Universidad Nacional de Loja

Facultad de la 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 ON VEHICLE ACCIDENT RATES IN THE URBAN AREA OF LOJA CANTON

Trabajo de Titulación previo a la obtención del título de Ingeniero en Sistemas

AUTOR:

Patricio Bolívar Benítez Lanche

DIRECTOR:

Ing. Edison Coronel Romero, Mg. Sc

Loja – Ecuador

2023

Certificación

Loja, 17 de marzo del 2023

Ing. Edison Leonardo Coronel Romero, Mg. Sc.
DIRECTOR DEL TRABAJO DE TITULACIÓN

CERTIFICO:

Que he revisado y orientado todo el proceso de elaboración del Trabajo de Titulación denominado: **"MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA"**, previo a la obtención del título de **Ingeniero en Sistemas**, de la autoría del estudiante **Patricio Bolívar Benítez Lanche**, con cédula de identidad **Nro. 1105665044**, una vez que el trabajo cumple con todos los requisitos exigidos por la Universidad Nacional de Loja para el efecto, autorizo la presentación para la respectiva sustentación y defensa.

Edison Leonardo Coronel Romero, Mg. Sc.
DIRECTOR DEL TRABAJO DE TITULACIÓN

Autoría

Yo, **Patricio Bolívar Benítez Lanche** declaro ser autor del presente Trabajo de Titulación y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos, de posibles reclamos o acciones legales, por el contenido del mismo. Adicionalmente acepto y autorizo a la Universidad Nacional de Loja la publicación de mi Trabajo de Titulación, en el Repositorio Digital Institucional - Biblioteca Virtual.

Firma: 

Cédula de identidad: 1105665044

Fecha: 31 de agosto del 2023

Correo electrónico: patricio.benitez@unl.edu.ec

Teléfono: 0992182051

Carta de autorización por parte del autor, para consulta, reproducción parcial y/o total, publicación electrónica del texto completo, del Trabajo de Titulación

Yo, **Patricio Bolívar Benítez Lanche** declaro ser autor del Trabajo de Titulación denominado: **"MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA"**, como requisito para optar por el título de **Ingeniero en Sistemas**, autorizo al sistema bibliotecario de la Universidad Nacional de Loja para que, con fines académicos, muestre la producción intelectual de la universidad, a través de la visibilidad de su contenido en el Repositorio Digital Institucional.

Los usuarios pueden consultar el contenido de este trabajo en el Repositorio Institucional, en las redes de información del país y del exterior, con las cuales tenga convenio la Universidad Nacional de Loja.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia del Trabajo de Titulación que realice un tercero.

Para constancia de esta autorización, en la Ciudad de Loja, a los treinta y un días del mes de agosto del dos mil veintitrés.

Firma: 

Autor: Patricio Bolívar Benítez Lanche

Cédula de identidad: 1105665044

Dirección: Loja, Barrio "El Dorado"

Correo electrónico: patricio.benitez@unl.edu.ec

Teléfono: 0992182051

DATOS COMPLEMENTARIOS:

Director de Trabajo de Titulación: Ing. Edison Coronel Romero, Mg. Sc.

Dedicatoria

Dedico este trabajo de titulación a toda mi familia, en especial a mis padres, por ser el motor que me ha impulsado a cumplir mis metas y sueños, gracias a su esfuerzo, dedicación y cariño he logrado todo lo que me he propuesto alcanzar. Asimismo, a mis hermanos por apoyarme en todo momento, por sus consejos y por ser partícipes de los momentos más especiales de mi vida.

Además, dedico este trabajo a todas las personas que alguna vez estuvieron, a las que ahora están y en especial a las que se mantuvieron firmes hasta el final, que sin ellos no estaría hoy en estas instancias y me han impulsado a seguir adelante y conseguir una meta más.

Patricio Bolívar Benítez Lanche

Agradecimiento

Mi eterna gratitud a la carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja y a mi director de Trabajo de Titulación Ing. Edison Coronel Romero, Mg. Sc., por contribuir en mi formación profesional y humana a lo largo de estos años; además por permitirme crecer y aprender mediante sus experiencias y por su apoyo continuó.

Agradezco a la Unidad de Control Operativo de Tránsito de cantón Loja, por su colaboración y su apoyo.

A todos ustedes infinitas gracias.

Patricio Bolívar Benítez Lanche

Índice de Contenido

Portada ..... **i**

Certificación ..... **ii**

Autoría ..... **iii**

Carta de autorización ..... **iv**

Dedicatoria ..... **v**

Agradecimiento ..... **vi**

Índice de Contenido ..... **vii**

Índice de tablas: ..... **x**

Índice de figuras: ..... **xiii**

Índice de anexos: ..... **xvii**

1. Título **1**

2. Resumen **2**

2.1. Abstract 3

3. Introducción **4**

4. Marco teórico **7**

4.1. Minería de datos 7

4.2. Descubrimiento de conocimiento en bases de datos o Proceso KDD 7

4.3. Técnicas de minería de datos 9

4.4. Técnicas Supervisadas o Predictivas 9

4.5. Técnicas No supervisadas o Descriptivas 10

4.6. Algoritmos de predicción o regresión 10

4.7. Árbol de decisión (AD) 10

4.8. Herramientas para minería de datos 11

4.9. Weka 12

4.10. Python 12

4.11. OpenRefine 12

4.12. Algoritmo CART 12
4.13. Algoritmo J48 13
4.14. Bibliotecas o librerías de python 13
4.15. Trabajos relacionados 14
5. Metodología 17
5.1. Área de estudio 17
5.2. Procedimiento 18
5.3. Recursos 19
5.3.1. Recursos Científicos 19
5.3.1.1. Método científico 19
5.3.1.2. Experimentación 19
5.3.2. Recursos Técnicos 19
5.3.2.1. Técnicas de investigación 19
5.3.3. Metodología de minería de datos: Metodología KDD 20
5.4. Participantes 23
6. Resultados 25
6.1. Objetivo 1: Obtener la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos. 25
6.1.1. Etapa 1: Integración y recopilación de datos 25
Tarea 1: Establecer las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja. 25
6.1.2. Etapa 2: Selección, limpieza y transformación de datos 26
Tarea 2: Obtener la base de datos de siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019 – 2020. 26
Tarea 3: Desarrollar la transformación y limpieza de la base de datos obtenida. 27
6.2. Objetivo 2: Implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos. 58

6.2.1. Etapa 3: Aplicación de la Minería de datos 58

Tarea 1: Identificar las variables que influyen en el cometimiento del accidente de tránsito. 58

Tarea 2: 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..... 59

6.3. Objetivo 3: Evaluación de la técnica de minería de datos propuesta. 72

6.3.1. Etapa 4: Interpretación y presentación de resultados 72

Tarea 1: Analizar los resultados obtenidos del modelo entrenado. 72

Tarea 2: Evidenciar el funcionamiento del modelo con datos actuales. 80

7. Discusión 101

7.1. Desarrollo de la propuesta alternativa 101

7.1.1. Objetivo 1: Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020. 101

7.1.2. Objetivo 2: Implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos. 102

7.1.3. Objetivo 3: Evaluación de la técnica de minería de datos propuesta. 104

7.2. Valoración técnica, económica, ambiental y social..... 106

7.2.1. Valoración técnica 106

7.2.2. Valoración económica 106

7.2.3. Valoración social 108

8. Conclusiones 109

9. Recomendaciones 111

10. Bibliografía 112

11. Anexos 117

Índice de tablas:

Tabla 1. Disponibilidad de algoritmos de modelado de software aplicados a los aprendizajes supervisados, no supervisados y extendidos 11

Tabla 2. Bibliotecas de maching learning para data mining con python 13

Tabla 3. Estudios relacionados al TT 14

Tabla 4. Exactitud de la predicción de los modelos del año 2021 23

Tabla 5. Criterios de inclusión de variables para selección de bases de datos 26

Tabla 6. Variables de la base de datos año 2018 27

Tabla 7. Variables de la base de datos año 2019 28

Tabla 8. Variables de la base de datos año 2020 29

Tabla 9. Variables de la base de datos año 2021 31

Tabla 10. Variables seleccionadas año 2018 40

Tabla 11. Variables seleccionadas año 2019 40

Tabla 12. Variables seleccionadas año 2020 41

Tabla 13. Variables seleccionadas año 2021 41

Tabla 14. Variables renombradas 42

Tabla 15. Reemplazo de caracteres 43

Tabla 16. Rango de datos de variable "hora" 53

Tabla 17. Tabla de datos actuales caso 1. 54

Tabla 18. Tabla de rangos de datos (dia, hora, tipologia) caso 2. 56

Tabla 19. Tabla de rangos de datos (dia, hora, tipologia) caso 3. 57

Tabla 20. Comparación de variables 58

Tabla 21. Variables seleccionadas para el estudio 59

Tabla 22. Resultados generales de exactitud del modelo mediante Google Colab 66

Tabla 23. Variables establecidas para aplicación de algoritmo J48 69

Tabla 24. Resultados de instancias clasificadas y métricas de algoritmo J48 a variables caso 1 70

Tabla 25. Resultados de instancias clasificadas correctamente y métricas de algoritmo J48 a variables caso 2 71

Tabla 26. Resultados de instancias clasificadas correctamente y métricas de algoritmo J48 a variables caso 3. 71

Tabla 27. Valores generados por la predicción de variable "dia" prueba 1 72

Tabla 28. Valores generados por la predicción de variable "hora" prueba 1 73

Tabla 29. Valores generados por la predicción de variable "tipologia" prueba 1 74

Tabla 30. Selección de mayores porcentajes de exactitud de modelos 75

Tabla 31. Resultados generales de instancias clasificadas correctamente 76

Tabla 32. Selección de mayor exactitud de instancias clasificadas correctamente 76

Tabla 33. Comparación de modelos con mayor promedio aplicados a variables 77

Tabla 34. Selección de modelos con mayor promedio aplicados a variables 78

Tabla 35. Revisión de métricas asociadas a la exactitud del modelo 79

Tabla 36. Datos comparativos de exactitud del modelo predictivo con sus respectivas métricas 80

Tabla 37. Representación de porcentajes de clases - variable "dia" 81

Tabla 38. Probabilidades de accidentes de tránsito variable "dia" - 2021 83

Tabla 39. Representación de porcentajes de clases - variable "hora" 83

Tabla 40. Probabilidades de accidentes de tránsito variable "dia" - 2021 85

Tabla 41. Representación de porcentajes de clases - variable "tipologia" 86

Tabla 42. Probabilidades de accidentes de tránsito variable "tipologia" - 2021 88

Tabla 43. Representación de porcentajes de clases - variable "parroquia_urbana" 89

Tabla 44. Probabilidades de accidentes de tránsito variable “parroquia_urbana” – 2021 91

Tabla 45. Representación de porcentajes de clases - variable "causas" 92

Tabla 46. Probabilidades de accidentes de tránsito variable "causas" - 2021 94

Tabla 47. Representación de porcentajes de clases - variable "nro_heridos" 95

Tabla 48. Probabilidades de accidentes de tránsito variable "nro_heridos" - 2021 97

Tabla 49. Representación de porcentajes de clases - variable "nro_fallecidos"..... 98

Tabla 50. Probabilidades de accidentes de tránsito variable "nro_fallecidos" - 2021..... 100

Tabla 51. Recursos de talento humano..... 106

Tabla 52. Recursos técnicos y tecnológicos..... 106

Tabla 53. Recursos de servicios 107

Tabla 54. Recursos totales utilizados en el TT..... 107

Índice de figuras:

Figura 1. El proceso de descubrimiento del conocimiento KDD [20] 8

Figura 2. Zona de aplicación de minería de datos - Plano urbano del Cantón Loja. Fuente: GAD Municipal de Loja 17

Figura 3. Metodología propuesta para el TT [41] 21

Figura 4. Resumen de variables no consideradas que no cuentan con un valor considerable de información referente al año 2020 39

Figura 5. Resumen de variables no consideradas que no cuentan con un valor considerable de información referente al año 2021 39

Figura 6. Carga de datos en la herramienta Software "OpenRefine" 43

Figura 7. Renombre de variables de las bases de datos 44

Figura 8. Transformación de registros de datos a minúsculas 44

Figura 9. Visualización de datos transformados 44

Figura 10. Transformación personalizada de datos a través del comando value.replace() 45

Figura 11. Creación de nueva variable "dia" 45

Figura 12. Transformación de datos de variable "dia" 46

Figura 13. Estandarización y limpieza de registros de la variable "hora" 46

Figura 14. Eliminación de caracteres “1899-12-31t”, “1990-01-01t” y “-05:00” 47

Figura 15. Limpieza de signo ortográfico (tildes) de registros de datos 47

Figura 16. Limpieza de carácter especial "ñ" del registro de datos 48

Figura 17. Faceta de texto de constancia de registros similares 48

Figura 18. Corrección de gramática existente de registros de datos 49

Figura 19. Corrección de gramática de registros de datos a través de comando value.replace() 49

Figura 20. Eliminación de datos redundantes de direcciones viales a través del comando vaule.replace() 50

Figura 21. Eliminación de espacios consecutivos en los registros de datos 50

Figura 22. Ingreso de valor “0” en registros vacíos de variables numéricas 51

Figura 23. Conversión de registros de texto a registros numéricos 51

Figura 24. Corrección e ingreso de datos en celdas vacías para la variable "zona"..... 52

Figura 25. Creación de variable "dia" 52

Figura 26. Transformación de rangos de variable "hora" 53

Figura 27. Eliminación de datos de variable "zona" (rural)..... 54

Figura 28. Carga de conjuntos de datos en entorno de Google Colab 60

Figura 29. Revisión de código de librerías y carga de Dataset 60

Figura 30. Presentación general de datos cargados en el entorno de Google Colab..... 61

Figura 31. Variables excluidas para el proceso de minería de datos 61

Figura 32. Resumen de datos incluidos para proceso de minería de datos 62

Figura 33. Verificación de registros como objeto de estudio..... 62

Figura 34. Visualización de datos contenidos en variable 62

Figura 35. Transformación de datos de texto a datos numéricos..... 63

Figura 36. Verificación de transformación de datos numéricos 63

Figura 37. Presentación de datos numéricos contenidos en el dataset..... 63

Figura 38. Matriz de correlación de variables de accidentes de tránsito 64

Figura 39. División de los datos..... 65

Figura 40. Configuración del modelado predictivo 65

Figura 41. Selección de explorador "Explorer" de WEKA..... 67

Figura 42. Selección de preprocesamiento para carga de dataset 67

Figura 43. Selección de dataset específico para minería de datos 68

Figura 44. Verificación de datos de muestra para minería de datos 68

Figura 45. Selección de algoritmo J48 árboles de decisión 69

Figura 46. Representación gráfica de datos predichos para la variable "dia" prueba 1 73

Figura 47. Representación gráfica de los datos predichos para variable “hora” prueba 1 74

Figura 48. Representación gráfica de los datos predichos para variable “tipologia” prueba 1 75

Figura 49. Representación de valores de instancias clasificadas correctamente 78

Figura 50. Representación de valores seleccionados de métrica de precisión 79

Figura 51. Representación de valores seleccionados de métrica recall 79

Figura 52. Datos originales de variable “dia” 2021 81

Figura 53. Datos de predicción de modelo para variable "dia" 2021 82

Figura 54. Árbol de decisión de la variable "dia" – 2021 82

Figura 55. Datos originales de variable “hora” 2021 84

Figura 56. Datos de predicción de modelo para variable "hora" 2021 84

Figura 57. Árbol de decisión de la variable "hora" – 2021 85

Figura 58. Datos originales de variable “tipologia” 2021 87

Figura 59. Datos de predicción de modelo para variable "tipologia" 2021 87

Figura 60. Árbol de decisión de la variable "tipologia" – 2021 88

Figura 61. Datos originales de variable “parroquia_urbana” 2021 90

Figura 62. Datos de predicción de modelo para variable "parroquia_urbana" 2021 90

Figura 63. Árbol de decisión de la variable "parroquia_urbana" – 2021 91

Figura 64. Datos originales de variable “causas” 2021 93

Figura 65. Datos de predicción de modelo para variable "causas" 2021 93

Figura 66. Árbol de decisión de la variable "causas" – 2021 94

Figura 67. Datos originales de variable “nro_heridos” 2021 96

Figura 68. Datos de predicción de modelo para variable "nro_heridos" 2021 96

Figura 69. Árbol de decisión de la variable "nro_heridos" – 2021 97

Figura 70. Datos originales de variable “nro_fallecidos” 2021 98

Figura 71. Datos de predicción de modelo para variable "nro_fallecidos" 2021..... 99

Figura 72. Árbol de decisión de la variable "nro_fallecidos" – 2021 99

Índice de anexos:

Anexo 1. Entrevista dirigida al jefe de la Unidad de Control Operativo de Tránsito del Cantón Loja..... 117

Anexo 2. Base de datos inicial de accidentes de tránsito en el Cantón Loja periodo 2018 – 2021. 117

Anexo 3. Diccionario de datos de la BD inicial..... 117

Anexo 4. Registro de variables no consideradas..... 117

Anexo 5. Limpieza de los registros de bases de datos. 117

Anexo 6. Conversión de registros de datos de texto a numéricos..... 117

Anexo 7. Casos para selección de rangos de variables. 117

Anexo 8. Pruebas desarrolladas mediante aplicación de entorno Google Colab. 117

Anexo 9. Pruebas desarrolladas mediante aplicación de entorno WEKA. 117

Anexo 10. Resultados de las pruebas realizadas por los modelos en Python. 117

Anexo 11. Resultados de los modelos predictivos en Google Colab del año 2021. 117

Anexo 12. Resultados de los modelos predictivos en WEKA del año 2021. 117

Anexo 13. Ensayo argumentativo del proyecto de trabajo de titulación. 117

Anexo 14. Plano de la ciudad de Loja. 117

Anexo 15. Solicitud de base de datos de UCOT. 117

Anexo 16. Anteproyecto de minería de datos en la accidentabilidad vehicular en la zona urbana del cantón Loja. 117

Anexo 17. Base de datos finales para minería de datos. 117

Anexo 18. Pruebas desarrolladas mediante aplicación de entorno Google Colab. 117

Anexo 19. Certificado de traducción del resumen. 117

1. Título

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

2. Resumen

Los estudios sobre accidentabilidad vehicular permiten identificar los factores que inciden en un siniestro vial; por lo tanto, es imprescindible realizar este tipo de estudios, motivo por el cual este trabajo tiene como objetivo aplicar la minería de datos en la accidentabilidad vehicular en la zona urbana del cantón Loja, mediante la implementación de la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) considerando cinco etapas: (i) integración y recopilación de datos; (ii) selección, limpieza y transformación; (iii) minería de datos, (iv) interpretación y presentación de resultados; y (v) difusión y uso. Los datos analizados se obtuvieron de los registros estandarizados de accidentes de tránsito que posee la Unidad de Control Operativo de Tránsito (UCOT) durante el periodo 2018 – 2021. Utilizando la herramienta OpenRefine se realizó la selección, limpieza y transformación de datos, como la comparación de variables más influyentes dentro de los registros de tránsito. Para aplicar la minería de datos se utilizó la técnica de árboles de decisión, usando los algoritmos J48 y CART, a través de las herramientas WEKA y Python respectivamente. Se realizaron 43 pruebas diferentes donde se compararon los modelos predictivos. La herramienta Python presentó mejores niveles de rendimiento y exactitud usando las variables hora (41,62%) y parroquia urbana (34,59%); mientras que la herramienta WEKA generó mayores resultados de instancias clasificadas correctamente para las variables "dia", "tipologia", "causas", "nro_heridos" y "nro_fallecidos" con el 36,21%, 58,37%, 38,10% y 98,64 % respectivamente. Se concluyó que se puede aplicar la minería de datos en la zona urbana del cantón Loja, a través de modelos predictivos capaces de predecir la probabilidad de un accidente de tránsito en la zona urbana del cantón Loja a través de los 370 registros del año 2021, lo que permitió generar 370 porcentajes de probabilidades resultantes y patrones distintos para cada una de los atributos de accidentabilidad vehicular.

Palabras clave: *Metodología KDD, Árboles de decisión, WEKA, Python, accidentes de tránsito.*

2.1. Abstract

Studies on vehicular accident rates allow identifying the factors that affect a road accident; therefore, it is essential to conduct this type of studies, which is why this work aims to apply data mining in vehicular accident rates in the urban area of Loja, through the implementation of the methodology of Knowledge Discovery in Databases (KDD) considering five stages: (i) integration and data collection; (ii) selection, cleaning and transformation; (iii) data mining, (iv) interpretation and presentation of results; and (v) dissemination and use: (i) data integration and collection; (ii) selection, cleaning and transformation; (iii) data mining, (iv) interpretation and presentation of results; and (v) dissemination and use. The analyzed data were obtained from the standardized traffic accident records held by the Operational Traffic Control Unit (UCOT) during the period 2018 - 2021. Using the OpenRefine tool, data selection, cleaning and transformation were performed, such as the comparison of the most influential variables within the traffic records. To apply data mining, the decision tree technique was used, using the J48 and CART algorithms, through WEKA and Python tools, respectively. Forty-three different tests were performed to compare the predictive models. The Python tool showed better levels of performance and accuracy using the variables hour (41.62%) and urban parish (34.59%); while the WEKA tool generated higher results of correctly classified instances for the variables "day", "typology", "causes", "nro_injured" and "nro_dead" with 36.21%, 58.37%, 38.10% and 98.64% respectively. It was concluded that data mining can be applied in the urban area of Loja Canton, through predictive models capable of forecasting the probability of a traffic accident in the urban area of Loja Canton based on the 370 records from the year 2021. This allowed generating 370 resulting probability percentages and distinct patterns for each of the vehicle accident attributes.

Keywords: *KDD Methodology, Decision trees, WEKA, Python, Traffic accident.*

3. Introducción

Actualmente los accidentes de tránsito representan una de las causas más 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; en Ecuador, 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, siendo las más comunes como arrollamientos, atropellos, caída de pasajero, choque frontal, choque lateral, choque posterior, colisión, perdida de carril, perdida de pista, rozamientos, volcamientos, atípicos entre otros, según registra el informe estadístico presentado por Agencia Nacional de Tránsito (ANT)[1] ; 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 2021, 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 tránsito 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” [2], 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.

Por otra parte, la información recopilada por la UCOT trae consigo la oportunidad de implementar análisis de datos y procesos de estudios de datos; 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 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. Sin duda, abre una puerta para estudiar y analizar todos estos datos para proporcionar

un apoyo en la toma de decisiones y análisis de futuros patrones de comportamientos viales de responsabilidad de los entes reguladores del tránsito para mitigar estos problemas. Por tanto, la mejor forma de desarrollar un estudio de siniestros viales es implementando a la minería de datos como solución para conocer el comportamiento de accidentabilidad vehicular, en donde se permite identificar las circunstancias que motivan la ocurrencia de accidentes de tránsito, fortaleciendo la seguridad vial mediante la reducción de la tasa de accidentes de tránsito haciendo uso de predicciones y software en beneficio de la toma de decisiones en torno a la movilidad vehicular. En consecuencia, para abordar el inconveniente de accidentabilidad vehicular dentro el cantón Loja se plantea el siguiente problema ¿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?, para dicho problema, se planteó el modelo predictivo de data mining de árbol de decisión con aplicación de las herramientas de Python y WEKA, de las cuales fueron las más utilizadas por los estudios [3], [4], [5], [6], [7], [8], [9] y [10] que realizan la minería de datos en diferentes países a través de estas herramientas; de manera que, este estudio es un referente sobre la aplicación de minería de datos, que si bien existen estudios en otras ciudades como el estudio realizado por [11], pero aún no se ha evidenciado un estudio dirigido a la problemática de la ciudad de Loja. Por consiguiente, se llegó a una solución viable la cual fue la implementación de un estudio de minería de datos enfocado al comportamiento de los patrones de los accidentes de tránsito en la zona urbana del cantón Loja, utilizando las técnicas de árboles de decisión, aplicada y validada mediante modelos predictivos.

En el presente Trabajo de Titulación (TT) se presentó como objetivo general el analizar la accidentabilidad vehicular mediante la aplicación del modelo predictivo de árboles de decisión como técnica de minería de datos en la zona urbana del Cantón Loja en el periodo 2019-2020; a su vez, tres objetivos específicos como son: obtener la fuente de datos reales de accidentes de tránsito registrados en la UCOT para la utilización de minería de datos, específicamente los registrados en el año 2019 – 2020, la implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos y finalmente, la evaluación de la técnica de minería de datos propuesta; la elaboración de este trabajo de titulación fue con la motivación de cumplir el objetivo general de analizar la accidentabilidad vehicular mediante la aplicación del modelo predictivo de árboles de decisión como técnica de minería de datos en la zona urbana del Cantón Loja en el periodo 2019-2020, para brindar un apoyo a la toma de decisiones de parte de la UCOT en beneficio de la ciudadanía en general.

El Trabajo de Titulación está conformado por distintas secciones, en donde se destacan las siguientes: el marco teórico, donde se definen los conceptos teóricos y técnicos sobre minería de datos, tales como la técnica de árboles de decisión, aprendizaje supervisado y no supervisado, algoritmos de clasificación, entre otras; además, metodología aplicadas al data mining, herramientas software destacadas para el desarrollo de modelos predictivos; así también como la evidencia de trabajos relacionados al tema de estudio. Posterior se señala la sección de metodología en donde resalta el aérea de estudio, el procedimiento de cada uno de los objetivos establecidos, los recursos científicos y técnicos y las diferentes etapas que conforman la metodología KDD. Más adelante se presenta la sección de resultados, en donde se evidencia la obtención de la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos, específicamente los registrados en el año 2019 – 2020, considerando principalmente las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja, en donde se desarrolló la transformación y limpieza de la base de datos obtenida, lo que permitió verificar que los datos sean coherentes y puedan ser medibles; además, en esta sección se aplicó la implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos, en donde se identificaron las variables que influyen en el cometimiento del accidente de tránsito, datos que luego fueron utilizados para el entrenamiento de los modelo de árboles de decisión; y por último se desarrolló la evaluación de la técnica de minería de datos propuesta, lo que significó el análisis de los resultados obtenidos del modelo entrenado, y se evidenció el funcionamiento del modelo con datos actuales, a través de la comparación de resultados de datos de accidentabilidad vehicular en el cantón Loja. A continuación, se presenta la sección de discusión en donde se hace énfasis al análisis crítico de los objetivos propuestos, señalando los aportes y limitantes durante el cumplimiento de cada etapa; a su vez, en esta sección se menciona la valoración técnica y social al que aporta el estudio realizado. Luego, se señala la sección de conclusiones que reflejan el criterio técnico del suscrito, para finalizar con la sección de recomendaciones en donde hace mención a las sugerencias que puede tomarse en cuenta para futuros trabajos que se enfocados a la minería de datos siguiendo la línea de la accidentabilidad vehicular.

4. Marco teórico

En esta sección se presenta conceptos relevantes que sustentan el desarrollo del Trabajo de Titulación (TT), que permiten establecer un concepto claro sobre el tema. La información recopilada se presenta a través del proceso de revisión bibliográfica, así también se presenta las principales metodologías y técnicas utilizadas para la minería de datos, para finalmente establecer un estudio con todos los trabajos implementados dentro de la línea de investigación.

4.1. Minería de datos

La Minería de Datos descubre relaciones, tendencias, desviaciones, comportamientos atípicos, patrones y trayectorias ocultas, con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento automatizando procesos [12], siendo utilizados en un sistema de apoyo para la toma de decisiones estratégicas de una organización, definiendo un conjunto de técnicas y herramientas software que establecen la identificación y reconocimiento de patrones y algoritmos utilizados conjuntamente con la inteligencia artificial, explorando patrones y reglas ocultos en los datos contenidos sea en un Data Warehouse o base de decisión DataMart o incluso un Big Data [13][14]. Estas reglas suelen ser implícitas, pero son críticas para la toma de decisiones, por lo que han desarrollado una multitud de algoritmos de aprendizaje estadístico y computacional [15][16].

Además, tiene una serie de tareas que pueden interpretarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra [17], por lo tanto, se han identificado diferentes etapas, como lo son: la recopilación de información, análisis, predicciones, la oportuna toma de decisiones, entrenamiento en sistemas de información y desarrollo de la inteligencia de negocios. Estos aplicativos han dado pauta a la creación de herramientas que permiten la recopilación y manejo de datos [13].

4.2. Descubrimiento de conocimiento en bases de datos o Proceso KDD

El proceso de KDD con su significado en inglés Knowledge Discovery Data o Descubrimiento de conocimiento en bases de datos, es un proceso de soporte de decisión, iterativo e interactivo que combina la experiencia en un problema con una variedad de técnicas de análisis de datos tradicionales y tecnológicas avanzadas de aprendizaje automático. El objetivo es descubrir patrones y relaciones en los datos que puedan ser usados para hacer predicciones válidas

[18][19]. Además, involucra la aplicación de varios procedimientos algorítmicos para la manipulación de datos, construcción de modelos desde los datos y la manipulación de los mismos [16]. Los diferentes pasos de este proceso se presentan en la Figura 1.

Figura 1. El proceso de descubrimiento del conocimiento KDD [20]

A su vez, su proceso semiautomático consta de varios pasos; lo primero es desarrollar una comprensión del dominio de la aplicación y el conocimiento previo relevante e identificar el objetivo del proceso KDD desde el punto de vista del cliente. Luego, es crear un conjunto de datos de destino, por lo que se procede a seleccionar un conjunto de datos o centrarse en un subconjunto de variables o muestras de datos, en el que se realizará el descubrimiento; no siempre es una tarea fácil reunir esta información en una base de datos centralizada, ya que esto puede llevar a conversiones de bajo nivel. A continuación, es la limpieza y el preprocesamiento de datos. Aquí, se incluye operaciones básicas, como eliminar ruido o valores atípicos si corresponde, recopilar la información necesaria para modelar o dar cuenta del ruido, decidir sobre estrategias para manejar los campos de datos faltantes y dar cuenta de la información de secuencia de tiempo y los cambios conocidos, así como decidir la base de datos. problemas del sistema de gestión, como tipos de datos, esquema y mapeo de valores faltantes y desconocidos. Posterior es la reducción y proyección de datos: encontrar características útiles para representar los datos según el objetivo de la tarea. Con reducción de dimensionalidad o transformación de métodos, se puede reducir el número efectivo de variables bajo consideración, o se pueden encontrar representaciones invariantes para los datos. El siguiente es la elección de la función de minería de datos; esto incluye decidir el propósito del modelo derivado por el algoritmo de minería de datos (por ejemplo, resumen, clasificación, regresión y agrupación). A continuación, se elige el algoritmo de minería de datos; en donde se incluye la selección de métodos que se utilizarán para buscar patrones en los datos, como decidir qué modelos y parámetros pueden ser apropiados y hacer coincidir un método de extracción de datos en particular con los criterios generales del proceso KDD. Posterior, es la minería de datos: la búsqueda de patrones de interés en una forma de representación particular o en un conjunto de tales representaciones, incluidas

las reglas de clasificación o los árboles, la regresión y la agrupación. Después, es interpretar patrones extraídos, incluye interpretar los patrones descubiertos y posiblemente volver a cualquiera de los pasos anteriores, así como la posible visualización de los patrones extraídos, eliminando patrones redundantes o irrelevantes y traduciendo los útiles a términos comprensibles para los usuarios. Por último, es actuar sobre el conocimiento descubierto: usar el conocimiento directamente, incorporar el conocimiento en otro sistema para acciones posteriores, o simplemente documentarlo y reportarlo a las partes interesadas. Este proceso también incluye verificar y resolver posibles conflictos con el conocimiento previamente creído o extraído [14][20][18], [21].

4.3. Técnicas de minería de datos

Las técnicas de data mining para obtener un acceso eficiente a los datos, agrupar y ordenar las operaciones al acceder a los datos y optimizar las consultas constituyen los elementos básicos para escalar los algoritmos a conjuntos de datos más grandes. A su vez, las técnicas de bases de datos crean modelos que son predictivos y/o descriptivos [22][19], en donde las técnicas descriptivas son aquellas que tienen como objetivo construir un modelo a partir de un conjunto de datos para tratar de describir el mundo real al cual corresponden dichos datos; mientras que, las técnicas predictivas son aquellas que tienen como objetivo construir un modelo a partir de un conjunto de datos para tratar de predecir cómo se comportará el mundo real bajo determinadas condiciones [23].

4.4. Técnicas Supervisadas o Predictivas

Es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado presente o futuro [24]. A su vez, los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de las bases de datos que se denominan variables independientes o predictivas [17].

El propósito de estos métodos es aprender una hipótesis la cual pueda clasificar a nuevos individuos. Los algoritmos principales son: Regresión y Clasificación, Árboles de Decisión, Clasificación Bayesiana, Redes Neuronales, Algoritmos Genéticos, Conjuntos y Lógica Difusa [19][25].

4.5. Técnicas No supervisadas o Descriptivas

De acuerdo a [22] las técnicas descriptivas o no supervisadas se enfocan en encontrar patrones interpretables por humanos que describen los datos. El objetivo de estos procedimientos es la búsqueda de la caracterización o discriminación de un conjunto de datos. Las técnicas más conocidas son: agrupamiento o clustering, reglas de asociación, análisis de patrones secuenciales, análisis de componentes principales, detección de desviación [25]; en general, se basa en descubrir regularidades en los datos de cualquier índole: agrupaciones, contornos, asociaciones, valores anómalos [19].

4.6. Algoritmos de predicción o regresión

La regresión es un método matemático utilizando para crear un modelo de relación entre una variable dependiente, un número finito de variables independientes y una constante. [19] intenta determinar los valores de una o varias variables, a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión [21]. Consiste en aprender una función real que asigna a cada instancia un valor real, es decir, permite corresponder un dato con un valor real de una variable, teniendo como objetivo, minimizar el error entre el valor predicho y el valor real [24].

4.7. Árbol de decisión (AD)

Son parte del modelo predictivo, permite la construcción de diagramas de forma lógica a partir de la información que contiene una base de datos, se utiliza en la minería de datos con la finalidad de tomar decisiones convenientes desde el punto de vista probabilístico; su estructura de árbol es similar a un diagrama de flujo, donde cada nodo interno denota una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo de hoja (nodo terminal) tiene una etiqueta de clase [26][27], el nodo superior que no es hoja se llama nodo raíz, que también es el origen de cada rama. Al seguir la ruta desde el nodo raíz superior hasta un nodo de hoja inferior vinculado por conexiones, las reglas implícitas de toma de decisiones se interpretan explícitamente como si-entonces [6].

Debido a que la esencia del modelo de árbol de decisiones es una relación de regla directa entre factores y resultados, no se necesita ningún cálculo de simulación para hacer predicciones y decisiones. Esto tiene las ventajas de velocidad de cálculo, conveniencia, eficiencia y economía. Además, los requisitos sobre el conocimiento relevante de los usuarios del modelo de árbol de

decisión son relativamente bajos, lo que conduce a la popularización y aplicación de dicho modelo en diferentes campos [6].

4.8. Herramientas para minería de datos

Es claro que ante la creciente importancia de la minería de datos se han introducido una cantidad increíble de nuevas herramientas y mejoras de software en el mercado, caracterizando las capacidades que facilitan la visualización de datos, admiten interfaces con formatos de base de datos estándar y por su implementación para el manejo de Big Data, lo que permite filtrar una gran cantidad de datos, descubrir datos ocultos, revelar nuevas relaciones y patrones y extraer información anticipada y útil implícita en grandes conjuntos de datos [15].

Sin embargo, seleccionar la herramienta de minería de datos adecuada se convierte en un procedimiento desafiante; por lo que, de acuerdo al análisis realizado por Vera, Galindo, Sánchez, Salazar, Moreno, Salazar-Villalva en el año 2021, como se presenta en la Tabla 1. Disponibilidad de algoritmos de modelado de software aplicados a los aprendizajes supervisados, no supervisados y extendidos, señala los resultados de acuerdo a las herramientas más utilizadas por los principales métodos de aprendizajes de minería de datos.

Tabla 1. Disponibilidad de algoritmos de modelado de software aplicados a los aprendizajes supervisados, no supervisados y extendidos

Algorith type	WEKA	Rapid miner	TIBCO Spotfire	Alteryx
Supervised	111	84	39	46
Unsupervisado	18	68	7	19
Extended	40+ (Pentaho Community)	100+ (Rapid Miner Community originarios WEKA)	+ Algorith R, Python	30 Algorith R, Python
Total	169+	252+	46+	65+

En general, las herramientas software aplicadas a la minería de datos analizan relaciones y patrones entre los datos generados, utilizando técnicas estadísticas, redes neuronales y aprendizaje automático; a su vez, el presente estudio sugiere que las herramientas software más aplicadas al método de aprendizaje supervisado son: WEKA, Rapidminer, así como la implementación de lenguaje de programación R y Python.

4.9. Weka

Weka es un software libre disponible bajo General Public License (GNU) basada en aprendizaje automático. Se caracteriza por tener una colección de herramientas de visualización y algoritmos para el análisis de datos y el modelado predictivo, junto con interfaces gráficas intuitivas (GUI) para el fácil acceso a sus funcionalidades de usuario [13]. También tiene la capacidad para que los desarrolladores creen sus propios algoritmos de aprendizaje automático [7]; además, se pueden implementar técnicas de clasificación, asociación, agrupamiento y predicción, siendo un software multiplataforma [24].

4.10. Python

Python se ha convertido en el lenguaje de programación de facto para el análisis de datos y el aprendizaje automático, tanto para fines de investigación como para el despliegue operativo en sistemas de producción a gran escala. Es un lenguaje de secuencias de comandos que se puede usar de forma interactiva y no requiere la compilación del código fuente en un ejecutable para ejecutarse, lo que facilita la transferencia de un programa Python entre computadoras y sistemas operativos [7]. Python es compatible con la programación de procedimientos y orientada a objetos, así como también ofrece cierto soporte para metodologías de programación funcional, pero no tan sólidamente como los lenguajes puramente funcionales como Lisp y Haskell. Python puede llamar a programas C o C++ externos, y puede integrarse en otros lenguajes para implementar la capacidad de secuencias de comandos C.

4.11. OpenRefine

Es una herramienta de código abierto que puede ayudar a convertir datos sucios en datos limpios y utilizables de manera sencilla [8]. Además, se ejecuta como un servidor web en el ordenador del usuario, de tal manera que no se consumen muchos recursos del mismo. Este utiliza un navegador web como su interfaz, para que los datos se guarden en el ordenador local. Una gran ventaja que posee es que es capaz de manejar volúmenes grandes de información. En consecuencia, es utilizado por científicos, investigadores de datos, analistas de negocios, periodistas de datos y administradores de repositorios digitales en una variedad de disciplinas, los cuales necesitan datos utilizables y limpios.

4.12. Algoritmo CART

"CART" es una sigla en inglés que significa análisis de árbol regresivo y de clasificación. Al igual que los análisis de árboles de decisión, organiza los datos según opciones que compiten.

Al contrario que los algoritmos de árboles de decisión, que sólo pueden clasificar una salida o una salida numérica basada en la regresión, el algoritmo CART puede usar los dos para predecir la probabilidad de un evento [28].

4.13. Algoritmo J48

El algoritmo J48 es una implementación del algoritmo C4.5, uno de los más utilizados en la minería de datos. Este algoritmo utiliza el enfoque de divide y vencerás para generar el clasificador, calcula la ganancia de cada atributo y elige el de mayor ganancia para que sea la raíz del árbol, los datos son divididos con base en el dominio del atributo, de manera iterativa, se repite el proceso con cada partición [29].

4.14. Bibliotecas o librerías de python

Son herramientas simples y eficientes para el análisis predictivo de datos, accesible para todos y reutilizable en varios contextos y sobre todo basado en NumPy, SciPy y matplotlib de código abierto, utilizable comercialmente - Licencia BSD [30] como se presenta en la Tabla 2.

Tabla 2. Bibliotecas de maching learning para data mining con python

Nro	Librerías o bibliotecas	Descripción	URL
1	plotly	Biblioteca de gráficos declarativa de alto nivel.	https://plotly.com/python/
2	numpy	Biblioteca de operaciones matemáticas rápidas sobre matrices.	https://numpy.org/
3	pandas	Biblioteca de análisis y manipulación de datos para Python.	https://pandas.pydata.org/
4	matplotlib.pyplot	Proporciona una forma de trazado similar a MATLAB, diseñado principalmente para gráficos interactivos.	https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html
5	seaborn	Biblioteca que permite generar fácilmente elegantes gráficos, proporciona una interfaz de alto nivel.	https://pypi.org/project/seaborn/
6	statsmodels.api	Biblioteca que proporciona clases y funciones para la estimación de modelos estadísticos.	https://www.statsmodels.org/stable/index.html
7	sklearn	Librería de aprendizaje automático para preprocesamiento de los datos, creación de	https://scikit-learn.org/stable/

Nro	Librerías o bibliotecas	Descripción	URL
		modelos y optimización de hiperparámetros de los modelos.	

4.15. Trabajos relacionados

Mediante la revisión de literatura realizada enfocada al objeto de estudio, se elabora la Tabla 3. Estudios relacionados al TT, en donde se presentan los trabajos o estudios relacionados que fueron encontrados y tomados como referencia para permitir sustentar el desarrollo del presente TT.

Tabla 3. Estudios relacionados al TT

Registro	Trabajos seleccionados	Ref.	Términos	Técnica
TS01	Impacto de la aplicación de algoritmos de minería de textos en los mensajes del medio social twitter para los eventos del tráfico vehicular de la ciudad de Cuenca.	[31]	Weka	C4.5 (Árboles de decisión)
TS02	Estimación del estado del flujo de tráfico mediante preprocesado y minería de datos. Aplicación de Dataset de posiciones GPS de taxis de Porto	[4]	Metodología KDD, Weka	Árboles de decisión, Hoeffding Tree, Random forest
TS03	Análisis y predicción de la lesividad en accidentes de tráfico mediante la aplicación de random forest.	[32]	Tráfico vehicular, Metodología KDD, lenguaje R	Árboles de decisión, Random Forest
TS04	Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study	[33]	Lenguaje R	Árboles de decisión, Random forest
TS05	Big Vehicular Traffic Data Mining: Towards Accident and Congestion Prevention.	[5]	Weka	Árboles de decisión, Random forest

Registro	Trabajos seleccionados	Ref.	Términos	Técnica
TS06	Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory.	[34]	Weka, Metodología KDD	Árboles de decisión,
TS07	Análisis y aplicación de algoritmos de minería de datos.	[35]	Weka	Árboles de decisión
TS08	Determination of the Severity of Motorcycle and Tricycle Accidents in Nigeria	[6]	Python	Árboles de decisión
TS09	Minería de datos para la toma de decisiones en la unidad de nivelación y admisión universitaria ecuatoriana.	[36]	Weka	Árboles de decisión, Reglas de asociación
TS10	Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador.	[11]	Metodología KDD, lenguaje R, RStudio	Árboles de decisión
TS11	Predicting Factors Contributing to Road Trafic Accident and Implying Driver's Driving Behavior in Addis Ababa City	[37]	Metodología KDD, Weka	Árboles de decisión, A priori algorithm.
TS12	Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France	[14]	Metodología KDD, RStudio	Arboles de decisión, Reglas de asociación
TS13	Traffic crashes prediction using machine learning models, case study: Rwanda	[8]	Python	Naïve Bayes, árboles de decisión, K-vecinos cercanos
TS14	Predicción del riesgo de un accidente de tránsito en Colombia por medio del software Weka.	[9]	Weka	Árboles de decisión

Registro	Trabajos seleccionados	Ref.	Términos	Técnica
TS15	Análisis de accidentes de tránsito en zonas urbanas y rurales usando minería de datos difusa	[3]	Metodología KDD, Weka	Árboles de decisión, K-means
TS16	Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates	[10]	Weka	Árboles de decisión, Naïve Bayes
TS17	DATA MINING FOR VEHICLE TELEMETRY	[38]	Weka	Árboles de decisión, Random Forest, Naive Bayes
TS18	Analysis of traffic injury severity: an application of non-parametric classification tree techniques	[39]		Árboles de decisión, Random Forest, Naive Bayes

5. Metodología

Para llevar a cabo el desarrollo del proyecto dentro del ámbito ingenieril es indispensable contar con una metodología adecuada para la minería de datos, dado que contará con un conjunto de actividades que cuenta con distintas fases que permitirán desarrollar con eficacia y eficiencia el desarrollo del mismo, considerando que se contó con todos los recursos adecuados para su realización.

5.1. Área de estudio

El presente Trabajo de Titulación se desarrolló dentro de la Universidad Nacional de Loja, en las instalaciones de la Facultad de la Energía, las Industrias y los Recursos Naturales No Renovables, de la Carrera de Ingeniería en Sistema/Computación, a través de la colaboración brindada por parte de la Unidad de Control Operativo de Tránsito (UCOT), perteneciente al GAD Municipal de Loja; de esta manera se indica que, la presente investigación fue dirigida a la aplicación de minería de datos enfocada hacia la zona urbana del Cantón Loja, en donde se proyectó el análisis de los datos, tomados, tal como lo presenta la Figura 2, para ver con mayor detalle revisar anexo 14.

Figura 2. Zona de aplicación de minería de datos - Plano urbano del Cantón Loja. Fuente: GAD Municipal de Loja

5.2. Procedimiento

Para el cumplimiento del objeto de estudio se definieron tres fases, en donde se establecieron sus respectivas actividades, para evidenciar en detalle cada una de las etapas, ver la sección 6.

1. Se obtuvo la fuente de datos reales de accidentes de tránsito registrados en la UCOT, para la utilización de minería de datos, específicamente los registrados en el año 2019 – 2020, tal como se presenta en la sección 6.1.
    a. Se establecieron las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja, reflejada en la sección 6.1.1.
    b. Se obtuvo la base de datos de siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019 – 2020, como se presenta en la sección 6.1.20.
    c. Se desarrolló la transformación y limpieza de la base de datos obtenida, reflejada en la documentación de la sección 6.1.2.
2. Se aplicó la implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos, tal como se establece en la sección 6.2.
    a. Se identificaron las variables que influyen en el cometimiento del accidente de tránsito, a través de las comparaciones de datos comunes dentro de los registros de datos sobre accidentabilidad vehicular, presentadas en la sección 6.2.1.
    b. Se entrenó el modelo de árboles de decisión, con la información de accidentes de tránsito ocurridos en la zona urbana del cantón Loja, periodo 2019 – 2020, como se muestra en la sección 6.2.1.
3. Se desarrollaron la evaluación de la técnica de minería de datos propuesta, como se puede ver la sección 6.3.
    a. Se analizaron los resultados obtenidos del modelo entrenado, verificando los mayores porcentajes de exactitud de los modelos para cada variable, presentados en la sección 6.3.1.
    b. Se evidenció el funcionamiento del modelo con datos actuales, a través de la comparación de resultados de datos de accidentabilidad vehicular en el cantón Loja, como se evidencia en la sección 6.3.1.

5.3. Recursos

5.3.1. Recursos Científicos

Se tomó en cuenta varios métodos y metodologías para la elaboración del presente TT, apegados al estudio de datos de accidentabilidad vehicular del cantón Loja; los cuales se presentan a continuación:

5.3.1.1. Método científico

Este método es el único procedimiento que no pretende obtener resultados definitivos y que se extiende a todos los campos del saber, por lo que abarca directamente a la predicción; es decir, constituye una de las esencias claves de la ciencia, de una teoría científica o de un modelo científico. Así, el éxito se mide por el éxito o acierto que tengan sus predicciones, adaptándose a lo que ocurrirá en determinadas condiciones especificadas (Juaréz, 2018; Zarate, 2009). Por lo tanto, para la aplicación de algoritmos de minería de datos se implementó este método para establecer la calidad de los datos, puesto que esto se realizó a través de experimentos que deben poder repetirse o mediante estudios observacionales rigurosos; así mismo, se generó diversas colecciones de datos, lo que permitió desarrollar una configuración más precisa de los algoritmos establecidos, a fin de permitir el análisis, generación y evaluación de los resultados de los modelos predictivos de las diferentes variables de accidentes de tránsito consideradas para la minería de datos, con el objeto de validar la eficiencia y eficacia de estos datos para poder generar conclusiones óptimas al objeto de estudio del presente TT.

5.3.1.2. Experimentación

Se implementó este método con el propósito de evidenciar los resultados establecidos dentro de las colecciones de datos proporcionadas, de forma que, los algoritmos aplicados CART y J48 presenten los patrones temporales significativos generando clasificaciones, caracterizaciones y pueden ser identificados estadísticamente, también se identificó la influencia que una variable ejerce sobre otra, lo que permitió establecer información más exacta de los resultados obtenidos, evidenciándolos dentro de las conclusiones de este TT.

5.3.2. Recursos Técnicos

5.3.2.1. Técnicas de investigación

Observación

Se estableció la técnica de observación para analizar las variables tomadas como referencia de la base de datos de los siniestros de tránsito del Cantón Loja; así mismo para interpretar cada

uno de los resultados obtenidos mediante la aplicación de los algoritmos de árboles de decisión de minería de datos, indispensable para realizar la selección y evaluación de los datos con mayor relevancia, y obtener información más exacta, referente a los accidentes de tránsito ocasionados dentro de la zona urbana de la ciudad, establecidos dentro de la etapa 1, como se presenta en la sección 6.1.1.

Entrevista

La entrevista de investigación permite la recopilación de información detallada en vista de que la persona que informa comparte oralmente con el investigador lo concerniente a un tema específico; además, se encuentra dirigida y registrada con el propósito de favorecer la producción de un discurso conversacional, continuo y con una cierta línea argumental, no fragmentada, segmentada, pre codificado y cerrado por un cuestionario previo del entrevistado sobre un tema definido en el marco de la investigación (Juaréz, 2018). Mediante la aplicación de esta técnica se sustentó el presente TT, debido a que fue dirigida hacia la institución pública del GAD Municipal de Loja, específicamente al Departamento de la Unidad de Control Operativo de Tránsito (UCOT), quien cuenta con el personal capacitado y experto. Esta entrevista fue dirigida al Jefe de la UCOT (E) Ing. Mercedes Victoria Torres Pereira disponible en el anexo 1, en donde fue indispensable en establecer los criterios técnicos en la materia de accidentes de tránsito, dentro de lo que concierne a la zona urbana del cantón Loja, obteniendo información en cuanto a problemas de planificación, regulación y control del tránsito se refiere y además velan por la seguridad vial, lo que permitió identificar variables especificas concerniente a la accidentabilidad vehicular, presentadas en la sección 6.1.2, siendo relevante para ser interpretadas dentro de los resultados del presente TT.

5.3.3. Metodología de minería de datos: Metodología KDD

Para la elaboración del presente TT, se utilizó a la metodología KDD como referencia para desarrollar cada uno de los objetivos propuestos, a través de sus cinco etapas con la que se encuentra estructurada, como se presenta en la Figura 3, que son señaladas a continuación:

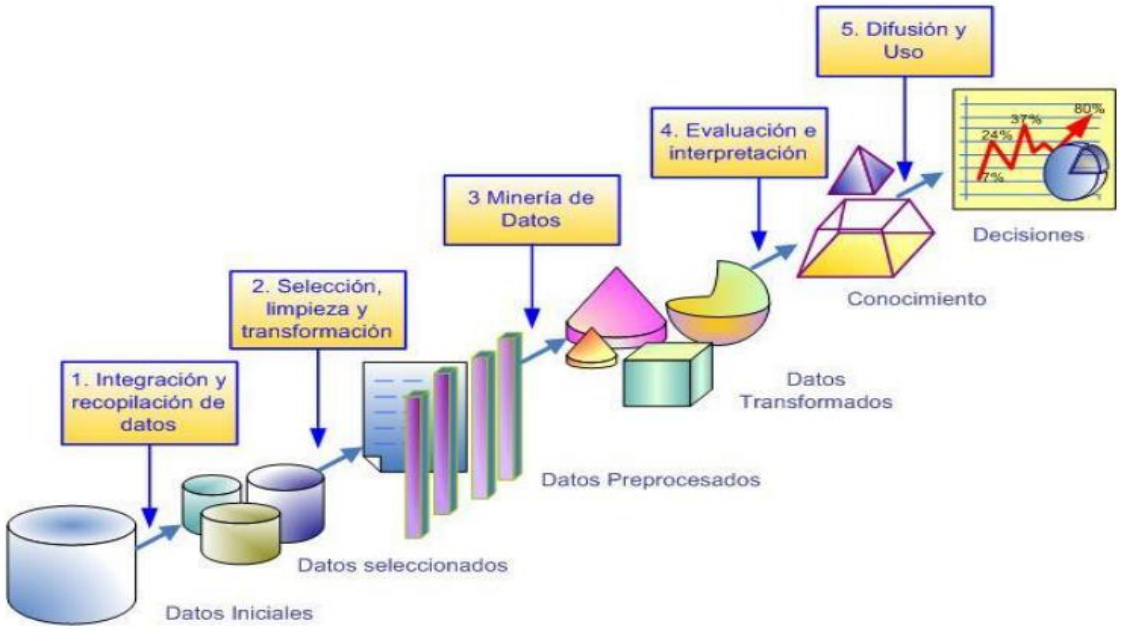


Figura 3. Metodología propuesta para el TT [41]

Etapa I: Integración y recopilación de datos

Para el desarrollo de esta etapa en el TT, se realizó la investigación acerca de los entes rectores encargados de la planificación, regulación y control del transporte terrestre y tránsito dentro del Cantón Loja, así como la búsqueda de datos abiertos, que proporcionen esta información, en donde se identificó a las entidades públicas como son: Agencia Nacional de Tránsito y GAD Municipal de Loja, con su departamento de la UCOT, como las instituciones más adecuadas para contener información veraz y confiable, enfocada hacia la accidentabilidad vehicular dentro del cantón, documentación que se evidencia en el anexo 15.

Etapa II: Selección, limpieza y transformación

En esta etapa se buscó la forma de acceder a la información recopilada por la Unidad de Control Operativo de Tránsito (UCOT) pertenecientes al GAD Municipal de Loja, para obtener las bases de datos relevantes relacionadas sobre accidentabilidad vehicular dentro de la zona urbana del cantón durante el periodo 2018 – 2021; este proceso contribuyó para conseguir una sola base de datos de acuerdo al periodo solicitado específicamente del cantón en donde se encuentra las competencias de la institución, tal como se muestra en el anexo 2; a su vez, se aplicó el proceso de cleaning, con la finalidad de identificar los datos erróneos o con menor relevancia, para poder sustituirlos o eliminarlos de la base de datos; para realizar con efectividad el Data cleaning, se desarrolló el filtrado de datos a través de la herramienta OpenRefine, siendo la que depuró, transformó variables y caracteres; así como datos que afectan al proceso de predicción

y que no están relacionados entre los cuatro periodos de accidentes de tránsito, desarrollando un registro estandarizado de variables idóneas para el proceso de data mining.

Etapa III: Minería de datos

En esta etapa se realizó la implementación del modelo se requirió el análisis de las variables contenidas en la base de datos, siendo comparados cada uno de los periodos de accidentabilidad vehicular en la zona urbana del cantón Loja, en donde coincidieron siete variables relevantes considerándolas como objeto de estudio. Las principales herramientas aplicadas al objeto de estudio fueron las herramientas de Python y WEKA; por lo tanto, para la herramienta Python se desarrolló dentro del entorno interactivo de Google Colab por facilidad de escribir y ejecutar código, en donde mediante la implementación de la librería sklearn, se aplicó el algoritmo predefinido CART, que establece la exactitud de la predicción, métricas de precisión y la métrica sensibilidad (recall) para identificar los valores positivos correctamente clasificados, que contrastan con los datos obtenidos por la herramienta WEKA, que se utilizó el algoritmo J48, que es la adaptación del algoritmo C4.5, en donde generó las instancias correctamente clasificadas, que representa la exactitud del modelo, así como los resultados de las métricas de precisión, recall y matriz de confusión; por lo que para la herramienta Python se realizaron un total de veintidós pruebas, a diferencia de la herramienta WEKA que se realizaron veintiuno pruebas, generando 43 modelos diferentes presentadas en la sección 6.2.1, siendo este proceso satisfactorio para verificar los resultados más óptimos en la toma de decisiones.

Etapa IV: Interpretación y presentación de resultados

En esta etapa se realizó el análisis de los resultados obtenidos del total de 43 pruebas realizadas para las siete variables del conjunto de datos desarrollados por las herramientas Python y Weka, para luego ser comparados los niveles de exactitud de cada predicción, eligiendo un solo modelo con mayor porcentaje de exactitud, métrica de precisión y métrica sensibilidad (recall) por cada variable, seleccionando 7 modelos predictivos resultantes; la finalidad es implementar los modelos y ejecutarlos con los nuevos registros de datos de accidentabilidad vehicular suscitados en el cantón Loja durante el año 2021; también se señala que, 2 de los modelos pertenecen a la ejecución en Python y 5 modelos fueron generados por la ejecución en WEKA.

Con el objeto de identificar las principales situaciones en las que puede suscitarse un accidente de tránsito, se determinó que las variables con mayor porcentaje de exactitud de los modelos, tal como se representa en la Tabla 4. Exactitud de la predicción de los modelos del año 2021.

Tabla 4. Exactitud de la predicción de los modelos del año 2021

Variable	Exactitud de la predicción	Métrica de precisión	Métrica recall
dia	36,21%	36,1%	36,2%
hora	41,62%	41,62%	41,62%
tipologia	58,37%	-	58,4%
parroquia_urbana	34,59%	34,59%	34,59%
causas	38,10%	-	38,1%
nro_heridos	64,59 %	-	64,6%
nro_fallecidos	98,64 %	-	98,6%

Etapa V: Difusión y uso

En esta etapa se pudo establecer que los días sábados y domingos son considerados días con mayor riesgo, puesto que presenta un porcentaje de probabilidad del 65,68%; el horario de 12:00 a 17:00 tiene el 49,5% de probabilidades en que se susciten mayor número de accidentes; así como un choque entre vehículos tiene el 74,59% de probabilidad que sea la tipología del siniestro; conducir en exceso de velocidad tiene el 37,57% de probabilidad que sea la causa de un accidente de tránsito; existe la probabilidad del 93,24% que en un accidente de tránsito no resulten personas heridas, con complicaciones físicas o algún tipo de lesión y el 100% de probabilidades que no existan personas fallecidas, todos estos datos recopilados dentro del casco urbano del cantón; consecuentemente se considera que a medida que los datos se incrementen en los registros, los porcentajes de exactitud del modelo serán más precisos; además los modelos aplicados son dinámicos y permiten adaptarse a nuevos registros de datos de accidentabilidad vehicular que en un futuro pueden implementarse como objeto de estudio.

5.4. Participantes

Para el desarrollo del presente TT enfocado a la línea de Minería de Datos, se contó con los siguientes participantes:

- Patricio Bolívar Benítez Lanche, como estudiante investigador y autor del presente TT. Inició sus actividades desde el planteamiento del tema del PTT, hasta el desarrollo y finalización de los diferentes objetivos establecidos en el presente TT, como se presenta en la sección 5.2 de Metodología.

- El Ing. Edison Leonardo Coronel Romero, Mg. Sc. como director del TT, quien supervisó los avances académicos y técnicos desarrollados por el autor del presente TT.
- La Ing. María Del Cisne Ruilova Sánchez, como tutor académico, quien supervisó los avances académicos desarrollados por el autor del presente TT.
- La Ing. Mercedes Victoria Torres Pereira, como principal actora entrevistada debido a que se desempeña como Jefe Operativa (E) de la institución pública UCOT, perteneciente al GAD Municipal de Loja, quien aportó la información y criterios técnicos sobre la accidentabilidad vehicular dentro del Cantón Loja, en donde la información proporcionada fue aplicada en la Etapa 2: Selección, limpieza y transformación de datos y Etapa 3: Aplicación de la Minería de datos de la metodología KDD, indispensable para el autor y ejecución del presente TT.

6. Resultados

En la presente sección se presenta en detalle los objetivos específicos del TT, incorporando cada una de las fases de la metodología de minería de datos establecida en la Figura 3, cumpliendo con cada uno de los objetivos propuestos.

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

En este objetivo se ejecutaron dos de las etapas de la metodología KDD, mismas que se detallan a continuación:

6.1.1. Etapa 1: Integración y recopilación de datos

Tarea 1: Establecer las directrices para la extracción de información referente a los accidentes de tránsito más frecuentes en la zona urbana del Cantón Loja.

Una vez aplicada la investigación de entes rectores encargados de la planificación, regulación y control del transporte terrestre y tránsito dentro del Cantón Loja tales como Agencia Nacional de Tránsito (ANT) y UCOT, se procedió a establecer criterios y lineamientos para obtener las bases de datos más adecuadas al presente TT, que a continuación se detallan:

Para establecer el primer criterio de inclusión, se tomaron en cuenta las variables relacionadas con accidentes de tránsito dentro de la zona geográfica urbana del Cantón Loja considerando la entrevista dirigida al jefe de la Unidad de Control Operativo de Tránsito del Cantón Loja, adjunta al anexo 1, tales como la fecha, hora, tipología del accidente, dirección del accidente, zona geográfica, parroquia en donde se produjo el accidente (urbana y rural), motivo o causa del accidente; esto tomado de referencia por los datos registrados desde el año 2018 por la UCOT, debido a que mencionada institución cuenta con información registrada desde ese año; a través de las cuales son indispensables para la identificación de patrones de accidentes de tránsito, siendo referente para el análisis y estudio de los factores que inciden en la accidentabilidad vehicular del cantón, tal como se presentan dentro de la Tabla 5. Criterios de inclusión de variables para selección de bases de datos.

Para el segundo criterio de inclusión se mencionó que, cada una de las bases de datos recopiladas debe tener libre acceso público, es decir, que se brinde la facilidad al investigador dentro los procesos administrativos para la obtención de la información.

Para el tercer criterio de inclusión se estableció que, dentro de las bases de datos a utilizar se identifiquen mínimo con siete variables como objeto a analizar, tomando como referente el estudio de investigación [9] en donde se toman como objeto de estudio a seis variables o atributos en total para la identificación de patrones para la minimización del impacto de un accidente de tránsito.

Para el cuarto y quinto criterio de inclusión se consideró que, los datos que se han recopilado se encuentren registrados desde el año 2018 al año 2021, basando su relación de variables para el análisis de datos de los años 2019 y 2020 para ser ejecutados con los datos del año 2021; además, esta selección de información debe provenir de fuentes oficiales, lo que permite el sustento y aval de los datos de las instituciones gubernamentales o públicas.

Respecto al primer y segundo criterio de exclusión se resolvió que, no se considera los datos que no aporten con contenido relacionado a accidentabilidad vehicular, así también que las bases de datos a obtener exijan procesos administrativos prolongados o impidan su conocimiento por confidencialidad de información. En cuanto al tercer y cuarto criterio de exclusión se indicó que, las bases de datos que contengan inferior a siete variables y no sean de los años especificados no serán consideradas para su proceso de estudio; a su vez, el quinto criterio de exclusión se estableció que, no será objeto de estudio toda la información que no sea obtenida a través de fuentes oficiales.

Tabla 5. Criterios de inclusión de variables para selección de bases de datos

Criterio de inclusión
Contenido relacionado
BD con libre acceso público.
Siete o más variables.
Años 2018 – 2019 – 2020 – 2021.
Fuentes oficiales.

6.1.2. Etapa 2: Selección, limpieza y transformación de datos

Tarea 2: Obtener la base de datos de siniestros de tránsito ocurridos en la zona urbana del Cantón Loja, periodo 2019 – 2020.

Selección de datos

Una vez establecidos los criterios de selección, se continuó con el proceso de obtención de las bases de datos, por lo que se tuvo únicamente acceso a cuatro bases de datos a través de la coordinación del GAD Municipal de Loja, con su departamento de la UCOT, puesto que son la única entidad que posee información fehaciente sobre la accidentabilidad vehicular dentro del cantón Loja, lo que no permitió que los protocolos establecidos en la etapa anterior pudiesen ser evaluados.

La información recopilada consta de cuatro archivos de los años 2018 al 2021 respectivamente; que se basa en los procedimientos o partes registrados de accidentes de tránsito dentro del cantón Loja por el personal uniformado de la UCOT. La información consta de entre 7 a 247 variables correspondiente a cada uno de los parámetros preestablecidos en los partes informativos realizados por la institución, acordes al diseño y aprobación por la ANT establecidos en 1960 registros, que se encuentran dentro de un repositorio en GitHub al cual se puede acceder desde el anexo 2.

Tarea 3: Desarrollar la transformación y limpieza de la base de datos obtenida.

Evaluación de las bases de datos

La base de datos obtenida de accidentes de tránsito contiene datos no procesados, así como un formato adecuado para la aplicación de los modelos necesarios para desarrollar la minería de datos. Cada base de datos contiene distintos parámetros con información relevante al accidente de tránsito, por tanto, considerando la información presentada dentro del conjunto de datos se presentan las variables como el número del siniestro, dirección del accidente, tipo del accidente, causa probable, fecha, hora, etc. Como se puede observar en la Tabla 6, Tabla 7, Tabla 8 y Tabla 9 se presentan las variables que contienen las bases de datos de los años 2018, 2019, 2020 y 2021 respectivamente.

Tabla 6. Variables de la base de datos año 2018

Nro.	Variable	Tipo de variable	Descripción
1	NRO	Numérico	Número de accidente
2	DIRECCIÓN DEL ACCIDENTE	Texto	Dirección del accidente suscitado

Nro.	Variable	Tipo de variable	Descripción
3	TIPO DE ACCIDENTE	Categórica	Tipo del accidente suscitado
4	CAUSA PROBABLE	Categórica	Causa por la que ocurrió el accidente
5	HERIDOS	Numérica	Número de personas heridas
6	PERSONAS FALLECIDAS	Numérica	Número de personas fallecidas
7	HORA	Numérica	Hora del accidente
8	FECHA	Numérica	Fecha del accidente
9	ZONA	Categórica	Región zonal en donde acontece el accidente

Tabla 7. Variables de la base de datos año 2019

Nro.	Variable	Tipo de variable	Descripción
1	NRO	Numérica	Número de accidente
2	HORA	Numérica	Hora del accidente
3	FECHA	Numérica	Fecha del accidente
4	TIPOLOGÍA	Categórica	Tipología del accidente
5	NRO. HERIDOS	Numérica	Número de personas heridas
6	NRO. FALLECIDOS	Numérica	Número de personas fallecidas
7	BARRIO	Texto	Barrio en donde ocurrió el accidente
8	PARROQUIA URBANA	Categórica	Zona parroquial en donde ocurrió el accidente
9	PARROQUIA RURAL	Categórica	Zona parroquial en donde ocurrió el accidente
10	UBICACIÓN	Texto	Dirección del accidente suscitado
11	LATITUD	Numérica	Coordenada geográfica

Nro.	Variable	Tipo de variable	Descripción
12	LONGITUD	Numérica	Coordenada geográfica
13	CAUSAS	Categórica	Causa por la que ocurrió el accidente

Tabla 8. Variables de la base de datos año 2020

Nro.	Variable	Tipo de variable	Descripción
1	N° ITEM	Numérica	Número de accidente
2	AÑO	Numérica	Año del accidente
3	FECHA	Numérica	Fecha del accidente
4	HORA	Numérica	Hora del accidente
5	LATITUD	Numérica	Coordenada geográfica
6	LONGITUD	Numérica	Coordenada geográfica
7	SEÑALIZACIÓN EXISTENTE	Categórica	Señalización en la vía
8	LUZ ARTIFICIAL	Categórica	Luz artificial
9	TIPOLOGÍA	Categórica	Tipología del accidente
10	CAUSAS	Categórica	Causa por la que ocurrió el accidente
11	PERSONAS	Numérica	Personas en el interior del vehículo
12	VEHÍCULOS RETENIDOS	Numérica	Número de vehículos retenidos
13-15	[PARTICULAR – PÚBLICO - COMERCIAL]	Numérica	Número de vehículos por tipo de servicio
16-23	[TIPO DE VEHÍCULO REGISTRADO 1 - TIPO DE VEHÍCULO REGISTRADO 5]	Categórica	Tipo de vehículos registrados

Nro.	Variable	Tipo de variable	Descripción
24	RESULTADO DE TIPO DE VEHÍCULOS	Numérica	Resultado de tipos de vehículos
25	NRO. HERIDOS	Numérica	Número de personas heridas
26	NRO. FALLECIDOS	Numérica	Número de personas fallecidas
27	GRAVEDAD DEL SINIESTRO	Categórica	Gravedad del siniestro
28-29	[NRO. PRUEBAS DE ALCOHOTEST - NRO. PRUEBAS PSICOSOMÁTICAS]	Numérica	Número de pruebas de alcohotest/psicosomáticas
30	DAÑOS MATERIALES	Categórica	Daños materiales ocasionados
31	BIEN PRIVADO	Categórica	Daños materiales al bien privado
32	DESCRIPCIÓN AL DAÑO BIEN PARTICULAR	Categórica	Descripción del daño
33	DAÑOS OCASIONADOS AL BIEN PÚBLICO	Categórica	Daños materiales al bien público
34	DESCRIPCIÓN AL DAÑO BIEN PÚBLICO	Categórica	Descripción del daño
35	RESULTADOS CONSECUENCIAS	Categórica	Resultados/consecuencias del accidente
36	ZONA	Categórica	Región zonal del accidente
37	BARRIO	Categórica	Barrio den donde ocurrió el accidente
38	PARROQUIA URBANA	Categórica	Zona parroquial en donde ocurrió el accidente
39	PARROQUIA RURAL	Categórica	Zona parroquial en donde ocurrió el accidente

Nro.	Variable	Tipo de variable	Descripción
40	DIRECCIÓN REGISTRADA COMPLETA	Texto	Dirección del accidente
41	REFERENCIA	Texto	Punto de referencia del accidente
42-135	COMPROBACIÓN INVENTARIO 1 – SERVICIO VEHICULO 3	Categórica	Comprobaciones y registros vehiculares

Tabla 9. Variables de la base de datos año 2021

Nro.	Variable	Tipo de variable	Descripción
1	N° ITEM	Numérica	Número de accidente
2	SINIESTROS	Categórica	Siniestro de tránsito
3	FECHA	Numérica	Fecha del accidente
4	AÑO	Numérica	Año del accidente
5	HORA	Numérica	Hora del accidente
6	PARTE DEL DÍA	Categórica	Noción temporal
7	LATITUD	Numérica	Coordenada geográfica
8	LONGITUD	Numérica	Coordenada geográfica
9	CONDICIÓN CALZADA	Categórica	Condición de la calzada
10	CONDICIÓN ATMOSFÉRICA SINET	Categórica	Condición atmosférica
11	CONDICIÓN VÍA	Categórica	Condición o estado de la vía
12	LUGAR EN LA VÍA	Categórica	Sección de la vía
13	SEÑALIZACIÓN EXISTENTE	Categórica	Señalización de tránsito

Nro.	Variable	Tipo de variable	Descripción
14	TIPOLOGÍA	Categórica	Tipología del accidente
15	CAUSAS	Categórica	Causa del accidente
16	PERSONAS DETENIDAS	Numérica	Número de personas detenidas
17	VEHÍCULOS RETENIDOS	Numérica	Número de vehículos retenidos
18	VEHÍCULOS INVOLUCRADOS EN EL SINIESTRO	Numérica	Número de vehículos involucrados en el accidente
19	SUMA PÚBLICOS Y COMERCIALES	Numérica	Número del servicio del vehículo
20	FILTRO PARA PÚBLICO Y COMERCIAL	Categórica	Novedades con los vehículos
21	PARTICULAR	Numérica	Número de vehículos de servicio particular
22	PÚBLICO	Numérica	Número de vehículos de servicio público
23	COMERCIAL	Numérica	Número de vehículos de servicio comercial
24	SUMA DE VEHÍCULOS POR SERVICIO	Numérica	Número de total de vehículos por servicio
25-32	[TIPO DE VEHÍCULO REGISTRADO 1 – TIPO DE VEHÍCULO REGISTRADO 5]	Categórica	Registro del vehículo
33	RESULTADO DE TIPO DE VEHÍCULOS	Numérica	Número de total de tipos de vehículos involucrados
34	RESULTADO	Categórica	Tipos de vehículos involucrados en el accidente

Nro.	Variable	Tipo de variable	Descripción
35	CUADRE DE VEHÍCULOS RETENIDOS	Categórica	Posición del vehículo
36	NRO. HERIDOS	Numérica	Número de personas heridas
37	NRO. FALLECIDOS	Numérica	Número de personas fallecidas
38	SUMATORIA PARA DETERMINAR LA GRAVEDAD DEL SINIESTRO	Numérica	Grado de gravedad del siniestro
39	GRAVEDAD DEL SINIESTRO	Categórica	Tipo de gravedad del accidente
40-41	[NRO. PRUEBAS DE ALCOHOTEST – NRO. PRUEBAS PSICOSOMÁTICAS]	Numérica	Número de pruebas de alcohotest/psicosomáticas
42	DAÑOS MATERIALES	Categórica	Verificación de daños materiales
43	BIEN PRIVADO	Categórica	Verificación de daños materiales al bien privado
44	DESCRIPCIÓN AL DAÑO BIEN PARTICULAR	Categórica	Descripción del daño al bien particular
45	DAÑOS OCASIONADOS AL BIEN PÚBLICO	Categórica	Verificación de daños materiales al bien público
46	DESCRIPCIÓN AL DAÑO BIEN PÚBLICO	Texto	Descripción del daño al bien público
47	CADENA DE CUSTODIA	Categórica	Existencia de cadena de custodia
48	NRO. CADENA DE CUSTODIA	Numérica	Número de registros de cadena de custodia

Nro.	Variable	Tipo de variable	Descripción
49	RESULTADOS CONSECUENCIAS	Categórica	Consecuencia del accidente de tránsito
50	ZONA	Categórica	Región zonal del accidente
51	BARRIO	Texto	Barrio den donde ocurrió el accidente
52	PARROQUIA URBANA	Categórica	Zona parroquial en donde ocurrió el accidente
53	PARROQUIA RURAL	Categórica	Zona parroquial en donde ocurrió el accidente
54	DIRECCIÓN REGISTRADA COMPLETA	Texto	Dirección del accidente
55-57	[CALLE / AV. PRINCIPAL (1) - CALLE / AV. PRINCIPAL (3)]	Categórica	Calles principales del accidente
58-60	[REFERENCIA FILTRO – REFERENCIA]	Categórica	Puntos de referencia del accidente
61-62	[RETENIDO POR SINIESTRO VEHÍCULO 1 – PLACA VEHÍCULO 1]	Categórica	Retención de vehículo e identificación vehicular
63	MARCA/MODELO VEHÍCULO 1	Categórica	Marca del vehículo
64	TIPO DE VEHÍCULO 1	Categórica	Tipo del vehículo involucrado
65-71	[SUBTIPO DE VEHÍCULO 1 – RETENIDO VEHÍCULO 1]	Categórica	Características del vehículo involucrado

Nro.	Variable	Tipo de variable	Descripción
72-81	[TIPO DOCUMENTO CONDUCTOR - VEHÍCULO 1]	Categórica	Características del conductor involucrado
82-84	[RETENIDO POR SINIESTRO VEHÍCULO 2 – PLACA VEHÍCULO 2]	Categórica	Retención de vehículo e identificación vehicular
85	MARCA/MODELO VEHÍCULO 2	Categórica	Marca del vehículo involucrado
86	TIPO VEHÍCULO 2	Categórica	Tipo del vehículo involucrado
87-93	[SUBTIPO DE VEHÍCULO 2 – VEHÍCULO RETENIDO 2]	Categórica	Características del vehículo involucrado
94-103	[TIPO DOCUMENTO CONDUCTOR VEHÍCULO 2 – ESTADO CONDUCTOR VEHÍCULO 2]	Categórica	Características del conductor involucrado
104-105	[RETENIDO POR SINIESTRO VEHÍCULO 3 - PLACA 3]	Categórica	Retención de vehículo e identificación vehicular
106	MARCA/MODELO VEHÍCULO 3	Categórica	Marca del vehículo
107	TIPO VEHÍCULO 3	Categórica	Tipo del vehículo involucrado
108-114	[SUBTIPO DE VEHÍCULO 3 – VEHÍCULO RETENIDO 3]	Categórica	Características del vehículo involucrado

Nro.	Variable	Tipo de variable	Descripción
115-124	[TIPO DOCUMENTO CONDUCTOR – CONDUCTOR VEHÍLO 3]	Categórica	Características del conductor involucrado
125-247	[INVENTARIO CRV VEHÍCULO 4 – SERVICIO VEHÍCULO 42]	Texto	Comprobaciones y registros vehiculares

Es necesario recalcar que, una vez verificadas las variables de la base de datos, fue necesario desarrollar un diccionario de datos en donde se presenta cada una de las variables conjuntamente con su descripción y la categoría a que pertenecen, tal y como se evidencia dentro del anexo 3.

Para la evaluación de cada una de las variables de las bases de datos se realizó un análisis exploratorio de datos en donde se consideró que los registros cuentan con variables irrelevantes, mismas que no son de gran influencia para la ocurrencia de un accidente de tránsito, de acuerdo a lo señalado por el jefe operativo de la UCOT.

Respecto a la data de accidentes de tránsito ocurridos en el cantón Loja durante el año 2018, se eliminó la variable "Nro", debido a que el número del accidente no es relevante para el cometimiento de un accidente de tránsito, también, se mantuvo las variables "FECHA", "HORA", "TIPOLOGÍA", "DIRECCIÓN", "ZONA", "CAUSA", "NRO. HERIDOS", "NRO. FALLECIDOS", inclusive se generó una nueva variable "PARROQUA URBANA", en donde sectoriza a través de la zona parroquial, permitiendo identificar los distintos factores para el cometimiento del siniestro.

Referencia a la data obtenida del año 2019, se elimina la variable "NRO", por lo que esta variable es irrelevante para identificar alguna situación para la ocurrencia del accidente de tránsito. Posterior, mediante el análisis exploratorio de datos se consideró mantener las siguientes variables que contienen la información de tiempo, sectorización, causas y riesgo humano, que son de relevancia para el cometimiento de un accidente de tránsito:

- "HORA"
- "FECHA"

- “TIPOLOGÍA”
- “NRO. HERIDOS”
- “NRO. FALLECIDOS”
- “PARROQUIA URBANA”
- “UBICACIÓN”
- “ZONA”
- “LATITUD”
- “LONGITUD”
- “CAUSAS”

Las variables “BARRIO” no se consideró por motivo que contiene información redundante con respecto a la variable “PARROQUIA URBANA”, en donde se encuentra ya especificada la zona parroquial urbana; a su vez, la variable “PAROQUIA RURAL” tampoco fue considerada por lo que no se apega al objeto de estudio de accidentabilidad vehicular en la zona urbana del cantón; también se procedió a eliminar las variables “LATITUD” y “LONGITUD” puesto que son irrelevantes para el cometimiento de un siniestro vial.

Respecto a la base de datos del año 2020 y 2021, mediante el análisis exploratorio de datos se estableció mantener las variables referentes a tiempo, sectorización, causas y riesgo humano, que son de relevancia para el cometimiento de un accidente de tránsito:

- “FECHA”
- “HORA”
- “LATITUD”
- “LONGITUD”
- “SEÑALIZACIÓN EXISTENTE”
- “TIPOLOGÍA”
- “CAUSAS”
- “VEHÍCULOS RETENIDOS”
- “NRO. HERIDOS”
- “NRO. FALLECIDOS”
- “GRAVEDAD”
- “ZONA”
- “PARROQUIA URBANA”
- “DIRECCIÓN REGISTRADA COMPLETA”

Además, no se consideró en la base de datos del año 2021 de las anteriormente señaladas las variables "CONDICIÓN CALZADA" y "CONDICIÓN ATMOSTERICA SINET", por motivo que los factores presentes dentro de un accidente de tránsito son: factor vía y factor entorno, además la variable "CONDICIÓN VÍA" es redundante, ya se encuentra establecida dentro de la variable "CONDICIÓN CALZADA".

La variable "N° ITEM" fue eliminada por lo que esta variable es irrelevante para identificar alguna situación para la ocurrencia del accidente de tránsito. De igual manera, se eliminó las variables "AÑO" por lo que contiene información redundante a la variable "FECHA", al igual que la variable "LUZ ARTIFICIAL" se encuentra descrita dentro de la variable "SEÑALIZACIÓN EXISTENTE". Tampoco se consideró las siguientes variables debido a que no aportan con información necesaria para identificar la razón suscitada de un accidente de tránsito, así mismo, poseen información redundante a la variable "VEHÍCULOS RETENDOS", "PARTE DEL DÍA", "PARTICULAR", "PÚBLICO", "COMERCIAL", "TIPO DE VEHÍCULO REGISTRADO" y "RESULTADO DE TIPOS DE VEHÍCULOS".

Del mismo modo, las siguientes variables fueron eliminadas ya que no aportan con información relevante para identificar el motivo causal del siniestro vial, tal como "NRO. PRUEBAS DE ALCOHOTEST", "NRO. PRUEBAS PSICOSOMÁTICAS", y "RESULTADOS CONSECUENCIAS"

Las variables "BARRIO" y "PARROQUIA RURAL" no fueron consideradas debido a que cuentan con información redundante con respecto a la variable "PARROQUIA URBANA", en donde ya consta la zona parroquial establecida.

Del mismo modo dentro de la base de datos del año 2021 las variables "CALLE/AV. PRINCIPAL (1)", "CALLE/AV. PRINCIPAL (2)" y "CALLE/AV. PRINCIPAL (3)" se optimizó creando una sola variable denominada "DIRECCIÓN REGISTRADA COMPLETA" para equilibrar su información con los anteriores registros de datos.

Dentro de las bases de datos desde la variable "REFERENCIA FILTRO" hasta la variable "SERVICIO VEHÍCULO 42", no se tomaron como referencia debido a que constan de comprobaciones y registros vehiculares, mismos que no cuentan con un valor considerable de información, por el hecho que existen variables que no superan el 60% de datos o en algunos casos con variables que se encontraban completamente vacías, como se presenta en la Figura 4 y Figura 5, lo que no aporta suficiente información para identificar las causas pertinentes sobre

el incidente de siniestros vehiculares de tránsito, y en los datos no considerados adjuntos en el **¡Error! No se encuentra el origen de la referencia.**.

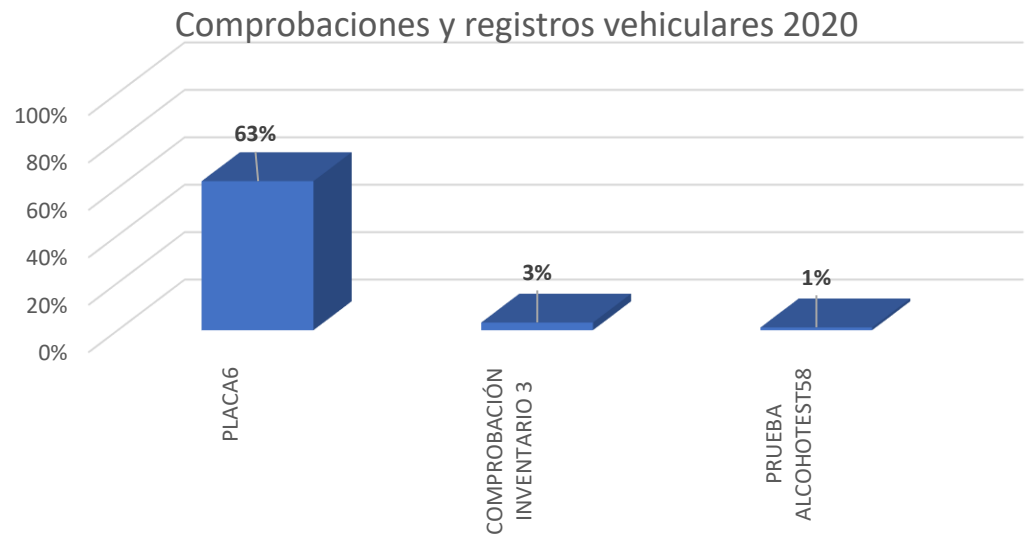


Figura 4. Resumen de variables no consideradas que no cuentan con un valor considerable de información referente al año 2020

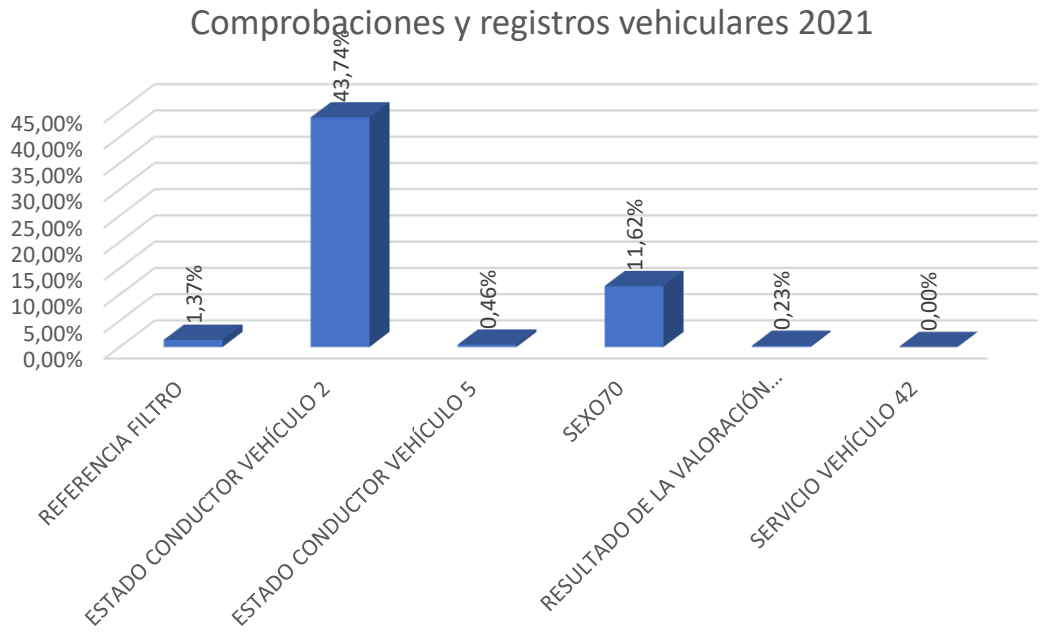


Figura 5. Resumen de variables no consideradas que no cuentan con un valor considerable de información referente al año 2021

Luego de haber realizado el análisis del registro de datos y no haber considerado variables que no aportan con información necesaria para el estudio de accidentabilidad vehicular, se generó la Tabla 10, Tabla 11, Tabla 12 y Tabla 13 con cada una de las variables seleccionadas para la aplicación de minería de datos con referencia a las investigaciones de [9] y obtener patrones que justifiquen los factores entorno y vía implicados en la incidencia de un siniestro de tránsito.

Tabla 10. Variables seleccionadas año 2018

Nro.	Variables seleccionadas	Tipo de variable
1	FECHA	Categórica
2	HORA	Categórica
3	TIPOLOGÍA	Categórica
4	DIRECCIÓN	Categórica
5	ZONA	Categórica
6	PARROQUIA URBANA	Categórica
7	CAUSA	Categórica
8	NRO. HERIDOS	Numérica
9	NRO. FALLECIDOS	Numérica

Tabla 11. Variables seleccionadas año 2019

Nro.	Variables seleccionadas	Tipo de variable
1	FECHA	Categórica
2	HORA	Categórica
3	LATITUD	Numérica
4	LONGITUD	Numérica
5	TIPOLOGÍA	Categórica
6	UBICACIÓN	Categórica
7	ZONA	Categórica
8	PARROQUIA URBANA	Categórica
9	CAUSAS	Categórica
10	NRO. HERIDOS	Numérica
11	NRO. FALLECIDOS	Numérica

Tabla 12. Variables seleccionadas año 2020

Nro.	Variables seleccionadas	Tipo de variable
1	FECHA	Categórica
2	HORA	Categórica
3	LATITUD	Numérica
4	LONGITUD	Numérica
5	TIPOLOGÍA	Categórica
6	DIRECCIÓN	Categórica
7	ZONA	Categórica
8	PARROQUIA URBANA	Categórica
9	CAUSAS	Categórica
10	GRAVEDAD	Categórica
11	NRO. HERIDOS	Numérica
12	NRO. FALLECIDOS	Numérica
13	VEHÍCULOS RETENIDOS	Numérica
14	SEÑALIZACIÓN EXISTENTE	Categórica

Tabla 13. Variables seleccionadas año 2021

Nro.	Variables seleccionadas	Tipo de variable
1	FECHA	Categórica
2	HORA	Categórica
3	LATITUD	Numérica
4	LONGITUD	Numérica
5	TIPOLOGÍA	Categórica
6	DIRECCIÓN REGISTRADA COMPLETA	Categórica
7	ZONA	Categórica
8	PARROQUIA URBANA	Categórica
9	CAUSAS	Categórica
10	GRAVEDAD	Categórica
11	NRO. HERIDOS	Numérica
12	NRO. FALLECIDOS	Numérica
13	VEHÍCULOS RETENIDOS	Numérica
14	SEÑALIZACIÓN EXISTENTE	Categórica
15	CONDICIÓN CALZADA	Categórica

Nro.	Variables seleccionadas	Tipo de variable
16	CONDICIÓN ATMOSTERICA SINET	Categórica

Limpieza de base de datos

Para obtener una imparcialidad de datos y considerando que los registros de datos cuentan con similitudes en sus variables, con el fin de optimizar la información para una correcta aplicación de minería de datos, se procedió a renombrarlas, tal como se muestra en la Tabla 14. Variables renombradas. Luego, se realizó un análisis de la información que presentaba las variables y su contenido, por lo que se evidenció que contenían datos con letras mayúsculas y minúsculas, lo que afectaría al reconocimiento de información al momento de realizar la aplicación del modelo de minería de datos, por lo que se realizó la transformación de datos a letras en minúsculas.

Tabla 14. Variables renombradas

Nombre de variable original	Nombre de variable renombrada
TIPOLOGÍA	tipologia
DIRECCIÓN/UBICACIÓN/DIRECCIÓN REGISTRADA COMPLETA	direccion
PARROQUIA URBANA	parroquia_urbana
CAUSA/CAUSAS	causas
NRO. HERIDOS	nro_heridos
NRO. FALLECIDOS	nro_fallecidos
VEHÍCULOS RETENIDOS	vehiculos_retenidos
SEÑALIZACIÓN EXISTENTE	senalizacion_existente
CONDICIÓN CALZADA	condicion_calzada
CONDICIÓN ATMOSTERICA SINET	condicion_atmosterica_sinet
CONDICIÓN VÍA	condicion_via

De la misma manera, ante la existencia de palabras que contenían signos ortográficos como la tilde, de igual manera similitud de caracteres que no contaban con este signo, generó cantidad de errores ortográficos; por lo que se desarrolló una estandarización de datos para aplicación de los algoritmos, reemplazando las vocales que contenían tilde por vocales sin tilde. A su vez, ante la existencia del carácter o letra consonante “ñ”, y su complicada visualización en la herramienta software “OpenRefine” considerándolo como carácter especial, se estandarizó

mencionado carácter para que sea reemplazado por la letra “n”, tal como se presenta en la Tabla 15. Reemplazo de caracteres.

Tabla 15. Reemplazo de caracteres

	Caracter original	**Caracter reemplazado**
Tildes	Á/á	a
	É/é	e
	Í/í	i
	Ó/ó	o
	Ú/ú	u
Letra	Ñ/ñ	n

Por lo tanto, para la limpieza de los datos se utilizó la herramienta Openrefine, a través del lenguaje General Refine Expressión Lenguage (GREL), en donde se implementaron todos los registros, tal como se presenta en la Figura 6.

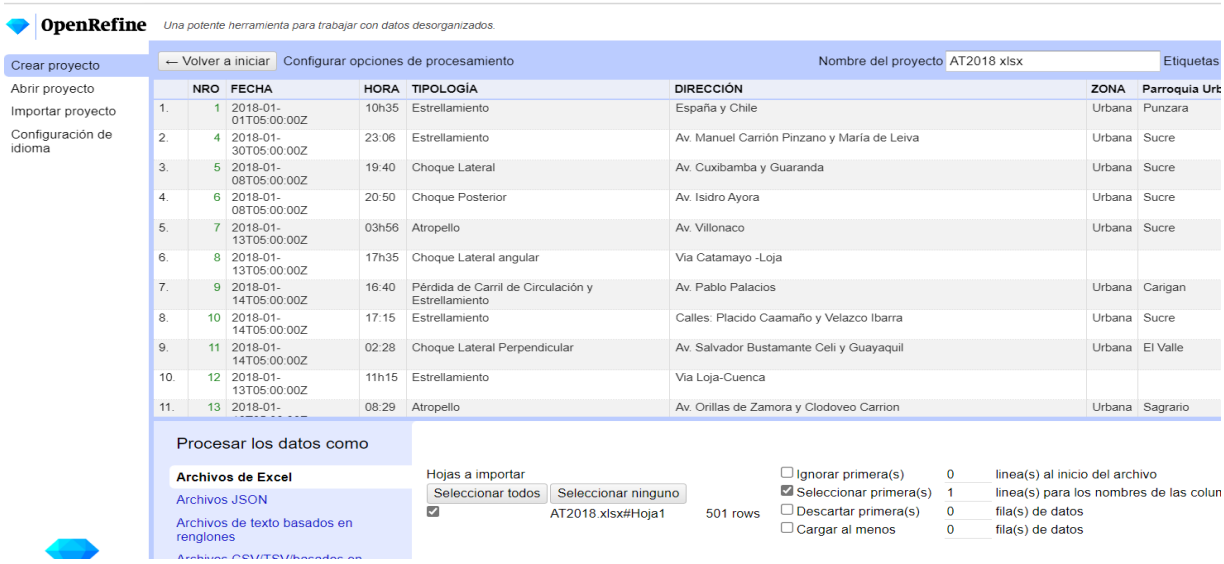


Figura 6. Carga de datos en la herramienta Software "OpenRefine"

Ya establecidos los datos dentro de la herramienta, se procedió a renombrar cada una de las variables señaladas en la Tabla 14. Variables renombradas, como se evidencia en la Figura 7.

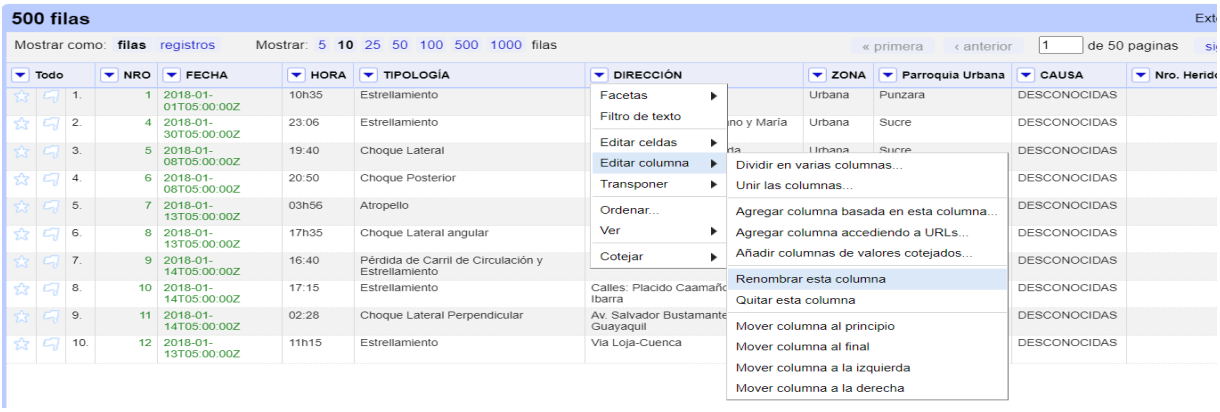


Figura 7. Renombre de variables de las bases de datos

Al renombrar las variables, se estableció el estándar de nomenclatura a utilizarse en los datos para reconocimiento de información; por tanto, en todos los registros se transformó cada uno de los datos de las variables a minúsculas, como muestra la Figura 8 y Figura 9.

Figura 8. Transformación de registros de datos a minúsculas

Figura 9. Visualización de datos transformados

Para la variable fecha, se estableció que solamente se utilizarían mes y día del accidente por lo que se reemplazó los caracteres “2018-”, tal como se evidencia en la Figura 10, así como los

caracteres “T05:00:00Z” a través del comando value.replace(“valor original”, “valor reemplazado”), como por ejemplo value.replace(“2018-”, “”).

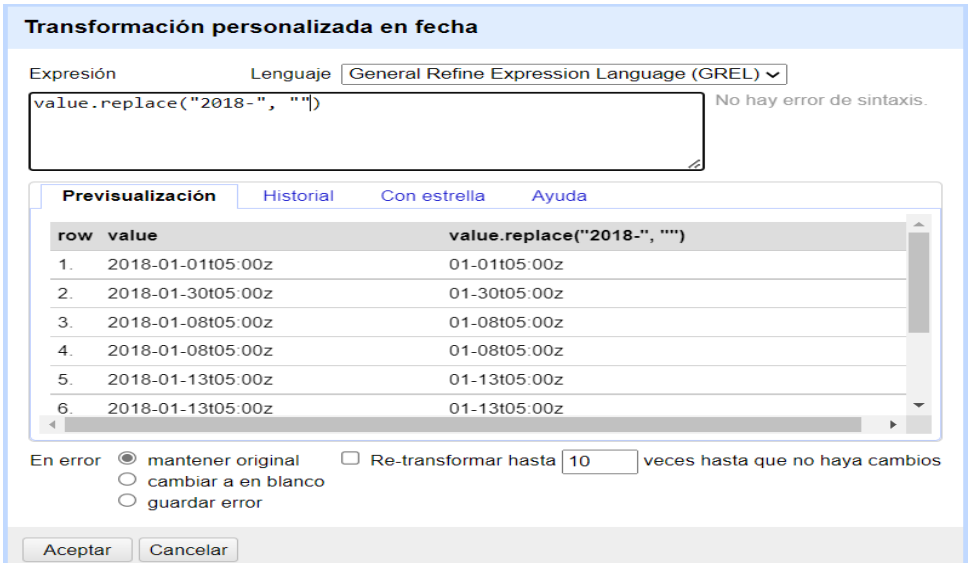


Figura 10. Transformación personalizada de datos a través del comando value.replace()

Para establecer datos más precisos, se implementó la variable “dia” en donde a través de los datos de fechas ingresadas, se generó los datos de los días de la semana en donde ocurrieron los accidentes, tal como lo indica la Figura 11 y Figura 12.

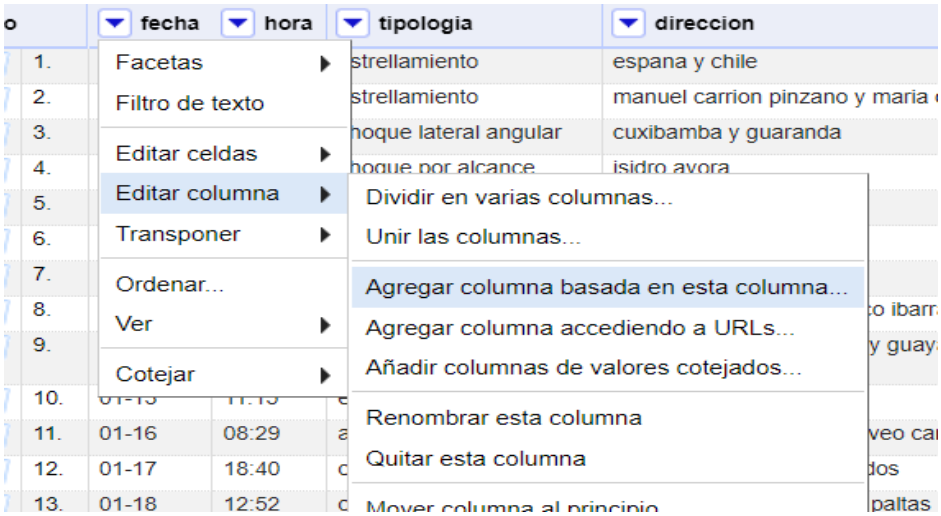


Figura 11. Creación de nueva variable "dia"

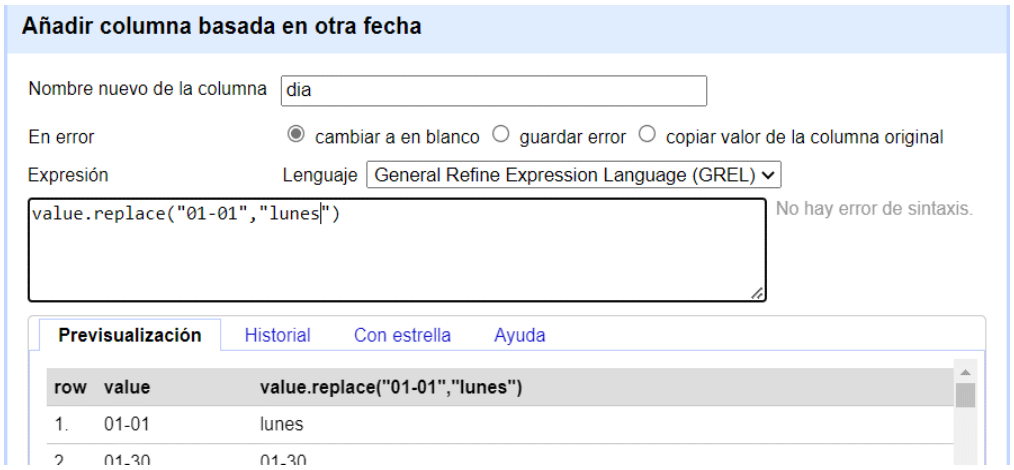


Figura 12. Transformación de datos de variable "dia"

A su vez, se estandarizó la variable “hora”, por lo que se reemplazó el carácter “h” por el carácter “:”, como se observa en la Figura 13; mientras que, se observó que dentro de esta variable se encontraban los caracteres “1899-12-31t”, “1990-01-01t” y “-05:00” por lo que se reemplazó estos datos de los registros, como se presenta en la Figura 14, así como la eliminación de espacios con el comando value.trim().

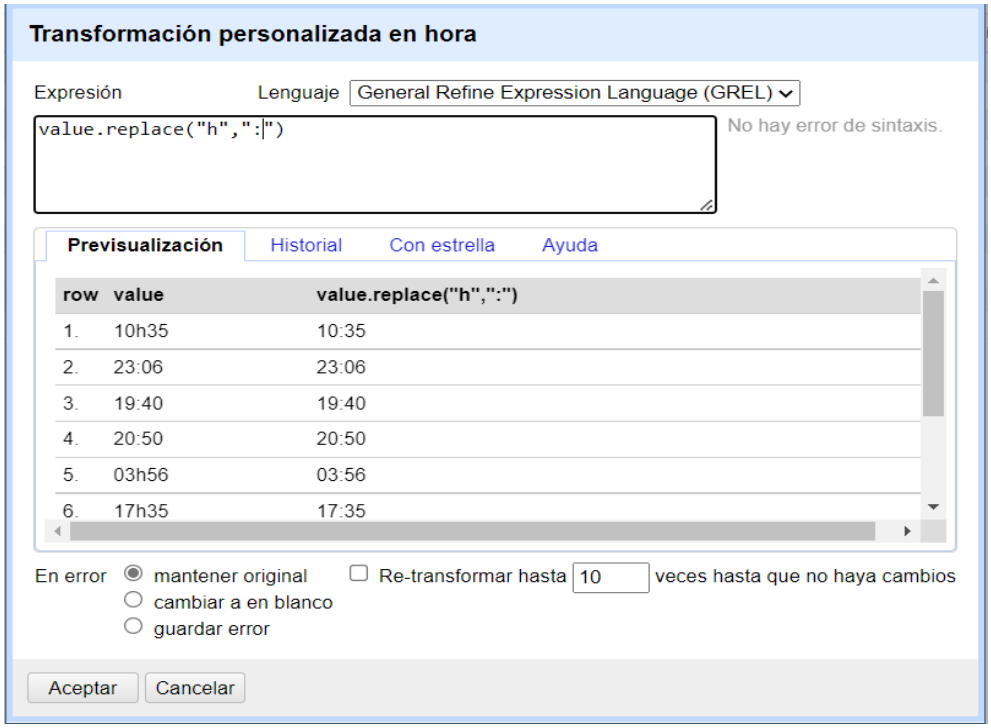


Figura 13. Estandarización y limpieza de registros de la variable "hora"

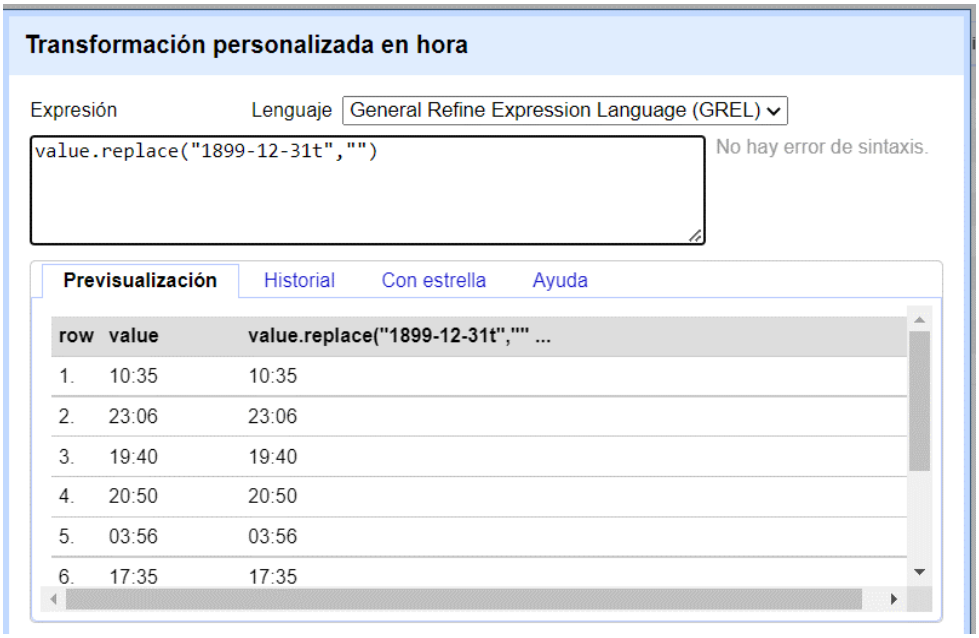


Figura 14. Eliminación de caracteres “1899-12-31t”, “1990-01-01t” y “-05:00”

Luego, se reemplazó cada una de las tildes en los registros de datos a través del comando value.replace(“valor original”, “valor reemplazado”), como por ejemplo: value.replace(“á”, “a”) como se muestra en la Figura 15; y se realizó el reemplazo del carácter “ñ” por el carácter “n”, utilizando el comando value.replace(“ñ”, “n”), establecida en la Tabla 15, presentado en la Figura 16.

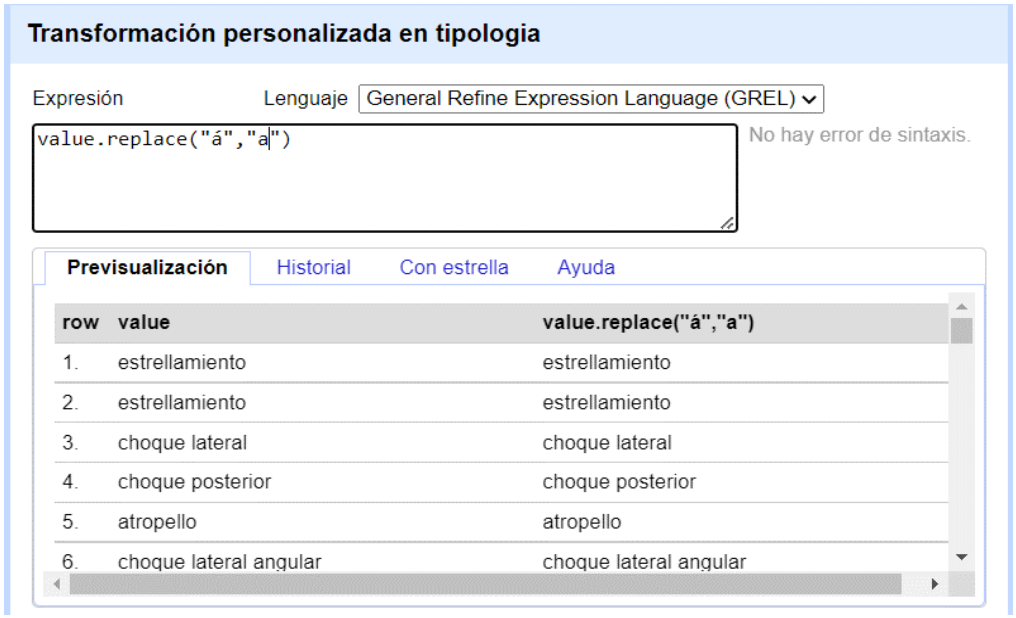


Figura 15. Limpieza de signo ortográfico (tildes) de registros de datos

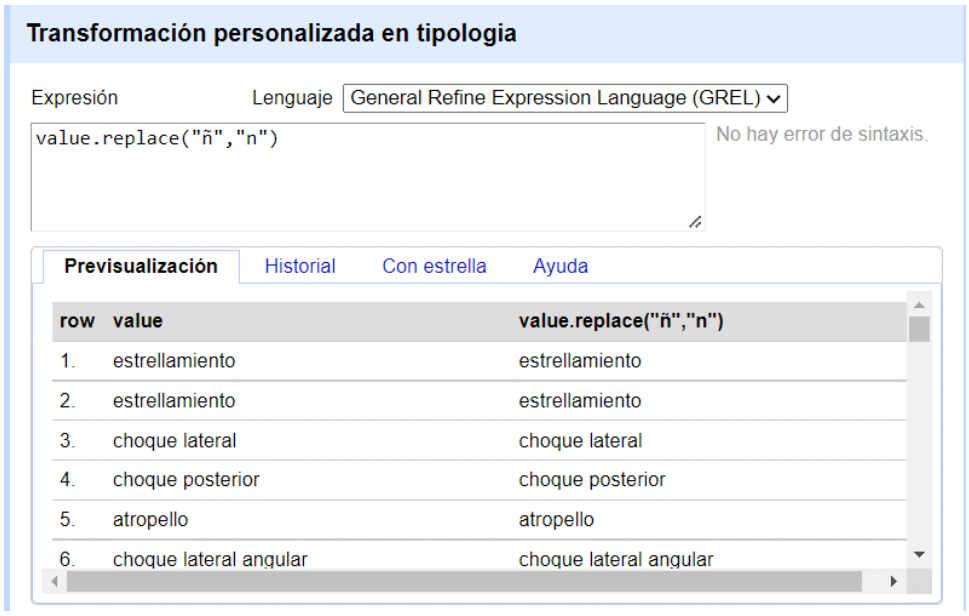


Figura 16. Limpieza de carácter especial "ñ" del registro de datos

Luego se verificó la gramática de forma manual de los datos para establecer si se encuentran similitudes en los mismos y evitar conflictos al momento de realizar la minería de datos, como por ejemplo el carácter “choquie” por “choque”, tal como se evidencia en la Figura 17 y Figura 18, así como, a través del ingreso del comando value.replace(“valor original”, “valor reemplazado”), como por ejemplo: value.replace(“alonso de mercadillo”, “mercadillo”), como se visualiza en la Figura 19.

Figura 17. Faceta de texto de constancia de registros similares

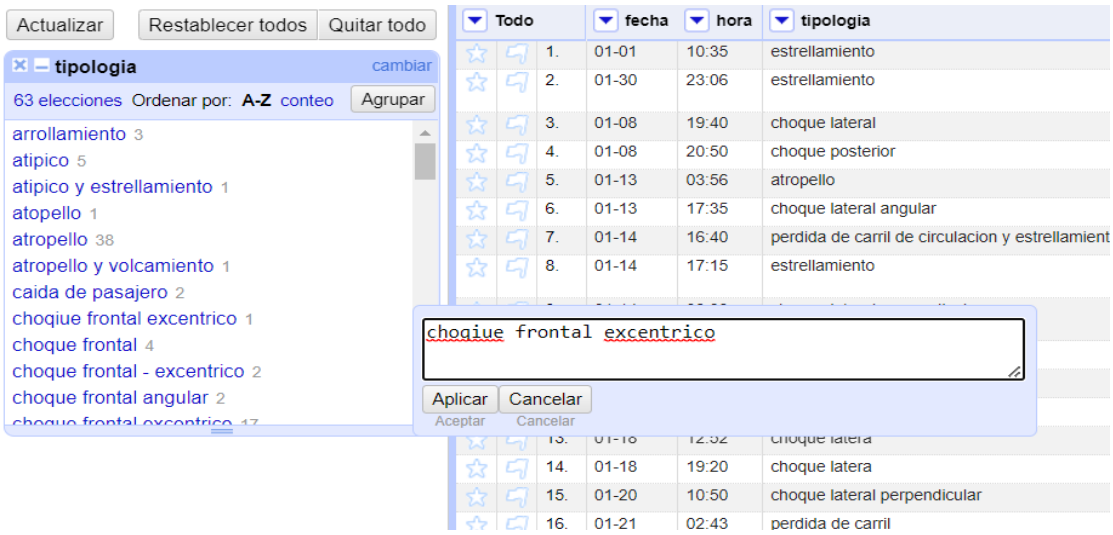


Figura 18. Corrección de gramática existente de registros de datos

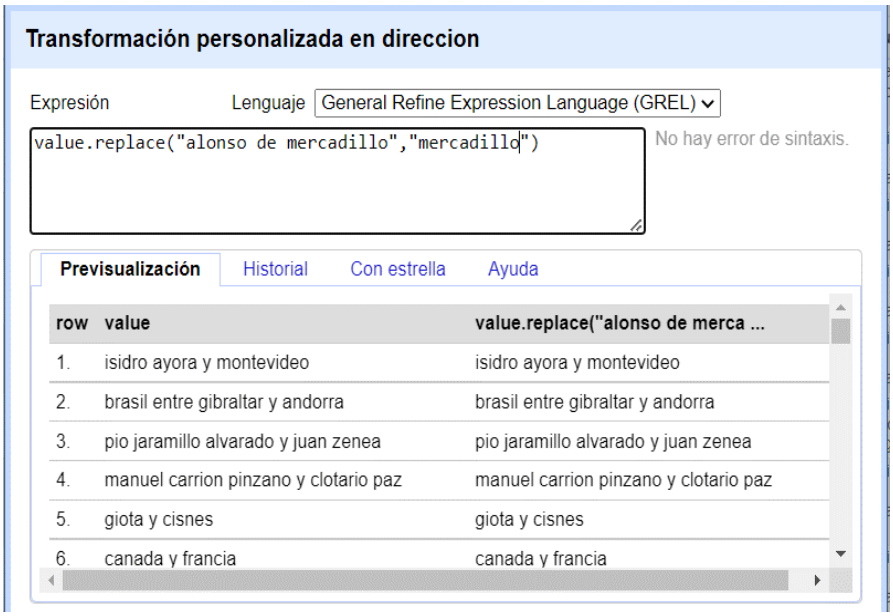


Figura 19. Corrección de gramática de registros de datos a través de comando value.replace()

Para criterios de ubicación se reemplazó en los datos de la variable “ubicacion” los registros que contenían los datos de “calle”, “avenida”, “av.”, utilizando nuevamente el comando value.replace(“valor original”, “valor reemplazado”), como por ejemplo value.replace(“calles”, “”), tal como se evidencia en la Figura 20, a su vez se implementó la contracción de espacios, tal como se observa en la Figura 21.

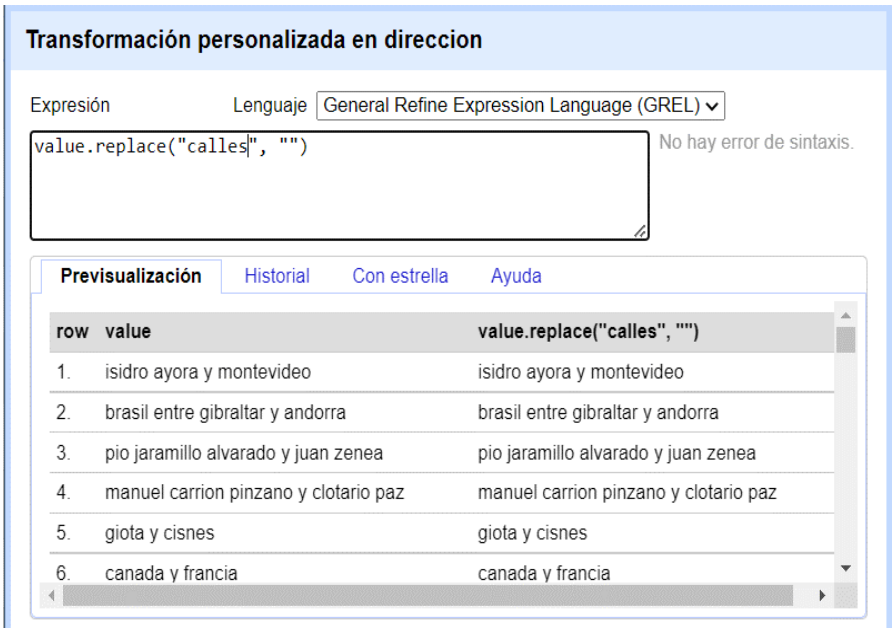


Figura 20. Eliminación de datos redundantes de direcciones viales a través del comando vaule.replace()

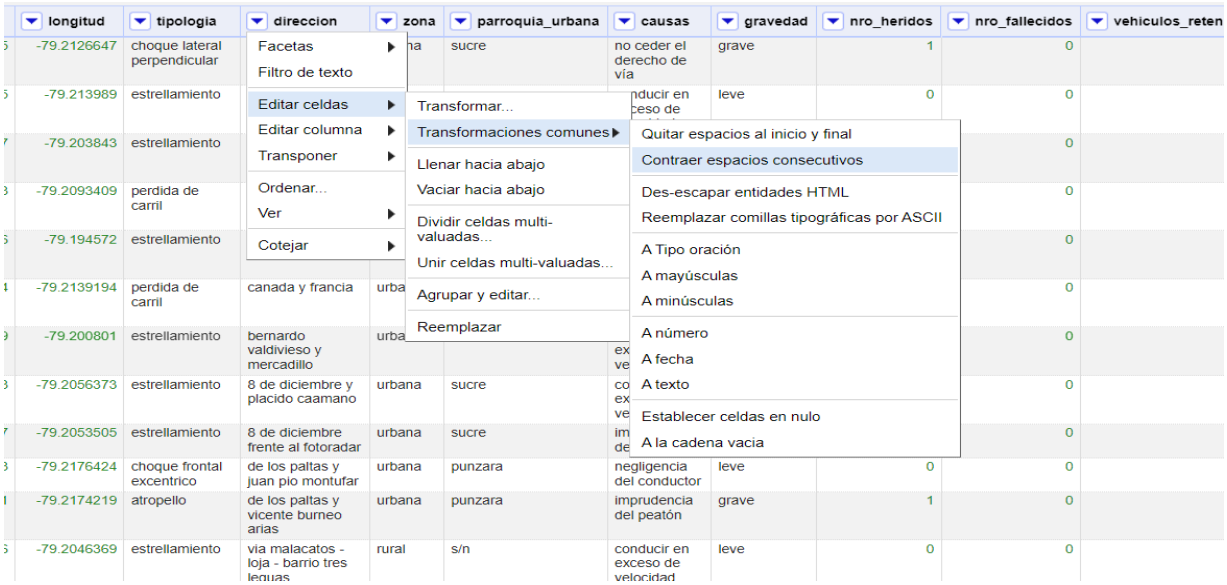
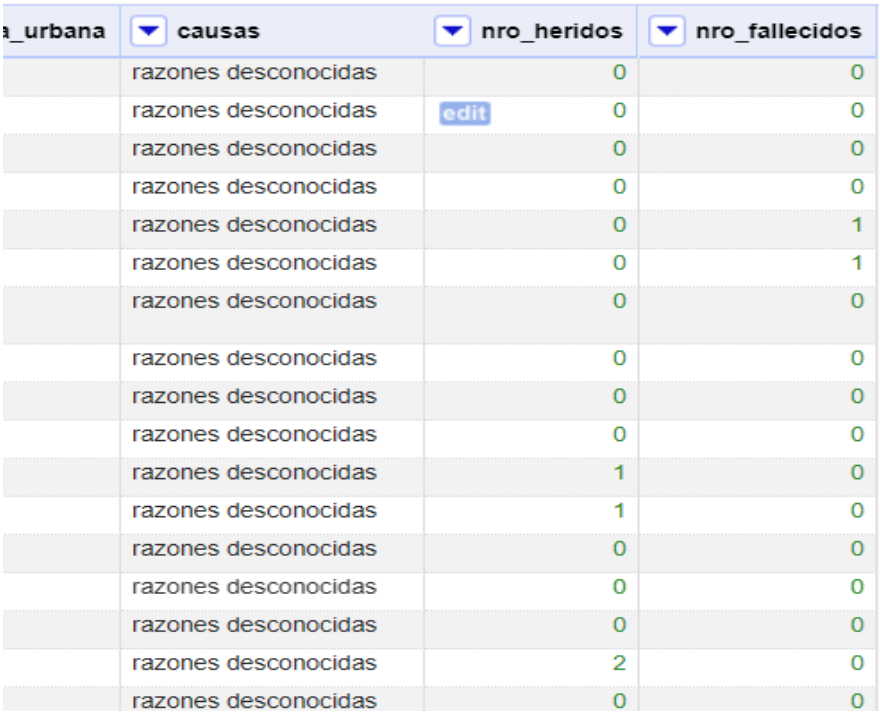


Figura 21. Eliminación de espacios consecutivos en los registros de datos

Mientras para las variables numéricas que constaban como datos vacíos o nulos se estableció ingresar el valor de 0 en formato de texto y estos datos se los transformó a número, tal como se presenta en la Figura 22 y Figura 23.



_urbana	causas	nro_heridos	nro_fallecidos
	razones desconocidas	0	0
	razones desconocidas	edit 0	0
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	0	1
	razones desconocidas	0	1
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	1	0
	razones desconocidas	1	0
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	0	0
	razones desconocidas	2	0
	razones desconocidas	0	0

Figura 22. Ingreso de valor "0" en registros vacíos de variables numéricas

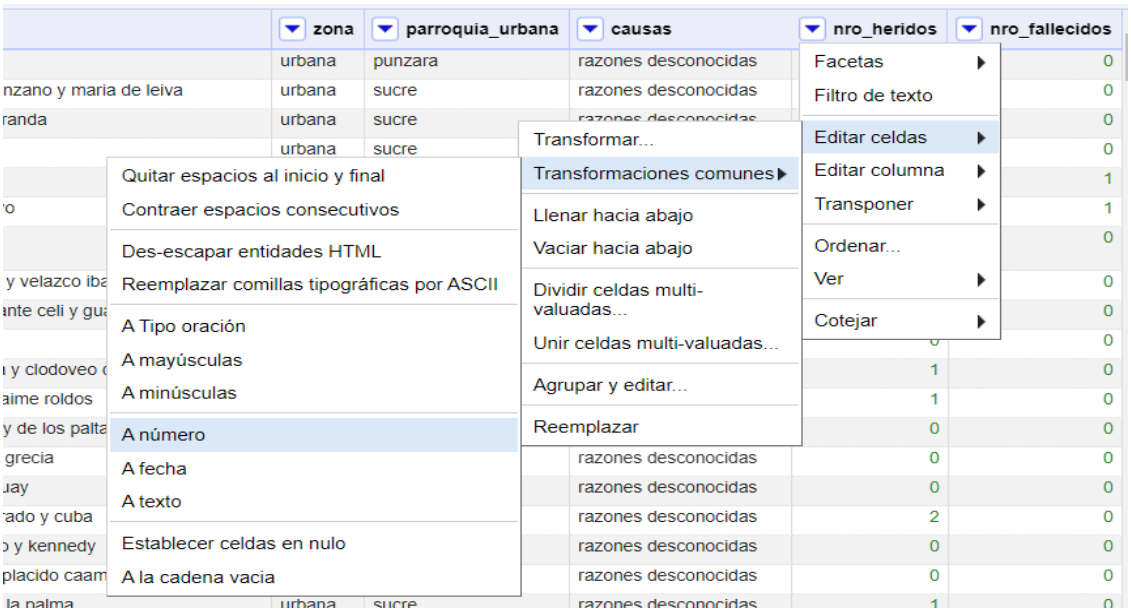


Figura 23. Conversión de registros de texto a registros numéricos

Con respecto a la zona, se analizó los datos faltantes por lo que se consideró la variable "parroquia_urbana", para establecer a que región zonal pertenecía el dato, por lo que se realizó de forma manual, tal como se presenta en la Figura 24, como por ejemplo variable "parroquia_urbana" con dato "san sebastian" se completó la variable "zona" con el dato "urbana".

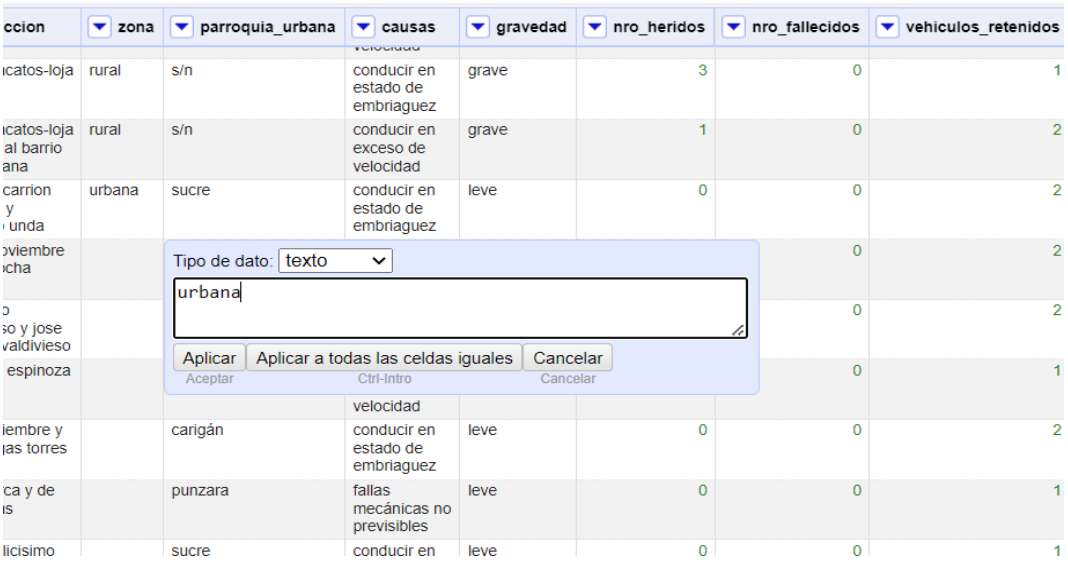


Figura 24. Corrección e ingreso de datos en celdas vacías para la variable "zona"

Para la transformación de los datos se consideró una nueva columna para generar el registro de datos de días, con el fin de optimizar la variable “fecha”, donde se colocó los datos acordes a los días señalados en el calendario, lo que generó la nueva variable “dia”, tal como se presenta en la Figura 25.

Figura 25. Creación de variable "dia"

Posterior, con respecto a los registros de datos de la variable “hora” y con fines representativos y visuales, se estableció generar rangos de horarios tal como se indican en la Tabla 16. Rango de datos de variable "hora", para luego realizar la transformación de los datos a través del comando value.replace(“valor original”, “valor reemplazado”), como por ejemplo: value.replace(“00:00”, “h00”), tal como se observa en la Figura 26.

Tabla 16. Rango de datos de variable "hora"

00:00 - 00:59	**01:00 - 01:59**	**02:00 - 02:59**	**03:00 - 03:59**	**04:00 - 04:59**	**05:00 - 05:59**	**06:00 - 06:59**	**07:00 - 07:59**	**08:00 - 08:59**	**09:00 - 09:59**	**10:00 - 10:59**	**11:00 - 11:59**
h00	h01	h02	h03	h04	h05	h06	h07	h08	h09	H10	H11
12:00 - 12:59	**13:00 - 13:59**	**14:00 - 14:59**	**15:00 - 15:59**	**16:00 - 16:59**	**17:00 - 17:59**	**18:00 - 18:59**	**19:00 - 19:59**	**20:00 - 20:59**	**21:00 - 21:59**	**22:00 - 22:59**	**23:00 - 23:59**
H12	h13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23

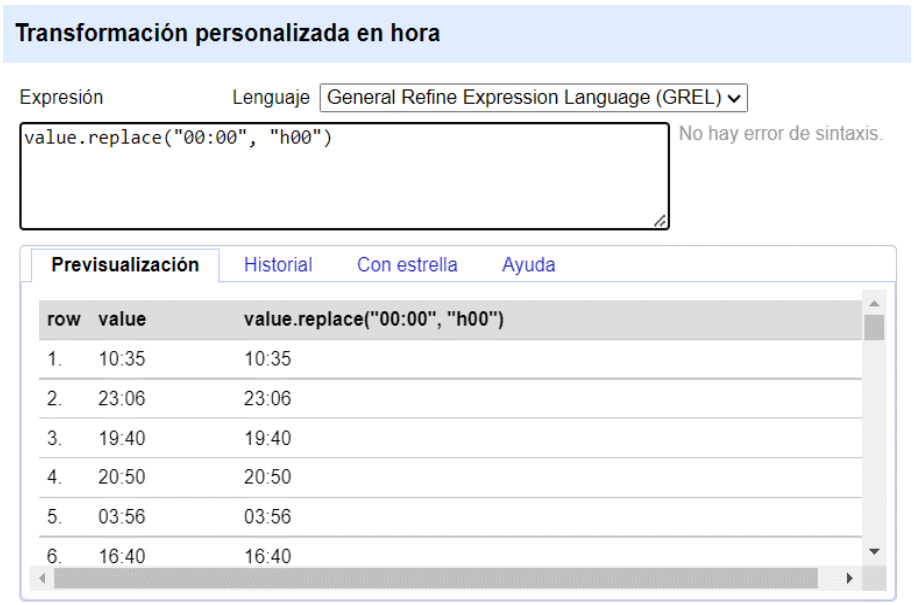


Figura 26. Transformación de rangos de variable "hora"

A su vez, con respecto a los registros de datos pertenecientes a la variable “zona”, se empleó el respectivo filtrado de datos, considerando únicamente los datos que se encuentran dentro de la zona urbana, excluyendo los datos que forman parte de la zona rural, tal como se presenta en la Figura 27.

Figura 27. Eliminación de datos de variable "zona" (rural)

Al culminar la limpieza de las bases de datos, se exportó los nuevos registros de datos en formato CSV, que fueron nombrados como “AT2018_NBD.csv”, “AT2019_NBD.csv”, “AT2020_NBD.csv” y “AT2021_NBD.csv”, debido a que estos nuevos registros se encuentran estandarizados, los cuales se utilizó para reemplazar los anteriores registros seleccionados, mismos que pueden ser visualizados en el **¡Error! No se encuentra el origen de la referencia.**.

Casos

Para la minería de datos se implementó tres casos de estudio para optimizar la aplicación del modelo de árboles de decisión y generar una exactitud considerable de predicción del modelo.

- **Caso 1**

Para el primer caso, se utilizó las variables “dia”, “hora” y “tipologia” con los datos actuales, tal como se muestra en la Tabla 17.

Tabla 17. Tabla de datos actuales caso 1.

Caso 1					
dia	**dia transformado**	**hora**	**hora transformada**	**tipologia**	**tipologia transformada**
lunes	1	h00	0	arrollamiento	1
martes	2	h01	1	atipico	2
miercoles	3	h02	2	atropello	3
jueves	4	h03	3	caida de pasajero	4

Caso 1					
dia	**dia transformado**	**hora**	**hora transformada**	**tipologia**	**tipologia transformada**
viernes	5	h04	4	choque frontal	5
sabado	6	h05	5	choque frontal excentrico	6
domingo	7	h06	6	choque frontal longitudinal	7
-	-	h07	7	choque lateral angular	8
-	-	h08	8	choque lateral perpendicular	9
-	-	h09	9	choque por alcance	10
-	-	h10	10	colision	11
-	-	h11	11	encunetamiento	12
-	-	h12	12	estrellamiento	13
-	-	h13	13	perdida de carril	14
-	-	h14	14	perdida de pista	15
-	-	h15	15	roce negativo	16
-	-	h16	16	roce positivo	17
-	-	h17	17	rozamiento	18
-	-	h18	18	volcamiento	19
-	-	h19	19	volcamiento lateral	20
-	-	h20	20	volcamiento longitudinal	21
-	-	h21	21	-	-
-	-	h22	22	-	-
-	-	h23	23	-	-

- **Caso 2**

En el segundo caso, se generó una lista de rangos, en donde la variable “dia” fue estructurada con tres rangos de datos, mientras la variable “hora” fue conformada por cuatro rangos de datos y para la variable “tipologia” se consideró los datos más relacionados para presentar un dato general, como se muestra en la Tabla 18.

Tabla 18. Tabla de rangos de datos (dia, hora, tipologia) caso 2.

Caso 2					
dia	**dia transformado**	**hora**	**hora transformada**	**tipologia**	**tipologia transformada**
lunes	1	h00	0	arrollamiento	1
martes	1	h01	0	atipico	2
miercoles	1	h02	0	atropello	3
jueves	2	h03	0	caida de pasajero	4
viernes	2	h04	1	choque frontal	5
sabado	3	h05	1	choque frontal excentrico	5
domingo	3	h06	1	choque frontal longitudinal	5
-	-	h07	1	choque lateral angular	5
-	-	h08	2	choque lateral perpendicular	5
-	-	h09	2	choque por alcance	5
-	-	h10	2	colision	6
-	-	h11	2	encunetamiento	7
-	-	h12	3	estrellamiento	8
-	-	h13	3	perdida de carril	9
-	-	h14	3	perdida de pista	9
-	-	h15	3	roce negativo	10
-	-	h16	4	roce positivo	10
-	-	h17	4	rozamiento	10
-	-	h18	4	volcamiento	11
-	-	h19	4	volcamiento lateral	11
-	-	h20	5	volcamiento longitudinal	11
-	-	h21	5	-	-
-	-	h22	5	-	-
-	-	h23	5	-	-

- **Caso 3**

Para el tercer caso, se estableció para la variable “hora” un conjunto de rangos de datos que contengan seis registros consecutivos para disminuir la cantidad de registros y establecer una predicción más exacta, en donde también se consideró los datos de las variables “dia” y “tipologia” del caso 2, presentados en la Tabla 19.

Tabla 19. Tabla de rangos de datos (dia, hora, tipologia) caso 3.

Caso 3					
dia	**dia transformado**	**hora**	**hora transformada**	**tipologia**	**tipologia transformada**
lunes	1	h00	0	arrollamiento	1
martes	1	h01	0	atipico	2
miercoles	1	h02	0	atropello	3
jueves	2	h03	0	caida de pasajero	4
viernes	2	h04	0	choque frontal	5
sabado	3	h05	0	choque frontal excentrico	5
domingo	3	h06	1	choque frontal longitudinal	5
-	-	h07	1	choque lateral angular	5
-	-	h08	1	choque lateral perpendicular	5
-	-	h09	1	choque por alcance	5
-	-	h10	1	colision	6
-	-	h11	1	encunetamiento	7
-	-	h12	2	estrellamiento	8
-	-	h13	2	perdida de carril	9
-	-	h14	2	perdida de pista	9
-	-	h15	2	roce negativo	10
-	-	h16	2	roce positivo	10
-	-	h17	2	rozamiento	10
-	-	h18	3	volcamiento	11
-	-	h19	3	volcamiento lateral	11

Caso 3					
dia	**dia transformado**	**hora**	**hora transformada**	**tipologia**	**tipologia transformada**
-	-	h20	3	volcamiento longitudinal	11
-	-	h21	3	-	-
-	-	h22	3	-	-
-	-	h23	3	-	-

6.2. Objetivo 2: Implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos.

En este objetivo se ejecutó la tercera etapa de la metodología KDD, que se detalla a continuación:

6.2.1. Etapa 3: Aplicación de la Minería de datos

Tarea 1: Identificar las variables que influyen en el cometimiento del accidente de tránsito.

De acuerdo a lo realizado en sección 6.1.2, de transformación y limpieza de la base de datos, como antecedente se consideró las variables de la base de datos de accidentes de tránsito suscitados en el cantón Loja, en el año 2018, como se muestran en la Tabla 20. Comparación de variables, para el entrenamiento del modelo de árboles de decisión, y la comparativa entre los datos de accidentes de tránsito del periodo 2019 – 2020.

Tabla 20. Comparación de variables

Nro.	Variables 2018	Variables 2019	Variables 2020
1	fecha	fecha	fecha
2	dia	dia	dia
3	hora	hora	hora
4	tipologia	latitud	latitud
5	direccion	longitud	longitud
6	zona	tipologia	tipologia
7	parroquia_urbana	direccion	direccion
8	causas	zona	zona
9	nro_heridos	parroquia_urbana	parroquia_urbana

Nro.	Variables 2018	Variables 2019	Variables 2020
10	nro_fallecidos	causas	causas
11	-	nro_heridos	gravedad
12	-	nro_fallecidos	nro_heridos
13	-	-	nro_fallecidos
14	-	-	vehiculos_retenidos
15	-	-	senalizacion_existente

En la Tabla 21. Variables seleccionadas para el estudio se presentan las variables similares de accidentabilidad vehicular, se considera siete variables del conjunto de datos que contienen información relevante para predecir un suceso de tránsito.

Tabla 21. Variables seleccionadas para el estudio

Nro.	Variables 2018	Variables 2019	Variables 2020
1	dia	dia	dia
2	hora	hora	hora
3	tipologia	tipologia	tipologia
4	parroquia_urbana	parroquia_urbana	parroquia_urbana
5	causas	causas	causas
6	nro_heridos	nro_heridos	nro_heridos
7	nro_fallecidos	nro_fallecidos	nro_fallecidos

Tarea 2: 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.

En esta tarea se presenta la carga del conjunto de datos, así como la creación y entrenamiento de los modelos de predicción mediante la aplicación de herramientas software de minería de datos, es importante resaltar que las variables se encuentran descritas con la misma semantica aplicada en los datos, presentes en el anexo 17.

Aplicación de herramientas software para minería de datos

Se desarrolla en el presente TT el estudio de los datos mediante las plataformas de software Python y Weka, considerando lo realizado en la sección 4.8 en donde refleja las herramientas software más utilizadas para la aplicación de minería de datos.

Aplicación de Python mediante entorno de Google Colab

Para el ejecución de esta herramienta se utilizó los conjuntos de datos “AT2018_NBD”, “AT2019_NBD” y “AT2020_NBD”, archivos que fueron cargados en el entorno, tal como se presenta en la Figura 28.

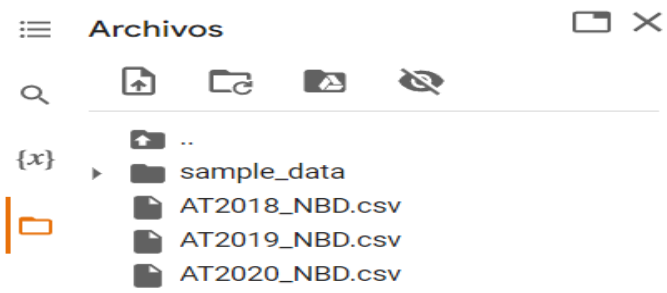


Figura 28. Carga de conjuntos de datos en entorno de Google Colab

Para el desarrollo de la minería de datos se utilizó bibliotecas y librerías de Python para la calidad de visualización y presentación estadística de resultados , tal como se presentan en la Tabla 2. Bibliotecas de maching learning para data mining con python, a su vez se importó los registros, presentando la información de los datos en el entorno, tal como se muestra en la Figura 29 y Figura 30.

```
!pip install plotly
import numpy as np #Operaciones matemáticas rápidas sobre matrices
import pandas as pd #biblioteca de análisis y manipulación de datos para Python
import plotly.express as px
import matplotlib.pyplot as plt #Proporciona una forma de trazado similar a MATLAB. py
import seaborn as sns #permite generar fácilmente elegantes gráficos, proporciona una
import statsmodels.api as sm

# Preprocesado y modelado
# ------------------------------------------------------------------------------
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.tree import export_graphviz
from sklearn.tree import export_text
from sklearn.model_selection import GridSearchCV
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix

# Configuración warnings
# ------------------------------------------------------------------------------
import warnings
warnings.filterwarnings('once')

df= pd.read_csv('AT2018_NBD.csv')
df1= pd.read_csv('AT2019_NBD.csv')
df2= pd.read_csv('AT2020_NBD.csv')
df=df.append(df1)
df=df.append(df2)
df.head()
```

Figura 29. Revisión de código de librerías y carga de Dataset

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: plotly in /usr/local/lib/python3.7/dist-packages (5.5.0)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from plotly) (1.15.0)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.7/dist-packages (from plotly) (8.0.1)
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions in the
  import pandas.util.testing as tm
```

	fecha	dia	hora	tipologia	direccion	zona	parroquia_urbana	causas	nro_heridos	nro_fallecidos	latitud	longitud	gravedad	ve
0	09-26	miercoles	h13	atropello	cuero y caicedo entre manuel angulo y carlos m...	urbana	punzara	imprudencia del peaton	1	0	NaN	NaN	NaN	
1	09-27	jueves	h20	atropello	mayas y juan jose samaniego	urbana	sucre	negligencia del conductor	1	0	NaN	NaN	NaN	
2	09-28	viernes	h18	choque lateral perpendicular	jose martinez ruiz y sixto duran romero	urbana	san sebastian	condiciones climaticas desfavorables	1	0	NaN	NaN	NaN	
3	09-28	viernes	h21	estrellamiento	pablo palacios y cesar p. ludena	urbana	carigan	fallas mecanicas no previsibles	0	0	NaN	NaN	NaN	

Figura 30. Presentación general de datos cargados en el entorno de Google Colab

Posterior se verificó dentro del conjunto de datos las variables no consideradas como objeto de estudio, por lo que através de la función de la biblioteca pandas, con la implementación del método “DataFrameloc.[]” , con el cual se localizó los datos de las filas y columna específicas, realizando la limpieza de información innecesaria del conjunto de datos, como por ejemplo “df=df.loc[:,df.columns!="fecha"]”, tal como se presenta en la Figura 31.

```
[4]  df=df.loc[:,df.columns!="fecha"]
     df=df.loc[:,df.columns!="zona"]
     df=df.loc[:,df.columns!="latitud"]
     df=df.loc[:,df.columns!="longitud"]
     df=df.loc[:,df.columns!="direccion"]
     df=df.loc[:,df.columns!="gravedad"]
     df=df.loc[:,df.columns!="vehiculos_retenidos"]
     df=df.loc[:,df.columns!="senalizacion_existente"]
     df=df.loc[:,df.columns!="condicion_calzada"]
     df=df.loc[:,df.columns!="condicion_atmosferica"]
```

Figura 31. Variables excluidas para el proceso de minería de datos

En la Figura 32 se presentó un resumen de los registros de datos establecidos utilizando el comando “.sample(10)”, información que fue empleada para la predicción de datos.

```
[6] df.sample(10)
```

	dia	hora	tipologia	parroquia_urbana	causas	nro_heridos	nro_fallecidos
109	miercoles	h15	estrellamiento	sucre	conducir en exceso de velocidad	0	0
392	sabado	h10	choque por alcance	san sebastian	no mantener la distancia reglamentaria	1	0
135	lunes	h05	choque lateral perpendicular	el sagrario	conducir en estado de embriaguez	2	0
140	miercoles	h18	atipico	san sebastian	imprudencia del conductor	1	0
373	jueves	h22	perdida de carril	sucre	conducir en exceso de velocidad	0	0
187	miercoles	h08	roce negativo	punzara	imprudencia del conductor	0	0
343	viernes	h18	choque lateral angular	punzara	conducir en estado de embriaguez	2	0
425	martes	h21	atropello	carigan	imprudencia del peaton	1	0
497	jueves	h11	choque lateral angular	carigan	no respetar las senales de transito	0	0
45	domingo	h06	choque lateral angular	sucre	imprudencia del conductor	0	0

Figura 32. Resumen de datos incluidos para proceso de minería de datos

Se evidenció que se utilizaron una cantidad de 1028 registros de datos como objeto de estudio, en donde cada uno de los registros cuenta con 7 variables, tal como se observa en la Figura 33.

```
ds=pd.DataFrame(df)
#Presenta el numero de filas
print("El numero de registros(observaciones) es: ",ds.shape[0])

#Presenta el numero de columnas
print("El numero de columnas(variables) es: ",len(ds.columns))
```

```
El numero de registros(observaciones) es:  1028
El numero de columnas(variables) es:  7
```

Figura 33. Verificación de registros como objeto de estudio

En la Figura 34, se analizó los datos que contiene cada una de las variables, por lo que se aplicó el preprocesamiento de datos, en donde a cada dato se asignó un identificador, acorde a lo establecido en el **¡Error! No se encuentra el origen de la referencia.**, por ejemplo: los datos de la variable "causas" fueron transformados a tipo numérico, tal como se evidencia en la Figura 35 y Figura 36; mientras tanto, los datos restantes se encuentran adjuntos en el **¡Error! No se encuentra el origen de la referencia.**.

```
[181] print(df['causas'].unique())# datos en texto
```

```
['imprudencia del peaton' 'negligencia del conductor'
 'condiciones climaticas desfavorables' 'fallas mecanicas no previsibles'
 'imprudencia del conductor' 'conducir en estado de embriaguez'
 'no ceder el derecho de via' 'cambio brusco e indebido de carril'
 'no mantener la distancia reglamentaria' 'desatento a la conduccion'
 'no respetar las senales de transito' 'imprudencia del pasajero'
 'cruce de via sin preferencia' 'perdida de pista'
 'conducir en exceso de velocidad' 'invadir carril de circulacion'
 'razones desconocidas' 'no respetar las ordenes del agente de transito'
 'impericia del conductor' 'inobservancia de leyes de transito']
```

Figura 34. Visualización de datos contenidos en variable

```
[182] df['causas']= df['causas'].apply(lambda x:
                                       1 if x == 'conducir en estado de embriaguez' else
                                       2 if x == 'imprudencia del conductor' else
                                       3 if x == 'no ceder el derecho de via' else
                                       4 if x == 'conducir en exceso de velocidad' else
                                       5 if x == 'cambio brusco e indebido de carril' else
                                       6 if x == 'fallas mecanicas no previsibles' else
                                       7 if x == 'no respetar las senales de transito' else
                                       8 if x == 'invadir carril de circulacion' else
                                       9 if x == 'imprudencia del peaton' else
                                       10 if x == 'no mantener la distancia reglamentaria' else
                                       11 if x == 'razones desconocidas' else
                                       12 if x == 'condiciones climaticas desfavorables' else
                                       13 if x == 'negligencia del conductor' else
                                       14 if x == 'no respetar las ordenes del agente de transito' else
                                       15 if x == 'impericia del conductor' else
                                       16)
```

Figura 35. Transformación de datos de texto a datos numéricos

```
[269] print(df['causas'].unique()) #datos a números

[ 9 13 12  6  2  1  3  5 10 16  7  4  8 11 14 15]
```

Figura 36. Verificación de transformación de datos numéricos

En la Figura 37, se presenta cada uno de los registros transformados a datos numéricos, en donde se utiliza cada una de las variables para el entrenamiento del modelo predictivo, por lo que se consideró todas las siete variables presentes en el conjunto de datos.

```
[283] df.head()
```

	dia	hora	tipologia	parroquia_urbana	causas	nro_heridos	nro_fallecidos
0	3	13	3	5	9	1	0
1	4	20	3	4	13	1	0
2	5	18	9	2	12	1	0
3	5	21	13	6	6	0	0
4	6	19	9	1	2	0	0

Figura 37. Presentación de datos numéricos contenidos en el dataset

Para la interpretación de datos se señala las siete variables que contiene el dataset, en donde para cada variable se señala cada uno de los casos utilizados en la predicción de modelo, a su vez, en la Figura 38 mediante la matriz de correlación se describen los datos de accidentes de tránsito identificando las variables que están más relacionadas.

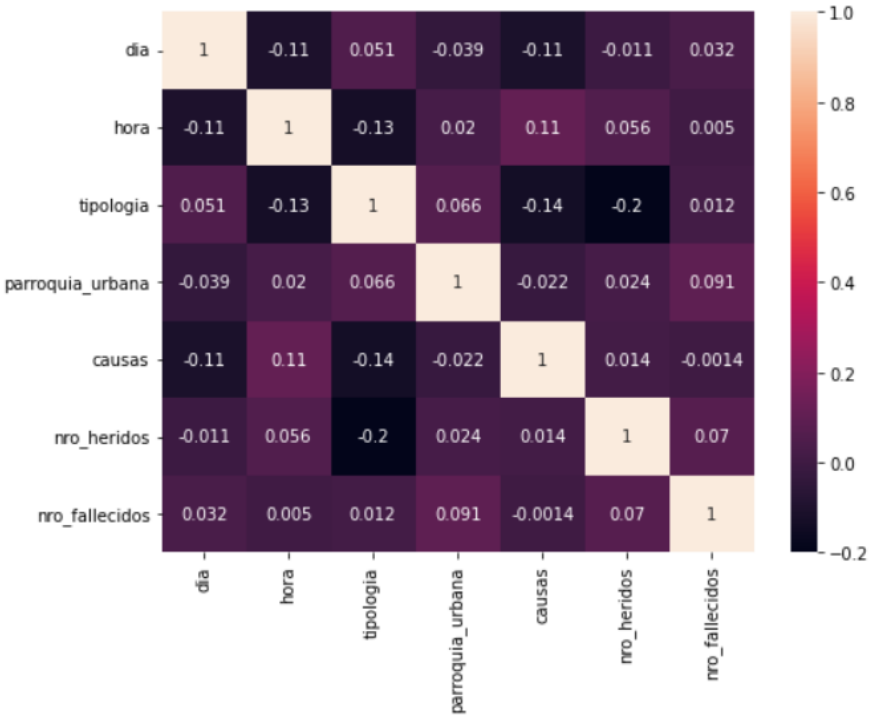


Figura 38. Matriz de correlación de variables de accidentes de tránsito

Aquí podemos ver que existe una correlación negativa de -0.11entre el día de la semana y la hora en que se ocasionaron los accidentes. Lo que sugiere que la frecuencia de los accidentes varía dependiendo del día de la semana y la hora del día, de manera que en el momento en que ocurre un siniestro de tránsito en general, tal como las horas pico, no son los mismos en todos los días de la semana. Referente a la variable hora y tipología se visualiza una correlación débil lo que indicó que los diferentes tipos de accidentes son menos probables en ocurrir durante ciertas horas del día; mientras que, existe la correlación negativa débil entre la variable tipología y causas con el -0.14 lo que nos indica que existe menos posibilidad de que ocurra diferentes tipos de accidentes por las mismas causas; por otro lado, la correlación negativa de -0.2 entre la variable tipología y número de heridos señaló que existen tipologías de accidentes de tránsito que son menos probables de causar un mayor número de heridos.

En la Figura 39 se indica la división de los datos en dos conjuntos: un conjunto de entrenamiento (train) y un conjunto de prueba (test). En este caso, se ha seleccionado el 20% de los datos para el conjunto de prueba, mientras que el 80% restante se usará para el conjunto de entrenamiento. Se proporcionó un valor específico (en este caso, 7) al parámetro random_state, por lo que siempre se obtendrá la misma división de datos cuando se ejecute el código varias veces.

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=7)# stratify=y
    print('Son {} datos para entrenamiento y {} datos para prueba'.format(X_train.shape[0], X_test.shape[0]))

    Son 822 datos para entrenamiento y 206 datos para prueba
```

Figura 39. División de los datos

En la Figura 40 se importó la clase DecisionTreeClassifier desde la biblioteca sklearn.tree para construir el modelo de árbol de decisión; también, se estableció la profundidad máxima del árbol de decisión en 3 mediante el parámetro max_depth=3, esto limita la cantidad de niveles que tendrá el árbol, lo que ayudó a evitar un sobreajuste del modelo. Además, se especificó que el criterio para evaluar las divisiones en el árbol fue la entropía mediante el parámetro criterion='entropy' que se utilizó para medir la impureza del nodo y determinar la mejor manera de dividir los datos.

```
#Cargamos la libreria DecisionTreeClasifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, classification_report


#Llamamos al constructor del arbol de decision
classifier = DecisionTreeClassifier(max_depth=3,criterion = 'entropy')

#Entrenamos el modelo
arbol_modelo = classifier.fit(X_train,y_train)

y_pred = classifier.predict(X_test)
```

Figura 40. Configuración del modelado predictivo

Pruebas

Aplicación de caso 1, caso 2 y caso 3.

Para el proceso de pruebas del algoritmo de predicción CART aplicado al modelo de árboles de decisión para la minería de datos se contó con 1028 registros, a su vez se crearon siete subconjuntos de datos referentes a las variables establecidas en la Tabla 21. Variables seleccionadas para el estudio, a fin de desarrollar las pruebas y obtener la generación de datos más óptimos que se detallan en la Tabla 22, en donde se identificó 22 pruebas para las 7 variables del Dataset, cada valor se encuentra referenciado con el número de prueba realizada por variable con la etiqueta “P”; así mismo, se establecieron los resultados generales del porcentaje de exactitud durante la ejecución del modelo predictivo.

Tabla 22. Resultados generales de exactitud del modelo mediante Google Colab

Número de Variable	Variable	Caso 1	Caso 2		Caso 3	
1	dia	P1: 19,90%	P2: 38,34%	-	P3: 36,40%	-
2	hora	P1: 3,88%	P2: 21,84%	P3: 21,35%	P4: 37,37%	P5: 36,89%
3	tipologia	P1: 37,37%	P2: 56,31%	P3: 56,31%	P4: 56,31%	-
4	parroquia_urbana	P1: 31,06%	P2: 27,66%	-	P3: 27,66%	-
5	causas	P1: 38,83%	P2: 35,43%	-	P3: 37,86%	-
6	nro_heridos	P1: 63,10%	P2: 63,59%	-	P3: 37,86%	-
7	nro_fallecidos	P1: 96,11%	-	-	-	-

Los códigos generados y pruebas detalladas de cada una de las variables con sus respectivos casos que fueron aplicados para la minería de datos en el entorno google colab se encuentra adjuntos dentro del **¡Error! No se encuentra el origen de la referencia.**.

Aplicación de herramienta WEKA para minería de datos

Para la ejecución y desarrollo de esta herramienta se utilizó el conjunto de datos "AT2018_2019_2020_NBD" que se encuentra en el repositorio de GitHub al cual se puede acceder desde el siguiente enlace: rangos, donde se realizó la selección de explorador, carga de registros de datos y preprocesamiento del Dataset, tal como se presenta en la Figura 41, Figura 42, Figura 43 y Figura 44.

Figura 41. Selección de explorador "Explorer" de WEKA

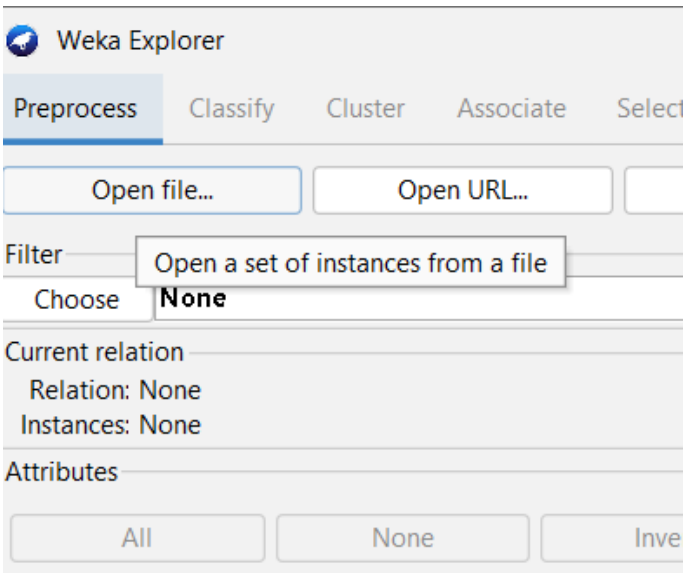


Figura 42. Selección de preprocesamiento para carga de dataset

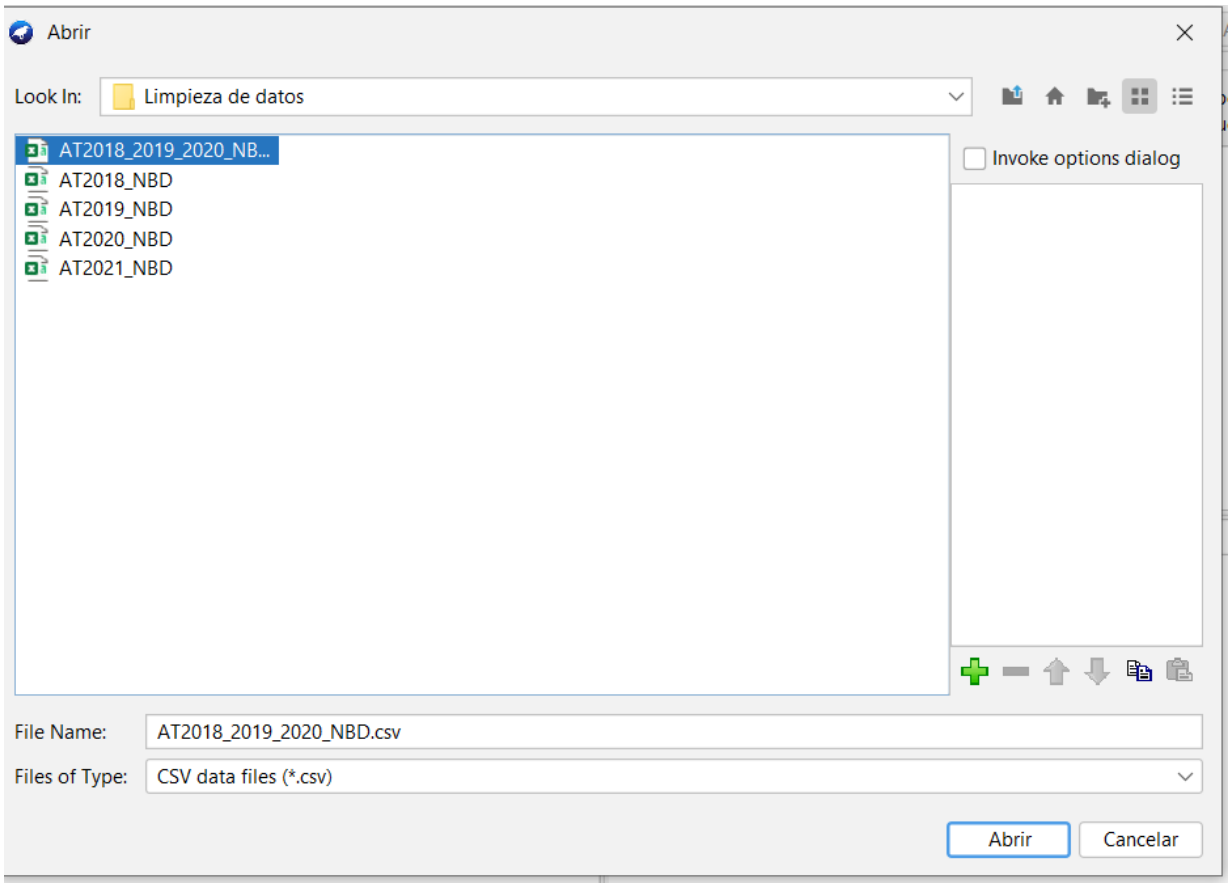


Figura 43. Selección de dataset específico para minería de datos

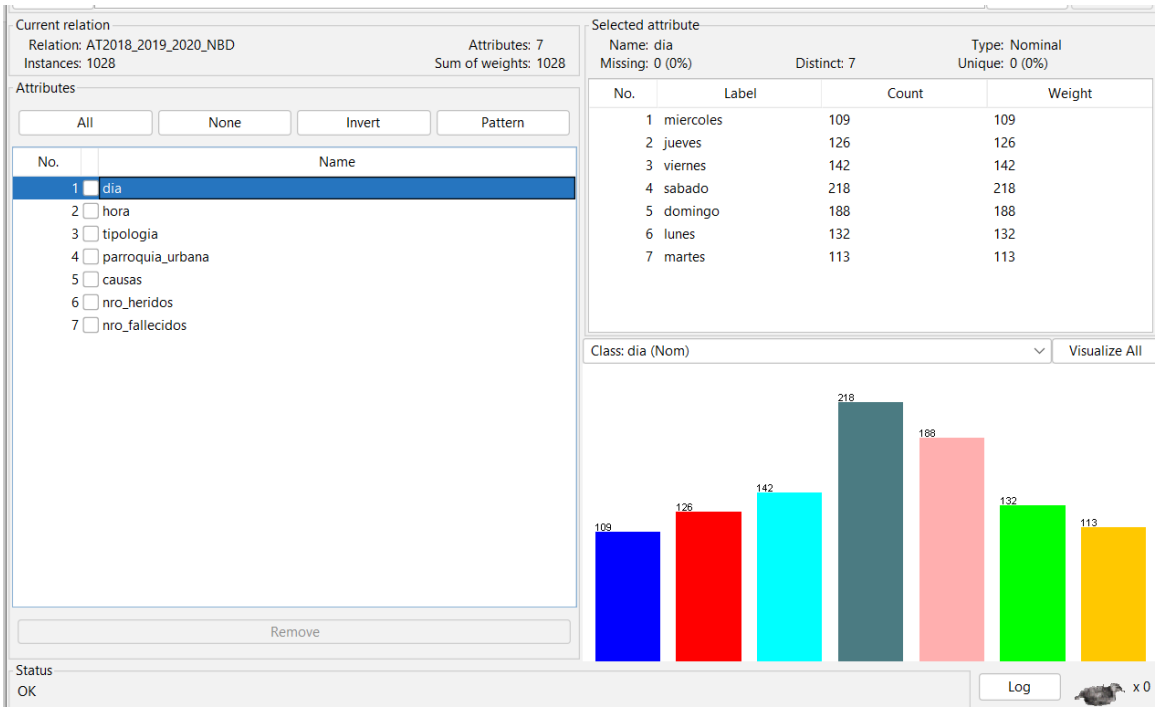


Figura 44. Verificación de datos de muestra para minería de datos

Algoritmo J48

En la aplicación de la clasificación de la información para la minería de datos, se seleccionó el algoritmo J48 apegada a su relación directo con el modelo de árboles de decisión, en donde se señaló la variable destinada para la predicción, siendo esta la variable “dia” tal como se muestra en la Figura 45.

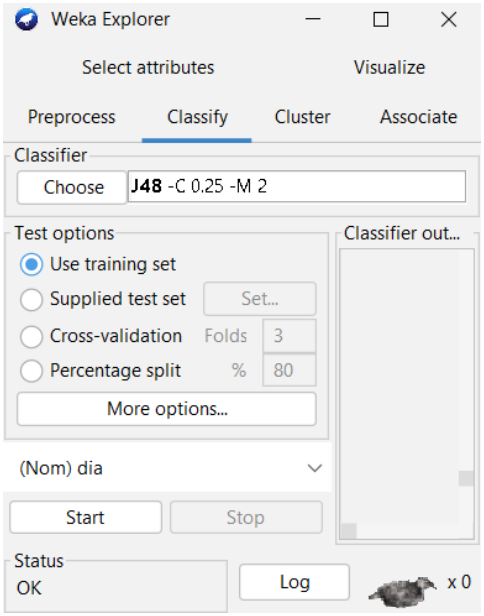


Figura 45. Selección de algoritmo J48 árboles de decisión

Pruebas

Para la aplicación del algoritmo de predicción J48 para la minería de datos se contó con 1028 registros, a su vez se crearon siete subconjuntos de datos referentes a las variables establecidas en la Tabla 23. Variables establecidas para aplicación de algoritmo J48, a fin de desarrollar las pruebas y obtener la generación de datos más óptimos que a continuación se detallan.

Tabla 23. Variables establecidas para aplicación de algoritmo J48

Nro.	Variable
1	dia
2	hora
3	tipología
4	parroquia_urbana
5	causas
6	nro_heridos
7	nro_fallecidos

Cade recalcar que, se consideró la eliminación de variables con el objetivo de mejorar el nivel de porcentajes de predicción de cada una de las variables, implementando hasta un mínimo de seis variables; a su vez, para la minería de datos se implementó tres casos de estudio para optimizar la aplicación del modelo de árboles de decisión y generar una exactitud considerable de predicción del modelo.

Caso 1

Para el primer caso, se utilizó las variables "dia", "hora" y "tipologia" con los datos actuales, tal como se muestra en el **¡Error! No se encuentra el origen de la referencia.**; estos datos fueron utilizados para identificar el porcentaje de las instancias clasificadas correctamente del algoritmo, tal como se presenta en la Tabla 24.

Tabla 24. Resultados de instancias clasificadas y métricas de algoritmo J48 a variables caso 1

Número de Prueba	Variable	Instancias clasificadas correctamente	Precision	Recall
1	dia	14,56%	0,12	0,14
2	hora	3,39%	-	0,03
3	tipologia	43,68%	-	0,43
4	parroquia_urbana	31,06%	0,31	0,31
5	causas	41,26%	-	0,41
6	nro_heridos	66,01%	-	0,66
7	nro_fallecidos	97,08 %	-	0,97

Caso 2

En el segundo caso, se generó una lista de rangos, en donde la variable "dia" fue estructurada con tres rangos de datos, mientras la variable "hora" fue conformada por cuatro rangos de datos y para la variable "tipologia" se consideró los datos más relacionados para presentar un dato general, datos necesarios para identificar el porcentaje de las instancias clasificadas correctamente del algoritmo tal como se muestra en la Tabla 25, para mayor verificación de datos ver el **¡Error! No se encuentra el origen de la referencia.**.

Tabla 25. Resultados de instancias clasificadas correctamente y métricas de algoritmo J48 a variables caso 2

Número de Prueba	Variable	Instancias clasificadas correctamente	Precision	Recall
1	dia	42,71 %	0,39	0,42
2	hora	20,87 %	0,19	0,20
3	tipologia	60,67 %	-	0,60
4	parroquia_urbana	21,84 %	-	0,21
5	causas	33,00 %	-	0,33
6	nro_heridos	65,53 %	-	0,65
7	nro_fallecidos	97,08 %	-	0,97

Caso 3

Para el tercer caso, se estableció para la variable "hora" desarrollar rangos de datos que contengan seis datos consecutivos, para disminuir la cantidad de registros y establecer una predicción más exacta, como se observa en la Tabla 26Tabla 26. Resultados de instancias clasificadas correctamente y métricas de algoritmo J48 a variables caso 3., en donde también se consideró los datos de las variables "dia" y "tipologia" del caso 3, presentados en el **¡Error! No se encuentra el origen de la referencia.**.

Tabla 26. Resultados de instancias clasificadas correctamente y métricas de algoritmo J48 a variables caso 3.

Número de Prueba	Variable	Instancias clasificadas correctamente	Precision	Recall
1	dia	41,26 %	0,40	0,41
2	hora	33,98 %	0,34	0,34
3	tipologia	62,13 %	-	0,62
4	parroquia_urbana	23,30 %	0,20	0,23
5	causas	34,95 %	-	0,35
6	nro_heridos	65,53 %	-	0,65
7	nro_fallecidos	97,08 %	-	0,97

Los códigos generados y pruebas detalladas de cada una de las variables con sus respectivos casos que fueron aplicados para la minería de datos en el entorno WEKA se encuentra adjuntos dentro del **¡Error! No se encuentra el origen de la referencia.**.

6.3. Objetivo 3: Evaluación de la técnica de minería de datos propuesta.

En este objetivo se ejecutó la cuarta etapa de la metodología KDD, que se detalla a continuación:

6.3.1. Etapa 4: Interpretación y presentación de resultados

Tarea 1: Analizar los resultados obtenidos del modelo entrenado.

A continuación, se presenta la interpretación de los resultados obtenidos por el modelo de predicción de cada variable, aplicada a cada una de las clases del conjunto de datos, mediante la utilización de las herramientas Python y WEKA.

Interpretación de resultados de Python entorno Google Colab

A continuación, se detalla cada una de las variables con sus respectivos resultados obtenidos por medio de la predicción del modelo.

- **Variable “dia”**

Resultados Prueba 1 - Caso 1

En la Tabla 27. Valores generados por la predicción de variable "dia" prueba 1, se verificó la muestra de las principales métricas de clasificación en el modelo aplicado, en donde se puede ver que la exactitud es de 19,90%, lo que representa que el nivel de porcentaje obtenido es relativamente bajo; además en la Figura 46, se evidenció los datos de las predicciones realizadas por el modelo, identificando a los días con mayor accidentabilidad vehicular pertenece al día sábado con un total del 77,2%, día domingo con el 17%, miércoles con un 3,4% y jueves con el 2,4% de accidentabilidad vehicular, tal como se evidencia en el anexo 18.

Tabla 27. Valores generados por la predicción de variable "dia" prueba 1

Clase	Porcentaje de predicción
sabado	77,2%
domingo	17%
miercoles	3,4%
jueves	2,4%

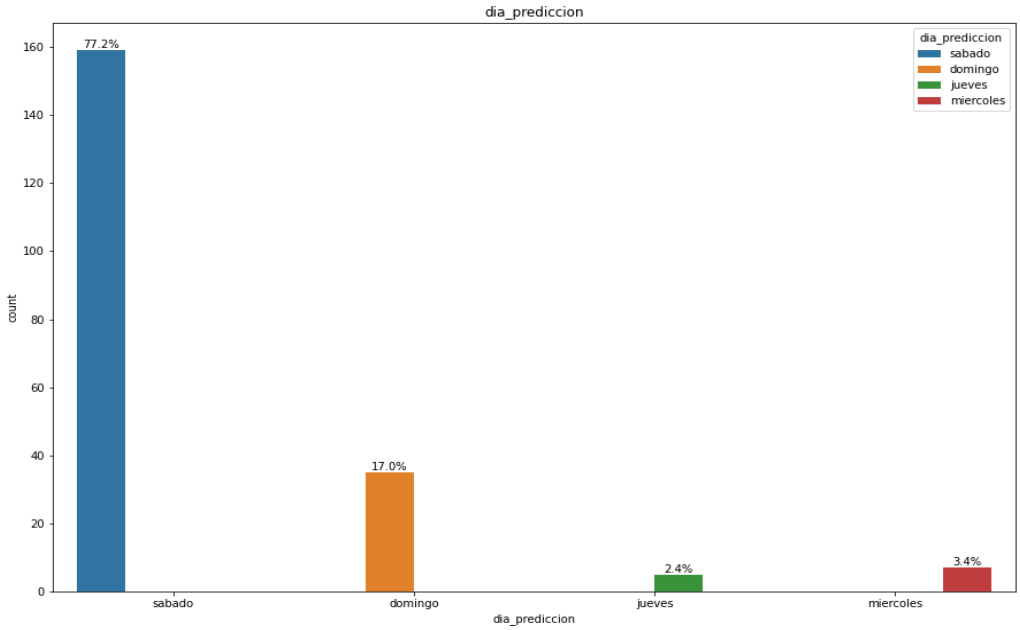


Figura 46. Representación gráfica de datos predichos para la variable "dia" prueba 1

- **Variable “hora”**

Resultados Prueba 1 - Caso 1

En la Tabla 28. Valores generados por la predicción de variable "hora" prueba 1, se verificó la muestra de las principales métricas de clasificación en el modelo aplicado, en donde se puede observar que la exactitud es de 3,88%, lo que representa que el valor obtenido es relativamente bajo; además, en la presente Figura 47, se evidenció los datos de las predicciones realizadas por el modelo, verificando las horas con mayor accidentabilidad vehicular en donde se identificó los horarios con mayor posibilidad de ocurrencia de un siniestro de tránsito indicando los siguientes horarios.

Tabla 28. Valores generados por la predicción de variable "hora" prueba 1

rango	porcentaje de predicción
h12	40,3%
h03	27,2%
h22	13,1%
h07	10,7%
h01	3,9%
h02	3,4%
h04	1,5%

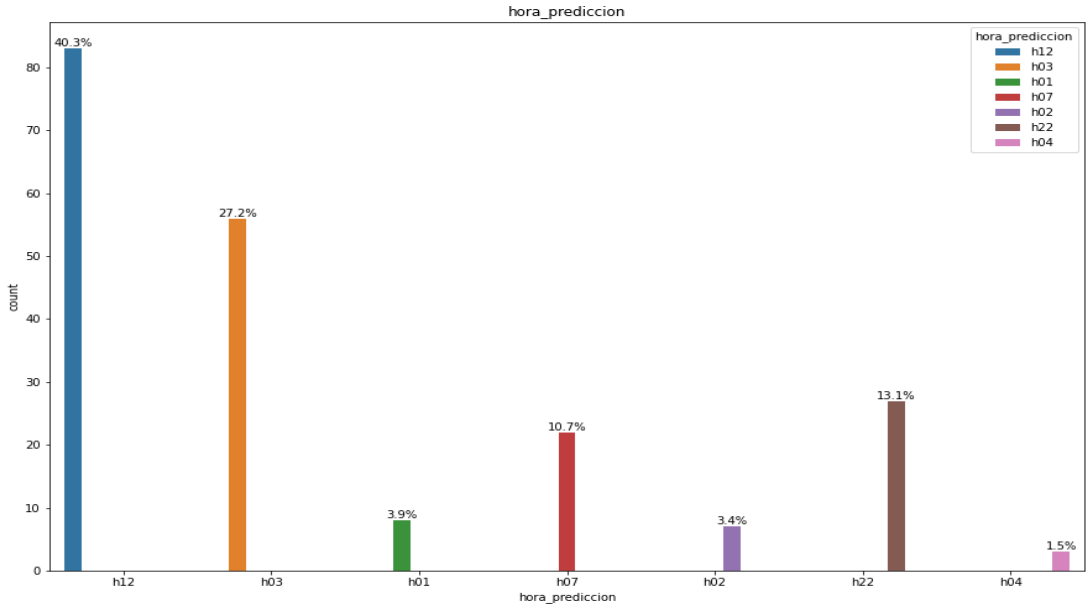


Figura 47. Representación gráfica de los datos predichos para variable “hora” prueba 1

- **Variable “tipologia”**

Resultados Prueba 1 - Caso 1

En la Tabla 29. Valores generados por la predicción de variable "tipologia" prueba 1, se verificó la muestra de las principales métricas de clasificación en el modelo aplicado, en donde se puede observar que la exactitud es de 37,37%, lo que representa que el valor obtenido es relativamente bajo; además, en la Figura 48, se evidenció los datos de las predicciones realizadas por el modelo, verificando las horas con mayor accidentabilidad vehicular en donde se identificó las tipologías viales con mayor posibilidad de ocurrencia de un siniestro de tránsito.

Tabla 29. Valores generados por la predicción de variable "tipologia" prueba 1

Rango	Porcentaje de predicción
choque lateral perpendicular	24,3%
estrellamiento	69,9%
atropello	2,9%
choque por alcance	2,9%

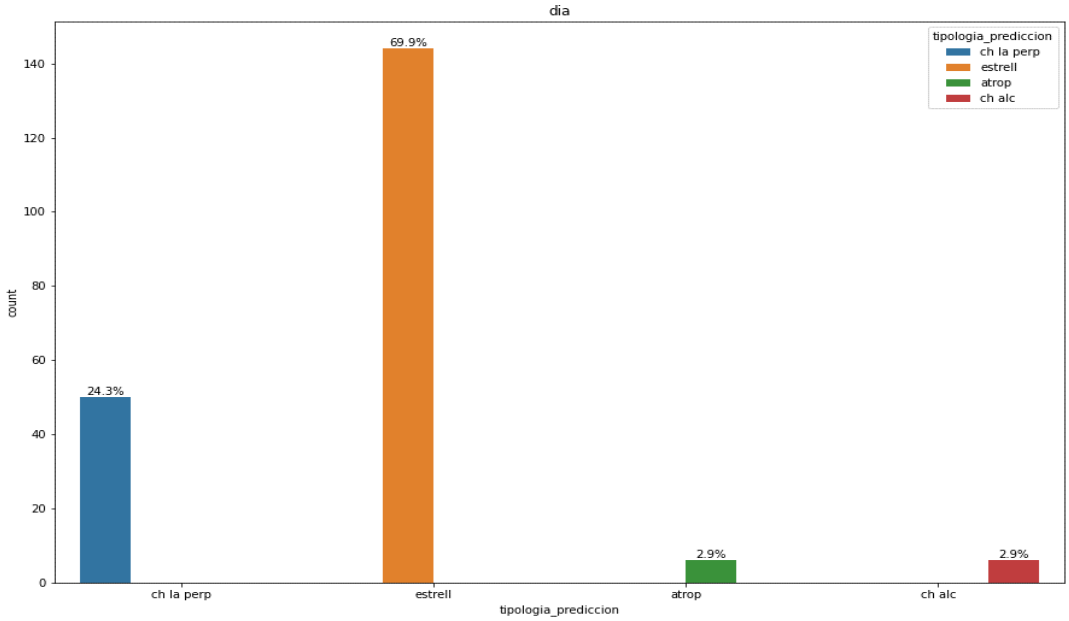


Figura 48. Representación gráfica de los datos predichos para variable "tipologia" prueba 1

Para la revisión detallada de cada una de las pruebas realizadas por los modelos de las variables, revisar el **¡Error! No se encuentra el origen de la referencia.** y anexo 18.

Selección de casos con superior porcentaje de exactitud del modelo.

En la Tabla 30. Selección de mayores porcentajes de exactitud de modelos, con los datos presentados en la Tabla 22 se presentan los resultados de las siete variables de cada uno de los modelos, optando por elegir los resultados con mayor índice de exactitud de los modelos; para esta selección se verificó mejor de los tres casos antes señalados en la sección 6.2.1. Es importante mencionar que los espacios ubicados en la tabla representan las pruebas que no fueron consideradas para la selección.

Tabla 30. Selección de mayores porcentajes de exactitud de modelos

Número de Variable		**Caso 1**	**Caso 2**		**Caso 3**	
1	dia	-	38,34%	-	-	-
2	hora		-	-	37,37%	-
3	tipologia	-	56,31%	-	-	-
4	parroquia_urbana	31,06%	-	-	-	-
5	causas	38,83%	-	-	-	-
6	nro_heridos	-	63,59%	-	-	-
7	nro_fallecidos	96,11%	-	-	-	-

Interpretación de resultados de herramienta WEKA

Para la interpretación de datos en la herramienta WEKA se señala las siete variables que contiene el dataset de accidentes de tránsito del periodo 2018 - 2020, en donde para cada variable se señala cada uno de los casos utilizados en la predicción de modelo, a su vez se describen el porcentaje de exactitud del modelo predictivo, especificando que para cada valor del rango se lo consideró como clase para la predicción.

Resultados de instancias clasificadas correctamente de caso 1, caso2 y caso 3.

En la Tabla 31. Resultados generales de instancias clasificadas correctamente, se identifican veintiuno pruebas para las siete variables del Dataset, en donde se establecieron los resultados de las instancias clasificadas correctamente con su respectivo porcentaje de exactitud durante la ejecución del modelo predictivo; a su vez en la Tabla 32. Selección de mayor exactitud de instancias clasificadas correctamente, se presentan los resultados en donde se seleccionó el mejor de los tres casos ejecutados para ejecución correcta del modelo.

Tabla 31. Resultados generales de instancias clasificadas correctamente

Número de Variables	Variable	Instancias clasificadas correctamente Caso 1	Instancias clasificadas correctamente Caso 2	Instancias clasificadas correctamente Caso 3
1	dia	14,56%	42,71%	41,26%
2	hora	3,39%	20,87%	33,98%
3	tipologia	43,68%	60,67%	62,14%
4	parroquia_urbana	31,06 %	21,84%	23,30%
5	causas	41,26%	33,01%	34,95%
6	nro_heridos	66,01%	65,53%	65,53%
7	nro_fallecidos	97,09 %	97,09%	97,09%

Tabla 32. Selección de mayor exactitud de instancias clasificadas correctamente

Número de Variables	Variable	Instancias clasificadas correctamente Caso 1	Instancias clasificadas correctamente Caso 2	Instancias clasificadas correctamente Caso 3
1	dia	-	42,71 %	
2	hora	-	-	33,98%

Número de Variables	Variable	Instancias clasificadas correctamente Caso 1	Instancias clasificadas correctamente Caso 2	Instancias clasificadas correctamente Caso 3
3	tipologia	-	-	62,14%
4	parroquia_urbana	31,06%	-	-
5	causas	41,26%	-	-
6	nro_heridos	66,01%	-	
7	nro_fallecidos	97,09%	-	-

Comparación entre exactitud de Python Google Colab e instancias clasificadas correctamente de WEKA

En la Tabla 33. Comparación de modelos con mayor promedio aplicados a variables, fueron seleccionados los modelos desarrollados con mayor porcentaje de clasificación correcta de cada una de las variables mediante la aplicación de la herramienta Python en el entorno Google Colab y en la herramienta WEKA.

Tabla 33. Comparación de modelos con mayor promedio aplicados a variables

Número de Variables	Variable	Python Google Colab	Precisión	Recall	WEKA	Precisión	Recall
1	dia	38,34%	0,38	0,38	42,71 %	0,39	0,42
2	hora	37,37%	0,37	0,37	33,98%	0,34	0,34
3	tipologia	56,31%	0,56	0,56	62,14%	-	0,62
4	parroquia_urbana	31,06%	0,31	0,31	31,06%	0,31	0,31
5	causas	38,83%	0,38	0,38	41,26%	-	0,41
6	nro_heridos	63,59%	0,63	0,63	66,01%	-	0,66
7	nro_fallecidos	97,09%	0,97	0,97	97,09%	-	0,97

Luego fueron comparados estos valores, donde se consideró el modelo que contenga mayor porcentaje de efectividad de los datos desarrollados, tal como se presenta en la Tabla 34. Selección de modelos con mayor promedio aplicados a variables, evidenciando estos valores en la Figura 49.

Tabla 34. Selección de modelos con mayor promedio aplicados a variables

Número de Variables	Variable	Promedio de exactitud del modelo seleccionado
1	dia	42,71 %
2	hora	37,37%
3	tipología	62,14%
4	parroquia_urbana	31,06%
5	causas	41,26%
6	nro_heridos	66,01%
7	nro_fallecidos	97,09%

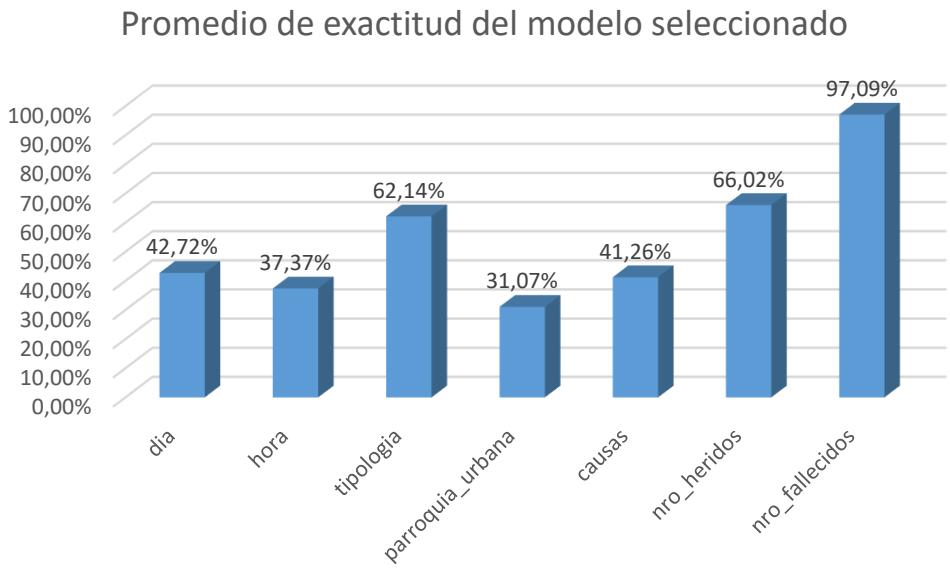


Figura 49. Representación de valores de instancias clasificadas correctamente

Resultado de métricas según la exactitud del modelo

Las métricas seleccionadas para el estudio de cada uno de los modelos pertenecen a los resultados de las exactitudes y las instancias seleccionadas en la Tabla 35, siendo consideradas específicamente las métricas de precisión y métricas recall para cada variable de los modelos realizados, evidenciando los valores de cada métrica en la Figura 50 y Figura 51.

Tabla 35. Revisión de métricas asociadas a la exactitud del modelo

Variable	Instancias clasificadas correctamente	Precisión	Recall
dia	42,72%	0,39	0,42
hora	37,37%	0,37	0,37
tipologia	62,14%	-	0,62
parroquia_urbana	31,06%	0,31	0,31
causas	41,26%	-	0,41
nro_heridos	66,02%	-	0,66
nro_fallecidos	97,09%	-	0,97

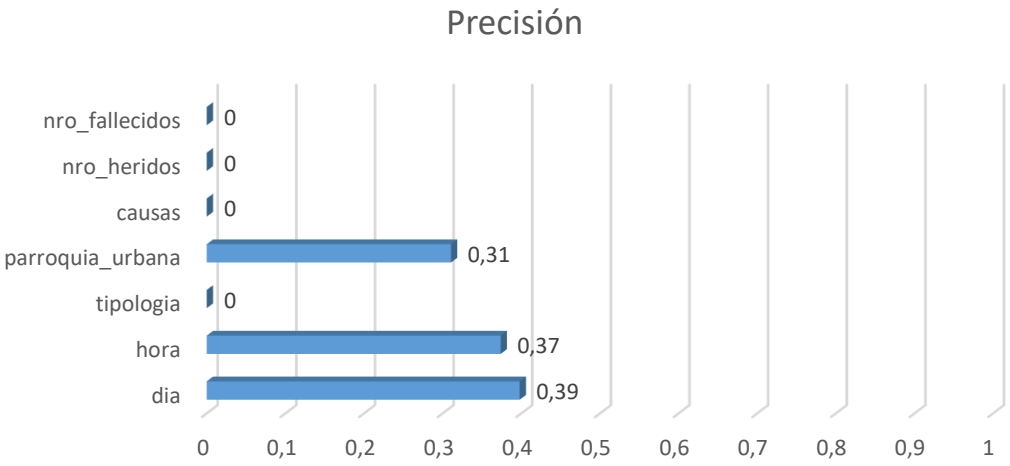


Figura 50. Representación de valores seleccionados de métrica de precisión

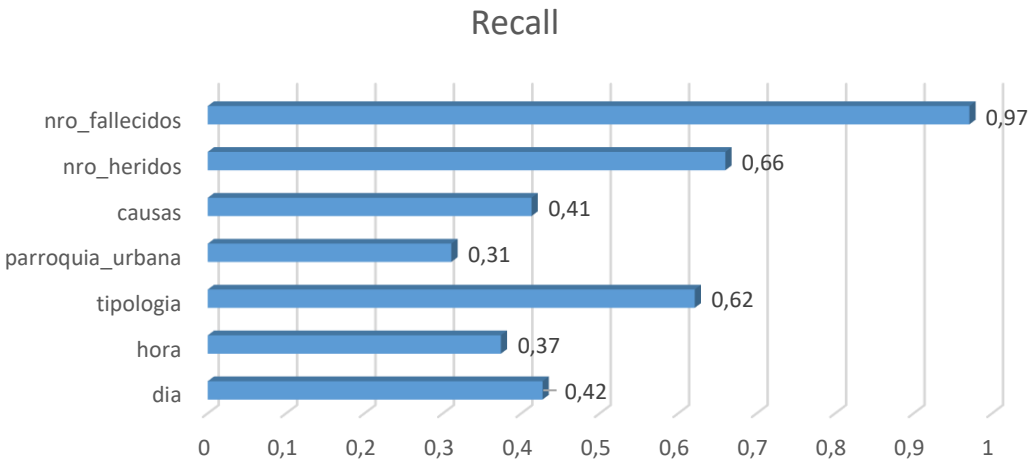


Figura 51. Representación de valores seleccionados de métrica recall

Estos modelos predictivos obtenidos en el proceso de análisis de resultados fueron utilizados y cargados en los entornos para desarrollar los nuevos modelos de predicción de Python[1] y modelos de predicción WEKA[2], los cuales fueron implementados en los registros de datos de los accidentes de tránsito del cantón Loja del año 2021.

Tarea 2: Evidenciar el funcionamiento del modelo con datos actuales.

En esta sección se realizó el estudio de los datos de accidentes de tránsito del Cantón Loja durante el año 2021, para la ejecución de la minería de datos se implementa los modelos seleccionados de cada una de las variables; a su vez, fueron seleccionadas 7 variables considerando su similitud, como se presenta en la Tabla 34. Selección de modelos con mayor promedio aplicados a variables; se cargan los respectivos dataset, mismos que contienen 370 registros de accidentes de tránsito, en donde se encuentran los datos transformados de acuerdo al número de caso y prueba del modelo predictivo obtenido, tal como se refleja en la sección 6.2.1; a continuación, en la Tabla 36. Datos comparativos de exactitud del modelo predictivo con sus respectivas métricas se presenta las predicciones realizadas a cada una de las variables, adjuntas con sus respectivas métricas.

Tabla 36. Datos comparativos de exactitud del modelo predictivo con sus respectivas métricas

Variable	Exactitud del Modelo	Exactitud de la predicción	Métrica de precisión	Métrica recall
dia	42,72%	36,21%	0,36	0,36
hora	37,37%	41,62%	0,41	0,41
tipologia	62,14%	58,37%	-	0,58
parroquia_urbana	31,06%	34,59%	0,34	0,34
causas	41,26%	38,10%	-	0,38
nro_heridos	66,01%	64,59 %	-	0,64
nro_fallecidos	97,09%	98,64 %	-	0,98

Se identificó los valores que contienen cada variable siendo asignadas como clase; cada clase presenta un valor independiente de la variable y de la exactitud del modelo predictivo. A

[1] DataMining_Accidentes_Transito_Canton_Loja/Python/modelo at main · Orixstranger/DataMining_Accidentes_Transito_Canton_Loja · GitHub
[2] DataMining_Accidentes_Transito_Canton_Loja/Weka/modelos at main · Orixstranger/DataMining_Accidentes_Transito_Canton_Loja · GitHub

continuación, se detalla los datos obtenidos para cada una de las variables y clases del dataset de accidentes de tránsito en la zona urbana del cantón Loja pertenecientes al año 2021.

Variable “dia”

En la Tabla 37. Representación de porcentajes de clases - variable "dia" se presenta los resultados a través de la predicción del modelo aplicado, en donde se puede evidenciar las predicciones de los días en que pueden ocurrir un siniestro de tránsito, indicando que para la clase “sabado_domingo” se presenta un porcentaje de predicción del 65,68% como probabilidad de ocurrencia de un accidente de tránsito; mientras que, para la clase “lunes_martes_miercoles” el porcentaje de predicción es del 21,89%, y para la clase “jueves_viernes” el porcentaje de predicción es de 12,43%; señalando que, los días sábados y domingos son considerados días con mayor riesgo de accidentabilidad vehicular en la zona urbana del cantón Loja, tal como se presenta en la comparación de datos originales presentes en la Figura 52 y los datos de predicción evidentes en la Figura 53. A su vez,

Tabla 37. Representación de porcentajes de clases - variable "dia"

Nro.	Clase	Datos Día - 2021	Predicción
1	lunes_martes_miercoles	35,4%	21,89%
2	jueves_viernes	32,2%	12,43%
3	sabado_domingo	32,4%	65,68%

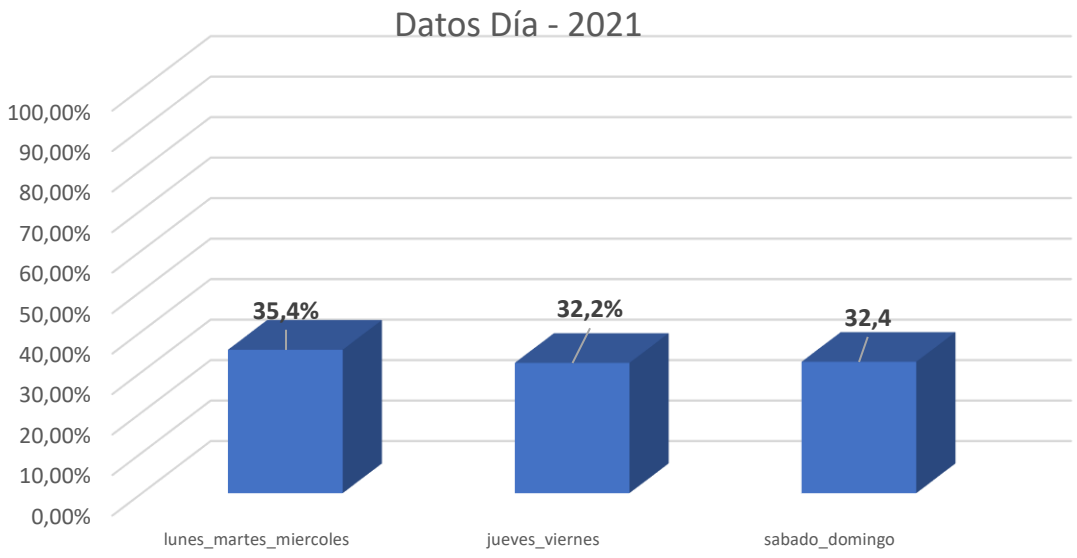


Figura 52. Datos originales de variable “dia” 2021

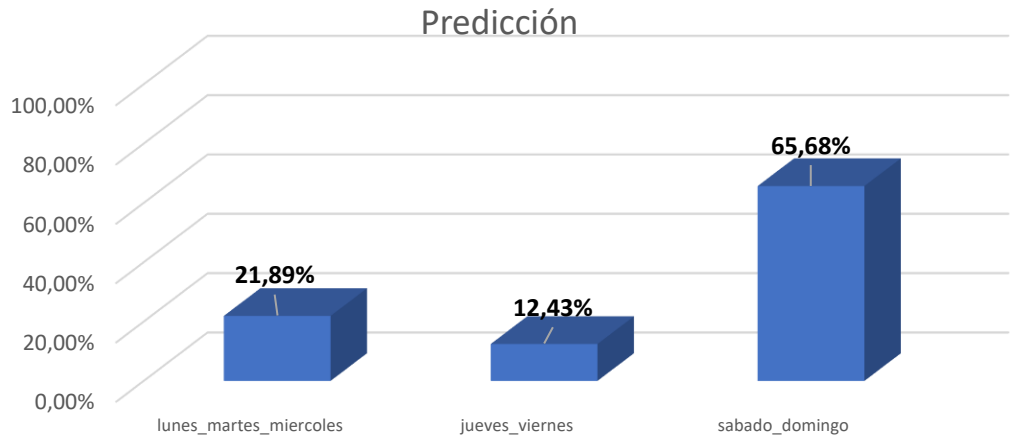


Figura 53. Datos de predicción de modelo para variable "dia" 2021

En la Figura 54 se establece la estructura del árbol de decisión generada por el modelo predictivo para la variable “dia”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[3].

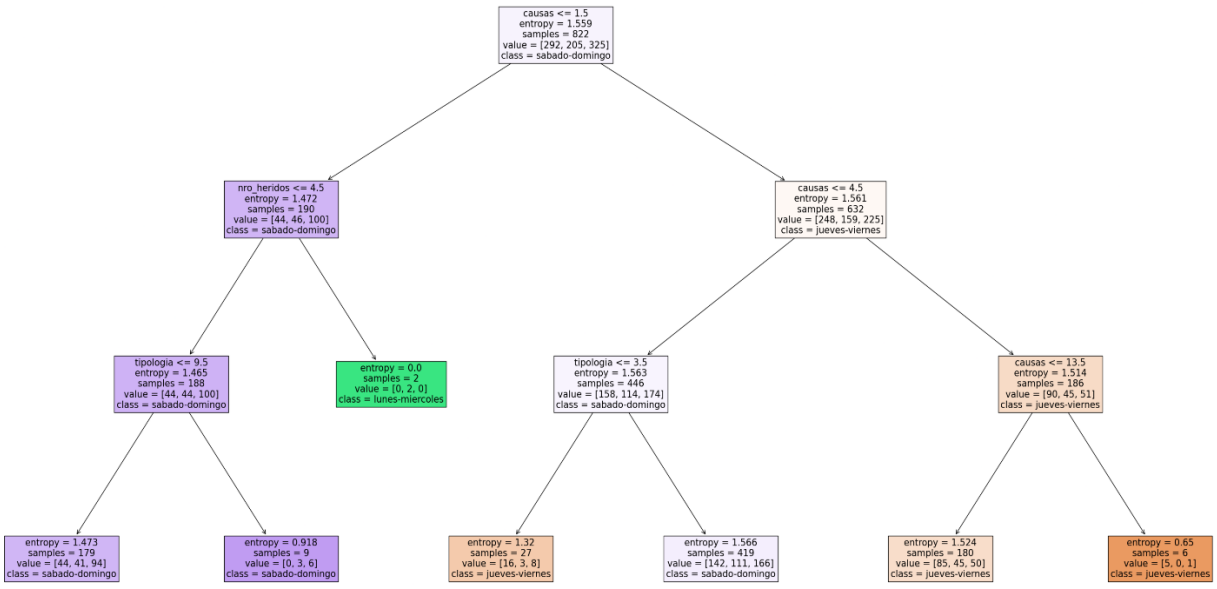


Figura 54. Árbol de decisión de la variable "dia" – 2021

3 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/dia/arbol_colab_dia_2021.png

En la Tabla 38 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[4].

Tabla 38. Probabilidades de accidentes de tránsito variable "dia" - 2021

Registro	Hora	Tipología	Parroquia _ urbana	Causas	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicción
5	h08-h11	choque	sucre	no mantener la distancia	0	0	25,00%	lunes-miercoles
171	h20-h23	choque	san sebastian	imprudencia del conductor	0	0	39,62%	sabado-domingo
360	h16-h19	roce	el sagrario	imprudencia del conductor	0	0	39,62%	sabado-domingo
50	h04-h07	estrellamiento	sucre	conducir en estado de embriaguez	0	0	52,51%	sabado-domingo

Variable “hora”

En la Tabla 39. Representación de porcentajes de clases - variable "hora" se presenta los resultados a través de la predicción del modelo aplicado a la variable “hora”, en donde se puede evidenciar las probabilidades de las horas en que pueden ocurrir un siniestro de tránsito, indicando que para la clase “h12_h17” se presenta un porcentaje de predicción del 49,5% como probabilidad de ocurrencia de un accidente de tránsito; mientras que, la clase “h00_h05” el porcentaje de predicción es del 37% y para la clase “h18_h23” el porcentaje de predicción es de 13,5%; lo que representa que, según el modelo de predicción, en el periodo de horas de 12:00 a 17:00 son considerados los horarios con mayor riesgo de accidentabilidad vehicular en la zona urbana del cantón Loja, tal como se presenta en la comparación de datos originales presentes en la Figura 55 y los datos de predicción evidentes en la Figura 56. Cabe señalar que no se ha considerado el horario de “h06_h11” debido a que el modelo no lo seleccionó al momento de realizar la predicción debido a que existen cuatro clases y lo manifestado en la sección 6.2.1.

Tabla 39. Representación de porcentajes de clases - variable "hora"

Nro.	Clase	Datos Hora - 2021	Predicción
1	h00_h05	17,8%	37%

[4] https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/dia/tabla_probabilidades_dia_2021.pdf

Nro.	Clase	Datos Hora - 2021	Predicción
2	h06_h11	21,6%	-
3	h12_h17	29,7%	49,5%
4	h18_h23	30,8%	13,5%

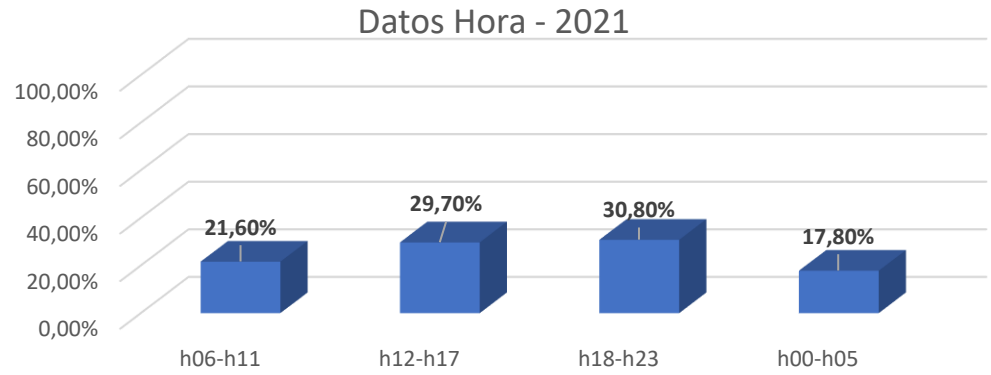


Figura 55. Datos originales de variable “hora” 2021

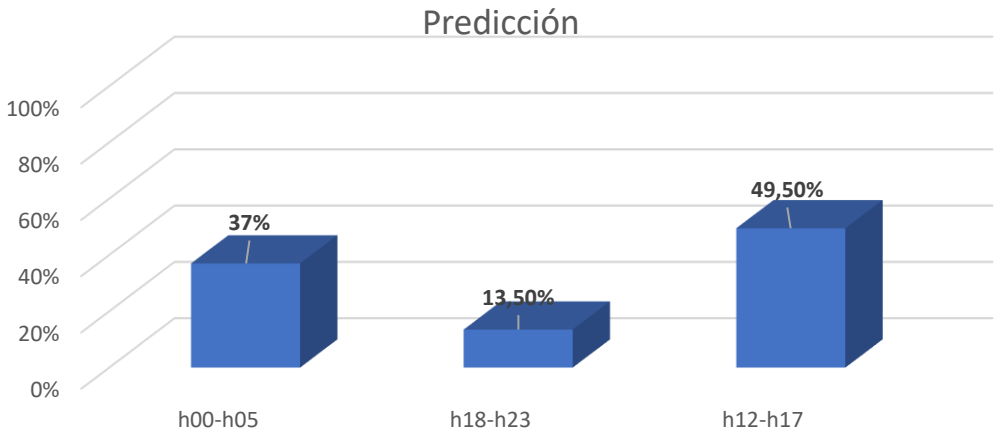


Figura 56. Datos de predicción de modelo para variable "hora" 2021

En la Figura 57 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable “hora”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[5].

5 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/hora/arbol_colab_hora_2021.png

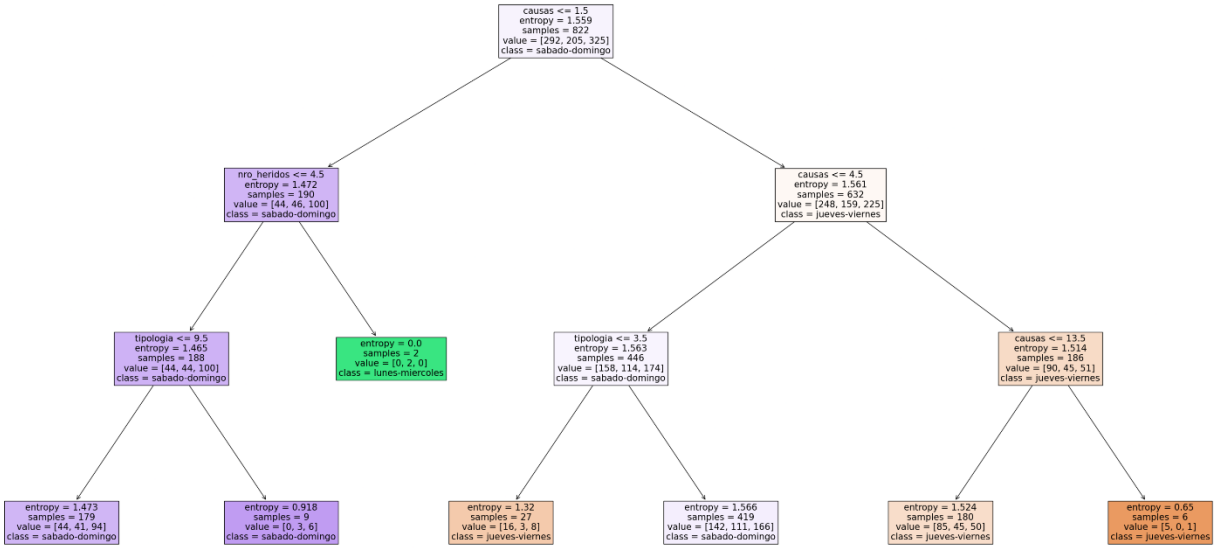


Figura 57. Árbol de decisión de la variable "hora" – 2021

En la Tabla 40 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[6].

Tabla 40. Probabilidades de accidentes de tránsito variable "dia" - 2021

Registro	Día	Tipología	Parroquia_ urbana	Causas	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
147	domingo	estrellamiento	el sagrario	imprudencia del conductor	1	0	39,65%	h00-h05
149	sabado	choque frontal excentrico	san sebastian	imprudencia del conductor	1	0	14,98%	h12-h17
223	jueves	choque lateral perpendicular	sucre	imprudencia del conductor	1	0	39,65%	h12-h17
160	domingo	choque lateral angular	punzara	conducir en estado de embriaguez	1	0	28,63%	h18-h23

6 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/hora/tabla_probabilidades_hora_2021.pdf

Variable “tipologia”

En la Tabla 41. Representación de porcentajes de clases - variable "tipologia" se presenta los resultados a través de la predicción del modelo aplicado a la variable “tipologia”, en donde se puede evidenciar las circunstancias en que produce un siniestro de tránsito, indicando que para la clase “choque” se presenta un porcentaje de predicción del 74,59% como probabilidad de circunstancia de un accidente de tránsito; mientras que, la clase “estrellamiento” presenta el 17,30% de porcentaje de predicción, la clase “atropello” el porcentaje de predicción es del 4,86% y para la clase “perdida de carril_pista” el porcentaje de predicción es del 3,24%; lo que representa que, según el modelo de predicción, la tipología “Choque” es la más probable a ocurrir en un accidente de tránsito en la zona urbana del cantón Loja, tal como se presenta en la comparación de datos originales presentes en la Figura 58 y los datos de predicción evidentes en la Figura 59.

Tabla 41. Representación de porcentajes de clases - variable "tipologia"

Nro.	Clase	Datos Tipología - 2021	Predicción
1	atropello	5,4%	4,86%
2	choque	46,2%	74,59%
3	estrellamiento	30,8%	17,30%
4	roce	3,8%	0%
5	colision	2,7%	0%
6	perdida de carril_pista	7%	3,24%
7	caida de pasajero	0,3%	0%
8	atipico	2,7%	0%
9	encunetamiento	0%	0%
10	arrollamiento	0,8%	0%
11	volcamiento	0,3%	0%

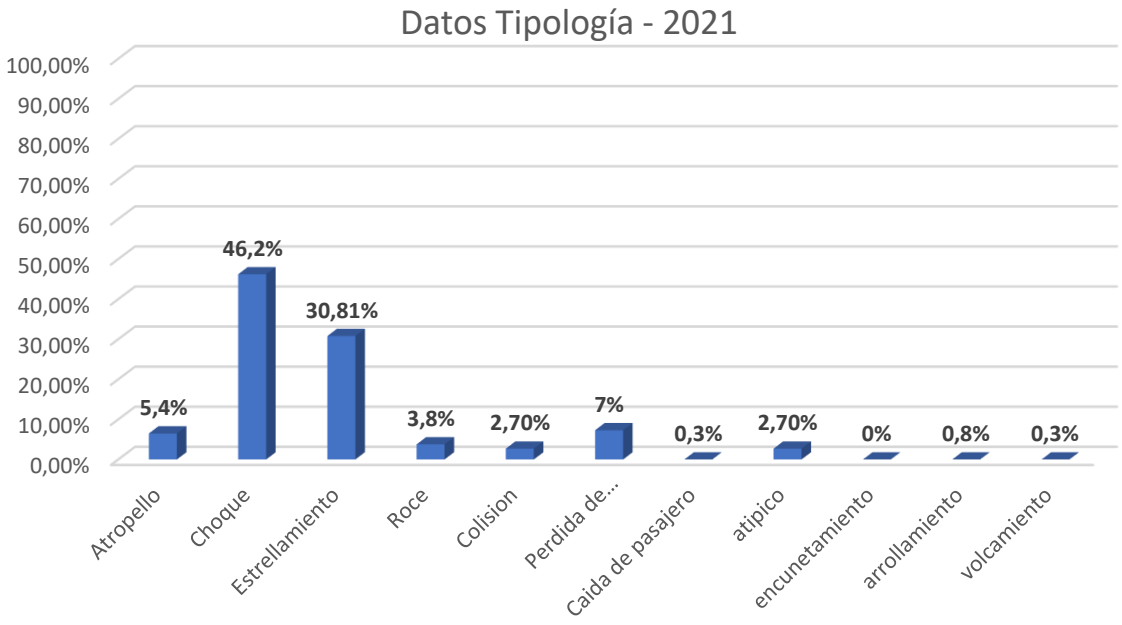


Figura 58. Datos originales de variable “tipologia” 2021

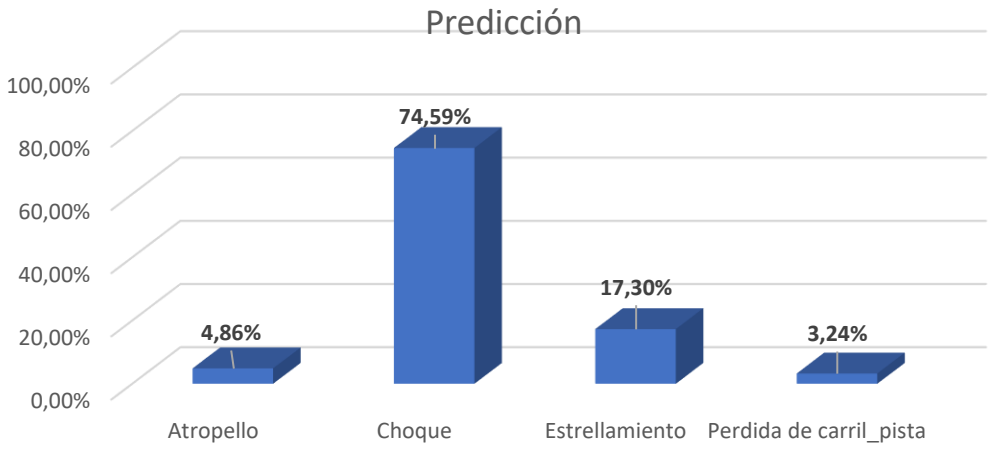


Figura 59. Datos de predicción de modelo para variable "tipologia" 2021

En la Figura 60 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable “tipología”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores

generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[7].

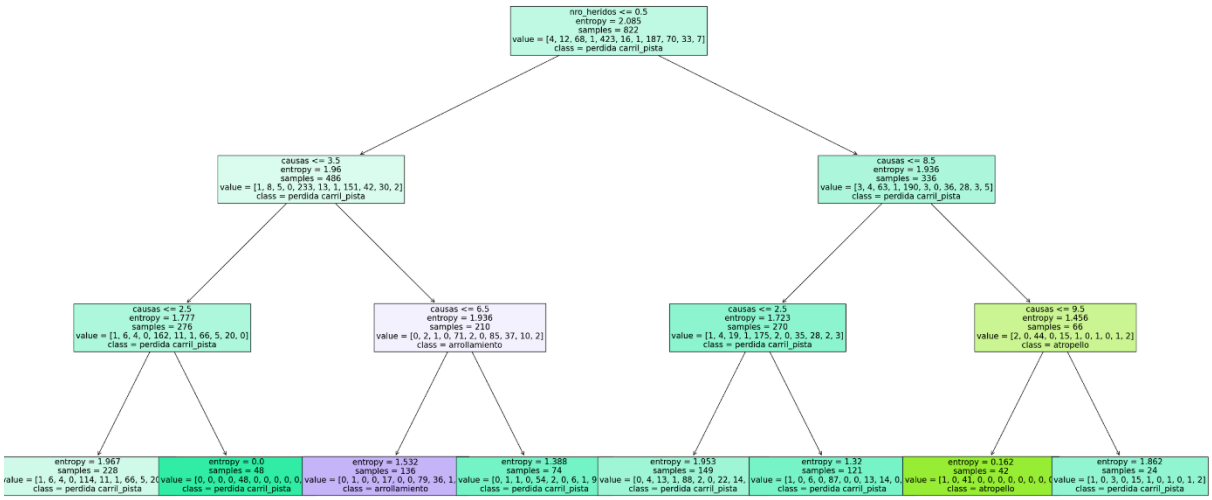


Figura 60. Árbol de decisión de la variable "tipologia" – 2021

En la Tabla 42 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[8].

Tabla 42. Probabilidades de accidentes de tránsito variable "tipologia" - 2021

Registro	Día	Hora	Parroquia_ urbana	Causas	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
356	lunes_ martes_ miercoles	h18-h23	sucre	imprudencia del conductor	3	0	1,34%	choque
97	sabado_ domingo	h06-h11	el sagrario	imprudencia del conductor	0	0	1,34%	choque
274	sabado_ domingo	h00-h05	el sagrario	conducir en exceso de velocidad	0	0	9,40%	estrellamiento

7 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/tipologia/arbol_colab_tipologia_2021.png

8 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/tipologia/tabla_probabilidades_tipologia_2021.pdf

Registro	Día	Hora	Parroquia_ urbana	Causas	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
115	jueves_ viernes	h06-h11	punzara	conducir en estado de embriaguez	0	0	1,34%	choque

Variable “parroquia_urbana”

En la Tabla 43 se presenta los resultados a través de la predicción del modelo aplicado a la variable “parroquia_urbana”, en donde se evidencia los sectores dentro del casco urbano del cantón Loja como factor para la ocurrencia de un siniestro de tránsito; se indica que para la clase “sucre” se presenta un porcentaje de predicción del 75,9% como parroquia con alta probabilidad de ocurrencia de un accidente de tránsito; mientras que, la clase “el sagrario” presenta el 22,7% de porcentaje de predicción y para la clase “punzara” el porcentaje de predicción es del 1,4%; lo que representa que, según el modelo de predicción, la parroquia “Sucre” es el sector con mayor probabilidad de ocurrencias de accidentes de tránsito dentro de la zona urbana del cantón Loja, tal como se presenta en la comparación de datos originales presentes en la Figura 61 y los datos de predicción evidentes en la Figura 62.

Tabla 43. Representación de porcentajes de clases - variable "parroquia_urbana"

Nro.	Clase	Datos Parroquia Urbana - 2021	Predicción
1	sucre	31,1%	75,9%
2	el sagrario	17%	22,7%
3	punzara	16,8%	1,4%
4	san sebastian	14,9%	0%
5	carigan	11,4%	0%
6	el valle	8,9%	0%

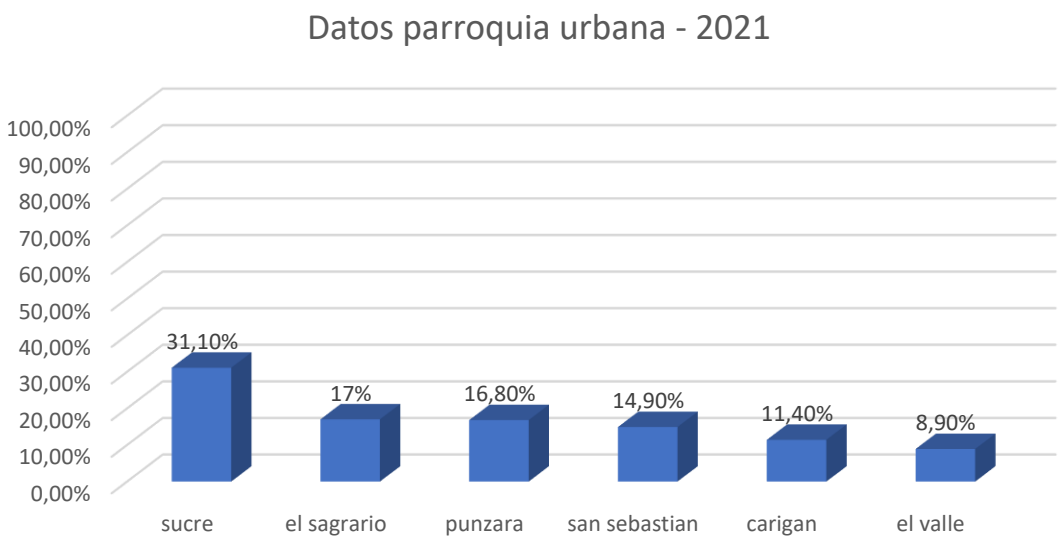


Figura 61. Datos originales de variable “parroquia_urbana” 2021

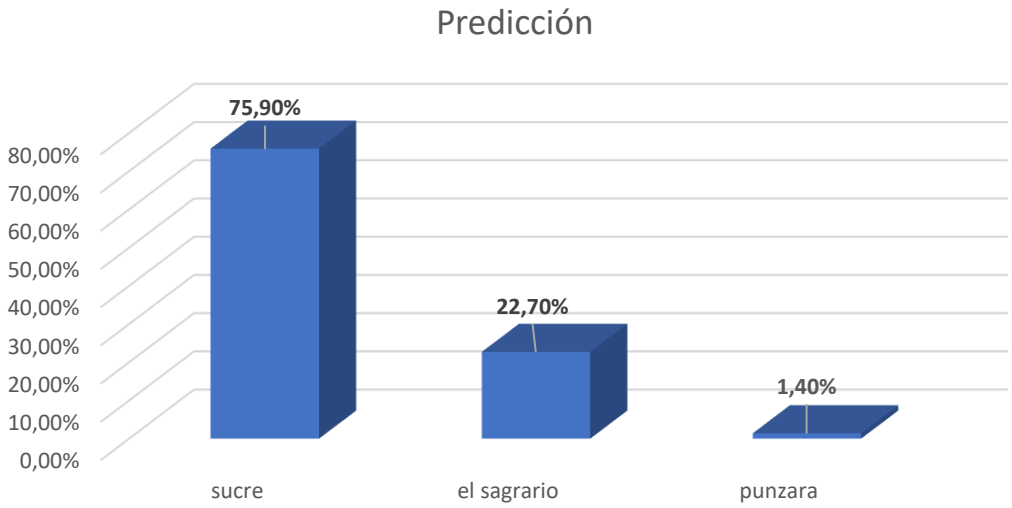


Figura 62. Datos de predicción de modelo para variable "parroquia_urbana" 2021

En la Figura 63 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable “parroquia_urbana”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[9].

9 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/parroquia_urbana/arbol_colab_parroquia_urbana_2021.png

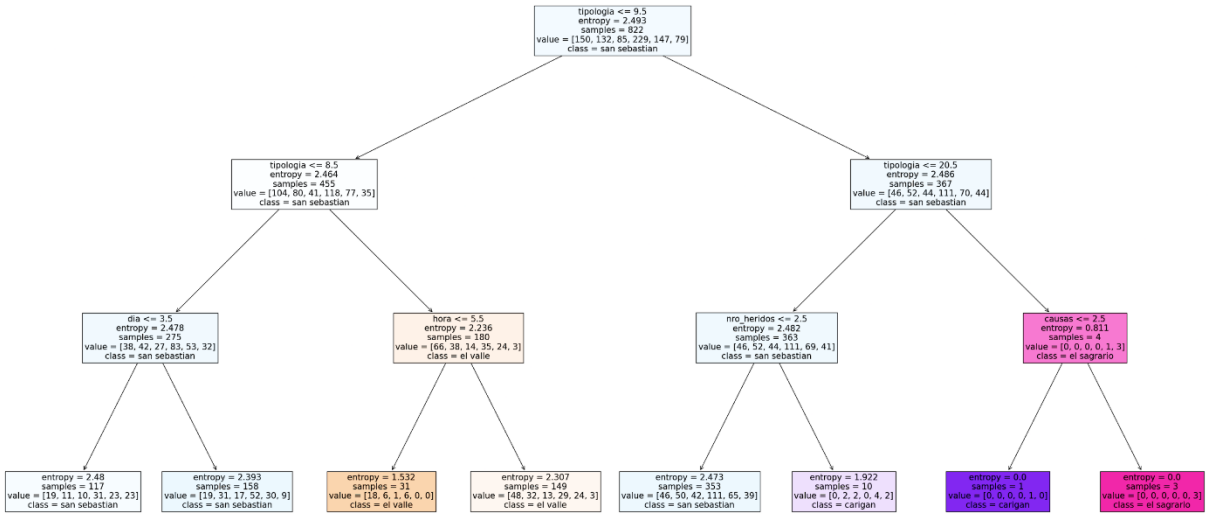


Figura 63. Árbol de decisión de la variable "parroquia_urbana" – 2021

En la Tabla 44 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[10].

Tabla 44. Probabilidades de accidentes de tránsito variable “parroquia_urbana” – 2021

Registro	Día	Hora	Tipología	Causas	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
50	domingo	h04	estrellamiento	conducir en estado de embriaguez	0	0	18,41%	sucre
312	viernes	h00	choque por alcance	conducir en estado de embriaguez	0	0	18,41%	sucre
249	miercoles	h22	choque lateral perpendicular	no respetar las senales de transito	1	0	14,16%	el sagrario
50	martes	h22	choque lateral perpendicular	imprudencia de conductor	2	0	14,16%	el sagrario

10 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/parroquia_urbana/tabla_probabilidades_parroquia_urbana_2021.pdf

Variable “causas”

En la Tabla 45. Representación de porcentajes de clases - variable "causas" se presenta los resultados a través de la predicción del modelo aplicado a la variable “causas”, en donde se puede evidenciar las principales causas implicadas en la ocurrencia de un siniestro de tránsito, indicando que dentro de este conjunto de clases, se identifica que la clase “conducir en exceso de velocidad ” se presenta un porcentaje de predicción del 37,8% como la causa con más alto porcentaje de probabilidad; lo que representa que, según el modelo de predicción, el conducir en exceso de velocidad es la causa más probable para producir un accidente de tránsito en la zona urbana del cantón Loja, tal como se presenta en la comparación de datos originales presentes en la Figura 64 y los datos de predicción evidentes en la Figura 65.

Tabla 45. Representación de porcentajes de clases - variable "causas"

Nro.	Clase	Datos Causa - 2021	Predicción
1	Conducir en exceso de velocidad	20,3%	37,8%
2	No ceder derecho de via	0,5%	37,3%
3	Imprudencia del conductor	43,5%	9,5%
4	Imprudencia del peaton	4,1%	7%
5	No mantener la distancia reglamentaria	0,5%	6,5%
6	Conducir en estado de embriaguez	23,8%	1,9%
7	No respetar las señales de tránsito	4,1%	0%
8	Condiciones climatológicas desfavorables	0,5%	0%
9	Impericia del conductor	0,5%	0%
10	Fallas mecánicas no previsibles	1,6%	0%
11	Cruce de animales	0,5%	0%

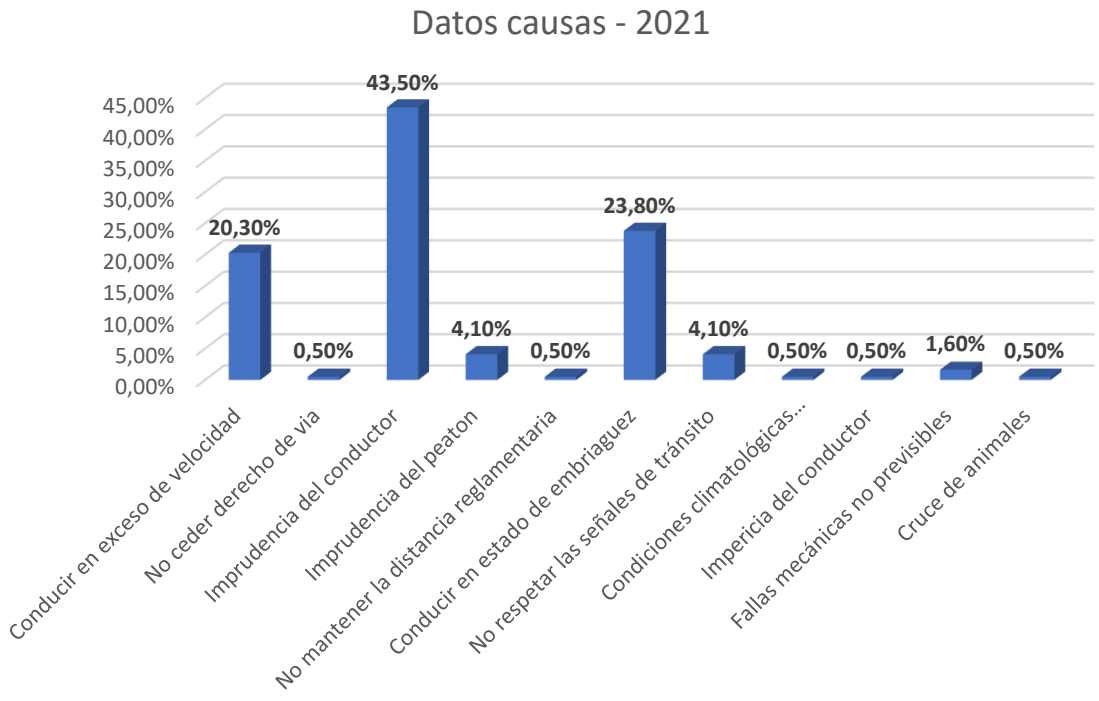


Figura 64. Datos originales de variable “causas” 2021

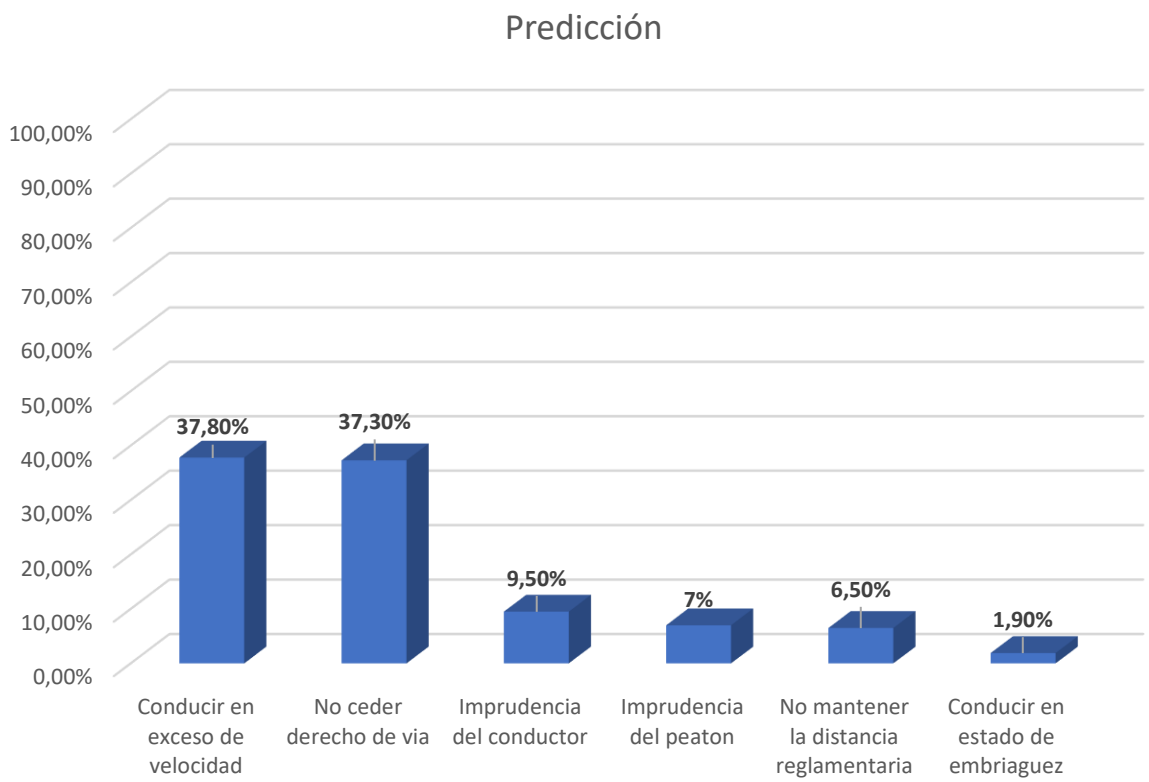


Figura 65. Datos de predicción de modelo para variable "causas" 2021

En la Figura 66 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable "causas", en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[11].

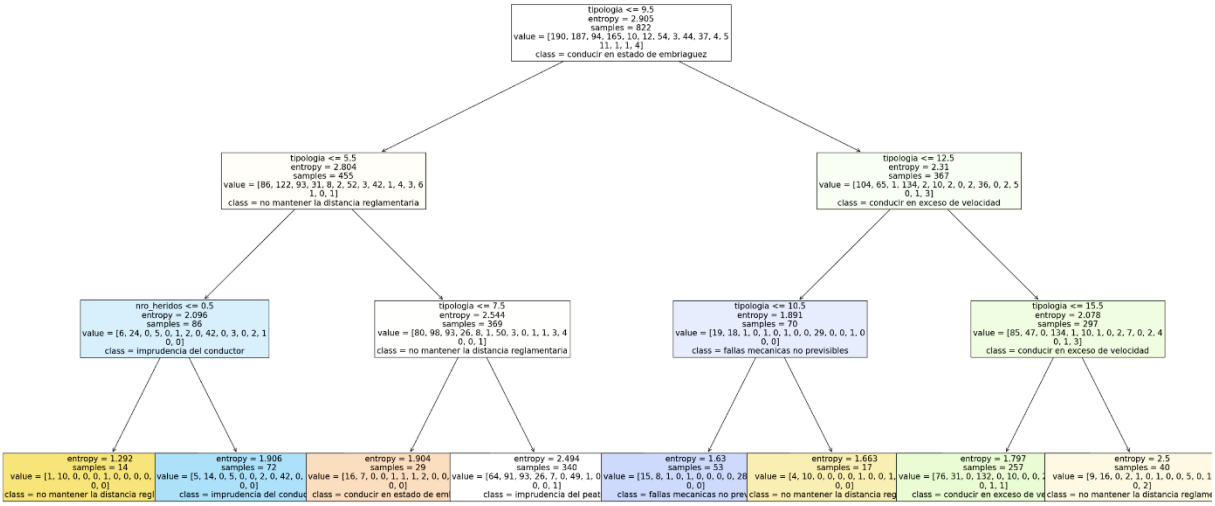


Figura 66. Árbol de decisión de la variable "causas" – 2021

En la Tabla 46 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[12].

Tabla 46. Probabilidades de accidentes de tránsito variable "causas" - 2021

Registro	Día	Hora	Tipología	Parroquia_ urbana	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
60	sabado	h12	choque lateral perpedicular	el sagrario	1	0	51,36%	no ceder derecho de via
300	sabado	h15	estrellamiento	carigan	0	0	0%	conducir en exceso de velocidad

11 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/causas/arbol_colab_causas_2021.png

12 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/causas/tabla_probabilidades_causas_2021.pdf

Registro	Día	Hora	Tipología	Parroquia_ urbana	Nro_ heridos	Nro_ fallecidos	Probabilidad de accidente	Predicciones
12	sabado	h20	choque lateral angular	sucre	0	0	51,36%	no ceder derecho de via
203	sabado	h07	atropello	sucre	1	0	0,78%	imprudencia del peaton

Variable "nro_heridos"

En la Tabla 47. Representación de porcentajes de clases - variable "nro_heridos" se presenta los resultados a través de la predicción del modelo aplicado a la variable "nro_heridos", en donde se evidencia registros de personas heridas con algún producto de lesión o complicación física producto de un siniestro de tránsito en el casco urbano del cantón Loja; se indica que para la clase "0h" se presenta un porcentaje de predicción del 92,4% con alta probabilidad de personas heridas en un accidente de tránsito y para la clase "1h" el porcentaje de predicción es del 7,6%; lo que representa que, según el modelo de predicción, la mayor probabilidad de personas que resulten con daños físicos producto de la ocurrencia de los accidentes de tránsito dentro de la zona urbana del cantón Loja es de cero personas; sin embargo, la predicción del 7,6% presentó que al menos una persona resulta con lesiones o complicaciones físicas, tal como lo presentan los registros originales de los datos del 2021, siendo este un número significativo, como se puede visualizar la comparación de datos originales presentes en la Figura 67 y los datos de predicción evidentes en la Figura 68.

Tabla 47. Representación de porcentajes de clases - variable "nro_heridos"

Nro.	Clase	Datos de personas heridas - 2021	Predicción
1	0h	58,9%	92,4%
2	1h	31,4%	7,6%
3	2h	7%	0%
4	3h	1,6%	0%
5	4h	0%	0%
6	5h	0,3%	0%
7	6h	0,2%	0%
8	7h	0,3%	0%

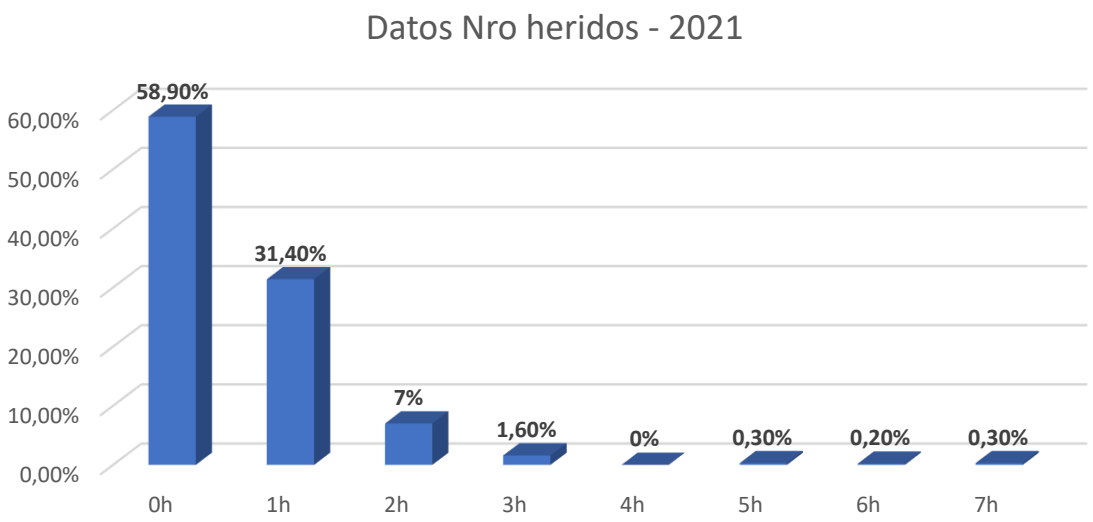


Figura 67. Datos originales de variable “nro_heridos” 2021

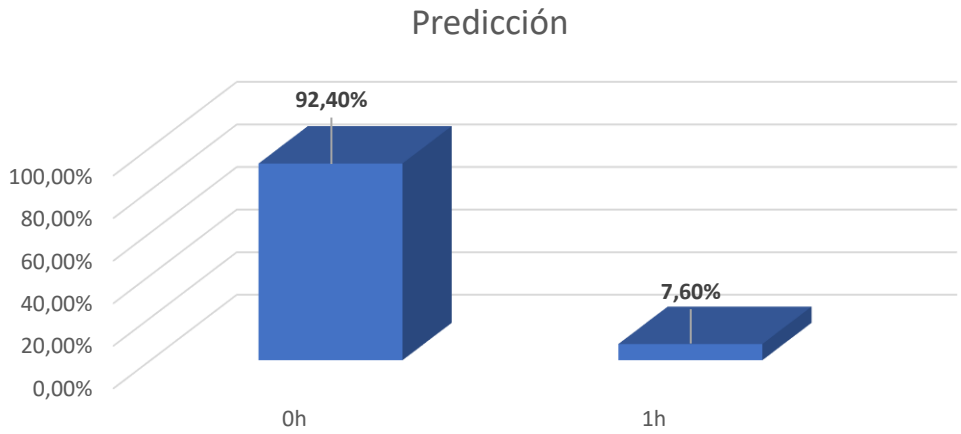


Figura 68. Datos de predicción de modelo para variable "nro_heridos" 2021

En la Figura 69 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable “causas”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[13].

13 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/nro_heridos/arbol_colab_nro_heridos_2021.png

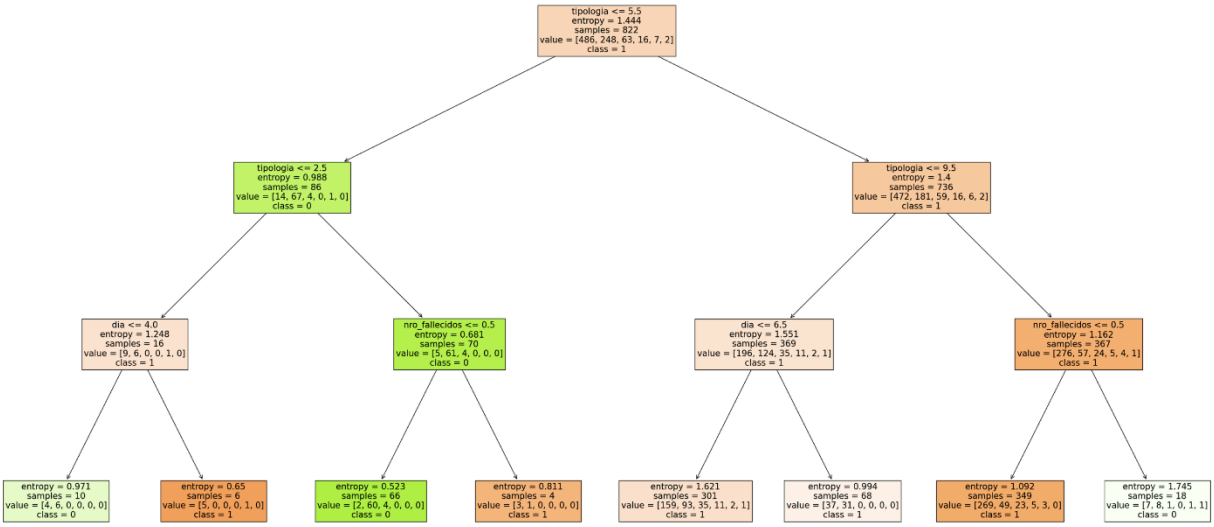


Figura 69. Árbol de decisión de la variable "nro_heridos" – 2021

En la Tabla 48 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[14].

Tabla 48. Probabilidades de accidentes de tránsito variable "nro_heridos" - 2021

Registro	Día	Hora	Tipología	Parroquia_ urbana	Causas	Nro_ fallecidos	Probabilidad de accidente	Predicciones
77	miercoles	h13	choque lateral perpendicular	el sagrario	imprudencia del conductor	0	77,08%	0
3	viernes	h10	atropello	sucre	imprudencia del peaton	0	14,04%	1
269	domingo	h11	choque lateral perpendicular	san sebastian	no respetar las señales de transito	0	77,08%	0
357	martes	h21	choque alcance	el valle	conducir en estado de embriaguez	0	77,08%	0

Variable “nro_fallecidos”

En la Tabla 49. Representación de porcentajes de clases - variable "nro_fallecidos" se presenta los resultados a través de la predicción del modelo aplicado a la variable “nro_fallecidos”, en

14 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/nro_heridos/tabla_probabilidades_nro_heridos_2021.pdf

donde se evidencia registros de personas que pierden la vida producto de un siniestro de tránsito en el casco urbano del cantón Loja; se indica que para la clase “0fa” se presenta un porcentaje de predicción del 99,7% en un accidente de tránsito y para la clase “1fa” el porcentaje de predicción es del 0,3%; lo que representa que, según el modelo de predicción, la mayor probabilidad de personas que pierdan la vida en la ocurrencia de los accidentes de tránsito dentro de la zona urbana del cantón Loja; sin embargo, la predicción del modelo aun obteniendo un alto nivel de exactitud no es confiable debido a la reducida cantidad de datos como objeto de estudio, si bien, el modelo predice con un 98,6% de exactitud, resultó evidente que existen personas fallecidas dentro de los dataset, pero al ser reducida cantidad de información el modelo no los selecciona en el proceso de la minería de datos, como se puede visualizar la comparación de datos originales presentes en la Figura 70 y los datos de predicción evidentes en la Figura 71.

Tabla 49. Representación de porcentajes de clases - variable "nro_fallecidos"

Nro.	Clase	Datos de personas fallecidas – 2021	Predicción
1	0fa	98,6%	99,7%
2	1fa	1,4%	0,3%

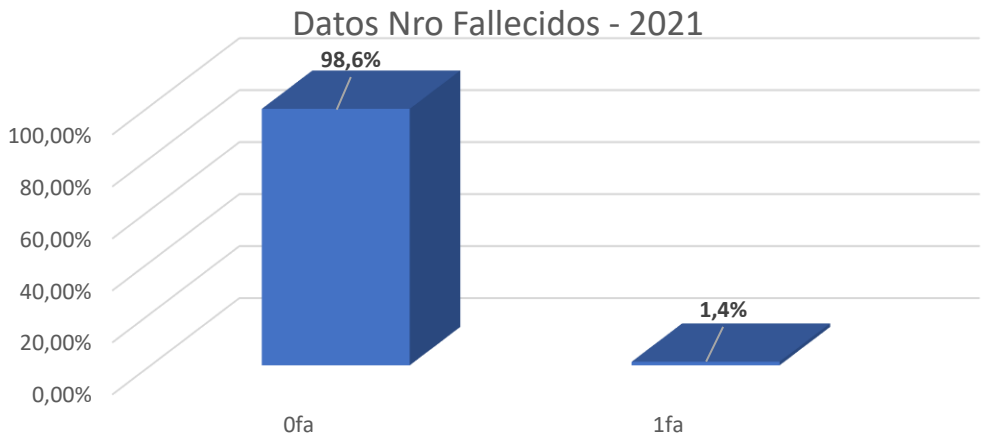


Figura 70. Datos originales de variable “nro_fallecidos” 2021

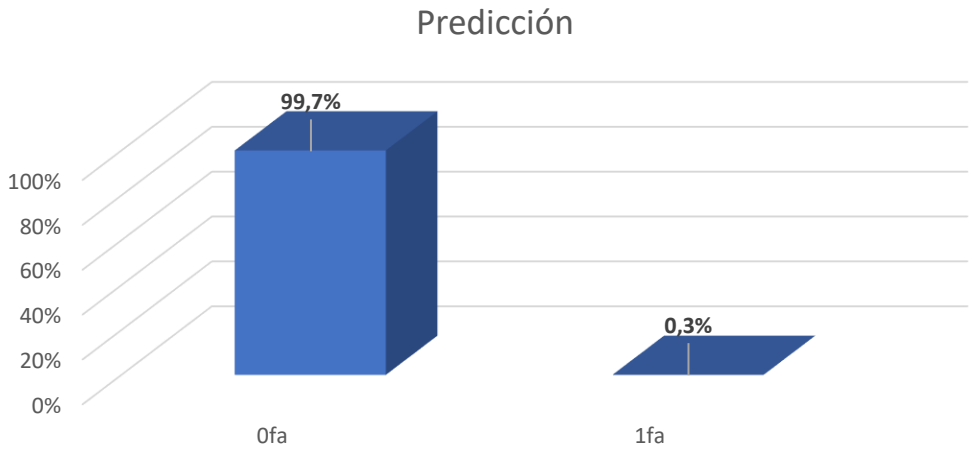


Figura 71. Datos de predicción de modelo para variable "nro_fallecidos" 2021

En la Figura 72 se estableció la estructura del árbol de decisión generada por el modelo predictivo para la variable “causas”, en este caso se visualiza los valores de los nodos resultantes que señalan la distribución de clases que contiene el modelo, así como los valores generados por la entropía y el número de muestras en cada uno, para finalmente señalar su clasificación, para mejor visualización de la figura, véase el siguiente enlace[15].

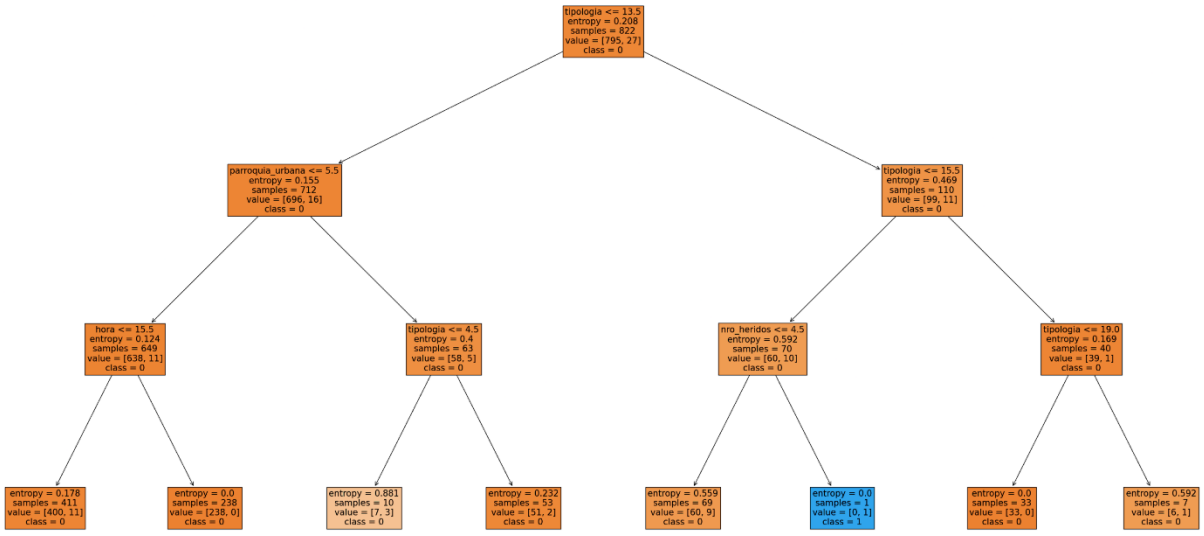


Figura 72. Árbol de decisión de la variable "nro_fallecidos" – 2021

15 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/nro_fallecidos/arbol_colab_nro_fallecidos_2021.png

En la Tabla 50 se observan un extracto de las variables, porcentajes de probabilidades y predicciones resultantes de los registros de accidentes de tránsito registrados en el año 2021, para poder evidenciar el total de probabilidades obtenidas véase el siguiente enlace[16].

Tabla 50. Probabilidades de accidentes de tránsito variable "nro_fallecidos" - 2021

Registro	Día	Hora	Tipología	Parroquia_ urbana	Causas	Nro_ heridos	Probabilidad de accidente	Predicciones
229	miercoles	h13	choque lateral perpedicular	el sagrario	imprudencia del conductor	1	97,32%	0
54	domingo	h21	perdida carril	sucre	fallas mecanicas no previsibles	6	2,68%	1
312	viernes	h00	choque alcance	san sebastian	conducir en estado de embriaguez	0	97,32%	0
47	domingo	h12	choque lateral angular	punzara	conducir en estado de embriaguez	2	97,32%	0

Los códigos generados, pruebas y resultados detallados de cada una de las variables con sus respectivos casos que fueron aplicados para la minería de datos en el entorno google colab se encuentra adjuntos dentro del anexo 11.

16 https://github.com/Orixstranger/DataMining_Accidentes_Transito_Canton_Loja/blob/main/Python/probabilidades%202021/nro_fallecidos/tabla_probabilidades_nro_fallecidos_2021.pdf

7. Discusión

El enfoque de la presente sección es ratificar los objetivos que conforman el Trabajo de Titulación (TT) con afán de dar cumplimiento al objetivo general basado en la implementación de minería de datos en la accidentabilidad vehicular en la zona urbana del cantón Loja, cuyo propósito fue conocer el comportamiento de accidentabilidad vehicular, colaborar en la toma de decisiones de las entidades de regulación y control del tránsito para fortalecer la seguridad vial en el cantón Loja y brindar un apoyo a la sociedad; además, solventar el siguiente problema de investigación: ¿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?. En base a los resultados obtenidos de los tres objetivos planteados, se ratifica la implementación de minería de datos, donde según las pruebas realizadas a los modelos predictivos permiten justificar e identificar los patrones de accidentabilidad vehicular y probabilidades de un accidente de tránsito dentro del cantón Loja.

7.1. Desarrollo de la propuesta alternativa

El desarrollo del presente TT fue basado en el desarrollo de tres objetivos específicos con su respectiva discusión frente a otros trabajos. Cada una de las etapas más importantes de los procesos realizados, actividades, resultados obtenidos y limitaciones se presentan a continuación.

7.1.1. Objetivo 1: Analizar los repositorios de datos referente a los accidentes de tránsito registrados en la UCOT en el año 2019 - 2020.

El primer objetivo consistió en el establecimiento de criterios y lineamientos para acceder a la información recopilada por la Unidad de Control Operativo de Tránsito (UCOT) pertenecientes al GAD Municipal de Loja, lo que permitió obtener las bases de datos relevantes relacionadas sobre accidentabilidad vehicular dentro de la zona urbana del cantón durante el periodo 2018 – 2021. La UCOT aportó el permiso para obtener la información relevante y ser analizada por el suscrito, por lo que, se adquirió información respecto a los datos únicos y específicos del cantón en donde se encuentra las competencias de la institución, esto en relación a los estudios [3], [6], [8], [10], [11], [14], [31], [32], [33] y [37] en los cuales se investigaron ciudades y países con mayor flujo vehicular y alto índice de siniestros de tránsito, cuyas bases de datos contienen registros de accidentabilidad vehicular aportados por las entidades de regulación de tránsito.

La base de datos inicialmente ayudó con la información de accidentabilidad vehicular, no obstante, al observar las variables, se notó que a medida que los datos eran más recientes, el registro contenía mayor cantidad de variables, por lo que se evidenció que algunas variables no se relacionaban con los periodos anteriores o incluso no estaban presentes en ellos; para abordar esto, se realizó la aplicación del Data Cleaning, con la finalidad de identificar los datos erróneos o con menor relevancia de la base de datos como se observan en los estudios [6], [8], [11], [14] y [32]; por lo tanto, a través de la herramienta OpenRefine permitió realizar con efectividad el filtrado de datos, ya que fue ideal para trabajar con datos desordenados ajustándose a la exploración, limpieza, transformación y coincidencia de los datos, depurando y estandarizando las variables para el proceso de data mining.

En este proceso se consideró como limitante importante para el presente TT, el reducido registro de información de accidentabilidad vial que posee la UCOT, que abarca desde el año 2018 hasta el año 2021. Lamentablemente, la información del año 2022 no se encuentra digitalizada; además, los registros de datos de accidentes pertenecientes a los años anteriores al 2018 no estuvieron disponibles, esto se debe a que la entidad no los registraba o no lo hacían con información detallada, datos que hubieren sido útiles como antecedentes para la predicción del modelo, lo que podría haber mejorado la exactitud y calidad de los modelos durante el proceso de data mining.

7.1.2. Objetivo 2: Implementación del modelo de árboles de decisión para desarrollar el análisis exploratorio de datos.

En la implementación del modelo se requirió el análisis de las variables contenidas en la base de datos, siendo comparados cada uno de los periodos de accidentabilidad vehicular en la zona urbana del cantón Loja, en donde coincidieron siete variables relevantes considerándolas como objeto de estudio, en contraste con los estudios [9], [37], [6], [10] y [39] quienes realizaron la predicción de datos de accidentes de tránsito a través de la utilización de mínimo seis variables.

El modelo de árboles de decisión fue implementado con sus respectivos algoritmos como eje fundamental para el estudio de accidentes de tránsito, considerando los datos generados en el cantón Loja, en relación a los estudios de [37], [33], [8], y [6] que indican la eficiente aplicación del modelo en las áreas, zonas y ciudades con mayor afluencia vehicular, quienes establecen la generación de datos mediante el desarrollo y estudio de predicciones como información primordial para la toma de decisiones.

Las principales herramientas aplicadas al objeto de estudio fueron Python y WEKA, como se menciona en los estudios [27] y [39], donde las señalan como herramientas adecuadas debido a sus algoritmos de modelado de software aplicados a los aprendizajes supervisados para la minería de datos; además, sus interfaces gráficas incorporadas son adaptables a los datos proporcionados; por lo tanto, para Python se implementó la librería sklearn y se utilizó el algoritmo predefinido CART, que presenta la exactitud de la predicción, métricas de precisión que señala el porcentaje de valores clasificados positivos y métrica recall para los valores de sensibilidad del modelo, que contrastan con los datos obtenidos por la herramienta WEKA, que se utilizó el algoritmo J48, siendo la adaptación del algoritmo C4.5, en donde generó las instancias correctamente clasificadas, que representa la exactitud del modelo, así como los resultados de las métricas de precisión y recall; estos dos algoritmos tuvieron una ventaja en la capacidad de predicción, estructura y división del árbol construido; además, se aplicó un conjunto de pruebas con el objetivo de obtener los niveles de predicción superiores y ser reflejada esta información en el TT, con afán de aportar datos concretos a las instituciones y entes responsables del ámbito del sistema de movilidad del cantón; por lo que para la herramienta Python se realizaron un total de 22 pruebas, a diferencia de la herramienta WEKA que se realizaron 21 pruebas, generando 43 modelos diferentes, siendo este proceso satisfactorio para verificar los resultados más óptimos en la toma de decisiones.

En este proceso se denotó como limitante al algoritmo de predicción de scikit-learn de Python, puesto que solo admite variables numéricas, por lo que para la predicción de datos dentro de la plataforma se transformó los registros categóricos a numéricos, lo que permitió que el modelo funcione generado los resultados; a su vez, los registros obtenidos por la predicción fueron nuevamente transformados a categóricos para ser evidenciados de manera más adecuada. Otra limitante fue la herramienta WEKA, puesto que si bien permite identificar el porcentaje del patrón generado por las 7 variables, no permiten identificar la probabilidad de los registros de accidentabilidad y realizar una comparativa; además su herramienta gráfica no permitió visualizar de manera adecuada los árboles de decisión generados, puesto que a pesar de podar el árbol de decisión a 3 nodos, no se pudo estudiar los datos por lo que estos gráficos eran extensos y de grandes dimensiones, siendo imposible guardar los archivos; por ello, se utilizó la herramienta Python para evidenciar las gráficas y las probabilidades de los modelos de predicción; a su vez, a pesar que se aporta con datos relevantes para la predicción de accidentes de tránsito, los modelos son adaptables a bases de datos futuras, en donde a través de la ejecución de pruebas realizadas a cada una de las variables pudo incrementar el nivel de

exactitud y precisión del modelo, generando nuevos resultados de datos siendo considerados para obtener resultados más fiables de accidentabilidad vehicular, así como que poder ser comparados desde otro enfoque por diferentes modelos predictivos.

7.1.3. Objetivo 3: Evaluación de la técnica de minería de datos propuesta.

En el proceso de evaluación de la técnica de minería de datos a través del modelo predictivo se realizó 43 pruebas para las siete variables del conjunto de datos desarrollados por las herramientas Python y Weka. Se eligió un solo modelo con mayor porcentaje de exactitud, métrica de precisión y métrica recall por cada variable, seleccionando 7 modelos predictivos resultantes, tal como lo realizan los estudios [27] y [39], que efectúan un análisis y comparación los porcentajes de exactitud de cada modelo generado por los distintos algoritmos, así como de los datos obtenidos en la predicción; la finalidad de implementar los modelos y ejecutarlos con registros de datos de accidentabilidad vehicular suscitados en el cantón Loja durante el año 2021 fue identificar si los niveles de exactitud y rendimiento varían para beneficiar la toma de decisiones de las instituciones encargadas de la movilidad vehicular y ejecutar acciones disuasivas para prevenir siniestros de tránsito.

Con el objeto identificar las principales situaciones en las que puede suscitarse un accidente de tránsito, se determinó que los modelos con mayor porcentaje de exactitud son: las variables referente a tipología con un 62,14%; la variable número de heridos posee un 66,01% de predicción; mientras que, la predicción del número de fallecidos presenta un porcentaje de 97,09% con métricas de precisión y sensibilidad del 98,64%; sin embargo al ejecutar los datos en las variables del año 2021, los resultado de la exactitud del modelo variaron, en donde el porcentaje de exactitud presentan que las variables referente a tipología con un 58,37% con una métrica de sensibilidad de 58% en donde el tipo de accidente con mayor predicción de ocurrencia es el choque con un 74,59%; la variable número de heridos posee un 64,59% con una métrica de precisión y sensibilidad de 64,6% señalando que existe la predicción del 92,40% que no exista heridos en un accidente de tránsito y el 7,60% de predicción que al menos una persona resulte herida; por último para el número de fallecidos se tiene un porcentaje de 98,64% de predicción con métricas de precisión y sensibilidad del 98,6% que menciona la predicción del 99,7% de que no resulte una persona fallecida en un siniestro vial y al menos un 0,3% de que una persona resulte fallecida. Por lo tanto, mediante el estudio realizado en el TT, se pudo establecer que existe la probabilidad del 77,08% de que el día domingo en el horario de 11:00 a 11:59, exista un choque lateral perpendicular como tipología de accidente de tránsito en la

parroquia de San Sebastián, cuyo causa sea la de no respetar las señales de tránsito y no existan personas heridas o fallecidas; al contrario, existe un 14,04% de probabilidad de los días viernes, en el horario de 10:00 a 10:59 se produzca un atropello como tipología de accidente de tránsito en la parroquia sucre, cuya causa es por imprudencia del peatón en donde resulte una persona herida y no se encuentren personas fallecidas, todos estos datos recopilados dentro del casco urbano del cantón; sin embargo, es importante destacar que, esta predicción del modelo aun obteniendo un alto nivel de exactitud, presenta reducida cantidad de datos como objeto de estudio, si bien, el modelo predice para la variable número de fallecidos un 98,64% de exactitud, resulta evidente que existe al menos un 2,68% de probabilidad de que una persona resulte fallecida en un siniestro de tránsito. Cabe mencionar que, dentro de los dataset al ser reducida cantidad de información de los registros aplicados en la minería de datos, los modelos poseen niveles de porcentajes bajos de predicción puesto a que no se logró poseer información o data de accidentes viales de años anteriores al 2018 en el cantón Loja; consecuentemente, a través de los estudios [3], [11], [14], [32], [34] y [37] en donde se implementó la metodología KDD para la minería de datos y modelado predictivo de accidentabilidad vehicular en diferentes ciudades del país con mayor flujo vehicular, se consideró que fue la más viable para la predicción de accidentabilidad vehicular dentro del cantón Loja, puesto que este cantón registra mayor índice de siniestralidad vehicular y tiene el mayor flujo vehicular en la provincia; así también a través de sus 5 etapas permitió analizar los datos de las 7 variables de siniestros de tránsito, generar los modelos de accidentes de tránsito, verificar y evaluar los resultados generados por la minería de datos y establecer las probabilidades de la ocurrencia de un siniestro de tránsito en donde a diferencia de los estudios [3], [6], [8], [10], [11], [14], [31], [32], [33] y [37] se obtuvo en un solo aporte de investigación la obtención de 7 modelos predictivos diferentes, dinámicos y que permiten adaptarse a nuevos registros de datos de accidentabilidad vehicular que en un futuro pueden implementarse como objeto de estudio de modo que a medida que los datos se incrementen en los registros, los porcentajes de exactitud del modelo serán más precisos.

Finalmente, al obtener toda esta información tal como lo realizaron los estudios se presentó al departamento UCOT, quienes actualmente utilizan estos resultados para la toma de decisiones enfocadas a la regulación, control y seguridad vehicular por medio de capacitaciones al cuerpo de agentes civiles de tránsito, así también como la ejecución de campañas educativas de conciencia vial en las principales vías rápidas de la ciudad para reducir el índice de siniestralidad; por lo tanto, por medio del estudio realizado se pudo establecer que a través de

los modelos predictivos de minería de datos y de las probabilidades generadas por cada una de las variables de accidentes de tránsito suscitados a nivel urbano del cantón, se pueden identificar los patrones de accidentabilidad vehicular en la zona urbana del cantón Loja, logrando responder la pregunta de investigación.

7.2. Valoración técnica, económica, ambiental y social

7.2.1. Valoración técnica

Los recursos técnicos aplicados para el TT se basan en la utilización de herramientas software, tales como OpenRefine para data cleaning de los datos, Python con aplicación de los algoritmos de clasificación CART y WEKA con implementación del algoritmo J48; los que sirvieron para realizar el proceso de minería de datos y generar los modelos con los que se realizaron las respectivas pruebas de predicción.

7.2.2. Valoración económica

Para el desarrollo del presente TT se utilizaron recursos que fueron necesarios para el desarrollo de la minería de datos, involucrando directamente el recurso humano, tal como se observa en la Tabla 51. Recursos de talento humano, recursos técnicos y tecnológicos especificados en la Tabla 52. Recursos técnicos y tecnológicos; por último, recursos para servicios, como se indican en la Tabla 53.

Tabla 51. Recursos de talento humano

Recurso Humano			
Responsable	**Número de horas**	**Costo por hora**	**Costo total**
Investigador	480	$ 4,00	$ 1920,00
Director TT	80	$ 10,47	$ 837,60
Docente	80	$ 10,47	$ 837,60
	TOTAL		**$ 3595,20**

Tabla 52. Recursos técnicos y tecnológicos

Recursos Técnicos y Tecnológicos		
Recursos de Software		
Nombre del Recurso	**Cantidad**	**Costo total**
Python	1	$ 0,00
WEKA	1	$ 0,00
OpenRefine	1	$ 0,00

Recursos Técnicos y Tecnológicos		
Recursos de Software		
Nombre del Recurso	**Cantidad**	**Costo total**
Mendeley	1	$ 0,00
Entorno Google Colab	1	$ 0,00
Google Drive	1	$ 0,00
Firma Electrónica	1	$ 30,00
Subtotal	**$ 30,00**	
Recursos Hardware		
Computador	1	$ 1400,00
Subtotal	**$ 1400,00**	
TOTAL	**$ 1430,00**	

Tabla 53. Recursos de servicios

Servicios			
Tipo de servicio	**Tiempo en meses**	**Costo unitario**	**Costo total**
Transporte	5	$ 20,00	$ 100,00
Internet	5	$ 22,00	$ 110,00
	TOTAL		**$ 210,00**

Finalmente se presenta la Tabla 54. Recursos totales utilizados en el TT, con el presupuesto aproximado que fue necesario para el desarrollo del TT.

Tabla 54. Recursos totales utilizados en el TT

Recursos Totales	
Recurso humano	$ 3595,20
Recurso técnico y tecnológico	$ 1430,00
Recurso de servicios	$ 210,00
Subtotal	$ 5235,20
Gastos imprevistos (10%)	$ 523,52
Presupuesto total del TT	**$ 5758,72**

El presupuesto de $ 5758,72 (Cinco mil setecientos cincuenta y ocho dólares con setenta y dos centavos), que fueron financiados por el autor y el recurso humano docente de la Universidad Nacional de Loja.

7.2.3. Valoración social

El presente TT, está orientado hacia la necesidad de la Unidad de Control Operativa de Tránsito (UCOT), de realizar un análisis de datos para determinar la focalización de zonas de mayor riesgo de siniestros viales, lo que promueve que a través de la información otorgada por la institución promueva predicciones de accidentes de tránsito para la toma de decisiones y futuros patrones de comportamientos viales, velando por la integridad de las personas, así como la 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, fue referente sobre la aplicación de minería de datos; que si bien existen estudios en otras ciudades del Ecuador, aún no se ha evidenciado un estudio dirigido a la problemática de la ciudad de Loja, lo que beneficia a la identificación de sectores críticos más propensos a derivar en accidentes de tránsito, patrones de accidentabilidad vial actualizados, tomas de decisiones ideales que fomentan la seguridad vial en acciones eficientes para reducir el índice de siniestros viales que se produce a diario en nuestro sector.

8. Conclusiones

Una vez culminado el Trabajo de Titulación, se obtiene las siguientes conclusiones:

- Con minería de datos, mediante la implementación de la técnica predictiva de árboles de decisión en base a los registros de accidentes de tránsito del año 2018 al 2020, permitió generar 370 porcentajes de probabilidades resultantes y patrones distintos para cada una de los atributos de accidentabilidad vehicular; esta información guarda relación a los registros de datos que posee la Unidad de Control Operativo de Tránsito del GAD Municipal de Loja, cuyos valores probabilísticos se pusieron a órdenes de la UCOT para la toma de decisiones, siendo este el primer estudio enfocado al análisis de la siniestralidad del cantón, lo que permitió la aplicación de acciones eficientes de seguridad vial con afán de reducir el índice de siniestros viales que se produce a diario en la ciudad.
- La obtención de la base de datos de accidentes de tránsito en la zona urbana del cantón Loja periodo 2028 - 2021, fue parte de la estrategia desarrollada en el presente TT para poder realizar los estudios predictivos, pues la colaboración por parte de las autoridades de la UCOT, permitió analizar los registros de siniestros de tránsito desde el año 2018, aportando con la verificación de variables y relación de clases predominantes e influyentes en la accidentabilidad vehicular de la ciudad, beneficiando exitosamente al proceso de Data Mining.
- Los patrones de predicción obtenidos mediante el desarrollo de las pruebas de minería de datos a través de la herramienta Python en el entorno de Google Colab con el uso del algoritmo CART y la herramienta WEKA con aplicación del algoritmo J48 son eficientes para la creación de modelos predictivos, puestos que entre las dos herramientas se ejecutaron un total de 43 pruebas, de las cuales generaron los porcentajes de instancias clasificadas correctamente, las métricas de precisión y métrica de sensibilidad; sin embargo, se consideró a la herramienta de Python en el entorno Google colab como el más adecuado para el análisis de datos debido a que permite interpretar de mejor manera las predicciones, árboles de decisión y gráficas de probabilidades de las 7 variables estudiadas, así también como sus métricas que son ideales para seleccionar los modelos con mejores niveles de predicción y generar comparativas de nuevos datos de accidentabilidad vehicular.
- La predicción de los modelos a través de la implementación del conjunto de datos de accidentes de tránsito de la zona urbana del cantón Loja pertenecientes al año 2021 en

las herramientas Python con algoritmo CART, presentaron mejores niveles de rendimiento y exactitud mediante la implementación de las siguientes variables hora y parroquia urbana con el 41,62% y 34,59% respectivamente; por el contrario, la aplicación del algoritmo J48 en la herramienta WEKA generaron mayores resultados de instancias clasificadas correctamente para las variables “dia”, “tipologia”, “causas”, “nro_heridos” y “nro_fallecidos” con el 36,21%, 58,37%, 38,10% y 98,64% respectivamente; sin embargo, la interfaz de WEKA a diferencia de Python es más compleja lo que no permitió identificar las probabilidades de los registros de datos, considerando a la información proporcionada por Python la más relevante para el análisis de patrones de accidentes de tránsito.

- Consecuentemente se considera que a medida que los datos se incrementen en los registros, los modelos aplicados a esta investigación serán más precisos y dinámicos, capaces de adaptarse a nuevos datos, lo que permite ejecutar acciones disuasivas para prevenir siniestros de tránsito, por lo que este trabajo de titulación actualmente se encuentra ejecutándose con el registro de datos de accidentabilidad vehicular de las zonas rurales, datos que son una puerta de acceso para mejorar la calidad de información relevante e importante para las instituciones encargadas de movilidad vehicular.

9. Recomendaciones

En base al desarrollo del Trabajo de Titulación, se obtiene las siguientes recomendaciones:

- Realizar un análisis de las bases de datos obtenidas, debido a que algunas variables pueden poseer información errónea, irrelevante o incompleta, lo que implica que los datos no serán fiables para desarrollar el proceso de minería de datos.
- Utilizar la herramienta de data cleaning denominada OpenRefine, puesto que esta herramienta es útil para la revisión de información contenida en las bases de datos, caracterizada por ser una herramienta para trabajar con datos desordenados, lo que permite verificar la calidad de los datos; siendo indispensable en la comparación de variables y clases contenidas en los registros para la correcta ejecución en los modelos predictivos; sin embargo, se pueden aplicar nuevas herramientas de data cleaning que pueden mejorar la calidad de los datos.
- Para el proceso de minería de datos, se recomienda las herramientas de Python y WEKA puesto que son gratuitas; además que, la interfaz gráfica de los resultados permite identificar de manera óptima la estructura del árbol generado, el porcentaje de exactitud del modelo o las instancias clasificadas correctamente, las predicciones realizadas y las matrices de confusión que desarrollan estos datos, según sea el caso; además poseen una serie de métricas que permiten definir de manera más exacta los modelos aplicados; sin embargo, para verificar las probabilidades de los datos se recomienda a Python puesto que su interfaz gráfica es más amigable con el usuario y permite establecer valores más precisos.
- Para mejorar la calidad de los modelos de predicción incorporar más resultados puesto que estos modelos son dinámicos; además, los modelos pueden ser utilizados en la predicción de accidentes de tránsito en otras ciudades en done requieran realizar un estudio de la información referente a siniestros de tránsito.
- Se recomienda el uso del algoritmo J48 en WEKA, este aplica la técnica de clasificación de árboles de decisión; también permite trabajar con datos nominales o categóricos, además presenta una interfaz gráfica de la estructura del del árbol de decisiones y permite exportar el modelo para ser utilizado en otras herramientas de minerías de datos.

10. Bibliografía

[1] Agencia Nacional de Tránsito, "Estadísticas siniestros de tránsito – Agencia Nacional de Tránsito del Ecuador – ANT," 2018. https://www.ant.gob.ec/?page_id=2670 (accessed Aug. 05, 2021).

[2] CONSEJO NACIONAL DE COMPETENCIAS, *Resolución Nro. 006-CNC-2012*. 2012.

[3] A. I. ARÁNGUIZ CASTRO, "ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS Y RURALES USANDO MINERÍA DE DATOS DIFUSA.," 2012. http://opac.pucv.cl/pucv_txt/txt-3500/UCF3892_01.pdf (accessed Aug. 10, 2021).

[4] A. del P. Saavedra, "Estimación del estado del flujo de tráfico mediante preprocesado y minería de datos. Aplicación de Dataset de posiciones GPS de taxis de Porto," Jul. 2016.

[5] H. Al Najada and I. Mahgoub, "Big vehicular traffic Data mining: Towards accident and congestion prevention," *2016 Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2016*, pp. 256–261, Sep. 2016, doi: 10.1109/IWCMC.2016.7577067.

[6] T. Simon Yange, O. Onyekwere, M. Adeiza Rufai, C. Ojochogwu Egbunu, and O. Rehoboth Ogboli, "Determination of the Severity of Motorcycle and Tricycle Accidents in Nigeria," *Adv. Appl. Sci.*, vol. 5, no. 2, p. 41, 2020, doi: 10.11648/J.AAS.20200502.14.

[7] A. Ali Endalkachew Mr, "Predicting Factors Contributing to Road Trafic Accident and Implying Driver's Driving Behavior in Addis Ababa City," May 2020, Accessed: May 11, 2022. [Online]. Available: https://digitalcommons.kennesaw.edu/acist.

[8] J. Mucyo Nzabambarirwa, "TRAFFIC CRASHES PREDICTION USING MACHINE LEARNING MODELS, CASE STUDY: RWANDA," Sep. 2020, Accessed: May 11, 2022. [Online]. Available: http://dr.ur.ac.rw/bitstream/handle/123456789/1452/Mucyo Nzabambarirwa James.pdf?sequence=1&isAllowed=y.

[9] O. D. Ph. D Castrillón, J. A. Ph. D Giraldo, and S. Ph. D Ruiz Herrera, "Predicción del riesgo de un accidente de tránsito en Colombia por medio del software Weka.," Jul. 2019. http://laccei.org/LACCEI2019-MontegoBay/work_in_progress/WP221.pdf (accessed May 21, 2022).

[10] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," *http://dx.doi.org/10.1080/19439962.2016.1152338*, vol. 9, no. 2, pp. 146–166, Apr.

2016, doi: 10.1080/19439962.2016.1152338.

[11] A. G. Pumares Romero, “Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador,” 2019, Accessed: Apr. 29, 2022. [Online]. Available: http://repositorio.uisrael.edu.ec/bitstream/47000/2299/1/UISRAEL-EC-MASTER - TELEM-378.242-2019-015.pdf.

[12] B. Beltrán Martínez, “MINERÍA DE DATOS,” 2022, Accessed: May 04, 2022. [Online]. Available: http://bbeltran.cs.buap.mx/NotasMD.pdf.

[13] J. Cendejas, M. Acuña, G. Cortes, and G. Bolaños, “El uso de modelos y metodologías de minería de datos para la inteligencia de negocios,” vol. 3, no. 8, pp. 54–63, 2017, Accessed: May 04, 2022. [Online]. Available: www.ecorfan.org/spain.

[14] F. Z. El Mazouri, M. C. Abounaima, and K. Zenkouar, “Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France,” *J. Big Data*, vol. 6, no. 1, pp. 1–30, Dec. 2019, doi: 10.1186/S40537-018-0165-0/TABLES/9.

[15] X. Du, “A Data Mining Methodology for Vehicle Crashworthiness Design - CORE Reader,” 2019. https://core.ac.uk/reader/270037520 (accessed May 11, 2022).

[16] G. Villarino Martínez, “Metodología de minería de datos para el estudio de tablas de siniestralidad vial,” 2015, Accessed: May 11, 2022. [Online]. Available: https://eprints.ucm.es/id/eprint/34870/1/TFM_GuillermoVillarino_Nov_2015.pdf.

[17] H. Camargo and M. Silva, “Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP Two paths in search of patterns through Data Mining: SEMMA and CRISP,” *Rev. Tecnol.-Journal Technol. • Vol.*, vol. 9, no. 1.

[18] N. L. Matsudo, “ÁRBOLES DE DECISIÓN, UNA TÉCNICA DE DATA MINING DESDE UNA PERSPECTIVA INFORMÁTICA Y ESTADÍSTICA,” 2001. http://gestion.dc.uba.ar/media/academic/grade/thesis/matsudo.pdf (accessed May 12, 2022).

[19] J. Imbaquingo, “Reconocimiento de patrones en el tráfico de Quito, para el diagnóstico y predicción de las posibles causas que origina el tráfico vehicular en el sector de las avenidas. Napo y Alpahuasi de la ciudad de Quito,” Sep. 09, 2020. http://repositorio.espe.edu.ec/bitstream/21000/22503/1/T-ESPE-043806.pdf (accessed

May 04, 2022).

[20] G. Mariscal, Ó. Marbán, and C. Fernández, “A survey of data mining and knowledge discovery process models and methodologies,” *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, Jun. 2010, doi: 10.1017/S0269888910000032.

[21] J. M. Molina López and J. García Herrero, “TÉCNICAS DE ANÁLISIS DE DATOS APLICACIONES PRÁCTICAS UTILIZANDO MICROSOFT EXCEL Y WEKA,” 2006. http://matema.ujaen.es/jnavas/web_recursos/archivos/weka master recursos naturales/apuntesAD.pdf (accessed May 16, 2022).

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” 1996, Accessed: May 13, 2022. [Online]. Available: https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

[23] S. P. Ing. Alvarez, “MINERIA DE CALIDAD DE DATOS: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA EVALUACIÓN DE LA CALIDAD DE LOS DATOS,” 2018. https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/25468/1/PIO18.pdf (accessed May 15, 2022).

[24] J. S. Alejandro Domínguez, “IMPLEMENTACIÓN DE UNA APLICACIÓN DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE PERFILES DE USUARIOS Y PATRONES DE CONSULTA DE RECURSOS BIBLIOGRÁFICOS,” 2018, Accessed: May 13, 2022. [Online]. Available: https://www.dspace.espol.edu.ec/retrieve/130439/D-106541.pdf.

[25] H. O. Nigro, S. González Císaro, and D. Xodo, “Ontologías en el Proceso de Descubrimiento de Conocimiento en Bases de Datos.” http://sedici.unlp.edu.ar/bitstream/handle/10915/20407/094.pdf?sequence=1 (accessed May 06, 2022).

[26] M. Ben Ahmed, A. A. Boudhir, D. Santos, M. El Aroussi, and I. R. Karas, “Innovations in Smart Cities Applications Edition 3 : the proceedings of the 4th International Conference on Smart City Applications,” p. 1284, Feb. 2020, Accessed: May 17, 2022. [Online]. Available: https://www.researchgate.net/publication/339032296_Innovations_in_Smart_Cities_Applications_Edition_3.

[27] 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.

[28] C. O. Zambrano Yont, “Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.,” Mar. 2021, Accessed: Aug. 30, 2022. [Online]. Available: http://repositorio.ucsg.edu.ec/bitstream/3317/16543/1/T-UCSG-PRE-ING-CIS-284.pdf.

[29] Y. D. García-Ramírez, M. S. Segarra-Morales, B. A. Zárate-Torres, and M. A. Cobos-Ramon, “Relación entre las restricciones del tránsito vehicular y las tendencias del COVID-19: un caso de estudio ecuatoriano,” *CienciAmérica*, vol. 9, no. 2, p. 176, Jun. 2020, doi: 10.33210/ca.v9i2.308.

[30] Scikit-learn community, “scikit-learn: aprendizaje automático en Python — documentación de scikit-learn - 1.3.0,” 2010. https://scikit-learn.org/stable/ (accessed Jul. 18, 2023).

[31] J. Ramírez, “Impacto de la aplicación de algoritmos de minería de textos en los mensajes del medio social twitter para los eventos del tráfico vehicular de la ciudad de Cuenca,” 2018. https://dspace.uazuay.edu.ec/bitstream/datos/8205/2/13928.pdf (accessed May 04, 2022).

[32] J. Herrera Briones, “Análisis y Predicción de la lesividad en accidentes de tráfico mediante la aplicación de random forest,” Jun. 2021, Accessed: May 04, 2022. [Online]. Available: https://oa.upm.es/67548/1/TFG_JUAN_HERRERA_BRIONES.pdf.

[33] J. Lee, T. Yoon, S. Kwon, and J. Lee, “Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study,” *Appl. Sci. 2020, Vol. 10, Page 129*, vol. 10, no. 1, p. 129, Dec. 2019, doi: 10.3390/APP10010129.

[34] 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.

[35] J. I. S. Torres and E. G. Cardenas, “Análisis y aplicación de algoritmos de minería de

datos," *Rev. Perspect.*, vol. 6, no. 21, pp. 71–88, Dec. 2021, doi: 10.26620/UNIMINUTO.PERSPECTIVAS.6.21.2021.71-88.

[36] M. I. Uvidia Fassler, A. S. Cisneros Barahona, P. M. Méndez Naranjo, and H. M. Villa Yánez, "Minería de datos para la toma de decisiones en la unidad de nivelación y admisión universitaria ecuatoriana - Dialnet," Sep. 2018. https://dialnet.unirioja.es/servlet/articulo?codigo=6836545 (accessed May 20, 2022).

[37] shimelis Deneke, S. DENEKE, A. endalkachew, S. Deneke, and A. Ali, "Predicting Factors Contributing to Road Traffic Accident and Implying Driver's Driving Behavior in Addis Ababa City," *African Conf. Inf. Syst. Technol.*, Jul. 2020, Accessed: Jun. 30, 2023. [Online]. Available: https://digitalcommons.kennesaw.edu/acist/2020/allpapers/1.

[38] P. Taylor *et al.*, "DATA MINING FOR VEHICLE TELEMETRY," 2016, Accessed: May 21, 2022. [Online]. Available: https://www.dcs.warwick.ac.uk/~nathan/resources/Publications/aii-2016.pdf.

[39] L. Y. Chang and H. W. Wang, "Analysis of traffic injury severity: an application of non-parametric classification tree techniques," *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 1019–1027, Sep. 2006, doi: 10.1016/J.AAP.2006.04.009.

[40] P. F. Juaréz, "La importancia de la técnica de la entrevista en la investigación en comunicación y las ciencias sociales. Investigación documental. Ventajas y limitaciones," *Sintaxis*, no. 1, pp. 78–93, Jul. 2018, doi: 10.36105/STX.2018N1.07.

[41] L. D. Zarate Manzaneda, "TÉCNICA KDD PARA LA OBTENCIÓN DE INFORMACIÓN EN EL SERVICIO DE MENSAJERIA CORTA," 2009. https://repositorio.umsa.bo/bitstream/handle/123456789/947/T.1786.pdf?sequence=3&isAllowed=y (accessed Jun. 09, 2022).

11. Anexos

Anexo 1. Entrevista dirigida al jefe de la Unidad de Control Operativo de Tránsito del Cantón Loja.

Anexo 2. Base de datos inicial de accidentes de tránsito en el Cantón Loja periodo 2018 – 2021.

Anexo 3. Diccionario de datos de la BD inicial.

Anexo 4. Registro de variables no consideradas.

Anexo 5. Limpieza de los registros de bases de datos.

Anexo 6. Conversión de registros de datos de texto a numéricos.

Anexo 7. Casos para selección de rangos de variables.

Anexo 8. Pruebas desarrolladas mediante aplicación de entorno Google Colab.

Anexo 9. Pruebas desarrolladas mediante aplicación de entorno WEKA.

Anexo 10. Resultados de las pruebas realizadas por los modelos en Python.

Anexo 11. Resultados de los modelos predictivos en Google Colab del año 2021.

Anexo 12. Resultados de los modelos predictivos en WEKA del año 2021.

Anexo 13. Ensayo argumentativo del proyecto de trabajo de titulación.

Anexo 14. Plano de la ciudad de Loja.

Anexo 15. Solicitud de base de datos de UCOT.

Anexo 16. Anteproyecto de minería de datos en la accidentabilidad vehicular en la zona urbana del cantón Loja.

Anexo 17. Base de datos finales para minería de datos.

Anexo 18. Pruebas desarrolladas mediante aplicación de entorno Google Colab.

Anexo 19. Certificado de traducción del resumen.

CERTIFICADO DE TRADUCCIÓN

La Srta. María Paulina Patiño Macas, identificada con cédula de ciudadanía número 1104533797, Licenciada en Ciencias de la Educación mención idioma Inglés.

CERTIFICO:

Que el texto traducido al idioma inglés que compone el **Resumen** del Trabajo de Titulación denominado **"MINERÍA DE DATOS EN LA ACCIDENTABILIDAD VEHICULAR EN LA ZONA URBANA DEL CANTÓN LOJA / DATA MINING ON VEHICLE ACCIDENT RATES IN THE URBAN AREA OF LOJA CANTON"** correspondiente al Sr. Patricio Bolívar Benítez Lanche, con cédula de ciudadanía número 1105665044, fue realizado y verificado bajo mi supervisión.

Eso es todo en cuanto puedo indicar en honor a la verdad, facultando al interesado a hacer uso del presente documento para los fines que se crean pertinentes.

Loja, 17 de marzo del 2023

Lic. María Paulina Patiño Macas
Senescyt Registration Number: 1031-2022-2563323
Número de registro Senescyt: 1031-2022-2563323