Modelo computacional de minería de datos, para un sistema de información geográfica de monitoreo de vehículos, que permita la predicción de eventos peligrosos

Andrés Felipe Guerrero Ramírez

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá D.C., Colombia

2014

Modelo computacional de minería de datos, para un Sistema de Información Geográfico de monitoreo de vehículos, que permita la predicción de eventos peligrosos

Andrés Felipe Guerrero Ramírez

Trabajo Final de Maestría presentado como requisito parcial para optar al título de:

MAGISTER EN INGENIERÍA - INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director:

Msc. Javier Francisco López Parra

Línea de Investigación:

Ciencias de la Información Geográfica e Inteligencia Artificial

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá D.C., Colombia

2014

Dedicatoria

A mi familia, que es el motor de mi vida y que sin su apoyo incondicional, nada de esto sería posible.

Agradecimientos

A la Universidad Nacional de Colombia y su plantel de profesores, por brindarme el conocimiento necesario para el desarrollo de este proyecto. A la empresa **Id Company S.A.**, por su apoyo económico y por el acceso a los datos de prueba del sistema de información geográfico **IdMaps**, que fueron parte fundamental en el desarrollo de este trabajo. A mi director de tesis, profesor Javier Francisco López Parra, por guiarme, corregirme, ayudarme en la generación de ideas y solución de dudas.

Resumen

Durante los últimos años, el auge de las Ciencias de la Información Geográfica, ha ido en aumento de una manera considerable, debido a diversos avances tecnológicos, tanto en Hardware, como en Software. Con la implementación de estos avances en el sector transporte, surgen varías inquietudes. Entre ellas, la manera en que el uso de estas tecnologías, permitan predecir y prevenir imprevistos, riesgos y demás factores que afecten la operación de vehículos en la vía. En este trabajo, se propone un modelo computacional, que pueda ser instanciado en un sistema de información geográfico de monitoreo de vehículos, que capture información de eventos de riesgo y de esta manera, poder predecirlos, aplicando técnicas de minería de datos (algoritmos de clasificación) a partir de una muestra de datos.

Palabras clave: SIG, Redes Bayesianas, Redes Neuronales Artificiales, Arboles de Decisión, clasificación, riesgos en vehículos, GPS, ciencias de la información geográfica.

Abstract

In recent years, the rise of Geographic Information Science, has been increasing in a considerable way, due to various technological advances in both Hardware and Software. With the implementation of these advances in the transport sector, several questions arise. These include how the use of these technologies to predict and prevent unforeseen, risks and other factors affecting the operation of vehicles on the road. In this paper, is proposed a computational model that can be instantiated into a geographic information system for monitoring vehicles, that capture information from event risk and

thus its prediction, using data mining techniques (classification algorithms is proposed) from a sample of data.

Keywords: GIS, Bayes Net, Artificial Neural Network**,** Decision Tree**,** classification, vehicle risks, GPS, geographic information science.

Contenido

Lista de figuras

Lista de tablas

Introducción

Actualmente, gracias a los avances tecnológicos en las ciencias de la información geográfica, es posible realizar un trazado del recorrido de los vehículos sobre las vías, almacenando los datos que genera cada uno de ellos. Dichos datos, normalmente son procesados por Sistemas de Información Geográficos (SIG o GIS por sus siglas en inglés) y permiten realizar análisis de estos datos a través de reportes, por medio de los cuales podemos observar la operación del vehículo, el comportamiento del conductor, el estado de la vía y además factores externos, que puedan ocasionar riesgos que afecten el vehículo, la carga o la integridad del conductor. ¿Pero qué pasaría, si en vez de analizar información de accidentes o situaciones de riesgo ya ocurridas, pudiéramos predecir y prevenir dichos eventos? Aquí surge la idea de hacer uso de la minería de datos espacial, para realizar un análisis de información ya existente, que permita llevar a cabo esta tarea. La minería de datos espacial (Wang Jinlin, 2008), se define como el proceso de extraer información verídica, innovadora, interesante, oculta, desconocida, potencialmente útil y por ultimo entendible de un conjunto de datos espaciales. En este caso, este conjunto de datos, contiene cartografías donde se incluyen países, departamentos, estados, provincias, municipios, localidades, barrios, vías principales y además es posible identificar con posición exacta, la ubicación en la que un vehículo presentó un evento de riesgo o accidente, asociado también a información temporal (hora, fecha, día de la semana) que podría identificar un patrón para eventos de riesgo, que podrían ocurrir a futuro. Continuando con el desarrollo de este trabajo final de maestría, es necesario identificar el contexto de este problema, que se desea resolver y cuáles son los objetivos propuestos, seguido de un marco teórico en el cual se identifican algunos conceptos básicos y se presentan algunos avances encontrados en la literatura.

1. Descripción General del Proyecto

Para el estudio de este problema y la implementación de este trabajo, se cuenta con una base de datos espacial de un Sistema de Información Geográfico (SIG), que contiene información de tracking de objetos espaciales (Vehículos) que se encuentran en constante movimiento, los cuales reportan ubicación y diversos eventos vía GPRS (3G) en tiempo real. Cada registro que es almacenado en esta base de datos, contiene datos de ubicación (Longitud y Latitud), información temporal (fecha y hora) y en los casos en que se generan eventos de riesgo, también contienen información relacionada con este (excesos de velocidad, excesos de velocidad en lluvia, botón de pánico, excesos de RPM, entre otros), además de la identificación del conductor que se encontraba en el vehículo, cuando ocurrió dicho evento.

En la industria Colombiana, existen muchos Sistemas de Información Geográficos (SIG por sus siglas en español), que permiten la captura de tracking y eventos en vehículos de carga, sin embargo, no se identificó en la literatura ninguno que permita de realizar predicción de dichos eventos. Para este trabajo, el objetivo es atacar este punto en particular. Se desea entonces, desarrollar un modelo computacional basado en minería de datos que permita realizar predicción de eventos que afecten la operación de los vehículos, llevando a cabo un proceso de minería de datos espacial, haciendo uso de algoritmos de clasificación (árboles de decisión, bayes, redes neuronales artificiales). Con esto, se espera identificar los sitios geográficos en donde la probabilidad de ocurrencia de los eventos de riesgo es más alta y las condiciones en que estos pueden ocurrir.

Como entregables del proyecto, se entregará un prototipo de un sistema que permita realizar la predicción de eventos y que además esté integrado con el Sistema de Información Geográfico que hace uso de la base de datos espacial con la que se va a desarrollar este trabajo.

1.1 Objetivo General

Desarrollar un modelo computacional de minería de datos para un sistema de información geográfico de monitoreo de vehículos, que permita la predicción de ocurrencia de situaciones de riesgo, relacionados con posiciones geográficas y diversas variables, haciendo uso de algoritmos de clasificación.

1.2 Objetivos Específicos

- Aplicar un análisis descriptivo y pre-procesamiento de los datos, para definir el conjunto de datos a que será minado con los diferentes algoritmos de clasificación.
- Definir el modelo de clasificación para la predicción de riesgos, estudiando y seleccionando algoritmos de clasificación de la literatura.
- Validar el modelo definido.
- Integrar modelo generado al sistema de información geográfico.

2. Marco Teórico

En este capítulo se definen conceptos necesarios para un total entendimiento del desarrollo del trabajo, además de presentar investigaciones previas en áreas relacionadas con el tema y avances identificados en la literatura.

2.1 Marco Conceptual

A continuación se presentan conceptos básicos, que se identificaron como necesarios para la compresión del trabajo desarrollado y expuesto en el capítulo 3.

2.1.1 Bases de Datos Espaciales

Una base de datos espacial es aquella que contiene información de datos geográficos. La característica principal que destaca este tipo de datos, es que no son datos atómicos (a cada variable X, le pertenece un valor Y único), sino que por el contrario, son datos estructurados que contienen distinta información, asociada a una sola variable. Un dato espacial contiene como mínimo, información de Longitud y Latitud, que determinan ubicación en un espacio geográfico, basándose en un sistema de coordenadas geográficas; también puede contener una variable de rumbo, que determina cardinalidad (Norte, Sur, Este, Oeste) en los casos en que el dato caracteriza un objeto que esta o puede estar en movimiento; además, puede contener información temporal (fecha y hora) (Xiaofang Zhou, 2000).

2.1.2 Objetos Geográficos

Un objeto geográfico, es un ente el cual interactúa de manera extrínseca o intrínseca en un determinado espacio. Este puede incluir componentes naturales y artificiales presentes en un espacio geográfico. Su localización es siempre expresada a través de una posición en un sistema de coordenadas. Puede contener una variable de tiempo, que indica el momento en el cual el objeto interactúa con el sistema o en el cual se hizo el

levantamiento del dato. Además, cuentan con atributos, que son aquellos que lo caracterizan (área de un lote, ancho de una calle, longitud de una vía, entre otros) (Xiaofang Zhou, 2000).

Existen tres tipos básicos de objetos geográficos. Puntos, líneas y polígonos, además de otros tipos más complejos que derivan de los anteriores, como multipuntos, multi-lineas y multi-poligonos. Un objeto Punto, contiene como mínimo un sistema de coordenadas para su ubicación geográfica, que poder ser representadas con un valor de Longitud y de Latitud. El objeto línea es una unión continua de puntos y a su vez, un objeto polígono es una unión continua de líneas que encierran un área.

Para los objetos geográficos derivados (multi-puntos, multi-líneas y multi-polígonos) son conjuntos de objetos básicos que no se encuentran obligatoriamente distribuidos de manera continua y que no necesariamente se intersectan.

2.1.3 SIG en Transporte

Un SIG en transporte, puede ser definida como una herramienta que permite el procesamiento de datos capturados de un vehículo o una flota de vehículos en tiempo real, usando un dispositivo de localización automática de vehículos (AVL, Automated Vehicle Location) (O. Aloquili, 2008). El AVL, permite a una empresa o compañía, coordinar el movimiento de una flota de vehículos y se utiliza con el propósito de conocer su localización constantemente y en tiempo real. Existen diversos medios, por medio de los cuales un dispositivo AVL transmite los datos a un SIG, que pueden ser usando tecnología celular (SMS), red móvil (GPRS, 3G, 3.5G, 4G, CDMA) o a través de un modem satelital (O. Aloquili, 2008).

Un sistema de monitoreo de vehículos, tiene características que hacen que el software sea más confiable. No únicamente se captura el posicionamiento de los vehículos y/o rutas recorridas, también se puede capturar otra información (dependiendo del dispositivo AVL) entre las cuales se encuentran las siguientes:

- Alarmas de velocidad
- Alarmas de temperatura del motor
- Consumo de combustible

- Tiempo de viaje en un recorrido
- Frenadas bruscas
- Condiciones de trafico
- Estado de la vía

La cantidad de características y datos puede llegar a ser muy voluminosa, pero hay que recordar que dicha información, depende de la capacidad que tenga dispositivo AVL, para capturarla.

A partir de la literatura y trabajo con un SIG, se han podido determinar ciertos eventos que podrían ser capturados por un dispositivo AVL, que son considerados riesgosos, tanto para la integridad del vehículo y sus ocupantes, como para la carga del vehículo, según sea el caso. En la sección 3.1 de este trabajo, se identifican y describen los eventos de riesgo.

2.1.4 Minería de Datos Espacial

La minería de datos es un proceso iterativo en el que el progreso se define por el descubrimiento, a través de cualquier método automático o manual. Es más útil en un escenario de análisis exploratorio en el que no hay nociones predeterminadas sobre lo que constituye un resultado "interesante". La minería de datos espacial se define como la búsqueda de nueva información, valiosa, y no trivial en grandes volúmenes de datos que contienen información espacial. Se trata de un esfuerzo cooperativo de los seres humanos y las computadoras. Los mejores resultados se logran mediante el equilibrio de los conocimientos de los expertos en los problemas descritos y metas con las capacidades de búsqueda de los computadores (Wang Jinlin, 2008) (Wang Peng, 2009) (Kantardzic, 2003).

2.1.5 Integración de la Minería de Datos Espacial con un SIG

Un sistema de minería de datos espacial puede ser integrado con un SIG, debido a la composición de sus estructuras de datos, a su habilidad única de análisis de información geográfica, su capacidad de hacer búsquedas espaciales de una manera eficiente y su complejo motor de consulta. Se debe tener en cuenta que para la consulta de dicha

información, se requieren funciones especiales del SIG que no pueden ser implementadas por métodos comunes de consulta (Soyoung Jung, 2014).

El pre-procesamiento de datos espaciales para el descubrimiento de conocimiento debe realizarse siempre, debido a que la gran mayoría de estos datos tienen información faltante, mucho ruido e inconsistencias, además, la extracción de información sobre los datos originales y sin procesar, podrían generar costos computacionales muy altos, produciendo errores en la información encontrada o simplemente no obteniendo ningún resultado, debido a incapacidad del CPU (Soyoung Jung, 2014).

Es necesario estudiar trabajos previos realizados en esta área. Por esta razón se continúa con la sección de antecedentes, en donde se exponen algunos trabajos representativos que han sido desarrollados en diversas partes del mundo.

2.2 Marco Contextual

A continuación se presentan algunas experiencias que se han desarrollado a nivel internacional sobre la prevención de eventos de riesgo en vehículo, que sirven para conocer el aporte que hace este trabajo sobre el tema en general. Se describen metodologías usadas y resultados obtenidos.

2.2.1 Descripción global de estudios previos con respecto a los eventos de riesgo

El autor Durduran, en el año 2010 a partir de estudios realizados y publicados a nivel internacional pudo determinar que la lluvia, es uno de los principales factores externos que pueden afectar la normal operación de un vehículo y puede ocasionar un accidente (Durduran, 2010). En condiciones de lluvia, la fricción de las llantas es menor que en condiciones de tiempo seco, la visibilidad disminuye y el conductor pierde microsegundos de reacción en caso de una eventualidad. Dicho estudio, examinó el impacto que tiene la lluvia en el área metropolitana de Melbourne – Australia desde 1897 hasta 2002 y se llegó a la conclusión que la presencia de precipitaciones, representa un riesgo en la conducción y se deben tener altas precauciones bajo estas condiciones. Pero no solo la

lluvia es un factor de riesgo a tener en cuenta en la conducción. En grandes autopistas, donde se pueden alcanzar altas velocidades, el estado de la vía, las curvas en suelo plano o en inclinaciones, deben ser tomadas en consideración como un punto importante para la ocurrencia de accidentes. En la autopista inter-estatal de Wisconsin – U.S.A., el factor más sobresaliente, además de las condiciones de lluvia, son las curvas existentes, tanto horizontales como verticales. Según información proporcionada por el WisDOT (Wisconsin Department of Transportation) 628 accidentes de auto que involucran al menos dos vehículos, ocurrieron en curvas de los cuales 196 involucraron heridos o víctimas fatales entre 1999 y 2005. Con suelo mojado, la fricción de las llantas contra el asfalto disminuye y aunque un vehículo que transporta carga pesada, frene con antelación, es muy probable que la viscosidad del suelo, provoque que este se deslice y ocasione accidentes con los vehículos que se encuentran en frente de sí. El tipo de suelo (estado de la vía) es también en cierto grado, un factor importante en el momento de analizar la ocurrencia de accidentes vehiculares.

La falta de atención y las distracciones durante la conducción son causantes de una gran cantidad de accidentes en la vía. El “National Highway Traffic Safety Administration” (NHTSA) y el “Virginia Tech Transportation Institute” (VTTI) reportaron que el 80% de las accidentes y el 65% de los que casi fueron accidentes, son provocados por la falta de atención del conductor justo antes de la colisión (3 segundos aproximadamente). En Taiwán los resultados estadísticos muestran que la mayoría de las accidentes de tráfico, son provocadas por un mal comportamiento de los conductores y la principal causa, es no prestar una debida atención al camino (Markus Deubleina, 2013).

En Colombia, según la Corporación Fondo de Prevención Vial (FPV Colombia), las cifras de accidentalidad no son muy alentadoras. Tan solo en el departamento de Cundinamarca, Departamento del cual hace parte la capital del país (Bogotá), las cifras de accidentalidad en el año 2012 superan los 3200 heridos y 480 víctimas fatales (estos son datos consolidados, no consideran estado de las vías, factores climáticos, ni sociales).

A los casos anteriormente descritos, no se involucran factores de parte del conductor. De su condición física, psicológica o laboral. (Markus Deubleina, 2013) Un conductor debe ser contratado, únicamente para conducir y no para ejecutar otras tareas que puedan distraerlo mientras opera un vehículo, a menos que estas tareas tengan un efecto positivo sobre la conducción, por ejemplo, si se conduce sobre una vía monótona con baja estimulación. Aquí se debe considerar el estado físico del conductor. ¿Ha

descansado lo suficiente? ¿Lo afecta algún tipo de molestia muscular o enfermedad? ¿Sus sentidos se encuentran funcionando de manera correcta? (Markus Deubleina, 2013) Los conductores se acostumbran a las acciones repetitivas que llevan a cabo en la operación y muchas veces no identifican de manera inmediata una situación de riesgo. A menudo ignoran cualquier señal de riesgo. Sin embargo el riesgo no solo se ve representado en clima, estado de la vía o condiciones del conductor, existen también factores sociales y zonas de alto riesgo en la vía que son vulnerables por altos índices de delincuencia y que pueden afectar al vehículo y la integridad del conductor y/o pasajeros.
Planteados los antecedentes y contextualizando la problemática a tratar, se presentarán avances y soluciones relacionados y así de esta manera, proponer un trabajo futuro basados en el tema.
A continuación se presentan algunas de las soluciones identificadas, que realizan un trabajo claro con respecto al problema y que hacen uso de herramientas de Minería de Datos y/o Inteligencia Artificial.

2.2.2 Identificación de Variables y Factores de Riesgo

Según Durduran, para la estimación de intensidad de las accidentes, estas se clasificaron en tres categorías independientes, con base en el tamaño de una muestra de datos para el análisis estadístico y las características de gravedad de cada evento. Sin heridos incapacitados (tipo B), con heridos incapacitados (tipo A) y accidentes mortales (tipo K) que incluyen heridas visibles que representan la gravedad más alta de las accidentes. Existe otra categoría que incluyen (tipo C) heridas invisibles y representan el segundo grado de gravedad de las accidentes. Además se incluye una categoría más en la que no existen heridos (PDO), los daños son únicamente materiales y esta representa el menor grado de gravedad en las accidentes (Durduran, 2010).
A continuación se listan algunas de las variables que influyen en la generación de accidentes, tomadas en una muestra de datos (Durduran, 2010).

- Sexo del conductor
- Conductor bajo efecto de drogas o alcohol
- Edad del conductor
- Acciones del conductor previos al accidente

- Tipo de vehículo
- Manera de la colisión
- Material de la superficie de la vía
- Condiciones de iluminación
- Temperatura (°C)
- Velocidad del viento
- Intensidad de lluvia

Haciendo uso de una muestra de datos de los accidentes ocurridos en la autopista interestatal de Wisconsin, se realizó un conteo, que permite estimar la frecuencia con la que ocurren, sabiendo que este valor es un número entero discreto y no negativo. La muestra de datos usada en este caso de estudio, no presenta muchos valores nulos o ceros en segmentos de una milla, pero si muestra datos muy dispersos, geográficamente hablando. Por esta razón se llevó a cabo una regresión binomial negativa para estimar la frecuencia con la cual se presentan accidentes entre más de un vehículo en esta zona de U.S.A. (Durduran, 2010).

2.2.3 Minería de Datos Espacial Aplicada en un sistema de análisis y prevención de riesgos

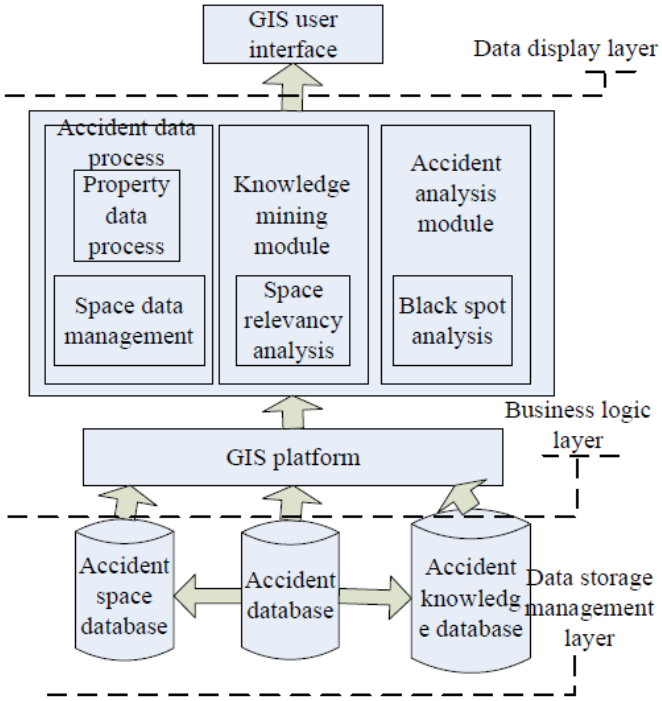
Los riesgos a los que está expuesto un conductor y un vehículo en la vía, es una problemática a nivel mundial y que requiere ser solucionada lo antes posible. Haciendo uso de la minería de datos, sobre los datos de un SIG, es posible que las autoridades de una ciudad, provincia o país, identifiquen aquellas zonas de alto riesgo para los conductores y en consecuencia se tomen medidas preventivas y NO correctivas. Haciendo uso del framework de desarrollo ArcGIS Engine y el lenguaje de programación C#, fue desarrollado un sistema de análisis de accidentes de tráfico que proporciona diferentes funciones (Wang Peng, 2009).

El diseño de este sistema, se basó en mapas y administró la información de accidentes de tráfico de una región determinada en una ciudad y combina la gestión de datos espaciales con la de datos transaccionales para mejorar la capacidad de gestión del sistema. Además de gestionar la información espacial y transaccional de la región, el sistema está en la capacidad de buscar e identificar la normatividad y leyes ocultas de una zona geográfica seleccionada, a través de un módulo de minería de datos espacial,

para que de esta manera las autoridades, puedan tomar decisiones de una forma más acertada y oportuna. Este sistema pudo ser alimentado manualmente con registros de accidentes y este, de manera interactiva, puede sugerir ubicaciones en el mapa, en donde existen probabilidades de generarse nuevos accidentes, con base en la información ya almacenada en su base de datos. El diagrama del diseño de la estructura del sistema, puede verse en la gráfica que se muestra en la Figura 1 (Wang Peng, 2009).

Haciendo uso de algoritmos de asociación de minería de datos, este sistema es capaz de generar reglas, con base en accidentes ocurridos y así identificar las relaciones entre cada variable y/o factor que pueda generar los accidentes. Una vez las reglas han sido establecidas, el sistema se encarga de establecer bajo que ítems y/o circunstancias se presentan la mayoría de los accidentes y poder generar alarmas de advertencia a los usuarios del sistema, con respecto a los vehículos que se encuentren sobre la vía.

Figura 1: (Wang Peng, 2009) Diagrama del diseño de la estructura del sistema



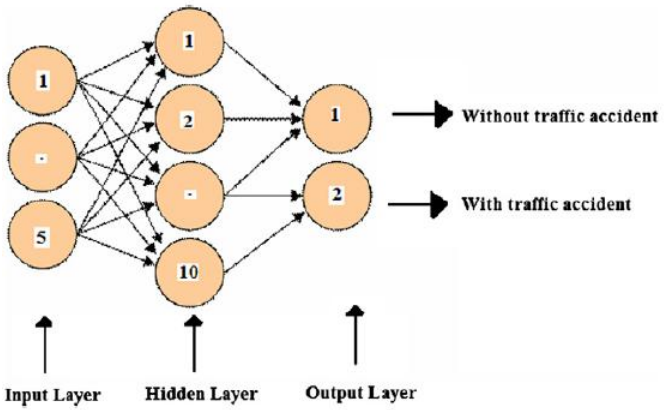
2.2.4 Uso de Redes Neuronales Artificiales para prevención de eventos de riesgo

En la actualidad, los SIG, se han convertido en la herramienta más importante para el análisis de accidentes de tráfico en las autopistas. Estos sistemas, tienen la habilidad de mantener y almacenar un volumen de datos considerable, que puede ser fácilmente compartido, analizado y administrado. Un SIG provee (o debe proveer) una plataforma para analizar y visualizar datos espaciales y explorar las relaciones de datos espaciales con datos no-espaciales (Yang-Kun Ou, 2013).

Un sistema fue implementado para la predicción de accidentes de tráfico en la ruta que comunica a Konya con Afyonkarahisar en Turquía (279 Kilómetros aproximadamente) usando cinco características de los accidentes documentados: día, temperatura, humedad, condiciones del clima y mes del año en el que ocurrió cada accidente y comprende 378 puntos críticos (179 sitios en donde ocurrieron accidentes de tráfico y 179 donde no). Es de considerar, que el mes del año es muy influyente en la ocurrencia de un accidente, puesto que las condiciones climáticas en Turquía, dependen de las estaciones del año (Yang-Kun Ou, 2013).

Fue implementada una red neuronal artificial que a su vez utiliza el algoritmo Levenberg Marquart con el fin de predecir la ocurrencia de los accidentes en esta autopista. En la Figura 2, se puede observar la estructura de la red (Yang-Kun Ou, 2013).

Figura 2: (Yang-Kun Ou, 2013) Estructura de la Red Neuronal con LM



Una vez entrenada la red, los resultados obtenidos no fueron tan satisfactorios como se esperaban, pues una vez se valida la red con el repositorio de datos obtenido, no se alcanzó ni siquiera 60% de clasificación correcta de las clases. En la Figura 3, se

muestran los resultados de la clasificación, usando redes neuronales y Maquinas de Soporte Vectorial.

Figura 3: (Yang-Kun Ou, 2013) Resultados de la clasificación

Classifier algorithm	Kernel type	Classification accuracy (%)	Sensitivity (%)	Specificity (%)
Artificial neural network	–	**53.93**[a]	**52.67**[a]	**57.44**[a]
Support vector machine	RBF kernel function	52.25	51.39	55.88
	Polynomial kernel function	37.64	37.78	37.50
	Linear kernel function	48.31	0.0	49.14

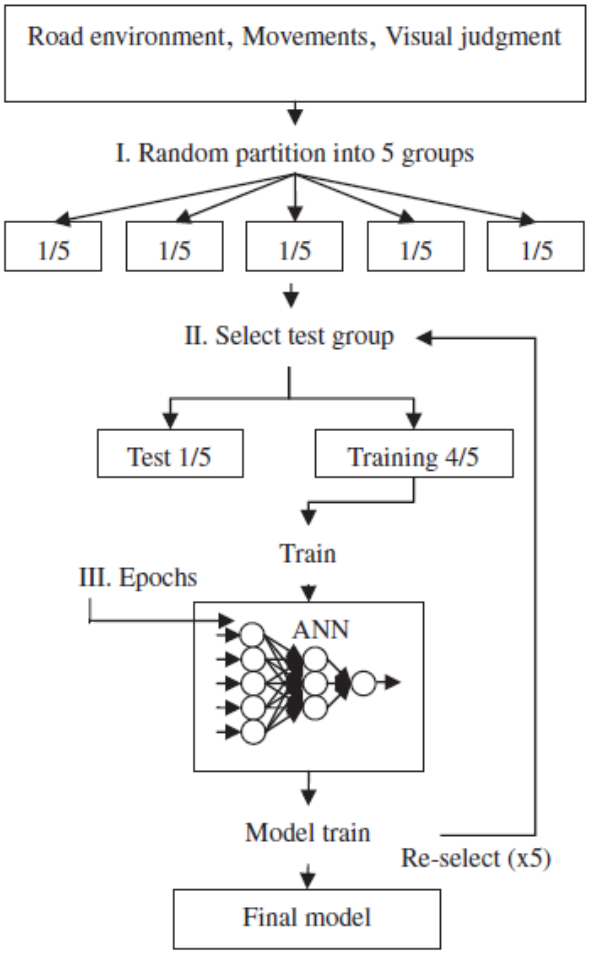
Considerando el comportamiento de los conductores, fue posible determinar que la mayoría de los accidentes de tránsito son provocados por su falta de atención y basados en esto, se podría modelar una red neuronal que permita relacionar dichos factores. Dado el efectivo resultado de la implementación de una red neuronal artificial para construir modelos predictivos, grandes cantidades de investigaciones han decidido hacer uso de esta herramienta para predecir comportamientos de conductores y accidentes de tráfico, pero en este caso, se usó para predecir dichos riesgos, cuando el conductor lleva a cabo también actividades secundarias durante la operación del vehículo (Yang-Kun Ou, 2013).

Para el desarrollo de esta actividad, se tomaron datos de 66 conductores voluntarios. Sin embargo, el estudio no se hizo sobre vehículos reales. Fue realizado sobre un simulador, que capturaba los datos de la operación en tiempo real. El simulador crea un ambiente de tráfico normal, con normas de tránsito y congestión vehicular. Entre los datos capturados, se tienen en cuenta los movimientos del conductor dentro de la cabina, tiempos de reacción, número de movimientos, la constancia visual al camino que se está recorriendo y además la magnitud de la velocidad. La captura de dicha información, se hace tres veces por segundo (Yang-Kun Ou, 2013).

El modelo de la red neuronal, fue dividido en tres capas. La de entrada, las capas ocultas y la capa de salida. Las capas ocultas, determinan el peso en las conexiones entre la capa de entrada y la capa de salida. El número de conexiones entre las capas ocultas, fue determinado a través de la experimentación. La ilustración de los procesos de

entrenamiento y análisis de la red neuronal con propagación hacia atrás, se puede observar en la Figura 4 (Yang-Kun Ou, 2013).

Figura 4: (Yang-Kun Ou, 2013) Ilustración de Entrenamiento y Análisis de la RN



El modelo propuesto, tuvo una precisión (*accuracy*) del 66% en donde el 93% de las predicciones realizadas, fueron sobre comportamientos seguros en la conducción y el 72% de comportamientos cautelosos.

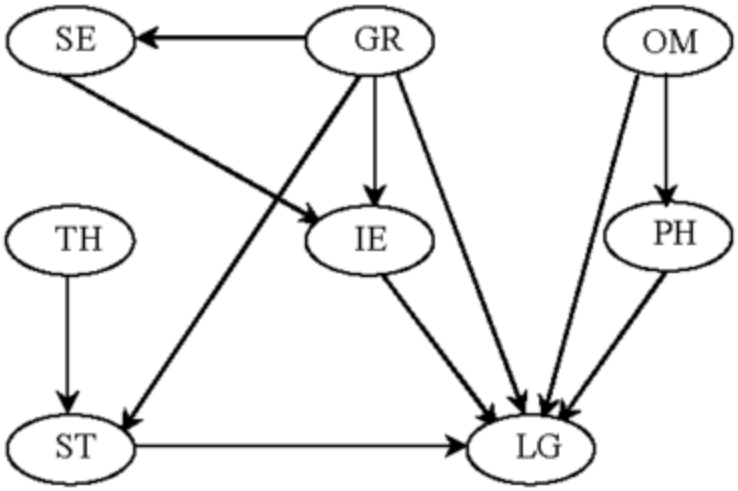
2.2.5 Uso de Algoritmos Bayesianos

En un trabajo realizado en el 2013 se propuso una metodología para determinar modelos que puedan ser usados para prevenir el número de accidentes, que puedan provocar heridos, tanto al conductor como a demás usuarios de la vía, en donde no existe mucha información por cada segmento de la vía en cuestión. La utilidad de la metodología propuesta fue demostrada en un caso de estudio, sobre la red vial de Austria. Las vías,

fueron segmentadas en tramos homogéneos y el modelo se formuló en términos de riesgos que indican variables que son utilizadas por una red bayesiana probabilística. Las redes fueron desarrolladas combinando análisis de regresión multi-variantes jerárquicos para la evaluación de inferencias de datos y técnicas de minería de datos modernas para la adaptación de las redes bayesianas a los datos disponibles. En el caso de estudio, el modelo reacciona a variables considerables y actualiza las tasas de ocurrencia de accidentes en donde se presentan heridos sin importar la gravedad de las heridas. Las variables que indican riesgo se seleccionan teniendo en cuenta características del tráfico y parámetros del diseño de la vía, como el volumen del tráfico, composición del tráfico, velocidad, curvas y numero de calzadas del camino. Esta metodología facilita el desarrollo de modelos precisos para predecir el número de accidentes con víctimas y usuarios de la carretera lesionados de manera diferente que pueden ocurrir en un tramo de carretera de concreto (Markus Deubleina, 2013).

Las redes bayesianas, pueden ser usadas en la minería de datos espacial, dada la capacidad de inferencia de clases, según el conjunto de datos que este procesando y además permite el razonamiento probabilístico en sistemas expertos. El uso de las redes bayesianas, no está limitado a la aplicación de la clasificación de clases, sino que también puede ser utilizada para representar gráficamente conocimiento espacial, llevar a cabo agrupación de objetos espaciales y realizar predicciones. En un trabajo se expone la aplicación de modelos que hacen uso de las redes bayesianas en la minería de datos espacial en un sistema que contiene datos de la gradación de terrenos agrícolas. El conjunto de datos, contiene 2160 registros de evaluación de recursos de los terrenos, donde el problema contiene 8 variables, que a su vez contienen gran cantidad de atributos. Una vez se definió el dominio de las variables y la preparación de los datos, se obtuvo la estructura del modelo de la red bayesiana, utilizando el modelo de construcción de los tres pasos (Jiejun Huang, 2007).

Figura 5: (Jiejun Huang, 2007) Modelo de Red Bayesiana para la gradación de terrenos agrícolas.



Fueron utilizados 200 casos del conjunto de datos, para la validación del modelo y se exponen los resultados, en donde se compara el uso de las redes bayesianas, contra el algoritmo Naive Bayes, obteniendo como resultados, una precisión del 89% del modelo de red bayesiana, contra un 78% del algoritmo Naive Bayes, en la clasificación.

Figura 6: (Jiejun Huang, 2007) Redes Bayesianas Vs. Naive Bayes

method	Testing set	correct	Evaluation accuracy
Bayesian Network	200	178	89.0%
Naive Bayes	200	157	78.5%

Los resultados mostrados, muestra la viabilidad que tiene el uso de modelos de redes bayesianas en la minería de datos espacial y deja abierto el tema de investigación, del uso de estos modelos, en otro tipo de conjunto de datos espaciales, como por ejemplo el propuesto en esta tesis de maestría, en donde se propone uno para la predicción de eventos de riesgo en vehículos.

3. Desarrollo

En los capítulos anteriores se desarrolló la etapa de presentación de conceptos necesarios para comprender el modelo y posteriormente se presentaron en el marco contextual algunas experiencias relacionadas con temas similares.

En el presente capítulo se expone la manera como se llegó al modelo computacional siguiendo una metodología. La metodología para producir este modelo se presenta en la figura 7. Está fundamentada en las siguientes fases:

Fase de Exploración e identificación de variables

- Identificación de Eventos de Riesgo
- Identificación de variables que influyen en la ocurrencia de eventos de riesgo.
- Diseño de modelo de bodega de datos.

Fase de preparación de los datos

- Desarrollo de ETL
- Análisis descriptivo
- Pre-procesamiento de datos
- Muestreo de datos

Fase de prueba de algoritmos de clasificación

- Aplicación de algoritmos de clasificación
- Algoritmos de Redes Bayesianas
- Algoritmos de Redes Neuronales Artificiales
- Algoritmos de Arboles de Decisión
- Comparación de modelos predictivos
- Selección de modelo predictivo

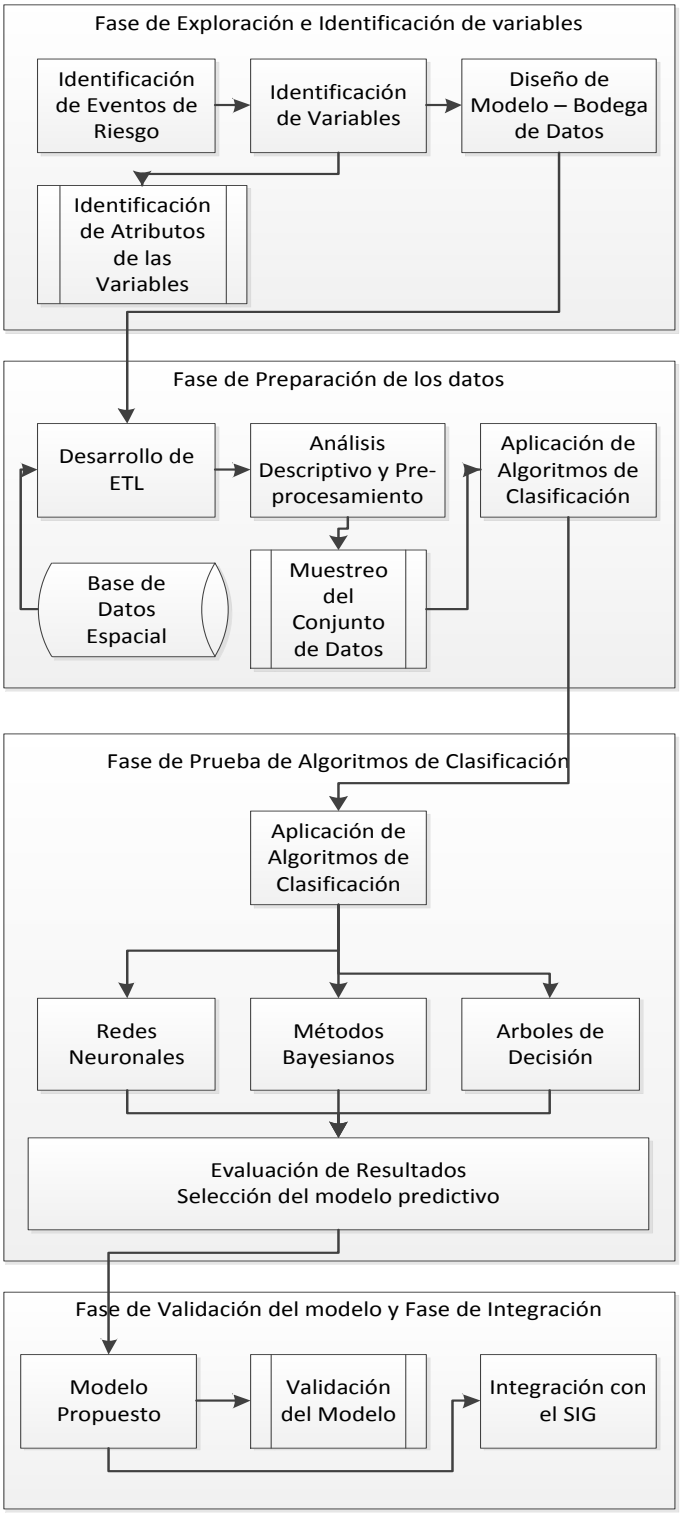
Fase de validación del Modelo

- Validación del modelo seleccionado

Fase de Integración

- Integración con el SIG

Figura 7: Metodología



3.1 Identificación de Eventos de Riesgo

Teniendo en cuenta la capacidad de un dispositivo AVL y un SIG, algunos de los eventos de riesgo que pueden ser capturados y procesados en un software de este tipo, se describen a continuación y se justifica por qué fueron seleccionados.

3.1.1 Exceso de velocidad

Como es conocido, en cada país, estado o ciudad, existen unos límites de velocidad máxima establecidos, ya sea por tramos en la vía o por sectores de una ciudad o barrio. En algunos casos, estos parámetros de seguridad, pueden ser establecidos por zonas en el SIG o directamente en el dispositivo AVL y el software debe estar en capacidad de detectar, si la velocidad registrada por un vehículo, sobrepasa dichos parámetros. Superar los límites de velocidad establecidos en una zona determinada, puede desencadenar un accidente, considerando así este exceso, como un evento de alto riesgo.

3.1.2 Exceso de velocidad en condiciones de lluvia

En la sección 2.2.1 de este trabajo, se exponen los riesgos que existen en la conducción en condiciones de lluvia, teniendo en cuenta que la fricción de las llantas contra el suelo, disminuye considerablemente, por la capa de agua que forma sobre ellas. También es de considerar, que el estado de la vía cambia bajo estas condiciones y si a eso, se le suma un exceso de velocidad del vehículo, las probabilidades de que se ocasione un accidente aumenta de manera considerable. Por esta razón, cuando el SIG o el AVL detectan un evento de estas características, es considerado riesgoso.

3.1.3 Exceso RPM (Revoluciones por Minuto)

Este evento de riesgo, afecta principalmente la integridad del vehículo, pues refleja un mal uso del motor, que con el paso del tiempo puede sufrir desgaste significativo y generar daños que podrían ser irreparables o costar mucho dinero. No se considera como un evento de alto riesgo, pero si es necesario tenerlo en cuenta.

3.1.4 Aceleración Brusca

Cuando un vehículo, acelera bruscamente, puede ocasionar un accidente, impactando a un vehículo o un peatón, además de afectar la integridad de sí mismo, ya que puede generar desajustes en la carrocería o el motor. Depende completamente de la manipulación del vehículo por parte del conductor y puede afectar tanto la integridad del vehículo como la del operador. Para la detección de este evento de riesgo, es necesario que el dispositivo AVL, cuente con un acelerómetro, que permita su captura.

3.1.5 Frenadas Bruscas

Las reducciones bruscas de velocidad son un evento de riesgo, que debe evitarse en lo posible durante la conducción. Puede provocar que otro vehículo impacte la parte posterior del vehículo y es consecuencia de excesos de velocidad o de no mantener la distancia con otro vehículo. Por transportadoras, es considerado un evento de alto riesgo y es consecuencia de una mala operación del vehículo por parte del conductor. Igual que el evento de Aceleración Brusca, es necesario que el dispositivo AVL, cuente con un acelerómetro o sensor de detección de gravedades para poder ser capturado y reportado.

3.1.6 Aceleraciones Laterales

Este evento, que depende de la capacidad del dispositivo AVL para ser detectado, se mide normalmente en gravedades y es generado cuando un vehículo toma una curva en la vía, en exceso de velocidad. Cuando se presentan aceleraciones laterales muy altas, existe una alta probabilidad de volcamiento del vehículo, que puede ocasionar víctimas fatales y/o afectación tanto al vehículo como a la carga que pueda transportar. Es considerado como un evento de muy alto riesgo.

3.1.7 Paradas NO Autorizadas

En un SIG de monitoreo de vehículos, es posible geo-referenciar sitios en donde un vehículo puede detenerse. Lo anterior puede ser por medidas de seguridad del conductor o por protección de la integridad del vehículo y su carga. Un dispositivo AVL, junto con el

SIG, debe estar en capacidad de detectar cuando un vehículo se detiene en un lugar no autorizado y considerarlo como un evento de riesgo para la operación. La detección de este evento permite evitar actos criminales y por esta razón, también es considerado como riesgoso.

3.1.8 Violación de la Carga

Cuando se pone en riesgo la carga de un vehículo, se pierde en si el propósito del recorrido. Algunos dispositivos AVL, permiten la detección de la apertura de compuertas o compartimientos de carga en un vehículo y es posible determinar si dicha manipulación, fue o no en un lugar autorizado, verificando su posición geográfica. Ya que los eventos de riesgo, no solo involucran al conductor o pasajeros, sino también, integridad del vehículo y su carga, es necesario que este evento, sea tenido en cuenta.

3.1.9 Exceso de Temperatura del Motor

Algunos dispositivos AVL, permiten la captura de información desde el módulo electrónico del motor de un vehículo y de esta manera, conocer cuál es la temperatura del motor en tiempo real. De esta manera, es posible detectar si durante un recorrido de un vehículo, la temperatura del motor se eleva a condiciones que puedan afectar la integridad del vehículo, pasajeros o carga, puesto que una situación de estas puede desembocar en un incendio y tragedia.

3.2 Identificación de Variables

En esta sección, se describen cada una de las variables que han sido identificadas, que influyen en la ocurrencia de los eventos de riesgo anteriormente descritos. Algunas de las variables han sido tomadas, basándose en la literatura y antecedentes que fueron descritos en el Marco Contextual (sección 2.2) de este trabajo. Otros han sido propuestos por el autor de este trabajo de grado.

3.2.1 Conductor

El comportamiento de un conductor en la vía, es uno de los factores más importantes en el momento de evaluar porque se produjo un evento que puso en riesgo el vehículo, pasajeros o la carga. De parte del conductor de un vehículo, son muchos los factores que

deben ser tenidos en cuenta y a continuación se proponen cuáles son los atributos que deberían tenerse en cuenta para esta variable, en la generación del modelo.

3.2.1.1.1 Edad

(Durduran, 2010)La edad es tenida en cuenta como un atributo del conductor, que puede influir de manera considerable en la ocurrencia de un evento de riesgo. El comportamiento de una persona madura (en un rango de 30 a 40 años), es diferente al de una persona joven (de 18 a 29 años) no solo en la vía, sino también en los social y personal. Según trabajos previos, algunos eventos de exceso de velocidad, son más frecuentes entre personas jóvenes y menos frecuentes en personas maduras.

3.2.1.1.2 Sexo

(Durduran, 2010) El sexo del conductor, es un atributo determinable en el momento de la conducción de un vehículo. Los comportamientos entre hombres o mujeres en la operación de un vehículo pueden variar y generar un evento de riesgo bajo diferentes circunstancias.

3.2.1.1.3 Actividades Secundarias

(Durduran, 2010) En estudios previos, se muestra la posibilidad de que se genere un evento de riesgo, si el conductor realiza algún otro tipo de actividad mientras opera un vehículo. Por eso, este atributo debe ser considerado para el estudio y se debe manejar como un valor *booleano*, que indica TRUE si el conductor llevaba a cabo una actividad secundaria o FALSE en el caso contrario.

3.2.2 Tiempo

Cuando se diseña una bodega de datos, una de las dimensiones que no debe faltar, es la de Tiempo, en donde se pueden tomar como atributos los días de la semana, horas del día, días del mes, el mes u otras caracterizaciones del tiempo, según se requieran en el modelo de negocio. En este trabajo, se propone el uso de esta variable para estudiar, bajo qué circunstancias del tiempo, existan probabilidades de ocurrencia de un evento de riesgo. A continuación se describen los atributos propuestos por el autor para esta variable.

3.2.2.1.1 Mes del año

Se quiere evaluar, si el mes del año infiere en la ocurrencia de un evento de riesgo en la conducción. En países que tienen estaciones de clima y las temperaturas varían considerablemente, es posible que el mes del año tenga una importancia con respecto a la manera en que se opera un vehículo, sin embargo en zonas del trópico puede verse influenciado también en este aspecto.

3.2.2.1.2 Día del Mes

Conocer que días del mes o rango de días en el que se presentan mayor cantidad de eventos de riesgo, puede ser de gran importancia para el modelo. Posteriormente en este trabajo, se implementará un caso de estudio en donde se podrá observar cómo influye este atributo de la variable Tiempo, para la predicción de los eventos de riesgo.

3.2.2.1.3 Día de la Semana

Es posible que los días de la semana (Lunes, martes, miércoles, etc), representen un factor importante, con respecto a la ocurrencia de ciertos eventos de riesgo. Por esa razón se ve la necesidad de incluir este atributo en el estudio.

3.2.2.1.4 Día Festivo

¿Qué tanto puede influir en la ocurrencia de un evento, que el día en el que ocurre sea un día festivo en el país en el que se generó? Se considera necesario incluir este atributo en el estudio y poder responder a esta pregunta. Se debe manejar como una variable booleana, en donde TRUE representa el hecho de que SI es un día festivo y viceversa.

3.2.2.1.5 Hora del día

Las jornadas en las que un conductor opera un vehículo, son consideradas un atributo importante, con respecto a la ocurrencia de estos eventos. En algunas horas del día, las condiciones de tráfico no son tan adversas como en otras, además de las temperaturas y condiciones climáticas. Por esta razón, dependiendo de la hora del día o jornadas, el tipo de eventos y la cantidad, pueden variar considerablemente.

3.2.3 Vehículo

Las características físicas y técnicas de un vehículo, pueden inferir el comportamiento del conductor en la operación del mismo. Dado que el dispositivo AVL, se instala

directamente en el vehículo, esta variable debe considerarse para su estudio. Sus atributos pueden afectar considerablemente la ocurrencia de ciertos eventos de riesgo, dependiendo del tipo de vehículo, su carga, motor y otras que se describen a continuación.

3.2.3.1.1 Tipo de Servicio

Cada vehículo tiene un objetivo principal, que normalmente debe ser registrado ante un ente regulador (ministerio, secretaria u otros). El uso que se le da al mismo, depende de su objetivo. En el desarrollo de este trabajo, se propone la clasificación de vehículos de uso particular, de servicio público y de transporte de carga. Con el modelo que se propone en este estudio, se ve necesaria dicha clasificación, puesto que los vehículos de servicio público y de carga, son los que permanecen mayor tiempo en la vía y por esta razón, no se pueden evaluar de la misma manera que los vehículos particulares.

3.2.3.1.2 Tipo de Carrocería

Determina el peso del vehículo, así como la cantidad de carga o pasajeros que puede transportar. Así mismo, depende el tipo de carrocería influye en la velocidad máxima que puede alcanzar, dependiendo del peso que está transportando, tanto en vacío, como en cupo completo. Algunos tipos de carrocería pueden ser, sin carrocería (motos), sedan, buses, utilitarios, tracto mulas, entre otros.

3.2.3.1.3 Cilindrada del Motor

El cilindraje del motor de un vehículo, es directamente proporcional a la velocidad que puede levantar y a sus caballos de fuerza (HP por sus siglas en ingles Horsepower). Puesto que uno de los eventos de riesgo que se desea predecir, es precisamente el de exceso de velocidad, se considera pertinente incluir en el modelo este atributo del vehículo.

3.2.3.1.4 Peso

Es importante considerar el peso de un vehículo, porque así mismo se deben tener en cuenta las consecuencias de ciertos eventos de riesgo. Cuando un vehículo muy pesado, alcanza altas velocidades, detenerse puede tomar más tiempo que un vehículo liviano, pudiendo producir eventos relacionados con el acelerómetro del dispositivo AVL. En el

diseño de este modelo, se propone la inclusión de este atributo, considerándolo como un factor influyente en la ocurrencia de los eventos de riesgo que se quieren predecir.

3.2.3.1.5 Potencia (Horsepower)

Aunque este atributo está relacionado con el cilindraje del vehículo, conocer la magnitud de los caballos de fuerza, nos permitirá así mismo, conocer que tan poderoso puede ser su motor y relacionar los valores de sus velocidades, con los eventos que se hayan podido generar.

3.2.3.1.6 Modelo (año)

Teniendo este dato, se puede conocer la antigüedad del vehículo y así mismo, poder identificar qué relación puede existir entre ciertos eventos de riesgo, con el estado del vehículo y sus condiciones físicas y mecánicas.

3.2.4 Ubicación

El modelo propuesto en este trabajo, se realiza por país únicamente, esto debido a que cada país tiene sus normas, leyes y geografía distinta, además de un parque automotor distinto. Cuando se habla de la ubicación, se pueden incluir zonas geográficas grandes y pequeñas, como por ejemplo estados o provincias y al mismo tiempo ciudades o distritos, según corresponda. También es necesario incluir circunstancias climáticas o sociales de una ubicación, ya que estas pueden afectar el comportamiento de un conductor al operar el vehículo y generar diversos eventos de riesgo, según en donde se encuentre. A continuación se listan y describen cada uno de los atributos que se proponen para el modelo.

3.2.4.1.1 Estado, Provincia o Departamento

Para poder hacer la caracterización geográfica de un país, es necesario conocer su división política y conocer cada uno de los estados, departamentos o provincias que agrupan un conjunto de ciudades o distritos. Esta información comúnmente se encuentra almacenada en el SIG que procesa la información transmitida por el dispositivo AVL y dicho sistema debe estar en capacidad de identificar en que ubicación ocurre un determinado evento de riesgo. Este atributo puede influir en el modelo, dado que cada estado puede tener distinta topografía, cambiando la manera en que se opera el vehículo.

3.2.4.1.2 Ciudad

Perteneciendo a un estado o provincia, las ciudades tienen su propia topografía o condiciones viales y sociales, por eso es necesario identificar cada ciudad en la que se genera un evento de riesgo. En algunos casos, una ciudad puede ser muy grande y tener ciertas características, que no comparte con otra ciudad contigua.

3.2.4.1.3 Altura sobre el nivel del mar

Este atributo, tiene asociado algunas condiciones climáticas, en especial en países que se encuentran cercanos a la línea ecuatorial (zona tropical). Bajo dichas condiciones, los comportamientos de los conductores pueden variar y por eso es necesario tener en cuenta este atributo (numérico).

3.2.4.1.4 Clima

Este atributo hace énfasis a las condiciones de temperatura de la ubicación en la que se genera un evento de riesgo. En zonas geográficas en las que existen estaciones del clima, es necesario que el SIG, este en capacidad de determinar las condiciones de temperatura, según la época del año. Los valores que puede tomar son: Frio, templado o cálido.

3.2.4.1.5 Tipo de Zona

Se refiere específicamente a la naturaleza de la zona, en la que se generó el evento. Estas zonas pueden ser Urbanas o Rurales.

3.2.4.1.6 Temperatura Promedio

En zonas geográficas cercanas al trópico, la temperatura promedio no varía, sin embargo en aquellos países en los que se presentan estaciones climáticas, las temperaturas promedio varían y el SIG debe estar en capacidad de detectar estos cambios. Se considera importante este atributo de la ubicación, puesto que es un factor que influye en la conducción y por lo tanto en la ocurrencia de eventos de riesgo. El valor de esta temperatura, puede tomarse en Celsius o Fahrenheit, según lo requiera el estudio. En ese trabajo, se tomarán en grados Celsius.

3.2.4.1.7 Estado de la Vía

Si el SIG del que se extraen los datos de eventos generados por un vehículo, contiene información sobre el estado de la vía, este dato debe ser incluido en el modelo, debido a que si el estado de la vía no es bueno, afectaría el funcionamiento del acelerómetro y podrían generarse falsos eventos de frenadas o aceleraciones bruscas. Por el contrario, en una vía que se encuentre en buenas condiciones, podrían generarse otros eventos de riesgo como excesos de velocidad.

3.3 Diseño de Modelo – Bodega de Datos de Datos

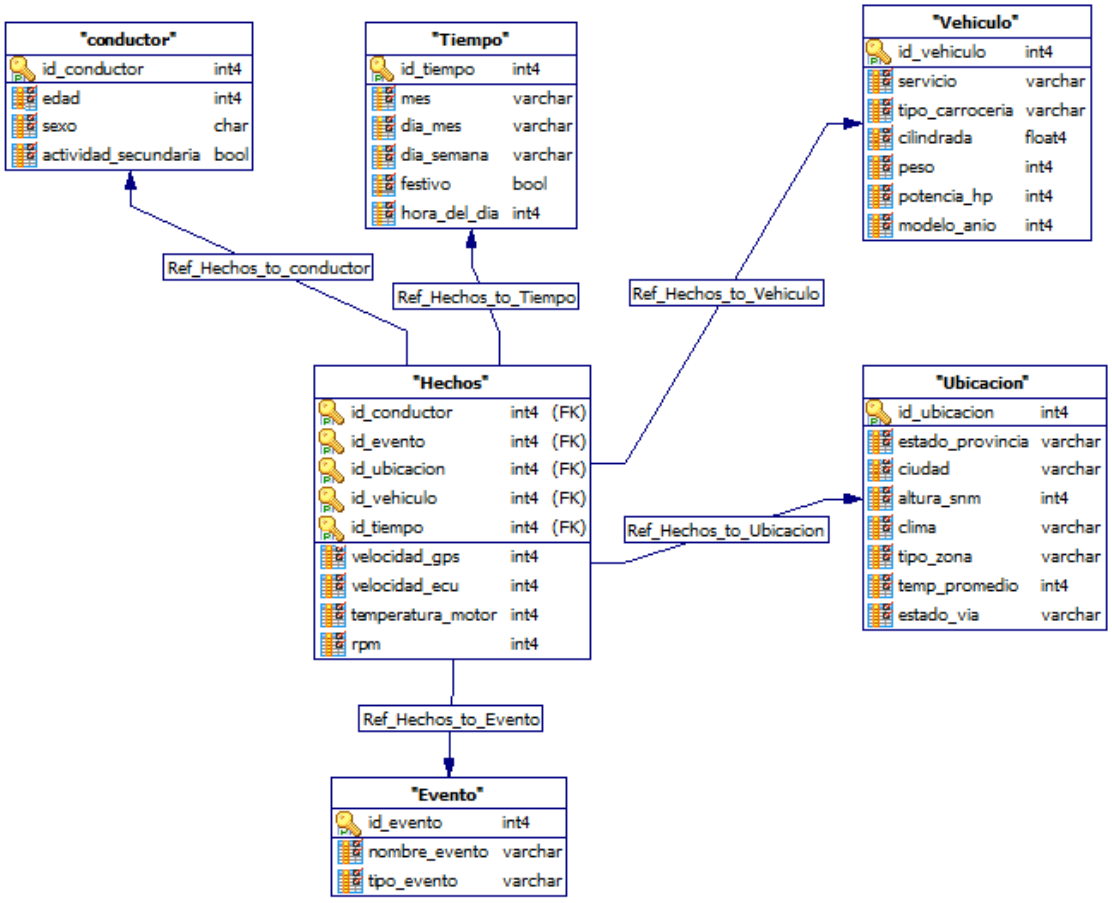
Una vez se han identificado las variables es importante encontrar la manera de integrarlas para poder ser pre-procesadas, analizadas y posteriormente minadas. Una manera para integrar información de manera que sea útil es el diseño e implementación de una bodega de datos. Una Bodega de Datos (o Data Warehouse en inglés) es un medio de almacenamiento que tiene por objetivo producir una información precisa y útil para soportar la toma de decisiones en una organización (Pocut Viqarunnisa, 2011). Para obtener información de alta calidad, es necesario que el diseño de la bodega sea lo mejor posible, puesto que la estructura de este, determina como se almacenaran los datos y como podrá ser consultada la información posteriormente.

En el desarrollo de este trabajo, no se realizó la implementación total de la bodega de datos espacial. Únicamente se plantea el diseño de esta, que servirá de base, para proponer el modelo computacional de minería de datos.

El esquema de Estrella (Star) es el patrón de modelo de datos de uso más común en el diseño de bodegas de datos. Este es un patrón de estructura orientada que representan los datos como hechos relacionados con las dimensiones (Pocut Viqarunnisa, 2011).

Para la implementación del modelo computacional de minería de datos, se propone el uso de los eventos de riesgo mencionados en la sección 3.1 y las variables descritas en la sección 3.2 de este trabajo, como las dimensiones de un esquema estrella, en donde también se incluye una tabla de hechos, que contiene valores relacionados a los eventos de riesgo. Todo esto para la integración y consolidación de los datos del SIG.

Figura 8: Bodega de Datos – Modelo Estrella



En la tabla de Hechos de este modelo estrella, se relacionan cada uno de los eventos de riesgo almacenados en la base de datos de un SIG para ser cargadas en la bodega de datos. Como se observa en la Figura 8, cada una de las dimensiones se encuentra relacionada a la tabla de Hechos y esta contiene a su vez, cuatro atributos que se describen a continuación.

3.3.1 Velocidad GPS

Es la velocidad que el dispositivo AVL captura, según el movimiento del vehículo con respecto a la triangulación de los satélites. Este valor puede estar en Kilómetros por Hora (KPH) o en Millas por Hora (MPH). Es necesario conocer este valor e identificar si el

vehículo se encontraba o no en movimiento en el momento en que se generó el evento de riesgo.

3.3.2 Velocidad ECU

Velocidad del vehículo, según la computadora o módulo electrónico del mismo. Para capturar este valor de velocidad, el dispositivo AVL debe estar en capacidad de conectarse a dicha computadora o módulo por medio de una interface, para obtener dicho valor de magnitud. Este valor puede estar en Kilómetros por Hora (KPH) o en Millas por Hora (MPH). Igual que la velocidad de GPS, es necesario conocer este valor e identificar si el vehículo se encontraba o no en movimiento en el momento en que se generó el evento de riesgo.

3.3.3 Temperatura de Motor

Valor capturado por el dispositivo AVL, que indica la temperatura del motor, en el momento en que se genera un evento de riesgo. Este valor es proporcionado por la computadora o módulo electrónico del vehículo y es necesario para saber si el motor se encontraba o no en una temperatura aceptable, en el momento en que ocurre el evento.

3.3.4 RPM

Revoluciones por Minuto (RPM por sus siglas) capturadas por el dispositivo AVL, directamente de la computadora del vehículo. Permiten determinar su valor en el momento en que se genera un evento de riesgo.

Finalmente, con el diseño de la bodega de datos, fue posible identificar todas las variables y atributos que se consideraron, tienen influencia en la ocurrencia de los eventos de riesgo, anteriormente mencionados. A continuación se listan todos los atributos del modelo de datos, indicando la variable a la que pertenecen y su tipo de dato.

Tabla 1: Atributos del Modelo

Atributo	**Variable**	**Tipo de Dato**
Mes	Tiempo	String (Cadena de Caracteres)
Día del Mes	Tiempo	Integer (Numérico Entero)

Día de la semana	Tiempo	String (Cadena de Caracteres)
Festivo	Tiempo	Boolean (Booleano)
Hora del día	Tiempo	Integer (Numérico Entero)
Edad	Conductor	Integer (Numérico Entero)
Sexo	Conductor	String (Cadena de Caracteres)
Actividad secundaria	Conductor	Boolean (Booleano)
Tipo de servicio	Vehículo	String (Cadena de Caracteres)
Tipo de carrocería	Vehículo	String (Cadena de Caracteres)
Cilindrada	Vehículo	Double (Numérico con valor decimal)
Peso	Vehículo	Double (Numérico con valor decimal)
Potencia	Vehículo	Integer (Numérico Entero)
Modelo	Vehículo	Integer (Numérico Entero)
Estado o Provincia	Ubicación	String (Cadena de Caracteres)
Ciudad	Vehículo	String (Cadena de Caracteres)
Altura Nivel del Mar	Vehículo	Double (Numérico con valor decimal)
Clima	Vehículo	String (Cadena de Caracteres)
Tipo de Zona	Vehículo	String (Cadena de Caracteres)
Temperatura Promedio	Vehículo	Double (Numérico con valor decimal)
Estado vía	Vehículo	String (Cadena de Caracteres)
Velocidad GPS	Hechos	Double (Numérico con valor decimal)
Velocidad ECU	Hechos	Double (Numérico con valor decimal)
Temperatura Motor	Hechos	Double (Numérico con valor decimal)
RPM	Hechos	Integer (Numérico Entero)
Evento	Evento	String (Cadena de Caracteres)

3.4 Desarrollo ETL

Para la implementación de la herramienta ETL (Extrac, Transform and Load) fue necesario contar un una base de datos espacial de un SIG, de la cual se pudieran extraer la mayor cantidad de datos posibles, con respecto al modelo de datos propuesto en la sección 3.3.

Para dicha implementación, se hizo uso de la base de datos del SIG ***idMaps***, perteneciente a la empresa de monitoreo y rastreo **Id Company S.A.**, una empresa colombiana y con sus oficinas principales ubicadas en la ciudad de Cali.
A continuación muestra un *checklist* realizado a esta base de datos, con respecto a los atributos que posee del modelo propuesto.

Tabla 2: Checklist idMaps

Atributo	**Está en idMaps**
Mes	X
Día del Mes	X
Día de la semana	X
Festivo	X
Hora del día	X
Edad	
Sexo	
Actividad secundaria	X
Tipo de servicio	X
Tipo de carrocería	
Cilindrada	
Peso	
Potencia	
Modelo	X
Estado o Provincia	X
Ciudad	X
Altura Nivel del Mar	X
Clima	X
Tipo de Zona	X
Temperatura Promedio	X
Estado vía	
Velocidad GPS	X
Velocidad ECU	X
Temperatura Motor	X

RPM	X
Evento	X

El modelo de datos planteado en la sección 3.3, es un modelo IDEAL, del que probablemente, no muchos SIG contienen todos los atributos. En el caso específico del sistema ***idMaps***, este cumple aproximadamente con el 73% de los atributos propuestos, de las variables identificadas. En este trabajo de maestría, únicamente se realizaron pruebas con este SIG, pero se esperaría que el modelo funcione, cuando el SIG con el que se vaya a instanciar dicho modelo, cumpla al menos con el 70% de los atributos, incluyendo obligatoriamente el evento de riesgo, dado que este es el atributo que se desea predecir.

La herramienta ETL, fue desarrollada utilizando el lenguaje de desarrollo C# y el IDE Visual Sudio 2012, bajo la licencia de la empresa ***Id Company S.A***. Como motor de base de datos, para la consolidación de los datos, se utilizó ***PostgreSQL*** en su versión 9.2, y el complemento de herramientas ***PostGIS***, para las operaciones GIS.

El propósito del desarrollo de la ETL, fue generar un conjunto de datos, que unificara todas los atributos y variables relacionadas con los eventos registrados en el sistema ***idMaps***, en una sola tabla, que fue nombrada “Hechos”, y que también fue almacenada en una base de datos ***PostgreSQL***. La herramienta ETL, se encarga de consultar y transformar todos los datos desde la base de datos de origen, hasta una base de datos de destino, de la cual posteriormente se extrajeron dichos datos para ser pre-procesados y analizados para después llevar a cabo el proceso de minería.

El código fuente de la ETL, se encuentra en el Anexo A, incluido en el disco que se entrega junto con este trabajo, el cual se encuentra comentado detalladamente.

3.5 Análisis Descriptivo y Pre-procesamiento de datos

En esta sección del trabajo, se explica cuáles fueron las metodologías y técnicas de pre-procesamiento y análisis descriptivo, utilizadas sobre el conjunto de datos resultante de la ejecución del *script* de la ETL.

3.5.1 Identificación de dimensiones

Inicialmente se identificaron aquellos atributos, que pueden aportar información y cuales no aportan, según la naturaleza de los datos.

En la sección 3.4, se describieron aquellos atributos que el conjunto de datos contiene a partir del modelo de bodega de datos planteado. En el conjunto de datos extraído del SIG, cada uno de estos atributos, contiene datos nulos o inválidos, por lo tanto fueron filtrados del mismo y posteriormente identificados aquellos atributos que si aportan al proceso de minería. De esta manera se elimina parte del ruido que contenían los datos. Una vez filtrados dichos atributos, también fueron identificados aquellos en los que su valor varia, dada la naturaleza del SIG. A continuación se presentan las justificaciones y los atributos que no fueron incluidos en el proceso de minería.

El SIG no cuenta con esta información. Por ende, no puede tenerse en cuenta en el proceso.

- Edad (Conductor)
- Sexo (Conductor)
- Tipo Carrocería (Vehículo)
- Cilindrada Motor (Vehículo)
- Peso(Vehículo)
- Potencia (Vehículo)
- Modelo (Vehículo)
- Estado de la vía (Ubicación)

Los datos no varían en el SIG, dada la naturaleza del mismo. Solo se monitorean vehículos de transporte de carga.

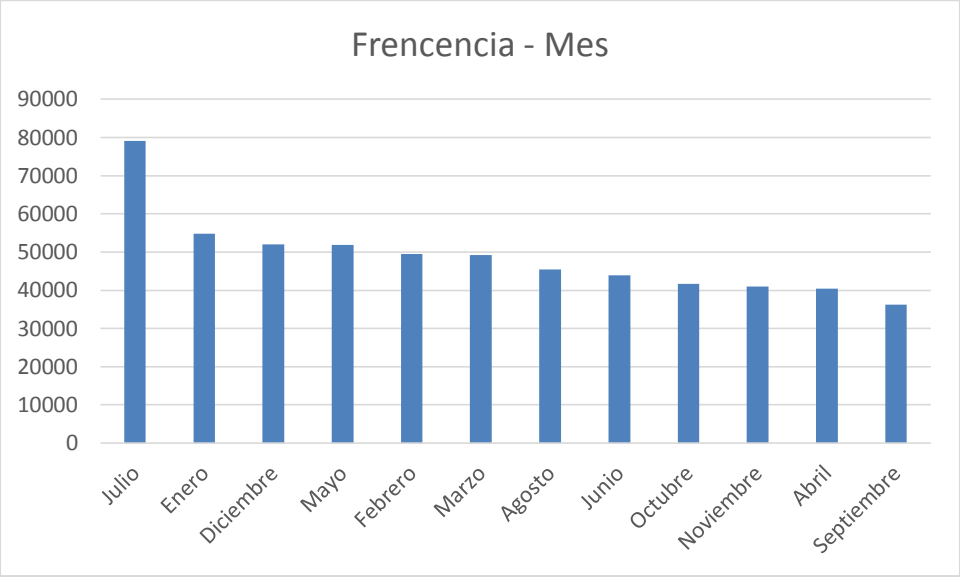
- Actividad Secundaria (Conductor)
- Tipo de Servicio (Vehículo)

3.5.2 Análisis Descriptivo

En esta sección, se explica el análisis descriptivo realizado al conjunto de datos, en donde se identifica ruido, datos fuera de rango y se explican los procesos llevados a cabo para la limpieza de datos numéricos.

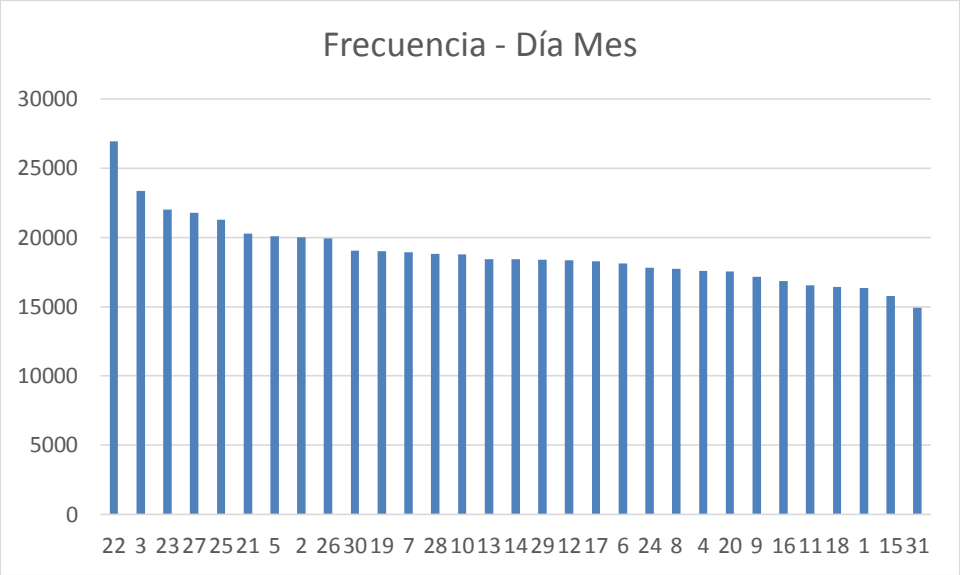
Antes de realizar esta tarea, primero fue necesario identificar aquellos atributos, a los que fue necesario aplicar técnicas cuantitativas para la detección de datos atípicos. Inicialmente se grafican histogramas de las frecuencias los datos nominales y se muestran a continuación.

Figura 9: Frecuencia – Mes del año



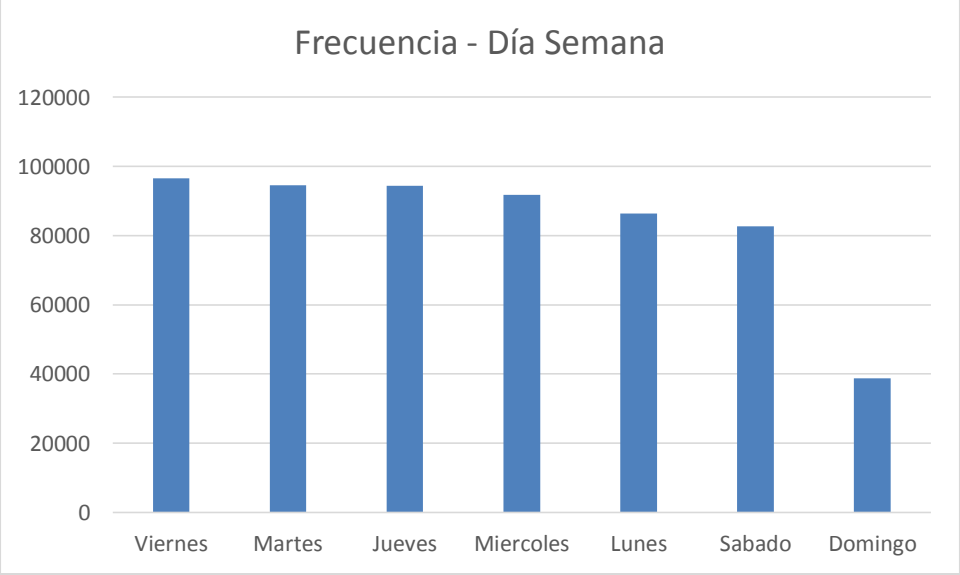
Del anterior histograma, se pueden identificar gráficamente, que el mes del año en el que más se presentaron eventos de riesgo, fue Julio, aunque para conocer las condiciones en que se presentaron estos eventos, se observan más adelante en los resultados del proceso de minería.

Figura 10: Frecuencia – Día Mes



El gráfico anterior, permite observar que el día 22 del mes, es aquel en el que más eventos de riesgo se generan y que el día 31, es aquel que menos eventos registra, dado que solo 7 meses del año, tienen día 31, así que este comportamiento, era tácito en los datos.

Figura 11: Frecuencia – Día Semana



La figura 11, muestra que el día domingo, es aquel en el que menos eventos se registran. Este comportamiento es normal, dado que el conjunto de datos, fue extraído de un SIG de vehículos de carga, los cuales trabajan normalmente en días de semana y en circunstancias normales, la gran mayoría de los conductores, descansan los domingos.

Figura 12: Frecuencia – Día Festivo

Se debe considerar, que el conjunto de datos extraídos del SIG, se encuentran ubicados temporalmente entre Septiembre 1 de 2013 y Agosto 31 de 2014 y son únicamente de Colombia en donde existen varios días festivos al año. Aunque gráficamente la cantidad de festivos se ve insignificante, la cantidad de eventos de riesgo registrados fueron 11.422 y se consideró importante observar el comportamiento de estos eventos en estos días.

Figura 13: Frecuencia – Hora del Día

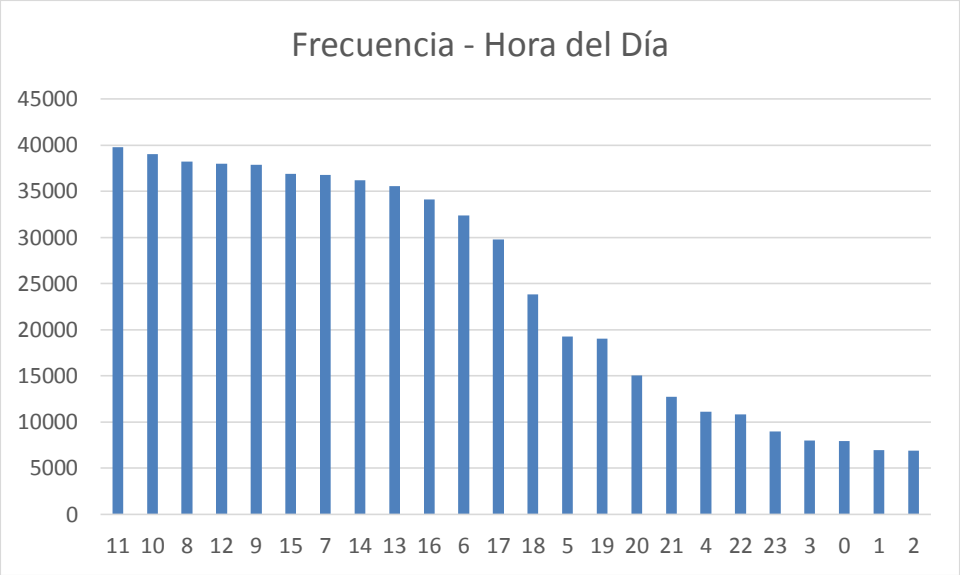
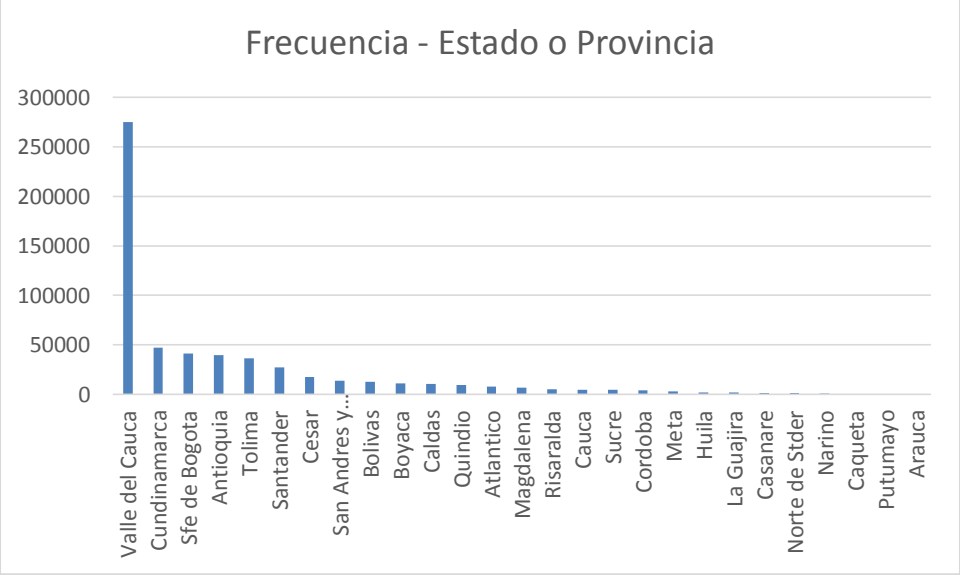
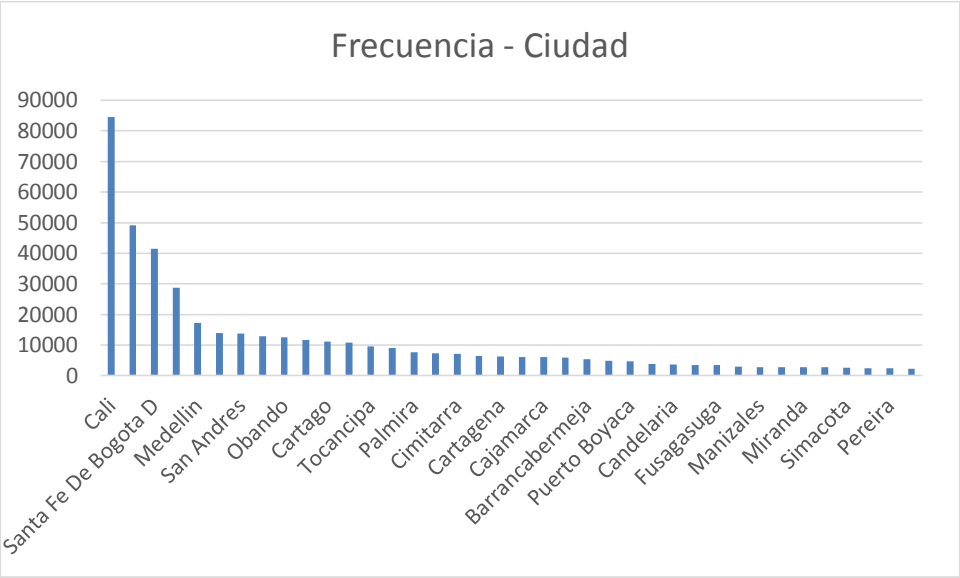


Figura 14: Frecuencia – Estado o Provincia



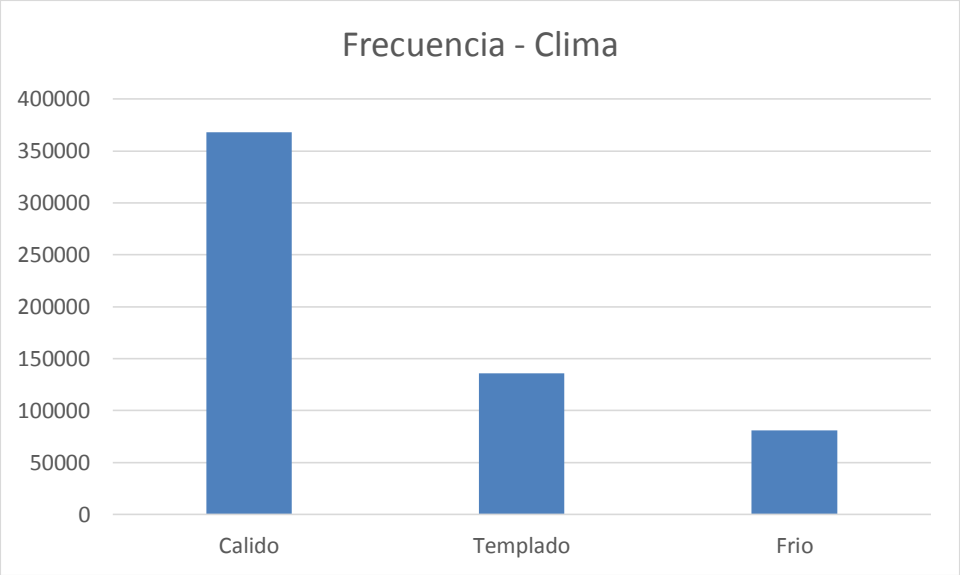
En el histograma de frecuencias del departamento, se puede observar que existe una diferencia bastante notoria entre el departamento del Valle del Cauca, con respecto a los demás. Basados en este gráfico, se podría deducir que es en esta ubicación, en donde más eventos de riesgo se producen, sin embargo, también podría determinar que en esta ubicación es donde se presenta la mayor concentración de vehículos. Por eso, para poder sacar conclusiones correctas, fue necesario llevar a cabo el proceso de minería.

Figura 15: Frecuencia – Ciudad



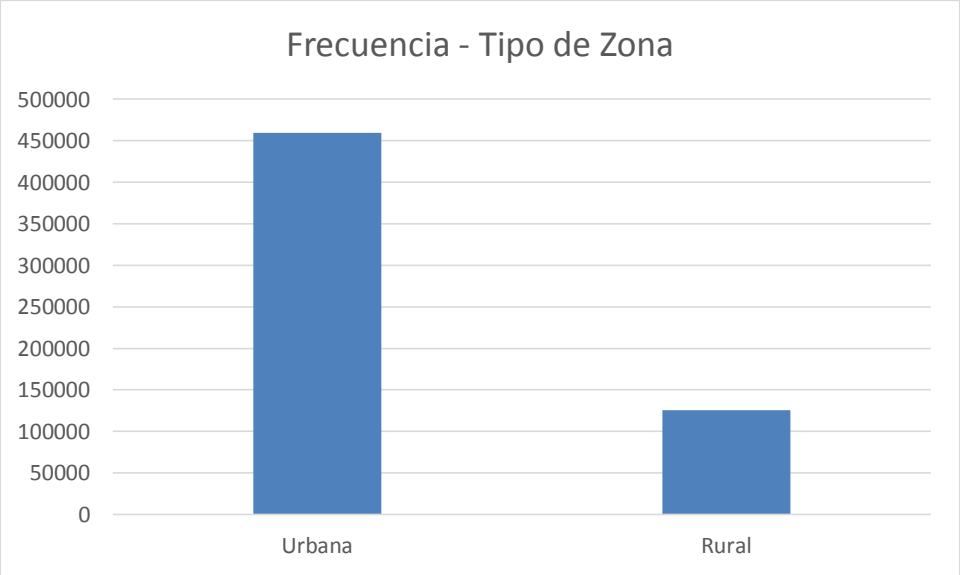
Los datos utilizados para graficar el histograma mostrado en la figura 15, no fueron completos, únicamente se tomaron valores significativos, puesto que existen ciudades en las que únicamente se ha generado un evento de riesgo en todo el conjunto de datos. Posteriormente en este trabajo, se describen las acciones que se tomaron para este tipo de datos. De la gráfica, se puede observar, que en donde más se generan eventos de riesgo, es en ciudades principales y siendo consecuente con la figura 14, la ciudad en la que mayor cantidad de eventos de riesgo se presentaron, fue en Cali, capital del departamento del Valle del Cauca.

Figura 16: Frecuencia – Clima



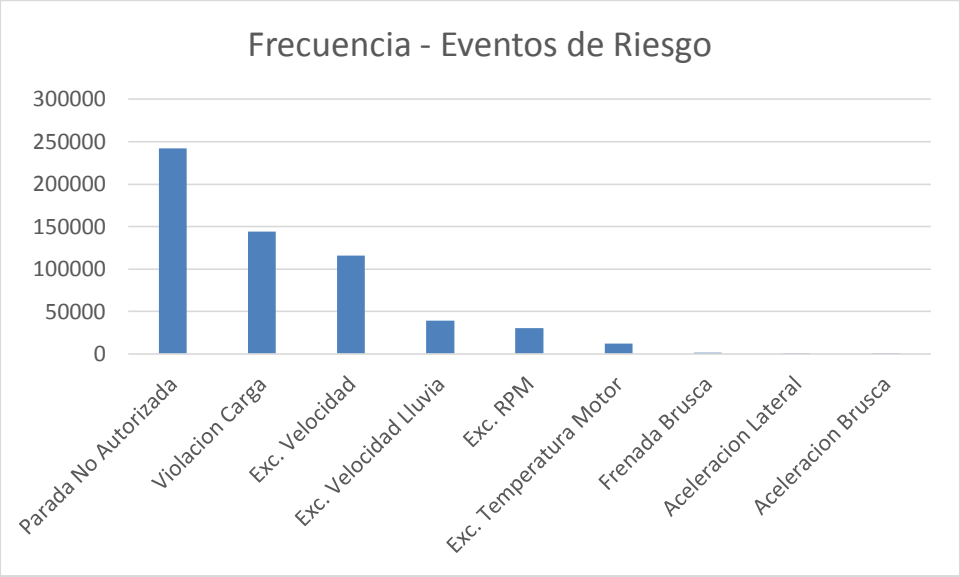
Observando esta gráfica, es posible concluir, que la mayor cantidad de eventos de riesgo registrados en el SIG, ocurren en climas cálidos y la menor cantidad en clima frio. Existe una alta diferencia entre la cantidad de eventos generados según el clima.

Figura 17: Frecuencia – Tipo de Zona



El SIG, permite conocer el tipo de zona en la que se genera un evento de riesgo. Para la muestra de datos, es posible conocer que la mas de la mitad de eventos de riesgo ocurridos, se generan en zonas urbanas, tal y como es de esperarse, en vehículos de carga pesada.

Figura 18: Frecuencia – Eventos de Riesgo



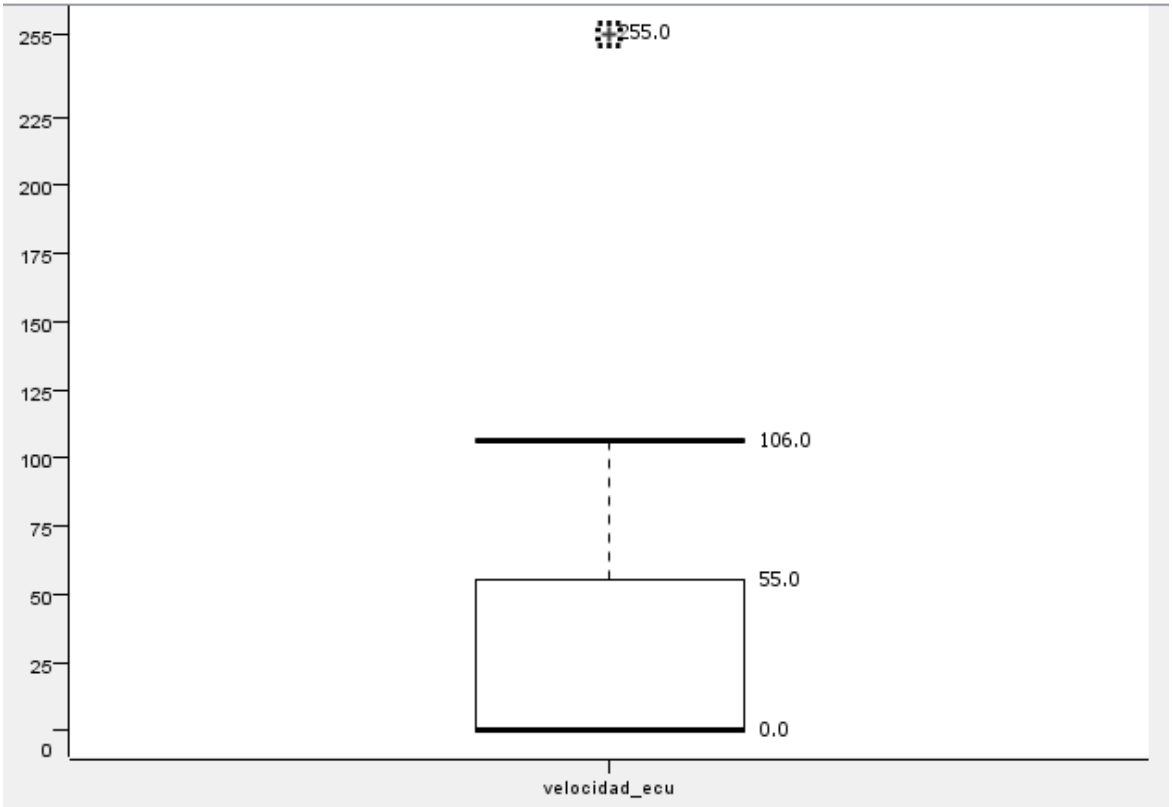
A partir de la gráfica 18, se pudo identificar, que para el caso de estudio, no es posible tener en cuenta los eventos de Aceleración Lateral y Aceleración Brusca, ya que en la muestra de datos que inicialmente cuenta con 585.254 registros, solamente se tienen 6 eventos de aceleración lateral y 4 de aceleración brusca. Con tan pocos registros para estos dos eventos, no es posible identificar las características que los producen, ni mucho menos predecirlos. Por esta razón, a partir de este momento, estos dos eventos, quedan por fuera del pre-procesamiento y estos 10 registros, son eliminados de la muestra de datos.

Posteriormente, se debe eliminar el ruido que se ha podido identificar en los datos, en algunos atributos en especial. Se conoce el ruido, como aquellos datos que han sido distorsionados o modificados debido al instrumento de medida con el que se capturan, en este caso, el dispositivo AVL. En algunos casos, ya sea porque la ECU del motor del

vehículo o fallas en el dispositivo AVL, se pueden presentar valores atípicos en la captura de la velocidad de la ECU o en temperatura del motor. Esto puede apreciarse, realizando un análisis estadísticos de los datos, en los que se encuentran velocidades hasta de 255 Km/H (valores no posibles en vehículos de carga) o temperaturas de 65534 grados (valor no posible en la temperatura de un motor) las cuales deben ser tratadas, ya que son datos fuera de rango, que pueden afectar posteriormente el proceso de minería. A continuación, se explican las acciones tomadas para estos dos atributos, con respecto a los registros mencionados.

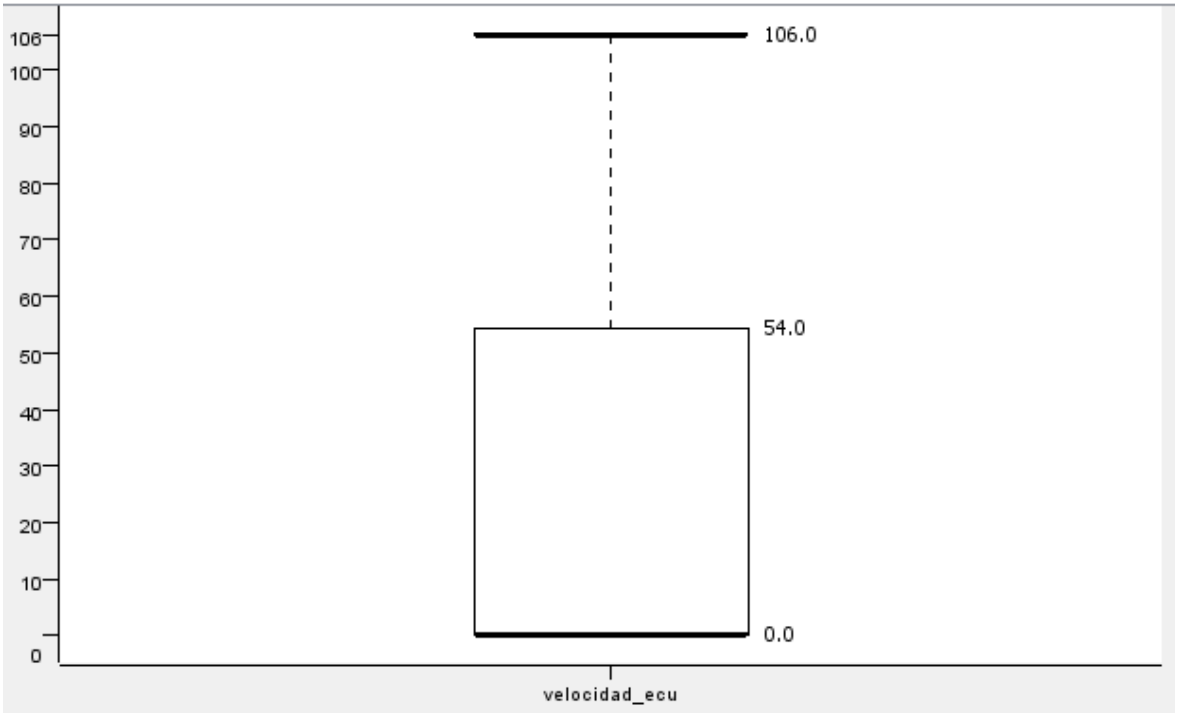
En los datos correspondientes a la velocidad ECU, es posible detectar que su valor mínimo y más frecuente es 0, y su valor más alto es 255. El valor de velocidad 255, es un valor considerado como ruido, ya que este es un error de medición del dispositivo AVL, que proviene desde la ECU del motor de un vehículo, cuando por algún error interno, no es posible reportar un valor real de velocidad.
Utilizando una gráfica de caja (Box Plot), puede observase que la mayor cantidad de los registros, cuentan con una velocidad 0, pero existen datos con valores fuera de rango. Aquellos en los que la velocidad registrada por la ECU, tienen una magnitud de 255 Km/H.

Figura 19: Gráfica de Caja – Velocidad ECU, versión inicial



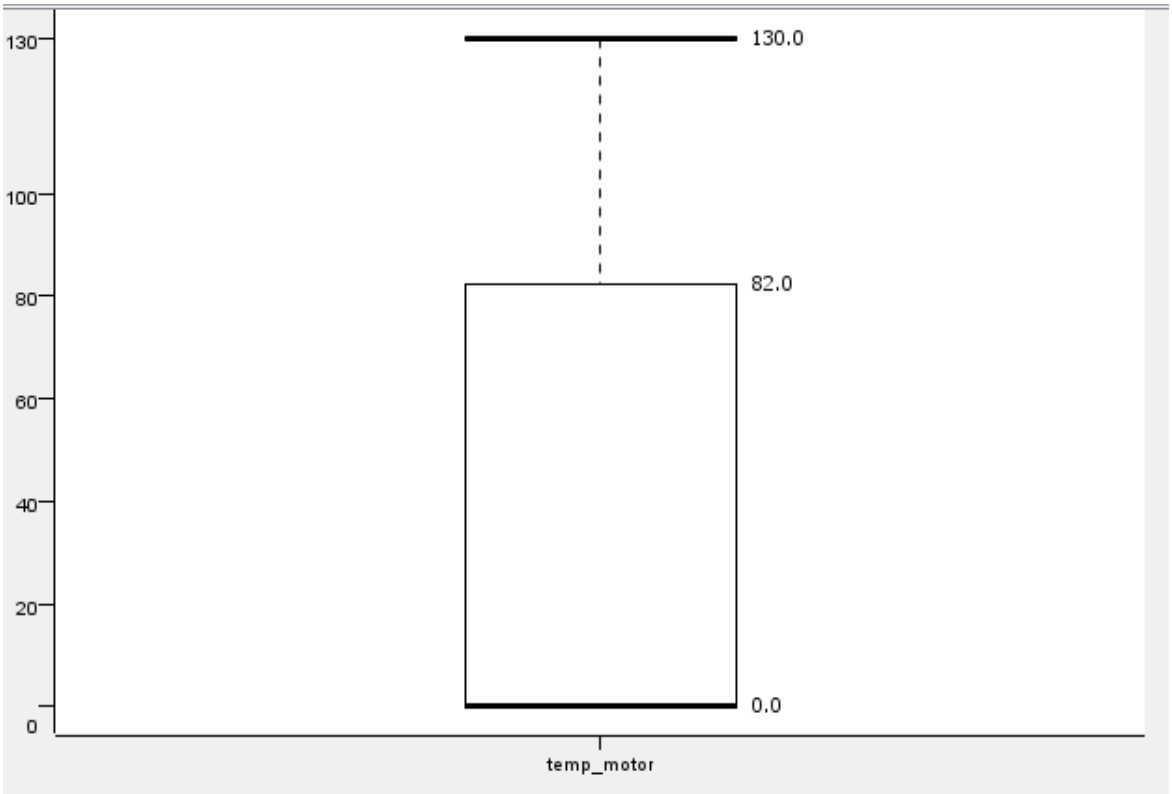
Ya identificado este valor fuera de rango, aquellos registros que contenían este dato, fueron modificados. El valor 255 Km/H, fue reemplazado por el valor de la velocidad GPS, que es un atributo similar a la velocidad ECU. No se reemplazó por la Moda, ya que esta es igual a cero y se requería balancear un poco la gráfica. Tampoco se reemplazó por la medía, ya que los registros en los que la velocidad ECU, registraban este valor, siempre estaban relacionados con el mismo evento y esto podría afectar posteriormente las predicciones. La figura 20, muestra una nueva gráfica de caja, con los valores fuera de rango, actualizados.

Figura 20: Gráfico de Caja - Velocidad ECU, versión final



Una vez corregido el ruido en el atributo de velocidad ECU, fue necesario hacer lo mismo con el atributo de temperatura del motor, el cual, tiene un problema similar en la captura del dato, desde el dispositivo AVL, pero en estos casos, la magnitud reportada es de 65.534. El grafico de caja, en esta ocasión, no permitió mucha claridad en los datos, ya que se encontraban muy dispersos, pues su dominio estaba en valores de 0 a 65.534, por esta razón, en el trabajo, únicamente se expone el resultado obtenido después de eliminar el ruido. En este caso, se aplicó la técnica de reemplazar el ruido por la moda de los datos sin tener en cuenta la magnitud 0, ya que con la media de los datos, se obtenía un valor, no típico para la temperatura normal de un vehículo encendido. La moda en este caso, fue de una magnitud de 82.

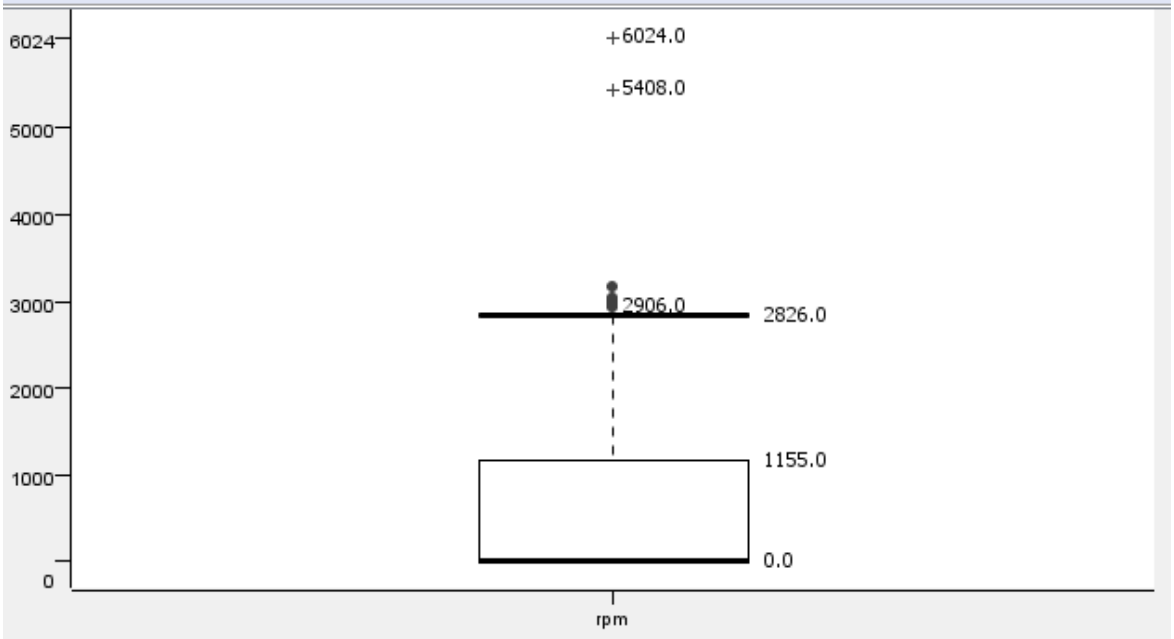
Figura 21: Gráfica de Caja - Temperatura de Motor



3.5.3 Detección de Datos Fuera de Rango

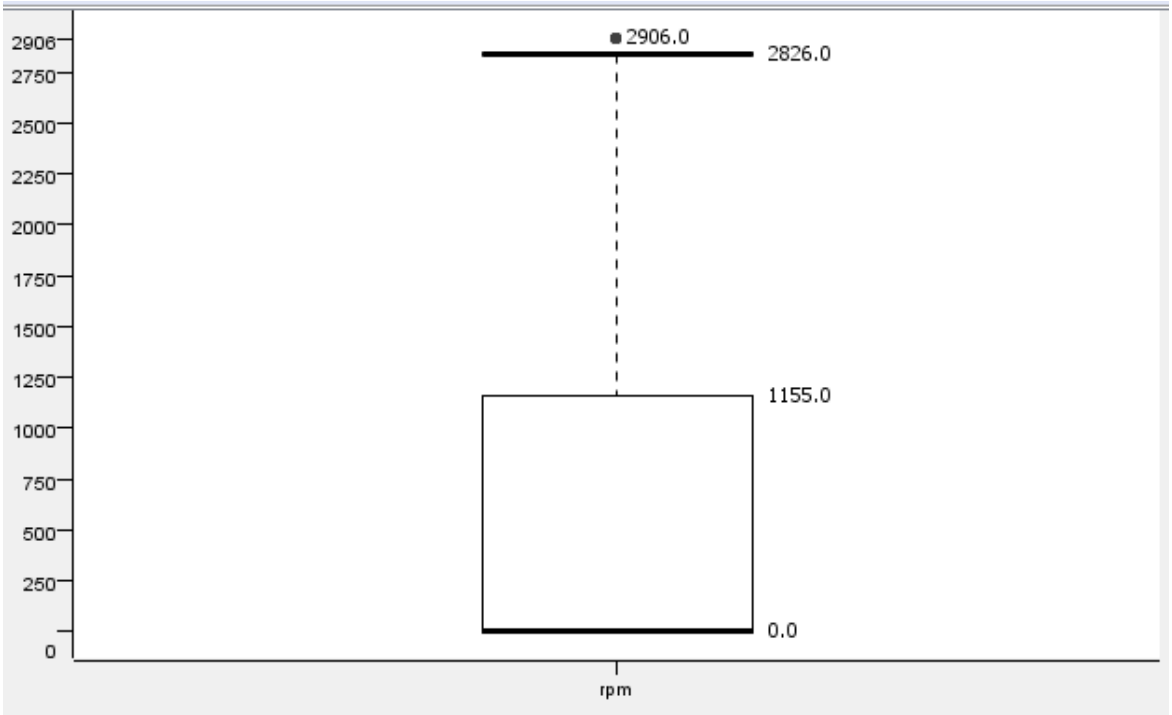
Realizado un análisis de las gráficas de caja de las demás variables numéricas, se identificó en el atributo RPM del vehículo, algunos registros, que aunque no pueden ser considerados como ruido (porque no existe un error en la medición), si afectan los datos estadísticos de dicha variable. A continuación se observa gráficamente la distribución de los datos de este atributo, en donde se pueden observar los datos fuera de rango existentes.

Figura 22: Gráfica de Caja - RPM



Observando la gráfica de caja, se hizo un filtro de los datos y se encontraron 7 registros que contienen los datos fuera de rango en el atributo RPM. El conjunto de datos, fue particionado en dos. En una parte, se obtuvieron los registros que contienen los datos fuera de rango y en la otra parte, los demás. A esta segunda parte, se le calculó nuevamente información estadística, para obtener una nueva media del atributo RPM, sin los datos fuera de rango. Posteriormente, a los 7 registros de la primera partición, se le sustituyó el dato de RPM, por el máximo valor de la segunda partición (2906), dado que estos registros, corresponden a eventos de riesgo de *exceso de RPM*. A continuación se muestra la nueva gráfica de caja, para el atributo RPM.

Figura 23: Gráfica de Caja – RPM, versión final



Cuando se identifican datos fuera rango, no solo es necesario realizar un análisis de atributos numéricos, pues dada la naturaleza de los datos, también es necesario identificar aquellos datos, que geográficamente, se encuentran fuera rango. Para la identificación de dichos datos, se tuvo en cuenta el atributo **Estado o Provincia**, que para el caso de la división política de Colombia, se toman como los departamentos del país. Para esto, es necesario retomar las frecuencias de los datos mostradas en la Figura 14: Frecuencia – Estado o Provincia, pero a continuación, se observa la tabla, que realiza el conteo de estos datos.

Tabla 3: Frecuencias Estado o Provincia

Estado	Frecuencia
Valle del Cauca	274817
Cundinamarca	47026

Sfe de Bogota	41476
Antioquia	39551
Tolima	36675
Santander	27385
Cesar	17523
San Andrés y Providencia	13821
Bolívar	12456
Boyacá	11361
Caldas	10402
Quindío	9589
Atlántico	7931
Magdalena	6672
Risaralda	5176
Cauca	4721
Sucre	4499
Córdoba	4153
Meta	2799
Huila	1998
La Guajira	1688
Casanare	1582
Norte de Stder	1149
Nariño	721
Caquetá	41
Putumayo	41
Arauca	1

Considerando la cantidad de registros que contiene el registro de datos en este momento (585.244), existen datos de algunos Departamentos, en los que la cantidad de eventos, no alcanzan ni siquiera el 1% del total de los datos, se decidió filtrar aquellos, en los que la cantidad de eventos, no alcanzan los 1.000 eventos por Departamento. Así, se eliminan aquellos registros, que representan geográficamente, datos fuera de rango. Por ende, la cantidad de eventos contenidos en los Departamentos de Nariño, Caquetá, Putumayo y Arauca, no serán tenidos en cuenta en el proceso de minería.

Finalmente, el conjunto de datos inicial, contiene 584.440 registros.

3.5.4 Muestreo de los Datos

Dada la cantidad de datos contenidos en el conjunto y la naturaleza del trabajo, fue necesario separar el conjunto en muestras distintas. El modelo generado y la minería de datos fue realizada con unas muestras y con las restantes, se realizó la validación del mismo. Para el caso de estudio, extrajeron cuatro muestras de datos. Dos para la generación del modelo y dos para su validación. Las muestras de datos se encuentran en el Anexo B, en el CD adjunto a este trabajo.

3.6 Prueba de Algoritmos de Clasificación

Una vez extraídas las muestras de datos del conjunto de datos original, se procedió a aplicar diversos algoritmos de clasificación tal y como se expuso en la metodología a llevar en el inicio del capítulo 3 de este trabajo. Dado que se obtuvieron cuatro muestras de datos, fueron llamadas *Muestra1*, *Muestra2*, *Muestra3* y *Muestra4* respectivamente. Para la generación de los modelos, fueron utilizados 3 algoritmos:

- Redes Bayesianas – Algoritmo K2 de construcción de la red
- Redes Neuronales Artificiales – Perceptron Multicapa
- Arboles de Decisión – Algoritmo J48

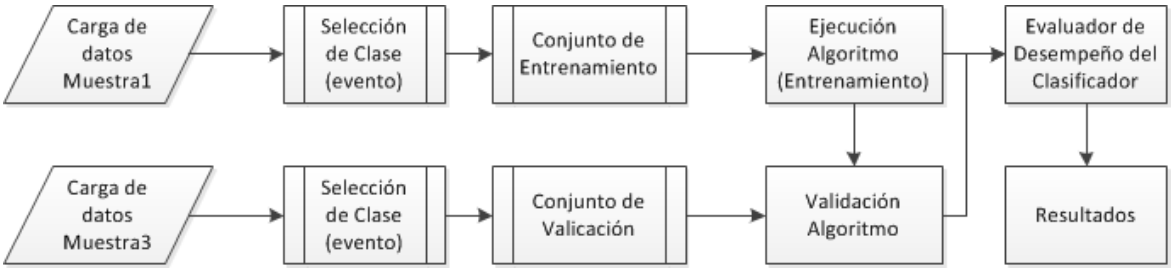
A continuación se exponen los resultados obtenidos con la ejecución de cada uno de los algoritmos.

3.6.1 Redes Bayesianas

Este algoritmo fue ejecutado con una muestra de datos y sus resultados, probados con una de las muestras restantes.

Inicialmente la red fue construida con la *Muestra1*, y validado con la *Muestra3*, siguiendo el siguiente proceso.

Figura 24: Proceso clasificación - BayesNet



La cantidad de registros, contenida en el conjunto de datos de entrenamiento, es la misma cantidad, contenida en el conjunto de datos de validación, que es de 146.110 registros. Una vez ejecutado el algoritmo, generado el primer modelo clasificador y validado el mismo, se obtuvo una precisión del 94.47%, con un total de 138.027 eventos correctamente clasificados y solo un 5.53% de error. La ejecución del algoritmo, tomó alrededor de 1 minuto, siendo su ejecución muy eficiente.

La red construida con el algoritmo K2, se muestra en la figura 25.

Figura 25: Red Bayesiana

A continuación se muestra la matriz de confusión obtenida después de la ejecución del algoritmo y generación del modelo predictivo.

Tabla 4: Matriz de Confusión – Red Bayesiana

A	B	C	D	E	F	G	← Clasificado como
54804	5043	0	12	0	305	0	**A = Parada NA**
551	35136	132	90	42	166	33	**B = Violación Carga**
44	270	28648	22	21	30	24	**C = Exc. Velocidad**
103	1	34	7032	43	247	2	**D = Exc. RPM**
10	35	46	32	9698	47	5	**E = Exc. Velocidad Lluvia**
441	10	41	135	55	2299	2	**F = Exc. Temp. Motor**
0	8	1	0	0	0	410	**G = Frenada Brusca**

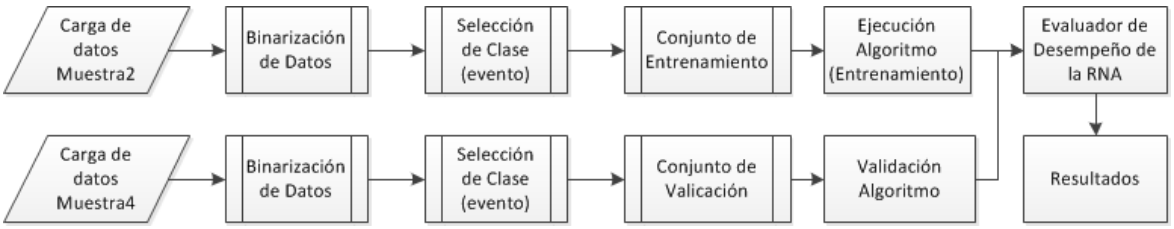
Observado los resultados obtenidos en la matriz de confusión, fue posible deducir, que en su diagonal, se encuentran siempre los valores más altos, lo que indica que en la mayoría de los casos, las predicciones fueron acertadas y la clasificación de los eventos fue realizada correctamente en un alto porcentaje, como se precisó anteriormente.

3.6.2 Redes Neuronales Artificiales

La red neuronal, fue entrenada con la *Muestra2* y probada con la *Muestra4*. Ambas muestras, contienen un total de 146.110 registros, extraídos del conjunto de datos original.

Para la implementación de la red neuronal y el modelo predictivo, se llevó a cabo un proceso similar al de las redes bayesianas. Sin embargo, es de recalcar, que debido a la naturaleza y funcionamiento del algoritmo, fue necesario binarizar los datos, para su ejecución. En la figura 26, se observa la metodología de dicho proceso.

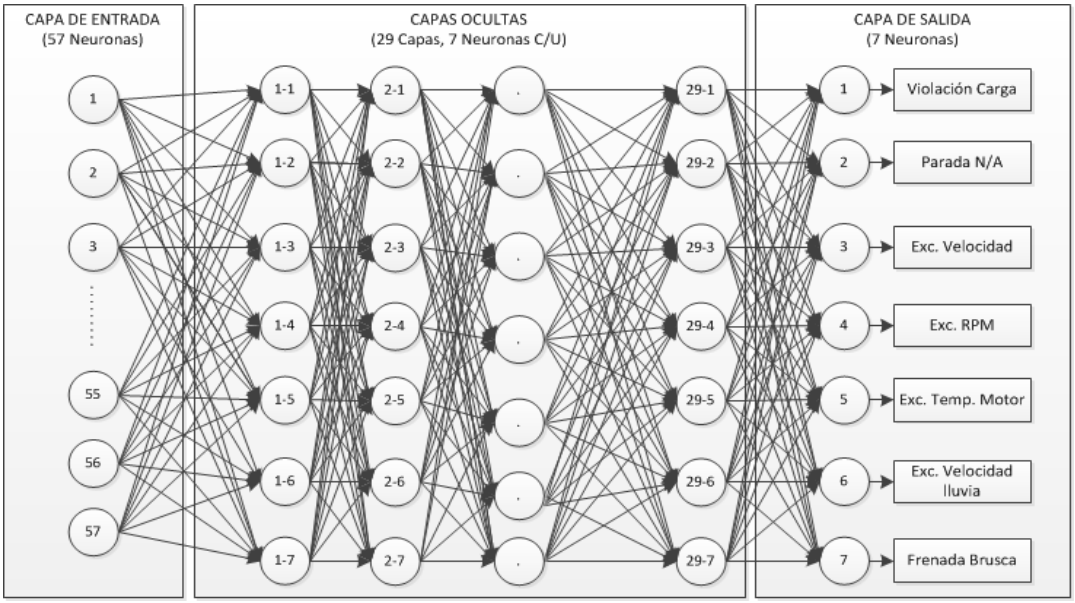
Figura 26: Proceso clasificación - RNA



Teniendo en cuenta la cantidad de atributos después de la binarización de los datos y la clase, así mismo se sugieren la cantidad de capas ocultas de la red neuronal y la cantidad de neuronas en cada capa. Son 7 valores que puede tomar la clase (7 eventos

de riesgo) y 57 atributos. Utilizando el software de minería de datos Weka, se sugiere la utilización de la cantidad de atributos + la clase / 2. Según este cálculo, la cantidad de capas utilizadas fue de 29, con 7 neuronas cada una. En la figura 27, se muestra un bosquejo de la RNA.

Figura 27: Bosquejo - Red Neuronal Artificial



Habiendo ejecutado el algoritmo, para la clasificación de eventos de riesgo, los resultados obtenidos, son un poco mejores, comparados con las redes bayesianas, pues un 95.1% de los eventos fueron clasificados correctamente contra un 4.9% de error. Sin embargo, el entrenamiento de la red, toma mucho más tiempo en comparación al algoritmo anterior. En este caso, utilizando la RNA mostrada en la figura 27, su entrenamiento tomo aproximadamente 3 horas y 30 minutos, tiempo relativamente alto en comparación a los 15 segundos que tomó la ejecución del clasificador que utiliza la red bayesiana.

En la tabla 5, se muestra la matriz de confusión obtenida, posterior al entrenamiento y validación de la RNA.

Tabla 5: Matriz de Confusión - RNA

A	B	C	D	E	F	G	← Clasificado como
34065	1595	122	118	121	72	11	**A = Violación Carga**
3169	56771	0	67	63	3	0	**B = Parada NA**
77	27	28793	39	21	36	14	**C = Exc. Velocidad**
0	114	64	7254	101	21	0	**D = Exc. RPM**
10	531	80	189	2223	54	1	**E = Exc. Temp. Motor**
25	18	48	187	97	9479	4	**F = Exc. Velocidad Lluvia**
16	92	51	0	0	0	267	**G = Frenada Brusca**

De la matriz de confusión es posible concluir, que el clasificador realiza un buen trabajo, utilizando la red neuronal, sin embargo sus costos computacionales son elevados ya que toma mucho tiempo en su ejecución.

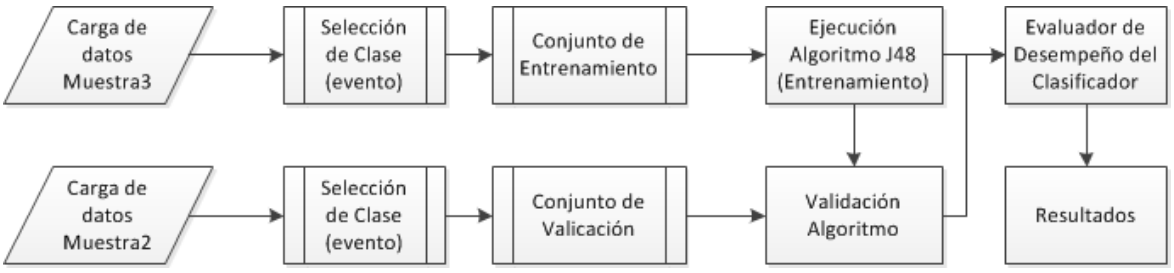
3.6.3 Arboles de Decisión

Como clasificador, se utilizó el algoritmo J48 de árboles de decisión. Igual que en los casos anteriores, el clasificador fue entrenado con una de las muestras y probado con otra. Para este caso, se entrenó con la *Muestra3* y se validó con la *Muestra2*.

El árbol generado posterior a la ejecución del algoritmo, es demasiado grande para ser adjuntado a este trabajo, por lo tanto, únicamente se mostrarán sus resultados de ejecución, precisión y matriz de confusión.

El proceso llevado a cabo para la generación del modelo, utilizando arboles de decisión, se muestra en la figura 28.

Figura 28: Proceso Clasificación - Arboles de Decisión



Los resultados obtenidos después de aplicar este algoritmo, son un poco mejores, comparados con los otros dos expuestos anteriormente. En este caso, el modelo pudo clasificar correctamente 142.037 eventos, equivalentes al 97.2% de los datos, con solo un 2.8% de error. Además, los tiempos de ejecución, son reducidos, similares a la ejecución de las redes bayesianas. Aplicando este algoritmo, se obtuvo un 100% de precisión clasificando el evento “Frenadas Bruscas”, característica que no se presentó con ningún otro evento de riesgo en esta prueba, ni en las pruebas realizados previamente con los otros algoritmos. A continuación se encuentra la tabla 6, en la cual es posible observar los datos de la matriz de confusión, obtenida como resultado de la aplicación de árboles de decisión a los datos.

Tabla 6: Matriz de Confusión – Arboles de Decisión

A	B	C	D	E	F	G	← Clasificado como
58786	1768	2	7	0	31	0	**A = Parada NA**
922	34566	95	39	52	40	44	**B = Violacion Carga**
9	48	28802	12	20	8	1	**C = Exc. Velocidad**
77	47	33	7304	25	102	0	**D = Exc. RPM**
5	14	17	34	9851	6	0	**E = Exc. Velocidad Lluvia**
360	30	35	106	84	2304	0	**F = Exc. Temp. Motor**
0	0	0	0	0	0	424	**G = Frenada Brusca**

3.7 Selección del Modelo Predictivo y Modelo Propuesto

Ya expuestos los tres modelos generados a partir de los algoritmos de clasificación y habiendo obtenido los resultados y pruebas de cada uno de ellos, se seleccionó el modelo explicado en la sección 3.6.3 de este trabajo, el cual fue generado utilizando arboles de decisión, específicamente el algoritmo J48. Puesto, que en la sección anterior, se documentaron los resultados de dicho modelo, a continuación se exponen elementos estadísticos del modelo teniendo en cuenta el valor bajo la curva ROC de los resultados de la clasificación de cada evento. Cabe recalcar, que las curvas ROC, muestran la precisión del algoritmo, para clasificar cada una de las clases (o eventos de riesgo, para este caso) del conjunto de datos, teniendo en cuenta los valores de la matriz de confusión (verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos).

Tabla 7: Resultados estadísticos

	#	%
Instancias Correctamente clasificadas	142037	97,20%
Instancias Incorrectamente clasificadas	4073	2,80%
Coeficiente Kappa		0,97
Media del Error Absoluto		0,02
Error Relativo Absoluto		6,3

En la tabla 8, es posible observar la precisión detallada del modelo, por cada una de las clases.

Tabla 8: Precisión Detallada por Clase

	Tasa Verdaderos Positivos	**Tasa Falsos Positivos**	**Precisión**	**Recall**	**Área ROC**	**Clase**
	0,97	0,016	0,977	0,97	0,996	Parada No Autorizada
	0,967	0,017	0,948	0,967	0,994	Violación Carga
	0,997	0,002	0,994	0,997	0,999	Exc. Velocidad
	0,963	0,001	0,974	0,963	0,993	Exc. RPM
	0,992	0,001	0,982	0,992	0,998	Exc. Velocidad Lluvia
	0,789	0,001	0,925	0,789	0,977	Exc. Temperatura Motor
	1	0	0,904	1	1	Frenada Brusca
Promedios	0,972	0,011	0,972	0,972	0,996	

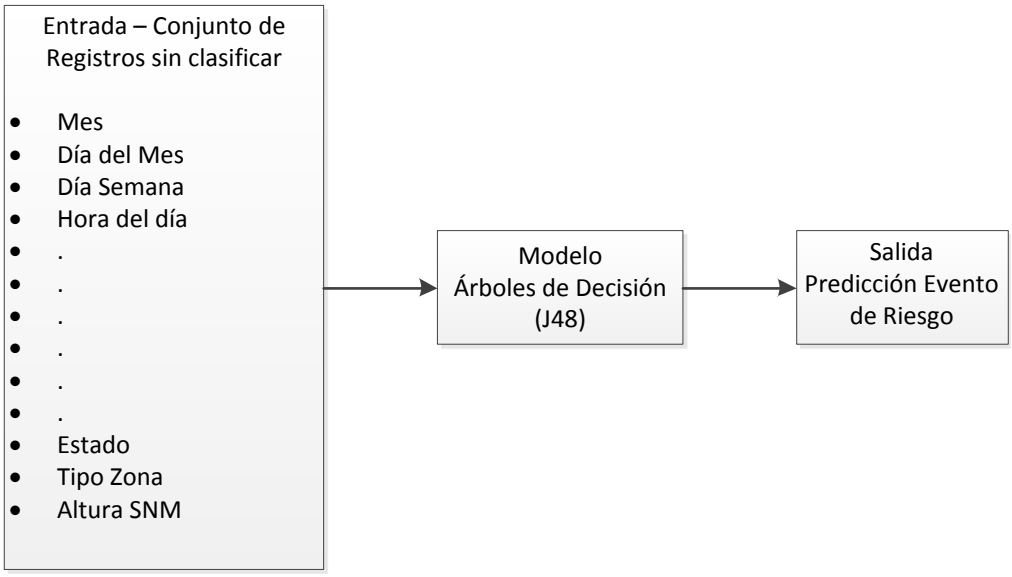
De la tabla 8, se pudo identificar la precisión con la que cada una de los eventos de riesgo fueron clasificados y de aquí se puede concluir que el evento “Frenada Brusca” es aquel con la mejor tasa de Verdaderos Positivos (VP o TP por sus siglas en inglés), con un 100% de VP en su clasificación. Por el contrario, el evento “Exceso de Temperatura del Motor”, es aquel con la menor tasa de VP, con un valor de 78.9%.

Una vez fue seleccionado el modelo, se procedió a la integración con un SIG de monitoreo de vehículos. El mismo del cual fue extraído el conjunto de datos con el cual se desarrolló el modelo predictivo.

El modelo predictivo propuesto, está basado en un árbol de decisión, implementado con el algoritmo J48, en el software de minería de datos Weka en su versión 3.7. El árbol, es demasiado voluminoso para anexarlo en este trabajo, sin embargo, el archivo del modelo

y su código fuente en JAVA, generados por Weka, se encuentran en el disco de datos que acompaña este trabajo, en la carpeta de nombre Anexo B.
La figura 29, muestra el funcionamiento del modelo.

Figura 29: Funcionamiento del Modelo



3.8 Integración del Modelo con el SIG

Para la generación del modelo predictivo, se utilizó el software Weka en su versión 3.7. Este software, además de leer los datos y prestar algunos servicios de pre-procesamiento, ejecuta los algoritmos, que permitió la construcción del árbol de decisión utilizado para la clasificación de los datos. El modelo generado por este software, es exportable a código fuente en JAVA. A continuación se explica el proceso que se realizó para la integración del modelo, con el SIG *idMaps*, de la empresa Id Company S.A.

- Paso 1: Generación Archivo

El objetivo principal de la integración con el SIG, es que realice la clasificación de los datos y sus predicciones en tiempo real. Lo que significa, que debe realizar una clasificación del conjunto de datos, que contiene los registros más recientes de cada uno de los vehículos.

Para esto, se desarrolló un script, que consulta de la base de datos del SIG, con el último reporte de cada uno de los vehículos, que se encuentren en el país (Colombia), obteniendo cada uno de los atributos que fueron utilizados para la generación del modelo, exceptuando el evento de riesgo (o la clase), en donde en vez de indicar un evento (que hasta ahora es desconocido) se utilizó un signo de interrogación, para que posteriormente sea clasificado. Una vez obtenido este conjunto de datos, es almacenado en un archivo de texto plano, que sigue la estructura de un archivo ARFF, formato en el que el software Weka, lee un conjunto de datos.

En la Figura 30, se muestra una imagen de ejemplo, del resultado de este primero paso de la integración.

Figura 30: Ejemplo archivo ARFF

```
@RELATION DataTable
@ATTRIBUTE mes  {Septiembre,Octubre,Noviembre,Diciembre,Enero,Febrero,Marzo,Abril,Mayo,Junio,Julio,Agosto}
@ATTRIBUTE dia_mes  {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}
@ATTRIBUTE dia_semana {Viernes,Miercoles,Sabado,Martes,Domingo,Jueves,Lunes}
@ATTRIBUTE dia_festivo  {f,t}
@ATTRIBUTE hora_del_dia {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23}
@ATTRIBUTE estado_provincia {'Sfe de Bogota','Valle del Cauca',Caldas,Boyaca,Cundinamarca,Tolima,Atlantico,Antioq
@ATTRIBUTE altura_snm INTEGER
@ATTRIBUTE clima  {Frio,Calido,Templado}
@ATTRIBUTE tipo_zona  {Urbana,Rural}
@ATTRIBUTE temp_promedio  INTEGER
@ATTRIBUTE velocidad_gps  REAL
@ATTRIBUTE velocidad_ecu  REAL
@ATTRIBUTE temp_motor INTEGER
@ATTRIBUTE rpm  INTEGER
@ATTRIBUTE evento {'Parada No Autorizada','Violacion Carga','Exc. Velocidad','Exc. RPM','Exc. Velocidad Lluvia','
@DATA
Noviembre,27,Martes,f,11,'Valle del Cauca',917,Calido,Urbana,23,0,0,0,0,?
Noviembre,27,Martes,f,11,Cundinamarca,1728,Templado,Urbana,19,25,25,0,0,?
Noviembre,27,Martes,f,11,'San Andres y Providencia',2,Calido,Urbana,28,58,56,61951,1227,?
Noviembre,27,Martes,f,11,Casanare,350,Calido,Urbana,26,0,0,80,0,?
Noviembre,27,Martes,f,11,Caldas,190,Calido,Urbana,36,0,0,0,0,?
Noviembre,27,Martes,f,11,'Sfe de Bogota',2630,Frio,Urbana,14,0,0,0,0,?
Noviembre,27,Martes,f,11,'Valle del Cauca',917,Calido,Urbana,23,0,0,0,0,?
Noviembre,27,Martes,f,10,'Sfe de Bogota',2630,Frio,Urbana,14,0,0,0,0,?
Noviembre,27,Martes,f,11,'Sfe de Bogota',2630,Frio,Urbana,14,0.1,0.1,0,0,?
Noviembre,27,Martes,f,11,Tolima,255,Calido,Rural,27,0,0,0,0,?
Noviembre,27,Martes,f,11,'Sfe de Bogota',2630,Frio,Urbana,14,35,0,0,0,?
Noviembre,27,Martes,f,11,'Valle del Cauca',969,Calido,Urbana,23,0,0,0,0,?
Noviembre,27,Martes,f,9,Tolima,323,Calido,Urbana,26,0,0,23295,599,?
Noviembre,27,Martes,f,11,'Valle del Cauca',1612,Templado,Urbana,20,55,55,21503,1273,?
```

- Paso 2: Implementación del Modelo

Como se mencionó en la sección 3.8 de este trabajo, el software Weka, generó el código fuente del modelo en el lenguaje de programación JAVA. Haciendo uso de este código y utilizando el IDE NetBeans, fue posible desarrollar otro script, que permitiera hacer uso del clasificador, con un conjunto de datos en formato ARFF de Weka. En este segundo paso de la integración, es ejecutada una tarea que hace un llamado al script generado

por Weka, la cual recibe como parámetro el conjunto de datos ya listo en el archivo ARFF obtenido como resultado del Paso 1. El resultado de la clasificación, es almacenado en un archivo de texto plano, el cual permite conocer, cual es el evento más probable a ocurrir, según las características del último registro de cada vehículo.

En la figura 31, se muestra el resultado de una de las clasificaciones en tiempo real, del modelo.

Figura 31: Ejemplo de datos del SIG clasificados

```
=== Predictions on test data ===

 inst#     actual  predicted error predi
     1        1:? 1:Parada N       1
     2        1:? 2:Violacio       1
     3        1:? 5:Exc. Vel       1
     4        1:? 1:Parada N       1
     5        1:? 1:Parada N       1
     6        1:? 1:Parada N       1
     7        1:? 1:Parada N       1
     8        1:? 1:Parada N       1
     9        1:? 1:Parada N       1
    10        1:? 1:Parada N       1
    11        1:? 7:Frenada        1
    12        1:? 1:Parada N       1
    13        1:? 6:Exc. Tem       1
    14        1:? 5:Exc. Vel       1
    15        1:? 6:Exc. Tem       1
    16        1:? 4:Exc. RPM       1
    17        1:? 1:Parada N       1
    18        1:? 1:Parada N       1
    19        1:? 7:Frenada        1
    20        1:? 3:Exc. Vel       1
    21        1:? 6:Exc. Tem       1
    22        1:? 7:Frenada        1
    23        1:? 3:Exc. Vel       1
    24        1:? 1:Parada N       1
    25        1:? 6:Exc. Tem       1
```

- Paso 3: Lectura de Resultados

Posterior a la ejecución del modelo con los datos del SIG y la generación del archivo de resultados, se ejecuta otro script que fue desarrollado, para la lectura e interpretación de dicho archivo. Este script, se encarga de leer los resultados y almacenarlos en una tabla, en la base de datos del SIG, para que posteriormente este pueda leer los resultados y mostrarlos en el sistema, por medio de un reporte.

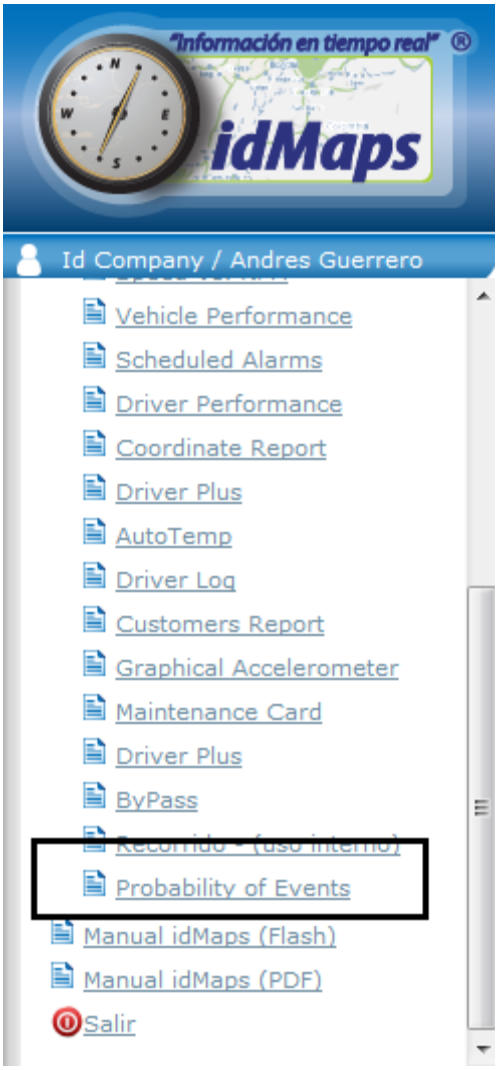
Figura 32: Datos Clasificados

id_vehiculo character var	fecha date	hora time without	fecha_hora timestamp wi	latitud real	longitud real	ubicacion character var	evento character varying
PFZ480	2014-11-27	14:45:40	2014-11-27	12.5708	-81.7005	Colombia, S	Parada No Autorizada
TFQ670	2014-11-27	14:24:52	2014-11-27	4.32157	-76.0911	Colombia ,	Parada No Autorizada
SZX908	2014-11-27	14:46:24	2014-11-27	4.60901	-74.0949	Colombia ,	Exc. Temperatura Motor
MQA912	2014-11-27	14:54:50	2014-11-27	5.4465	-74.6744	Colombia ,	Parada No Autorizada
THU843	2014-11-27	13:56:27	2014-11-27	5.30931	-72.6987	Colombia ,	Violacion Carga
SWM832	2014-11-27	14:49:05	2014-11-27	4.64631	-74.1492	Colombia ,	Parada No Autorizada
SVA942	2014-11-27	14:51:35	2014-11-27	4.56287	-75.9831	Colombia ,	Exc. Velocidad
SWO253	2014-11-27	14:53:06	2014-11-27	4.63452	-74.1031	Colombia ,	Exc. Temperatura Motor
TAM163	2014-11-27	14:43:27	2014-11-27	6.25366	-74.4961	Colombia ,	Exc. Velocidad
SYM421	2014-11-27	14:37:34	2014-11-27	4.63516	-74.1044	Colombia ,	Parada No Autorizada
OCK595	2014-11-27	11:03:31	2014-11-27	4.64664	-74.1295	Colombia ,	Frenada Brusca
TTZ353	2014-11-27	14:45:53	2014-11-27	4.63446	-74.1031	Colombia ,	Parada No Autorizada

- Paso 4: Reporte

Para finalizar la integración, se implementó un reporte, que hace parte del SIG, en el cual es posible visualizar los datos ya clasificados e indicando un posible evento a generarse en cada uno de los vehículos, según su último reporte. Para esto se agregó una opción de acceso a dicho reporte, que hasta el momento se encuentra en un usuario de pruebas, en donde constantemente se pueden observar las predicciones realizadas por el modelo. El paso 1, 2 y 3 de la integración del modelo al SIG, se ejecutan cada 5 minutos en el servidor en el que se encuentra el SIG, mostrando siempre información actualizada en el reporte del paso 4.

Figura 33: Acceso al reporte en el SIG



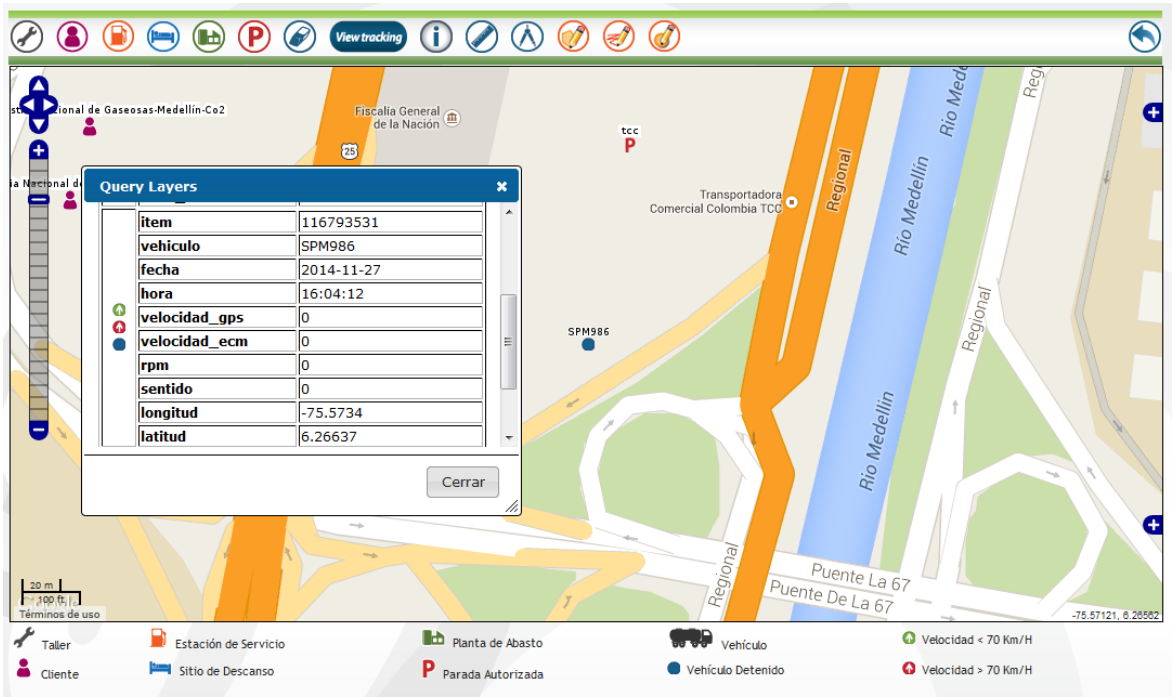
En la figura 34, es posible observar el reporte integrado con el SIG. Se muestran los datos de los últimos registros de cada vehículo (un único registro por vehículo) ya clasificados. En la última columna del reporte, se observa el evento de riesgo que el modelo predijo para cada vehículo, según los datos del registro. Además, en la primera columna, que indica la placa o identificador del conductor, se encuentra habilitado un hipervínculo al módulo geográfico del sistema, que nos permite ver sobre una cartografía, donde se encuentra ubicado exactamente el vehículo e incluso, consultar datos adicionales. Esta parte, se visualiza en la figura 35.

Figura 34: Reporte en el SIG

Predicción de eventos en Tiempo Real

Vehículo	Fecha Ultimo Reporte	Hora Ultimo Reporte	Altura S.N.M. (Mts)	Clima	Tipo Zona	Temp. Promedio (°C)	Última Ubicación	Evento Probable
CAI538	2014-11-27	15:49:12	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, PUENTE ARANDA, Parqueadero TIX Pte Aranda	Parada No Autorizada
CEL424	2014-11-27	15:48:44	995	Calido	Urbana	23	Colombia , Valle del Cauca , Cali, PARCELACIONES PANCE	Violacion Carga
CEL426	2014-11-27	15:49:37	995	Calido	Urbana	23	Colombia , Valle del Cauca , Cali, PARCELACIONES PANCE	Violacion Carga
CEO350	2014-11-27	15:56:58	995	Calido	Urbana	23	Colombia , Valle del Cauca , Cali	Violacion Carga
KUN690	2014-11-27	15:55:49	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, SAN RAFAEL	Parada No Autorizada
KUN969	2014-11-27	15:57:25	995	Calido	Urbana	23	Colombia , Valle del Cauca , Cali, PARCELACIONES PANCE, KR 125	Violacion Carga
MQA912	2014-11-27	15:59:50	190	Calido	Urbana	36	Colombia , Caldas , La Dorada	Parada No Autorizada
OCK595	2014-11-27	11:03:31	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, VILLA ALSACIA	Frenada Brusca
PFZ480	2014-11-27	15:57:40	2	Calido	Urbana	28	Colombia, San Andres y Providencia, San Andres	Violacion Carga
PFZ481	2014-11-27	15:53:06	2	Calido	Urbana	28	Colombia, San Andres y Providencia, San Andres	Parada No Autorizada
SKR789	2014-11-27	15:55:08	1538	Templado	Urbana	22	Colombia , Antioquia , Medellin, TERMINAL DE TRANSPORTES TERPEL TERMINAL	Parada No Autorizada
SMB655	2014-11-27	15:53:30	5	Calido	Urbana	28	Colombia , Atlantico , Barranquilla	Violacion Carga
SMW477	2014-11-27	15:56:36	1001	Calido	Urbana	23	Colombia , Valle del Cauca , Palmira	Exc. Velocidad

Figura 35: Módulo Geográfico del SIG



Finalmente, es necesario realizar un análisis, de los resultados obtenidos de la integración del modelo con el SIG, el cual se realizó en el siguiente capítulo de este trabajo.

4. Análisis de resultados

Los resultados obtenidos posterior a la integración del modelo predictivo con el SIG, permiten conocer cuál es el evento de riesgo más probable a ocurrir en cada uno de los vehículos de la flota. Sin embargo, es necesario analizar casos específicos de eventos predichos por el modelo y que son visualizados en el reporte del SIG.

En esta sección, se analizarán algunos casos puntuales observando sus resultados y lo acertado del modelo.

- Caso 1:

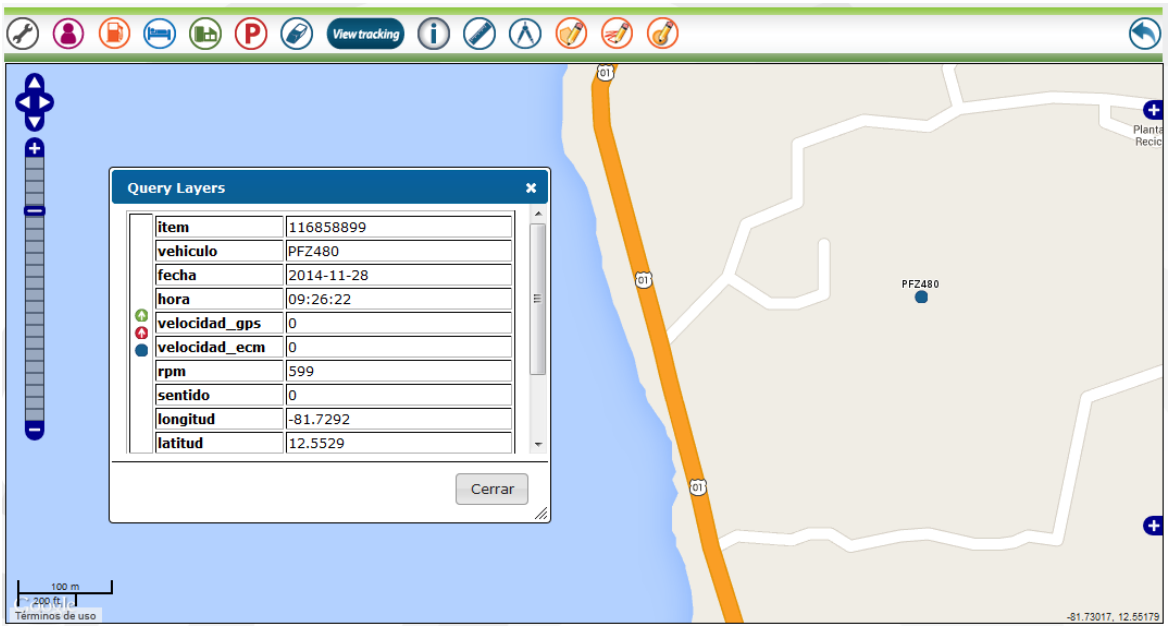
Figura 36: Caso 1 – Parada No Autorizada

Predicción de eventos en Tiempo Real

Vehículo	Fecha Ultimo Reporte	Hora Ultimo Reporte	Altura S.N.M. (Mts)	Clima	Tipo Zona	Temp. Promedio (°C)	Última Ubicación	Evento Probable
KUN690	2014-11-28	09:24:25	350	Calido	Rural	27	Colombia , Casanare , Villanueva	Exc. Temperatura Motor
OCK595	2014-11-28	00:52:06	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, VILLA ALSACIA	Frenada Brusca
PFZ480	2014-11-28	09:26:22	2	Calido	Urbana	28	Colombia, San Andres y Providencia, San Andres	Parada No Autorizada
PFZ481	2014-11-28	09:26:16	2	Calido	Urbana	28	Colombia, San Andres y Providencia, San Andres	Parada No Autorizada
SKR789	2014-11-28	09:25:11	1538	Templado	Urbana	22	Colombia , Antioquia , Medellin, OLEODUCTO	Parada No Autorizada
SMB655	2014-11-28	09:05:23	5	Calido	Urbana	28	Colombia , Atlantico , Barranquilla	Parada No Autorizada

Se observa uno de los vehículos de la flota, en donde según el modelo, el evento más probable a ocurrir, según las características de su último reporte es una “Parada No Autorizada” (evento de riesgo descrito en la sección 3.1.7). Dando clic sobre el hipervínculo que nos permite ubicar esta posición en el mapa (Figura 37), se visualiza que el vehículo efectivamente, se encuentra detenido (con su velocidad en 0) en una zona del mapa, en donde no existe ningún lugar autorizado para realizar una parada. Así que fue posible determinar que el evento ocurrió y el modelo obtuvo un resultado acertado.

Figura 37: Caso 1 - Mapa



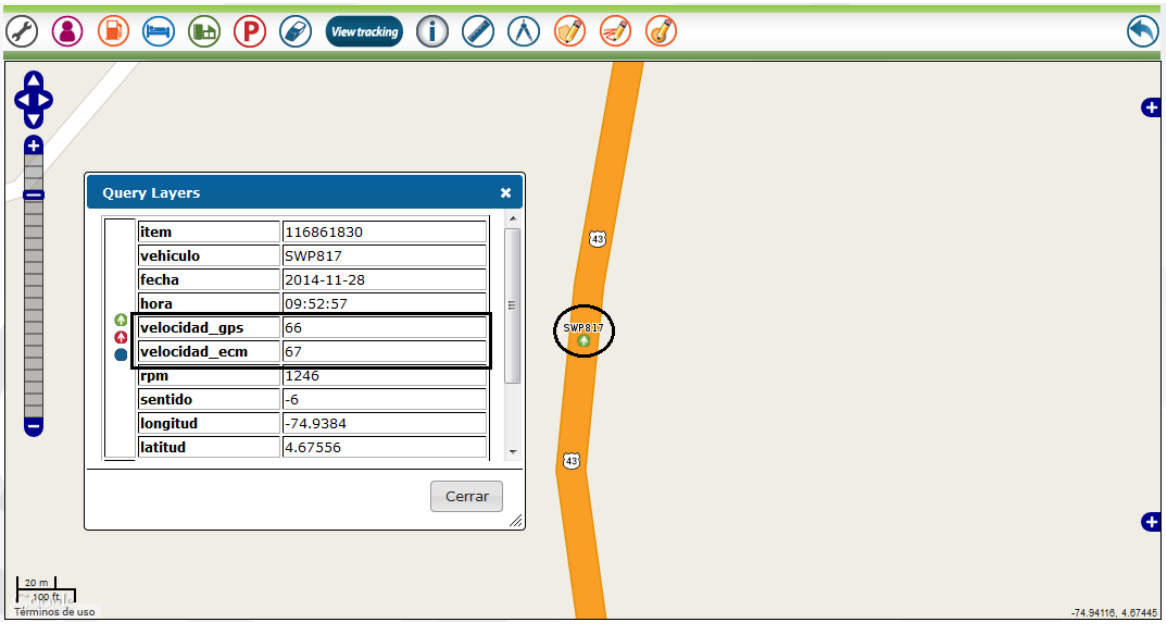
- Caso 2

Figura 38: Caso 2 – Exceso de Velocidad

SWN157	2014-11-28	09:46:10	1895	Templado	Rural	12	Colombia , Quindio , Salento	Exc. Temperatura Motor
SWO253	2014-11-28	09:58:41	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, CIUDAD SALITRE SUR-ORIENTAL	Exc. Temperatura Motor
SWP816	2014-11-28	09:56:29	213	Calido	Urbana	28	Colombia , Sucre , Sincelejo	Exc. Velocidad
SWP817	2014-11-28	09:52:57	348	Calido	Urbana	26	Colombia , Tolima , Venadillo	Exc. Velocidad
SWP820	2014-11-28	09:56:49	2170	Frio	Urbana	16	Colombia , Caldas , Aguadas	Exc. Velocidad Lluvia

En este segundo caso, se observó un vehículo, que bajo ciertas características, el modelo predijo un evento de “Exceso de Velocidad (en tiempo seco)” en la ciudad de Venadillo, Tolima. Ubicando este vehículo en el módulo geográfico y obteniendo información más detallada sobre su reporte, en la figura 39, se pudo valorar el riesgo que existe en la generación de este evento, ya que, siendo el límite de velocidad para esta flota, de 70 Km/H, el vehículo reporta velocidades cercanas a dicho valor (67 Km/H).

Figura 39: Caso 2 - Mapa



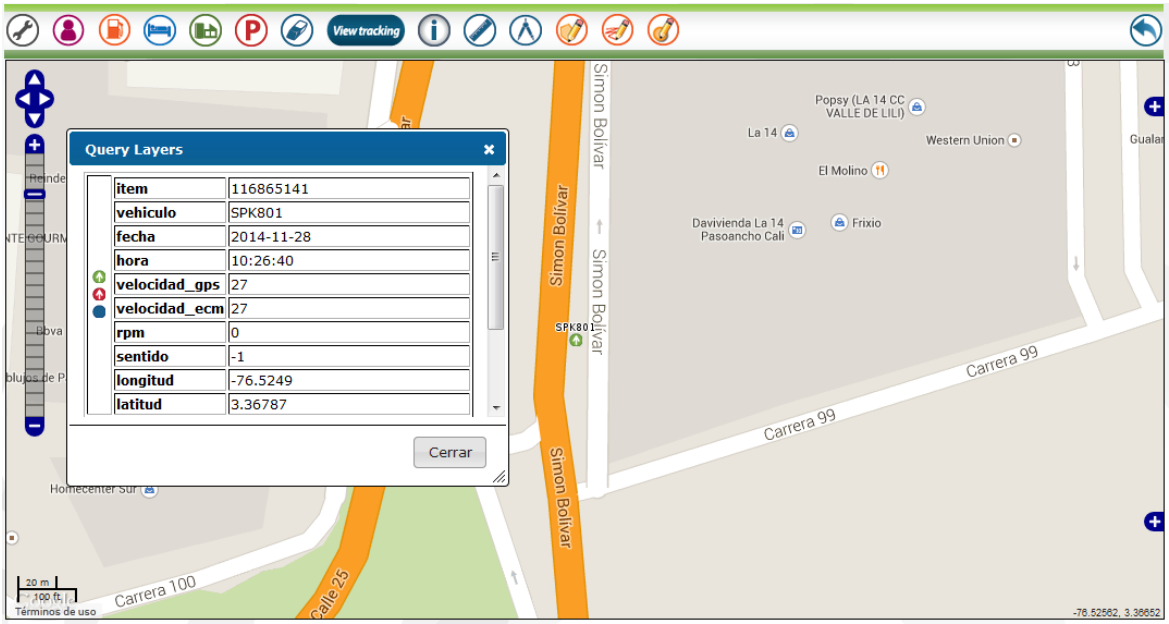
- Caso 3

Figura 40: Caso 3 – Violación de Carga

SOP583	2014-11-28	10:07:19	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, LA ESPERANZA	Exc. Temperatura Motor
SOP602	2014-11-28	09:51:25	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, PUENTE ARANDA, Parqueadero Planta Llenadero	Parada No Autorizada
SPK801	2014-11-28	10:26:40	995	Calido	Urbana	23	Colombia , Valle del Cauca , Cali, LILI	Violacion Carga
SPM986	2014-11-28	10:21:08	1538	Templado	Urbana	22	Colombia , Antioquia , Medellin, condor san javier	Parada No Autorizada
SPO022	2014-11-28	01:50:08	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, EL TINTAL	Parada No Autorizada

El caso 3 muestra una predicción del evento de riesgo “Violación de carga”, la cual implica la manipulación y/o apertura de la carga que transporta el vehículo. Primero, se debe tener en cuenta si el vehículo en cuestión, tiene o no un AVL que permita realizar dicho control. En este caso, aplica, pues este vehículo si cuenta con un control de la carga en el momento que realizó capturo este dato y se generó el reporte. En la figura 41, se muestra la ubicación del vehículo en el mapa en el momento de la predicción y se observa, que no se encuentra cerca de ninguna ubicación autorizada para la manipulación de la carga.

Figura 41: Caso 3 - Mapa



- Caso 4

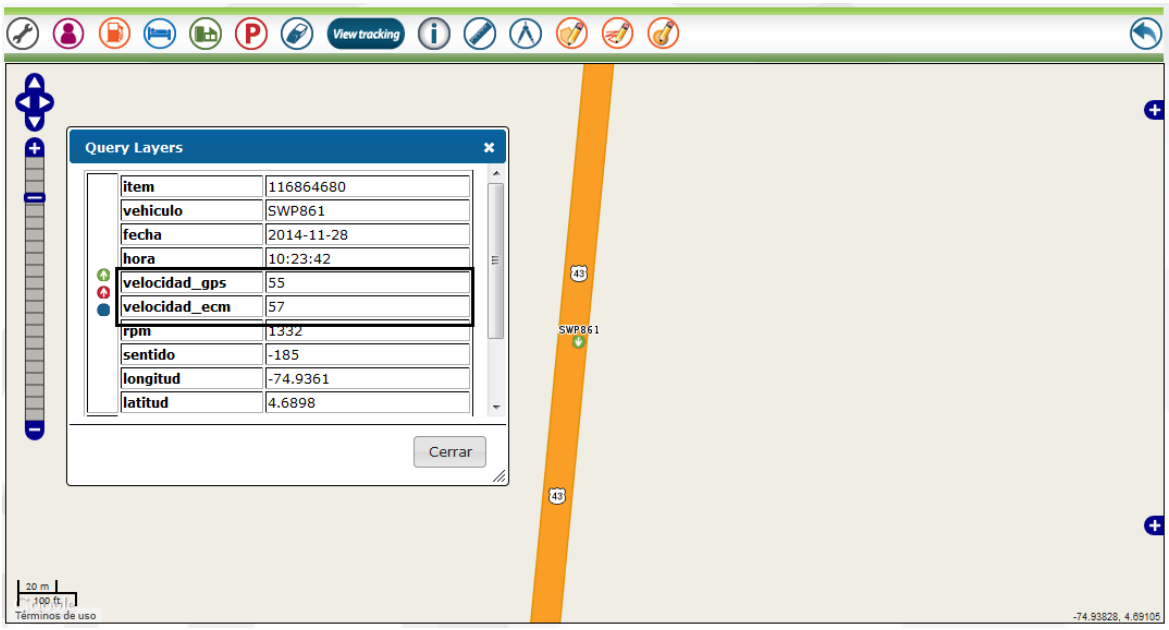
Figura 42: Caso 4 – Exceso de Velocidad en Lluvia

SWP816	2014-11-28	10:42:43	55	Calido	Urbana	28	Colombia , Sucre , San Onofre	Exc. Velocidad
SWP817	2014-11-28	10:50:57	500	Calido	Rural	28	Colombia , Tolima , Armero (Guayabal)	Exc. Temperatura Motor
SWP820	2014-11-28	10:51:19	1800	Templado	Rural	19	Colombia , Antioquia , Santa Barbara	Exc. Temperatura Motor
SWP861	2014-11-28	10:23:42	348	Calido	Urbana	26	Colombia , Tolima , Venadillo	Exc. Velocidad Lluvia
SWP862	2014-11-28	10:22:37	980	Calido	Urbana	23	Colombia , Valle del Cauca , San Pedro, E/S San Pedro	Parada No Autorizada

25 Por Página Página 2 de 5

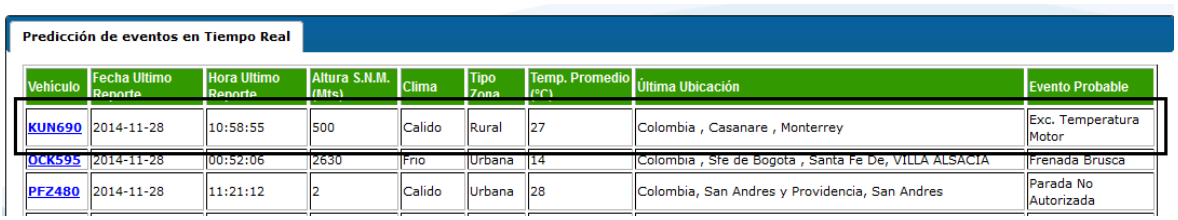
Para que se genere un evento de riesgo “Exceso de Velocidad en Lluvia”, no es necesario que el vehículo supere alcance los 70 Km/H como en el evento de velocidad en tiempo seco. El límite de velocidad cuando el AVL detecta lluvia, se reduce a flota a 50 Km/H, por el alto riesgo que se puede producir al conducir a altas velocidades bajo estas circunstancias. Verificando la veracidad de esta predicción, en el mapa se puede observar que este vehículo ya superaba dicho límite y estaba próximo a generar el evento de riesgo.

Figura 43: Caso 4 - Mapa



- Caso 5

Figura 44: Caso 5 - Exceso Temperatura de Motor



Predicción de eventos en Tiempo Real

Vehiculo	Fecha Ultimo Reporte	Hora Ultimo Reporte	Altura S.N.M. (Mts)	Clima	Tipo Zona	Temp. Promedio (°C)	Última Ubicación	Evento Probable
KUN690	2014-11-28	10:58:55	500	Calido	Rural	27	Colombia , Casanare , Monterrey	Exc. Temperatura Motor
OCK595	2014-11-28	00:52:06	2630	Frio	Urbana	14	Colombia , Sfe de Bogota , Santa Fe De, VILLA ALSACIA	Frenada Brusca
PFZ480	2014-11-28	11:21:12	2	Calido	Urbana	28	Colombia, San Andres y Providencia, San Andres	Parada No Autorizada

El caso 5, refleja un problema que afecta el modelo. Dado que los datos en tiempo real, pueden contener ruido o datos fuera rango, esto se ve reflejado en posibles errores que se pueden producir en la predicción de un evento de riesgo.

En este caso particular, observando sus registros en la base de datos, se observa que la temperatura de motor, presenta el dato de 255 °C. Este dato, lo reporta el AVL, cuando no ha sido posible comunicarse con el modulo electrónico del vehículo y no se pudo capturar el dato real.

Estos problemas, que afectan la clasificación de eventos de riesgo, se puede presentar también en el atributo que determina el valor de las RPM del vehículo.

5. Conclusiones y recomendaciones

5.1 Conclusiones

La base de datos de un SIG, puede contener muchas dimensiones y estas a su vez, atributos que nos pueden proporcionar más información para hacer más acertado un modelo predictivo, sin embargo, así mismo, puede contener mucho ruido y datos fuera de rango, que pueden entorpecer la construcción de este, dado que las herramientas de medida (AVL, GPS, ECU del motor y demás) con las que se capturan estos datos, pueden fallar o capturar datos equivocados. Por lo anterior, como resultado de este trabajo de grado se pudo demostrar que llevar a cabo el análisis descriptivo y pre-procesamiento del conjunto de datos de un SIG es un paso imprescindible para la construcción de un modelo predictivo robusto, ya que de estos depende en gran medida su construcción, para que permita clasificar los eventos de riesgo de una flota de una con una precisión aceptable.

El uso de varios algoritmos de clasificación para la construcción del modelo, fue acertada, puesto que permitió obtener distintas opciones que pudieron ser comparadas y medidas unas con otras, no únicamente evaluando sus resultados, sino también su costo computacional y tiempos de ejecución. Fue importante, contar con distintos conjuntos de datos o muestras de datos, para realizar una correcta validación de los modelos.

Finalmente, con los resultados obtenidos, se pudieron conocer aquellos atributos que se consideran importantes, para la clasificación de eventos de riesgo en un SIG de monitoreo de vehículos, sin dejar por fuera, lo más trascendentales, que son los que proporcionan información espacial o geográfica (Estado, Ciudad, clima, altura sobre el nivel del mar, entro otros), además de re-afirmar la posibilidad existente, de realizar una integración de un modelo de este tipo, con un SIG, para lograr una interpretación más sencilla de sus resultados, a manera de reportes y/o advertencias en el sistema

5.2 Recomendaciones

El uso de software para la construcción de un modelo predictivo (Weka, KNime, RapidMiner, entre otros) podría ser opcional, pues no se descarta la idea de realizar dicha implementación en el mismo lenguaje de programación en el que este desarrollado el SIG, para que los resultados de este, puedan ser leídos de manera nativa y los datos sean mucho más sencillo de manipular, aunque al mismo tiempo, se debería tener en cuenta su costo computacional y tiempos de respuesta.

En este trabajo, se identificaron eventos de riesgo, que dependen mucho de la operación del conductor y estado del vehículo, pero estos no son los únicos riesgos a los que están expuestos. Como trabajo futuro, podrían plantearse otro de tipo de eventos, que sean externos a la operación, ya sean sociales o naturales, y de esta manera conocer, si para este tipo de eventos, influyen de la misma manera los atributos aquí identificados.

Como recomendación final para trabajo futuro, se propone el uso de otras herramientas para realizar las predicciones de eventos de riesgo en los vehículos, ya sea por medio de otras técnicas de minería de datos distintas a la clasificación, haciendo uso de modelos matemáticos, algoritmos de toma de decisiones o análisis estadísticos a la implementación de modelos estrella de bodegas de datos.

A. Anexo: Implementaciones

Implementación de la ETL. Se adjunta en CD, código fuente y archivo Ejecutable.

Integración con el SIG. Se adjunta en CD, código fuente de scripts.

B. Anexo: Conjuntos de datos

Se adjunta en CD, archivos del conjunto de datos, en su versión nominal y binarizada.

Bibliografía

Durduran, S. S. (2010). A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform. *Expert Systems with Applications*, 8.

Jiejun Huang, Y. Y. (2007). Construction and Application of Bayesian Network Model for Spatial Data Mining. *2007 IEEE International Conference on Control and Automation*, 2802-2805.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* John Wiley & Sons.

Markus Deubleina, M. S. (2013). Prediction of road accidents: A Bayesian hierarchical approach. *Accident Analysis and Prevention*, 18.

O. Aloquili, A. E.-A. (2008). Automatic vehicle location tracking system based on GIS environment. *IET Software*, 9.

Pocut Viqarunnisa, H. L. (2011). Generic Data Model Pattern for Data Warehouse. *International Conference on Electrical Engineering and Informatics*, 8.

Soyoung Jung, K. J. (2014). Contributing factors to vehicle to vehicle crash frequency and severity under rainfall. *Journal of Safety Research*, 10.

Wang Jinlin, C. X. (2008). Application of Spatial Data Mining in Accident Analysis System. *Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on (Volume 1)*, 4.

Wang Peng, M. L. (2009). Research on Logistics Oriented Spatial Data Mining Techniques. *Management and Service Science, 2009. MASS '09. International Conference on*, 1-4.

Xiaofang Zhou, Y. Z. (2000). On Spatial Information Retrieval and Database Generalization. *Digital Libraries: Research and Practice, 2000 Kyoto, International Conference on.*, 328-334.

Yang-Kun Ou, Y.-C. L.-Y. (2013). Risk prediction model for drivers' in-vehicle activities – Application of task analysis and back-propagation neural network. *Transportation Research Part F*, 83-93.