

Ing. Pablo F. Ordoñez Ordoñez, Mg. Sc.

Encargado de la Gestión de la CIS/C-UNL

Loja, 25 de octubre del 2021

De nuestra consideración:

De conformidad a lo preceptuado en los Arts: 123 y 128 del Reglamento de Régimen Académico de la Universidad Nacional de Loja, régimen 2009 y con la finalidad de optar el grado en Ingeniería en Sistemas de la Facultad de la Energía, las Industrias y los Recursos Naturales No Renovables, adjuntamos un ejemplar del Proyecto de Trabajo de Titulación, el mismo que versa sobre **Minería de datos en la accidentabilidad vehicular en la zona urbana del Cantón Loja** a efecto de que el responsable de la línea de investigación, emita, para su pertinencia respectiva, un informe sobre la coherencia, estructura y pertinencia académica del proyecto; y, si está de acuerdo a su trascendencia, que amerite ser investigado.

En la seguridad de que su autoridad, se digne ordenar el trámite legal y reglamentario a la presente petición, anticipo mis más sinceros agradecimientos.

Patricio Bolívar

Benítez Lanche

CI: 1105665044

C/ Calle Calicuchima

110102

Ecuador

☎ 0992182051

✉ patricio.benitez@unl.edu.ec



f: Patricio Bolívar Benítez Lanche

CI: 1105665044

*Anexos: Proyecto de Trabajo de Titulación - Patricio Benítez,
periodo Abril-Septiembre 2021*



Universidad
Nacional
de Loja

Facultad de Energía, las Industrias y los Recursos Naturales no Renovables

CARRERA DE INGENIERÍA EN SISTEMAS

MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA

DATA MINING IN VEHICLE ACCIDENT RATES IN THE URBAN
AREA OF CANTÓN LOJA

Línea de investigación: Energía, Industrias y Recursos Naturales no
Renovables

PROYECTO DE TITULACIÓN EN
INGENIERÍA EN SISTEMAS.

Autor:

Patricio Bolívar Benítez Lanche , patricio.benitez@unl.edu.ec

Tutor:

- Pablo F. Ordoñez-Ordoñez, Mg.Sc.

Cotutor:

- Edison L. Coronel-Romero, Mg.Sc.



LOJA - ECUADOR
2021

Certificación del Tutor

Pablo F. Ordoñez-Ordoñez, Mg.Sc. en calidad de tutor voluntario del proyecto de trabajo de titulación, certifico la tutela a **Patricio Bolívar Benítez Lanche**, con el tema **MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA**, quien ha cumplido con todas las observaciones requeridas. Es todo cuanto puedo decir en honor a la verdad, facultando al interesado hacer uso de la presente, así como el trámite de pertinencia del presente proyecto.

Loja, 28 de septiembre de 2021



Firmado digitalmente por
PABLO FERNANDO
ORDONEZ ORDONEZ
Razón: He revisado este
documento
Ubicación: Loja

Atentamente,

Pablo F. Ordoñez-Ordoñez, Mg.Sc.
TUTOR



Firmado electrónicamente por:
**EDISON LEONARDO
CORONEL ROMERO**

Edison L. Coronel-Romero, Mg.Sc.
COTUTOR

Certificación de Autoría del Proyecto

Yo, Patricio Bolívar Benítez Lanche , estudiante de la Universidad Nacional de Loja, declaro en forma libre y voluntaria que el presente proyecto de trabajo de titulación que versa sobre **MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA**, así como la expresiones vertidas en la misma son autoría del compareciente, quien ha realizado en base a recopilación bibliográfica primaria y secundaria. En consecuencia asumo la responsabilidad de la originalidad de la misma y el cuidado al remitirse a las fuentes bibliográficas respectivas para fundamentar el contenido expuesto.



Atentamente,
Patricio Bolívar Benítez Lanche

Índice general

1. Problemática	1
1.1. Situación Problemática	1
1.2. Problema de Investigación	4
2. Justificación	5
3. Objetivos	8
3.1. General	8
3.2. Específicos	8
4. Alcance	9
5. Marco Teórico	11
5.1. Antecedentes	11
5.1.1. Conceptos y categorización de accidentes de tránsito	11
5.1.2. Causas de los accidentes de tránsito	12
5.1.3. Categorización de los accidentes de tránsito	12
5.1.4. Tipos de accidentes de tránsito	13
5.1.5. Minería de datos	13
5.1.6. Base de datos	14
5.1.7. Técnicas de minería de datos	15
5.1.8. Análisis de Discriminación lineal	16
5.1.9. Árboles de decisión	17
5.1.10. Metodología usadas en la minería de datos	20
5.1.11. KDD ó Knowledge Discovery in Databases	21
5.1.12. CRISP-DM (Cross Industry Standard Processfor Data Mining)	22
5.1.13. Proceso SEMMA	25
5.1.14. Trabajos relacionados	26
5.1.15. Organismos de Control de Tránsito en el Cantón Loja	29
6. Metodología	32
7. Cronograma	34
8. Presupuesto	36
8.1. Recursos Humanos	36

8.2. Recursos Software, Hardware y TICs	37
8.3. Insumos	37
8.4. Recursos Totales	38
Bibliografía	39
A. Anexo I	42
B. Revisión sistemática de literatura	49
C. Anexo II	51
D. Certificado de acceso a los datos	55

Índice de figuras

5.1. Tipos de accidentes de tránsito	12
5.2. Modelo de base de datos.	15
5.3. Vectores del Análisis de discriminación lineal	17
5.4. Matriz de confusión	18
5.5. Jerarquía del conocimiento	22
5.6. Proceso de Descubrimiento del Conocimiento (KDD)	23
5.7. Metodología CRISP-DM	24
6.1. Metodología del PTT.	33
7.1. Cronograma del PTT.	35

Índice de tablas

5.1. Trabajos relacionados a Minería de datos aplicada a siniestros de tránsito.	27
5.2. Trabajos relacionados a Minería de datos aplicada a siniestros de tránsito.	28
8.1. Recursos humanos empleados para el desarrollo del proyecto.	36
8.2. Recursos Software, Hardware, TICs a utilizar para el desarrollo del proyecto.	37
8.3. Insumos a utilizar para el desarrollo del proyecto.	37
8.4. Presupuesto total para la elaboración del proyecto.	38

**MINERÍA DE DATOS EN LA
ACCIDENTABILIDAD
VEHICULAR EN LA ZONA
URBANA DEL CANTÓN LOJA**
Linea de investigación: Energía, Industrias y
Recursos Naturales no Renovables

1. Problemática

1.1. Situación Problemática

Actualmente los accidentes de tránsito representan una de las causas mas comunes de muerte en países en desarrollo, siendo estos sucesos imprevistos producidos por la participación de un vehículo o más en las vías o carreteras, implicando pérdidas económicas incluso pérdidas humanas, así como daños a bienes públicos y privados; por ende, ante estos sucesos se genera un gran número de registros que pueden contener información valiosa que aporte en mejores decisiones y discusión sobre los accidentes de tránsito de manera fundamentada[1].

En la actualidad en Ecuador los accidentes de tránsito cada vez van en aumento, considerando que durante el año 2020 se registraron 16.972 accidentes de tránsito, mientras que durante los seis meses del año 2021 se han registrado aproximadamente 26.322 accidentes de tránsito, evidenciando un incremento del 55% de siniestros viales por diferentes causas más comunes como Arrollamientos, atropellos, caída de pasajero, choque frontal, choque lateral, choque posterior, colisión, pérdida de carril, pérdida de pista, rozamientos, volcamientos, atípicos entre otros, según registra el informe estadístico presentado por Agencia Nacional de Tránsito (ANT)[2]; hasta la elaboración de este documento tan solo en la Provincia de Loja el número de accidentes asciende a 552 accidentes en lo que va del año, en donde uno de los cantones más afectados es el Cantón Loja debido a que cuenta con un mayor flujo vehicular. En el Cantón Loja las competencias de transito se encuentran asignadas al Gobierno Autónomo Descentralizado Municipal de Loja según lo indica el Consejo Nacional de Competencias mediante resolución 006-CNC-2012 publicada en el Registro Oficial el 29 de mayo de

2012, quien transfiere a los Gobiernos Autónomos Descentralizados, Metropolitanos y Municipales del país la competencia exclusiva de planificar, regular y controlar el transporte terrestre, el tránsito y seguridad vial; ubicándolo además al Gobierno Municipal de Loja, dentro del modelo de gestión “A”^[3], organismo que cuenta con un departamento específico denominado “Unidad de Control Operativa de Tránsito” (UCOT), quien es el único ente autorizado para cumplir las funciones de regulación y control de tránsito dentro de las vías que se encuentran en la jurisdicción del Cantón Loja.

Actualmente UCOT lleva el proceso de registro de siniestros de tránsito a través de un software de ofimática (Hoja de cálculo), mismos que son almacenados dentro del correo institucional otorgado por el GAD Municipal de Loja, además cuentan con una oficina de archivo en donde toda la documentación es almacenada de forma física mediante archivadores, en donde se encuentran clasificados por el mes y año del suceso, presentando algunas dificultades durante el proceso de búsqueda de partes por accidentes de tránsito, en donde muchas de las veces no se logra encontrar la documentación y conlleva a invertir mucho tiempo del personal humano de la institución a realizar indagaciones que toman varios días. Por tanto se entrevisto a la Ing. Victoria Torres Pereira, Jefe Operativo de la UCOT con respecto al registro actual de los partes por accidentes de tránsito y las funciones que cumple esta documentación, en donde se determina que son elaborados de forma manual a través de un documento digital y físico elaborado en el departamento de redacción de partes; y su importancia se basa en que son documentos necesarios para exponer la posible circunstancia en la que se produjo el accidente, utilizado para la investigación y estudio de los factores y causas que incidieron para el cometimiento del accidente de tránsito, además de ser necesario en instancias judiciales.(C. ANEXO II Pregunta 1-4). Además, en cuanto a la obtención de la información de siniestros de tránsito suscitados en la localidad, se indica que los agentes de tránsito una vez de que llega al lugar de los hechos, lo que hace es recabar la información, implicando que tome contacto directamente con los

afectados o con personas que estuvieron en el lugar de los hechos y brindan información, respecto a cómo se suscitó en este caso el accidente de tránsito; se considera que, desde que asume las competencias el Gobierno Autónomo Descentralizado de Loja, los agentes de tránsito, todo lo que se refiere dentro del cantón, están autorizados para recabar este tipo de información, considerando en este caso todo lo que respecta a direcciones, nombres de conductores, tipología del accidente, placas vehiculares, títulos habilitantes, identificación del ingreso vehicular, información del Agente que tomó procedimiento; factores climatológicos, y como se encontraba la vía al momento del incidente; toda esta información y se respalda en documentos que eleva a conocimiento de la autoridad competente (C. ANEXO II Pregunta 2-4). En cuanto al estudio y análisis de los lugares donde existe mayor número de accidentes de tránsito en el cantón se indica que la Unidad de Control Operativo de Tránsito, a su documentación es registrada en una base de datos, poseendo una matriz de un documento excel, registrada en la base de datos, almacenada dentro de la información institucional, la cual después semanalmente y de forma mensual, se emite un respaldo de información estadística, la cual es enviada hacia la administración municipal y hacia la dirección de la Agencia Nacional de Tránsito ubicada aquí en Loja. Se cuenta además con una oficina de archivo, en este espacio se almacenan cada una de las informaciones como respaldo de los procedimientos que toma nuestro personal uniformado. (C. ANEXO II Pregunta 3-4). También se indagó referente a la posibilidad de existir una base de datos digital en donde toda la información se almacene en la Nube capaz de existir un adecuado respaldo de información, manifestando que no poseen este respaldo de base de datos en la nube, más bien los tiene registrados en documentos de ofimática (excel) en las computadoras como soporte son las unidades de almacenamiento, con la finalidad sobre todo a dar la atención y administración de la documentación, contemplando las leyes así lo requieren (C. ANEXO II Pregunta 4-4).

En resumen, en esta institución gubernamental actualmente no cuenta con un estudio adecuado para el análisis de datos de accidentes de tránsito, presentando inconvenientes

al momento de identificar cuáles son las principales causas por las que se producen los accidentes lo que produce inconvenientes al momento de identificar en que sectores existen mayor flujo de movilidad vehicular y con que frecuencia se producen los siniestros de tránsito, lo que conlleva al desconocimiento de las zonas de alto riesgo vial exponiendo tanto a peatones, conductores, así como a la propiedad pública y privada.

Ante esta situación, es posible aplicar sobre los datos técnicas de Minería de Datos, entre ellos los árboles de decisión, para predecir patrones y resolver problemas de distribución no lineal haciendo uso de variables con valores discretos o continuos[4], con el propósito de conocer el comportamiento de accidentabilidad vehicular, en donde se permita identificar los lugares con alto riesgo de accidentes de tránsito con el objetivo de fortalecer la seguridad vial, de tal forma que se brinda un apoyo a la toma de decisiones en el Cantón Loja, enfocada específicamente en la zona urbana de la ciudad.

1.2. Problema de Investigación

Dentro del enfoque de este proyecto se delimita a la necesidad de análisis de datos para identificar los focos de accidentabilidad dentro de la zona urbana del cantón Loja, debido a que en varios puntos de esta ciudad los siniestros de tránsito son frecuentes creando sectores o zonas de alto riesgo, exponiendo tanto a peatones como a conductores, propiedad pública y privada.

Por tanto, ¿se puede establecer el patrón de accidentabilidad vehicular con la aplicación de Minería de datos o Data Mining en la zona urbana del cantón Loja?

2. Justificación

La importancia del presente tema de investigación se enfoca en el problema de los focos de accidentabilidad vehicular de la zona urbana del cantón Loja, que esto preocupa tanto a conductores, peatones, responsables de la propiedad pública y privada, así como entidades de control vehicular, lo que ha conllevado a un desconocimiento de las zonas donde se producen o suscitan con mayor frecuencias los siniestros viales, por lo que debido a ello no se puede establecer un patrón para tomar las medidas preventivas y evitar un suceso que exponga las integridad de personas como bienes.

En la actualidad, la minería de datos se encarga de preparar, sondear y explorar los datos para sacar la información oculta y útil de ellos [5]. Con el estudio y análisis de la información nos abre una puerta para proporcionar una ayuda en la toma de decisiones y generar futuros patrones de comportamientos a partir de las bases de datos. De este modo los patrones generados por el data mining será de información única y verifica, que permitirá colaborar a las instituciones gubernamentales como el GAD Municipal de Loja, sin la necesidad de estar realizando procesos obsoletos de búsqueda de papel, brindando un conocimiento más óptimo sobre la hora y lugar de un accidente de tránsito, desarrollando focos de accidentabilidad minimizando la inversión en tiempo y brindando un dato más real.

Hoy en día existen un sinnúmero de estudios sobre modelos de Data Mining aplicados al sector empresarial; quienes tienen como finalidad extraer conocimiento de los datos para generar estrategias empresariales. Por ello, uno de los sectores en donde se ha vuelto indispensable realizar estudios de Minería de datos es en el sector del transporte terrestre, enfocado específicamente en la accidentabilidad vehicular, debido a que ante la disponibilidad de grandes volúmenes de datos, se aplican técnicas para descubrir

patrones, algoritmos y otras técnicas avanzadas para predecir los comportamientos futuros.

El desarrollo de este estudio esta orientado hacia la necesidad de la Unidad de Control Operativa de Tránsito (UCOT), por la falta de un análisis exhaustivo de datos para determinar la focalización de zonas de mayor riesgo de siniestros viales, deficiente análisis de la toma de decisiones y futuros patrones de comportamientos viales, en donde se puede poner en riesgo la integridad de las personas, asi como la poca implementación de infraestructura de señalización tanto horizontal y vertical en seguridad vial; de manera que este estudio con los permisos respectivos del manejo de la data operacional de la institución (D. Anexo Certificado de acceso a datos) será un referente sobre la aplicación de minería de datos, que si bien existen estudios en otras ciudades, aún no se ha evidenciado un estudio dirigido a la problemática de la ciudad de Loja, obteniendo los sectores críticos más propensos a derivar en accidentes de tránsito, patrones de accidentología vial actualizados, tomas de decisiones ideales fomentando la seguridad vial, aplicando acciones eficientes para reducir el índice de siniestros viales que se produce a diario en nuestro sector.

Este proyecto de titulación se justifica en los siguientes aspectos: académico, tecnológico y social; establecido en la línea de investigación priorizada por: el SENESCYT referente a “TIC para la educación e inclusión social”[?], además por medio de la Carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja con su vinculación con la Sociedad mediante aportes de investigación tecnológica, bajo los principios de calidad, planificación, gestión, desarrollo, responsabilidad, honestidad e interculturalidad. Priorizando las zona urbana del Cantón Loja, fomentando el control y seguridad vial al crear un estudio que esté al alcance y servicio de las instituciones gubernamentales y de la ciudadanía. [6]

Además, se aspira cubrir las demanda de TIC con la finalidad de brindar un aporte a las instituciones gubernamentales locales con la finalidad de velar por la integridad de las personas como se menciona en el Libro Blanco de Líneas de Investigación, Desarrollo

e Innovación y Transferencia de Conocimiento en TIC. Exactamente en las Línea de Investigación de Ciudades Inteligentes, Sostenibles e Inclusivas para la temáticas de Servicios Inteligentes donde se centra en mejorar la calidad de vida de la población. [7]

3. Objetivos

3.1. General

Analizar la accidentabilidad vehicular mediante la aplicación del modelo predictivo de árboles de decisión en la zona urbana del Cantón Loja en el periodo 2019-2020.

Analyze vehicle accident rates through the application of the decision tree predictive model in the urban area of Canton Loja in the 2019-2020 period.

3.2. Específicos

- Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020.

Analyze the data repositories regarding traffic accidents registered at UCOT in the year 2019-2020.

- Implementar el modelo de árboles de decisión para establecer las principales causas sobre los accidentes de tránsito.

Implement the decision tree model to establish the main causes of traffic accidents.

- Evaluar la exactitud del modelo predictivo de árboles de decisión para determinar las incidencias que ocasionan los siniestros de tránsito en la zona urbana del Cantón Loja.

Evaluate the accuracy of the decision tree predictive model to determine the incidents that cause traffic accidents in the urban area of Cantón Loja.

4. Alcance

El presente proyecto de titulación abarca la elaboración de un análisis de accidentabilidad vehicular en la zona urbana del cantón Loja a través del desarrollo y modelado, en los intervalos establecidos en el cronograma y respetando los recursos económicos determinados en la sección de presupuesto.

A continuación, se detalla el alcance por las etapas del Proyecto de Titulación:

Etapas 1: Obtener la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos.

- Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020.
 - a) Establecer las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja.
 - b) Obtener la base de datos siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019-2020.
 - c) Desarrollar la transformación y limpieza de la base de datos obtenida.

Etapas 2: Implementación del modelo de árboles de decisión.

- Desarrollar el análisis exploratorio de datos empenado el modelo de árboles de decisión.
 - a) Identificar las variables que influyen en el cometimiento del accidente de tránsito.

- b) Entrenar el modelo de árboles de decisión, con la información de accidentes de tránsito ocurridos en la zona urbana del cantón Loja, periodo 2019 - 2020.

Etapas 3: Evaluación de la Técnica de minería de datos.

- Evaluar del modelo de árboles de decisión.
 - a) Analizar los resultados obtenidos del modelo entrenado.
 - b) Evidenciar el funcionamiento del modelo con datos actuales.
 - c) Elaborar el documento final.
-

5. Marco Teórico

5.1. Antecedentes

5.1.1. Conceptos y categorización de accidentes de tránsito

Un accidente de tránsito es un suceso eventual o acción voluntaria que da como resultado pérdidas económicas, lesiones e incluso pérdidas humanas, a cuya ocurrencia contribuye por lo menos a un vehículo de tracción mecánica, animal o humana en circulación por una vía; mientras que, el sitio de suceso son vías o lugares públicos o privados abiertos al tránsito vehicular y peatonal. Además, este suceso eventual e imprevisto interactúan tres factores básicos conocidos como “trilogía vial”, que se interrelacionan y lo cual la seguridad vial depende de la correcta armonización de los mismos[8].

La trilogía vial se encuentra considerada por tres factores fundamentales: humano, vehículo y vial-ambiental.

1. Factor Humano: Interactúan en función de percepción de la información del entorno, su capacidad de procesar dicha información y emitir respuestas acertadas y oportunas.
2. Factor Vehículo: considerado el vehículo con todos sus mecanismos y componentes
3. Factor Vial - Ambiental: se considera a la vía y las condiciones ambientales.

5.1.2. Causas de los accidentes de tránsito

Existen distintas causas que pueden producir un accidente de tránsito; por ende, existen dos grupos importantes que son: causa basal y causa concurrente. Por lo tanto, al hablar de causa basal se refiere a la acción necesaria y suficiente que interviene en forma directa en la producción de un accidente de tránsito y sin la cual no se hubiera producido el mismo, mientras que causa concurrente señala que son los factores que contribuyen a que se produzca el accidente de tránsito[8].

5.1.3. Categorización de los accidentes de tránsito

Existen muchas formas de categorizar los accidentes de tránsito dependiendo de cada autor, en este caso lo vamos a realizar de acuerdo a dos factores relevantes, el factor humano y el factor vehículo[5.1]. Las definiciones de esta categorización fueron redactadas en base al Reglamento de la Ley Orgánica de Transporte[8].

Factor	Clasificación	Subclasificación
a) Humano	Atropello Arrollamiento Caída de pasajero	
b) Vehículo	Choque	Choque posterior o por alcance Choque frontal longitudinal Choque frontal excéntrico Choque lateral angular Choque lateral perpendicular
	Estrellamiento Colisión Pérdida de pista	
	Volcamiento	Volcamiento lateral Volcamiento longitudinal
	Roce	Roce positivo Roce negativo
	Rozamiento	

Figura 5.1: Tipos de accidentes de tránsito

5.1.4. Tipos de accidentes de tránsito

5.1.5. Minería de datos

La minería de datos consiste en la extracción de información sensible que reside de manera implícita en los datos. Dicha información es previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta de ellos, de modo que un pequeño hallazgo, una pequeña relación que se descubre, puede ser un hecho de gran impacto en las ventas de una compañía. También se denomina a la minería de datos como un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio, hecho que permite nutrirse de todas ellas sin prejuicios. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones [9].

La gran aplicación de la minería de datos hace que sea posible incorporar técnicas de otros campos de estudio, como por ejemplo la estadística, sistemas de bases de datos, machine learning, reconocimiento de patrones, entre otros. Un proceso típico de minería de datos consta de los siguientes pasos generales:

- Selección del conjunto de datos, tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
- Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

- Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como pre-procesamiento de los datos.
- Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.
- Extracción de conocimiento, mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.
- Interpretación y evaluación de datos, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias [4]

5.1.6. Base de datos

Un sistema de base de datos es una herramienta software para almacenar de forma ordenada los datos o registros. Es posible considerar a la propia base de datos como una especie de armario electrónico para archivar; es decir, es un depósito o contenedor de una colección de archivos de datos computarizados. Los usuarios del sistema pueden realizar una variedad de operaciones sobre dichos archivos [8], por ejemplo:

- Agregar nuevos archivos vacíos a la base de datos.
 - Insertar datos dentro de los archivos existentes.
 - Recuperar datos de los archivos existentes.
-

- Modificar datos en archivos existentes.
- Eliminar datos de los archivos existentes
- Eliminar archivos existentes de la base de datos

En la figura 5.2 se presenta una muestra de una base de datos de las estadísticas mensuales de accidentes de tránsito ocurridos en el año 2019 en el cantón Loja, cuyo certificado se encuentra anexado a este proyecto (D. Certificado a acceso a los datos).

#Siniestro	Fecha	Hora	Latitud	Longitud	Provincia	Cantón	Parroquia	Zona	Dirección	#Fallecidos	#Lesionados	TipoDeSiniestro	CausaDelProbable
512	2019-08-11	23:10	-3.9764656	-79.204476	LOJA	LOJA	SUCRE	URBANA	CALLE JUAN J	0	0	CHOQUE FRONTAL EXCÉNTRICO	CONDUCE BAJO LA INFLUENCIA DE
2503	2019-11-01	09:23	-3.979637	-79.232238	LOJA	LOJA	SUCRE	URBANA	AV ISIDRO AY	0	1	ATROPELLO	CONDUCE BAJO LA INFLUENCIA DE

Figura 5.2: Modelo de base de datos.

5.1.7. Técnicas de minería de datos

Al describir las técnicas de minería de datos se nos presentan un sinnúmero de herramientas; sin embargo, consideramos las técnicas predictivas como Regresión, Métodos Bayesianos, y de clasificación como el análisis de discriminación lineal, árboles de decisión y redes neuronales. [4]

La regresión logística es un procedimiento cuantitativo que puede usarse como método descriptivo, donde la variable dependiente toma valores en un conjunto finito.

Para construir un modelo de regresión logística se requiere de

- Un conjunto de variables independientes o predictoras y una variable respuesta dicotómica.
- Los parámetros del modelo se calculan usando una estimación de máxima verosimilitud.
- No incluir variables innecesarias al modelo.
- Ninguna variable relevante debe ser excluida.
- La colinealidad es un problema como ocurre en la regresión lineal múltiple. [4]

El modelo de regresión logística [5.1] puede describirse como:

$$\log \frac{p}{1-p} = b_0 + b_1x_1 + \dots + b_nx_n \quad (5.1)$$

Donde, p es la probabilidad (riesgo) de que ocurra el evento de interés, x como las variables independientes, b como los coeficientes asociados a cada variable.

La regresión logística puede usarse para fines explicativos, donde se busca explicar la relación de una variable con otras; o para fines predictivos, que es cuando se estima el comportamiento de una variable en función de las otras variables que influyen su comportamiento.

5.1.8. Análisis de Discriminación lineal

El análisis discriminante es una técnica de clasificación cuyo objetivo es obtener una función capaz de clasificar a un nuevo individuo a partir del conocimiento de los valores de ciertas variables discriminadoras, dentro de grupos previamente conocidos.

El análisis de discriminación lineal corresponde a un análisis de regresión donde la variable dependiente es una variable categórica y tiene como categorías la etiqueta de cada uno de los grupos, y las variables independientes son continuas y determinan a qué grupos pertenecen los objetos [4] .

Una vez identificadas las funciones discriminantes a partir de las variables explicativas, similares a las ecuaciones de regresión lineal múltiple, se puede realizar la clasificación en dos grupos o clasificación en mas de dos grupos (análisis discriminante múltiple). El modelo de análisis discriminante múltiple [5.2] puede describirse como: Cada una de las funciones discriminantes D_i , se obtiene como función lineal de las k variables explicativas [4] .

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \dots + u_{ik}X_k \quad i = 1, \dots, G-1 \quad (5.2)$$

Donde, los $G-1$ ejes vienen definidos respectivamente por los vectores u_1, u_2, \dots, u_{G-1}

$$u_1 = \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1k} \end{pmatrix}, \quad u_2 = \begin{pmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2k} \end{pmatrix}, \dots, \quad u_{G-1} = \begin{pmatrix} u_{G-1,1} \\ u_{G-1,2} \\ \vdots \\ u_{G-1,k} \end{pmatrix}$$

Figura 5.3: Vectores del Análisis de discriminacion lineal

El número máximo de ejes discriminantes ($G - 1, k$) ($G =$ número de categorías). Se obtiene $G - 1$ ejes discriminantes si el número de variables explicativas es mayor o igual que $G - 1$.

5.1.9. Árboles de decisión

Los árboles de decisión son utilizados para representar y categorizar una serie de condiciones que ocurre de forma sucesiva para la resolución de problemas. Se pueden definir como un mapa de posibles resultados de una serie de decisiones relacionadas, que comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados posibles crea nodos adicionales que se ramifican en otras posibilidades [\[4\]](#).

Existe tres diferentes nodos: nodos de probabilidad, nodos de decisión y nodos terminales. Los nodos de probabilidad muestran las probabilidades de ciertos resultados. Un nodo de decisión muestra una decisión que se tomará y el nodo terminal muestra el resultado definitivo de una ruta de decisión.

Los elementos que componen la estructura de un árbol de decisión y adicionalmente los pasos para la elaboración de este:

Existen 4 tipos de nodos en un árbol de decisión.

1. Nodo raíz: inicio del árbol de decisión, se plantea la decisión entre alternativas con resultados desconocidos.
2. Nodos de decisión: caminos de acción que pueden ser elegidos por el tomador de

decisión después de haber analizado los resultados de decisiones previas.

3. Nodos de eventos: representan los posibles resultados en una decisión. Es necesario determinar los posibles resultados y la probabilidad de ocurrencia de cada uno basados en la información disponible al momento de plantear el árbol de decisión. Puede ocurrir que después de nodos de eventos continúe el proceso de selección entre alternativas probables.
4. Nodos finales: resultados finales generados por la serie de decisiones y resultados previos.

Para determinar que tan correcta es la predicción o clasificación realizada por modelos como los arboles de decisión, se emplean métricas en el campo de la búsqueda y análisis de información para determinar la eficacia y si es correcto el resultado obtenido, como la exactitud y precisión, en donde la primera representa cuantos casos sobre el total han sido clasificados correctamente en la clase a la que pertenecen, y la segunda mide cuantos casos que han sido clasificados en una clase pertenecen realmente a esa clase.

Una forma de visualizar cuantos casos se clasifican correctamente es a través de la matriz de confusión, donde las columnas representa los datos predichos y las filas representa la clase a la que pertenece los datos [4] .

Matriz de confusión		Predicho		
		Negativo	Positivo	
Real	Negativo	a	b	Verdadero Negativo
	Positivo	c	d	Exactitud
		Sensibilidad	Especificidad	Precisión

Figura 5.4: Matriz de confusión

De acuerdo a la matriz de confusión tenemos que:

- a es el número de predicciones correctas de clase negativa (negativos reales)
- b es el número de predicciones incorrectas de clase positiva (falsos positivos)

- c es el número de predicciones incorrectas de clase negativa (falsos negativos)
- d es el número de predicciones correctas de clase positiva (positivos reales)

De acuerdo a los valores de la matriz de confusión se define las métricas de exactitud (P) y precisión (AC). La Precisión o “Accuracy” (AC) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones, y se calcula mediante la ecuación 5.3:

$$AC = \frac{a + d}{a + b + c + d} \quad (5.3)$$

La Exactitud o, en inglés, “Precision” se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. También se conoce como Verdadero Positivo (o “True positive rate”). Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos, y se calcula con la siguiente ecuación 5.4:

$$P = \frac{d}{b + d} \quad (5.4)$$

- Si todos los casos en un nodo tiene un valor idéntico, el nodo no se dividirá.
- Si se llega a la profundidad máxima del árbol o el tamaño del nodo a dividir es el mínimo entonces no se sigue dividiendo el nodo y se deja como hoja.

El algoritmo emplea unidades de medida de la entropía o ganancia de la información para determinar que atributo es mejor para dividir la información en cada nodo. La entropía se utiliza para medir la cantidad de información útil y se calcula como:

Para la construcción del árbol de decisión se utilizó el algoritmo de clasificación y regresión de arboles CART, donde solo necesita datos de entrada, se ajustan parámetros y devuelve un árbol como resultado; este algoritmo se decide cual atributo es el que mejor divide los datos, ignoran los valores vacios y usan reglas de parada para terminar la ejecución [5.5](#):

$$H(S) = - \sum_{x \in X} p(x) * \log_2 p(x) \quad (5.5)$$

Donde, $H(S)$ es la entropía del conjunto de datos S . X es el conjunto de clases en S . $p(x)$ la proporción de elementos de la clase x sobre el cardinal de S .

La ganancia de información determina cuanta información se gana escogiendo el atributo como candidato a ser el nodo del árbol, siendo la diferencia de la entropía calculada anterior y la entropía resultante de escoger un atributo concreto por tanto un valor grande de IG implica que la entropía habiendo escogido un atributo A que nos indica que el conjunto de datos está mejor clasificado con el atributo A . y se presenta como [5.6](#):

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) * H(t) \quad (5.6)$$

Donde, $IG(A, S)$ es la Ganancia de la información del conjunto S dividiéndolo mediante el atributo A .

$H(S)$ es la entropía del conjunto de datos S .

T son los subconjuntos de clases de S dividido por el atributo A .

$p(t)$ la proporción de elementos de la clase t sobre el cardinal de S . $H(t)$ es la entropía del subconjunto t .

5.1.10. Metodología usadas en la minería de datos

Se ha investigado los principales métodos existentes que son los más adecuados para la realización de proyectos enfocados hacia la minería de datos como són: KDD

(Knowledge Discovery in Databases) , CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess).

5.1.11. KDD ó Knowledge Discovery in Databases

De manera general, los datos son considerados como la materia prima en bruto. Una vez que las personas o usuarios involucrados le atribuyen algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento. En lo alto de la pirámide se encuentra la inteligencia, la cual tiene relación con la aplicación del conocimiento en beneficio de buenas decisiones y resultados. Lo anterior se encuentra reflejado en la Figura 5.5, la cual ilustra la jerarquía que existe entre los datos, la información, el conocimiento y la inteligencia; los datos están en gran volumen y poseen poco valor, mientras que el conocimiento está en bajo volumen y posee un alto valor para las personas o usuarios involucrados, sobre esto último se aplica la inteligencia orientada a tomar buenas decisiones[10].

El KDD apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos y, de esta manera, permitir al usuario el uso de información valiosa para su conveniencia. Además, al hablar del concepto KDD se debe considerar el proceso no trivial existente para identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos[10]. Por lo tanto, KDD se clasifica en diferentes etapas como son: Selección, Preprocesamiento, Transformación, Minería de datos e Interpretación/Evaluación.

Selección: Esta etapa consiste en crear un conjunto de datos de destino, o centrarse en un subconjunto de variables o muestras de datos, en el que se va a realizar el descubrimiento.

Preprocesamiento: Esta etapa consiste en la limpieza y preprocesamiento de los datos de destino para obtener datos consistentes.



Figura 5.5: Jerarquía del conocimiento

Fuente: [11]

Transformación: Esta etapa consiste en la transformación de los datos utilizando métodos de reducción de la dimensionalidad o transformaciones sobre los mismos.

Minería de datos: Esta etapa consiste en la búsqueda de patrones de interés en una forma representacional particular, dependiendo del objetivo de la gestión de datos (generalmente, predicción).

Interpretación/Evaluación: Esta etapa consiste en la interpretación y evaluación de los patrones minados[4]

Tal como se indica en la figura 5.6:

5.1.12. CRISP-DM (Cross Industry Standard Processfor Data Mining)

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), consta básicamente de seis etapas flexibles: Entendimiento del negocio o análisis del problema, Análisis o entendimiento de los datos, Preparación de los datos, Modelado,

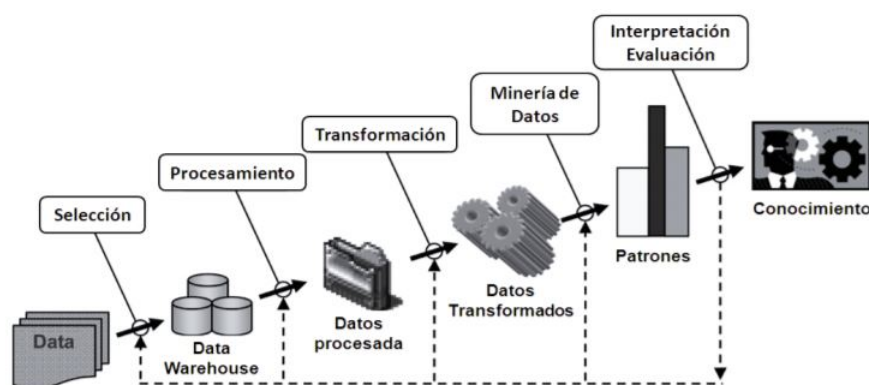


Figura 5.6: Proceso de Descubrimiento del Conocimiento (KDD)

Fuente: [10]

Evaluación e Implementación/Explotación. A diferencia del proceso KDD, presenta dos etapas adicionales; Entendimiento del negocio e Implementación/Explotación. Tal como se muestra en la Figura 5.7, las etapas de la metodología CRISP-DM en conjunto se comportan de manera cíclica para la mejora continua en la calidad de los resultados. Cabe destacar que una vez finalizado el ciclo, se puede comenzar nuevamente con el análisis del problema para obtener mayor conocimiento a partir de la implementación o explotación de los resultados obtenidos.

Por tanto, se evidencia las etapas de la metodología en la figura 5.7

A continuación se describe cada una de las fases:

Fase I. Comprensión del negocio: Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos[4].

Fase II. Estudio y comprensión de los datos: La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta[4].

Fase III. Preparación de los datos: La fase de preparación de datos cubre todas

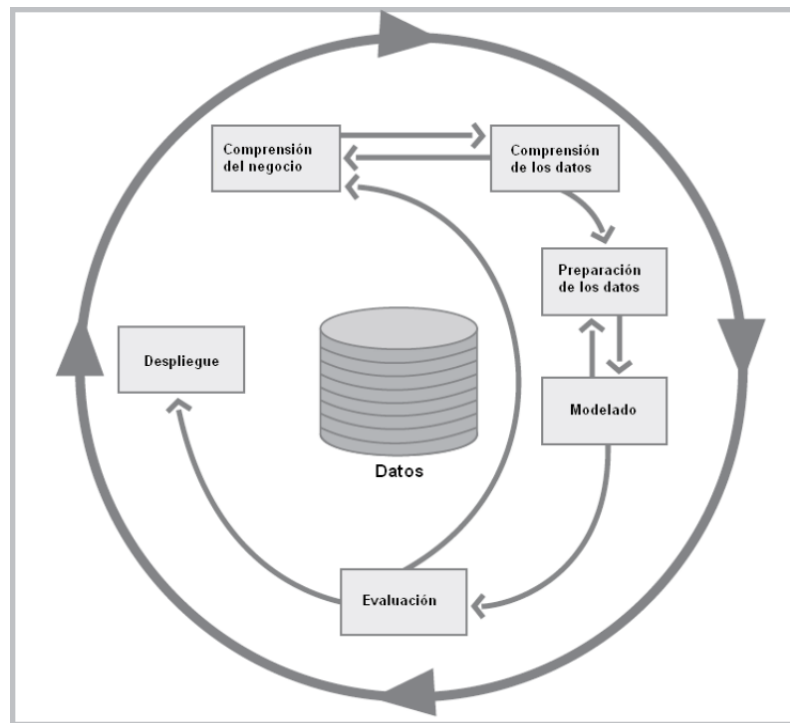


Figura 5.7: Metodología CRISP-DM

Fuente: [10]

las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan[4].

Fase IV. Modelado: En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos[4].

Fase V. Evaluación (obtención de resultados): En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo,

es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos [4].

Fase VI. Despliegue (puesta en producción): Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización [4].

5.1.13. Proceso SEMMA

El proceso SEMMA (Sample, Explore, Modify, Model, Assess) se basa en una serie de etapas para conducir proyectos de minería de datos. El SAS Institute lo ha desarrollado, considerando un ciclo compuesto de cinco etapas: Sample consiste en obtener una muestra de datos a través de la extracción de una porción suficientemente grande como para contener la información significativa, pero no tan elevada para que sea fácil de manipular; Explore, es la etapa para explorar los datos buscando, de forma anticipada, tendencias y anomalías para lograr el entendimiento e ideas sobre los datos; Modify se centra en la transformación o modificación de los datos a través de la creación, selección y tratamiento de variables orientados a la selección de un modelo; Model es la etapa para el modelamiento de los datos a través de la aplicación de distintos algoritmos, buscando combinaciones de datos que sean útiles y confiables para predecir resultados esperados; finalmente, Assess consiste en la evaluación de los datos mediante la valoración de la utilidad y confiabilidad del conocimiento descubierto a partir del proceso de minería de datos [10].

Muestreo: Esta etapa consiste en muestrear los datos extrayendo una porción de un gran conjunto de datos lo suficientemente grande como para contener la información significativa, pero lo suficientemente pequeño como para manipularla rápidamente[4].

Exploración: Esta etapa consiste en la exploración de los datos buscando tendencias y anomalías imprevistas con el fin de obtener comprensión e ideas[4].

Modificación: Esta etapa consiste en la modificación de los datos creando, seleccionando y transformando las variables para enfocar el proceso de selección del modelo[4].

Modelado: Esta etapa consiste en modelar los datos permitiendo que el software busque automáticamente una combinación de datos que prediga de manera confiable un resultado deseado[4].

Evaluación: Esta etapa consiste en evaluar los datos, evaluando la utilidad y confiabilidad de los hallazgos del proceso de gestión de la movilidad y estimar qué tan bien se desempeña[4].

Aunque el proceso SEMMA ofrece una metodología fácil de entender y seguir, carece de ciertas etapas que estén relacionadas, principalmente, con la contextualización y entendimiento del negocio desde el cual se obtiene la muestra de datos.

5.1.14. Trabajos relacionados

A través de la realización de una revisión sistematica de literatura entorno a los temas de interés en relación al TT, se ha obtenido un conjunto de trabajos o estudios relacionados al objeto de estudio (Tablas 5.1 y 5.2), los mismos que fundamentan el presente TT y sirven de ayuda para comprender la forma en la que se desarrollan temas similares. El proceso de su obtención se detalla en el Anexo I.

Código	Título	Palabras Clave
Rsl-01	Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa [1]	Minería de datos, Lógica difusa.
Rsl-02	Análisis de accidentalidad vehicular usando técnicas de minería de datos[4]	Minería de datos, Técnicas de data mining.
Rsl-03	Aplicación de técnicas de minería de datos en accidentes de tráfico[9].	Tenicas de minería de datos, Accidentes de tráfico.
Rsl-04	Análisis de accidentes de tránsito en Zonas urbanas y rurales usando minería De datos difusa[10].	Tenicas de minería de datos, Minería de datos difusa.
Rsl-05	Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellin. [12]	Big data, algoritmos predictivos, minería de datos
Rsl-06	Analysis of accident times for highway locations using K-means clustering and decision rules extracted from decision trees [13]	Decision trees techniques, Data mining, traffic accident
Rsl-07	Analysis of road accident data and determining affecting factors by using regression models and decision trees[14]	Data mining, traffic accidents, desicion trees techniques.
Rsl-08	Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India[15]	Analysis traffic accidents, data mining prediction decision trees techniques.
Rsl-09	Road traffic accidents: An overview of data sources, analysis techniques and contributing factors[16]	An overview of data sources, analysis techniques and contributing.
Rsl-10	Risk analysis of traffic accidents' severities: An application of three data mining models.[17]	Data minig model, traffic accidents.
Rsl-11	Machine Learning Based Analysis for Road Accident Prediction[18]	Machine Learning, Road Accident Prediction.
Rsl-12	Road-user specific analysis of traffic accident using data mining techniques[19]	Data mining techniques, analysis of traffic accident
Rsl-13	Review of the machine learning techniques in road crashes[20]	Machine learning techniques, road crashes, data mining.
Rsl-14	Road Accident Analysis System using Data Mining[21]	Data mining, Accident analysis system.
Rsl-15	An Empirical Study Road Accident In India With Data Mining Approach[21]	Data mining, Study Road Accident.

Tabla 5.1: Trabajos relacionados a Minería de datos aplicada a siniestros de tránsito.

Código	Título	Palabras Clave
Rsl-16	Statistical analysis and classifier accuracy improvements models for road accident issues on national highways in India. [22]	Decision strees, models for road accident.
Rsl-17	A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines [23]	Data mining, Decision trees model, road accident.
Rsl-18	A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity [24]	Association rule mining, decision trees techniques, traffic accident
Rsl-19	A Conjoint Analysis of Road Accident Data using K-modes Clustering and sayesian Networks (Road Accident Analysis using clustering and classification) [25]	K-modes clustering, Data mining, classification decision trees.
Rsl-20	Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques [26]	Prediction of road accident, Machine learnig techniques, decision tree model.
Rsl-21	Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique [27]	Classification technique, road accident, FRAM, Data mining.
Rsl-22	Predicting the Impact of Text-Reading Using Decision Trees [28]	Data mining, Decision trees, Predicting.
Rsl-23	A Radical Approach to Forecast the Road Accident Using Data Mining Technique. [29]	road accident, Data mining techniques.
Rsl-24	Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. [30]	Data mining, algorithms classification, Predicting road traffic.
Rsl-25	Severity Prediction with Machine Learning Methods. [31]	Machine learning techniques, Severity prediction.
Rsl-26	Collision analysis of driving scenarios. [32]	Data mining techniques, analysis of driving.

Tabla 5.2: Trabajos relacionados a Minería de datos aplicada a siniestros de tránsito.

Una vez desarrollada la Revisión Sistemática de Literatura(RSL), se establece que se han obtenido resultados óptimos de documentación en donde se enfoca a la minería de datos, mediante la aplicación de la técnica de árboles de decisión, mismos que serán referentes para el desarrollo y elaboración del presente proyecto. Estos trabajos principalmente se ejecutan con el estudio de siniestros viales, desarrollando un análisis con algoritmos predictivos de distintas técnicas de minería de datos [Rsl-02], [Rsl-03], [Rsl-04], [Rsl-06], [Rsl-09], [Rsl-12], [Rsl-21], [Rsl-22]; en donde una de las técnicas o modelos considerados como base de estudio son los árboles de decisión, desarrollando diferentes predicciones con varios conjuntos de datos recopilados y se enfoca a establecer que causas inciden en el cometimiento de los accidentes de tránsito, tal como el incremento poblacional, factores estructurales o humanos, así como fallas tecnicomecánicas del vehículo [Rsl-05], [Rsl-07], [Rsl-01], [Rsl-11], [Rsl-14], [Rsl-15], [Rsl-16], [Rsl-17], [Rsl-18], [Rsl-23], [Rsl-24], [Rsl-25], [Rsl-26]. Además, realiza una comparación con algunas otras técnicas de minería de datos como los son K-modes clustering y regresión, aplicadas principalmente en Machine Learning.

Por tanto, la gran parte de los trabajos destacados en el RSL, emplean el modelo de árboles de decisión para el desarrollo de la minería de datos, estableciendo como un modelo óptimo y de buen rendimiento frente a diferentes modelos predictivos, justificando que más del 80% de los trabajos ejecutan este modelo en donde sus resultados son más exactos y de mayor fiabilidad con respecto a la accidentología vial.

5.1.15. Organismos de Control de Tránsito en el Cantón Loja

En el cantón Loja intervienen entes que regulan, planifican y controlan el tránsito, transporte y la seguridad vial en redes estatales y troncales. A continuación se describe cada uno de ellos.

Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial

La Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial (ANT), es el ente encargado de la regulación, planificación y control del transporte terrestre, tránsito y seguridad vial en el territorio nacional, en el ámbito de sus competencias, con sujeción a las políticas emanadas de los Ministerios Rector y Coordinador del Sector; así como del control del tránsito en las vías de la Red Estatal-troncales nacionales, en coordinación con los Gobiernos Autónomos descentralizados (GAD), la Comisión de Tránsito del Ecuador y la Policía Nacional de Tránsito. Este organismo de control se creó a partir de la reforma a la antigua ley de tránsito en 2011, el cuál sustituye a la Comisión Nacional de Tránsito. Su fin es garantizar la libre y segura movilidad terrestre, prestando servicios de calidad que satisfagan la demanda ciudadana; coadyuvando a la preservación del medio ambiente y contribuyendo al desarrollo del país, en el ámbito de su competencia [33][8].

Unidad de Control Operativa de Tránsito del Cantón Loja

La Unidad de Control Operativa de Tránsito del Cantón Loja (UCOT) regula, dirige y controla cada una de las operaciones de transporte terrestre tránsito y seguridad vial en el Cantón Loja; fomentar conciencia ciudadana en el respeto de la ley, con objeto de disminuir los índices de siniestros de tránsito en el Cantón Loja en base a los controles planificados, actuar dentro de nuestra jurisdicción con honestidad y profesionalismo, incrementar la calidad y cobertura del servicio de transporte terrestre, incrementar la eficiencia operacional y el desarrollo del talento humano, mantener la capacitación constante del personal para proporcionar un alto nivel de seguridad vial. Esta institución, al igual que los organismos anteriormente citados, se creó por Ley en 2011, luego de que se delegara todas las competencias de control de tránsito a los gobiernos autónomos descentralizados. Comenzaron sus funciones dos años después de haber sido emitida la nueva Ley, específicamente el 28 de agosto del 2013 y ejercen sus funciones en todo el cantón Loja, hasta la fecha[33][3].

Jefatura de Accidentología Vial

El departamento de Jefatura de Accidentología Vial (JAV) de la Policía Nacional

del Ecuador, es un pilar fundamental para la administración de justicia en delitos de tránsito al determinar las causas científicas que generaron un percance vehicular. El trabajo en el ámbito técnico-científico ejecutado por los agentes permite determinar la causa o los orígenes de todos los accidentes de tránsito. El JAV cuenta con personal especializado en el área de investigación técnica-científica y los trabajos los coordina directamente con la Fiscalía a la que entrega informes como los peritajes en donde se evidencian las causas que generaron el percance. Remiten, también, los avalúos de los daños materiales públicos o privados; ejecuta los reconocimientos del lugar de los hechos, de las huellas y de los vestigios; y, además, reconstruye los accidentes de tránsito, a través, de una simulación virtual[8].

6. Metodología

El presente trabajo de titulación denominado “Análisis de la accidentabilidad vehicular en la zona Urbana del Cantón Loja con aplicación de las técnicas de Minería de Datos”, corresponde a un proyecto de análisis de patrones y generación de predicciones, el cual tiene un enfoque poder predecir un posible siniestro vial y tomar las medidas adecuadas para evitar sucesos que expongan la humanidad de las personas, poniendo en riesgo su vida así como bienes materiales, brindando un beneficio e integridad de todos los participantes viales y a la ciudadanía en general. Por consiguiente la metodología que se busca abordar para el presente Proyecto de Trabajo de Titulación se describe a continuación en la figura 6.1:

Objetivos Específicos	Actividades/Tareas	Métodos / Técnicas	Materiales	Lugar	Responsable	Producto
1. Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020.	a) Establecer las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja. b) Obtener la base de datos siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019-2020. c) Desarrollar la transformación y limpieza de la base de datos obtenida.	Deductivo, Inductivo, Empírico, Científico, KDD	PC, Repositorio de datos de Instituciones de Control Operativo de Tránsito, Herramientas para el preprocesamiento de datos, Librerías Office.	Domicilio del tesista	Patricio Benítez	Obtener la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos.
2. Implementar el modelo de árboles de decisión para establecer las principales causas sobre los accidentes de tránsito.	a) Identificar las variables que influyen en el cometimiento del accidente de tránsito. b) Entrenar el modelo de árboles de decisión, con la información de accidentes de tránsito ocurridos en la zona urbana del cantón Loja, periodo 2019 - 2020.	Deductivo, Inductivo, Empírico, Científico, Modelo de árbol de decisión, KDD	PC, Repositorio de datos de Instituciones de Control Operativo de Tránsito, Herramientas para el preprocesamiento de datos, Librerías Office, Mendeley Desktop.	Domicilio del tesista	Patricio Benítez	Modelo Entrenado
3. Evaluar la exactitud del modelo predictivo de árboles de decisión para determinar las incidencias que ocasionan los siniestros de tránsito en la zona urbana del Cantón Loja.	a) Analizar los resultados obtenidos del modelo entrenado. b) Evidenciar el funcionamiento del modelo con datos actuales. c) Elaborar el documento final.	Deductivo, Inductivo, Empírico, Científico, Modelo de árbol de decisión, KDD	PC, Repositorio de datos de Instituciones de Control Operativo de Tránsito, Herramientas para el preprocesamiento de datos, Librerías Office, Mendeley Desktop.	Domicilio del tesista,	Patricio Benítez	Documentación final del análisis

Figura 6.1: Metodología del PTT.

7. Cronograma

El trabajo de titulación denominado “Análisis de la accidentabilidad vehicular en la zona Urbana del Cantón Loja con aplicación de las técnicas de Minería de Datos”, es un proyecto de análisis de datos; por lo tanto, requiere de una buena planificación de las actividades, partiendo de esto los autores del presente proyecto se plantean seguir el siguiente cronograma descrito en la figura 7.1, para llevar a cabo la tareas/actividades en los tiempos establecidos y llevar un control acorde. Con esto se busca mitigar riesgos que pueden derivar en retrasos de la entrega del proyecto.

Objetivos Específicos	Actividades/Tareas	Producto	Fecha inicio	Fecha fin	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6
1. Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020.	a) Establecer las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja.	Obtener la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos.	25/10/2021	2/12/2021						
	b) Obtener la base de datos siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019-2020.									
	c) Desarrollar la transformación y limpieza de la base de datos obtenida.									
2. Implementar el modelo de árboles de decisión para establecer las principales causas sobre los accidentes de tránsito.	a) Identificar las variables que influyen en el cometimiento del accidente de tránsito.	Modelo Entrenado	2/12/2021	20/1/2022						
	b) Entrenar el modelo de árboles de decisión, con la información de accidentes de tránsito ocurridos en la zona urbana del cantón Loja, periodo 2019 - 2020.									
3. Evaluar la exactitud del modelo predictivo de árboles de decisión para determinar las incidencias que ocasionan los siniestros de tránsito en la zona urbana del Cantón Loja.	a) Analizar los resultados obtenidos del modelo entrenado.	Documentación final del análisis.	20/1/2022	31/3/2022						
	b) Evidenciar el funcionamiento del modelo con datos actuales.									
	c) Elaborar el documento final.									

Figura 7.1: Cronograma del PTT.

8. Presupuesto

8.1. Recursos Humanos

Talento Humano	Justificación	Número de Horas	Valor por Hora	SUBTOTAL
Autores del Proyecto	2 Estudiantes a cargo de la ejecución del proyecto	480	\$ 10,00	\$ 4800,00
Personal de Apoyo	Docente director en la elaboración y supervisión del proyecto.	80	\$ 10,47	\$ 837,60
	Docente guía de la materia Trabajo de Titulación del X de la Carrera de Ingeniería en Sistemas	45	\$ 12,50	\$ 562,50
	Profesional Directivo UCOT	15	\$ 38,00	\$ 570,00
Total				\$ 6.770,10

Tabla 8.1: Recursos humanos empleados para el desarrollo del proyecto.

8.2. Recursos Software, Hardware y TICs

Recursos HW/SW	Justificación	Número de Meses	Valor por Mes	SUBTOTAL
Internet	Revisión de literatura, consultas, comunicación	6	\$ 40.00	\$ 240,00
impresora	Para la documentación de la data	6	\$ 30,00	\$ 180,00
Pendrive	Para el almacenamiento de la data e informes digitales	6	\$ 0,00	\$ 0,00
2 Computadoras portátiles (Propias)	Herramienta básica para el trabajo a desarrollar	6	\$ 0,00	\$ 0,00
Total				\$ 420,00

Tabla 8.2: Recursos Software, Hardware, TICs a utilizar para el desarrollo del proyecto.

8.3. Insumos

Insumos	Justificación	Número de Meses	Valor por Mes	SUBTOTAL
Papeleria	Impresiones, anillados y empastados	6	\$ 2	\$ 12,00
Transporte	Movilización al campus Universitario	6	\$ 30	\$ 180,00
Viáticos	Alimentación y Hospedaje	6	\$ 400	\$ 2400,00
Total				\$ 2592,00

Tabla 8.3: Insumos a utilizar para el desarrollo del proyecto.

8.4. Recursos Totales

RECURSOS	SUBTOTAL
Recursos Humanos	\$ 6.770,10
Recursos Software, Hardware, TICs	\$ 420,00
Insumos	\$ 2592,00
SubTotal	\$ 9782,10
Imprevistos (+10% del Subtotal)	\$ 978,21
Presupuesto total del Proyecto	\$ 10.760,31

Tabla 8.4: Presupuesto total para la elaboración del proyecto.

El presupuesto de \$ 10.760,31 (Diez mil setecientos sesenta dólares con treinta y uno centavos), será financiado por el autor y el recurso humano docente de la Universidad Naciona de Loja.

Bibliografía

- [1] J. Sanabria Garzón, “”Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa”,”
- [2] Agencia Nacional de Tránsito, “Estadísticas siniestros de tránsito – Agencia Nacional de Tránsito del Ecuador – ANT.”
- [3] CONSEJO NACIONAL DE COMPETENCIAS, “Resolución Nro. 006-CNC-2012,” 2012.
- [4] D. Hasbleidy, C. Díaz, D. Felipe, and S. Vargas, *Análisis de accidentalidad vehicular usando técnicas de minería de datos*. PhD thesis, UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS, 2019.
- [5] Y. B. Monjas, “MINERÍA DE DATOS,”
- [6] UNL, “Computación | Universidad Nacional de Loja,” 2021.
- [7] MINTEL and SENESCYT, “Líneas de investigación, desarrollo e innovación y transferencia del conocimiento en TIC 2019,” *Mintel*, p. 110, 2019.
- [8] C. De Ingeniería, M. Automotriz, S. Cabrera, M. Esteban, V. Figueroa, and L. Henry, “UNIVERSIDAD POLITÉCNICA SALESIANA SEDE CUENCA,”
- [9] D. María José Ramírez Quintana, “UNIVERSIDAD POLITÉCNICA DE VALENCIA,” 2014.
- [10] A. I. ARÁNGUIZ CASTRO, “ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS Y RURALES USANDO MINERÍA DE DATOS DIFUSA.”
- [11] B. García Moraima and M. Delgado Fernández, “Gestión y Generación de Conocimientos a partir de la información de patentes.,” feb 2012.
- [12] Cristian Andrés Escobar Canizales, Sergio Andrés Rubiano Rodríguez, and Juan Camilo Vega Figueroa, “Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellín.”
- [13] A. M. Aljofey and K. Alwagih, “Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees,” *International Journal of Computer Applications Technology and Research*, vol. 7, pp. 1–011, 2018.

-
- [14] A. Nazeri, H. GharehGozlu, F. Faraji, and S. Asakareh, "Analysis of road accident data and determining affecting factors by using regression models and decision trees," *International Journal of Business Intelligence and Data Mining*, vol. 18, no. 4, pp. 449–471, 2021.
 - [15] S. Kumar and D. Toshniwal, "Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India," *European Transport Research Review 2017 9:2*, vol. 9, pp. 1–10, may 2017.
 - [16] A. Chand, S. Jayesh, and A. Bhasi, "Road traffic accidents: An overview of data sources, analysis techniques and contributing factors," *Materials Today: Proceedings*, jun 2021.
 - [17] S. AlKheder, F. AlRukaibi, and A. Aiash, "Risk analysis of traffic accidents' severities: An application of three data mining models," *ISA Transactions*, vol. 106, pp. 213–220, nov 2020.
 - [18] A. Venkat, G. K. Vijey, and I. Susan Thomas, "MACHINE LEARNING BASED ANALYSIS FOR ROAD ACCIDENT PREDICTION," *International Journal of Emerging Technology and Innovative Engineering*, vol. 6, pp. 2394–6598, 2020.
 - [19] P. Tiwari, S. Kumar, and D. Kalitin, "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques," *Communications in Computer and Information Science*, vol. 776, pp. 398–410, 2017.
 - [20] S. Srivastava, A. Bagga, and R. S. Shekhawat, "Review of the machine learning techniques in road crashes," *Proceedings of International Conference on Computation, Automation and Knowledge Management, ICCAKM 2020*, pp. 373–376, jan 2020.
 - [21] M. A. W. Chanchal Sharma, "Road Accident Analysis System using Data Mining," *International Journal on Future Revolution in Computer Science Communication Engineering*, vol. 4, pp. 846–848, apr 2018.
 - [22] Chitnis SD, "Design Engineering Statistical Analysis and Classifier Accuracy Improvements Models for Road Accident Issues on National Highways in India,"
 - [23] J. R. Asor, G. M. B. Catedrilla, and J. E. Estrada, "A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines," *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, vol. 2018-January, pp. 96–101, apr 2018.
 - [24] M. Gupta, V. Kumar Solanki, and V. Kumar Singh, "A Novel Framework to Use Association Rule Mining for classification of traffic accident severity," *Ingeniería Solidaria*, vol. 13, pp. 37–44, jan 2017.
-

- [25] S. Kumar, V. S. Semwal, V. K. Solanki, P. Tiwari, and D. Kalitin, “A Conjoint Analysis of Road Accident Data using K-modes Clustering and Bayesian Networks (Road Accident Analysis... prayag tiwari Related papers Performance Evaluation of Lazy, Decision Tree Classifier and Multilayer Perceptron on Traffic... prayag tiwari Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques prayag tiwari A Conjoint Analysis of Road Accident Data using K-modes Clustering and Bayesian Networks (Road Accident Analysis using clustering and classification),”
 - [26] J. Paul, Z. Jahan, K. F. Lateef, M. R. Islam, and S. C. Bakchy, “Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques,” *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, vol. 2020-December, dec 2020.
 - [27] K. P. Hlaing, N. T. T. Aung, S. Z. Hlaing, and K. Ochimizu, “Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique,” *2019 International Conference on Advanced Information Technologies, ICAIT 2019*, pp. 256–261, nov 2019.
 - [28] E. G. Nathanail, P. D. Prevedouros, M. Mintu Miah, and R. D. Melo Barros, “Predicting the Impact of Text-Reading Using Decision Trees,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11804 LNAI, pp. 685–696, sep 2019.
 - [29] A. Makkar and H. S. Gill, “A Radical Approach to Forecast the Road Accident Using Data Mining Technique,” *International Journal of Innovative Science and Research Technology*, vol. 2, no. 8, 2017.
 - [30] T. K. Bahiru, D. Kumar Singh, and E. A. Tessfaw, “Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity,” *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, pp. 1655–1660, sep 2018.
 - [31] B. Geyik and M. Kara, “Severity Prediction with Machine Learning Methods,” *HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, jun 2020.
 - [32] E. Watkins, M. Kloc, S. Weerasuriya, and M. El-Hajj, “Collision analysis of driving scenarios,” *2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017*, mar 2017.
 - [33] GAD Municipal de Loja, “UCOT | Municipio de Loja,” 2013.
-

A. Anexo I

REVISIÓN SISTEMÁTICA DE LITERATURA SOBRE MODELO DE ÁRBOLES DE DECISIÓN COMO TÉCNICA DE MINERÍA DE DATOS PARA LA APLICACIÓN EN SINIESTROS DE TRÁNSITO

Metodología

La presente Revisión Sistemática de Literatura (RSL) establece el análisis de las principales técnicas, herramientas y metodologías existentes de aplicadas a minería de datos con el objetivo de resumir la información existente respecto a Accidentología vial. Siguiendo los lineamientos de Kitchenham y Charters[1] se debe seleccionar bases de datos científicas multidisciplinarias de gran impacto mundial consideradas para la investigación; además, la actividad para realizar una RSL implica planificar, ejecutar y reportar la revisión, siendo aplicados estos parámetros en la investigación.

Proceso de la RSL

1. Planificación

Establecemos la importancia de la presente revisión, por tanto, es indispensable cumplir con los objetivos planteados en el proyecto de investigación detallado a continuación.

1.1. Objetivo

El objetivo planteado para la elaboración de la RSL es:

- Analizar la accidentabilidad vehicular mediante la aplicación del modelo predictivo de árboles de decisión como técnica de minería de datos.

1.2. Preguntas de Investigación

Para llevar a cabo la investigación se plantean las siguientes preguntas de investigación, con la finalidad de encontrar términos claros y precisos, para identificar las técnicas, herramientas y metodologías existentes para la aplicación de minería de datos en siniestros de tránsito, como lo presenta la Tabla 1:

Tabla I.

Preguntas de Investigación.

ID	Preguntas de Investigación
PI1	¿Cuáles son las principales técnicas de minería de datos aplicadas en accidentes de tránsito?
PI2	¿Cuáles son las principales herramientas de minerías de datos aplicadas en accidentes de tránsito?
PI3	¿Qué metodologías son las más utilizadas en la minería de datos con aplicación a accidentes de tránsito?

Es importante destacar que se ha usado los criterios PICOC (Población, Intervención, Comparación, Resultado, Contexto) para establecer las preguntas de investigación.

1.3. Estrategias de búsqueda

a. Scripts de búsqueda. Con base en la sección anterior, se presenta la lista de posibles scripts de búsqueda que se utilizarían para ser ejecutadas de acuerdo con cada base de datos científica, según se presentan en la *Tabla II*.

Tabla II.
Scripts de Búsqueda.

ID	Script de Búsqueda
SB1	("data mining") AND ("decision trees") AND ("analysis") AND ("prediction") AND ("road fatalities" OR "traffic accidents" OR "road traffic crashes" OR "road traffic accidents") ("minería de datos AND ("árboles de decisión") AND ("análisis") AND ("predicción") AND ("accidente de tránsito")

Es importante especificar que se estableció una búsqueda general sistemática, para trabajar con las bibliotecas virtuales mostradas en la *Tabla III*, además se utiliza una sintaxis de búsqueda lo más similar posible en las seis bases de datos y se adhiere las reglas establecidas para cada una de ellas.

b. Bibliotecas Virtuales. Las fuentes bibliográficas seleccionadas para la realización de la RSL son las descritas en la *Tabla III*:

Tabla III.
Bibliotecas virtuales

Bibliotecas Virtuales	URL
IEEE Digital Library	https://ieeexplore.ieee.org
Science Direct	https://www.sciencedirect.com
Springer Link	https://www.springerlink.com
Google Scholar	https://scholar.google.com

c. Criterios de Selección. Se utilizó la herramienta Mendeley para gestionar las referencias bibliográficas, de la ejecución de los scripts se obtuvo un total de 82 trabajos con formato de referencia ".bib"; mismos que serán evaluados con los criterios de inclusión y exclusión. A continuación, se especifican los criterios de inclusión y exclusión considerando las preguntas de investigación.

- **Criterios de Inclusión.** Los criterios de inclusión que se utilizó se describen en la *TABLA IV*.

Tabla IV.
Criterios de Inclusión.

Criterio	Descripción
Contenido	Los estudios deben contener relación con el objeto de investigación al menos en su título, resumen, introducción o desarrollo.
Fecha de publicación	Estudios publicados desde el año 2017 hasta el presente año
Bibliotecas Virtuales	IEEE Digital Library, Science Direct, Springer, Google Scholar
Idioma	Se consideran artículos escritos solo en Inglés y Español

- **Criterios de Exclusión.** No serán considerados los estudios que no cuenten con información relacionada al objeto de estudio, aquellos estudios que no ayuden a responder las preguntas de investigación, estudios duplicados y los que no cumplan con los criterios de inclusión que se muestran en la *Tabla IV*.

2. Ejecución y Resultados

Terminada de realizar el análisis a través de la cadena de búsqueda, así como también aplicados los criterios de inclusión y exclusión se obtuvo un total de 26 trabajos relacionados, seleccionados de todas las fuentes bibliográficas, en la *Tabla V* se detalla el número de trabajos relacionados más relevantes, encontrados en cada una de las fuentes bibliográficas entorno a las cadenas de búsqueda dadas anteriormente.

Tabla V.
Trabajos relacionados seleccionados.

Código	Título	Ref.	Buscador
TR01	Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa	[2]	Google Scholar
TR02	Análisis de accidentalidad vehicular usando técnicas de minería de datos	[3]	Google Scholar
TR03	Aplicación de técnicas de minería de datos en accidentes de tráfico	[4]	Google Scholar
TR04	Análisis de accidentes de tránsito en Zonas urbanas y rurales usando minería De datos difusa.	[5]	Google Scholar
TR05	Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellin.	[6]	Google Scholar
TR06	Analysis of accident times for highway locations using K-means clustering and decision rules extracted from decision trees	[7]	Google Scholar
TR07	Analysis of road accident data and determining affecting factors by using regression models and decision trees	[8]	Google Scholar
TR08	Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India	[9]	Google Scholar

TR09	Road traffic accidents: An overview of data sources, analysis techniques and contributing factors	[10]	Google Scholar
TR010	Risk analysis of traffic accidents' severities: An application of three data mining models.	[11]	ScienceDirect
TR11	Machine Learning Based Analysis for Road Accident Prediction	[12]	Google Scholar
TR12	Road-user specific analysis of traffic accident using data mining techniques	[13]	Google Scholar
TR13	Review of the machine learning techniques in road crashes	[14]	Google Scholar
TR14	Road Accident Analysis System using Data Mining	[15]	Google Scholar
TR15	An Empirical Study Road Accident In India With Data Mining Approach	[16]	Google Scholar
TR16	Statistical analysis and classifier accuracy improvements models for road accident issues on national highways in India	[17]	Google Scholar
TR17	A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines	[18]	Google Scholar
TR18	A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity	[19]	Google Scholar
TR19	A Conjoint Analysis of Road Accident Data using K-modes Clustering and sayesian Networks (Road Accident Analysis using clustering and classification)	[20]	Google Scholar
TR20	Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques	[21]	Google Scholar
TR21	Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique	[22]	Google Scholar
TR22	Predicting the Impact of Text-Reading Using Decision Trees	[23]	Spring erlink
TR23	A Radical Approach to Forecast the Road Accident Using Data Mining Technique.	[24]	Google Scholar
TR24	Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity	[25]	IEEE
TR25	Severity Prediction with Machine Learning Methods	[26]	IEEE
TR26	Collision analysis of driving scenarios	[27]	IEEE

2.1. Análisis de los datos

Se ha realizado un análisis individual para cada uno de los trabajos relacionados propuestos, los mismos que aportaron con información relevante entorno a la pregunta de investigación definida en el presente TT, dicho análisis se redacta en la sección de Marco Teórico en el apartado de Trabajos Relacionados.

Referencias

- [1] B. Kitchenham, “Procedures for Performing Systematic Reviews,” 2004.
- [2] J. Sanabria Garzón, “Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa.”
- [3] D. Hasbleidy, C. Díaz, D. Felipe, and S. Vargas, “Análisis de accidentalidad vehicular usando técnicas de minería de datos,” UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS, 2019.
- [4] Ing. Alvaro Galilei Pumares Romero, “Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador.” <http://repositorio.uisrael.edu.ec/bitstream/47000/2299/1/UISRAEL-EC-MASTER-TELEM-378.242-2019-015.pdf> (accessed Jun. 22, 2021).
- [5] A. I. ARÁNGUIZ CASTRO, “ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS Y RURALES USANDO MINERÍA DE DATOS DIFUSA.” http://opac.pucv.cl/pucv_txt/txt-3500/UCF3892_01.pdf (accessed Aug. 10, 2021).
- [6] Cristian Andrés Escobar Canizales, Sergio Andrés Rubiano Rodríguez, and Juan Camilo Vega Figueroa, “Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellín.” https://repository.ucc.edu.co/bitstream/20.500.12494/28567/4/2020_Modelo_BigData_computacional.pdf (accessed Jun. 22, 2021).
- [7] A. M. Aljofey and K. Alwagih, “Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees,” *Int. J. Comput. Appl. Technol. Res.*, vol. 7, pp. 1–011, 2018.
- [8] A. Nazeri, H. GharehGozlu, F. Faraji, and S. Asakareh, “Analysis of road accident data and determining affecting factors by using regression models and decision trees,” *Int. J. Bus. Intell. Data Min.*, vol. 18, no. 4, pp. 449–471, 2021, doi: 10.1504/IJBIDM.2021.115480.
- [9] S. Kumar and D. Toshniwal, “Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India,” *Eur. Transp. Res. Rev.* 2017 92, vol. 9, no. 2, pp. 1–10, May 2017, doi: 10.1007/S12544-017-0242-Z.
- [10] A. Chand, S. Jayesh, and A. B. Bhasi, “Road traffic accidents: An overview of data sources, analysis techniques and contributing factors,” *Mater. Today Proc.*, Jun. 2021, doi: 10.1016/J.MATPR.2021.05.415.
- [11] S. AlKheder, F. AlRukaibi, and A. Aiash, “Risk analysis of traffic accidents’ severities: An application of three data mining models,” *ISA Trans.*, vol. 106, pp. 213–220, Nov. 2020, doi: 10.1016/J.ISATRA.2020.06.018.
- [12] A. Venkat, G. K. Vijey, and I. Susan Thomas, “MACHINE LEARNING BASED ANALYSIS FOR ROAD ACCIDENT PREDICTION,” *Int. J. Emerg. Technol. Innov. Eng.*, vol. 6, pp. 2394–6598, 2020, Accessed: Aug. 30, 2021. [Online]. Available: <https://ssrn.com/abstract=3550462>.
- [13] P. Tiwari, S. Kumar, and D. Kalitin, “Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques,” *Commun. Comput. Inf. Sci.*, vol. 776, pp. 398–410, 2017, doi: 10.1007/978-981-10-6430-2_31.
- [14] S. Srivastava, A. Bagga, and R. S. Shekhawat, “Review of the machine learning techniques in road crashes,” *Proc. Int. Conf. Comput. Autom. Knowl. Manag. ICCAKM 2020*, pp. 373–376, Jan. 2020, doi: 10.1109/ICCAKM46823.2020.9051506.

- [15] M. A. W. Chanchal Sharma, "Road Accident Analysis System using Data Mining," *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol. 4, no. 4, pp. 846–848, Apr. 2018, Accessed: Aug. 30, 2021. [Online]. Available: <http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1624>.
- [16] L. Boyapati, R. Kamla Kuchipudi, and G. Gurukesava Dasu, "AN EMPIRICAL STUDY ROAD ACCIDENT IN INDIA WITH DATA MINING APPROACH."
- [17] Chitnis SD, "Design Engineering Statistical Analysis and Classifier Accuracy Improvements Models for Road Accident Issues on National Highways in India," Accessed: Aug. 30, 2021. [Online]. Available: <https://www.researchgate.net/publication/354022068>.
- [18] J. R. Asor, G. M. B. Catedrilla, and J. E. Estrada, "A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, pp. 96–101, Apr. 2018, doi: 10.1109/ICOIACT.2018.8350662.
- [19] M. Gupta, V. Kumar Solanki, and V. Kumar Singh, "A Novel Framework to Use Association Rule Mining for classification of traffic accident severity," *Ing. Solidar.*, vol. 13, no. 21, pp. 37–44, Jan. 2017, doi: 10.16925/IN.V13I21.1726.
- [20] S. Kumar, V. S. Semwal, V. K. Solanki, P. Tiwari, and D. Kalitin, "A Conjoint Analysis of Road Accident Data using K-modes Clustering and sayesian Networks (Road Accident Analysis... prayag tiwari Related papers Performance Evaluat ion of Lazy, Decision Tree Classifier and Mult ilayer Percept ron on Traffic... prayag t iwa."
- [21] J. Paul, Z. Jahan, K. F. Lateef, M. R. Islam, and S. C. Bakchy, "Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques," *IEEE Reg. 10 Humanit. Technol. Conf. R10-HTC*, vol. 2020-Decem, Dec. 2020, doi: 10.1109/R10-HTC49770.2020.9356987.
- [22] K. P. Hlaing, N. T. T. Aung, S. Z. Hlaing, and K. Ochimizu, "Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique," *2019 Int. Conf. Adv. Inf. Technol. ICAIT 2019*, pp. 256–261, Nov. 2019, doi: 10.1109/AITC.2019.8921119.
- [23] E. G. Nathanail, P. D. Prevedouros, M. Mintu Miah, and R. De Melo Barros, "Predicting the Impact of Text-Reading Using Decision Trees," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11804 LNAI, pp. 685–696, Sep. 2019, doi: 10.1007/978-3-030-30241-2_57.
- [24] A. Makkar and H. S. Gill, "A Radical Approach to Forecast the Road Accident Using Data Mining Technique," *Int. J. Innov. Sci. Res. Technol.*, vol. 2, no. 8, 2017, Accessed: Aug. 30, 2021. [Online]. Available: www.ijisrt.com.
- [25] T. K. Bahiru, D. Kumar Singh, and E. A. Tessfaw, "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity," *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018*, pp. 1655–1660, Sep. 2018, doi: 10.1109/ICICCT.2018.8473265.
- [26] B. Geyik and M. Kara, "Severity Prediction with Machine Learning Methods," *HORA 2020 - 2nd Int. Congr. Human-Computer Interact. Optim. Robot. Appl. Proc.*, Jun. 2020, doi: 10.1109/HORA49412.2020.9152601.
- [27] E. Watkins, M. Kloc, S. Weerasuriya, and M. El-Hajj, "Collision analysis of driving scenarios," *2017 IEEE 7th Annu. Comput. Commun. Work. Conf. CCWC 2017*, Mar. 2017, doi: 10.1109/CCWC.2017.7868413.

B. Revisión sistemática de literatura

Código	Título	Palabras Clave
Rsl-01	Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa [1]	Minería de datos, Lógica difusa.
Rsl-02	Análisis de accidentalidad vehicular usando técnicas de minería de datos[4]	Minería de datos, Técnicas de data mining.
Rsl-03	Aplicación de técnicas de minería de datos en accidentes de tráfico[10].	Tenicas de minería de datos, Accidentes de tráfico.
Rsl-04	Análisis de accidentes de tránsito en Zonas urbanas y rurales usando minería De datos difusa[11].	Tenicas de minería de datos, Minería de datos difusa.
Rsl-05	Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellin. [13]	Big data, algoritmos predictivos, minería de datos
Rsl-06	Analysis of accident times for highway locations using K-means clustering and decision rules extracted from decision trees [14]	Decision trees techniques, Data mining, traffic accident
Rsl-07	Analysis of road accident data and determining affecting factors by using regression models and decision trees[15]	Data mining, traffic accidents, desicion trees techniques.
Rsl-08	Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India[16]	Analysis traffic accidents, data mining prediction decision trees techniques.
Rsl-09	Road traffic accidents: An overview of data sources, analysis techniques and contributing factors[17]	An overview of data sources, analysis techniques and contributing.
Rsl-10	Risk analysis of traffic accidents' severities: An application of three data mining models.[18]	Data minig model, traffic accidents.
Rsl-11	Machine Learning Based Analysis for Road Accident Prediction[19]	Machine Learning, Road Accident Prediction.
Rsl-12	Road-user specific analysis of traffic accident using data mining techniques[20]	Data mining techniques, analysis of traffic accident
Rsl-13	Review of the machine learning techniques in road crashes[21]	Machine learning techniques, road crashes, data mining.
Rsl-14	Road Accident Analysis System using Data Mining[22]	Data mining, Accident analysis system.
Rsl-15	An Empirical Study Road Accident In India With Data Mining Approach[22]	Data mining, Study Road Accident.

Código	Título	Palabras Clave
Rsl-16	Statistical analysis and classifier accuracy improvements models for road accident issues on national highways in India.[23]	Decision strees, models for road accident.
Rsl-17	A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines[24]	Data mining, Decision trees model, road accident.
Rsl-18	A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity[25]	Association rule mining, decision trees techniques, traffic accident
Rsl-19	A Conjoint Analysis of Road Accident Data using K-modes Clustering and sayesian Networks (Road Accident Analysis using clustering and classification) [26]	K-modes clustering, Data mining, classification decision trees.
Rsl-20	Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques [27]	Prediction of road accident, Machine learnig techniques, decision tree model.
Rsl-21	Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique[28]	Classification technique, road accident, FRAM, Data mining.
Rsl-22	Predicting the Impact of Text-Reading Using Decision Trees [29]	Data mining, Decision trees, Predicting.
Rsl-23	A Radical Approach to Forecast the Road Accident Using Data Mining Technique.[30]	road accident, Data mining techniques.
Rsl-24	Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity.[31]	Data mining, algorithms classification, Predicting road traffic.
Rsl-25	Severity Prediction with Machine Learning Methods.[32]	Machine learning techniques, Severity prediction.
Rsl-26	Collision analysis of driving scenarios.[33]	Data mining techniques, analysis of driving.

C. Anexo II



UNIVERSIDAD NACIONAL DE LOJA

Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables

Carrera de Ingeniería en Sistemas

Entrevista

Entrevista realizada con el propósito de sustentar y justificar la realización del presente trabajo de titulación, tanto a nivel social como académico. Como apoyo de la información captada en la entrevista, se adjunta la grabación de la misma, en el siguiente enlace:

https://drive.google.com/file/d/1h97fkUCJ46Pov4VEOnz9_tuus79Xpfdm/view?usp=sharing.

A continuación, se redacta fielmente las respuestas vertidas por su autor.

Cargo: Jefe Operativo de la Unidad de Control Operativo de Tránsito de Loja.

Nombre: Ing. Mercedes Victoria Torres Pereira.

Fecha: 23 de julio del 2021.

Descripción:

1. ¿Actualmente cómo es el registro actual de los partes por accidentes de tránsito y que funciones cumplen estos documentos?

Los partes que realizamos los agentes civiles de tránsito, estos se los realiza a través de un documento digital y físico, el cual es elaborado en el departamento de redacción de partes por el personal uniformado, estos son de vital importancia porque en sí, reflejan las posibles circunstancias en la que se produjo el siniestro o lo que conocemos como accidente de tránsito; este documento tiene además una base legal puesto que el mismo es utilizado para la investigación y estudio de los factores que incidieron en este caso para que ocurra el accidente de tránsito, además de ser necesario en las diferentes instancias judiciales.

2. ¿Cómo obtienen la información con respecto a los accidentes suscitados en la ciudad de Loja?

Bueno, el agentes de tránsito una vez de que llega al lugar de los hechos, lo que hace es recabar la información, cuando hablamos de recabar la información esto implica de que toma contacto directamente con los afectados o con personas que estuvieron en el lugar de los hechos y nos pueden dar información, respecto a cómo se suscitó en este caso el accidente de tránsito; tomemos en consideración que, desde que el Gobierno Autónomo Descentralizado de Loja, tomó competencia, los agentes de tránsito, todo lo que se refiere dentro del cantón,



están autorizados para recabar esta información, considerando en este caso todo lo que respecta a por ejemplo direcciones, nombres de conductores , tipología del accidente, es muy importante tener placas vehiculares, títulos habilitantes como la matrícula, identificación del ingreso vehicular, y en este caso se refleja lo que es la información del Agente que tomó procedimiento; todo esto en razón incluso a factores climatológicos, y como se encontraba la vía al momento del incidente; toda esta información y se respalda en documentos que eleva a conocimiento de la autoridad competente.

3. ¿Cómo realizan el estudio y análisis de los lugares donde existe mayor número de accidentes de tránsito en el cantón?

Al ser una Unidad de Control Operativo de Tránsito, toda la documentación en este caso el parte, este es registrado en una base de datos, en una matriz de un documento excel, esto nos permite cada uno de los accidentes que se suscitan en los diferentes lugares del cantón, esta misma base de datos se almacena dentro de la información institucional, la cual después semanalmente y de forma mensual, se emite un respaldo de información estadística, la cual es enviada hacia la administración municipal y hacia la dirección de la Agencia Nacional de Tránsito ubicada aquí en Loja. Se cuenta además con una oficina de archivo, en este espacio se almacenan cada una de las informaciones como respaldo de los procedimientos que toma nuestro personal uniformado.

4. ¿Poseen un respaldo de la base de datos operacional de accidentes de tránsito en la nube?


En la nube no tenemos en realidad este respaldo de base de datos, directamente se los tiene registrados como se había mencionado en excel, en las computadoras, se tiene también de soporte las unidades de almacenamiento, con la finalidad sobre todo a dar la atención y administración de la documentación, contemplando las leyes así lo requieren.



UNIVERSIDAD NACIONAL DE LOJA
Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables

Carrera de Ingeniería en Sistemas

Entrevista realizada por: Patricio Bolívar Benítez Lanche.

Nombre	Institución/Empresa	Firma
Mercedes Victoria Tones Pereira	Unidad de Control Operativo de Tránsito	

D. Certificado de acceso a los datos

Loja, 21 de septiembre de 2021

Crnl. Paúl Aguilar Sotomayor

DIRECTOR ESTRATÉGICO DE LA UNIDAD DE CONTROL OPERATIVO DE TRÁNSITO

Presente

Yo, Patricio Bolívar Benítez Lanche, con cédula de identidad número 1105665044, estudiante del **NOVENO CICLO** de la Universidad Nacional de Loja, en la carrera de **INGENIERÍA EN SISTEMAS**, me dirijo a usted en esta oportunidad para solicitarle tenga a bien aceptar la solicitud de brindar el acceso a los archivos en formato abierto y reutilizable (xlsx o csv) de los accidentes de tránsito ocurridos en el cantón pertenecientes a los años 2019 y 2020 en el que consten la información detallada de cada uno de los siniestros de tránsito, con la finalidad de poder realizar un estudio e implementación en mi trabajo de titulación referente a la Minería de Datos.

Además, declaro conocer y aceptar las Condiciones de Uso de las Bases de Datos de la Unidad de Control Operativo de Tránsito - UCOT y por medio de este instrumento me hago responsable de asegurar su buen uso.

Atentamente



Patricio Bolívar Benítez Lanche
Estudiantes de CIS – UNL
patricio.benitez@unl.edu.ec
0992182051

