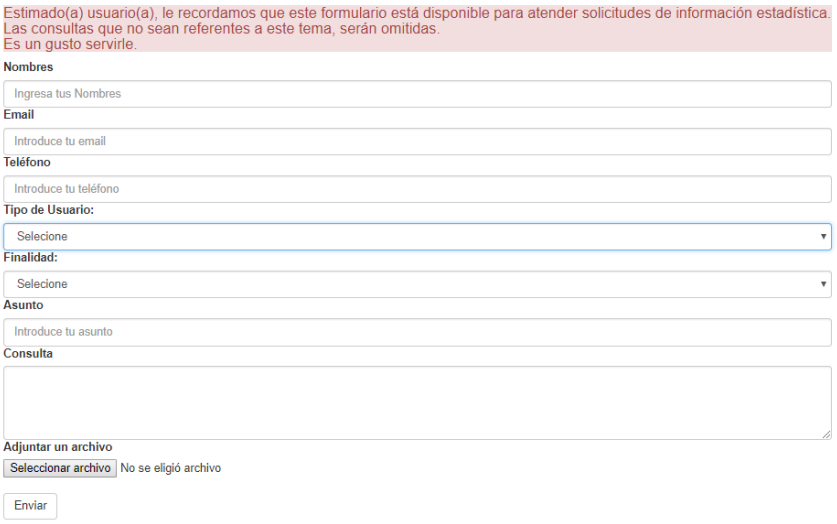


Figura 13. Formulario de Solicitud de Información Estadística
Fuente: (ANT, 2018)

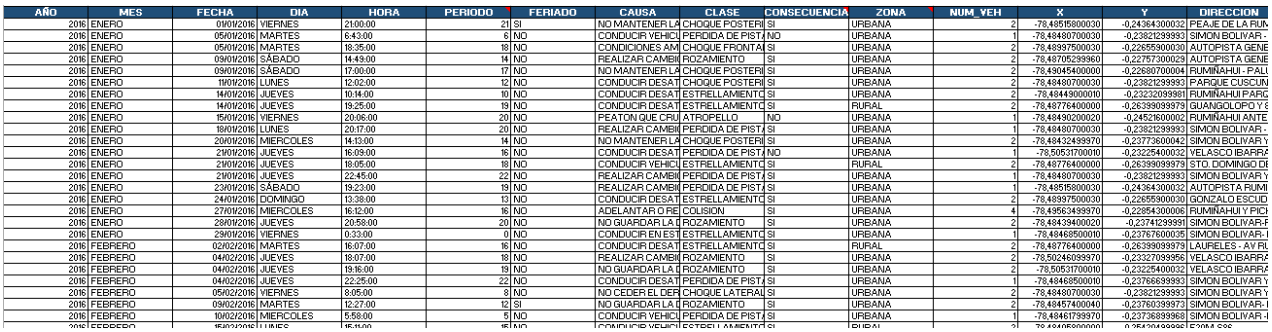
En esta solicitud se especificó que la información requerida debía pertenecer específicamente de la autopista General Rumiñahui. Al enviar la solicitud el tiempo de entrega fue de una semana, esta se entregó a través de correo electrónico en un archivo de Excel, en el cual se detallaban los años, días, horas, coordenadas, entre otras variables, de todos los siniestros ocurridos en esta vía

4.2.2. Interpretación de la Información

Aquí se procedió a obtener el diccionario de datos de la información obtenida, con el fin de entender el significado de cada variable dentro de la tabla.

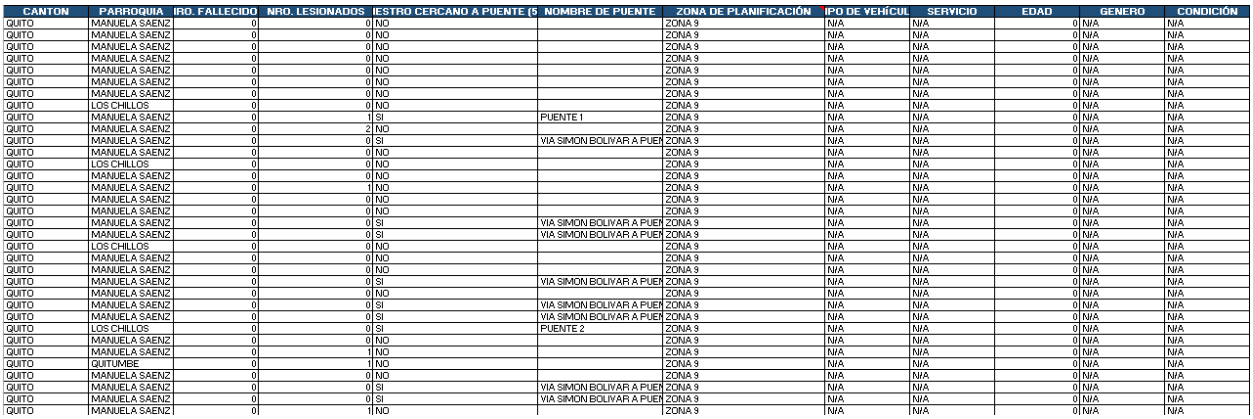
Esta tabla recolecta información en un periodo entre Enero 2016 y Julio 2018, esto se debe a que antes de estos años no se consideraban detalles específicos del siniestro (ANT, n.d.). La tabla

contiene 27 atributos, como se muestra en la Figura 14 y 15, los cuales permitieron llevar a cabo su correcto análisis.



AÑO	MES	FECHA	DIA	HORA	PERIODO	FERIADO	CAUSA	CLASE	CONSECUENCIA	ZONA	NUM_VEH	X	Y	DIRECCION
2016	ENERO	01/01/2016	VIERNES	21:00:00	21	SI	NO MANTENER LA	CHOQUE POSTERI	SI	URBANA	2	-78,48515000030	-0,24364300032	PEAJE DE LA RUN
2016	ENERO	05/01/2016	MARTES	6:43:00	6	NO	CONDUCIR VEHICU	PERDIDA DE PISTA	NO	URBANA	1	-78,48480700030	-0,23821299993	SIMON BOLIVAR -
2016	ENERO	05/01/2016	MARTES	18:35:00	18	NO	CONDICIONES AM	CHOQUE FRONTAL	SI	URBANA	2	-78,48397500030	-0,22655900030	AUTOPISTA GENE
2016	ENERO	09/01/2016	SÁBADO	14:49:00	14	NO	REALIZAR CAMBI	ROZAMIENTO	SI	URBANA	2	-78,48705299960	-0,22757300029	AUTOPISTA GENE
2016	ENERO	09/01/2016	SÁBADO	17:00:00	17	NO	NO MANTENER LA	CHOQUE POSTERI	SI	URBANA	2	-78,49045400000	-0,22680700004	RUMIÑAHUI - PALO
2016	ENERO	11/01/2016	LUNES	12:02:00	12	NO	CONDUCIR DESAT	CHOQUE POSTERI	SI	URBANA	2	-78,48480700030	-0,23821299993	PARQUE CUSCUN
2016	ENERO	14/01/2016	JUEVES	10:14:00	10	NO	CONDUCIR DESAT	ESTRELLAMIENTO	SI	URBANA	2	-78,48449000010	-0,23232099981	RUMIÑAHUI PARQ
2016	ENERO	14/01/2016	JUEVES	19:25:00	19	NO	CONDUCIR DESAT	ESTRELLAMIENTO	SI	RURAL	2	-78,48776400000	-0,26399099979	GUANGOLOFO Y S
2016	ENERO	15/01/2016	VIERNES	20:06:00	20	NO	PEATON QUE CRU	ATROPELLO	NO	URBANA	1	-78,48498200020	-0,24521600002	RUMIÑAHUI ANTE
2016	ENERO	18/01/2016	LUNES	20:17:00	20	NO	REALIZAR CAMBI	PERDIDA DE PISTA	SI	URBANA	1	-78,48480700030	-0,23821299993	SIMON BOLIVAR -
2016	ENERO	20/01/2016	MIERCOLES	14:13:00	14	NO	NO MANTENER LA	CHOQUE POSTERI	SI	URBANA	2	-78,48432499970	-0,23773600042	SIMON BOLIVAR Y
2016	ENERO	21/01/2016	JUEVES	16:09:00	16	NO	CONDUCIR DESAT	PERDIDA DE PISTA	NO	URBANA	1	-78,50531700010	-0,23225400032	VELASCO IBARRA
2016	ENERO	21/01/2016	JUEVES	18:05:00	18	NO	CONDUCIR VEHICU	ESTRELLAMIENTO	SI	RURAL	2	-78,48776400000	-0,26399099979	STO. DOMINGO DE
2016	ENERO	21/01/2016	JUEVES	22:45:00	22	NO	REALIZAR CAMBI	PERDIDA DE PISTA	SI	URBANA	1	-78,48480700030	-0,23821299993	SIMON BOLIVAR Y
2016	ENERO	23/01/2016	SÁBADO	19:23:00	19	NO	REALIZAR CAMBI	PERDIDA DE PISTA	SI	URBANA	1	-78,48515000030	-0,24364300032	AUTOPISTA RUMI
2016	ENERO	24/01/2016	DOMINGO	13:38:00	13	NO	CONDUCIR DESAT	ESTRELLAMIENTO	SI	URBANA	2	-78,48997500030	-0,22655900030	GONZALO ESCUD
2016	ENERO	27/01/2016	MIERCOLES	16:12:00	16	NO	ADELANTAR O RE	COLISION	SI	URBANA	4	-78,49563499970	-0,22854300006	RUMIÑAHUI Y PICH
2016	ENERO	28/01/2016	JUEVES	20:58:00	20	NO	NO GUARDAR LA D	ROZAMIENTO	SI	URBANA	2	-78,48439400020	-0,23741299991	SIMON BOLIVAR-R
2016	ENERO	29/01/2016	VIERNES	0:33:00	0	NO	CONDUCIR EN EST	ESTRELLAMIENTO	SI	URBANA	1	-78,48468500010	-0,23767600035	SIMON BOLIVAR-
2016	FEBRERO	02/02/2016	MARTES	16:07:00	16	NO	CONDUCIR DESAT	ESTRELLAMIENTO	SI	RURAL	2	-78,48776400000	-0,26399099979	LAURELES - AV RU
2016	FEBRERO	04/02/2016	JUEVES	18:07:00	18	NO	REALIZAR CAMBI	ROZAMIENTO	SI	URBANA	2	-78,50246099970	-0,23327099956	VELASCO IBARRA
2016	FEBRERO	04/02/2016	JUEVES	19:16:00	19	NO	NO GUARDAR LA D	ROZAMIENTO	SI	URBANA	2	-78,50531700010	-0,23225400032	VELASCO IBARRA
2016	FEBRERO	04/02/2016	JUEVES	22:25:00	22	NO	CONDUCIR DESAT	PERDIDA DE PISTA	SI	URBANA	1	-78,48468500010	-0,23766699993	SIMON BOLIVAR Y
2016	FEBRERO	05/02/2016	VIERNES	8:05:00	8	NO	NO CEDER EL DER	CHOQUE LATERAL	SI	URBANA	2	-78,48480700030	-0,23821299993	SIMON BOLIVAR Y
2016	FEBRERO	09/02/2016	MARTES	12:27:00	12	SI	NO GUARDAR LA D	ROZAMIENTO	SI	URBANA	2	-78,48457400040	-0,23760399973	SIMON BOLIVAR-I
2016	FEBRERO	10/02/2016	MIERCOLES	5:58:00	5	NO	CONDUCIR VEHICU	PERDIDA DE PISTA	SI	URBANA	1	-78,48461799970	-0,23736899968	SIMON BOLIVAR-I
2016	FEBRERO	15/02/2016	LUNES	15:15:00	15	NO	CONDUCIR VEHICU	ESTRELLAMIENTO	SI	RURAL	2	-78,48495000000	-0,25429499996	ZONA SUR

Figura 14. Siniestros 2016-2018 Parte 1



CANTON	PARROQUIA	IRO. FALLECIDO	NRO. LESIONADOS	IESTRO CERCANO A PUENTE (S	NOMBRE DE PUENTE	ZONA DE PLANIFICACIÓN	IPO DE VEHÍCUL	SERVICIO	EDAD	GENERO	CONDICIÓN
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	LOS CHILLOS	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	1	SI	PUENTE 1	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	2	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	LOS CHILLOS	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	1	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	LOS CHILLOS	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	LOS CHILLOS	0	0	SI	PUENTE 2	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	1	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	QUITUMBE	0	1	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	NO		ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	0	SI	VIA SIMON BOLIVAR A PUEN	ZONA 9	N/A	N/A	0	N/A	N/A
QUITO	MANUELA SAENZ	0	1	NO		ZONA 9	N/A	N/A	0	N/A	N/A

Figura 15. Siniestros 2016-2018 Parte 2

4.3 ANÁLISIS DE LA INFORMACIÓN

Para una mejor visualización de los datos se procedió a usar la herramienta de R Studio, con el fin de generar un análisis netamente estadístico, de los datos entregados por la A.N.T.

Dentro del primer gráfico estadístico, figura 16, podemos notar que la mayor cantidad de siniestros de transito son producidos por personas del sexo Masculino con un total de 87,69% mientras que las personas de sexo Femenino producen un aproximado del 12,31% de los siniestros.

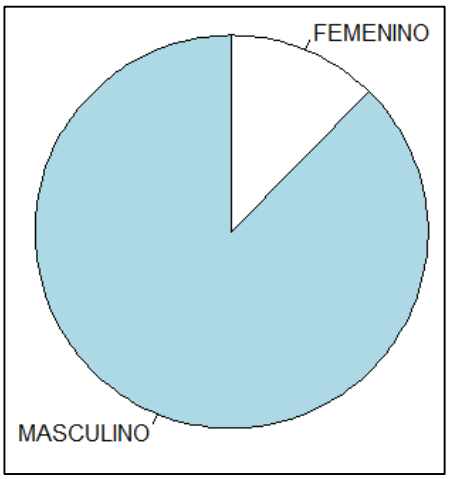


Figura 16. Siniestros por Género

Los vehículos que más siniestros causaron fueron los automóviles, seguidos de las motocicletas, camiones y buses; estos primeros indicios permitieron enfocar la recolección de información, bajo ciertos aspectos.

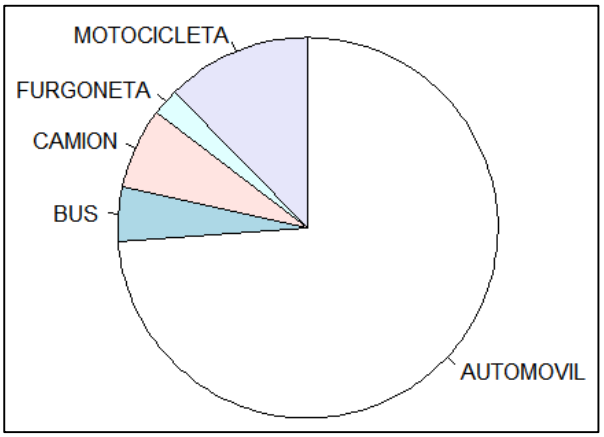


Figura 17. Índice de Siniestralidad en Vehículos

Dentro de los meses con más siniestralidad se obtuvieron los meses de Julio, Abril, Mayo y Junio, respecto a los días son Sábado, Domingo y Viernes.

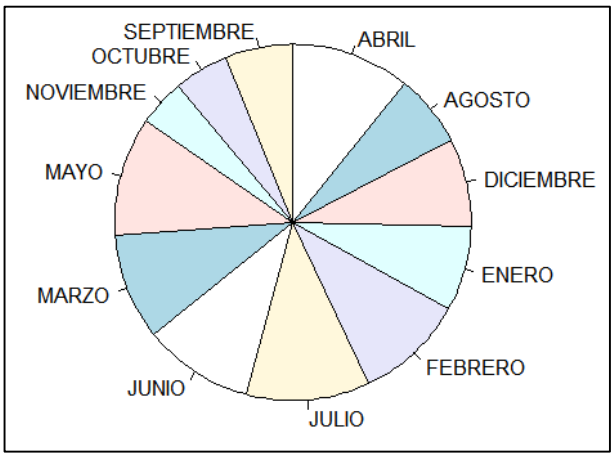


Figura 18. Siniestros Por Mes

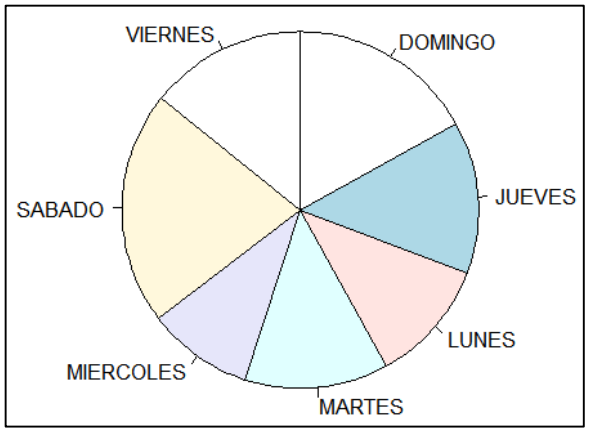


Figura 19. Siniestros Por Días

Las horas con mayor ocurrencia de siniestros de transito según la figura 20, es el periodo entre las 12 de la noche y las 5 de la mañana, seguido de las 3 de la tarde a las 11 de la noche.

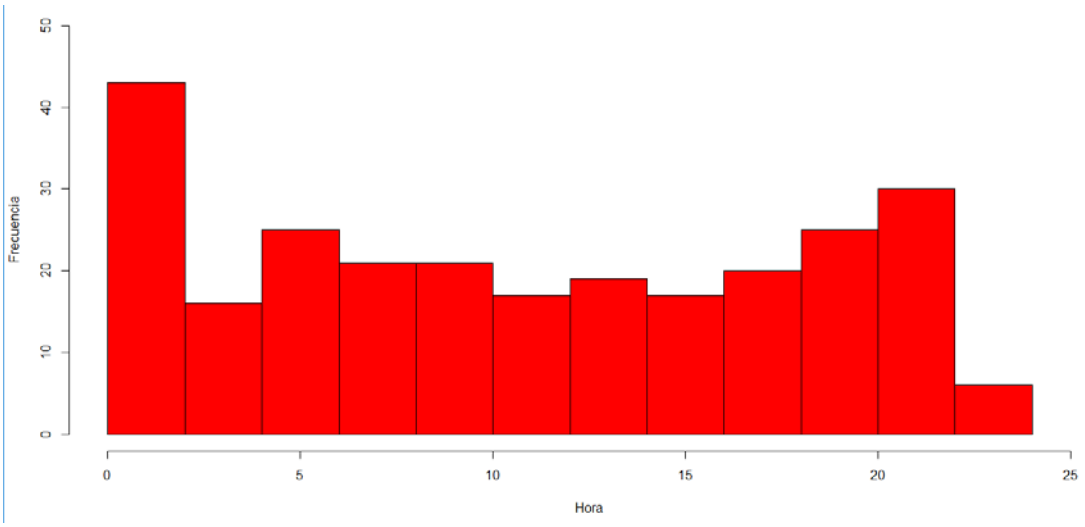


Figura 20. Índice de Siniestralidad por Hora

La principal causa según los datos de la A.N.T. para el periodo antes mencionado fue el conducir un vehículo superando los límites de velocidad (exceso de velocidad), seguido de conducir desatento, no tener prudencia frente a otros conductores, conducir bajo influencia de alcohol, etcétera.

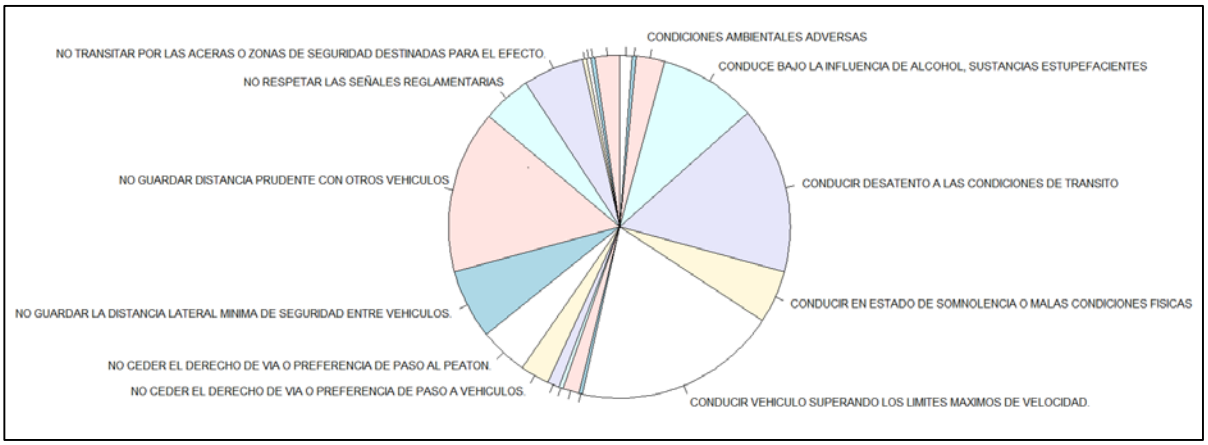


Figura 21. Principales Causas de Siniestros

Los principales efectos resultantes de un siniestro son choques, pérdidas de pista, estrellamiento y atropello, como se puede ver en la Figura 22.

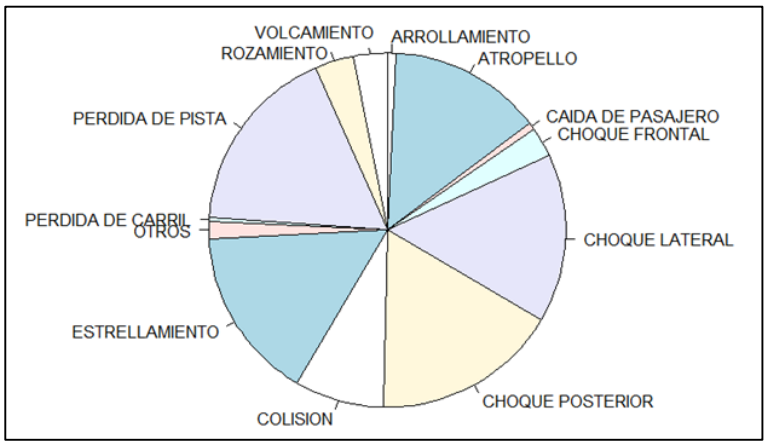


Figura 22. Principales Efectos de Siniestros

En este periodo los siniestros dieron como resultado un alto porcentaje de personas ilesas, seguido de un menor número de personas lesionadas y una pequeña cantidad de fallecidos, Figura 23.

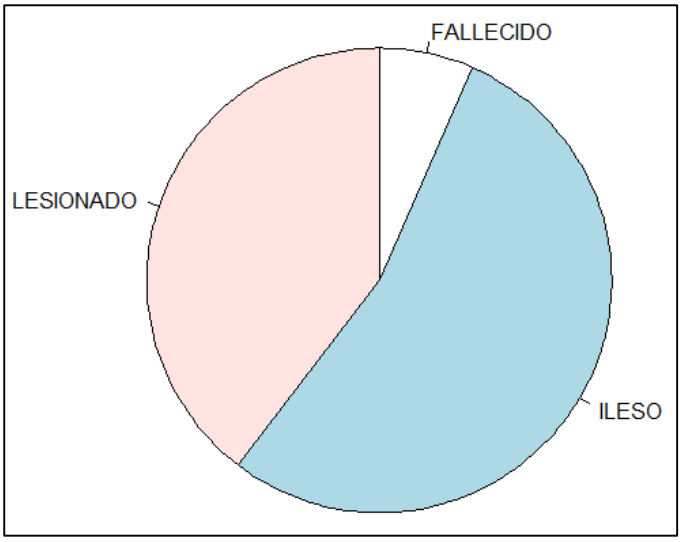


Figura 23. Condición de la Persona Dentro del Siniestro

CAPÍTULO V

PROTOTIPADO Y MODELO DE PREDICCIÓN

5.1 DISEÑO DE LA ARQUITECTURA DE LA APLICACIÓN PARA RECOPILAR DATOS

Para la obtención de datos, se escogió desarrollar una aplicación para dispositivos Android, basada en la nube, con servicios proporcionados por Google Maps, esto para el despliegue del mapa de la autopista y la obtención de coordenadas, Figura 24.

Figura 24. Arquitectura de la Aplicación

La aplicación accede mediante el uso de GPS y redes móviles a los servicios de Google Maps para obtener datos en tiempo real de la ubicación de un usuario, a la vez que transmite la información a una Base de Datos MySQL, contenida en la nube.

5.2 MODELADO DE LA BASE DE DATOS

Se procedió al modelado de una Base de Datos relacional MySQL, en la cual se consideró las siguientes variables: Código, ID, edad del conductor, genero del conductor, tipo de licencia del

conductor, tipo de vehículo, año del vehículo, clima, fecha, hora, temperatura, coordenadas. Estas relaciones se pueden observar en la figura 25.

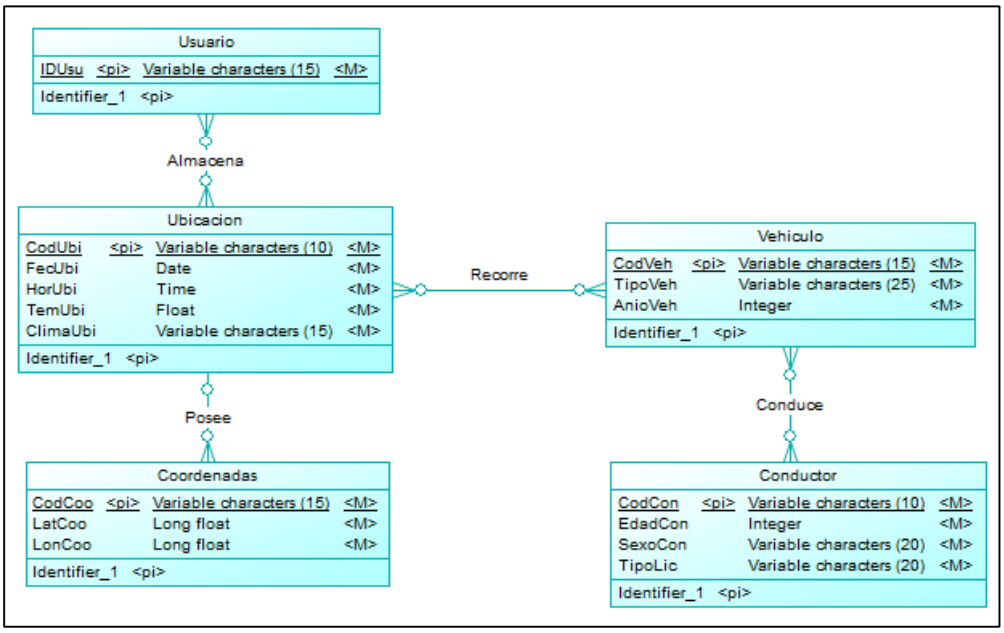


Figura 25. Modelo Conceptual de la Base de Datos

Las relaciones entre cada tabla señalan que: un usuario puede almacenar una o varias ubicaciones, pero estas ubicaciones pueden ser almacenadas por uno o varios usuarios a la vez. Las ubicaciones poseen un conjunto de coordenadas, pero una coordenada solo pertenece a una ubicación. Las ubicaciones son recorridas por uno o varios vehículos es así que un vehículo puede recorrer una o varias ubicaciones. Un vehículo puede ser conducido por uno o varios conductores y un conductor puede conducir uno o varios vehículos.

5.3 DISEÑO DEL PROTOTIPO DE LA APLICACIÓN

El software para recolección de datos, fue desarrollado en el IDE de Android Studio, utilizando la API 16 perteneciente a la versión de Android 4.1 (Jelly Bean), brindando una mayor compatibilidad para diferentes dispositivos. Fueron requeridas dos pantallas dentro de esta (Main y Maps).

Para que la aplicación sea amigable con el usuario, se incluyeron elementos de tipo Spinner, para reducir la cantidad de texto ingresada de forma manual, evitando datos erróneos por una digitación incorrecta, Figura 26.

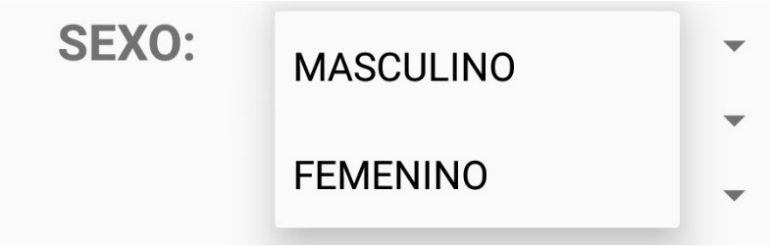


Figura 26. Elemento Spinner Para Variable Sexo

El único campo considerado como texto de entrada fue el ID del usuario (Cedula o ID Mi ESPE), Figura 27, el resto de variables son obtenidas a través del dispositivo de manera automática, mediante la implementación de funciones como la mostrada en la Figura 28.

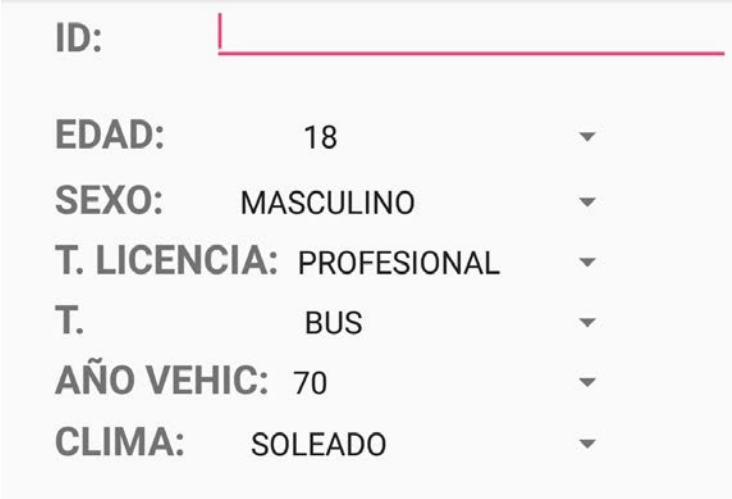


Figura 27. Variables de la Pantalla Principal

```
Calendar c=Calendar.getInstance();
int mes=c.get(Calendar.MONTH)+1;
String sDate=c.get(Calendar.YEAR)+"/"+mes +"/"+ c.get(Calendar.DAY_OF_MONTH);
String sTime=c.get(Calendar.HOUR_OF_DAY) + ":" + c.get(Calendar.MINUTE)+":"
        +c.get(Calendar.SECOND);
```

Figura 28. Clase Para la Obtención de Día y Hora

Dentro de la segunda interfaz, de tipo Mapa, se indica la ubicación del usuario, y al mismo tiempo se extraen las coordenadas, Figura 29.

Figura 29. Implementación del Mapa de Google

Para el primer prototipo de la aplicación que se muestra en la figura 30, se dispuso de un diseño sencillo, enfocado en el desarrollo de pruebas de funcionalidad.

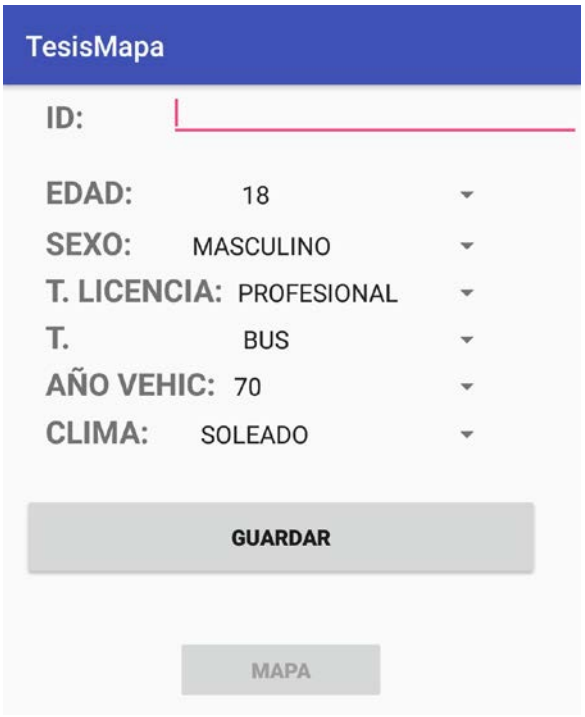


Figura 30. Prototipo de la Pantalla Principal

Las primeras pruebas de la aplicación presentaron un problema de ralentización, debido a que los datos obtenidos a través de Google Maps eran retornados en un periodo de entre 3 a 5 segundos, resultando en un cálculo de velocidades con un grado de error elevado. Como solución se realizó una modificación en el código, implementando una clase de tipo LocationManager (Android Developers, 2018), que permite obtener datos directamente de la señal de GPS obtenida a través de la red telefónica. Sin embargo esta solución acarreo otro problema, puesto que no todos los dispositivos Android disponen de un GPS de alto rendimiento y precisión, por esta razón la recolección de datos, se enfocó en el uso teléfonos de gama media - alta. Es así que las pruebas funcionales fueron llevadas a cabo dentro de 6 dispositivos Android: LG G5, LG G7, Samsung J7, Samsung S8, Huawei P10 y Motorola G5S.

```
LocationManager locationManager = (LocationManager) getSystemService(Context.LOCATION_SERVICE);


Location location = locationManager.getLastKnownLocation(LocationManager.GPS_PROVIDER);
actualizarUbicacion(location);
locationManager.requestLocationUpdates(LocationManager.GPS_PROVIDER,  minTime: 10,  minDistance: 0, locListener);
```

Figura 31. Implementación de Clase LocationManager

Cuando las pruebas finalizaron, se modificó el diseño de la aplicación, para otorgar una interfaz visualmente agradable y organizada, Figura 32.

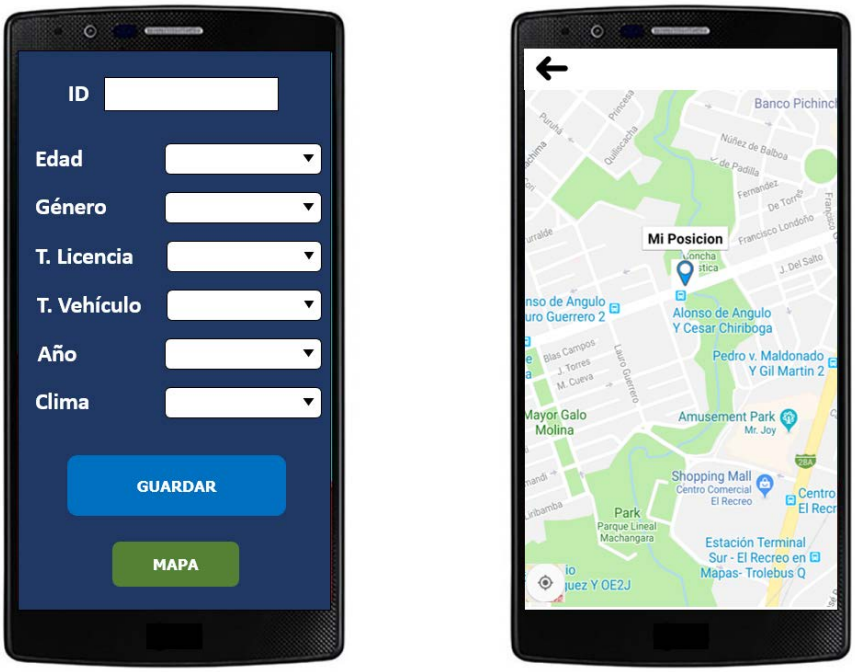


Figura 32. Versión Final de la Aplicación

5.4 PROCESO DE RECOLECCIÓN DE DATOS

Se entregó la aplicación a 10 usuarios, que faciliten la recolección de información, en dos horarios, Tabla 5. Durante 2 semanas consecutivas.

Tabla 5.
Horas de Recolección de Datos

Periodo	Horario 1	Periodo	Horario 2
6	6:00 – 6:59 am	14	2:00 – 2:59 pm
7	7:00 – 7:59 am	15	3:00 – 3:59 pm
8	8:00 – 8:59 am	16	4:00 – 4:59 pm
9	9:00 – 9:59 am	17	5:00 – 5:59 pm
10	10:00 – 10:59 am	18	6:00 – 6:59 pm
11	11:00 – 11:59 am	19	7:00 – 7:59 pm

12	12:00 – 12:59 pm	20	8:00 – 8:59 pm
13	1:00 – 1:59 pm	21	9:00 – 10:00 pm

Cada viaje tuvo un aproximado de entre 35 y 45 minutos, dependiendo el tipo de vehículo y la hora de movilidad. Con esto se obtuvieron alrededor de 1.000 datos por viaje, con un total de 16 viajes diarios por los 5 días de la semana, consiguiendo poco más de 80.000 registros.

5.5 PROCESO DE FILTRACIÓN Y VALIDACIÓN DE DATOS

Previo a su análisis se realizó un proceso de ETL (Extract, Transform and Load), utilizando el software de RapidMiner Studio, Figura 33, según Muñoz, Trujillo, & Mazón (2011) el proceso de ETL suele ser pieza fundamental en el éxito de proyectos de data mining, aunque su principal desventaja es su complejidad y su excesivo tiempo al ser aplicado.

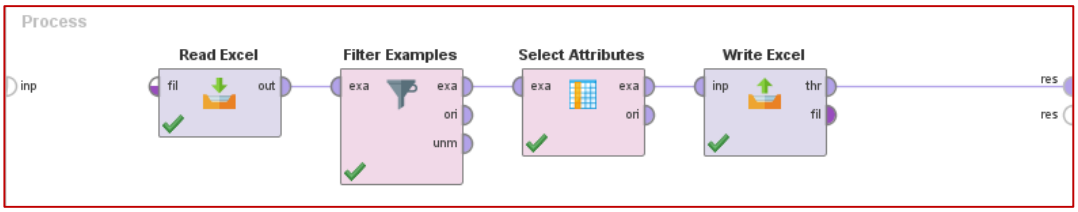


Figura 33. Proceso de Filtrado de Datos

En la figura 33 se puede observar el uso de 4 operadores principales: Read Excel, permite leer el archivo de datos incial; Filter Examples, filtra datos que contegan características particulares, en este caso velocidades mayores a 190 Km/h y distancias mayores a 53 metros; Select Attributes, permite delimitar los atributos requeridos para su posterior estudio; y Write Excel: genera un documento de tipo Excel con todos los datos procesados.

Como segunda etapa de este proceso, se realizó una segmentación por zonas cada 1 Km. entre el Trébol y la Universidad de la Fuerzas Armadas ESPE, obteniendo un total de 16 Zonas, estas a su vez conformadas por 5 subzonas de 200 metros respectivamente, Figura 34 - 49.

Figura 34. Zona 1 (SubZona 1- 5)

Figura 35. Zona 2 (SubZona 6- 10)

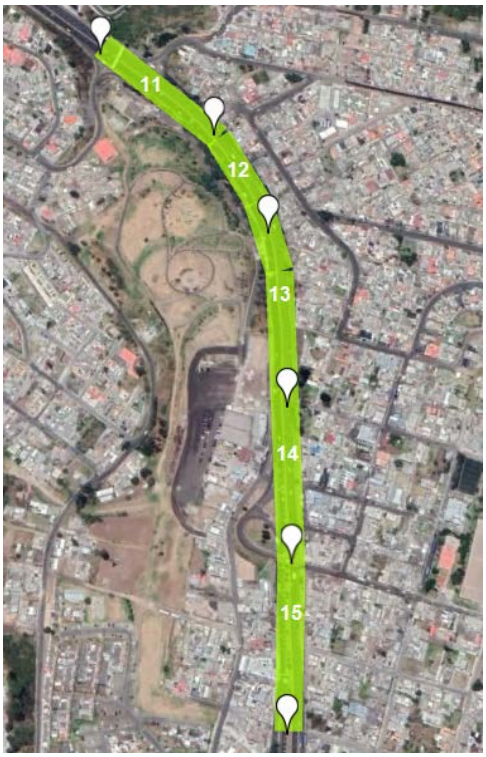


Figura 36. Zona 3 (SubZona 11- 15)

Figura 37. Zona 4 (SubZona 16- 20)

Figura 38. Zona 5 (SubZona 21- 25)

Figura 39. Zona 6 (SubZona 26- 30)

Figura 40. Zona 7 (SubZona 31- 35)

Figura 41. Zona 8 (SubZona 36- 40)

Figura 42. Zona 9 (SubZona 41- 45)

Figura 43. Zona 10 (SubZona 46- 50)

Figura 44. Zona 11 (SubZona 51- 55)

Figura 45. Zona 12 (SubZona 56- 60)

Figura 46. Zona 13 (SubZona 61- 65)

Figura 47. Zona 14 (SubZona 66- 70)

Figura 48. Zona 15 (SubZona 71- 75)

Figura 49. Zona 16 (SubZona 76- 78)

5.6 DISEÑO DEL MODELO PREDICTIVO

En esta sección de la investigación se pusieron a prueba técnicas predictivas (operadores predictivos) que formen parte del software RapidMiner, describiendo su funcionamiento (ventajas y desventajas), como podemos observar en la tabla 6.

Tabla 6.
Operadores Predictivos en RapidMiner

Técnica	Funcionalidad
K-NN	**1.** Permite el uso de atributos polinomiales y numéricos. **2.** Requiere de un Label (atributo a predecir). **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de seleccionar la medida de los K vecinos cercanos
Naive Bayes	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), de tipo polinomial. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión.
Arboles de Decisión	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), para obtener reglas de decisión, de tipo polinomial. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar su profundidad y su confianza.

Reglas de Inducción	**1.** Permite el uso de atributos numéricos y polinomial **2.** Requiere de un Label (atributo a predecir), este debe ser polinomial. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar el criterio de selección de atributos, relación de muestra, pureza y la cantidad mínima de beneficio.
Redes Neuronales	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir). **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar los ciclos de entrenamiento, taza de aprendizaje y su momentum.
Deep Learning	**1.** Permite el uso de atributos numéricos y polinomial **2.** Requiere de un Label (atributo a predecir). **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de seleccionar la función de activación, la cantidad y capacidad de las capas ocultas y el número de iteraciones
Regresión Lineal	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), de tipo numérico. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar su tolerancia.
Regresión Logística	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), de tipo binomial. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar su capacidad máxima de iteraciones y su solucionador (funciones).
SVM	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), de tipo numérico. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión. **5.** Capacidad de modificar el tipo de Kernel, kernel gamma, kernel cache, convergencia de épsilon y el máximo de iteraciones.
Análisis de Discriminante	**1.** Permite el uso de atributos numéricos **2.** Requiere de un Label (atributo a predecir), de tipo polinomial. **3.** Capacidad de medir su rendimiento. **4.** Validación a través de una matriz de confusión.

De lo obtenido de la tabla anterior, se pudo identificar aquellas técnicas que permitan obtener un modelo de predicción acorde a lo especificado. Para lo cual se procedió a descartar aquellas técnicas que presentan restricciones dentro de sus Labels y operadores que no dispongan

de atributos (marcadas en rojo), ya que esto limita el tipo de variable que se quiera predecir y el mejoramiento del modelo. Como resultado, se obtuvieron un total de 4 técnicas, estas se evaluaron considerando lo siguiente:

Se aplicó un nuevo filtro al conjunto de datos, para generar predicciones más precisas, se escogieron aquellos datos con velocidades mayores a 30 Km/h, considerando a esta velocidad como el punto de partida para la ocurrencia de siniestros (ANT, n.d.). Se generaron nuevos atributos, basados en la Tabla 7.

Tabla 7.

Diccionario de Datos De Atributos Generados

Atributos Iniciales	Atributos Generados	Nombre de los Atributos
Edad	1. Conductores con edad menor o igual a 30 años	**6.** Edad<=30
	2. Conductores con edad mayor a 30 años	**7.** Edad>30
Tipo de Licencia	**1.** Licencia Profesional	**5.** LicPro
	2. Licencia Sportman	**6.** LicSpo
Tipo de Vehículo	**1.** Bus	**6.** Bus
	2. Auto	**7.** Auto
Año del Vehículo	**1.** Vehículos del años menores o iguales al 2010	**6.** AniVeh<=2010
	2. Vehículos del años mayores al 2010	**7.** AniVeh>2010
Clima	**1.** Templado	**6.** ClimTemp
	2. Soleado	**7.** ClimSolea
	3. Lluvia	**8.** ClimLluv
Días	**1.** Lunes	**6.** Lunes
	2. Martes	**7.** Martes
	3. Miércoles	**8.** Miercoles
	4. Jueves	**9.** Jueves
	5. Viernes	**10.** Viernes
Hora	**1.** Hora entre las 6 y 8 am	**6.** Hora6-8
	2. Hora entre las 9 y 12 am	**7.** Hora9-12
	3. Hora entre las 1 y 3 pm	**8.** Hora13-15
	4. Hora entre las 4 y 5 pm	**9.** Hora16-17
	5. Hora entre las 6 y 9 pm	**10.** Hora18-21
Ruta Trébol - ESPE	**1.** Ruta desde el Trébol hacia la ESPE	**6.** Treb_Espe
	2. Ruta desde la ESPE hacia el Trébol	**7.** Espe_Treb
Puntos	**1.** Zona 1 y 2 (10 Subzonas)	**5.** Pun1-10
	2. Zona 3 y 4 (10 Subzonas)	**6.** Pun11-20

	3. Zona 5 y 6 (10 Subzonas) **4.** Zona 7 y 8 (10 Subzonas) **5.** Zona 9 y 10 (10 Subzonas) **6.** Zona 10 y 11 (10 Subzonas) **7.** Zona 12 y 13 (10 Subzonas) **8.** Zona 14 , 15 y 16 (13 Subzonas)	**7.** Pun21-30 **8.** Pun31-40 **9.** Pun41-50 **10.** Pun51-60 **11.** Pun61-70 **12.** Pun71-80
Rango de Velocidad	**1.** Rango de velocidad excedido **2.** Rango de Velocidad normal	**1.** RanExe **2.** RanNor
Temperatura	**1.** Temperatura menor o igual a 15 grados centígrados **2.** Temperatura mayor a 15 grados centígrados	**1.** Temp<=15 **2.** Temp>15

5.7 EVALUACIÓN DE TÉCNICAS PREDICTIVAS

Las técnicas fueron evaluadas bajo parámetros similares, tabla 8. Esto permitió considerar aquella técnica con menor error cuadrático y absoluto.

Tabla 8.
Parámetros Para Prueba de Técnicas

Parámetros	**Valores**
Cantidad de Atributos	• 29
Split Data	• 70% para Entrenamiento • 30% para Prueba
Label	• Velocidad

5.7.1. K-NN

Para el uso de K-NN, se siguió la siguiente estructura.

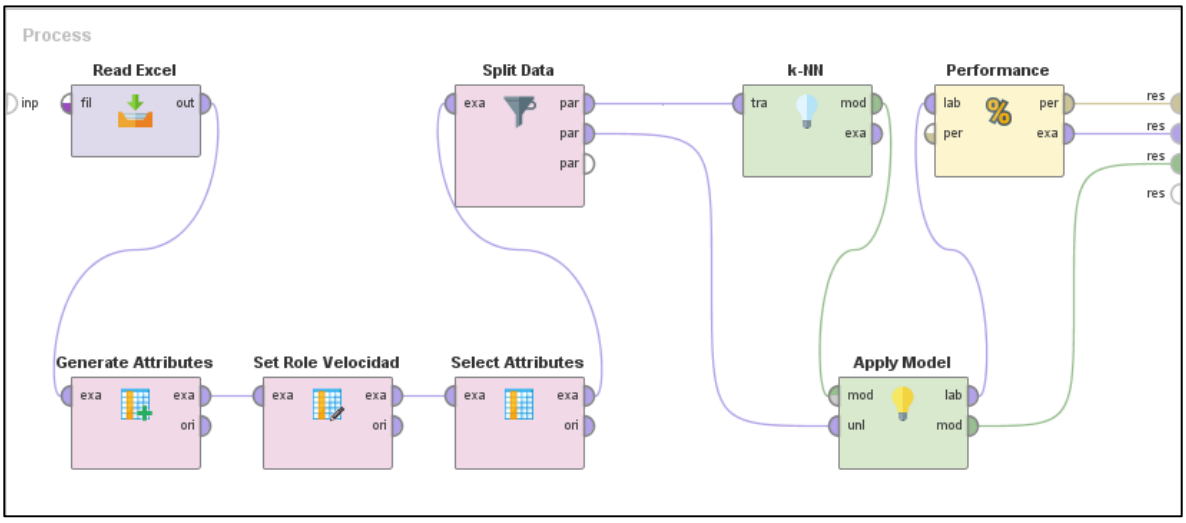


Figura 50. Proceso de K-NN

El operador de K-NN, fue configurado con un k=80 ya que permitió obtener la menor cantidad de error en la prueba.

5.7.2 Redes Neuronales

El proceso para Red Neuronal es:

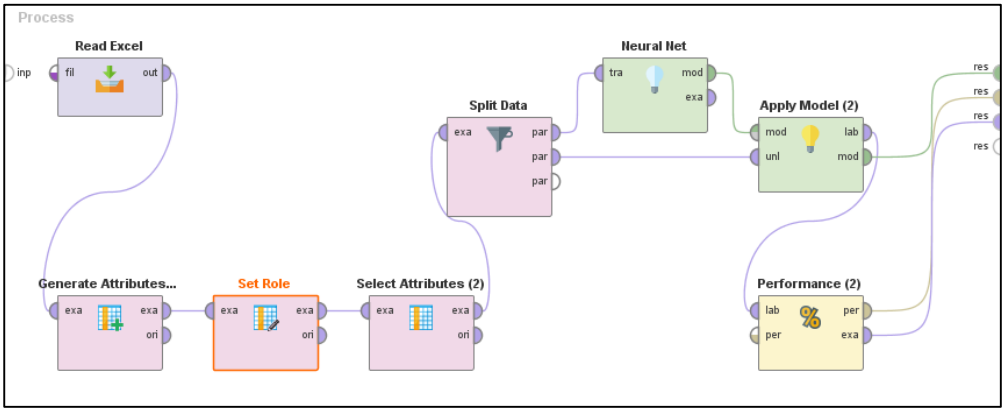


Figura 51. Proceso de Red Neuronal

Se configuro el operador Neural Net con un total de 500 iteraciones, una taza de aprendizaje de 0.01, un momentum de 0.75 y 4 capas ocultas con un tamaño de 16, 17, 15 y 16.

5.7.3 Deep Learning

Deep Learning usó la siguiente estructura.

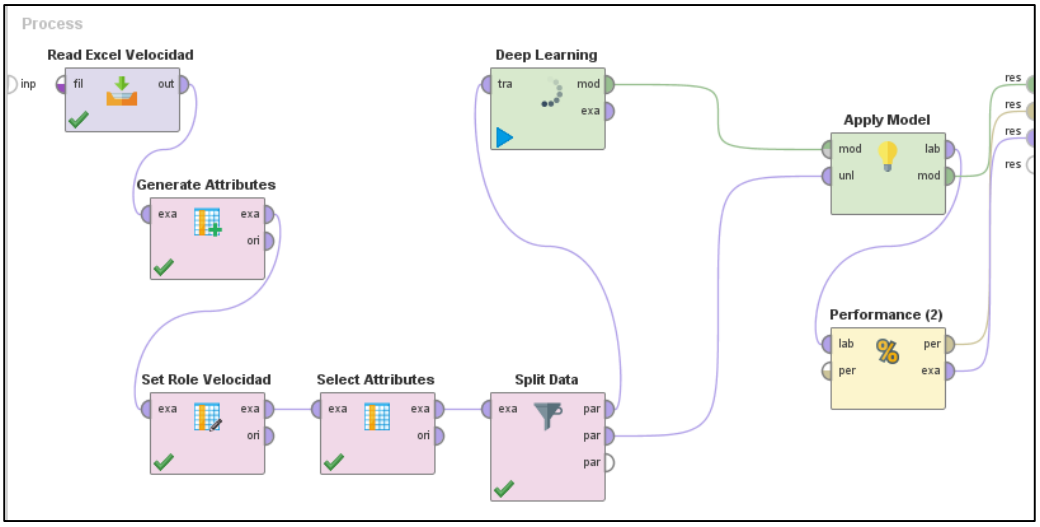


Figura 52. Proceso de Deep Learning

El operador Deep Learning tiene 50 iteraciones y dos capas ocultas con un tamaño de 75 cada una.

5.7.3 SVM

La estructura de SVM es:

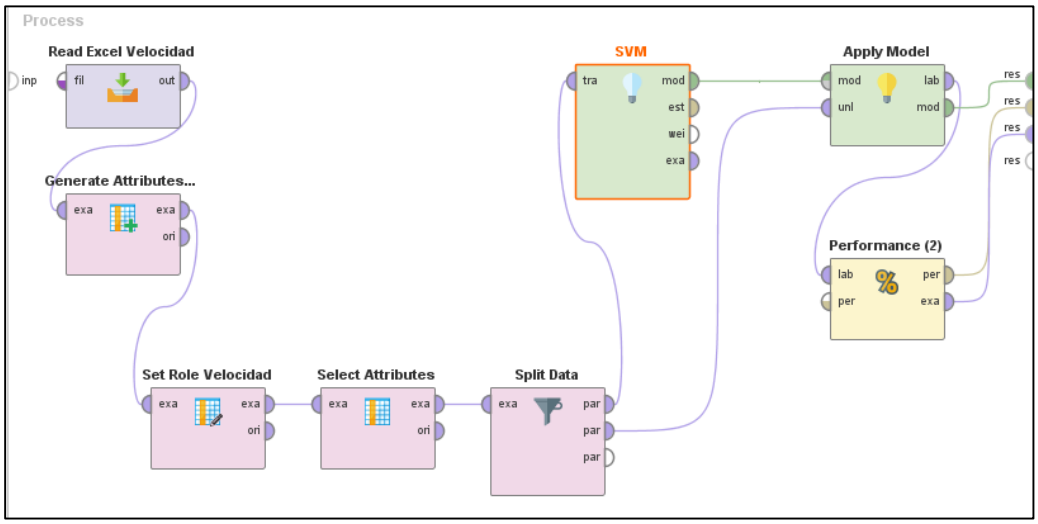


Figura 53. Proceso de SVM

El operador se configuro con un kernel gamma de 1, kernel cache de 200 y 1000 iteraciones.

Al finalizar de evaluar cada técnica con un operador de desempeño, se obtuvieron sus respectivos errores absolutos, contemplados en la tabla 9.

Tabla 9.

Desempeño de las Técnicas Predictivas

Técnica	Error Absoluto
K-NN	8,685
Redes Neuronales	7,675
Deep Learning	7,905
SVM	8,817

Con estos resultados se consideró usar la técnica de Redes Neuronales, al ser esta la que entrega un menor error absoluto.

5.8 IMPLEMENTACIÓN DEL MODELO PREDICTIVO

Se reestructuro la técnica de Redes Neuronales, implementado la opción de predicción de siniestros, basándose en la información provista por la ANT. Esta información recopilaba siniestros ocurridos a lo largo Autopista General Rumiñahui en un periodo de tres años 2016 – 2018. Gracias a esto se pudo determinar los lugares con mayor incidencia de siniestros, así como los días y horas en las cuales ocurrieron.

Los datos considerados se muestran en la figura 54, los cuales refieren a un mismo comportamiento, enfocado en registros en los cuales un siniestro sea causado por exceso de velocidad, por no mantener una distancia prudente con los demás conductores tanto de manera frontal como lateral.

AÑO	DIA	PERIODO	CAUSA	PUNTOS	TREB-ESPE	VEHÍCULO	EDAD	GENERO
2018	VIERNES	11	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	16,17	SI,NO	AUTO	21	MASCULINO
2018	MARTES	17	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	16,17	SI,NO	AUTO	28	MASCULINO
2018	MIÉRCOLES	16	CONDUCIR VEHÍCULO SUPERANDO LOS LÍMITES MÁXIMOS	16,17	SI,NO	AUTO	0	MASCULINO
2018	MIÉRCOLES	10	CONDUCIR VEHÍCULO SUPERANDO LOS LÍMITES MÁXIMOS	16,17	SI,NO	AUTO	21	MASCULINO
2018	MIÉRCOLES	6	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	8,9	SI	AUTO	0	MASCULINO
2018	MARTES	9	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	14,15	SI,NO	AUTO	36	MASCULINO
2018	JUEVES	6	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	20	SI,NO	AUTO	62	MASCULINO
2018	VIERNES	7	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	1	NO	AUTO	44	MASCULINO
2018	MARTES	18	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	1	NO	AUTO	45	MASCULINO
2017	MIÉRCOLES	8	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	8,9	SI	BUS	35	MASCULINO
2017	JUEVES	11	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	1	NO	AUTO	32	MASCULINO
2017	MARTES	7	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	16,17	SI,NO	AUTO	35	MASCULINO
2017	MARTES	13	CONDUCIR VEHÍCULO SUPERANDO LOS LÍMITES MÁXIMOS	16,17	SI,NO	AUTO	30	MASCULINO
2017	JUEVES	20	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	16,17	SI,NO	AUTO	23	MASCULINO
2017	JUEVES	6	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	16,17	SI,NO	BUS	32	MASCULINO
2017	VIERNES	8	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	16,17	SI,NO	AUTO	30	MASCULINO
2017	VIERNES	20	NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDA	16,17	SI,NO	AUTO	20	MASCULINO
2017	VIERNES	9	NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO	16,17	SI,NO	AUTO	47	MASCULINO

Figura 54. ANT, Siniestros años 2016-2018

Dentro del conjunto de datos que sirvió para la determinación del modelo de predicción, se implementó un nuevo atributo, denominado **Probabilidad Ocurrencia de Siniestro**, esta especifica si un siniestro puede o no ocurrir bajo determinados parámetros.

Aquí se reutilizo la estructura del proceso de la Figura 51, la Red Neuronal está presente en la Figura 55.

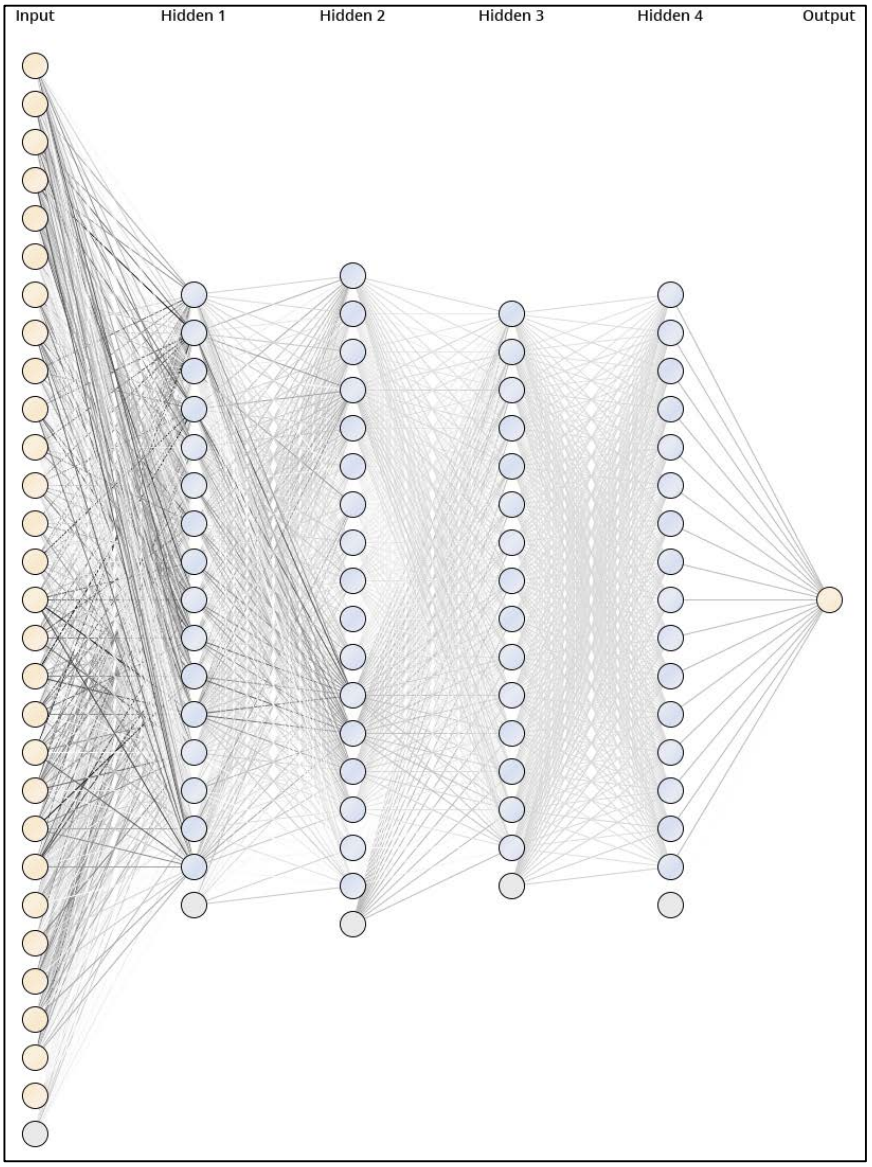


Figura 55. Red Neuronal

CAPÍTULO VI

VALIDACIÓN DEL MODELO

6.1 PRUEBAS DEL MODELO

Para el proceso de pruebas se propusieron dos escenarios:

Escenario 1

En el primer escenario se procedió a realizar una validación de predicción aplicando una división de los datos de manera recurrente, utilizando el operador Split, figura 56. Esto se realizó para verificar la exactitud de la predicción en cada uno de los cluster, Tabla 10.

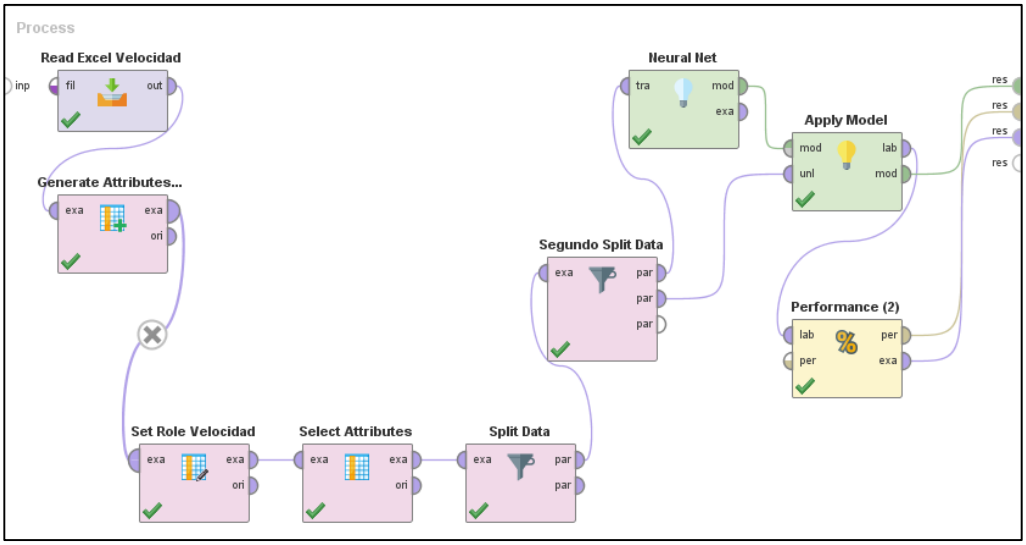


Figura 56. Proceso de Pruebas Red Neuronal

Tabla 10.
División de Datos Para Prueba

Split 1	Porcentaje	Split 2	Porcentaje
Entrenamiento 1	70%	Entrenamiento 2	70%
		Prueba 2	30%
Prueba 1	30%	Entrenamiento 3	70%
		Prueba 3	30%

Una vez obtenido los diferentes grupos de datos se procedió a evaluar cada nuevo Split, verificando cada uno de los errores absolutos para cada caso, Tabla 11.

Tabla 11.
Calculo de Error Absoluto para Splits

Split	Nombre	Error Absoluto
2	Entrenamiento 2	8,085
	Prueba 2	
3	Entrenamiento 3	8,644
	Prueba 3	

Escenario 2

Para el escenario 2 se volvieron a recolectar cerca de 600 datos utilizando la aplicación, estos sirvieron para verificar la exactitud del modelo, se incluyeron en el Split de entrenamiento, con la siguiente estructura, figura 57.

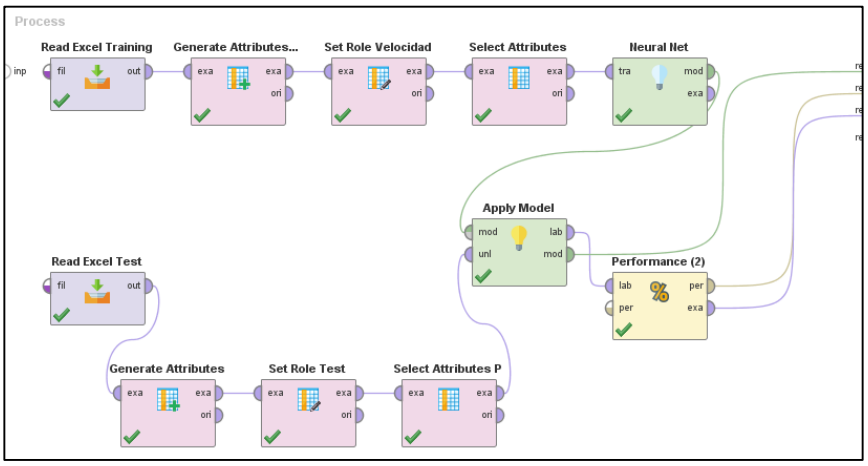


Figura 57. Prueba Red Neuronal

Se pudo predecir los nuevos valores con un error absoluto de 15,419. Con lo cual el modelo ofrece un comportamiento predictivo, cercano a la realidad.

6.2 VALIDACIÓN DEL MODELO

Dentro del proceso de validación se procedió a obtener el modelo de Kernel, Figura 58, en el cual se expresa el soporte de cada atributo, y a su vez nos brinda una breve descripción de aquellos atributos que pudieron ser descartados (valores que tienda a 0), ya que no producen ningún efecto sobre el modelo, como se puede observar en el atributo sexo y en cierto grado el atributo Pun71-80.

```
Total number of Support Vectors: 37881
Bias (offset): 0.257

w[Velocidad] = 1059.408
w[Edad<=30 = if(Edad<=30,1,0)] = 561.550
w[LicPro = if(Tlicencia=="PROFESIONAL",1,0)] = 616.853
w[Bus = if(Tvehiculo=="BUS",1,0)] = 568.446
w[AniVeh<=2010 = if(Anio<=2010,1,0)] = 596.232
w[ClimTemp = if(Clima=="TEMPLADO",1,0)] = 555.761
w[ClimSolea = if(Clima=="SOLEADO",1,0)] = 308.637
w[ClimLluv = if(Clima=="LLUVIA",1,0)] = 548.700
w[Lunes = if(Dia=="LUNES",1,0)] = 568.406
w[Miercoles = if(Dia=="MIERCOLES" ,1,0)] = 442.826
w[Viernes = if(Dia=="VIERNES" ,1,0)] = 492.678
w[Hora6-8 = if(Hora<=8,1,0)] = 330.551
w[Hora9-12 = if(Hora<=12 && Hora >=9 ,1,0)] = 701.666
w[Hora13-15 = if(Hora<=15 && Hora >=13 ,1,0)] = 434.522
w[Hora16-17 = if(Hora<=17 && Hora >=16 ,1,0)] = 297.852
w[Treb-Espe = if([Treb-ESPE]=="SI" ,1,0)] = 634.318
w[Pun1-10 = if(Puntos<11 ,1,0)] = 158.055
w[Pun11-20 = if(Puntos<=20 && Puntos>=11 ,1,0)] = 1169.244
w[Pun21-30 = if(Puntos<=30 && Puntos>=21 ,1,0)] = 252.522
w[Pun31-40 = if(Puntos<=40 && Puntos>=31 ,1,0)] = 255.613
w[Pun41-50 = if(Puntos<=50 && Puntos>=41 ,1,0)] = 521.877
w[Pun51-60 = if(Puntos<=60 && Puntos>=51 ,1,0)] = 293.177
w[Pun61-70 = if(Puntos<=70 && Puntos>=61 ,1,0)] = 187.317
w[Pun71-80 = if(Puntos>70 ,1,0)] = 92.809
w[Martes = if(Dia=="MARTES" ,1,0)] = 529.346
w[Jueves = if(Dia=="JUEVES" ,1,0)] = 288.847
w[Temp<=15 = if(Temperatura<=15 ,1,0)] = 419.177
w[Hora18-21 = if(Hora<=21 && Hora >=18 ,1,0)] = 429.163
w[Masculino = if(Sexo=="MASCULINO",1,0)] = 0.000
```

Figura 58. Modelo Kernel

Se validó el modelo haciendo uso de una Matriz de Confusión dentro de la herramienta de RapidMiner, se implementó el operador Performance Vector, Figura 59, para testear el modelo con falsos positivos y con falsos negativos.

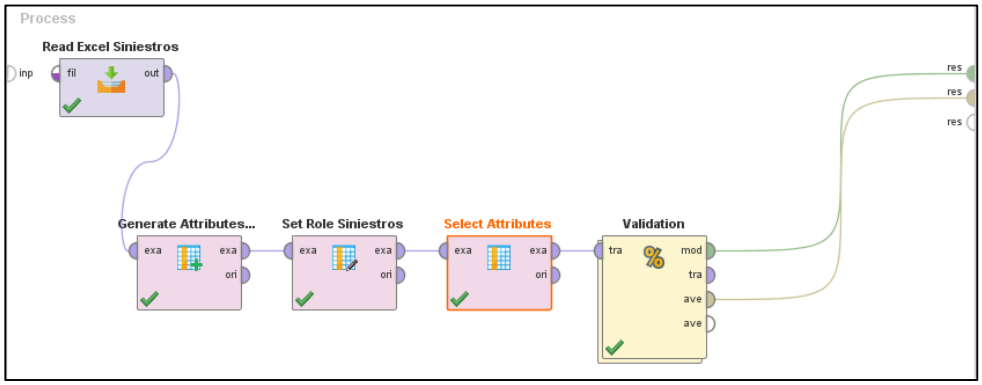


Figura 59. Validación de Red Neuronal

Como resultado podemos observar la matriz de confusión, mostrada en la Figura 60.

	true NO	true SI	true SI/NO	class precision
pred. NO	9902	0	160	98.41%
pred. SI	0	1298	4	99.69%
pred. SI/NO	0	0	0	0.00%
class recall	100.00%	100.00%	0.00%	

Figura 60. Matriz de Confusión Red Neuronal

De la matriz de confusión destacamos que: del total de no ocurrencias de siniestros, el 9902 fue clasificado como no siniestros, del total de ocurrencias de siniestros 1298 fueron clasificados como un posible siniestro, y del total de posibles siniestros 160 fueron clasificados como no ocurrencia de accidentes y 4 como posible ocurrencia de accidentes, de esto obtenemos un 98,5% de precisión general.

CAPÍTULO VII

CONCLUSIONES Y RECOMENDACIONES

7.1 CONCLUSIONES

En la actualidad la Agencia Nacional de Tránsito, se ha preocupado por recabar la información de siniestros de transito de una manera adecuada, considerando la mayor cantidad de atributos relacionados a este evento, esto es de gran importancia ya que es un tema crítico para el País.

Se evidencia que las estadísticas aplicadas a los datos iniciales otorgan datos muy interesantes como que los automóviles particulares son los que más accidentes causan.

En el análisis inicial de los datos proporcionados por la ANT, se puede notar que el mes con mayor incidencia de siniestros es el mes de Diciembre, y estos están ubicados principalmente en las provincias de Pichincha y Guayas.

Dentro de la información perteneciente a la Autopista General Rumiñahui, se pudo notar que los siniestros ocurridos en el año 2017-2018, fueron ocasionados por Hombres, que conducían automóviles, los días Viernes, Sábado y Domingo, los meses de Abril y Julio como principal causa el exceso de velocidad.

La granularidad dentro de los datos a ser analizados, son de gran importancia ya que de esto depende la precisión del modelo a obtener, para esta investigación fue necesario realizar varios filtrados y disminuir la granularidad de la información obtenida.

En esta investigación la técnica de Redes Neuronales ofreció el menor error absoluto, de esta manera se puede confirmar el supuesto de (Lippman & Lippman, 1987), ya que este tipo de perceptrón multicapa permite mejorar la precisión en las predicciones realizadas.

Para la predicción de posible siniestros, se requirió de información proporcionada por la Agencia Nacional de Tránsito, en un rango de años 2016-2018. Aquí se conocieron los lugares, horas y días con mayor ocurrencia de los mismos.

El modelo predictivo obtenido en esta investigación, puede verse afectado en años o meses, ya que pueden existir varios cambios, como por ejemplo nuevas normas y leyes dependiendo de las administraciones de tránsito o por decretos dentro del estado Ecuatoriano.

7.2 RECOMENDACIONES

Al obtener los datos desde una aplicación, se debe considerar la evaluación de cada uno de ellos, ya que pueden existir datos erróneos o redundantes ocasionados por diversos factores tanto internos como externos.

Es necesario el uso de varias técnicas al realizar un modelo de predicción, con el fin de verificar aquella que nos entregue la menor cantidad de error posible al ser usada, ayudando así a la validación de dicho modelo.

Al usar técnicas de minería de datos es recomendable, varias los valores ofrecidos por estas, ya que al combinar diferentes equivalencias dentro de los operadores de predicción, pueden resultar en la mejora considerable de dicha técnica.

Para mejorar notablemente la capacidad predictiva del modelo, se podría tomar en cuenta información de más años, de manera que el proceso de entrenamiento contemple más escenarios en los cuales la información aún puede ser incierta.

Para futuros trabajos se pueden considerar nuevas variables o factores, humanos y tecnológicos, como puede ser el estado de ánimo, salud del conductor, o eliminando aquellos atributos que no intervengan en el comportamiento final de los datos, de manera que permita obtener un mayor grado de precisión y confiabilidad del modelo.

La metodología aplicada en esta investigación, ayudo en la concepción de la solución al problema, sin embargo para futuros trabajos se podría usar una metodología enfocada a minería de datos.

REFERENCIAS

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Acuna, E. (2010). Mineria De Datos. *Universidad Cesar Vallejo*. Retrieved from http://www.slideshare.net/janettejf/mineria-de-datos-3582262

Android Developers. (2018). LocationManager. Retrieved November 22, 2018, from https://developer.android.com/reference/android/location/LocationManager?hl=ar

ANT. (n.d.). Estadísticas sobre Siniestros de Tránsito - Agencia Nacional de Tránsito del Ecuador - ANT. Retrieved May 30, 2018, from https://www.ant.gob.ec/index.php/noticias/estadisticas#siniestros-2017

Beltrán Martínez, M. B. (2014). Minería de datos. *Benemérita Universidad Autónoma de Puebla*, *1*(5), 67. Retrieved from http://bbeltran.cs.buap.mx/NotasMD.pdf

Cáceres, J. H. (2016). Clustering technique based on k- means algorithm for the identification of clusters of surgical patients. *Universidad Santo Tomás, Seccional Bucaramanga*, 1–8.

Choquehuanca-Vilca, V., Cárdenas-García, F., Collazos-Carhuay, J., & Mendoza-Valladolid, W. (2010). Epidemiological profile of road traffic accidents in Peru, 2005-2009. *Revista Peruana de Medicina Experimental y Salud Publica*, *27*(2), 162–169. https://doi.org/10.1590/S1726-46342010000200002

EDIMEC. (2014). *Manejo de emergencia a víctimas de accidentes de tránsito*.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). KDD for Science Data Analysis: Issues and Examples. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 50–56.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

Gandrud, C. (2015). Reproducible research with R and R studio, Second edition. *Chapman & Hall/CRC The R Series*, 323. https://doi.org/10.1201/b18546

García, C., & Gómez, I. (2006). Algoritmos de aprendizaje: knn & kmeans. *Universidad Carlos III de Madrid*, 1–8. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf

González, M. V. (2014). El uso del Perceptrón Multicapa para la clasificación de patrones en conductas adictivas, 1–55. Retrieved from http://dspace.uib.es/xmlui/bitstream/handle/11201/1126/TFG Marta Vidal González.pdf?sequence=1

Hall Mark, Eibe Frank, G. H. (1997). The WEKA Data Mining Software: An Update. *Marine Policy*, *7*(January 1996), 13–21.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining : concepts and techniques*. Burlington : Elsevier Science.

Hand, D. J. (2007). Principles of data mining. *Drug Safety*, *30*(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010

Híjar-Medina, M. C., Carrillo-Ordaz, C. E., Flores-Aldana, M. E., Anaya, R., & López-López, M. V. (1999). Factores de riesgo de lesión por accidentes de tráfico y el impacto de una intervención sobre la carretera. *Revista de Saúde Pública*, *33*(5), 505–512. https://doi.org/10.1590/S0034-89101999000500011

Jin, S.-B., & Lee, J.-W. (2017). Study on Accident Prediction Models in Urban Railway Casualty Accidents Using Logistic Regression Analysis Model. *Journal of the Korean Society for Railway*, *20*(4), 482–490. https://doi.org/10.7782/JKSR.2017.20.4.482

Kaur, G., & Kaur, E. H. (2017). Prediction of the cause of accident and accident prone location on roads using data mining techniques. *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. https://doi.org/10.1109/ICCCNT.2017.8204001

Lippman, R. P., & Lippman, R. P. (1987). An Intrduction to Computing with Neural Nets. *IEEE ASSP Magazine*, (April), 4–22. https://doi.org/10.1109/MASSP.1987.1165576

Mayou, R., Bryant, B., & Duthie, R. (1993). Psychiatric consequences of road traffic accidents. *Bmj*, *307*(6905), 647–651. https://doi.org/10.1136/bmj.307.6905.647

Muñoz, L., Trujillo, J., & Mazón, N. (2011). ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study. *Photonics Spectra*, *40*(1), 144. https://doi.org/10.1109/TLA.2011.5893784

OISEVI. (2018). Definiciones de Siniestro viales y víctimas. Retrieved October 25, 2018, from http://www.oisevi.org/a/index.php/estadisticas/glosario

Peng, R. D. (2011). Reproducible Research in Computational Science. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1126/science.1213847.Reproducible

Prasad Patil. (2018). K Means Clustering : Identifying F.R.I.E.N.D.S in the World of Strangers. Retrieved October 23, 2018, from https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d

RAMOS, ARAM; SILVA, ELID;AGUIRRE, A. (2015). Accidentes automovilísticos fatales en la Zona Metropolitana de la Ciudad de México: una perspectiva en el espacio y en el tiempo. Retrieved from http://www.redalyc.org/articulo.oa?id=11243018009

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). *Minería de datos: Conceptos y tendencias*. *Inteligencia Artificial* (Vol. 10). Asociación Española para la Inteligencia Artificial (AEPIA). Retrieved from https://idus.us.es/xmlui/handle/11441/43290

Rodriguez, Y., & Díaz, A. (2009). *Herramientas de Mineria de Datos*. *Revista Cubana de Ciencias Informáticas* (Vol. 3). [Universidad de las Ciencias Informatica]. Retrieved from http://www.redalyc.org/html/3783/378343637009/

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in …. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Saini, I. (2013). QRS detection using K -Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research*, *4*(4), 331–344. https://doi.org/10.1016/j.jare.2012.05.007

Science, M. T. C., Sakhare, A. V, & Science, C. (2017). A Review On Road Accident Data Analysis Using Data Mining Techniques, 2–6.

Shiau, Y.-R., Tsai, C.-H., Hung, Y.-H., & Kuo, Y.-T. (2015). The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents. *Mathematical Problems in Engineering*, *2015*, 1–8. https://doi.org/10.1155/2015/170635

Shin, H., Yoo, K.-H., & Nasridinov, A. (2016). An Accident Prediction in Military Barracks Using Data Mining (pp. 683–688). Springer, Singapore. https://doi.org/10.1007/978-981-10-1536-6_89

Taamneh, M., Alkheder, S., & Taamneh, S. (2017). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety and Security*, *9*(2), 146–166. https://doi.org/10.1080/19439962.2016.1152338

Venkat. (2017). Classification Series 5 – K-Nearest Neighbors (knn) – Data Science and Analytics. Retrieved October 23, 2018, from https://dslytics.wordpress.com/2017/11/16/classification-series-5-k-nearest-neighbors-knn/

Wang, J., Gu, Q., Wu, J., Liu, G., & Xiong, Z. (2016). Traffic Speed Prediction and Congestion Source Exploration:

A Deep Learning Method. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 499–508). IEEE. https://doi.org/10.1109/ICDM.2016.0061

Wieringa, R. (2010). *Design science methodology*. *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - ICSE '10* (Vol. 2). https://doi.org/10.1145/1810295.1810446

Yadav, A. K., Malik, H., & Chandel, S. S. (2015). Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India. *Renewable and Sustainable Energy Reviews*, *52*, 1093–1106. https://doi.org/10.1016/j.rser.2015.07.156

You, J., Wang, J., & Guo, J. (2017). Real-time crash prediction on freeways using data mining and emerging techniques. *Journal of Modern Transportation*, *25*(2), 116–123. https://doi.org/10.1007/s40534-017-0129-7

Zhexue, H. (2011). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *1*(3), 284–290. https://doi.org/10.3923/ajbmb.2011.284.290