



Universidad
Nacional
de Loja



Carrera de Ingeniería en
Sistemas/Computación.

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

CARRERA DE INGENIERÍA EN SISTEMAS

Minería de datos para determinar los factores más influyentes en la ocurrencia de Siniestros de Tránsito en Ecuador en el año 2020

Línea de investigación: Sistemas Inteligentes

**TESIS DE GRADO PREVIA A LA
OBTENCIÓN DEL TÍTULO DE
INGENIERO EN SISTEMAS.**

Autor:

- Yulissa Stefania Torres Quezada.

Director:

- Ing. Oscar Miguel Cumbicus Pineda, Mg.Sc.

LOJA - ECUADOR

2022

CERTIFICACIÓN

Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.
DIRECTOR DEL TRABAJO DE TITULACIÓN

CERTIFICA:

Que la egresada **Yulissa Stefania Torres Quezada** autora del presente trabajo de titulación, cuyo tema versa sobre “**MINERÍA DE DATOS PARA DETERMINAR LOS FACTORES MÁS INFLUYENTES EN LA OCURRENCIA DE SINIESTROS DE TRÁNSITO EN ECUADOR EN EL AÑO 2020**”, ha sido dirigido, orientado, discutido bajo mi asesoramiento y ha sido culminado al 100%, reúne a satisfacción los requisitos exigidos en una investigación de este nivel por lo cual autorizo su presentación y sustentación.

Loja, 17 de septiembre del 2021

Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.
DIRECTOR DEL TRABAJO DE TITULACIÓN

AUTORÍA

Yo **Yulissa Stefania Torres Quezada**, 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 Institucional - Biblioteca Virtual.

Firma:

Cédula: 1106035536

Fecha: Loja, 13 de enero del 2022

CARTA DE AUTORIZACIÓN POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO

Yo **YULISSA STEFANIA TORRES QUEZADA**, declaro ser autor del trabajo de titulación que versa: "**MINERÍA DE DATOS PARA DETERMINAR LOS FACTORES MÁS INFLUYENTES EN LA OCURRENCIA DE SINIESTROS DE TRÁNSITO EN ECUADOR EN EL AÑO 2020**", como requisito para optar al grado de: **INGENIERO EN SISTEMAS**; autorizo al Sistema Bibliotecario de la Universidad Nacional de Loja para que, con fines académicos, muestre al mundo la producción intelectual de la Universidad, a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional: Los usuarios pueden consultar el contenido de este trabajo en el (RDI), en las redes de información del país y del exterior, con los cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia de la tesis que realice un tercero.

Para constancia de esta autorización, en la ciudad de Loja, a los 13 días del mes de enero del 2022.

Firma:

Autor: Yulissa Stefania Torres Quezada

Cédula: 1106035536

Dirección: Loja – Macará, Luz de América (Avd. Panamericana y Reinaldo Celi)

Correo Electrónico: yulissa.torres@unl.edu.ec

Celular: 0969899025

Teléfono: 072695122

DATOS COMPLEMENTARIOS

Director de Tesis: Ing. Oscar Miguel Cumbicus Pineda, Mg. Sc.

Tribunal de Grado: Ing. Luis Antonio Chamba Eras, PhD.

Ing. Roberth Gustavo Figueroa Díaz, Mg. Sc.

Ing. Genoveva Jackelinne Suing Albito, Mg. Sc.

AGRADECIMIENTO

Agradezco primeramente a Dios por iluminarme cada día para seguir adelante, por acompañarme en el camino y brindarme la sabiduría necesaria para haber llegado hasta este momento tan importante de mi formación profesional.

De corazón agradezco a mis padres, Luis Torres y Yolanda Quezada, por haberme forjado como la persona que soy en la actualidad, por brindarme incondicionalmente todo su amor, respeto, cariño, comprensión, apoyo y motivación durante todos estos años de estudio, por el enorme sacrificio que han realizado para que yo pueda alcanzar esta meta y lo único que puedo decir es un enorme gracias, de todo corazón, los amo. Un agradecimiento para toda mi familia, en especial a mis hermanas Katherine y Verónica, que desde pequeña me impulsaron a cumplir todo lo que me proponga y a nunca darme por vencida.

También a la Universidad Nacional de Loja por darme la oportunidad de ser una profesional, a los distinguidos docentes de la Carrera de Ingeniería en Sistemas, quienes, a lo largo de estos periodos de estudio, con su cordialidad y entusiasmo depositaron en mí, sus valiosos conocimientos y de manera especial al Ing. Oscar Cumbicus Pineda, en su calidad de Director, por haberme guiado para culminar con éxito este Trabajo de Titulación, gracias por su tiempo y paciencia.

Finalmente, un agradecimiento a mis compañeros y amigos, en especial a Alex, Raquel y Joel, por todos los buenos momentos compartidos a lo largo de estos años, por las risas, las motivaciones, la bondad de sus corazones, por su ayuda y por siempre estar ahí presentes conmigo, enserio muchas gracias.

Yulissa Stefania Torres Quezada.

DEDICATORIA

Dedico este trabajo a Dios por permitirme tener vida y darme salud para poder alcanzar uno más de mis propósitos, él es ser ingeniera, por bendecirme sabiamente, en todo momento y en todo lugar de mi vida.

De igual forma, dedico esta tesis de manera especial a mis padres, quienes han sido la fortaleza, razón de mi vida y motivo de superación, por siempre creer en mí y por darme las fuerzas para seguir adelante; muchos de mis logros se los debo a ustedes entre los que se incluye este.

Dedico también esta tesis a todas las personas que han hecho que culmine mi carrera con éxito, a mi familia, mis hermanas, amigos y docentes, que de una u otra manera me han contribuido para alcanzar mis objetivos académicos, brindándome sus conocimientos, consejos y palabras de aliento día a día para culminar esta meta y logro profesional.

Yulissa Stefania Torres Quezada.

ÍNDICE DE CONTENIDOS

ÍNDICE GENERAL

CERTIFICACIÓN.....	II
AUTORÍA	III
CARTA DE AUTORIZACIÓN POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.....	IV
AGRADECIMIENTO.....	V
DEDICATORIA.....	VI
ÍNDICE DE CONTENIDOS.....	VII
ÍNDICE GENERAL	VII
ÍNDICE DE FIGURAS.....	XI
ÍNDICE DE TABLAS.....	XIV
1. TÍTULO	1
2. RESUMEN.....	2
SUMMARY	3
3. INTRODUCCIÓN.....	4
4. REVISIÓN DE LITERATURA.....	6
4.1. Conceptos preliminares	6
Minería de Datos	6
Descubrimiento de Conocimiento en Bases de Datos (KDD).....	6
Análisis Predictivo	7
Algoritmos de clasificación.....	8
Depuración de datos.....	8
Siniestro de Tránsito.....	8
Redes Neuronales Artificiales (RNA)	9
Redes Bayesianas (RB)	9
Árboles de Decisión (AD)	10

Reglas de Asociación (RA)	10
OpenRefine	10
R Studio.....	11
IBM SPSS Statistics	11
Weka	11
4.2. Trabajos relacionados	11
5. MATERIALES Y MÉTODOS	18
5.1. Tipo de investigación.....	18
5.2. Métodos de investigación	18
Método inductivo	18
Método deductivo	18
Método científico	19
Método sistémico.....	19
5.3. Técnicas de investigación.....	19
Observación	19
Entrevista	20
5.4. Metodología.....	20
Fase I: Búsqueda de información	21
Fase II: Obtención de datos.....	21
Fase III: Depuración de la base de datos.....	21
Fase IV: Aplicación de técnicas de minería de datos	21
Fase V: Interpretación y presentación de resultados	22
6. RESULTADOS.....	23
6.1. Objetivo 1: Identificar los repositorios donde se encuentra almacenada la información sobre los siniestros de tránsito en Ecuador en el año 2020.	23
Fase I: Búsqueda de información	23
Tarea 1: Establecer los lineamientos, para la búsqueda de la información relevante sobre los siniestros de tránsito en Ecuador en el año 2020.....	23
Fase II: Obtención de datos	24
Tarea 2: Obtener la base de datos con la información relevante sobre los	

siniestros de tránsito en Ecuador en el año 2020.	24
Fase III: Depuración de la base de datos.....	25
Tarea 3: Depurar la base de datos obtenida	25
6.2. Objetivo 2: Aplicar técnicas de minería de datos a la base de datos obtenida para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.....	38
Fase IV: Aplicación de técnicas de minería de datos	38
Tarea 1: Extraer la información relevante para transformarla	38
Tarea 2: Aplicar las técnicas de árboles de decisión, redes neuronales artificiales y redes bayesianas para el análisis de la información almacenada.	
43	
6.3. Objetivo 3: Interpretar y presentar los resultados obtenidos sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.....	69
Fase V: Interpretación y presentación de resultados	69
Tarea 1: Evaluar e identificar las mejores técnicas para la obtención de los resultados.....	69
Tarea 2: Presentar e Interpretar los resultados obtenidos.....	73
Tarea 3: Elaborar el documento final	99
7. DISCUSIÓN.....	100
7.1. Desarrollo de la propuesta alternativa	100
Objetivo 1: Identificar los repositorios donde se encuentra almacenada la información sobre los siniestros de tránsito en Ecuador en el año 2020. 100	
Objetivo 2: Aplicar técnicas de minería de datos a la base de datos obtenida para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.....	101
Objetivo 3: Interpretar y presentar los resultados obtenidos sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.	102
7.2. Valoración técnica, económica, ambiental y social	103
Valoración técnica	103
Valoración económica	104
Valoración ambiental	105

Valoración social	105
8. CONCLUSIONES.....	107
9. RECOMENDACIONES.....	109
10. BIBLIOGRAFÍA.....	111
11. ANEXOS	116
Anexo 1	116
Anexo 2	120
Anexo 3	121
Anexo 4	127
Anexo 5	133
Anexo 6	140
Anexo 7	141
Anexo 8	148
Anexo 9	150

ÍNDICE DE FIGURAS

Fig. 1 Fases de la metodología KDD	7
Fig. 2 Metodología aplicada en el TT	20
Fig. 3 Número total de registros de las variables Tipo de Vehículo 1 – Tipo de Vehículo 10	29
Fig. 4 Valores de los registros totales	30
Fig. 5 Conjunto de datos obtenido después de la evaluación.....	31
Fig. 6 Cargar conjunto de datos en OpenRefine	33
Fig. 7 Cambio de nombre de las variables	34
Fig. 8 Conversión de texto a mayúsculas.....	34
Fig. 9 Estandarización de las tildes.....	35
Fig. 10 Eliminación de la letra “Ñ”	35
Fig. 11 Depuración de la base de datos.....	37
Fig. 12 Número de registros por la sentencia de depuración	38
Fig. 13 Transformación de los registros de la variable “HORA”	42
Fig. 14 Transformación de los registros de la variable “CAUSA_PROBABLE”	42
Fig. 15 Diagrama de flujo para la aplicación de las técnicas de minería de datos	44
Fig. 16 Interfaz principal de SPSS Statistics	46
Fig. 17 Interfaz de SPSS Statistics para subir archivos a procesar.....	46
Fig. 18 Configurar la lectura del archivo CSV en SPSS Statistics	47
Fig. 19 Conjunto de datos cargado en SPSS Statistics.....	47
Fig. 20 Interfaz principal de Weka.....	48
Fig. 21 Interfaz de Weka para subir archivos a procesar	48
Fig. 22 Conjunto de datos cargado en Weka	49
Fig. 23 Crear árboles de decisión en SPSS Statistics	50
Fig. 24 Configuración de variables independientes y dependiente en CHAID	50
Fig. 25 Configuración de criterios del algoritmo CHAID	51
Fig. 26 Configuración de variables independientes y dependiente en CHAID Exhaustivo	52
Fig. 27 Configuración de criterios del algoritmo CHAID Exhaustivo	52
Fig. 28 Configuración de variables independientes y dependiente en CRT.....	53
Fig. 29 Configuración de criterios del algoritmo CRT	54
Fig. 30 Crear RN Perceptrón Multicapa en SPSS Statistics.....	54
Fig. 31 Configuración de los factores y de la variable dependiente en RN Perceptrón Multicapa	55
Fig. 32 Configuración de las particiones en la RN Perceptrón Multicapa	55

Fig. 33 Crear RN Función de Base Radial en SPSS Statistics	56
Fig. 34 Configuración de los factores y de la variable dependiente en RN Función de Base Radial	57
Fig. 35 Configuración de las particiones en la RN Función de Base Radial	57
Fig. 36 Configuración de la RB Naive Bayes	58
Fig. 37 Configuración de la RB BayesNet.....	59
Fig. 38 Número de registros por cada subconjunto generado.....	61
Fig. 39 Clasificación global correcta de instancias de cada algoritmo	70
Fig. 40 Mejores resultados de clasificación correcta y precisión de acuerdo a cada algoritmo.....	73
Fig. 41 Nodo raíz del algoritmo CHAID Exhaustivo.....	74
Fig. 42 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral.....	76
Fig. 43 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Estrellamientos	76
Fig. 44 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Atropellos.....	77
Fig. 45 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Posterior.....	77
Fig. 46 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Arrollamientos.....	78
Fig. 47 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral.....	79
Fig. 48 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Frontal.....	79
Fig. 49 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Rozamientos	80
Fig. 50 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Otros.....	81
Fig. 51 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Caída de Pasajero.....	82
Fig. 52 Resultados de las probabilidades de ocurrencia de las causas probables para cada clase de siniestro de tránsito	93
Fig. 53 Categorización en factores a las causas probables de los siniestros de tránsito	95
Fig. 54 Resultados de las probabilidades de ocurrencia en relación a los tipos de factores	98
Fig. 55 Factores influyentes para la ocurrencia de siniestros de tránsito en Ecuador..	98
Fig. 56 Conjunto de datos obtenido, aún sin procesar	120
Fig. 57 Índice de Siniestralidad por Provincia	133
Fig. 58 Índice de Siniestralidad por Día	134

Fig. 59 Siniestros por Hora	134
Fig. 60 Índice de Siniestralidad en Vehículos.....	135
Fig. 61 Siniestros por Tipo de Servicio del Vehículo	135
Fig. 62 Siniestros por Zona	136
Fig. 63 Índice de Siniestralidad por Participante	136
Fig. 64 Siniestros por Sexo	137
Fig. 65 Principales Causas Probables de Siniestros	138
Fig. 66 Clases de Siniestros	138
Fig. 67 Condición del Participante del Siniestro	139

ÍNDICE DE TABLAS

TABLA I	12
TABLA II	24
TABLA III	25
TABLA IV	30
TABLA V	32
TABLA VI	33
TABLA VII	36
TABLA VIII	38
TABLA IX	39
TABLA X	39
TABLA XI	45
TABLA XII	45
TABLA XIII	60
TABLA XIV	61
TABLA XV	62
TABLA XVI	62
TABLA XVII	63
TABLA XVIII	64
TABLA XIX	64
TABLA XX	64
TABLA XXI	65
TABLA XXII	65
TABLA XXIII	66
TABLA XXIV	66
TABLA XXV	67
TABLA XXVI	67
TABLA XXVII	68
TABLA XXVIII	68
TABLA XXIX	71
TABLA XXX	72
TABLA XXXI	72
TABLA XXXII	74
TABLA XXXIII	75
TABLA XXXIV	80
TABLA XXXV	82

TABLA XXXVI.....	84
TABLA XXXVII.....	85
TABLA XXXVIII.....	86
TABLA XXXIX.....	86
TABLA XL.....	87
TABLA XLI.....	87
TABLA XLII.....	88
TABLA XLIII.....	88
TABLA XLIV	89
TABLA XLV	89
TABLA XLVI	90
TABLA XLVII	92
TABLA XLVIII	94
TABLA XLIX	96
TABLA L	104
TABLA LI	104
TABLA LII	105
TABLA LIII	105
TABLA LIV.....	121
TABLA LV.....	127
TABLA LVI.....	140
TABLA LVII.....	141
TABLA LVIII.....	142
TABLA LIX.....	143
TABLA LX.....	144
TABLA LXI.....	145
TABLA LXII.....	146
TABLA LXIII.....	147
TABLA LXIV	148

1. TÍTULO

**Minería de datos para determinar los factores
más influyentes en la ocurrencia de Siniestros de
Tránsito en Ecuador en el año 2020.**

2. RESUMEN

La ocurrencia de siniestros de tránsito representa un problema de salud pública a nivel nacional y regional, ocasionando pérdidas humanas, además de que cada día va en aumento a nivel mundial, es por ello que resulta oportuno plantear un estudio que permita determinar cuáles son los factores que ocasionan la ocurrencia de siniestros de tránsito. El objetivo del presente Trabajo de Titulación (TT) es aplicar minería de datos para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador en el año 2020, esto se llevó a cabo mediante cinco fases de la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) constituida por: búsqueda de información, obtención de datos, depuración de la base de datos, aplicación de técnicas de minería de datos e interpretación y presentación de resultados, por lo cual, por medio del establecimiento de lineamientos para la búsqueda de información se obtuvo el conjunto de datos recolectado por la Agencia Nacional de Tránsito (ANT), en el que consta la recopilación de los partes policiales, diseñados y aprobados por cada uno de los entes de control, bajo los parámetros técnicos establecidos por la misma institución, que reposa en su sitio web oficial. Utilizando las herramientas OpenRefine y RStudio se realizó la depuración del conjunto de datos obtenido, evaluando y determinando las variables más útiles y relevantes para el objeto de estudio.

Las herramientas de software empleadas para la aplicación de los algoritmos de minería de datos fueron SPSS Statistics y Weka. Se aplicaron siete técnicas predictivas de minería de datos: CHAID, CHAID Exhaustivo, CRT, Perceptrón Multicapa, Función de Base Radial, Naive Bayes y BayesNet. La evaluación de estos algoritmos se realizó comparando los resultados obtenidos por cada uno, en relación a métricas de rendimiento con respecto a porcentajes de clasificación correcta de las instancias y de precisión. El algoritmo CHAID Exhaustivo fue el que obtuvo los mejores resultados con un porcentaje de clasificación correcta del 58,38% y 44,60% de precisión, con el cual se identificó los patrones más importantes en los datos y se evaluó las posibles asociaciones entre las variables recogidas. Finalmente, se determinó que el factor humano es el factor más influyente con una probabilidad de ocurrencia del 69,64%.

Palabras clave: Minería de datos, metodología KDD, Árboles de Decisión, Redes Neuronales, Redes Bayesianas, Siniestros de Tránsito en Ecuador.

SUMMARY

The occurrence of traffic accidents represents a public health problem at national and regional level, causing human losses, in addition to the fact that every day is increasing worldwide, which is why it is appropriate to propose a study to determine what are the factors that cause the occurrence of traffic accidents. The objective of this thesis is to apply data mining to determine the most influential factors in the occurrence of traffic accidents in Ecuador in the year 2020, this was carried out through five phases of the methodology of Knowledge Discovery in Databases (KDD) consisting of: search for information, data collection, database cleaning, application of data mining techniques and interpretation and presentation of results, whereby, through the establishment of guidelines for the search for information, the set of data collected by the National Transit Agency (ANT) was obtained, which includes the collection of police reports, designed and approved by each of the control entities, under the technical parameters established by the same institution, which is available on its official website. Using the OpenRefine and RStudio tools, the obtained data set was debugged, evaluating and determining the most useful and relevant variables for the object of study.

The software tools used for the application of the data mining algorithms were SPSS Statistics and Weka. Seven predictive data mining techniques were applied: CHAID, Exhaustive CHAID, CRT, Multilayer Perceptron, Radial Basis Function, Naive Bayes and BayesNet. The evaluation of these algorithms was performed by comparing the results obtained by each one, in relation to performance metrics with respect to percentages of correct classification of instances and accuracy. The CHAID Exhaustive algorithm was the one that obtained the best results with a percentage of correct classification of 58.38% and 44.60% accuracy, with which the most important patterns in the data were identified and the possible associations between the variables collected were evaluated. Finally, the human factor was determined to be the most influential factor with a probability of occurrence of 69.64%.

Keywords: Data mining, KDD methodology, Decision Trees, Neural Networks, Bayesian Networks, Traffic Accidents in Ecuador.

3. INTRODUCCIÓN

De acuerdo al Informe sobre la Situación Mundial de la Seguridad Vial del año 2018, realizado por la Organización Mundial de la Salud (OMS), los siniestros de tránsito se han convertido en una de las principales causas de muertes violentas de la población, y a su vez convirtiéndose en un problema de salud pública [1]. Así mismo, estos se han catalogado como un problema social debido al daño que produce en las personas, las familias y la comunidad, ya que al ocurrir este, el impacto que genera es devastador y lleva mucho tiempo el superarlo, afectando así fundamentalmente la calidad de vida de las personas involucradas. Esto ha llevado a que cada país reconozca el costo económico y social que representa estar afectados por este enorme fenómeno. En el caso de Ecuador, según datos del año 2019 presentados en el Anuario de Estadísticas de Transporte (ANET) ocurren once siniestros por cada mil vehículos que circulan en el país, esto lo ubica en el segundo país en Latinoamérica con mayor índice de ocurrencia de siniestros de tránsito [2].

En la actualidad con los avances en el campo de la inteligencia artificial, ha sido posible la explotación de datos generados por distintas entidades públicas, permitiendo el manejo de grandes volúmenes de información disponibles en bases de datos, para que de esta manera sea posible disminuir el tiempo de análisis e interpretación de los datos, además de obtener información que no es visualizada a simple vista, este es el caso de los datos recolectados por la ANT a través de la recopilación de los partes policiales recabados por cada uno de sus entes de control, este conjunto de datos tiene una gran cantidad de información que fue minada para determinar información útil que se encontraba de manera implícita, debido a que cada siniestro es el resultado de una cadena de eventos que es, en su totalidad único, pero algunos factores son comunes a varias circunstancias del accidente, y la identificación de estos factores y sus interdependencias se llevó a cabo mediante el uso de técnicas que brinda la inteligencia artificial [3]. Por tal motivo, se planteó aplicar técnicas de minería de datos a la información recolectada por la ANT con el propósito de determinar cuáles son los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

En el contexto del presente Trabajo de Titulación, se estableció tres objetivos específicos, la identificación de repositorios donde se encontraba almacenada la información sobre siniestros de tránsito en Ecuador en el año 2020, la aplicación de técnicas de minería de datos a la base de datos obtenida para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador, y finalmente se

planteó la interpretación y presentación de los resultados obtenidos; con el fin de cumplir el objetivo general: Aplicar Minería de Datos para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador en el año 2020.

El presente Trabajo de Titulación está conformado por varias secciones, entre ellas la de Revisión de Literatura la cual abarca temáticas con respecto a la parte técnica del estudio como minería de datos, algoritmos de clasificación, análisis predictivo, técnicas de minería de datos y todo lo necesario para realizar cada una de las fases de la metodología, en cuanto a la parte de la problemática se encuentra la definición de siniestro de tránsito. En la misma sección se habla de los trabajos relacionados tomados como referencia para el desarrollo del presente trabajo. Más adelante, en la sección de Materiales y Métodos se definió el tipo de investigación, los métodos de investigación aplicados, las técnicas utilizadas para la recolección de información y se estableció la metodología para el desarrollo del presente trabajo, en este caso la metodología KDD. Posteriormente, se encuentra la sección de Resultados estructurada a partir de los objetivos específicos, en los cuales se establecieron tareas para su cumplimiento. Para el primero objetivo se estableció los lineamientos para la búsqueda y obtención de la información relevante sobre los siniestros de tránsito en Ecuador en el año 2020 y además se realizó la depuración del conjunto de datos obtenido, el cual consistió en evaluarlo para establecer las variables útiles y relevantes para la ejecución del trabajo; y también en la limpieza de información innecesaria y con inconsistencias que pueda afectar a los resultados, en el segundo objetivo se extrajo la información, se la transformó y cargó en el software para aplicar los algoritmos de árboles de decisión, redes neuronales y redes bayesianas. Finalmente, en el tercer objetivo se realizó una evaluación e identificación de las mejores técnicas para la obtención de los resultados esto con el fin de presentar a través de gráficos la interpretación de los mismos. La siguiente, es la sección de Discusión en donde se explica y contrasta los resultados que se obtuvieron con estudios relaciones al presente TT y como aportan al cumplimiento de las fases planteadas para cada objetivo específico, así como también se especifica la valoración técnica, económica ambiental y social de este trabajo, en la sección Conclusiones se plasmaron las deducciones alcanzadas a partir de las experiencias obtenidas durante el proceso de cumplimiento de los objetivos. Por último, en la sección de recomendaciones se detallan sugerencias para un mejor desarrollo de trabajos similares al presente trabajo y para trabajos futuros.

4. REVISIÓN DE LITERATURA

En esta sección se presentan conceptos relacionados con la temática, los cuales sustentan el presente Trabajo de Titulación (TT), dicha información ha sido recopilada a través de un proceso de revisión bibliográfica, agregando a lo anterior se muestra una tabla con todos los estudios desarrollados dentro de la línea de investigación, que han sido seleccionados durante el proceso de revisión bibliográfica.

4.1. Conceptos preliminares

Minería de Datos

Es un área de estudio que posibilita la exploración de los datos extrayendo información que no es detectada a simple vista, a través del descubrimiento y cuantificación de relaciones predictivas en los datos [4], es decir mediante la combinación de técnicas semiautomáticas de Inteligencia Artificial, análisis estadístico, bases de datos y visualización gráfica se obtiene información que no está representada explícitamente en grandes cantidades de datos [5].

Esta herramienta también es conocida como el Descubrimiento de Conocimiento de Bases de datos (Knowledge Discovery in Databases, o KDD), debido a que es un conjunto de técnicas y tecnologías, que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos [5], que de otra manera permanecería oculta, con el fin de obtener información útil a través de programas de búsqueda e identificación de patrones y la descripción de las tendencias, desviaciones y correlaciones entre los datos, comportamientos atípicos y trayectorias ocultas, facilitando así la toma de decisiones [4][6].

Además, haciendo uso de diferentes algoritmos a partir de datos, esta resuelve problemas de agrupamiento automático, clasificación, predicción, asociación y detección de patrones secuenciales proporcionando nuevos conocimientos [7].

Descubrimiento de Conocimiento en Bases de Datos (KDD)

El proceso de KDD persigue la extracción automatizada de conocimiento no trivial, implícito, antes desconocido y potencialmente útil a partir de grandes volúmenes de datos [8], para la identificación de patrones a partir de datos, válidos, novedosos, potencialmente útiles, y en última instancia comprensibles [9].

La implementación del proceso KDD tiene un procedimiento complejo que implica la consecución de cinco fases, esquematizados en la Fig. 1.

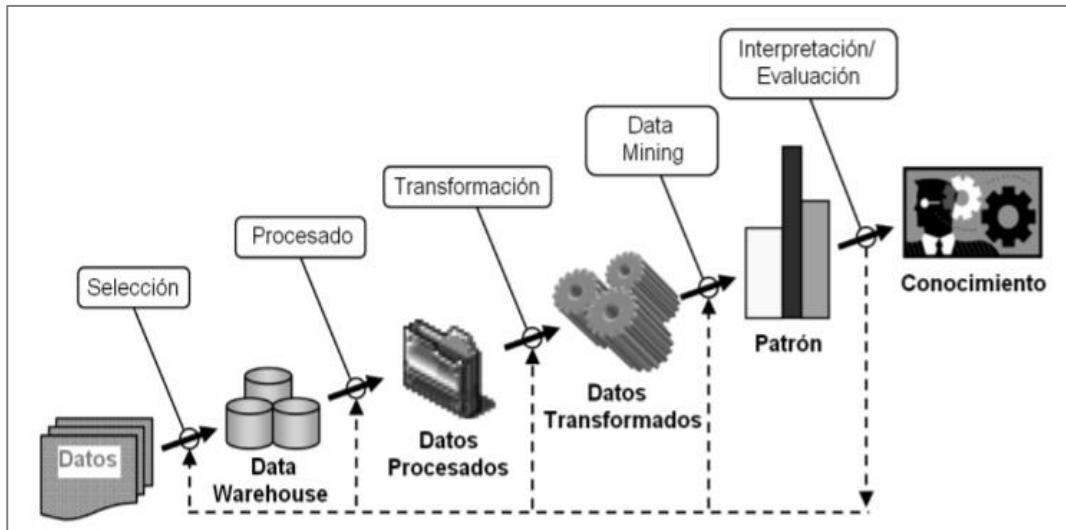


Fig. 1 Fases de la metodología KDD [10]

En la fase de selección se determinan las fuentes de datos y el tipo de información a utilizar, creando un conjunto de datos de destino, o centrándose en un subconjunto de variables o muestras de datos, en el que se va a realizar el descubrimiento; en el preprocessamiento se prepara y se depura los datos de destino, desde las distintas fuentes de datos en una forma manejable, para obtener datos consistentes, necesarios para las fases posteriores; la fase de transformación consiste en el tratamiento y trasformación de los datos utilizando métodos de reducción de la dimensionalidad, realizando operaciones de agregación o normalización, consolidando los datos con una estructura apropiada; la minería de datos es la fase en la que se construye el modelo a partir de búsqueda y extracción de patrones de interés previamente desconocidos, nuevos, potencialmente útiles y comprensibles, en una forma representacional particular, dependiendo del objetivo de la gestión de datos; y por último en la fase de interpretación y evaluación se identifican los patrones que tienen un alto grado de relevancia, además de interpretarlos y evaluarlos basándose en algunas medidas [9], [11], [12].

Análisis Predictivo

De acuerdo a Timón y Fontes [13], el análisis predictivo es un área de la minería de datos, la cual se basa en la extracción de información existente en los datos y su implementación para predecir tendencias y patrones de comportamiento, logrando aplicarse sobre cualquier evento desconocido, así sea en el pasado, presente o futuro, es decir este se basa en la identificación de relaciones entre variables en eventos pasados, para después explotar dichas relaciones y predecir posibles resultados en futuras situaciones. Para llevar a cabo el análisis predictivo es

imprescindible disponer de una considerable cantidad de datos, para lograr establecer patrones de comportamiento y de esta forma inducir conocimiento.

Algoritmos de clasificación

Uno de los conceptos más fundamentales relacionado con las técnicas de minería de datos, es el concepto de aprendizaje automático, cuya finalidad es inducir conocimiento a partir de datos [14]. Las técnicas de aprendizaje automático se utilizan constantemente para resolver problemas de clasificación en diversas áreas de manera rápida y precisa, proporcionando una significativa ventaja para los objetos de estudio [15]. Estas técnicas se clasifican en supervisadas y no supervisadas [16], donde la primera es más precisa debido a que los clasificadores trabajan con datos ya entrenados, es decir tienen ya una clase o etiqueta aginada de forma correcta. Los algoritmos no supervisados muestran un menor desempeño al no presentar en sus datos una etiqueta, pero minimizan el trabajo de clasificación, trabajando por medio de palabras semilla y calculando la orientación semántica de las frases [17].

Depuración de datos

También conocido como limpieza de datos, el cual consiste en un proceso de detección y corrección de una base de datos, “utilizada primordialmente cuando se tienen datos incorrectos, incompletos, inexactos o irrelevantes. Con esta técnica es posibles corregir, completar, cambiar o eliminar esos datos” [18], mediante el uso de herramientas que facilitan la depuración de la información [6].

Siniestro de Tránsito

Un siniestro de tránsito o accidente de tránsito en términos generales se refiere a todo suceso eventual, imprevisto o acción involuntaria que se presenta inesperadamente [19], que como efecto de una o más causas y con independencia del grado de estas, ocurre en vías o lugares destinados al uso público o privado, ocasionando pérdidas prematuras de vidas humanas, individuos con lesiones de diversa gravedad o naturaleza, con secuelas físicas o psicológicas, perjuicios materiales y daños a terceros en vehículos, vías o infraestructura, con la participación de los usuarios de la vía y/o entorno, vehículo o más [2].

Redes Neuronales Artificiales (RNA)

La Red Neuronal Artificial es un modelo estadístico clásico cuya estructura y operación se asemeja mucho a las redes neuronales biológicas del cerebro humano [9]. Una RNA puede verse como un grafo dirigido formado por un conjunto interconectado de elementos simples de procesamiento, unidades o nodos. El modelo de una RNA es una técnica poderosa con su capacidad de procesamiento almacenada en las fuerzas de conexión entre las unidades, o pesos, obtenidos por un proceso de aprendizaje a partir de un conjunto de patrones de entrenamiento para representar la compleja relación entre las variables de entrada y la variable objetivo [9], [3].

La RNA proporciona un mejor coeficiente de correlación entre las variables reales y las predichas, de tal forma que el modelo de RNA puede detectar las relaciones entre las variables de entrada y generar una salida basada en los patrones observados inherentes a los datos, todo esto debido a las unidades simples llamadas neuronas que la componen, las cuales tienen asociada una función matemática o de transferencia que genera la salida de la neurona a partir de las señales de entrada, esta función tiene como entrada la suma de todas las señales de entrada por el peso asociado a la conexión de entrada de la señal, de esta manera, la función de transferencia es la relación entre la señal de salida y de entrada [9], [20].

Redes Bayesianas (RB)

La Red Bayesiana es un modelo probabilístico que relaciona un conjunto de variables aleatorias mediante un grafo dirigido acíclico en el que a través de cada uno de los nodos se representan dichas variables aleatorias y cada arco simboliza las relaciones de dependencia probabilística directa que existe entre ellas, lo que permite conseguir soluciones a problemas de decisión en casos de incertidumbre [3]. De esta forma, la estructura de red es un modo compacto de representar el conocimiento que aporta información sobre las dependencias probabilísticas entre las variables y sobre las independencias condicionales ya sea de una o varias variables dadas a otras variables [9]. Es decir, las RB “modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; esto es, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas” [21]. Estos modelos bayesianos tienen diferentes aplicaciones para

diagnóstico, clasificación y decisión que brinde información fundamental referente a cómo se relacionan las variables, las cuales pueden ser interpretadas como relaciones de causa efecto [3].

Árboles de Decisión (AD)

Un Árbol de Decisión es un conjunto de condiciones organizadas en una estructura jerárquica de tal forma que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen a partir de la raíz del árbol hasta alguna de sus hojas [3]. La estructura de los AD está representada mediante grafos direccionalados formados por variables de entradas simbolizadas en nodos, también por ramas que están asociadas a los valores de la variable que forma el nodo y además por hojas, nodos hoja o nodos terminales, estos representan los valores de la variable de salida [9].

Los sistemas de aprendizaje basados en AD son tal vez el método más fácil de utilizar y aprender, debido a que destacan por su sencillez y transparencia, y se explican por sí mismos, además, su representación gráfica como una estructura jerárquica hace que sean fácilmente comprensibles, y por consiguiente, más fáciles de interpretar que otras técnicas [9], [3].

Reglas de Asociación (RA)

Según Maldonado [9], las RA constituyen un mecanismo de representación del conocimiento muy simple y útil para caracterizar las regularidades que se pueden encontrar en grandes bases de datos, estas reglas se utilizan cuando el resultado de interés no es conocido y el sistema debe aprender directamente de los datos disponibles.

OpenRefine

Es una potente herramienta desarrollada por Google, es de código abierto, escrita en Java y puede descargarse gratuitamente del sitio web, esta sirve para el tratamiento de datos desordenados, se encarga de limpiarlos y transformarlos de un formato a otro. También es de ayuda para explorar grandes conjuntos de datos con facilidad y permite vincular y ampliar el conjunto de datos con varios servicios web [22], [23].

R Studio

Es un entorno de desarrollo integrado (IDE) de código abierto para el lenguaje R, el cual cuenta con un conjunto de herramientas para el historial, depuración y administración de datos [3], [24]. El software R proporciona una amplia variedad de técnicas estadísticas, es simple y efectivo y ampliamente utilizado entre los estadísticos y los mineros de datos para el desarrollo de software estadístico y análisis de datos. Entre otras características, R dispone de almacenamiento y manipulación efectiva de datos, posee una amplia, coherente e integrada colección de herramientas y posibilidades gráficas utilizadas para la ejecución de análisis de datos [13].

IBM SPSS Statistics

Es una plataforma de software que ofrece análisis estadísticos avanzados, una amplia biblioteca de algoritmos de aprendizaje automático, análisis de texto, extensibilidad de código abierto, integración con big data y capacidad para trabajar con grandes bases de datos, para ello permite crear gráficos, tablas, redes neuronales, árboles de decisión, etc. Posee una interfaz sencilla, de fácil uso, flexible y escalable, haciéndolo de esta manera accesible para usuarios de todos los niveles [25].

Weka

Es considerada como una plataforma de software para el aprendizaje automático y la minería de datos, contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades [3], además de herramientas para hacer el procesamiento previo de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización [13].

4.2. Trabajos relacionados

De acuerdo a la revisión bibliográfica realizada, la cual está planificada a través del establecimiento de objetivos, preguntas de investigación y las estrategias de búsquedas que incluyen la estructuración de los scripts de búsqueda, la elección de los repositorios virtuales y el establecimiento de los criterios de inclusión y exclusión para identificar, analizar e interpretar toda la información disponible en base al objeto de estudio, conforme a lo antes expuesto en la TABLA I, se presentan

los trabajos o estudios relacionados con el presente TT, los cuales fueron tomados como referencia para el desarrollo del mismo.

TABLA I
ESTUDIOS RELACIONADOS CON EL TT.

Nº	Estudios seleccionados	Ref.	Términos	Técnica
ES01	Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador.	[8]	Metodología KDD, R, Python.	Árboles y Reglas de Decisión.
ES02	Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellín.	[26]	Metodología KDD.	Árboles de Decisión.
ES03	Analysis and Predict the Nature of Road Traffic Accident Using Data Mining Techniques in Maharashtra, India.	[27]	Weka, RapidMiner, R.	Árboles de Decisión, Redes Bayesianas, Apriori, Reglas de Asociación.
ES04	Predicción de accidentes viales en Cartagena, Colombia, con Árboles de Decisión y Reglas de Asociación.	[28]	Metodología KDD, Weka.	Árboles de Decisión, Reglas de Asociación,
ES05	Road Accident Prediction Using Data Mining Techniques.	[29]	Weka.	Redes Bayesianas, Reglas de Asociación.
ES06	Brazilian Federal Roads: Identifying Patterns in traffic Accidents using Data Mining	[30]	Python, Weka.	Apriori, Reglas de Asociación.

	Techniques with Apriori Algorithm.			
ES07	Risk analysis of traffic accidents' severities: An application of three data mining models.	[31]	Weka.	Árbol de decisión, Redes Bayesianas.
ES08	Traffic Accidents Prediction Using Ensemble Machine Learning Approach.	[32]	Metodología KDD, Python.	Árboles de Decisión.
ES09	Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study.	[33]	Metodología KDD.	Árboles de Decisión, Redes Neuronales.
ES10	Data mining applied for accident prediction model in Indonesia toll road.	[34]	R.	Redes Neuronales Artificiales
ES11	Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset.	[35]	Metodología KDD, Weka.	Redes Neuronales Artificiales, Redes Bayesianas, Apriori, Reglas de Asociación.
ES12	A Radical Approach to Forecast the Road Accident Using Data Mining Technique.	[36]	Weka	Árboles de Decisión, Redes Bayesianas, CHAID, Naives Bayes
ES13	A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction.	[37]	R, Weka.	Redes Neuronales
ES14	Comparison of Machine Learning Algorithms for Predicting Traffic Accident	[38]	Weka.	Árboles de Decisión, Redes

	Severity.			Bayesianas.
--	-----------	--	--	-------------

En el estudio ES01 el autor realizó un análisis de las causas fundamentales de accidentes de tránsito en el Ecuador, empleando la metodología de extracción del conocimiento KDD, y el algoritmo de minería de datos C4.5 que le permitió realizar árboles de decisiones los cuales le ayudaron a organizar y clasificar la información sobre los accidentes de tránsito, extrayendo reglas relacionadas a las causas , esto a través de la creación de un modelo tradicional, que le ayudó a identificar los patrones más importantes en los datos y evaluó las posibles asociaciones entre las variables recogidas, es decir al final el autor identificó las reglas de mayor utilidad para que puedan ser utilizadas por el personal encargado de analizar los datos de seguridad vial.

Escobar, Rubiano y Vega autores del estudio ES02 utilizaron la metodología KDD para realizar el análisis de accidentalidad en la ciudad de Medellín durante los periodos 2018 y 2019, con la cual siguieron una ruta secuencial para la investigación, permitiéndoles tener un procedimiento estructurado para la selección de los datos, preprocesamiento, limpieza, transformación, aplicación de técnicas para finalmente obtener los resultados.

Con esto ellos describieron y demostraron el impacto que genera el análisis de datos, encontrando patrones y correlaciones entre los datos y brindando soluciones a los mismos, además plantearon algoritmos de machine learning para detectar futuros eventos clasificando los tipos de accidentes que le pueden ocurrir a los ciudadanos.

El estudio ES03, realiza un análisis riguroso de los datos de los accidentes de tráfico en Maharashtra, utilizando algoritmos de clasificación y minería de reglas de asociación, tales como los árboles de decisión y redes bayesianas, esto debido a que mostraron un mejor rendimiento en análisis anteriores. Adicionalmente, emplea el algoritmo de reglas de asociación Apriori para poder determinar la relación entre las variables independientes y la naturaleza de los accidentes. Finalmente, a través del uso de la minera de datos se descubre patrones novedosos que aún no fueron descubiertos en el ámbito de los siniestros de tránsito utilizando diversas herramientas de código abierto, estas fueron Weka, RapidMiner y R.

El estudio ES04, manifiesta el proceso estructurado de predecir los factores asociados con la severidad en los accidentes viales de Cartagena utilizando una metodología basada en técnicas de minería de datos, tales como los árboles de

decisión y reglas de asociación. La investigación fue desarrollada con 10.053 registros de accidentes de tráfico entre 2016 y 2017, por medio del uso del Software WEKA. Además, el autor obtuvo resultados que demuestran que, a través de la utilización de las reglas de decisión se evidencia que más de 50% de las reglas definidas están relacionados con usuarios de motocicletas, todo esto con el fin de ayudar a promover contramedidas para mejorar la seguridad vial de la ciudad.

El ES05, propone la aplicación de un análisis estructurado de los datos de tráfico para descubrir las variables que están estrechamente relacionadas con la ocurrencia de los accidentes de tránsito, se utiliza el análisis de probabilidad y algoritmos de minería de datos como las redes bayesianas, para poder determinar la relación entre la tasa de estos accidentes y otros atributos como la forma de colisión, el tiempo, el estado de la luz y el conductor ebrio. Los resultados de sus análisis realizados a través de la herramienta Weka, incluyen reglas de asociación entre las variables.

El estudio ES06 analiza los datos de accidentes de tránsito ocurridos en 2017 en las carreteras federales brasileñas, utilizando técnicas de minería de datos en conjunto con el algoritmo de asociación Apriori, estas técnicas mostraron su eficacia en la detección

de patrones a partir de diferentes variables, siendo útiles para tratar problemas relacionados con la seguridad vial. Este empezó con una limpieza de los datos a analizar, ejecutada en Python, con los datos limpios, procedieron a analizados mediante el algoritmo Apriori a través del software Weka con el fin de descubrir relaciones y patrones entre las variables de la base de datos.

Los autores del estudio ES07 aplican tres modelos de minería de datos, esto para proporcionar un análisis exhaustivo de los factores de riesgo que contribuyen a la gravedad de los accidentes de tránsito, de los datos recogidos por los departamentos de tráfico de Abu Dhabi, incluyendo 5740 accidentes notificados entre 2008 y 2013, se eligieron las variables con valores de importancia de los factores establecidos a través de las redes bayesianas, finalmente para comprender la correlación entre dichas variables utilizaron el árbol de decisión para poder examinar las correlaciones entre los posibles factores de riesgo.

El ES08, destaca la importancia de utilizar varias estrategias de clasificación para determinar los sucesos de accidentes de tráfico, a través de las directrices de la metodología KDD se concentran principalmente en la fase de implementación, para interpretar el efecto de la agrupación de los accidentes de tránsito en la precisión

del modelo de predicción, implementan a través del lenguaje Python dos modelos de aprendizaje automático para la predicción y se compara el rendimiento para seleccionar qué modelo se ajusta mejor, al final el modelo más destacado es el que implementa árboles de decisión.

El estudio ES09 presenta nuevas exploraciones sobre técnicas efectivas para abordar los desafíos relacionados con los accidentes de tránsito para obtener mejores resultados de predicción, se recogió una gran cantidad de datos que incluyen todos los siniestros de tránsito ocurridos desde el 2006 al 2013 en el estado de Iowa. Se evalúan cuatro modelos, además de abordar el problema de las clases desequilibradas utilizando un enfoque de muestreo negativo que ayuda a mejorar eficazmente el rendimiento de todos

los modelos. En concreto con el uso de árboles de decisión y las redes neuronales se obtienen mejores resultados.

El estudio ES10, propone un modelo de predicción de accidentes de tránsito en la carretera de peaje de Yakarta, esto para identificar las causas más importantes de los accidentes y así mismo desarrollar modelos de predicción a través de la utilización de técnicas de minería de datos (redes neuronales artificiales) para modelar y obtener patrones útiles a partir de todos los datos históricos existentes. Basándose en la herramienta R, se pudo determinar y revelar la superioridad de las redes neuronales artificiales en la predicción e identificación de los factores subyacentes a los accidentes en carretera.

El ES11, consiste en la utilización de la herramienta de minería de datos Weka, en la cual se evalúan los clasificadores de redes neuronales, árboles de decisión y redes bayesianas sobre el conjunto de datos de los accidentes de tránsito, los resultados mostraron que el clasificador de redes neuronales tuvo el más alto nivel de rendimiento en comparación con los demás, por lo tanto, la implementación de redes neuronales como clasificador o algoritmo de minería de datos es eficiente para tareas de predicción, finalmente se encontró las dos mejores reglas para la minería de reglas de asociación utilizando el algoritmo Apriori.

El estudio ES12, utiliza técnicas de minería de datos para analizar un conjunto de datos de accidentes de tránsito proponiendo un enfoque para crear un modelo mediante el uso de combinaciones de algoritmos de minería de datos, tales como redes bayesianas y árboles de decisión, los resultados manifiestan que el algoritmo con más precisión es el de los árboles de decisión para predecir la ocurrencia de accidentes de tráfico en el futuro.

El estudio ES13 recopila grandes bases de datos de accidentes de tránsito, mediante la realización de análisis de los patrones espaciales y temporales de la frecuencia de los accidentes de tránsito, en base a los patrones encontrados se propone un modelo de aprendizaje de alta precisión basado en una red neuronal para la predicción del riesgo de accidentes de tránsito, basado en el resultado del análisis de patrones, se puede observar que el riesgo de accidente de tráfico no está distribuido uniformemente en el espacio y el tiempo.

El ES14, establece modelos para seleccionar un conjunto de factores influyentes, con el fin de construir un modelo el cual clasifica la gravedad de las lesiones, estos modelos se formulan mediante diversas técnicas de minería de datos, incluidas los árboles de decisión y redes bayesianas los cuales se implementan en el análisis de datos de accidentes de tránsito. Los resultados de las pruebas mostraron que los árboles de decisión presentan un mejor rendimiento que los otros modelos y se recomienda su aplicación para estudiar el impacto de la ocurrencia de los siniestros de tránsito.

De acuerdo al análisis de cada uno de los trabajos relacionados al objeto de estudio, se evidencia que en la gran mayoría de estos se utilizan las técnicas de Árboles de Decisión, Redes Neuronales y Redes Bayesianas para analizar conjuntos de datos relacionados a eventos de siniestros de tránsito.

5. MATERIALES Y MÉTODOS

5.1. Tipo de investigación

De acuerdo al enfoque del presente TT, se empleó el tipo de investigación cuantitativa-cualitativa, siendo principalmente experimental, debido a que fue evaluada la base datos obtenida con el propósito de determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador, con la finalidad de indagar en su significado profundo, considerando las diferentes condiciones en los que ocurren cotidianamente. Una investigación cualitativa produce a través de la interpretación, datos descriptivos que permiten hacer conclusiones generalizadas que pueden ser proyectadas en el tiempo [39].

Además, fue una investigación de campo ya que fueron aplicadas entrevistas a profesionales en el área de regulación y control de tránsito y seguridad vial, con el objetivo de justificar y sustentar las razones por las que fue importante el desarrollo del presente TT [40].

5.2. Métodos de investigación

Método inductivo

El método inductivo tiene un razonamiento ascendente que fluye de casos particulares o individuales hasta conocimientos generales, planteando la observación, estudio y conocimiento de propiedades genéricas, habituales o comunes partiendo de pautas existentes en los datos, se razona que la premisa inductiva es una reflexión enfocada en el fin [39], [41]. En la ejecución del presente TT, fue utilizado para la aplicación y evaluación de los algoritmos de minería de datos, analizando los resultados obtenidos por cada uno de ellos, para así poder definir cual cuales fueron los algoritmos con los resultados más óptimos.

Método deductivo

El método deductivo va de lo general a lo particular, parte de conocimientos generales e información recopilada para generalizar un caso de estudio particular [41]. En la realización de este TT, fue empleado en la definición de los siniestros de tránsito como problema global y al determinar los factores más influyentes en la ocurrencia de los mismos como parte específica. Así mismo se aplicó en la revisión de literatura, su planificación fue realizada partiendo de conceptos generales hacia los conceptos específicos.

Método científico

Es el procedimiento mediante el cual se desarrolla una investigación a través de la observación sistemática, medición y experimentación, la formulación y la interpretación; permitiendo alcanzar conocimiento objetivo de la realidad, tratando de dar respuesta a las interrogantes acerca del orden de la investigación, concretamente es un método que relaciona la ciencia con el conocimiento científico [42]. En el proceso de aplicación de los algoritmos de minería de datos se empleó este método para obtener el mejor conjunto de datos, del mismo modo, en el establecimiento de los aspectos de experimentación a través de la selección de los parámetros de configuración de los algoritmos aplicados, así mismo para el análisis de datos y la evaluación de los resultados siendo necesario definir métricas para cuantificar la eficacia con el fin de llegar a conclusiones finales en relación al objeto de estudio del presente TT.

Además, con este método se buscaron, analizaron y sintetizaron los conceptos presentes en la revisión de literatura que dieron fundamento teórico al proceso investigativo.

Método sistémico

El método sistémico analiza el problema en su complejidad, por medio de un proceso basado en la totalidad y la forma de interactuar entre las partes y sus propiedades emergentes resultantes, este viene a ser un orden manifestado por reglas, que permiten llegar a tener una comprensión sistémica de una situación dada [43]. Este método fue considerado para el presente TT, ya que mediante la aplicación de los algoritmos de minería de datos se analizó los datos recolectados previamente, con el propósito de determinar los factores más influyentes en la ocurrencia de siniestros de tránsito.

5.3. Técnicas de investigación

Observación

Se utilizó esta técnica para efectuar la evaluación de los resultados y de los algoritmos de minería de datos, fue de ayuda para seleccionar los mejores resultados, esto para obtener los factores más influyentes en la ocurrencia de siniestros de tránsito.

Entrevista

La entrevista es una técnica que posibilita obtener información acerca de las características de un problema de un informante clave. Dicha información puede ser novedosa o complementaria y ayuda a cuantificar características y la naturaleza del objeto de estudio [44]. Esta técnica se utilizó para adquirir información que sustente y justifique el presente TT; la misma que fue dirigida a autoridades de la Unidad de Control Operativa de Tránsito (UCOT) del GAD Municipal de Loja que son profesionales en cuanto a problemas de planificación, regulación y control del tránsito se refiere y además velan por la seguridad vial, estos son el Abg. Cnel. Paul Aguilar en calidad de Director Estratégico de la UCOT y el Sr. Patricio Benítez Agente Civil de Tránsito de la misma institución, dichas entrevistas (ver Anexo 1) aportaron con información que enfatizan la importancia de los resultados del presente TT.

5.4. Metodología

Para el presente TT se tomó como referencia las fases de la metodología KDD, adaptando cinco fases relacionadas, las cuales son: búsqueda de información, obtención de datos, depuración de base de datos, aplicación de técnicas de minería de datos y por último la interpretación y presentación de los resultados, tal como se muestra en la Fig. 2. Estas fases se describen a continuación:

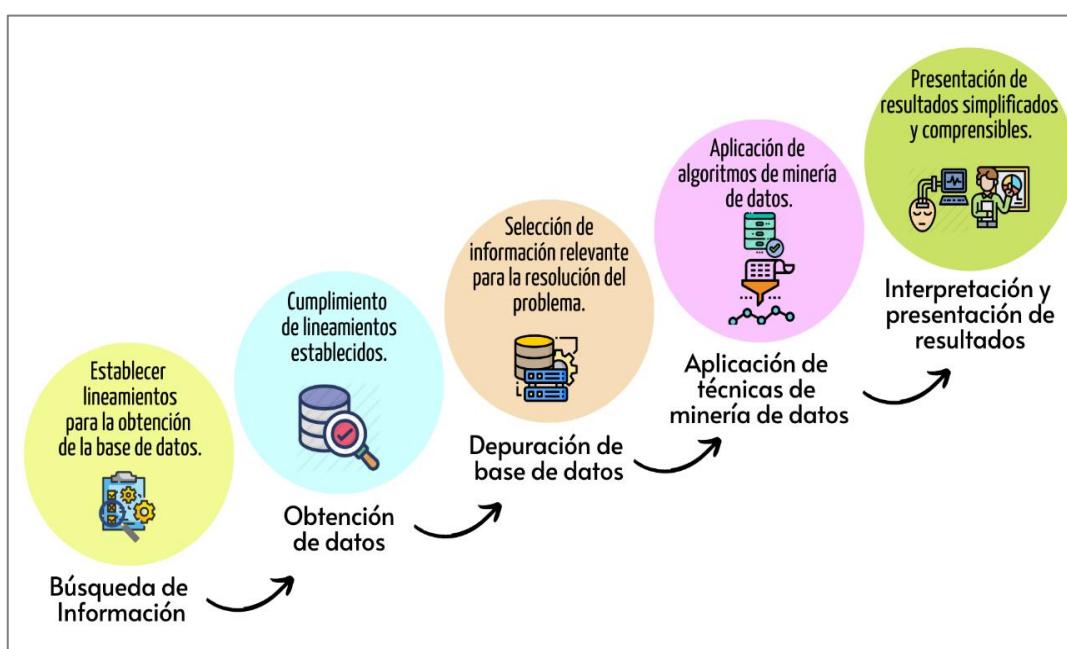


Fig. 2 Metodología aplicada en el TT

Fase I: Búsqueda de información

En esta fase se hizo una investigación que proporcionó información acerca del problema, se identificaron las instituciones públicas del Ecuador que tienen como finalidad la planificación, regulación y control del tránsito. Una vez seleccionadas las instituciones que velan por la seguridad vial, se identificaron aquellas que proporcionaban información acerca de los eventos de siniestros de tránsito. Al final se establecieron lineamientos que facilitaron la obtención de bases de datos confiables.

Fase II: Obtención de datos

Para el desarrollo del presente TT, fue necesario la obtención de datos referentes a Siniestros de Tránsito en Ecuador, por consiguiente, se procedió a seleccionar las bases de datos que cumplían con los lineamientos establecidos en la fase anterior. Se delimitó que la información obtenida debía ser recopilada del año 2020, correspondiente al Ecuador y sus provincias, además debía ser obtenida de fuentes oficiales y fidedignas. La base de datos seleccionada fue la realizada por la Agencia Nacional de Tránsito a partir de los partes policiales que son diseñados y aprobados por cada uno de los entes de control, bajo los parámetros técnicos establecidos por la misma institución [45].

Fase III: Depuración de la base de datos

Dentro de esta fase, primero se realizó la evaluación del conjunto de datos, esto con el fin de seleccionar las variables relevantes para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador, después se procedió a la limpieza del conjunto de datos que contiene las variables antes seleccionadas a través de la herramienta OpenRefine, con la cual se estandarizaron el nombre de las variables y los datos contenidos en los registros transformándolos a mayúsculas, además se procedió a eliminar y remplazar tildes y otros caracteres que distorsionaban la detección de patrones y no aportaban al descubrimiento de conocimiento, y por último a través de la herramienta de RStudio se realizó la eliminación de registros que presentaban inconsistencias relacionada a leyes de tránsito existentes en Ecuador.

Fase IV: Aplicación de técnicas de minería de datos

Más adelante, en esta fase se aplicaron siete algoritmos predictivos al conjunto de datos depurado, los algoritmos seleccionados de acuerdo a la literatura encontrada

fueron AD CHAID, AD CHAID Exhaustivo, AD CRT, RN Perceptrón Multicapa, RN de Función de Base Radial, RB Naive Bayes y RB BayesNet, para la aplicación de los algoritmos de árboles de decisión y redes neuronales se utilizó la herramienta SPSS Statistics y para los algoritmos de redes bayesianas se usó la herramienta Weka, al aplicar los algoritmos de clasificación la “CLASE_FINAL” fue establecida como variable objetivo.

Fase V: Interpretación y presentación de resultados

Finalmente en esta fase se eligió el algoritmo que proyectó mejores resultados, comparándolos con respecto a medidas de precisión y porcentajes de clasificación correcta de los datos, el algoritmo CHAID Exhaustivo fue el algoritmo con mejores porcentajes presentados y en base a este se generaron reportes con las reglas de asociación de patrones, gráficos y tablas que facilitaron la interpretación de las mismas para así determinar los factores más influyentes en la ocurrencia de siniestros de tránsito y de esta manera obtener conclusiones finales del análisis realizado.

6. RESULTADOS

En esta sección se detallan los resultados de cada uno de los objetivos específicos del presente TT, obtenidos a través de la aplicación de la metodología planteada en la Fig. 2, esto en el marco del cumplimiento de los objetivos propuestos.

6.1. Objetivo 1: Identificar los repositorios donde se encuentra almacenada la información sobre los siniestros de tránsito en Ecuador en el año 2020.

En la ejecución del presente objetivo se cumplieron tres fases de la metodología KDD especificadas a continuación:

Fase I: Búsqueda de información

Tarea 1: Establecer los lineamientos, para la búsqueda de la información relevante sobre los siniestros de tránsito en Ecuador en el año 2020.

Una vez analizadas las instituciones gubernamentales encargadas del control vial, así como aquellas que brindan información acerca de la ocurrencia de siniestros de tránsito. Se procedió a establecer los lineamientos para la selección de las bases de datos con la finalidad de identificar la más optima, los cuales son establecidos a continuación.

En la definición del primer criterio de inclusión para la selección de las bases de datos (ver TABLA II), se tiene en consideración el contenido de las variables relacionadas a los eventos de siniestros de tránsito como la provincia, cantón, mes, el día y hora de la ocurrencia, las causas probables, tipología o clase del siniestro, los tipos de vehículos involucrados, el sexo, edad, tipo y condición de los involucrados, esto, basado en los datos presentados en el “Reporte Nacional de Siniestros de Tránsito”, elaborado por la ANT [46], otro referente para este criterio se basa en la investigación [30] realizada para identificar patrones en accidentes de tránsito, en este estudio se analizan varios factores para la detección de patrones a partir de diferentes variables, la mayoría tratados a través de las variables anteriormente mencionadas.

Para el segundo criterio se toma en cuenta que la base de datos seleccionada debía contar como mínimo con nueve variables como objeto a analizar, esto, en concordancia con la investigación [8] en la cual se analizan nueve atributos en total para la identificación de patrones en la ocurrencia de accidentes de tránsito. En cuanto al primer y segundo criterio de exclusión se define que, no deben ser tomadas en cuenta bases de datos encontradas que contengan variables con contenido no relacionado y que cuenten con menos de nueve variables, ya que de

acuerdo a [28], para analizar siniestros de tránsito y detectar patrones de ocurrencia deben ser analizados varios factores relacionados.

En el tercer y cuarto criterio de inclusión se establece que, los datos que se obtengan hayan sido recabados en el año 2020 y sean provenientes de fuentes oficiales respectivamente, lo cual resulta imprescindible para la selección de información sustentada y avalada por instituciones gubernamentales.

Por último, el quinto criterio a considerar para seleccionar la base de datos, menciona que se debe tener acceso a la información sin realizar procesos administrativos prolongados para su obtención, es decir el acceso debe ser público.

TABLA II
CRITERIOS PARA LA SELECCIÓN DE LA BASE DE DATOS

Criterios de Inclusión	Criterios de Exclusión
Contenido Relacionado	Contenido no Relacionado
9 o más variables	Menos de 9 variables
Año 2020	Otros años
Fuentes oficiales	Fuentes no oficiales
Acceso público	Solicitud para obtener información

Fase II: Obtención de datos

Tarea 2: Obtener la base de datos con la información relevante sobre los siniestros de tránsito en Ecuador en el año 2020.

Ya establecidos los criterios para la selección de las bases de datos mostrados en la TABLA II, se inició el proceso para la obtención de las mismas, acotando que solo fue adquirido un conjunto de datos alojado en la página oficial de la ANT [45], esto no permitió que el protocolo de búsqueda establecido en la fase anterior fuera evaluado.

En el conjunto de datos obtenido, se encuentra almacenada la información recopilada a partir de datos de los partes policiales del año 2020. La base de datos cuenta con 418 variables correspondientes a las categorías incluidas en dichos partes policiales, diseñados y aprobados por cada uno de los entes de control, bajo los parámetros técnicos establecidos por la ANT [45] y 16972 registros (ver [Repositorio](#)¹⁾ de eventos ocurridos sobre siniestros de tránsito en Ecuador.

¹https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/datos_inicial.xlsx

Considerando lo establecido en la Tabla II, esta base de datos cumple con los cinco criterios establecidos. A través de la obtención de información se adquirió el conjunto de datos el cual fue el principal insumo para trabajar durante el desarrollo de las siguientes etapas.

Fase III: Depuración de la base de datos

Tarea 3: Depurar la base de datos obtenida

Evaluación del conjunto de datos

El conjunto de datos inicial contiene información sin tratar, en formatos no óptimos para que sea posible aplicar modelos de algoritmos de minería de datos. En cada una de las variables del conjunto de datos se muestran las características relacionadas a cada uno de los eventos de siniestros de tránsito ocurridos en Ecuador en el año 2020, como el mes, código, ente de control, provincia, causa probable, etc. En la TABLA III se presentan todas las variables contenidas en el conjunto de datos; además, en el Anexo 2 se presenta parte del conjunto de datos con la información aún sin procesar.

TABLA III

LISTA DE VARIABLES DEL CONJUNTO DE DATOS OBTENIDO

Nº	Variable	Tipo de variable	Descripción
1	MES	Categórica	Mes de ocurrencia del siniestro
2	AÑO	Numérica	Año de ocurrencia del siniestro
3	CÓDIGO	Categórica	Ente de control que registró el siniestro
4	ENTE DE CONTROL	Categórica	Ente de control
5	PROVINCIA	Categórica	Provincia de ocurrencia del siniestro
6	ZONA PLANIFICACIÓN	Categórica	Zona de planificación de ocurrencia del siniestro
7	DÍA	Categórica	Día de ocurrencia del siniestro
8	FECHA	Numérica	Fecha de ocurrencia del siniestro
9	HORA	Numérica	Hora de ocurrencia del siniestro

10	PERIODO	Numérica	Código para la hora de ocurrencia del siniestro
11	FERIADO	Categórica	Feriado en el día de la ocurrencia del siniestro
12	CÓDIGO CAUSA	Categórica	Código para la causa probable del siniestro
13	CAUSA PROBABLE	Categórica	Causa probable del siniestro
14	CAUSA FINAL	Categórica	Clase del siniestro
15	ZONA	Categórica	Zona de ocurrencia del siniestro
16	LATITUD (X)	Numérica	Latitud de ocurrencia del siniestro
17	LONGITUD (Y)	Numérica	Longitud de ocurrencia del siniestro
18	DIRECCIÓN	Texto	Dirección de ocurrencia del siniestro
19	CANTÓN	Categórica	Cantón de ocurrencia del siniestro
20	PARROQUIA	Categórica	Parroquia de ocurrencia del siniestro
21 – 30	[TIPO DE VEHÍCULO 1 – TIPO DE VEHÍCULO 10]	Categórica	Tipo de vehículo involucrado en el siniestro
31 – 40	[SERVICIO 1 – SERVICIO 10]	Categórica	Servicio del vehículo involucrado en el siniestro
41	AUTOMÓVIL	Numérica	Tipo de vehículo involucrado en el siniestro
42	BICICLETA	Numérica	Tipo de vehículo involucrado en el siniestro
43	BUS	Numérica	Tipo de vehículo involucrado en el siniestro
44	CAMIÓN	Numérica	Tipo de vehículo involucrado en el siniestro
45	CAMIONETA	Numérica	Tipo de vehículo involucrado

			en el siniestro
46	EMERGENCIAS	Numérica	Tipo de vehículo involucrado en el siniestro
47	ESPECIAL	Numérica	Tipo de vehículo involucrado en el siniestro
48	FURGONETA	Numérica	Tipo de vehículo involucrado en el siniestro
49	MOTOCICLETA	Numérica	Tipo de vehículo involucrado en el siniestro
50	NO IDENTIFICADO	Numérica	Tipo de vehículo no identificado
51	VEHÍCULO DEPORTIVO UTILITARIO	Numérica	Tipo de vehículo involucrado en el siniestro
52	SUMA DE VEHÍCULOS	Numérica	Suma de los vehículos involucrado en el siniestro
53	NUM_FALLECIDO	Numérica	Número de lesionados en el siniestro
54	NUM_LESIONADO	Numérica	Número de fallecidos en el siniestro
55 – 106	[TIPO DE IDENTIFICACIÓN 1 – TIPO DE IDENTIFICACIÓN 52]	Categórica	Tipo de identificación de la víctima
107 – 158	[EDAD 1 – EDAD 52]	Numérica	Edad de la víctima
159 – 210	[SEXO 1 – SEXO 52]	Categórica	Sexo de la víctima
211 – 262	[CONDICIÓN 1 – CONDICIÓN 52]	Categórica	Condición de la víctima
263	[PARTICIPANTE 1 –	Categórica	Tipo de participante en el

– 314	PARTICIPANTE 52]		siniestro
315 – 366	[USO DE CASCO 1 – USO DE CASCO 52]	Categórica	Uso de caso en el siniestro
367 – 418	[USO DE CINTURÓN DE SEGURIDAD 1 – USO DE CINTURÓN DE SEGURIDAD 52]	Categórica	Uso de cinturón de seguridad en el siniestro

Para la comprensión y entendimiento del significado de cada variable del conjunto de datos fue necesario la obtención de un diccionario de datos (ver [Repositorio](#)²), en el cual se describieron cada una de las variables y sus categorías existentes, en el Anexo 3 se presenta el diccionario de datos obtenido.

Mediante el análisis exploratorio de los datos se evaluó el conjunto de datos, esto permitió considerar a la variable “AÑO” irrelevante, ya que todos los registros del contenidos son sobre siniestros de tránsito ocurridos en el año 2020, por lo tanto, se procedió a eliminarlo.

A las variables “CAUSA PROBABLE” y “CLASE FINAL” se los mantiene en el conjunto de datos, para que sea posible identificar qué clase de siniestro de tránsito ocurrió y cuál fue la causa más probable por el que se suscitó. Posterior a ello también se procedió a eliminar las siguientes variables por ser irrelevantes, ya que por medio del análisis exploratorio de los datos se visualizó que estas no son de gran influencia para que ocurra un siniestro de tránsito:

- MES
- CÓDIGO
- ENTE DE CONTROL
- ZONA DE PLANIFICACIÓN
- FECHA
- LATITUD (X)
- LONGITUD (Y)

²https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/diccionario_datos_siniestros_transito_2020.pdf

- DIRECCIÓN
- CANTÓN
- PARROQUIA

Las variables “PERIODO” y “CÓDIGO CAUSA” son eliminadas debido a que almacenaban a través de códigos información redundante con las variables “HORA” y “CAUSA PROBABLE”.

En el conjunto de datos la variable llamada “TIPO DE VEHÍCULO 1” tiene relación directa con la variable “SERVICIO 1”, este caso se repite hasta llegar a la variable “TIPO DE VEHÍCULO 10” y “SERVICIO 10”, en estas variables fueron almacenados datos correspondientes a los diferentes tipos de vehículos involucrados en un siniestro de tránsito y el tipo de servicio que les corresponde respectivamente, tomando en cuenta que la cantidad de vehículos involucrados en el suceso varía, se observó que no todos los registros de estas variables están llenos completamente, en la Fig. 3 se muestra el número total de registros para cada una de las variables antes mencionadas, observando que las únicas variables completadas al 100%, con un total de 16972 registros, son las variables “TIPO DE VEHÍCULO 1” y “SERVICIO 1”, por este motivo estas variables fueron tomadas como las principales, de esta manera se procedió a eliminar las variables restantes, tales como:

- [TIPO DE VEHÍCULO 2 – TIPO DE VEHÍCULO 10]

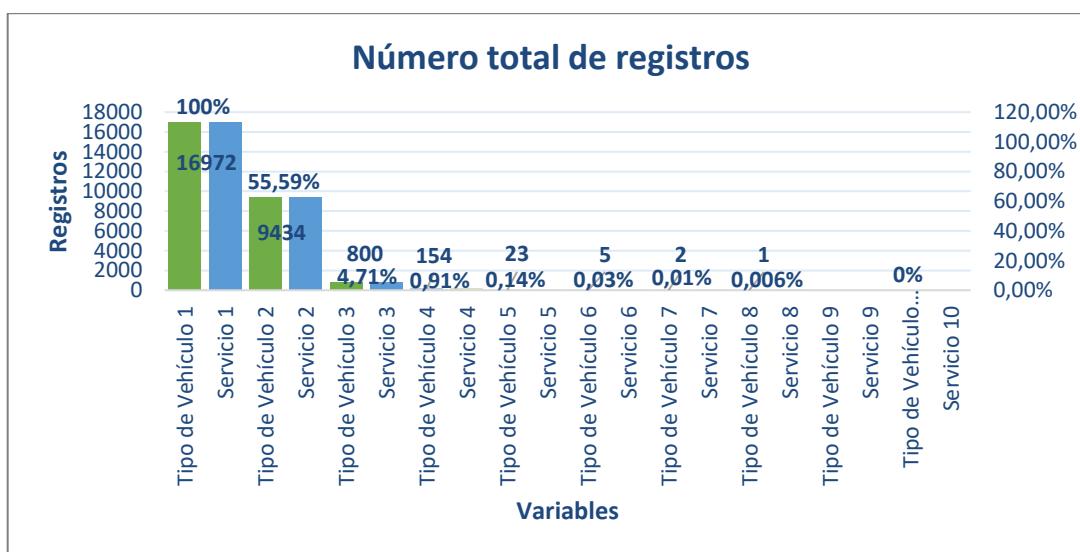


Fig. 3 Número total de registros de las variables Tipo de Vehículo 1 – Tipo de Vehículo 10

Las variables que muestran los diferentes tipos de vehículos involucrados como son “AUTOMÓVIL”, “BICICLETA”, “CAMIÓN”, “CAMIONETA”, “EMERGENCIAS”,

“ESPECIAL”, “FURGONETA”, “MOTOCICLETA”, “NO IDENTIFICADO” y “VEHÍCULO DEPORTIVO UTILITARIO”, son consideradas irrelevantes y no útiles debido a que más de la mitad de sus registros totales almacenan valores iguales a “0”, tal como se visualiza en la Fig. 4 , por tal motivo se procedió a eliminarlas ya que de la manera en la que se presentan no aportan al análisis de los siniestros de tránsito.

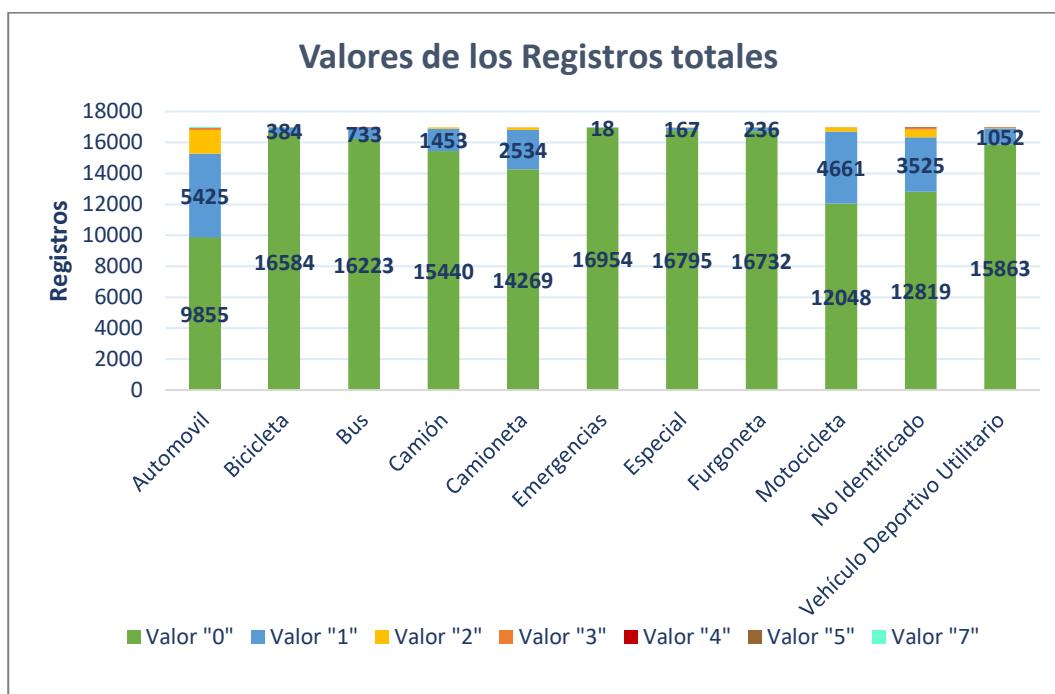


Fig. 4 Valores de los registros totales

En el caso del conjunto de variables [EDAD 1 – EDAD 52], [SEXO 1 – SEXO 52], [CONDICIÓN 1 – CONDICIÓN 52] y [PARTICIPANTE 1 – PARTICIPANTE 52], se vio necesario tomar la primera variable como la principal, debido a que las variables restantes no tenían un valor considerable de registros y en algunos casos se encontraban completamente vacíos y de esta manera no son útiles, en la TABLA IV se presentan las variables consideradas para poder determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador, estas variables seleccionadas con referencia a las investigaciones [8], [28] y [30] que tratan de la utilización de minería de datos para el análisis de los siniestros de tránsito.

TABLA IV

LISTA DE VARIABLES RELEVANTES SELECCIONADAS

Nº	Variables Relevantes	Tipo de variable
1	PROVINCIA	Categórica
2	DÍA	Categórica

3	HORA	Categórica
4	FERIADO	Categórica
5	CAUSA PROBABLE	Categórica
6	CLASE FINAL	Categórica
7	ZONA	Categórica
8	TIPO DE VEHÍCULO 1	Categórica
9	SERVICIO 1	Categórica
10	EDAD 1	Numérica
11	SEXO 1	Categórica
12	CONDICIÓN 1	Categórica
13	PARTICIPANTE 1	Categórica

Una vez realizada la evaluación del conjunto de datos utilizando el análisis exploratorio de los datos, el cual proporcionó un conocimiento específico de todo el conjunto de datos, se eliminó las variables irrelevantes, dejando así las variables mostradas en la TABLA IV, para así obtener el archivo “dataset_eval” y proceder a la depuración del conjunto de datos, el conjunto de datos obtenido se lo visualiza en la Fig. 5, donde se muestran las trece variables seleccionadas con 16972 registros (ver [Repositorio³](#)) para cada una de ellas.

PROVINCIA	DIA	HORA	FERIADO CAUSA PROBABLE	CLASE FINAL	ZONA	TIPO DE VEHÍCULO 1	SERVICIO 1	EDAD 1	SEXO 1	CONDICIÓN 1	PARTICIPANTE 1
GUAYAS	miércoles	1:10:00 SI	CONDUCIR DESAT ATROPELLOS	URBANA	MOTOCICLETA	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
MANABÍ	miércoles	5:15:00 SI	NO RESPETAR LAS CHOQUE LATI RURAL	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	LESIONADO PASAJERO				
LOS RÍOS	miércoles	3:10:00 SI	NO MANTENER LA CHOQUE POS RURAL	MOTOCICLETA	PARTICULAR	-1 NO IDEN	LESIONADO CONDUCTOR				
GUAYAS	miércoles	3:10:00 SI	CONDUCIR DESAT ATROPELLOS	RURAL	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
SANTO DOMI	miércoles	7:30:00 SI	CONDUCIR EN SEN CHOQUE FRO RURAL		VEHÍCULO DEPORTIV	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
GUAYAS	miércoles	4:30:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL		AUTOMÓVIL	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
GUAYAS	miércoles	4:00:00 SI	CONDUCIR EN SEN CHOQUE FRO RURAL		NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
GUAYAS	miércoles	7:00:00 SI	NO GUARDAR LA CZ ROZAMIENTO RURAL	BUS	PÚBLICO	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	7:00:00 SI	CONDUCIR DESAT ESTRELLAMIE URBANA	MOTOCICLETA	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	9:00:00 SI	CONDUCIR DESAT ATROPELLOS	URBANA	AUTOMÓVIL	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
SANTO DOMI	miércoles	4:00:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL		CAMIÓN	PARTICULAR	50 HOMBRE	LESIONADO CONDUCTOR			
EL ORO	miércoles	7:20:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL	MOTOCICLETA	PARTICULAR	48 HOMBRE	LESIONADO CONDUCTOR				
GUAYAS	miércoles	10:30:00 SI	CONDUCIR EN SEN CHOQUE FRO RURAL	CAMIIONETA	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	12:30:00 SI	BAJARSE O SUBIRS CAÍDA DE PAS RURAL	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	2:01:00 SI	CONDUCIR DESAT ARROLLAMIEI URBANA	AUTOMÓVIL	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
AZUAY	miércoles	3:40:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL	MOTOCICLETA	PARTICULAR	-1 HOMBRE	FALLECIDO CONDUCTOR				
GUAYAS	miércoles	13:40:00 SI	NO MANTENER LA CHOQUE POS URBANA	AUTOMÓVIL	PARTICULAR	48 HOMBRE	ILESO CONDUCTOR				
LOS RÍOS	miércoles	8:10:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL	AUTOMÓVIL	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	1:30:00 SI	CONDUCIR DESAT ESTRELLAMIE RURAL	CAMIIONETA	PARTICULAR	68 HOMBRE	ILESO CONDUCTOR				
SANTO DOMI	miércoles	14:00:00 SI	NO MANTENER LA CHOQUE POS RURAL	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
SANTA ELENA	miércoles	9:40:00 SI	CONDUCIR DESAT PÉRDIDA DE C RURAL	AUTOMÓVIL	PARTICULAR	27 HOMBRE	FALLECIDO CONDUCTOR				
GUAYAS	miércoles	17:20:00 SI	CONDUCIR DESAT ATROPELLOS	URBANA	AUTOMÓVIL	PARTICULAR	54 HOMBRE	ILESO CONDUCTOR			
SANTA ELENA	miércoles	9:45:00 SI	CONDUCIR EN SEN CHOQUE FRO RURAL	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
GUAYAS	miércoles	3:40:00 SI	CONDUCIR DESAT ATROPELLOS	URBANA	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR			
SANTA ELENA	miércoles	9:00:00 SI	NO MANTENER LA CHOQUE POS URBANA	NO IDENTIFICADO	PARTICULAR	-1 NO IDEN	NO IDENTIFIC/ CONDUCTOR				
AZUAY	miércoles	15:08:00 SI	CONDUCIR EN SEN CHOQUE FRO RURAL	CAMIIONETA	PARTICULAR	-1 NO IDEN	ILESO CONDUCTOR				

Fig. 5 Conjunto de datos obtenido después de la evaluación

³https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/dataset_eval.csv

Eliminación de información innecesaria

La limpieza de información innecesaria del conjunto de datos, consiste en limpiar información no útil que pueda afectar las relaciones entre las palabras, esta se la realizó en todas las variables del conjunto de datos, porque, se observó qué caracteres pueden incidir para una mala aplicación de los algoritmos de minería de datos, y se llegó a determinar lo siguiente:

- Los nombres de las variables de conjunto de datos estaban escritos con palabras con tildes y espacios, por ejemplo, la variable “TIPO DE VEHÍCULO 1”, esto representa un problema al tratar de leer este tipo de variables en R Studio, por tal motivo se procedió a renombrar las variables tal como se muestra en la TABLA V.

TABLA V
VARIABLES RENOMBRADAS

Nombre de la Variable Original	Nombre de la Variable Actual
CAUSA PROBABLE	CAUSA_PROBABLE
CLASE FINAL	CLASE_FINAL
TIPO DE VEHÍCULO 1	TIPO_DE_VEHICULO_1
SERVICIO 1	SERVICIO_1
EDAD 1	EDAD_1
SEXO 1	SEXO_1
CONDICIÓN 1	CONDICION_1
PARTICIPANTE 1	PARTICIPANTE_1

- El conjunto de datos contaba con variables que almacenaban información con letras en mayúsculas y minúsculas, esto afectaba notablemente al no ser reconocidos los caracteres como iguales cuando se compara una cadena. Se estandarizó el conjunto de datos, a todas las letras se las convirtió a mayúsculas.
- Las tildes en las palabras afectan de igual forma al reconocer un carácter con tilde y otro sin tilde como diferentes. En el conjunto de datos existían palabras con errores ortográficos, se las estandarizó reemplazando todas las vocales con tilde por vocales sin tilde.
- La letra “ñ” forma parte en algunos registros de las variables, se considera que esta letra representa en R Studio un problema al no poder visualizarse

correctamente, esta fue reemplazada por la letra “n”.

En la TABLA VI se muestran todos los caracteres que se reemplazaron en el conjunto de datos:

TABLA VI
CARACTERES REEMPLAZADOS O ELIMINADOS

	Carácter Original	Carácter Actual
Tildes	Á	A
	É	E
	Í	I
	Ó	O
	Ú	U
Letras	Ñ	N

Para limpiar el conjunto de datos se utilizó la herramienta de software OpenRefine, usando el lenguaje GREL⁴. Se cargó el conjunto de datos en OpenRefine, seleccionando la codificación UTF-8, tal como se muestra en la Fig. 6.

Fig. 6 Cargar conjunto de datos en OpenRefine

Una vez se cargó el conjunto de datos, se renombró las variables mostradas en la TABLA V, en la Fig. 7 se muestra que para esto solo se eligió la opción “Cambiar el nombre de esta columna”.

⁴ Lenguaje de Expresión de Refinamiento General

16972 filas											
Mostrar como: registros de filas Mostrar: 5 10 25 50 filas											
«Primera < anterior 1 - 10 siguiente > última»											
Todas	PROVINCIA	DIA	HORA	FERIADO	CAUSA PROBAB	CLASE FINAL	ZONA	TIPO DE VEHICL	SERVICIO 1	EDAD_1	SEXO 1
1.	GUAYAS	miércoles	01:10:00	SI	Faceta	TROPELLOS	URBANA	MOTOCICLETA	ESPECIAL	-1	NO IDENTIFICADO
					Filtro de texto						
					Editar celdas						
					Editar columna						
					Dividido en varias columnas ...						
					Transponer						
					Unir columnas ...						
					Clasificar...						
					Vista						
					Conciliar						
					DEL SEMAFORO, ETC)						
					NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO AL VEHICULO QUE LE ANTECEDE.						
					CONDUCIR DESATENTO A LAS CONDICIONES DE TRANSITO (USAR PANTALLAS DE VIDEO, COMIDA, MAQUILLAJE O CUALQUIER OTRO ELEMENTO						
					A						
					Mover columna a la izquierda						
					Mover columna a la derecha						
2.	MANABÍ	miércoles	05:15:00	SI							
3.	LOS RÍOS	miércoles	03:10:00	SI							
4.	GUAYAS	miércoles	03:10:00	SI							

Fig. 7 Cambio de nombre de las variables

Antes de reemplazar las tildes fue necesario estandarizar la nomenclatura de los registros de todas las variables, convirtiendo todos los caracteres a mayúsculas, para este caso solo se eligió la opción “A mayúsculas”, tal como se indica en la Fig. 8.

16972 filas											
Mostrar como: registros de filas Mostrar: 5 10 25 50 filas											
«Primera < anterior 1 - 10 siguiente > última»											
Todas	PROVINCIA	DIA	HORA	FERIADO	CAUSA_PROBAB	CLASE_FINAL	ZONA	TIPO_DE_VEHIC	SERVICIO_1	EDAD_1	SEXO_1
1.	Faceta	días	01:10:00	SI	CONDUCIR DESATENTO A LAS CONDICIONES DE TRANSITO (CELULAR, PANTALLAS DE	ATROPELLOS	URBANA	MOTOCICLETA	ESPECIAL	-1	NO IDENTIFICADO
	Filtro de texto				Recorre los espacios en blanco iniciales y finales						
	Editar celdas				Contraer espacios en blanco consecutivos						
	Editar columna				Entidades HTML de Unescape						
	Transformaciones comunes				Reemplazar comillas inteligentes con ascii						
	Transponer				Al título						
	Llenar hacia abajo				A mayúsculas						
	Clasificar...				A minúsculas						
	Vista				Al numero						
	Dividir celdas de varios valores ...				Hasta la fecha						
	Unir celdas de varios valores ...				Textear						
					Anular						
					Para vaciar la cuerda						
					VIDEO, COMIDA, MAQUILLAJE O CUALQUIER OTRO ELEMENTO DISTRATORIO)						
					CONDUCIR EN SENTIDO CONTRARIO A LA	FRONTAL CHOQUE	RURAL	VEHICULO DEPORTIVO (TITANICO)	ESPECIAL	-1	NO IDENTIFICADO
2.	LOS RÍOS	miércoles									
3.	GUAYAS	miércoles	03:10:00	SI							
4.	SANTO DOMINGO DE LOS	miércoles	07:30:00	SI							

Fig. 8 Conversión de texto a mayúsculas

Para el reemplazo de las tildes de los registros de todas las variables, se utilizó el comando `value.replace("valor actual", "valor nuevo")`, mostrado en la Fig. 9, poniendo, por ejemplo: `value.replace("É", "E")`.

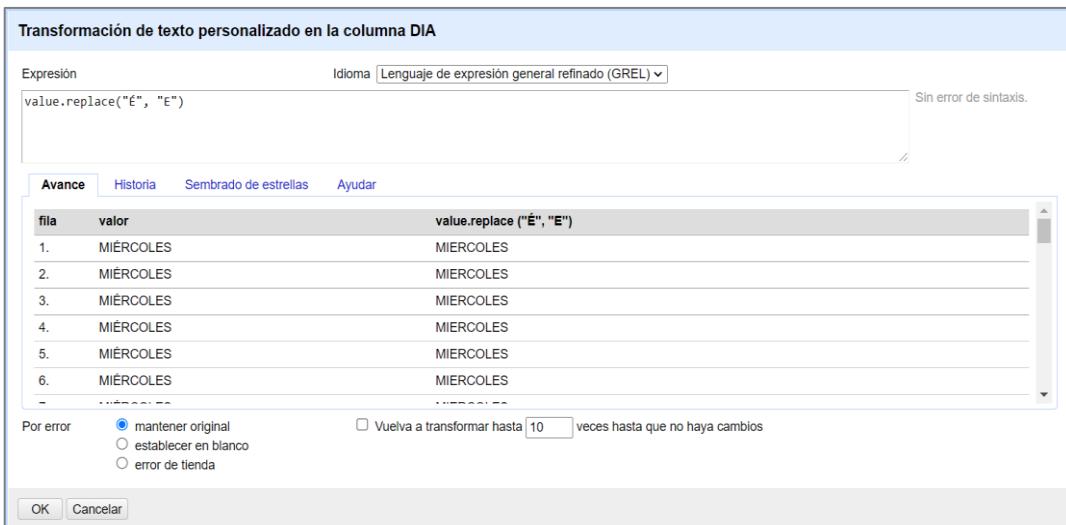


Fig. 9 Estandarización de las tildes

Por último, se procedió a la eliminación de la letra “Ñ”, reemplazándola por la letra “N”, utilizando nuevamente el comando `value.replace("valor actual", "valor nuevo")`, expresándolo de la siguiente manera: `value.replace("Ñ", "N")`, tal como se observa en la Fig. 10.

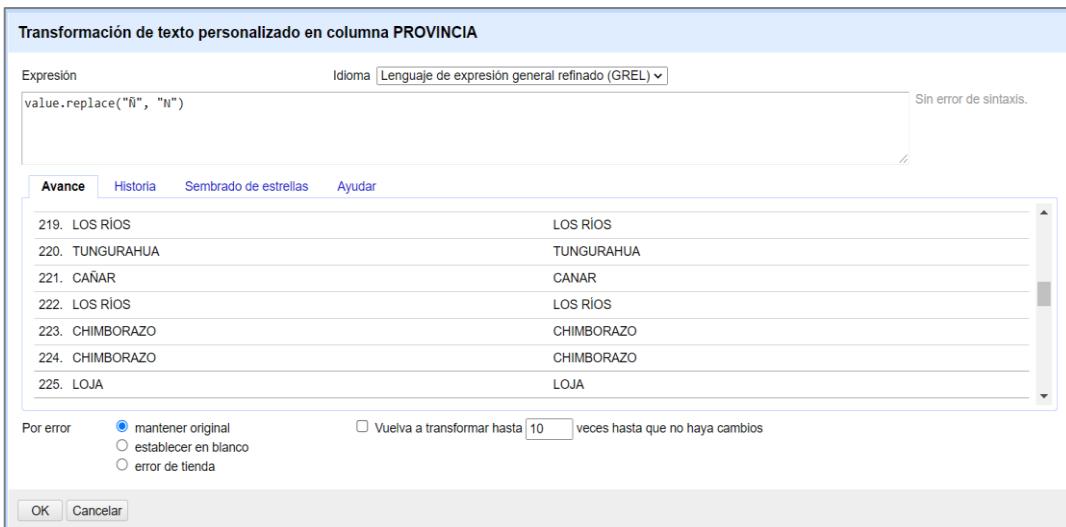


Fig. 10 Eliminación de la letra “Ñ”

Al terminar la limpieza del conjunto de datos, este fue exportado en un archivo con formato CSV, este archivo reemplazó al archivo anterior puesto que se encuentra estandarizado siendo nombrado como “dataset_estandarizado.csv” (ver [Repositorio](#)⁵).

Continuando la depuración de la base de datos utilizando el software Rstudio se

⁵https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/dataset_estandarizado.csv

procedió a eliminar información inconsistente en el conjunto de datos estandarizado, esta información se presentó en relación a la edad de los conductores de los diferentes tipos de vehículos involucrados en los siniestros de tránsito, de tal manera que existían registros que contenían datos de conductores con una edad no acorde a la establecida en las leyes vigentes en Ecuador. La eliminación de esta información se basó en el Art. 125 del Reglamento a Ley De Transporte Terrestre Tránsito y Seguridad Vial en donde se establece que “Ninguna persona podrá conducir vehículos a motor dentro del territorio nacional sin poseer los correspondientes títulos habilitantes otorgados por las autoridades competentes de tránsito, o un permiso de conducción, en el caso de menores adultos que hayan cumplido los 16 años de edad quienes deberán estar acompañados por un mayor de edad que posea licencia de conducir vigente, o algún documento expedido en el extranjero con validez en el Ecuador, en virtud de la ley, de tratados o acuerdos internacionales suscritos y ratificados por el Ecuador” [47], dicho permiso llamado “Permiso Menor Adulto” con el cual solo se autoriza la conducción de vehículos previstos en la licencia tipo B [48].

En la TABLA VII se detallan los registros que fueron eliminados en relación a las variables presentes en el conjunto de datos:

TABLA VII
VARIABLES PARA LA DEPURACIÓN DE LA BASE DE DATOS

PARTICIPANTE_1	TIPO_DE_VEHICULO_1	SERVICIO_1	EDAD_1
CONDUCTOR	AUTOMÓVIL CAMIONETA	PARTICULAR	<16
		CUENTA PROPIA	
		COMERCIAL	
		ESTADO	
		GOBIERNOS SECCIONALES	
		PÚBLICO	
	BUS		<18
	CAMIÓN		
	EMERGENCIAS		
	ESPECIAL		
	FURGONETA		
	MOTOCICLETA		

	NO IDENTIFICADO		
	VEHÍCULO		
	DEPORTIVO		
	UTILITARIO		

Para la lectura del conjunto de datos fue utilizada la librería “readr”, la cual es empleada para importación de archivos que contengan la información separada por comas cuya extensión es .csv (ver Fig. 11)

```
setwd("C:/Users/Yulissa Stefania/OneDrive/Escritorio/TESIS/Base de datos en R")
library(readr)
datos_siniestros <- read.csv("dataset_estandarizado.csv")
dep_dataset <- datos_siniestros[ !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='AUTOMOVIL' &
  (datos_siniestros$SERVICIO_1=='PARTICULAR' & datos_siniestros$SERVICIO_1=='CUENTA PROPIA')) &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='CAMIONETA' &
  (datos_siniestros$SERVICIO_1=='PARTICULAR' & datos_siniestros$SERVICIO_1=='CUENTA PROPIA')) &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='MOTOCICLETA') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='NO IDENTIFICADO') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='CAMION') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='VEHICULO DEPORTIVO UTILITARIO') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='BUS') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='FURGONETA') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='ESPECIAL') &
  (datos_siniestros$EDAD_1>16 & !datos_siniestros$EDAD_1==1)) &
  !(datos_siniestros$PARTICIPANTE_1=='CONDUCTOR' & datos_siniestros$TIPO_DE_VEHICULO_1=='EMERGENCIAS') &
  (datos_siniestros$EDAD_1>18 & !datos_siniestros$EDAD_1==1)),]
```

Fig. 11 Depuración de la base de datos

Se eliminaron los registros controlando que en la variable “PARTICIPANTE_1” los registros que tengan como valor “CONDUCTOR”, esto en combinación con la variable “TIPO_DE_VEHICULO_1” en la cual los registros con valor “AUTOMOVIL” y “CAMIONETA”, unido a que el valor de los registros de la variable “SERVICIO_1” sea “PARTICULAR” y “CUENTA PROPIA”, esto unido finalmente a que los registros de la variable “EDAD_1” sean valores mayores a 16, para los demás valores registrados en las variables “TIPO_DE_VEHICULO_1” y “SERVICIO_1” fueron eliminados los registros que contengan en la variable “EDAD_1” valores menores a 18, luego de ejecutar esta sentencia se obtuvo que, de los 16972 registros de siniestros de tránsito, un total de 32 registros fueron eliminados, los cuales representan el 0,19% del total de registros del conjunto de datos estandarizado, finalmente quedando un total de 16940 registros.

En la Fig. 12 se puede verificar en la sección “Global Environment” de Rstudio donde se muestra el número de registros, luego de ejecutar la sentencia ya mencionada ([Repositorio](#)⁶).

⁶https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/depuracion_dataset.R

Data		
• datos_siniestros	16972 obs. of 13 variables	
• dep_dataset	16940 obs. of 13 variables	

Fig. 12 Número de registros por la sentencia de depuración

Al finalizar la depuración de la base de datos, se exporta un nuevo archivo con formato CSV, remplazando al anterior, siendo nombrado como “dataset_dep.csv”.

6.2. Objetivo 2: Aplicar técnicas de minería de datos a la base de datos obtenida para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

En la realización del segundo objetivo se cumplió la cuarta fase de la metodología KDD especificada a continuación:

Fase IV: Aplicación de técnicas de minería de datos

Tarea 1: Extraer la información relevante para transformarla

Al finalizar la limpieza de la base de datos, se obtuvo un conjunto de datos estandarizado y depurado que contiene la información relevante al objeto de estudio, a este se procedió a transformarlo.

Transformación del conjunto de datos

En la TABLA VIII se muestran las dos variables que fueron transformadas, la principal razón de este cambio es debido a que afectan en la visualización de los gráficos resultantes al aplicar los algoritmos de minería de datos, ya que en sus registros almacenan datos muy extensos o contienen información que puede ser mejor visualizada y tratada a través de códigos.

TABLA VIII
LISTA DE VARIABLES A TRANSFORMAR

Variables a transformar
HORA
CAUSA_PROBABLE

Los registros de la variable “HORA” fueron transformados, tomando en cuenta el periodo de tiempo que tiene una hora, en la TABLA IX se muestran todos los valores que se reemplazaron en el conjunto de datos con respecto a la variable “HORA”:

TABLA IX
VALORES TRANSFORMADOS EN LA VARIABLE “HORA”

Valor original	Valor Actual
00:00:00 A 00:59:00	P00
01:00:00 A 01:59:00	P01
02:00:00 A 02:59:00	P02
03:00:00 A 03:59:00	P03
04:00:00 A 04:59:00	P04
05:00:00 A 05:59:00	P05
06:00:00 A 06:59:00	P06
07:00:00 A 07:59:00	P07
08:00:00 A 08:59:00	P08
09:00:00 A 09:59:00	P09
10:00:00 A 10:59:00	P10
11:00:00 A 11:59:00	P11
12:00:00 A 12:59:00	P12
13:00:00 A 13:59:00	P13
14:00:00 A 14:59:00	P14
15:00:00 A 15:59:00	P15
16:00:00 A 16:59:00	P16
17:00:00 A 17:59:00	P17
18:00:00 A 18:59:00	P18
19:00:00 A 19:59:00	P19
20:00:00 A 20:59:00	P20
21:00:00 A 21:59:00	P21
22:00:00 A 22:59:00	P22
23:00:00 A 23:59:00	P23

En la TABLA X se muestran los valores que se reemplazaron en los registros correspondientes a la variable “CAUSA_PROBABLE”:

TABLA X
VALORES TRANSFORMADOS EN LA VARIABLE “CAUSA_PROBABLE”

Valor original	Valor Actual
CASO FORTUITO O FUERZA MAYOR (EXPLOSIÓN DE	CP01

NEUMÁTICO NUEVO, DERRUMBE, INUNDACIÓN, CAÍDA DE PUENTE, ÁRBOL, PRESENCIA INTEMPESTIVA E IMPREVISTA DE SEMOVIENTES EN LA VÍA, ETC.).	
PRESENCIA DE AGENTES EXTERNOS EN LA VÍA (AGUA, ACEITE, PIEDRA, LASTRE, ESCOMBROS, MADEROS, ETC.).	CP02
CONDUCIR EN ESTADO DE SOMNOLENCIA O MALAS CONDICIONES FÍSICAS (SUENO, CANSANCIO Y FATIGA).	CP03
DAÑOS MECÁNICOS PREVISIBLES.	CP04
FALLA MECÁNICA EN LOS SISTEMAS Y/O NEUMÁTICOS (SISTEMA DE FRENS, DIRECCIÓN, ELECTRÓNICO O MECÁNICO).	CP05
CONDUCE BAJO LA INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS.	CP06
PEATÓN TRANSITA BAJO INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS.	CP07
PESO Y VOLUMEN – NO CUMPLIR CON LAS NORMAS DE SEGURIDAD NECESARIAS AL TRANSPORTAR CARGAS.	CP08
CONDUCIR VEHÍCULO SUPERANDO LOS LÍMITES MÁXIMOS DE VELOCIDAD.	CP09
CONDICIONES AMBIENTALES Y/O ATMOSFÉRICAS (NIEBLA, NEBLINA, GRANIZO, LLUVIA).	CP10
NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO AL VEHÍCULO QUE LE ANTECEDE.	CP11
NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDAD ENTRE VEHÍCULOS.	CP12
CONDUCIR DESATENTO A LAS CONDICIONES DE TRÁNSITO (CELULAR, PANTALLAS DE VIDEO, COMIDA, MAQUILLAJE O CUALQUIER OTRO ELEMENTO DISTRATOR).	CP13
DEJAR O RECOGER PASAJEROS EN LUGARES NO PERMITIDOS.	CP14

NO TRANSITAR POR LAS ACERAS O ZONAS DE SEGURIDAD DESTINADAS PARA EL EFECTO.	CP15
BAJARSE O SUBIRSE DE VEHÍCULOS EN MOVIMIENTO SIN TOMAR LAS PRECAUCIONES DEBIDAS.	CP16
CONDUCIR EN SENTIDO CONTRARIO A LA VÍA NORMAL DE CIRCULACIÓN.	CP17
REALIZAR CAMBIO BRUSCO O INDEBIDO DE CARRIL.	CP18
MAL ESTACIONADO – EL CONDUCTOR QUE DETENGA O ESTACIONE VEHÍCULOS EN SITIOS O ZONAS QUE ENTRAÑEN PELIGRO, TALES COMO ZONA DE SEGURIDAD, CURVAS, PUENTES, TÚNELES, PENDIENTES.	CP19
MALAS CONDICIONES DE LA VÍA Y/O CONFIGURACIÓN. (ILUMINACIÓN Y DISEÑO).	CP20
ADELANTAR O REBASAR A OTRO VEHÍCULO EN MOVIMIENTO EN ZONAS O SITIOS PELIGROSOS TALES COMO: CURVAS, PUENTES, TÚNELES, PENDIENTES, ETC.	CP21
NO RESPETAR LAS SEÑALES REGLAMENTARIAS DE TRÁNSITO. (PARE, CEDA EL PASO, LUZ ROJA DEL SEMÁFORO, ETC).	CP22
NO RESPETAR LAS SEÑALES MANUALES DEL AGENTE DE TRÁNSITO.	CP23
NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO A VEHÍCULOS.	CP24
NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO AL PEATÓN.	CP25
PEATÓN QUE CRUZA LA CALZADA SIN RESPETAR LA SEÑALIZACIÓN EXISTENTE (SEMÁFOROS O SEÑALES MANUALES).	CP26

Para transformar los registros de las variables del conjunto de datos antes mencionadas, se utilizó nuevamente la herramienta de software OpenRefine. La Fig. 13 muestra que para reemplazar los valores de los registros de la variable “HORA” se utilizó el comando `value.replace("valor actual", "valor nuevo")`.

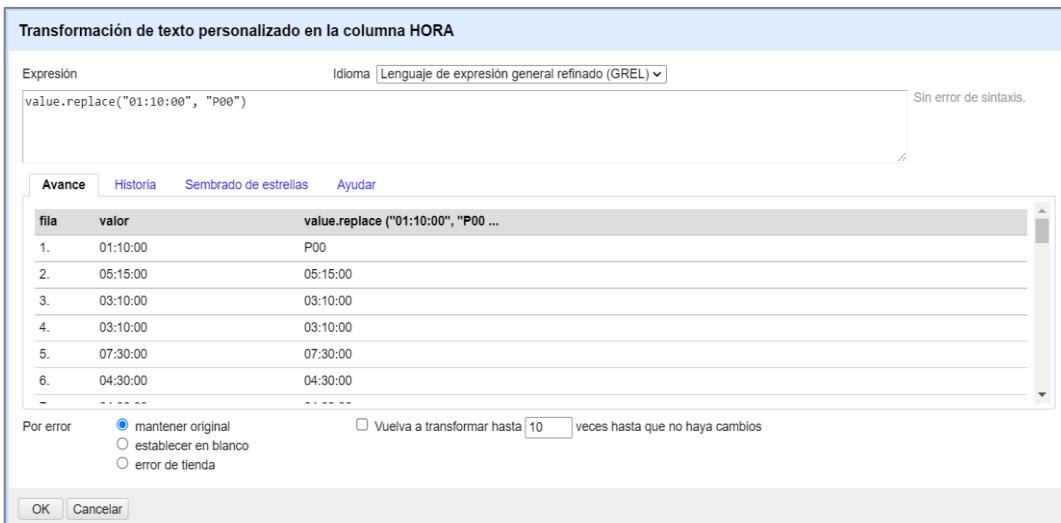


Fig. 13 Transformación de los registros de la variable “HORA”

Finalmente, se procedió a reemplazar los valores de los registros de la variable “CAUSA_PROBABLE”, utilizando nuevamente el comando `value.replace("valor actual", "valor nuevo")` tal como se muestra en la Fig. 14.

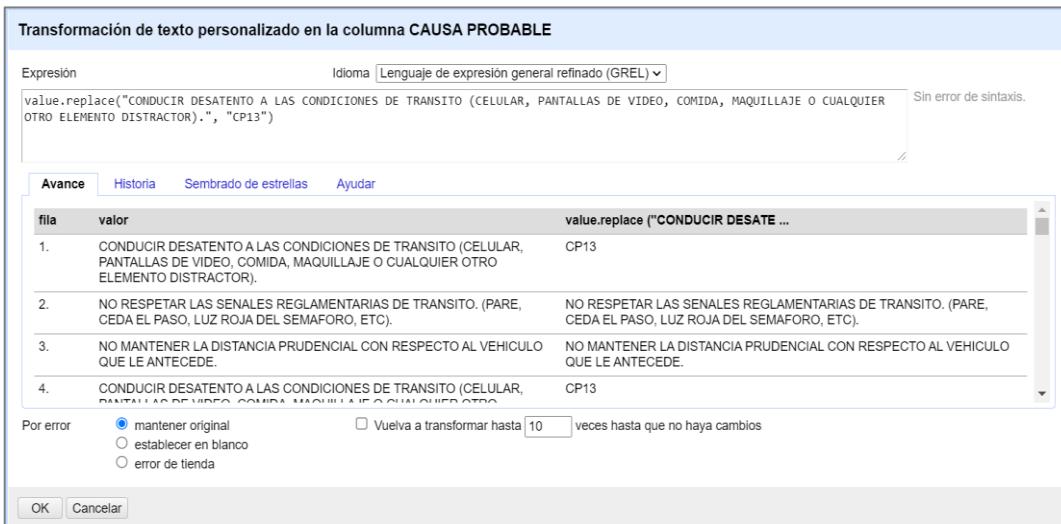


Fig. 14 Transformación de los registros de la variable “CAUSA_PROBABLE”

Al finalizar la transformación del conjunto de datos, se procedió a exportar un nuevo archivo con formato CSV, este archivo reemplazó al archivo anterior puesto que se encuentra depurado y transformado, y conservó el mismo nombre “dataset_dep.csv” (ver [Repositorio](#)⁷), esto para asegurar la compatibilidad con las herramientas SPSS Statistics, las cuales han sido seleccionadas para la aplicación de los algoritmos de minería de datos.

⁷https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/dataset_dep.csv

Análisis Exploratorio de datos

En esta sección se procedió a obtener el nuevo diccionario de datos (ver [Repository⁸](#)), después de la transformación del conjunto de datos, con el fin de entender el significado de cada valor reemplazado, en el Anexo 4 se presenta el diccionario de datos obtenido.

Para conocer la forma en la que se estructura la información, se procedió a generar un análisis exploratorio del conjunto de datos transformado, limitado a realizar deducciones directamente de los datos y parámetros obtenidos. Se explora los datos para poder comprender mejor la información contenida y poder aplicar correctamente los algoritmos de las técnicas de minería de datos seleccionada. A través del conjunto de datos analizado se deduce que en el año 2020 se registraron más siniestros de tránsito de clase final Choque Lateral con un total de 4851 ocurrencias correspondiente al 28,64% del total, además la causa probable por la cual se produjeron más siniestros fue la CP13 que corresponde a conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor) con un total de 5151 registros que es igual al 30,41% del total de siniestros registrados, mientras que la causa probable que generó menos siniestros fue la CP23 con solamente un único siniestro registrado del total de 16940. También se analizó que los automóviles fueron los tipos de vehículos más involucrados en los siniestros registrados con un total de 5358 registros correspondientes al 31,63% del total, el análisis exploratorio más detallado se puede visualizar en el Anexo 5.

Tarea 2: Aplicar las técnicas de árboles de decisión, redes neuronales artificiales y redes bayesianas para el análisis de la información almacenada.

Los técnicas seleccionados para la aplicación al conjunto de datos transformado, fueron establecidos en el Anteproyecto del presente TT, esto de acuerdo con los trabajos relacionados encontrados en los cuales según Hassinger [3] y Maldonado [9], destacan que las técnicas de minería de datos más utilizadas en el campo de la seguridad vial analizando siniestros de tránsito son los algoritmos predictivos de Árboles de Decisión, las Redes Neuronales Artificiales y las Redes Bayesianas, con los cuales se extrae información interesante, novedosa, que no se visualiza a simple vista y que es utilizada como soporte para la toma de decisiones.

⁸https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/diccionario_datos_siniestros_transito_2020_v2.0.pdf

Todo el proceso de la aplicación de dichos algoritmos predictivos de minería de datos se sintetiza en el diagrama de flujo expuesto en la Fig. 15:

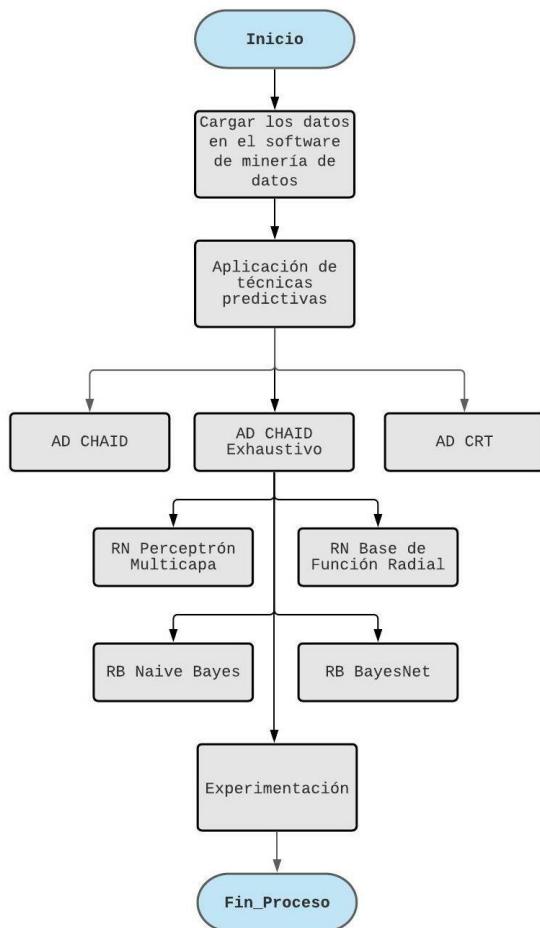


Fig. 15 Diagrama de flujo para la aplicación de las técnicas de minería de datos
 La Fig. 15 muestra que inicialmente se deben cargar los datos en el software de minería de datos para proceder a la aplicación de las técnicas predictivas seleccionadas, aplicando tres tipos de algoritmos de árboles de decisión: CHAID, CHAID Exhaustivo y CRT; dos tipos de algoritmos de redes neuronales: Perceptrón multicapa y de Función de Base Radial; y por último dos algoritmos de Redes Bayesianas: Naive Bayes y BayesNet (ver [Repositorio⁹](#)).

A continuación, se detalla más a profundidad cada una de las etapas del proceso de aplicación de las técnicas de minería de datos.

Cargar los datos en el software de minería de datos

El conjunto de datos “dataset_dep.csv” fue cargado en las herramientas elegidas, utilizando la herramienta SPSS Statistics para la aplicación de los algoritmos de

⁹https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/modelos

Árboles de Decisión y Redes Neuronales, esto debido a que están disponibles dentro de los múltiples algoritmos de minería de datos que esta herramienta ofrece, sin embargo SPSS Statistics no brinda la aplicación de los algoritmos de Redes Bayesianas, debido a esto dichos algoritmos fueron aplicados en la herramienta Weka, es decir en SPSS Statistics se aplican los siete algoritmos predictivos mostrados en la TABLA XI y en Weka los dos algoritmos predictivos que se visualizan en la TABLA XII.

TABLA XI

ALGORITMOS APLICADOS MEDIANTE LA HERRAMIENTA SPSS STATISTICS

Nº	Algoritmos
Árboles de Decisión	
1	CHAID
2	CHAID Exhaustivo
3	CRT
Redes Neuronales	
4	Perceptrón Multicapa
5	Función de Base Radial

TABLA XII

ALGORITMOS APLICADOS MEDIANTE LA HERRAMIENTA WEKA

Nº	Algoritmos
Redes Bayesianas	
1	Naïve Bayes
2	BayesNet

Para la aplicación de los algoritmos de árboles de decisión y redes neuronales en la herramienta SPSS Statistics, se importa el conjunto de datos en archivo .csv con la data de todas las variables, este proceso mostrado a continuación:

Al ejecutar SPSS Statistics se muestra una ventana como se ve en la Fig. 16.

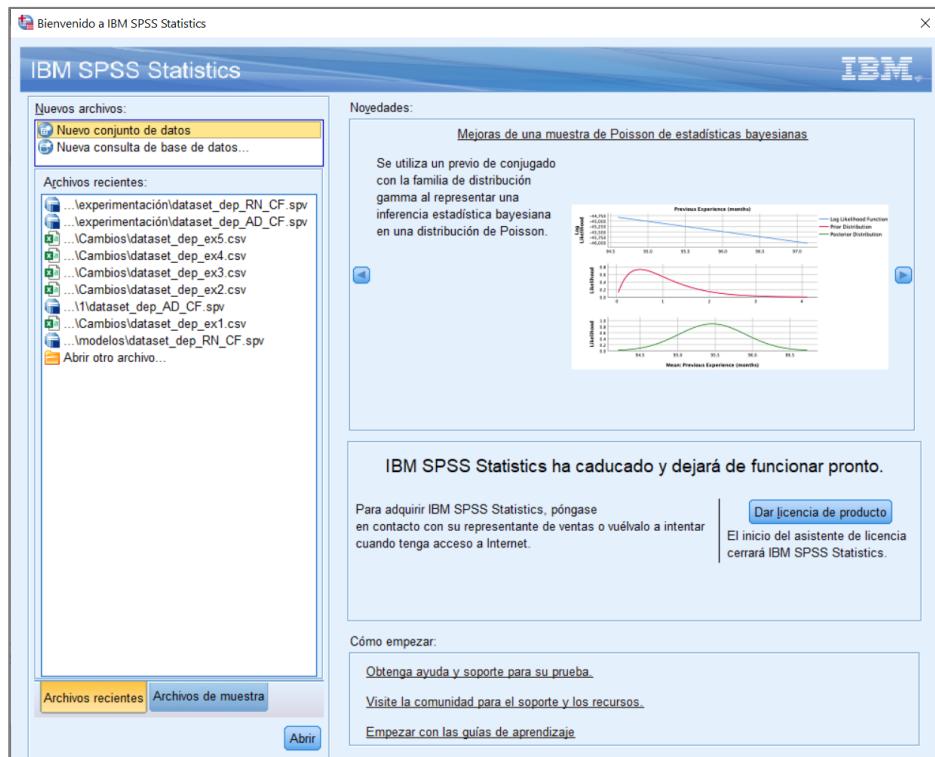


Fig. 16 Interfaz principal de SPSS Statistics

Dentro de la interfaz, se selecciona “Nuevo conjunto de datos”, con lo cual se despliega una nueva ventana, en la que se selecciona la opción “Archivo”, lo que permite visualizar “Importar datos” y seleccionarla, seguidamente se debe escoger la opción de “Datos CSV” tal como se muestra en la Fig. 17, esto permitirá desplegar una nueva ventada que permite configurar la lectura el archivo CSV para ejecutarlo.

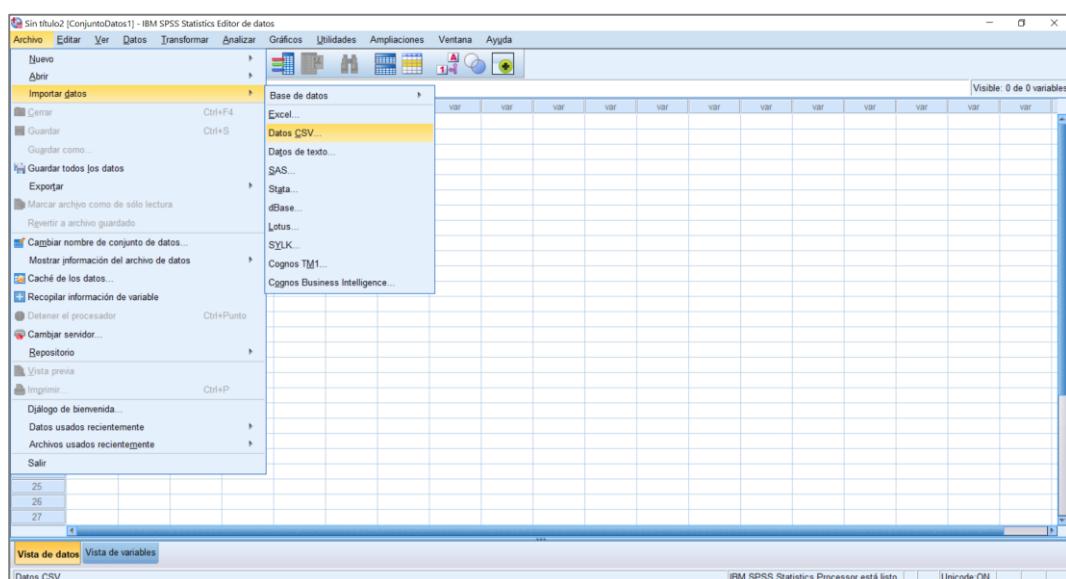


Fig. 17 Interfaz de SPSS Statistics para subir archivos a procesar.

En la interfaz mostrada en la Fig. 18 se configura el archivo CSV importado, de acuerdo a las características que presenta el mismo, tales como el delimitador entre valores, símbolo decimal, el calificador de texto, etc.

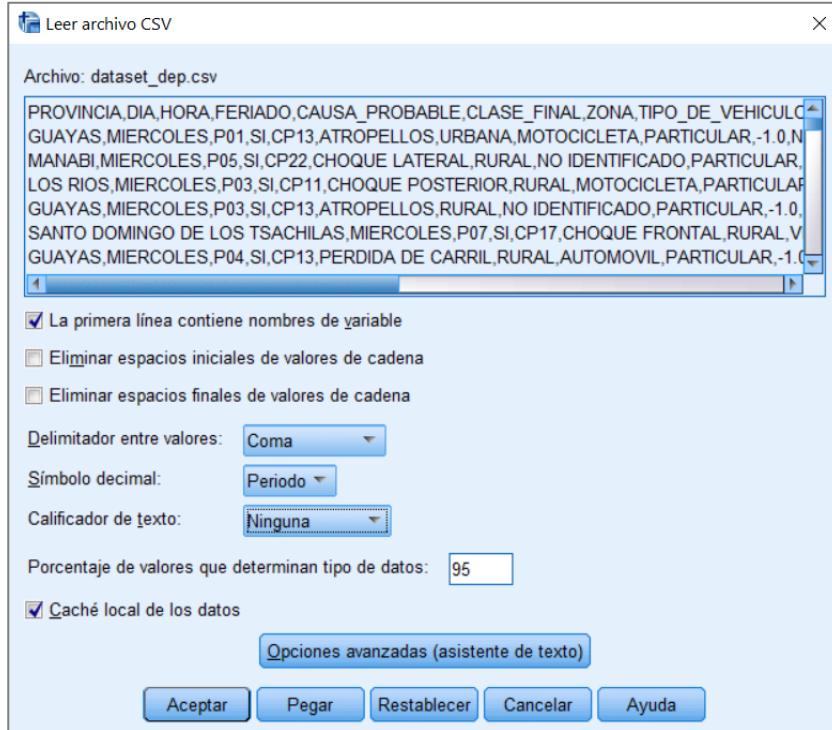


Fig. 18 Configurar la lectura del archivo CSV en SPSS Statistics

La Fig. 19 muestra el conjunto de datos cargado en la herramienta SPSS Statistics apto para aplicación de los algoritmos de Árboles de Decisión y Redes Neuronales.

	PROVINCIA	DIA	HORA	FERIADO	CAUSA_PROBABLE	CLASE_FINAL	ZONA	TIPO_DE_VEHICULO_1	SERVICIO_1	EDAD_1	SEXO_1	CONDICION_1	PARTICIPANTE_1
1	GUAYAS	MIERCOLES	P01	SI	CP13	ATROPELLOS	URBANA	MOTOCICLETA	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
2	MANABI	MIERCOLES	P05	SI	CP22	CHOQUE LATERAL	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	LESIONADO	PASAJERO
3	LOS RIOS	MIERCOLES	P03	SI	CP11	CHOQUE POSTE	RURAL	MOTOCICLETA	PARTICULAR	-1.0	NO IDENTIFICADO	LESIONADO	CONDUCTOR
4	GUAYAS	MIERCOLES	P03	SI	CP13	ATROPELLOS	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
5	SANTO DOMI.	MIERCOLES	P07	SI	CP17	CHOQUE FRONTAL	RURAL	VEHICULO DEPORTIVO ...	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
6	GUAYAS	MIERCOLES	P04	SI	CP13	PERDIDA DE CAR.	RURAL	AUTOMOVIL	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
7	GUAYAS	MIERCOLES	P04	SI	CP17	CHOQUE FRONTAL	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
8	GUAYAS	MIERCOLES	P07	SI	CP12	ROZAMIENTOS	RURAL	BUS	PUBLICO	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
9	GUAYAS	MIERCOLES	P07	SI	CP13	ESTRELLAMIENTO	URBANA	MOTOCICLETA	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
10	GUAYAS	MIERCOLES	P09	SI	CP13	ATROPELLOS	URBANA	AUTOMOVIL	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
11	SANTO DOMI.	MIERCOLES	P04	SI	CP13	PERDIDA DE CAR.	RURAL	CAMION	PARTICULAR	50.0	HOMBRE	LESIONADO	CONDUCTOR
12	EL ORO	MIERCOLES	P07	SI	CP13	PERDIDA DE CAR.	RURAL	MOTOCICLETA	PARTICULAR	48.0	HOMBRE	LESIONADO	CONDUCTOR
13	GUAYAS	MIERCOLES	P10	SI	CP17	CHOQUE FRONTAL	RURAL	CAMIONETA	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
14	GUAYAS	MIERCOLES	P12	SI	CP16	CAIDA DE PASAJ.	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
15	GUAYAS	MIERCOLES	P02	SI	CP13	ARROLLAIMIENTOS	URBANA	AUTOMOVIL	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
16	AZUAY	MIERCOLES	P03	SI	CP13	PERDIDA DE CAR.	RURAL	MOTOCICLETA	PARTICULAR	-1.0	HOMBRE	FALLECIDO	CONDUCTOR
17	GUAYAS	MIERCOLES	P13	SI	CP11	CHOQUE POSTE	URBANA	AUTOMOVIL	PARTICULAR	48.0	HOMBRE	ILESO	CONDUCTOR
18	LOS RIOS	MIERCOLES	P08	SI	CP13	PERDIDA DE CAR.	RURAL	AUTOMOVIL	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
19	GUAYAS	MIERCOLES	P01	SI	CP13	ESTRELLAMIENTO	RURAL	CAMIONETA	PARTICULAR	68.0	HOMBRE	ILESO	CONDUCTOR
20	SANTO DOMI.	MIERCOLES	P14	SI	CP11	CHOQUE POSTE	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
21	SANTA ELENA	MIERCOLES	P09	SI	CP13	PERDIDA DE CAR.	RURAL	AUTOMOVIL	PARTICULAR	27.0	HOMBRE	FALLECIDO	CONDUCTOR
22	GUAYAS	MIERCOLES	P17	SI	CP13	ATROPELLOS	URBANA	AUTOMOVIL	PARTICULAR	54.0	HOMBRE	ILESO	CONDUCTOR
23	SANTA ELENA	MIERCOLES	P09	SI	CP17	CHOQUE FRONTAL	RURAL	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
24	GUAYAS	MIERCOLES	P03	SI	CP13	ATROPELLOS	URBANA	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
25	SANTA ELENA	MIERCOLES	P09	SI	CP11	CHOQUE POSTE	URBANA	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR
26	AZUAY	MIERCOLES	P15	SI	CP17	CHOQUE FRONTAL	RURAL	CAMIONETA	PARTICULAR	-1.0	NO IDENTIFICADO	ILESO	CONDUCTOR
27	GUAYAS	MIERCOLES	P19	SI	CP13	ATROPELLOS	URBANA	NO IDENTIFICADO	PARTICULAR	-1.0	NO IDENTIFICADO	NO IDENTIFICADO	CONDUCTOR

Fig. 19 Conjunto de datos cargado en SPSS Statistics

Por el contrario, para la aplicación de los algoritmos de redes bayesianas se carga el conjunto de datos en la herramienta Weka, en la cual, en el proceso de ejecutarla se despliega la interfaz principal mostrada en la Fig. 20.



Fig. 20 Interfaz principal de Weka

En la interfaz mostrada en la se selecciona en el ícono “Explorar”, desplegándose una nueva interfaz, en la cual se selecciona en la parte superior izquierda “Open File”, esto permite subir el archivo CSV para procesarlo.

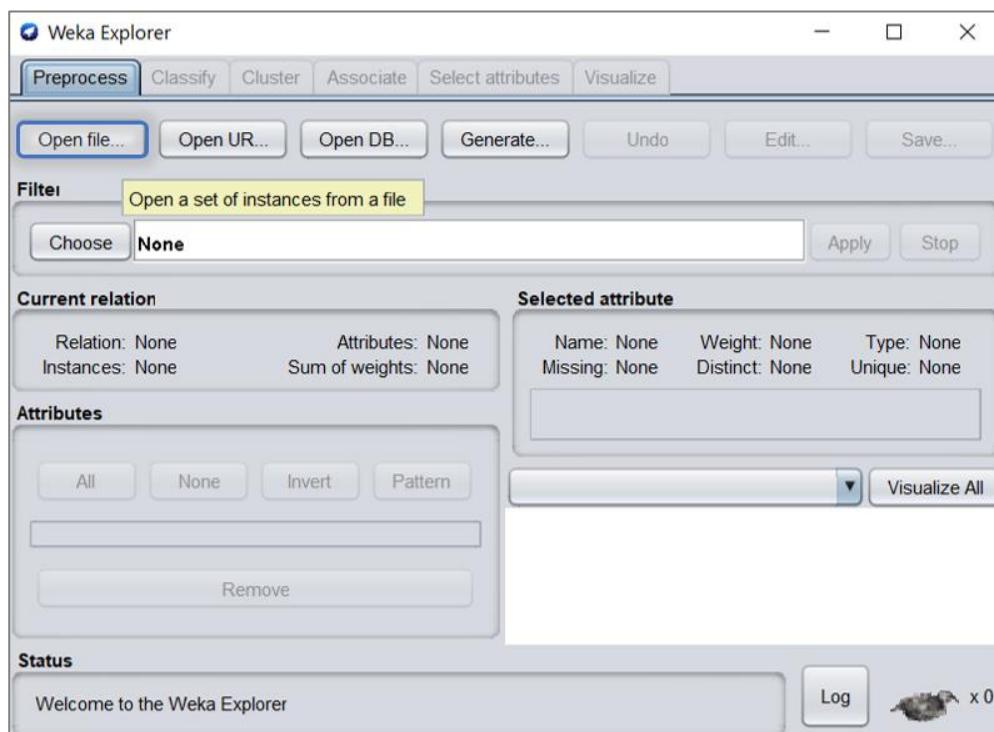


Fig. 21 Interfaz de Weka para subir archivos a procesar

En la Fig. 22 se muestra el conjunto de datos cargado en la herramienta Weka, en la cual se indica las características a procesar en la herramienta, y con las mismas se procede a la aplicación de los algoritmos de Redes Bayesianas.

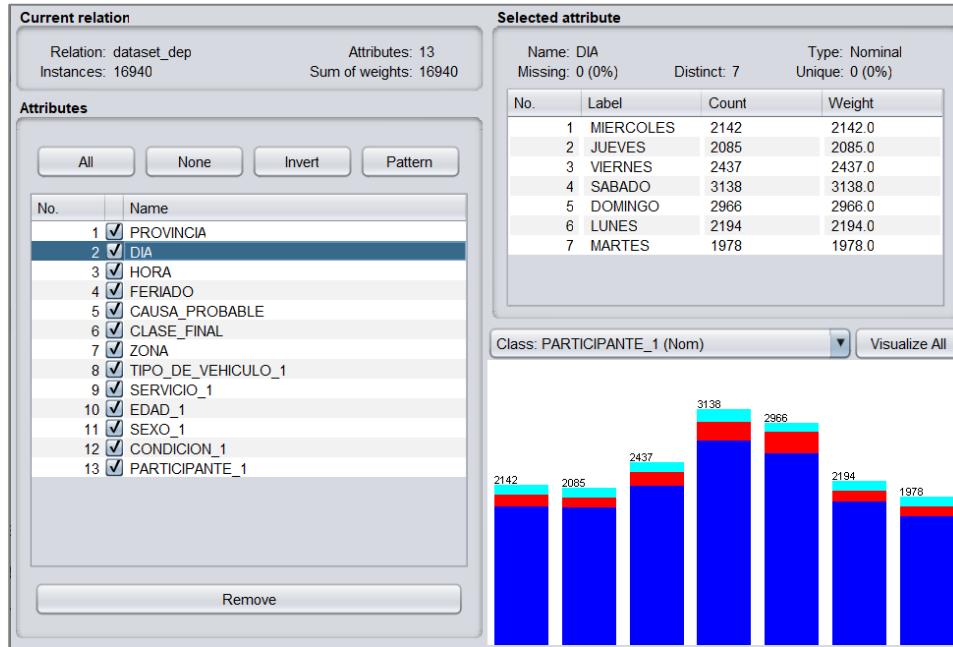


Fig. 22 Conjunto de datos cargado en Weka

Aplicación de técnicas predictivas

Las técnicas predictivas seleccionadas fueron establecidas en el Anteproyecto del presente TT (ver **¡Error! No se encuentra el origen de la referencia.**) , esto de acuerdo con los trabajos relacionados encontrados en los cuales según Hassinger [3] y Maldonado [9], destacan que las técnicas de minería de datos más utilizadas en el campo de la seguridad vial analizando siniestros de tránsito son los Árboles de Decisión, las Redes Neuronales Artificiales y las Redes Bayesianas

Es por lo antes mencionado que se aplicaron tres tipos de algoritmos de árboles de decisión: CHAID, CHAID Exhaustivo y CRT; dos tipos de algoritmos de redes neuronales: Perceptrón multicapa y de Función de Base Radial; y por último los algoritmos de Redes Bayesianas: Naive Bayes y BayesNet (ver [Repositorio¹⁰](#)).

A continuación, se detalla la aplicación de los algoritmos predictivos a nivel de cada herramienta.

¹⁰https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/modelos

Árbol de decisión CHAID

En la herramienta SPSS Statistics, ya cargado el conjunto de datos, a través de la opción “Analizar” → “Clasificar” → “Árbol” (ver Fig. 23), se aplicó el árbol de clasificación CHAID.

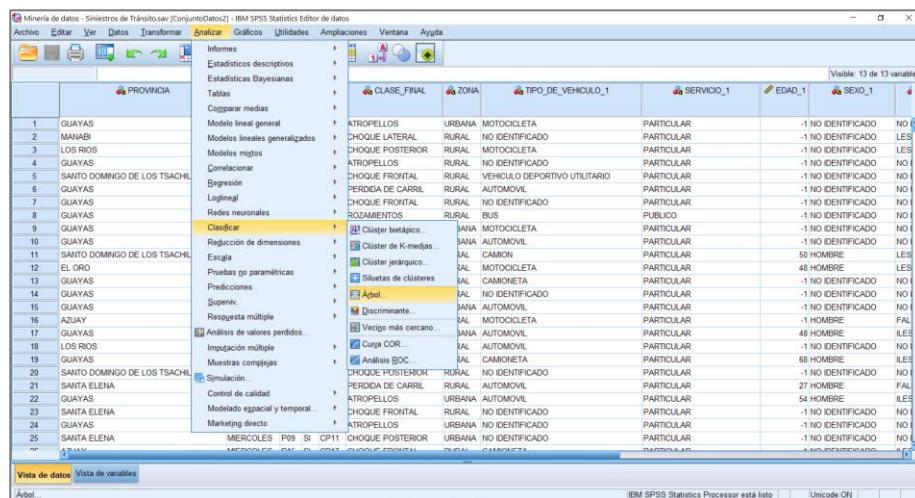


Fig. 23 Crear árboles de decisión en SPSS Statistics

Al configurar el AD CHAID se especifica la variable “CLASE_FINAL” como variable dependiente y el resto de variables como independientes tal como se muestra en la Fig. 24.

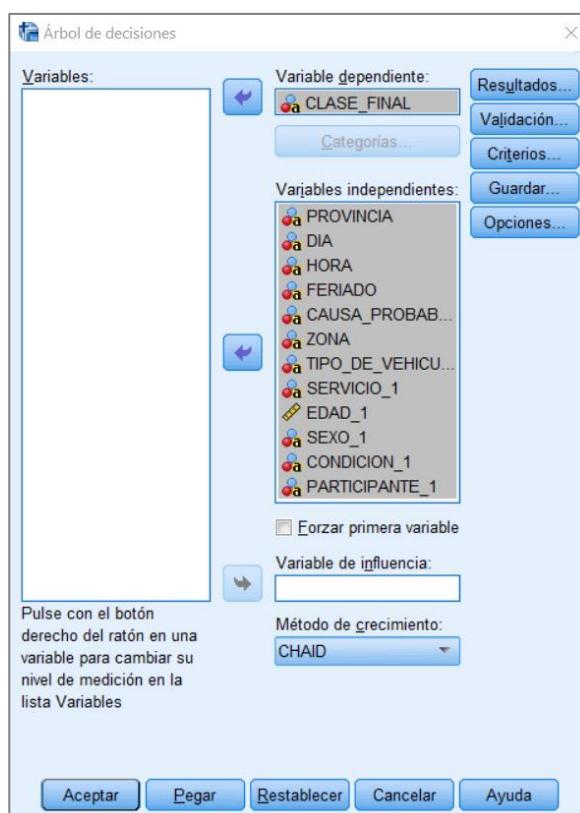


Fig. 24 Configuración de variables independientes y dependiente en CHAID

Además se realizó las configuraciones de los criterios del algoritmo mostradas en la Fig. 25, dando el valor de 0,05 y 0,001 a la división de nodos y el cambio mínimo de frecuencias respectivamente, también se estableció un valor de 100 para el número máximo de iteraciones, esto para controlar el crecimiento del árbol hasta haber alcanzado el número de iteraciones y se seleccionó en el estadístico de chi-cuadrado la opción de Pearson, ya que esta brinda cálculos más rápidos que la de Razón de verosimilitud [49]. Al finalizar se establece el criterio de ajuste de valores que permite corregir los valores de los criterios de división de nodos y fusión de categorías en caso de ser necesario.

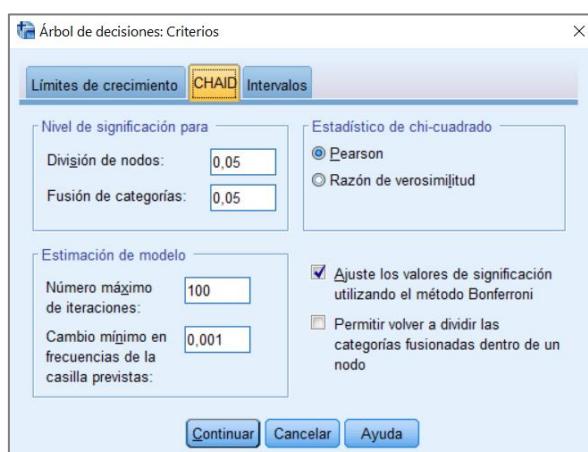


Fig. 25 Configuración de criterios del algoritmo CHAID

El árbol CHAID (ver [Repositorio](#)¹¹) obtenido cuenta con un porcentaje global de clasificación del 58,08% y 44,57% de precisión global, además con una profundidad máxima de 5 nodos, con 174 nodos de los cuales 117 son nodos terminales.

Árbol de decisión CHAID Exhaustivo

La aplicación del árbol CHAID Exhaustivo, comenzó de la misma forma que el algoritmo anterior (ver Fig. 23), especificando la variable “CLASE_FINAL” como variable dependiente y las demás asignadas como variables independientes, esto se observa en la Fig. 26.

¹¹https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/diagrama/arbol_de_clasificacion_CHAID.pdf

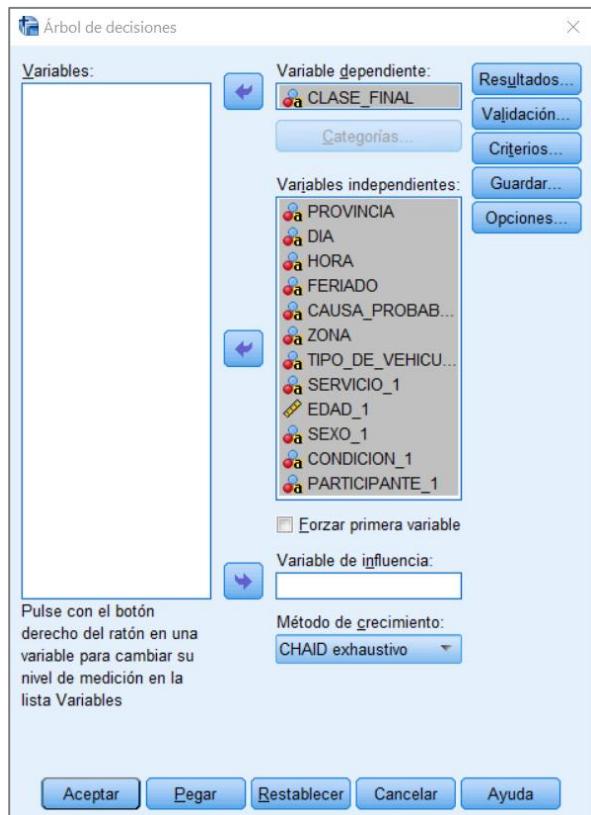


Fig. 26 Configuración de variables independientes y dependiente en CHAID
Exhaustivo

Seguidamente fueron configurados los criterios del algoritmo como se muestra en la Fig. 27, estos criterios son los mismos aplicados al algoritmo CHAID.

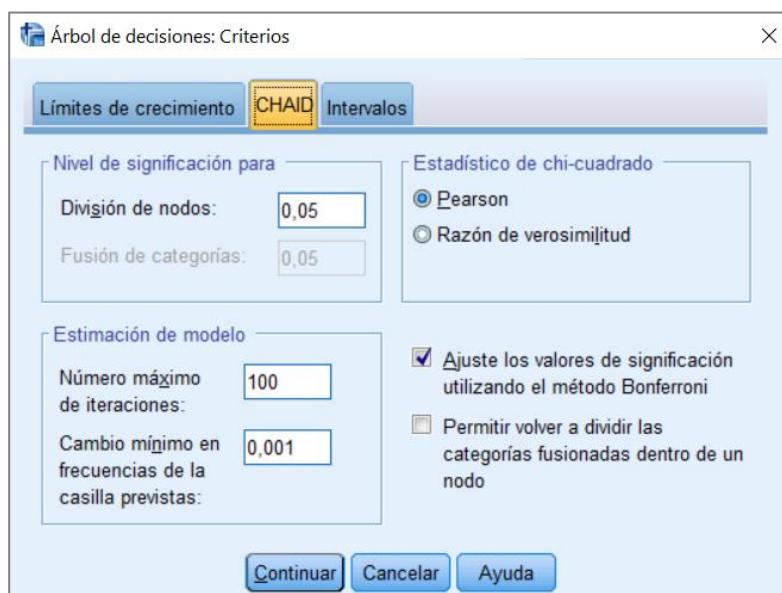


Fig. 27 Configuración de criterios del algoritmo CHAID Exhaustivo

El árbol CHAID Exhaustivo (ver [Repositorio](#)¹²) generado cuenta con un porcentaje global de clasificación del 58,38% y con 44,60% de precisión global, con una profundidad máxima de 5 nodos, con 216 nodos de los cuales 148 son nodos terminales.

Árbol de decisión CRT

La ejecución del presente algoritmo dentro de la herramienta SPSS Statistics (ver Fig. 23), se inició con la especificación de la variable dependiente “CLASE_FINAL” y las otras variables especificadas como variables independientes, en la Fig. 28 se presenta lo explicado.



Fig. 28 Configuración de variables independientes y dependiente en CRT

Seguidamente fueron configurados los criterios del algoritmo, tal como se muestran en la Fig. 29, seleccionando la opción Gini basada en el cuadrado de las probabilidades de pertenencia de cada categoría de la variable dependiente [50], también se estableció un valor de 0,0001 para el cambio mínimo en la mejora, esto necesario para la división de un nodo.

¹²https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/diagrama/arbol_de_clasificacion_CHAID_Exhaustivo.pdf

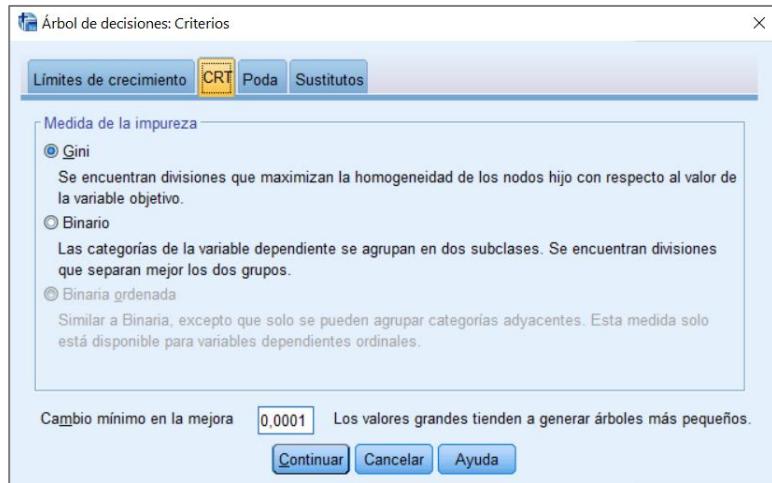


Fig. 29 Configuración de criterios del algoritmo CRT

Como resultado se obtuvo el árbol CRT (ver [Repositorio](#)¹³) el cual cuenta con un porcentaje global de clasificación del 45,33% y un porcentaje de precisión global del 28,38% con una profundidad de 5 nodos, con 25 nodos generados de los cuales 15 son nodos terminales.

Red Neuronal Perceptrón Multicapa

Para este algoritmo, como ya se tiene cargado previamente el conjunto de datos (ver Fig. 19) mediante la selección de la opción “Analizar” → “Redes Neuronales” → “Perceptrón Multicapa”, se aplicó la red neuronal Perceptrón Multicapa, tal como se muestra en la Fig. 30.

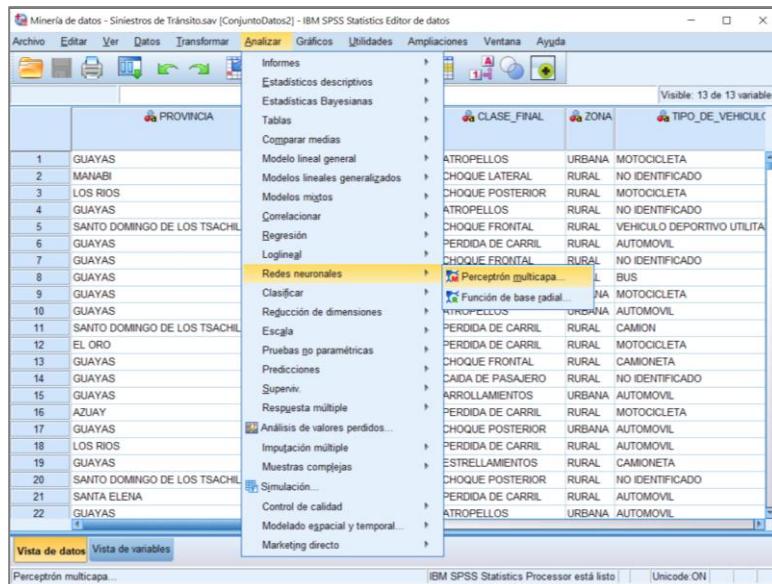


Fig. 30 Crear RN Perceptrón Multicapa en SPSS Statistics

¹³https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/diagrama/arbol_de_clasificacion_CRT.pdf

Dicha aplicación, inició especificando la variable “CLASE_FINAL” como variable dependiente y el resto de variables especificadas como factores, esto se muestra en la Fig. 31.

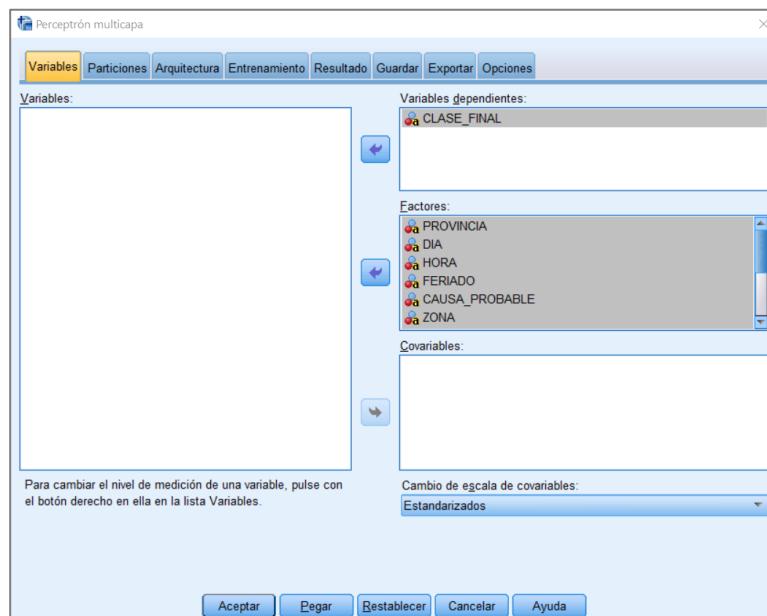


Fig. 31 Configuración de los factores y de la variable dependiente en RN
Perceptrón Multicapa

También fue configurado el método de crear particiones al conjunto de datos utilizado, se asignó el 80% a la muestra de entrenamiento, esto con el propósito de entrenar a la red neuronal y una muestra del 20% para pruebas, tal como se muestra en la Fig. 32.

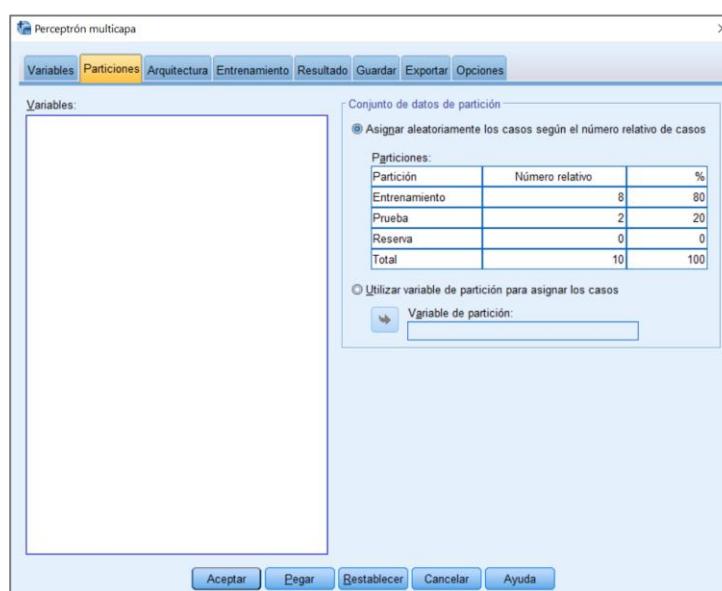


Fig. 32 Configuración de las particiones en la RN Perceptrón Multicapa

La RN Perceptrón Multicapa (ver [Repositorio](#)¹⁴) que fue generada cuenta con un porcentaje global de clasificación del 57,33% y un 44,52% de precisión global.

Red Neuronal Función de Base Radial

Para la ejecución del presente algoritmo en la herramienta SPSS Statistics, mediante la selección de la opción “Analizar” → “Redes Neuronales” → “Perceptrón Multicapa”, se aplicó la red neuronal de Función de Base Radial, esto mostrado en la Fig. 33.

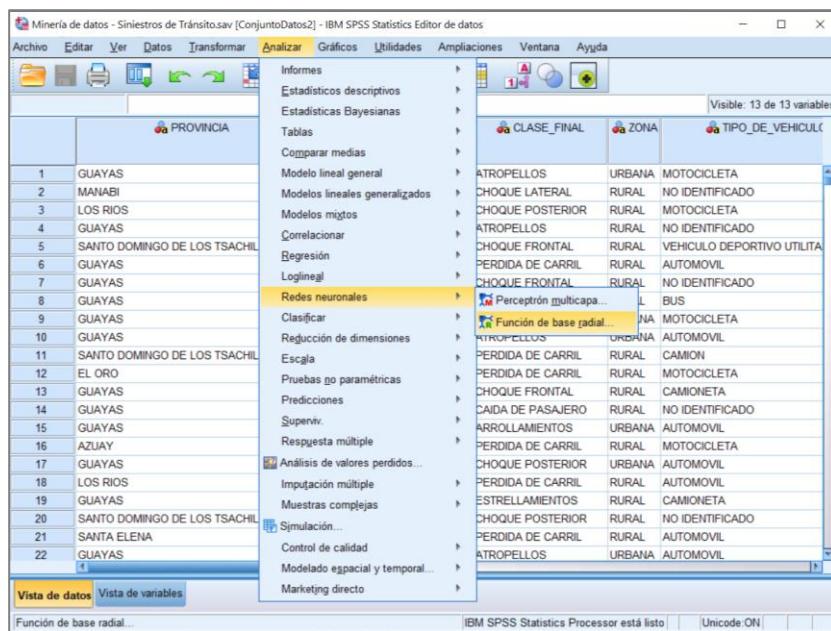


Fig. 33 Crear RN Función de Base Radial en SPSS Statistics

Al configurar la RN Función de Base Radial, primeramente, se especificó la variable dependiente “CLASE_FINAL” y se asignó al resto de variables especificadas como factores, en la Fig. 34 se presenta lo explicado.

¹⁴https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/diagrama/red_neuronal_Perceptron_Multicapa.pdf

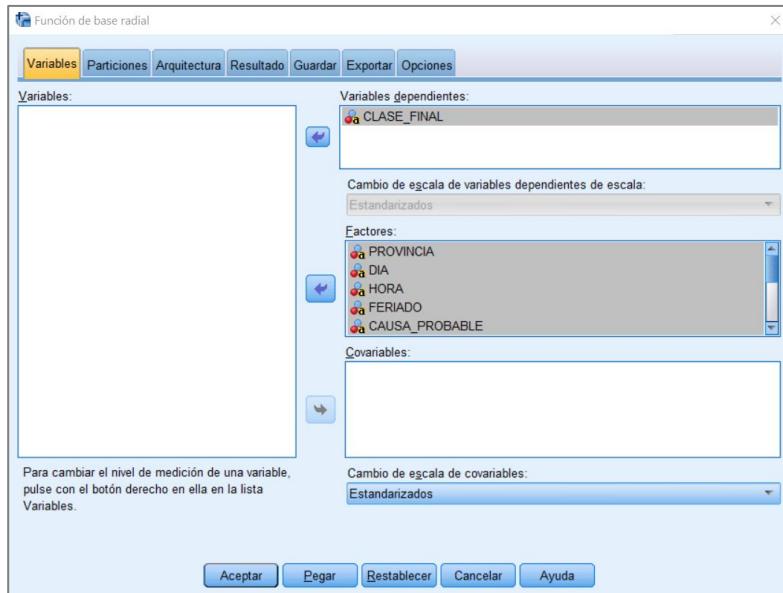


Fig. 34 Configuración de los factores y de la variable dependiente en RN Función de Base Radial

Al igual que la RN Perceptrón Multicapa fue configurado el método de crear particiones al conjunto de datos utilizado, asignando los mismos porcentajes a la muestra de entrenamiento y de prueba, dándoles un valor del 80% y 20% respectivamente, esto mostrado en la Fig. 35.

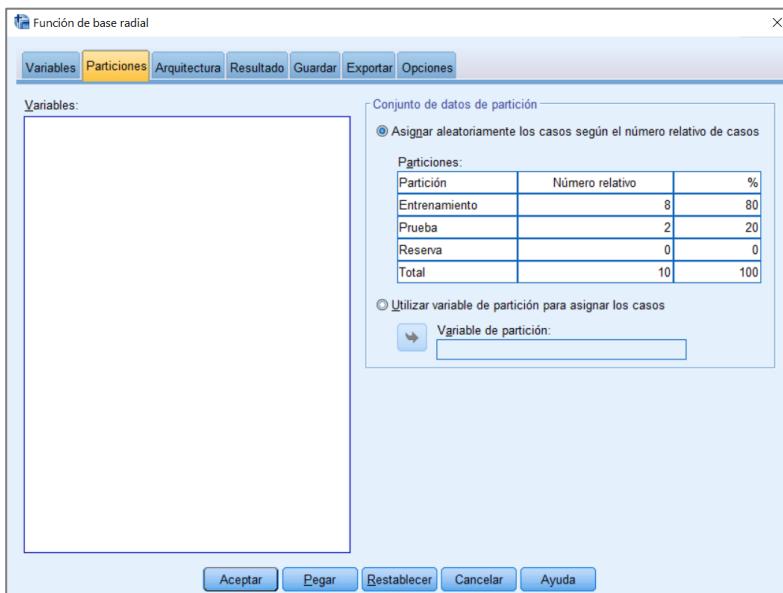


Fig. 35 Configuración de las particiones en la RN Función de Base Radial

El algoritmo generó una RN Función de Base Radial (ver [Repositorio](#)¹⁵) cuenta con un porcentaje global de clasificación del 43,38% y con 20,48% de precisión global.

¹⁵https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/diagrama/red_neuronal_Funcion_de_Base_Radial.pdf

Red Bayesiana Naive Bayes

En Weka, cargado el conjunto de datos (ver Fig. 22), se procedió a ingresar a la opción “Classify”, en donde se eligió el algoritmo Naive Bayes, configurando la variable “CLASE_FINAL” como la variable objetivo, y utilizando el conjunto de entrenamiento como opción de prueba (ver Fig. 36), con la cual se entrena al algoritmo con todos los datos disponibles y luego lo aplica otra vez sobre los mismos [51].

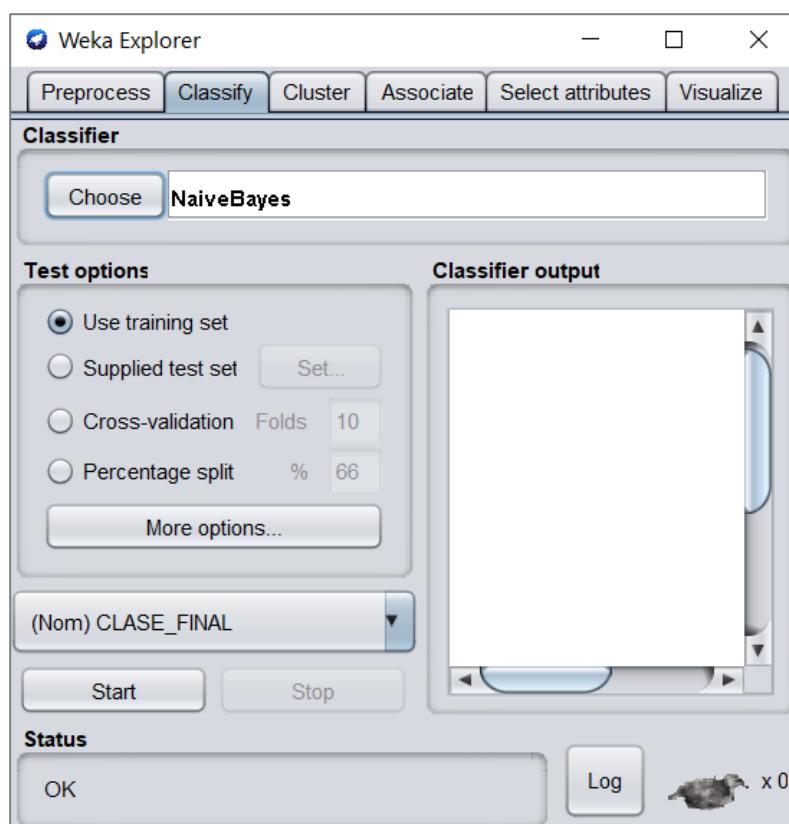


Fig. 36 Configuración de la RB Naive Bayes

La RB Naive Bayes (ver [Repositorio¹⁶](#)) que fue generada cuenta con un porcentaje global de clasificación del 55,27% y un 44,25% de precisión global. También se realizaron pruebas seleccionando la opción de validación cruzada, con la cual se obtuvo un porcentaje menor al seleccionar el uso de un conjunto de prueba.

¹⁶https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/modelos/dataset_dep_RB1_CF.model

Red Bayesiana BayesNet

La aplicación del presente algoritmo procedió inicialmente seleccionando la opción “Classify” y escogiendo el algoritmo BayesNet, seguido se especificó la variable dependiente “CLASE_FINAL” y se asignó al resto de variables especificadas como factores, tal como se muestra en la Fig. 37.

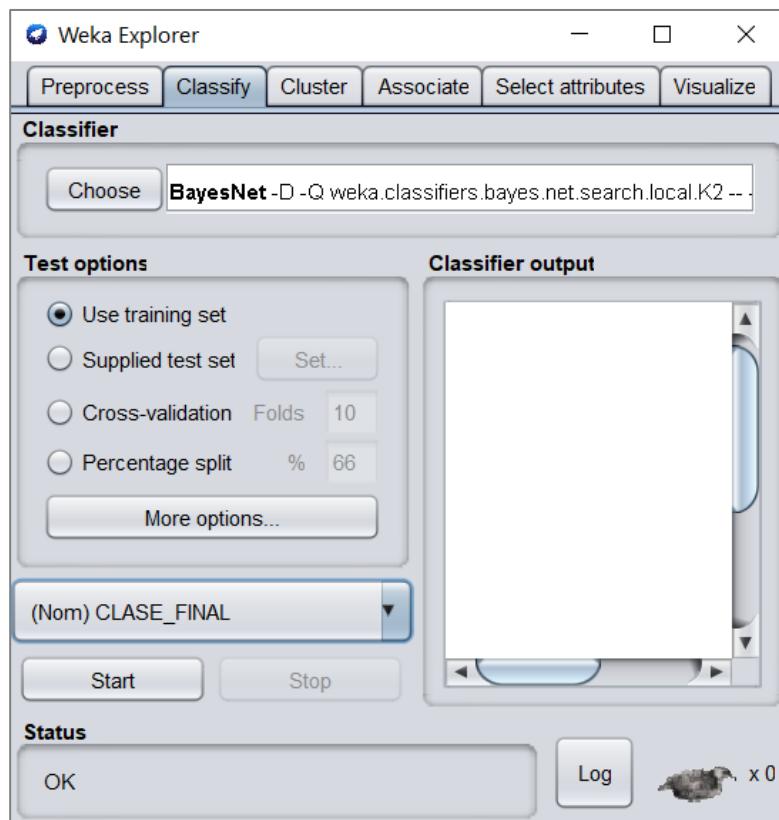


Fig. 37 Configuración de la RB BayesNet

El algoritmo generó una RB BayesNet (ver [Repositorio](#)¹⁷), la cual cuenta con un porcentaje global de clasificación del 55,42% y con un 44,34% de precisión global.

Experimentación

La aplicación de los algoritmos predictivos de minería de datos seleccionados pasaron por un proceso de experimentación (ver [Repositorio](#)¹⁸), esto para poder obtener los mejores resultados con respecto a la aplicación de cada uno de los algoritmos antes mencionados, esto a través del tratamiento del conjunto de datos. Como primera instancia se crearon trece subconjuntos de la base de datos ya depurada, con la finalidad de que cada nuevo conjunto de datos generado abarque

¹⁷https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/modelos/dataset_dep_RB2_CF.model

¹⁸https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion

la información orientada a las trece categorías existentes en la variable “CLASE_FINAL” en donde se especifica que clases de siniestros de tránsitos fueron los ocurridos. Dichos subconjuntos están estructurados de acuerdo a la TABLA XIII.

TABLA XIII
SUBCONJUNTOS CREADOS

Nº	Descripción	Número de registros
1	Choque Lateral	4851
2	Estrellamientos	2137
3	Atropellos	2001
4	Pérdida de Pista	1551
5	Choque Posterior	1320
6	Arrollamientos	1118
7	Pérdida de Carril	1037
8	Choque Frontal	877
9	Rozamientos	734
10	Caída de Pasajero	363
11	Colisión	341
12	Volcamientos	244
13	Otros	366

Los resultados de esta experimentación no fueron los más óptimos (ver Anexo 6), debido a que algunos de los nuevos subconjuntos generados contenían un número bajo de registros, aunque si bien es cierto no existe un número en particular como regla definido para efectuar técnicas de minería de datos, es necesario contar con varios registros los cuales ayudan a conocer la realidad de mejor manera y con mayor exactitud, en la Fig. 38 se observa que varios subconjuntos de datos generados cuentan con un numero de registros menor a 400 observaciones.



Fig. 38 Número de registros por cada subconjunto generado

Además, se recalca que al momento de aplicar los algoritmos predictivos a los subconjuntos generados se definió a la variable “CAUSA_PROBABLE” como variable objetivo, lo cual no aporta al cumplimiento del objetivo del presente TT, ya que la aplicación de la minería de datos no se orienta específicamente a las clases de siniestros de tránsito debido a que la variable “CLASE_FINAL” que contiene dichas clases, no fue configurada como la variable objetivo.

Por otra parte, con el fin de incrementar los porcentajes resultantes de la aplicación de los algoritmos predictivos al conjunto de datos depurado inicialmente, se procedió a experimentar con las variables contenidas en dicho conjunto de datos, probando que si al eliminar variables específicas del conjunto de datos sea posible el incremento de dichos porcentajes. De tal manera que se probó cinco veces, esto a través de la eliminación de una, dos, tres y cuatro variables, destacando que en este experimento la variable “CLASE_FINAL” fue establecida con la variable objetivo.

En la primera prueba, al conjunto de datos que contenía las variables mostradas en la TABLA IV, se le eliminó la variable “HORA”, quedando el conjunto de datos con las doce variables que se muestran en la TABLA XIV.

TABLA XIV

VARIABLES RESTANTES DEL CONJUNTO DE DATOS – PRUEBA 1

Nº	Variables Seleccionadas
1	PROVINCIA
2	DÍA

3	FERIADO
4	CAUSA PROBABLE
5	CLASE FINAL
6	ZONA
7	TIPO DE VEHÍCULO 1
8	SERVICIO 1
9	EDAD 1
10	SEXO 1
11	CONDICIÓN 1
12	PARTICIPANTE 1

Los resultados con respecto al porcentaje global de instancias clasificadas correctamente y el porcentaje de precisión global de la aplicación de los algoritmos correspondientes (ver [Repositorio](#)¹⁹), se muestran ordenados de manera descendente en la TABLA XV y TABLA XVI.

TABLA XV

RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN CORRECTA - PRUEBA 1

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	58.10%
2	Árbol de Decisión CHAID	58.08%
3	Red Neuronal Perceptrón Multicapa	57,80%
4	Naïve Bayes	55,14%
5	BayesNet	55,08%
6	Árbol de Decisión CRT	45,30%
7	Red Neuronal Base de Función Radial	41.90%

TABLA XVI

RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE PRECISIÓN – PRUEBA 1

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,47%

¹⁹https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion/1_una %20variable

2	Red Neuronal Perceptrón Multicapa	44,43%
3	Naive Bayes	44,40%
4	BayesNet	44,25%
5	Árbol de Decisión CHAID	44,23%
6	Árbol de Decisión CRT	28,38%
7	Red Neuronal Base de Función Radial	19.10%

Para la segunda prueba, se eliminó nuevamente una variable, en este caso fue la variable “DÍA”, de tal manera que el nuevo conjunto de datos quedó con doce variables, estas mostradas en la TABLA XVII.

TABLA XVII
VARIABLES RESTANTES DEL CONJUNTO DE DATOS – PRUEBA 2

Nº	VARIABLES SELECCIONADAS
1	PROVINCIA
2	HORA
3	FERIADO
4	CAUSA PROBABLE
5	CLASE FINAL
6	ZONA
7	TIPO DE VEHÍCULO 1
8	SERVICIO 1
9	EDAD 1
10	SEXO 1
11	CONDICIÓN 1
12	PARTICIPANTE 1

Los porcentajes de los resultados obtenidos, después de la aplicación de los algoritmos seleccionados (ver [Repositorio](#)²⁰), se visualizan los resultados con respecto al porcentaje global de instancias clasificadas correctamente en la TABLA XVIII y en la

TABLA XIX los porcentajes de precisión global.

²⁰https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion/1_una %20variable

TABLA XVIII
RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN
CORRECTA - PRUEBA 2

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	58,38%
2	Árbol de Decisión CHAID	58,10%
3	Red Neuronal Perceptrón Multicapa	58,10%
4	BayesNet	55,47%
5	Naïve Bayes	55,31%
6	Árbol de Decisión CRT	45,30%
7	Red Neuronal Base de Función Radial	43,00%

TABLA XIX
RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE
PRECISIÓN – PRUEBA 2

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,60%
2	Red Neuronal Perceptrón Multicapa	44,38%
3	BayesNet	44,25%
4	Árbol de Decisión CHAID	44,00%
5	Naive Bayes	43,56%
6	Árbol de Decisión CRT	28,38%
7	Red Neuronal Base de Función Radial	23,21%

En la siguiente prueba, fueron eliminadas las variables “DIA” y “FERIADO”, con lo cual, el nuevo conjunto de datos quedó con once variables, las cuales se muestran en la TABLA XX.

TABLA XX
VARIABLES RESTANTES DEL CONJUNTO DE DATOS – PRUEBA 3

Nº	Variables Seleccionadas
1	PROVINCIA
2	HORA
3	CAUSA PROBABLE
4	CLASE FINAL
5	ZONA

6	TIPO DE VEHÍCULO 1
7	SERVICIO 1
8	EDAD 1
9	SEXO 1
10	CONDICIÓN 1
11	PARTICIPANTE 1

Al finalizar la aplicación de los algoritmos correspondientes (ver [Repositorio²¹](#)), se obtuvieron los resultados con respecto al porcentaje global de instancias clasificadas correctamente (ver TABLA XXI) y el porcentaje de precisión global (ver TABLA XXII).

TABLA XXI
RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN
CORRECTA - PRUEBA 3

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	58,38%
2	Árbol de Decisión CHAID	58,10%
3	Red Neuronal Perceptrón Multicapa	57,90%
4	BayesNet	55,37%
5	Naïve Bayes	55,20%
6	Árbol de Decisión CRT	45,33%
7	Red Neuronal Base de Función Radial	41,38%

TABLA XXII
RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE
PRECISIÓN – PRUEBA 3

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,55%
2	Red Neuronal Perceptrón Multicapa	44,25%
3	BayesNet	44,22%
5	Naive Bayes	44,20%
4	Árbol de Decisión CHAID	44,00%
6	Árbol de Decisión CRT	28,38%

²¹https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion/3_dos_variables

7	Red Neuronal Base de Función Radial	21,80%
----------	-------------------------------------	--------

En la cuarta prueba, al conjunto de datos inicial, se le eliminaron las variables “DÍA”, “FERIADO” y “EDAD 1”, quedando el conjunto de datos con las diez variables mostradas en la TABLA XXIII.

TABLA XXIII
VARIABLES RESTANTES DEL CONJUNTO DE DATOS – PRUEBA 4

Nº	VARIABLES SELECCIONADAS
1	PROVINCIA
3	HORA
4	CAUSA PROBABLE
5	CLASE FINAL
6	ZONA
7	TIPO DE VEHÍCULO 1
8	SERVICIO 1
10	SEXO 1
11	CONDICIÓN 1
12	PARTICIPANTE 1

Después de la aplicación de los algoritmos seleccionados (ver [Repositorio](#)²²), los porcentajes de los resultados obtenidos se visualizan en la TABLA XXIV con respecto al porcentaje global de instancias clasificadas correctamente y en la TABLA XXV los porcentajes de precisión global.

TABLA XXIV
RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN CORRECTA - PRUEBA 4

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	58,38%
2	Árbol de Decisión CHAID	58,10%
3	Red Neuronal Perceptrón Multicapa	57,90%
4	BayesNet	55,37%
5	Naïve Bayes	55,20%

²²https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion/4_tres_variables

6	Árbol de Decisión CRT	45,33%
7	Red Neuronal Base de Función Radial	41,38%

TABLA XXV
RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE
PRECISIÓN – PRUEBA 4

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,55%
2	Red Neuronal Perceptrón Multicapa	44,25%
3	BayesNet	44,22%
5	Naive Bayes	44,20%
4	Árbol de Decisión CHAID	44,00%
6	Árbol de Decisión CRT	28,38%
7	Red Neuronal Base de Función Radial	21,80%

Finalmente, en la quinta prueba, al conjunto de datos se le eliminó cuatro variables, estas fueron la variable “DÍA”, “FERIADO”, “EDAD 1”, “CONDICIÓN 1”, quedando el conjunto de datos con las nueve variables que se visualizan en la TABLA XXVI.

TABLA XXVI
VARIABLES RESTANTES DEL CONJUNTO DE DATOS – PRUEBA 5

Nº	VARIABLES SELECCIONADAS
1	PROVINCIA
2	DÍA
3	HORA
3	FERIADO
4	CAUSA PROBABLE
5	CLASE FINAL
6	ZONA
7	TIPO DE VEHÍCULO 1
8	SERVICIO 1
9	EDAD 1
10	SEXO 1
11	CONDICIÓN 1
12	PARTICIPANTE 1

Los resultados con respecto al porcentaje global de instancias clasificadas correctamente y el porcentaje de precisión global, después de la aplicación de los algoritmos correspondientes (ver [Repositorio](#)²³), se muestran ordenados de manera descendente en la TABLA XXVII y TABLA XXVIII.

TABLA XXVII

RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN
CORRECTA - PRUEBA 5

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	57,90%
2	Árbol de Decisión CHAID	57,50%
3	Red Neuronal Perceptrón Multicapa	57,20%
4	BayesNet	55,54%
5	Naïve Bayes	55,52%
6	Red Neuronal Base de Función Radial	50,70%
7	Árbol de Decisión CRT	45,30%

TABLA XXVIII

RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE
PRECISIÓN – PRUEBA 5

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,36%
2	BayesNet	44,25%
3	Árbol de Decisión CHAID	44,30%
4	Naive Bayes	44,18%
5	Red Neuronal Perceptrón Multicapa	43,92%
6	Árbol de Decisión CRT	28,38%
7	Red Neuronal Base de Función Radial	28,33%

Al finalizar el proceso de experimentación, se evidencia que no fue posible incrementar los porcentajes obtenidos inicialmente, por lo tanto, dichos porcentajes iniciales fueron los utilizados para trabajar en la siguiente fase, ya que fueron lo que presentaron mejores resultados.

²³https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/experimentacion/5_cuatro_variables

6.3. Objetivo 3: Interpretar y presentar los resultados obtenidos sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

En la ejecución del tercer objetivo se cumplió la última fase de la metodología KDD especificada a continuación:

Fase V: Interpretación y presentación de resultados

Tarea 1: Evaluar e identificar las mejores técnicas para la obtención de los resultados

En esta sección se realizó el análisis de los resultados obtenidos después de la aplicación inicial de los algoritmos de minería de datos. El análisis estuvo dado principalmente entorno a métricas de rendimiento basadas en la matriz de confusión generada por cada algoritmo (las matrices se presentan en el Anexo 7), estas métricas fueron el porcentaje global de instancias clasificadas correctamente y el porcentaje de precisión global especificado para cada categoría de la variable objetivo.

Mediante esta aplicación se realizó la clasificación del conjunto de datos sobre siniestros de tránsito registrados en Ecuador en el año 2020, en el que se aplicaron siete algoritmos que de acuerdo a la revisión bibliográfica son los más relevantes, estos son:

- Árbol de Decisión CHAID
- Árbol de Decisión CHAID Exhaustivo
- Árbol de Decisión CRT
- Red Neuronal Perceptrón Multicapa
- Red Neuronal Base de Función Radial
- Naive Bayes
- BayesNet

A continuación, en la Fig. 39 se muestran los resultados correspondientes a los porcentajes globales de clasificación correcta obtenidos de la aplicación de los algoritmos al conjunto de datos para medir la eficiencia de cada modelo. Analizando los porcentajes, se puede observar que el algoritmo con el valor más alto es el árbol de decisión CHAID Exhaustivo con un valor del 58,38%.

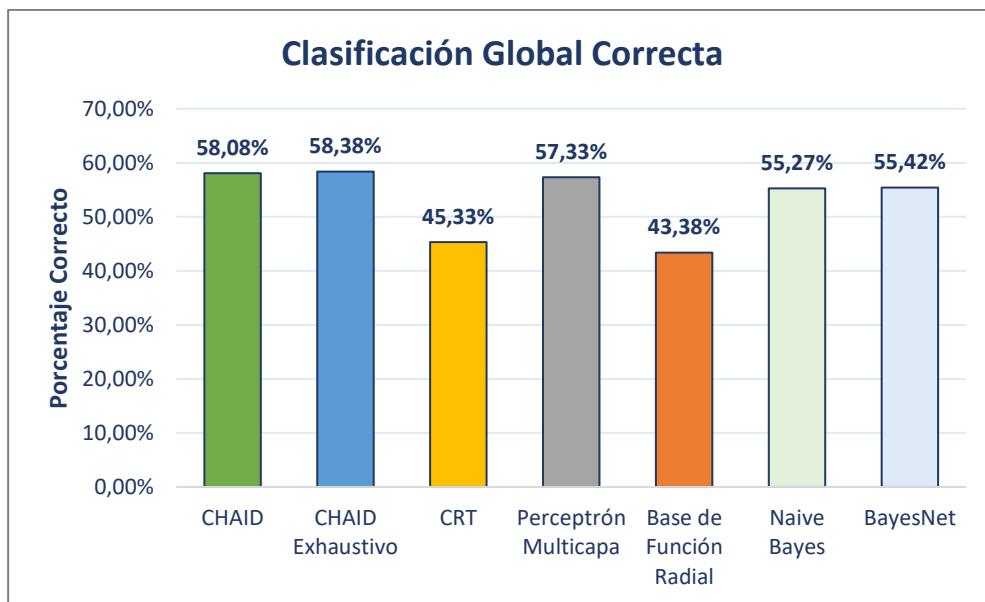


Fig. 39 Clasificación global correcta de instancias de cada algoritmo

Analizando los porcentajes de precisión visualizados en la TABLA XXIX, se observa que el resultado del árbol de clasificación CHAID Exhaustivo es mayor al de los otros algoritmos con un valor total de 44,60%, respaldando el número de predicciones correctas tomando la exactitud de los datos clasificados correctamente y comparándolos con el total del conjunto de datos.

TABLA XXIX
PRECISIÓN GLOBAL DE CADA ALGORITMO

Observado	CHAID	CHAID Exhaustivo	CRT	Perceptrón Multicapa	Función de Base Radial	Naive Bayes	BayesNet
ARROLLAMIENTOS	49,82%	49,82%	0,00%	49,57%	0,00%	63,00%	61,90%
ATROPELLOS	60,67%	62,97%	33,83%	64,83%	67,62%	68,10%	67,80%
CAÍDA DE PASAJERO	79,06%	79,06%	0,00%	75,00%	0,00%	68,70%	68,40%
CHOQUE FRONTAL	44,93%	44,93%	44,93%	55,17%	0,00%	70,80%	70,60%
CHOQUE LATERAL	86,11%	85,38%	68,67%	80,92%	87,60%	32,20%	30,90%
CHOQUE POSTERIOR	61,06%	61,06%	61,06%	58,43%	41,35%	81,70%	82,40%
COLISIÓN	0,00%	0,00%	0,00%	0,00%	0,00%	41,50%	41,50%
ESTRELLAMIENTOS	40,85%	41,79%	92,14%	46,52%	24,36%	82,10%	82,20%
OTROS	18,31%	18,31%	0,00%	13,92%	0,00%	34,20%	34,30%
PÉRDIDA DE CARRIL	12,25%	14,27%	0,00%	10,18%	0,00%	17,10%	17,20%
PÉRDIDA DE PISTA	53,77%	53,77%	0,00%	48,44%	45,31%	19,40%	19,20%
ROZAMIENTOS	68,39%	68,39%	68,39%	73,42%	0,00%	32,20%	32,30%
VOLCAMIENTOS	0,00%	0,00%	0,00%	0,00%	0,00%	38,60%	41,30%
Porcentaje de Precisión	44,25%	44,60%	28,39%	44,34%	20,48%	44,52%	44,57%

De acuerdo al análisis entorno a las métricas de rendimiento antes mencionadas, se identificó a los mejores algoritmos para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador en el año 2020, dichos algoritmos se muestran ordenados de manera descendente en la TABLA XXX y

TABLA XXXI.

TABLA XXX
RESULTADOS DE ALGORITMOS CON RESPECTO A LA CLASIFICACIÓN
CORRECTA

Nº	Algoritmos	Clasificación correcta
1	Árbol de Decisión CHAID Exhaustivo	58,38%
2	Árbol de Decisión CHAID	58,10%
3	Red Neuronal Perceptrón Multicapa	57,33%
4	BayesNet	55,42%
5	Naïve Bayes	55,27%
6	Árbol de Decisión CRT	45,33%
7	Red Neuronal Base de Función Radial	43,38%

TABLA XXXI
RESULTADOS DE ALGORITMOS CON RESPECTO A LA MÉTRICA DE
PRECISIÓN

Nº	Algoritmos	Precisión
1	Árbol de Decisión CHAID Exhaustivo	44,60%
2	BayesNet	44,57%
3	Naive Bayes	44,52%
4	Red Neuronal Perceptrón Multicapa	44,34%
5	Árbol de Decisión CHAID	44,25%
6	Árbol de Decisión CRT	28,39%
7	Red Neuronal Base de Función Radial	20,48%

El algoritmo de árbol de decisión CHAID Exhaustivo con un 58,38% y 44,60% de clasificación correcta y precisión respectivamente fue seleccionado como el de mejor rendimiento para la presentación de los resultados de la minería de datos aplicada.

En la Fig. 40 se presentan los mejores resultados tanto de clasificación correcta como de precisión de acuerdo a los algoritmos CHAID, CHAID Exhaustivo, CRT,

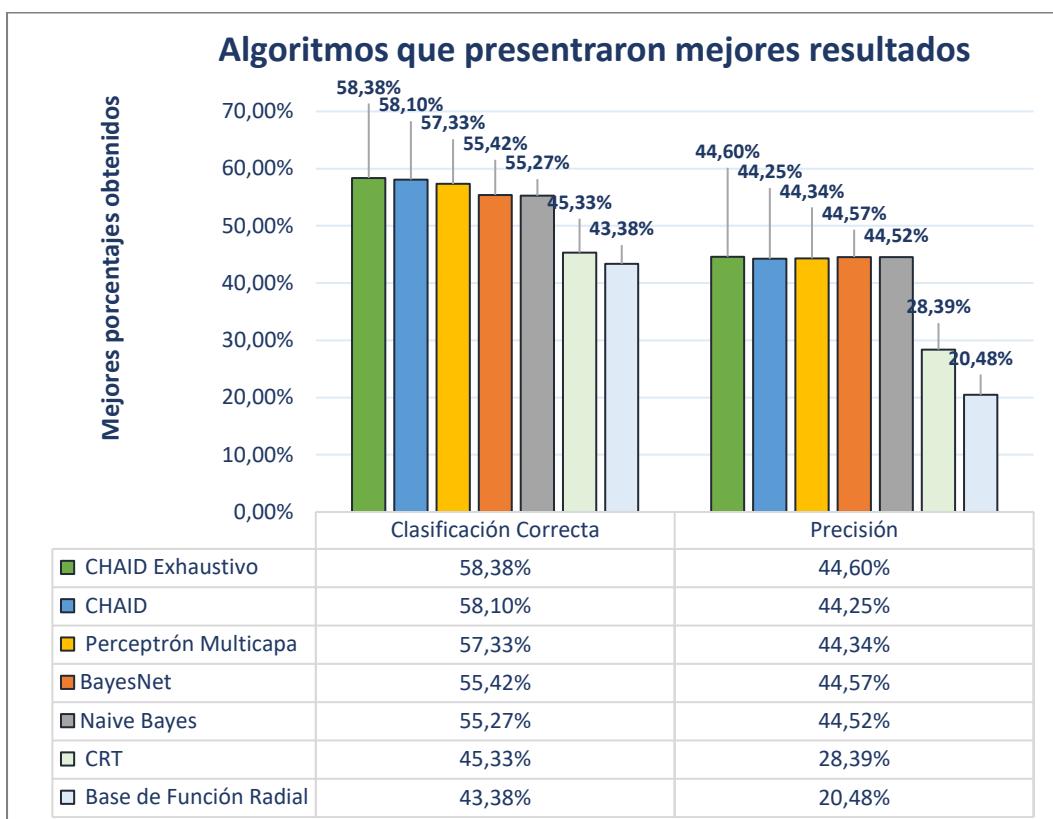


Fig. 40 Mejores resultados de clasificación correcta y precisión de acuerdo a cada algoritmo

Tarea 2: Presentar e Interpretar los resultados obtenidos

Los resultados que se presentan en esta tarea fueron obtenidos mediante la selección del algoritmo CHAID Exhaustivo, en concordancia a los porcentajes de clasificación correcta y de precisión, a través de dicho algoritmo se presenta el conjunto de reglas de asociación de patrones obtenida, estas interpretadas mediante la utilización de gráficos en los cuales se muestra el contexto de la ocurrencia de las diferentes clases de siniestros de tránsito en Ecuador.

Para el orden de presentación se toma como referencia la Fig. 31 correspondiente al nodo raíz del algoritmo seleccionado, en el cual muestra los resultados con respecto a las categorías presentes en la variable objetivo “CLASE_FINAL”.

Nodo 0		
Categoría	%	n
ARROLLAMIENTOS	6,6	1118
ATROPELLOS	11,8	2001
CAIDA DE PASAJERO	2,1	363
CHOQUE FRONTAL	5,2	877
CHOQUE LATERAL	28,6	4851
CHOQUE POSTERIOR	7,8	1320
COLISION	2,0	341
ESTRELLAMIENTOS	12,6	2137
OTROS	2,2	366
PERDIDA DE CARRIL	6,1	1037
PERDIDA DE PISTA	9,2	1551
ROZAMIENTOS	4,3	734
VOLCAMIENTOS	1,4	244
Total	100,0	16940

Fig. 41 Nodo raíz del algoritmo CHAID Exhaustivo

En la TABLA XXXII se presentan los resultados de las categorías de la variable “CLASE_FINAL” ordenadas de manera descendente con respecto al porcentaje de probabilidad de ocurrencia de cada clase de siniestro de tránsito analizado.

TABLA XXXII
OCURRENCIA DE SINIESTROS DE TRÁNSITO

Nº	Clase de Siniestros	Probabilidad de Ocurrencia
1	Choque Lateral	28,60%
2	Estrellamientos	12,60%
3	Atropellos	11,80%
4	Pérdida de Pista	9,20%
5	Choque Posterior	7,80%
6	Arrollamientos	6,60%
7	Pérdida de Carril	6,10%
8	Choque Frontal	5,20%
9	Rozamientos	4,30%
10	Otros	2,20%
11	Caída de Pasajero	2,10%
12	Colisión	2,00%
13	Volcamientos	1,40%

A continuación, en la TABLA XXXIII se presenta la clasificación de los siniestros de tránsito según los datos recolectados en Ecuador en el año 2020 mostrados en la TABLA XXXII.

TABLA XXXIII
CLASIFICACIÓN DE LOS SINIESTROS DE TRÁNSITO

Clase de Siniestros	Descripción
Choque Lateral	Es el que se da cuando la parte frontal de un vehículo impacta con la parte lateral de otro [52], [53].
Estrellamiento	Es el impacto que se produce entre un vehículo en movimiento contra un vehículo que este en reposo o contra un objeto fijo [52], [53].
Atropello	Es el impacto de un vehículo en movimiento a un peatón o animal [52], [53].
Pérdida de Pista	Es la salida del vehículo de la calzada normal de circulación [52], [53].
Choque Posterior	Es el impacto que se produce cuando un vehículo en movimiento se impacta con la parte frontal en la parte posterior de otro vehículo también en movimiento [52], [53].
Arrollamiento	Es la acción por la cual un vehículo pasa con su rueda o ruedas por encima del cuerpo de una persona o animal [52].
Pérdida de Carril	Es la salida del vehículo hacia el carril contrario al de su circulación [52], [53].
Choque Frontal	Es el impacto frontal entre dos vehículos [52], [53].
Rozamiento	Es el contacto o fricción de la parte lateral de un vehículo en movimiento con vehículo en reposo o un objeto fijo [52], [53].
Caída de Pasajero	Es la pérdida de equilibrio del pasajero que produce su descenso violento desde el estribo o del interior del vehículo hacia la calzada [52], [53].
Colisión	Choque de más de dos vehículos en movimiento [52], [53].
Volcamiento	Es el que se produce por la inversión de la posición de un vehículo, ya sea realizando giros por la parte lateral derecha o izquierda o por la parte frontal o posterior del vehículo [52].

La Fig. 42 muestra que la principal causa probable para que ocurran siniestros de tránsito de clase choque lateral es en la que el conductor no respeta las señales reglamentarias de tránsito (pare, ceda el paso, luz roja del semáforo, etc.), con una probabilidad del 87,00%, esta probabilidad aumenta a un 96,40% al darse el

siniestro en una zona urbana y al final la probabilidad incrementa al 97,30% si este siniestro ocurre en la provincia de Guayas, Loja, Morona Santiago o Napo.

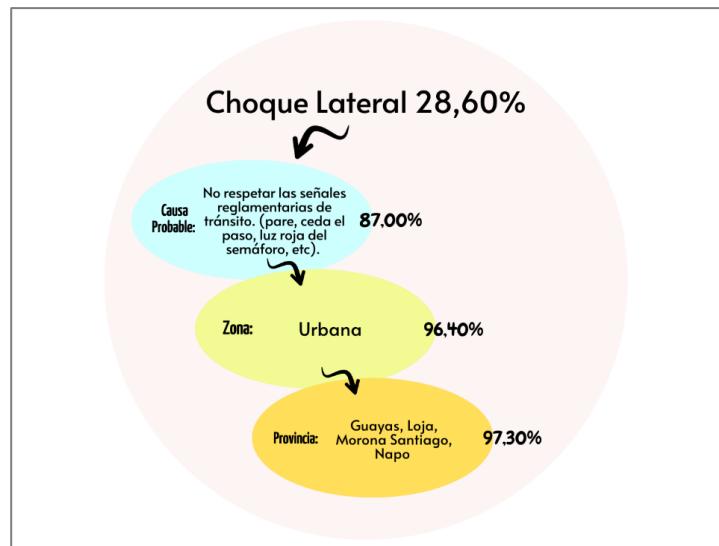


Fig. 42 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral

Tal como se muestra en la Fig. 43, el número de siniestros de tránsito registrados referentes a la clase estrellamientos es de 2137 que corresponde al 12,60% del total de registros, la causa probable más usual para que ocurra esta clase de siniestros de tránsito es debido a que el vehículo involucrado presenta una falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico), con una probabilidad de ocurrencia del 51,40%.

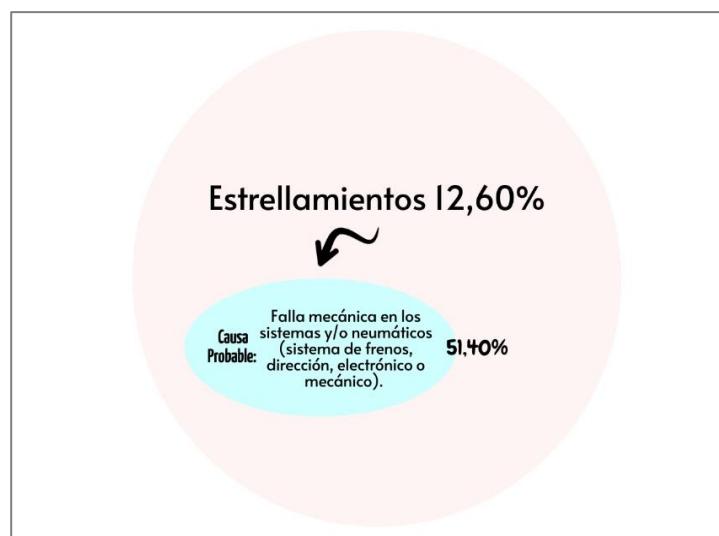


Fig. 43 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Estrellamientos

La principal causa probable para que ocurran siniestros de tránsito de clase

atropellos es debido a que el peatón no transita por las aceras o zonas de seguridad destinadas para el efecto con una probabilidad de ocurrencia del 92,40%, esto se muestra en la Fig. 44.

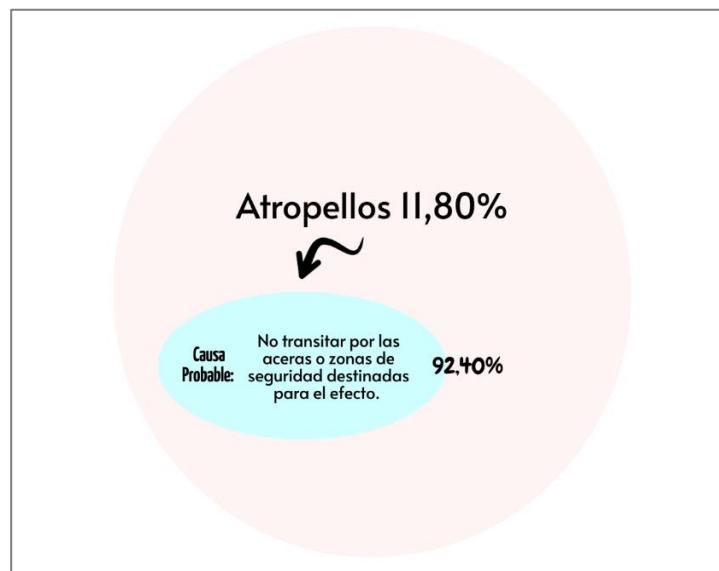


Fig. 44 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Atropellos

Como se muestra en la Fig. 45, la principal causa probable para que ocurran siniestros de tránsito de clase choque posterior es debido a que el conductor no mantiene la distancia prudencial con respecto al vehículo que le antecede, con una probabilidad del 68,20% de ocurrencia y además si el tipo de vehículo involucrado sea una motocicleta o bicicleta la probabilidad de ocurrencia aumenta a un 77,20%.

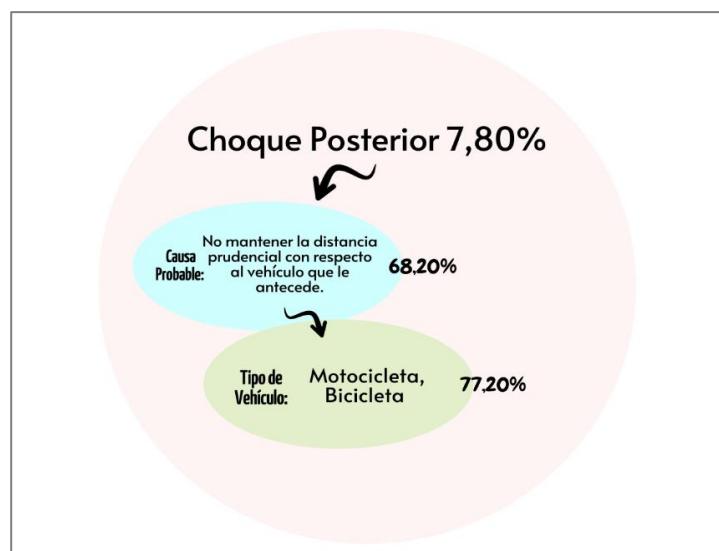


Fig. 45 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Posterior

En la Fig. 46 se muestra que para que ocurran siniestros de tránsito de clase arrollamientos, la causa probable más habitual es conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor), con una probabilidad de ocurrencia del 20,50%, además a que estos siniestros ocurran en la provincia de pichincha con una probabilidad del 42,20%, y que tipo de vehículo involucrado ya sea automóvil y especial con una probabilidad del 59,00% de ocurrencia y finalmente se de en una zona urbana con un 67,70% de probabilidad de ocurrencia.

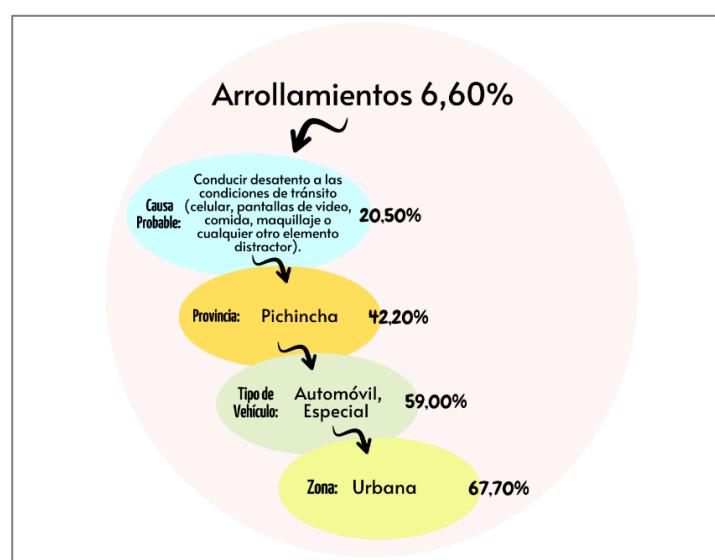


Fig. 46 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Arrollamientos

Para que ocurran siniestros de tránsito de clase pérdida de carril, de acuerdo a la Fig. 47, se obtuvo que la principal causa probable de ocurrencia es debido a la presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.) con una probabilidad del 44,40% y además que la condición del involucrado sea de lesionado con una probabilidad del 66,00%.

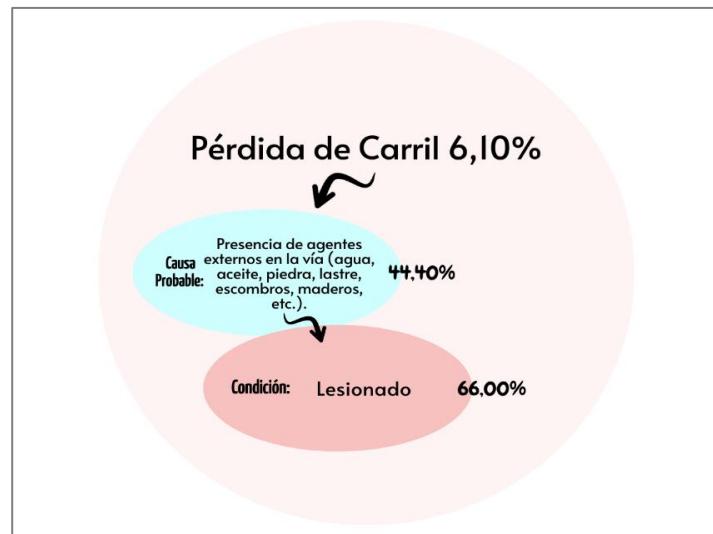


Fig. 47 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral

En la mayoría de siniestros de tránsito de clase frontal como se muestra en la Fig. 48, la principal causa probable fue el conducir en sentido contrario a la vía normal de circulación con un 88,30% de probabilidad de ocurrencia y esta probabilidad aumenta al 95,20% debido a que este ocurre en las provincias de Santa Elena, Loja, Morona Santiago, Pastaza o Cañar.

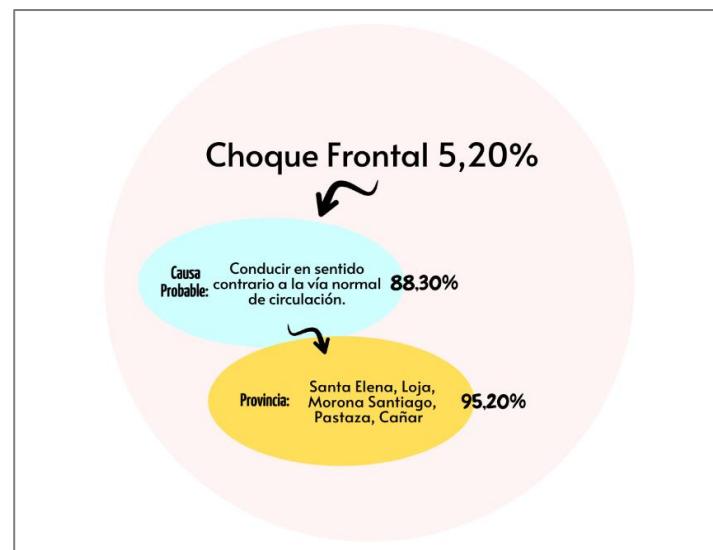


Fig. 48 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Frontal

De acuerdo a la Fig. 49, la principal causa probable para que ocurran siniestros de tránsito de clase rozamientos es debido a que el conductor no guarda la distancia lateral mínima de seguridad entre vehículos y no respeta las señales manuales del

agente de tránsito, con una probabilidad del 76,00% de ocurrencia, además a que estos siniestros ocurran en la provincia de Guayas o Santa Elena con un 95,70% de probabilidad de ocurrencia y que finalmente se den en los períodos de tiempo de las horas mostradas en la TABLA XXXIV con una probabilidad del 100% de ocurrencia.

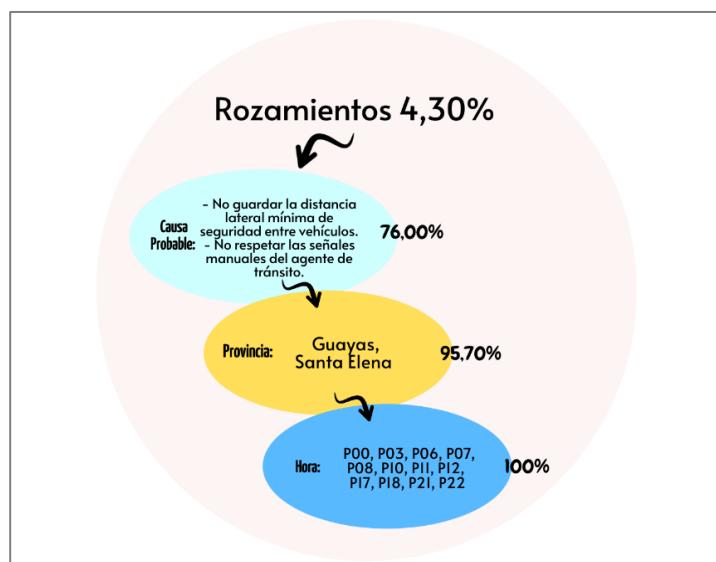


Fig. 49 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Rozamientos

TABLA XXXIV

HORAS DE OCURRENCIA DEL SINIESTRO DE TRÁNSITO ROZAMIENTOS

Hora
00:00:00 A 00:59:00 AM
03:00:00 A 03:59:00 AM
06:00:00 A 06:59:00 AM
07:00:00 A 07:59:00 AM
08:00:00 A 08:59:00 AM
10:00:00 A 10:59:00 AM
11:00:00 A 11:59:00 AM
12:00:00 A 12:59:00 PM
17:00:00 A 17:59:00 PM
18:00:00 A 18:59:00 PM
21:00:00 A 21:59:00 PM
22:00:00 A 22:59:00 PM

En la Fig. 50 se muestra que, la principal causa probable para que ocurran otras clases de siniestros de tránsito se da por un caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.), con una probabilidad de ocurrencia del 43,20%.

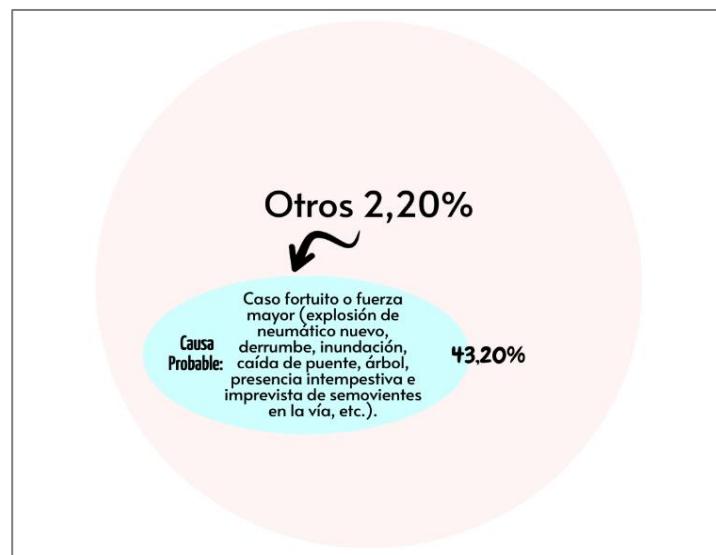


Fig. 50 Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Otros

De acuerdo a lo expuesto en la Fig. 51, la principal causa probable para que ocurran los siniestros de tránsito de clase caída de pasajero está determinada en su mayoría debido a que los pasajeros se bajan o suben de vehículos en movimiento sin tomar las precauciones debidas con una probabilidad del 93,80% y además que estos ocurrían en los períodos de tiempo de las horas mostradas en la TABLA XXXV con una probabilidad del 100% de ocurrencia.

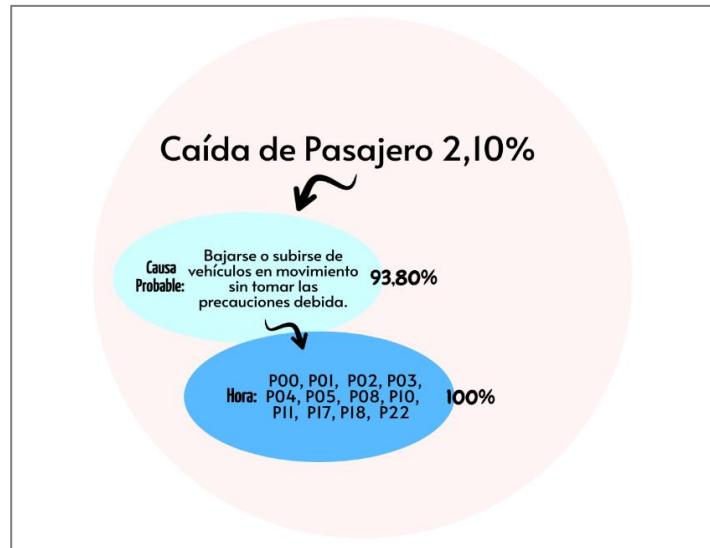


Fig. 51 Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Caída de Pasajero

TABLA XXXV
HORAS DE OCURRENCIA DEL SINIESTRO DE TRÁNSITO CAÍDA DE PASAJERO

Hora
00:00:00 A 00:59:00 AM
01:00:00 A 01:59:00 AM
02:00:00 A 02:59:00 AM
03:00:00 A 03:59:00 AM
04:00:00 A 04:59:00 AM
05:00:00 A 05:59:00 AM
08:00:00 A 08:59:00 AM
10:00:00 A 10:59:00 AM
11:00:00 A 11:59:00 AM
17:00:00 A 17:59:00 PM
18:00:00 A 18:59:00 PM
22:00:00 A 22:59:00 PM

Con la finalidad de que la presentación de los resultados tenga una orientación más técnica, se procedió a crear con las reglas obtenidas, perfiles de conductores y peatones involucrados en la ocurrencia de siniestros de tránsitos, presentados a través de tablas, estas especificadas para cada clase de siniestros de tránsito, tales como: Choque Lateral (ver TABLA XXXVI), Estrellamientos (ver TABLA XXXVII)

Atropellos (ver TABLA XXXVIII), Choque Posterior (ver TABLA XXXIX), Arrollamientos (ver TABLA XL), Pérdida de Carril (ver TABLA XLI), Choque Frontal (ver TABLA XLII), Rozamientos (ver TABLA XLIII), Otros (ver TABLA XLIV) y Caída de Pasajero (ver TABLA XLV).

TABLA XXXVI

PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE CHOQUE
LATERAL

Participante	Causa Probable	Provincia	Zona	Tipo de Vehículo	Probabilidad de Ocurrencia
Conductor	No respeta las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).	Guayas, Loja, Morona Santiago, Napo	Urbana		97,30%
Conductor	Realiza cambio brusco o indebido de carril.	Guayas, Morona Santiago			97,20%
Conductor	No cede el derecho de vía o preferencia de paso a vehículos.			Automóvil	93,30%

TABLA XXXVII
PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE ESTRELLAMIENTOS

Participante	Causa Probable	Provincia	Hora	Tipo de Vehículo	Condición	Probabilidad de Ocurrencia
Conductor	Conduce con falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).					51,40%
Conductor	Conduce en estado de somnolencia o malas condiciones físicas (sueno, cansancio y fatiga).					50,70%
Conductor	Conduce con daños mecánicos previsibles.					47,40%
Conductor	Conduce en condiciones ambientales y/o atmosféricas (niebla, neblina, granizo, lluvia).			Automóvil, Furgoneta, Bicicleta		43,90%
Conductor	Conduce vehículo superando los límites	Pichincha	01:00:00 a 01:59:00 am, 02:00:00 a 02:59:00 am,	Automóvil, Camión,	No identificado	43,00%

	máximos de velocidad.		03:00:00 a 03:59:00 am, 11:00:00 a 11:59:00 am, 15:00:00 a 15:59:00 pm	Camioneta, Furgoneta, Especial		
--	-----------------------	--	--	--------------------------------------	--	--

TABLA XXXVIII

PERFIL DE PEATONES Y CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE ATROPELLOS

Participante	Causa Probable	Probabilidad de Ocurrencia
Peatón	No transita por las aceras o zonas de seguridad destinadas para el efecto.	92,40%
Conductor	No cede el derecho de vía o preferencia de paso al peatón	92,10%
Peatón	Transita bajo influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos.	90,80%

TABLA XXXIX

PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE CHOQUE POSTERIOR

Participante	Causa Probable	Tipo de Vehículo	Probabilidad de Ocurrencia
Conductor	No mantiene la distancia prudencial con respecto al vehículo que le antecede.	Motocicleta, Bicicleta	77,20%

TABLA XL
PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE ARROLLAMIENTOS

Participante	Causa Probable	Provincia	Zona	Tipo de Vehículo	Probabilidad de Ocurrencia
Conductor	Conduce desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distracto).	Pichincha	Urbana	Automóvil, Especial	67,70%

TABLA XLI
PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE PÉRDIDA DE CARRIL

Participante	Causa Probable	Condición	Probabilidad de Ocurrencia
Conductor	Conduce en presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).	Lesionado	67,00%
Conductor	Conduce en malas condiciones de la vía y/o configuración. (iluminación y diseño).		34,40%

TABLA XLII
PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE CHOQUE FRONTAL

Participante	Causa Probable	Provincia	Probabilidad de Ocurrencia
Conductor	Conduce en sentido contrario a la vía normal de circulación.	Elena, Loja, Morona Santiago, Pastaza, Cañar	95,20%
Conductor	Adelanta o rebasa a otro vehículo en movimiento en zonas o sitios peligrosos tales como: curvas, puentes, túneles, pendientes, etc.		47,10%

TABLA XLIII
PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE ROZAMIENTOS

Participante	Causa Probable	Provincia	Hora	Probabilidad de Ocurrencia
Conductor	No guarda la distancia lateral mínima de seguridad entre vehículos. No respeta las señales manuales del agente de tránsito.	Guayas, Santa Elena	00:00:00 a 00:59:00 am, 03:00:00 a 03:59:00 am, 06:00:00 a 06:59:00 am, 07:00:00 a 07:59:00 am, 08:00:00 a 08:59:00 am, 10:00:00 a 10:59:00 am, 11:00:00 a 11:59:00 am,	100.00%

			12:00:00 a 12:59:00 pm, 17:00:00 a 17:59:00 pm, 18:00:00 a 18:59:00 pm, 21:00:00 a 21:59:00 pm, 22:00:00 a 22:59:00 pm	
--	--	--	--	--

TABLA XLIV

PERFIL DE CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE OTROS

Participante	Causa Probable	Probabilidad de Ocurrencia
Conductor	Conduce y sucede un caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).	43,20%

TABLA XLV

PERFIL DE PEATONES Y CONDUCTORES INVOLUCRADOS EN LA OCURRENCIA DEL SINIESTRO DE TRÁNSITO DE CLASE CAÍDA DE PASAJERO

Participante	Causa Probable	Hora	Probabilidad de Ocurrencia
Peatón	Baja o sube de vehículos en movimiento sin tomar las precauciones debida.	00:00:00 a 00:59:00 am, 01:00:00 a 01:59:00 am, 02:00:00 a 02:59:00 am, 03:00:00 a 03:59:00 am, 04:00:00 a 04:59:00 am,	43,20%

		05:00:00 a 05:59:00 am, 08:00:00 a 08:59:00 am, 10:00:00 a 10:59:00 am, 11:00:00 a 11:59:00 am, 17:00:00 a 17:59:00 pm, 18:00:00 a 18:59:00 pm, 22:00:00 a 22:59:00 pm	
Conductor	Deja o recoge pasajeros en lugares no permitidos.		28,30%

Según los gráficos y tablas presentadas anteriormente se identifica que la variable más influyente para que ocurran siniestros de tránsito es la “CAUSA PROBABLE”, es por esto que se realiza el análisis de las mismas, tomando en cuenta la información proporcionada por la ANT, entidad que lleva las estadísticas de accidentes y siniestralidad vial en Ecuador, la cual ha categorizado para el año 2020 las causas probables de los siniestros de tránsito tal como se muestra en la TABLA XLI.

TABLA XLVI
CAUSAS PROBABLES DE SINIESTROS DE TRÁNSITO EN EL ECUADOR

Nº	Causa probable
1	CP01: Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).
2	CP02: Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).
3	CP03: Conducir en estado de somnolencia o malas condiciones físicas (sueño, cansancio y fatiga).

4	CP04: Daños mecánicos previsibles.
5	CP05: Falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).
6	CP06: Conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos.
7	CP07: Peatón transita bajo influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos.
8	CP08: Peso y volumen - no cumplir con las normas de seguridad necesarias al transportar cargas.
9	CP09: Conducir vehículo superando los límites máximos de velocidad.
10	CP10: Condiciones ambientales y/o atmosféricas (niebla, neblina, granizo, lluvia).
11	CP11: No mantener la distancia prudencial con respecto al vehículo que le antecede.
12	CP12: No guardar la distancia lateral mínima de seguridad entre vehículos.
13	CP13: Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).
14	CP14: Dejar o recoger pasajeros en lugares no permitidos.
15	CP15: No transitar por las aceras o zonas de seguridad destinadas para el efecto.
16	CP16: Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.
17	CP17: Conducir en sentido contrario a la vía normal de circulación.
18	CP18: Realizar cambio brusco o indebido de carril.
19	CP19: Mal estacionado - el conductor que detenga o estacione vehículos en sitios o zonas que entrañen peligro, tales como zona de seguridad, curvas, puentes, túneles, pendientes.
20	CP20: Malas condiciones de la vía y/o configuración. (iluminación y diseño).
21	CP21: Adelantar o rebasar a otro vehículo en movimiento en zonas o sitios peligrosos tales como: curvas, puentes, túneles, pendientes, etc.
22	CP22: No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).
23	CP23: No respetar las señales manuales del agente de tránsito.
24	CP24: No ceder el derecho de vía o preferencia de paso a vehículos.

25	CP25: No ceder el derecho de vía o preferencia de paso al peatón.
26	CP26: Peatón que cruza la calzada sin respetar la señalización existente (semáforos o señales manuales).

En la TABLA XLVII se muestran los porcentajes de probabilidad de ocurrencia de las principales causas probables que intervienen para que se produzca cada una de las clases de siniestros de tránsito registrados en Ecuador en el año 2020.

TABLA XLVII
PROBABILIDADES DE OCURRENCIA DE LAS CAUSAS PROBABLES PARA
CADA CLASE DE SINIESTRO DE TRÁNSITO

Nº	Clase de siniestro	Causa probable	Probabilidad de ocurrencia
1	Choque Lateral	CP22: No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).	87,00%
2	Estrellamientos	CP05: Falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).	51,40%
3	Atropellos	CP15: No transitar por las aceras o zonas de seguridad destinadas para el efecto.	92,40%
4	Choque Posterior	CP11: No mantener la distancia prudencial con respecto al vehículo que le antecede.	69,20%
5	Arrollamientos	CP13: Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).	20,50%
6	Pérdida de Carril	CP02: Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).	44,40%
7	Choque Frontal	CP17: Conducir en sentido contrario a la vía normal de circulación.	95,20%
8	Rozamientos	CP12: No guardar la distancia lateral mínima de seguridad entre vehículos. CP23: No respetar las señales manuales	76,00%

		del agente de tránsito.	
9	Otros	CP01: Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).	43,20%
10	Caída de Pasajero	CP16: Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.	93,80%

La causa probable con mayor probabilidad de ocurrencia con un 95,20% es la de conducir en sentido contrario a la vía normal de circulación, provocando el siniestro de tránsito de clase choque frontal, mientras que conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractivo), es la causa probable con menor probabilidad de ocurrencia con un total del 20,50% provocando en su mayoría siniestros de tránsito de clase arrollamientos, esto mostrado en la Fig. 52.

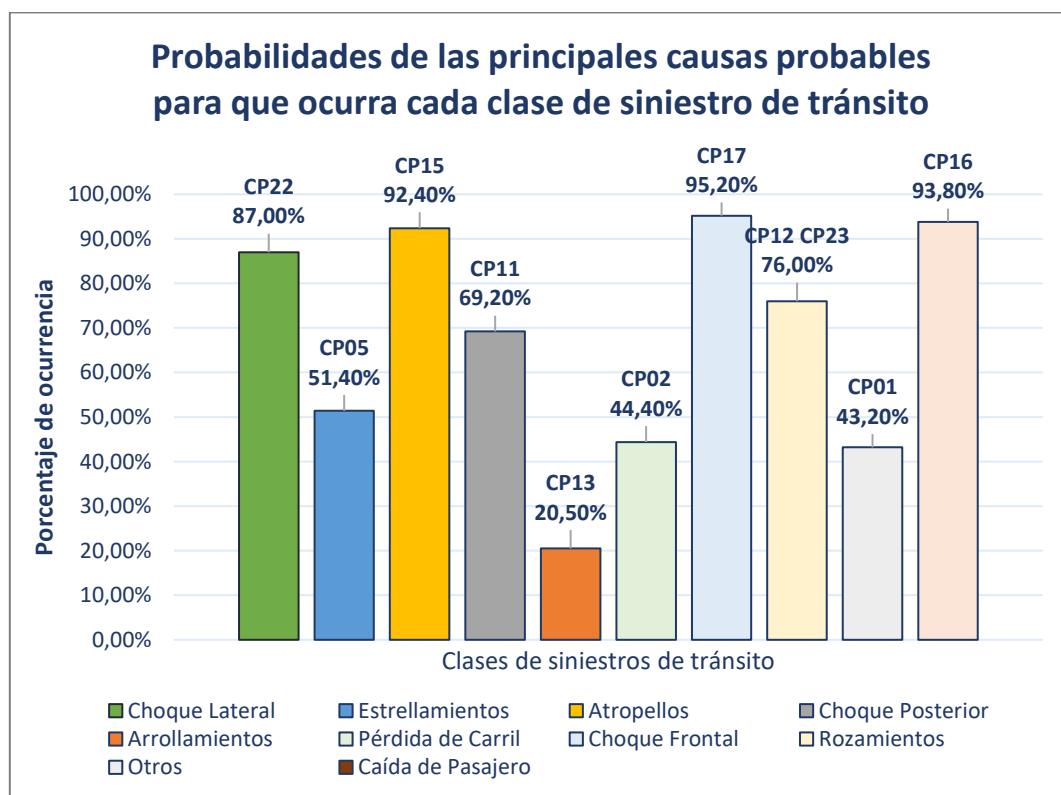


Fig. 52 Resultados de las probabilidades de ocurrencia de las causas probables

para cada clase de siniestro de tránsito

Para cumplir con el objeto de estudio las causas probables mostradas en la TABLA XLVII se las categorizó en tres factores que son: Factor Humano, Factor Vehículo y Factor Entorno, tomando como referencia los trabajos [52], [54] y [55]; esto con la finalidad y el afán de llegar a las causas probables más concretas, especificando el factor relacionado por el cual se producen o se ocasionan los siniestros de tránsito, para al final poder determinar cuáles son los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador, quedando la categorización de las causas probables tal como se muestra en la TABLA XLVIII (la categorización completa se puede visualizar en el Anexo 8.

TABLA XLVIII

CATEGORIZACIÓN DE LAS CAUSAS PROBABLES DE LOS SINIESTROS DE TRÁNSITO

Nº	Causa Probable	Categoría
1	CP11: No mantener la distancia prudencial con respecto al vehículo que le antecede.	Factor Humano
2	CP12: No guardar la distancia lateral mínima de seguridad entre vehículos.	
3	CP13: Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractivo).	
4	CP17: Conducir en sentido contrario a la vía normal de circulación.	
5	CP22: No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).	
6	CP23: No respetar las señales manuales del agente de tránsito.	
7	CP05: Falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).	Factor Vehículo
8	CP01: Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).	Factor Entorno
9	CP02: Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).	

10	CP15: No transitar por las aceras o zonas de seguridad destinadas para el efecto.	
11	CP16: Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.	

Como muestra la Fig. 53, al categorizar las once causas probables mostradas en la TABLA XLVIII, cinco de estas causas fueron categorizadas dentro del factor humano, una causa probable fue asignada al factor vehículo y por último cuatro causas categorizadas en el factor entorno.

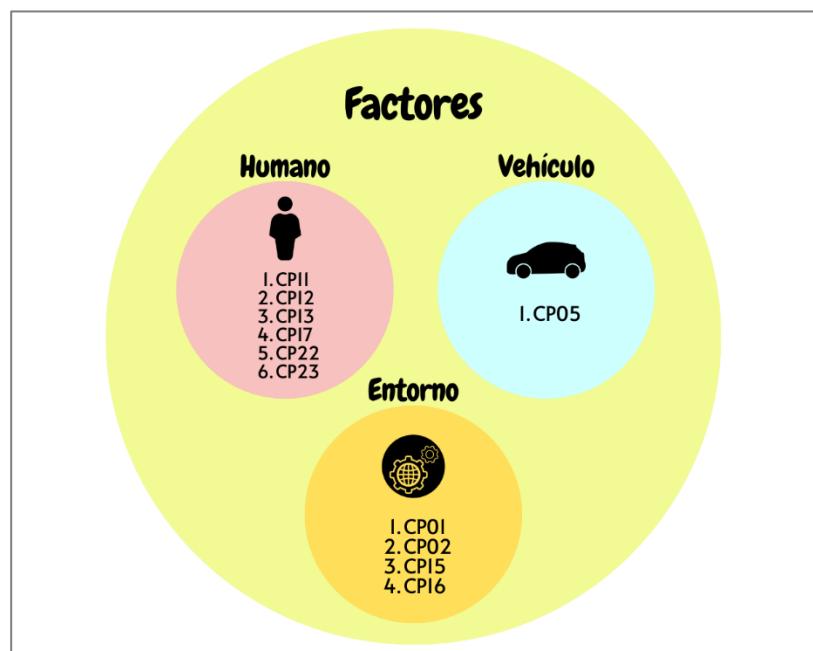


Fig. 53 Categorización en factores a las causas probables de los siniestros de tránsito

Como se puede ver en la Fig. 53, se separó específicamente las causas probables generadas por el factor humano, vehículo y entorno, que son en los cuales se basó para determinar los factores más influyentes para la ocurrencia de siniestros de tránsito en el Ecuador, lo cual es lo que se está planteando en el presente trabajo. Es así que a partir de esta diferenciación de las causas probables de los siniestros de tránsito en Ecuador, se analiza las cifras del año 2020, estas obtenidas del conjunto de reglas generadas a partir de la aplicación del algoritmo CHAID Exhaustivo, de tal forma que se realizó la ponderación de las probabilidades de ocurrencia de cada una de la causas probables correspondientes a los tres tipos de factores establecidos en la ocurrencia de las diferentes clases de siniestros de tránsito, esto mostrado en la TABLA XLIX.

TABLA XLIX
PROMEDIO DE LAS PROBABILIDADES DE OCURRENCIA DE LAS CAUSAS
PROBABLES PARA CADA TIPO DE FACTOR

Clase de siniestro	Causa probable	Tipo de Factor	Probabilidad de Ocurrencia	Media
Choque Frontal	CP17: Conducir en sentido contrario a la vía normal de circulación.	Factor Humano	95,20%	69,64%
Choque Lateral	CP22: No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).		87,00%	
Rozamientos	CP12: No guardar la distancia lateral mínima de seguridad entre vehículos. CP23: No respetar las señales manuales del agente de tránsito.		76,00%	
Choque Posterior	CP11: No mantener la distancia prudencial con respecto al vehículo que le antecede.		69,20%	
Arrollamientos	CP13: Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).		20,50%	
Caída de Pasajero	CP16: Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.	Factor Entorno	93,80%	68,45%

Atropellos	CP15: No transitar por las aceras o zonas de seguridad destinadas para el efecto.		92,40%	
Pérdida de Carril	CP02: Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).		44,40%	
Otros	CP01: Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).		43,20%	
Estrellamientos	CP05: Falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).	Factor Vehículo	51,40%	51,40%

La Fig. 54 muestra el contexto en el cual se determina la probabilidad de ocurrencia para cada tipo de factor establecido, involucrando en el Factor Humano a cinco clases de siniestros, para el Factor Entorno cuatro clases de siniestros y finalmente una clase de siniestro para el Factor Vehículo, acotando que, cada una de estas clases presenta la causa principal por la cual ocurre cada siniestro y además el porcentaje de probabilidad de ocurrencia para cada una de estas, los cuales sirvieron para ponderar el porcentaje global de probabilidad de ocurrencia para su respectivo tipo de factor.

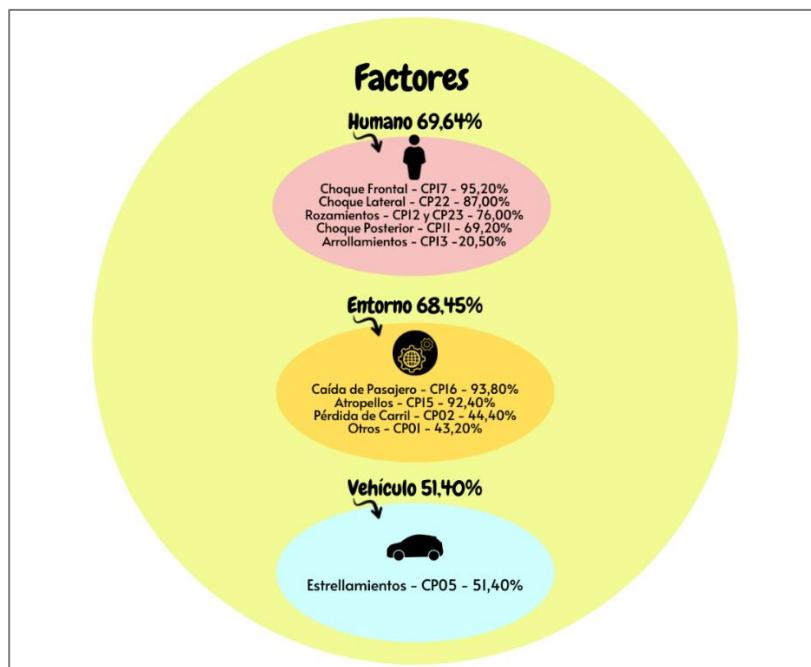


Fig. 54 Resultados de las probabilidades de ocurrencia en relación a los tipos de factores

En la Fig. 55 se muestra que los factores más influyentes para la ocurrencia de siniestros de tránsito en Ecuador en el año 2020 son los factores humanos con una probabilidad de ocurrencia del 69,64%, esto seguido del factor entorno con una probabilidad del 68,45% y por último el factor vehículo con un total del 51,40% de probabilidad de ocurrencia.

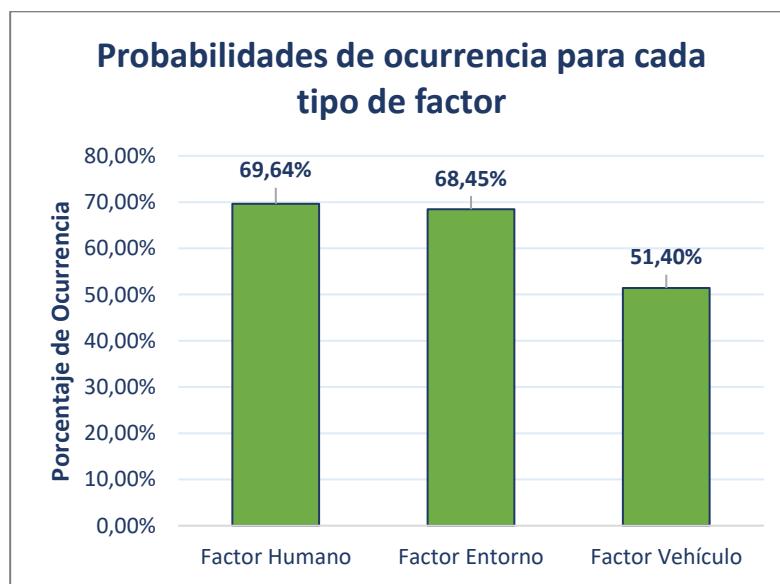


Fig. 55 Factores influyentes para la ocurrencia de siniestros de tránsito en Ecuador

Tarea 3: Elaborar el documento final

En esta tarea se elaboró el documento final a través de la documentación de todo el proceso para el desarrollo del trabajo de titulación “Minería de datos para determinar los factores más influyentes en la ocurrencia de Siniestros de Tránsito en Ecuador en el año 2020” y del mismo modo se llevó a cabo la realización de un informe(ver [Repositorio²⁴](#)) en donde se incluyen los resultados obtenidos después de haber desarrollado el presente trabajo de titulación, dicho documento fue entregado al Director Estratégico Municipal del Cuerpo de Agentes Civiles de Tránsito de la Unidad de Control Operativa de Tránsito (UCOT) del GAD municipal de Loja (ver Anexo 9), esto con el fin de que la información obtenida, sea aprovechada en beneficio de la sociedad y además le sirva como apoyo para el desempeño de sus objetivos como institución.

²⁴ https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/tree/main/informe

7. DISCUSIÓN

A través de la realización del presente TT se propuso el uso de minería de datos como apoyo para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

7.1. Desarrollo de la propuesta alternativa

Objetivo 1: Identificar los repositorios donde se encuentra almacenada la información sobre los siniestros de tránsito en Ecuador en el año 2020.

Para el cumplimiento del primer objetivo, lo primero que se realizó fue proponer un protocolo de búsqueda para identificar las bases de datos relacionadas al objeto de estudio, esto a través del establecimiento de lineamientos relacionados especificados mediante criterios, para poder obtener información relevante sobre siniestros de tránsito en Ecuador en el año 2020. Este protocolo inicialmente ayudó a enfocar la búsqueda de las bases de datos a utilizar, sin embargo, no tuvo ningún contribución en la selección de la base de datos acorde al cumplimiento de los criterios establecidos, debido a que dicho protocolo no fue ejecutado, ya que no fue posible evaluar y comparar cada una de las características establecidas en los criterios, pues solo se encontró una base de datos relacionada al objeto de estudio. A pesar de que no fue evaluado el protocolo, se obtuvo una base de datos consolidada con información optima, confiable, precisa y actualizada la cual permitió el cumplimiento del objeto de estudio. Estos resultados difieren a los del estudio [8], en el cual se realizó la integración de múltiples fuentes de datos separadas, a través del establecimiento de un proceso que lea los datos de dichas diferentes fuentes, los limpie y los aadecue a la estructura que tiene el data warehouse²⁵ para su almacenamiento, este tipo de proceso fue llevado a cabo mediante un sistema conocido como sistema ETL²⁶.

Con respecto a la depuración de la base de datos obtenida, esta fue realizada con el objetivo de filtrar aquellos datos que no fueron relevantes para el análisis posterior, inicialmente se filtraron atributos mediante el análisis de las variables en base al análisis exploratorio de los datos, como resultado fueron seleccionadas trece variables relevantes, de igual manera por medio de la herramienta OpenRefine se estandarizó el conjunto de datos y en R se realizó un filtrado de

²⁵ Data Warehouse.- es un sistema que apoya el procesamiento de la información proporcionando una sólida plataforma de datos integrados e históricos a partir de los cuales se puede realizar un análisis [56].

²⁶ Extraction-Transformation-Load (Extracción-Transformación-Carga)

registros con el fin de eliminar registros almacenados que afectan al proceso de descubrimiento del conocimiento. Esto en contraste con el trabajo relacionado [30], en el cual se ejecutó un proceso similar al aplicado en el presente TT, en relación a la depuración del conjunto de datos utilizado, realizando un análisis para seleccionar los atributos relacionados a las principales causas de los accidentes, esto con el fin de identificar patrones en la ocurrencia de siniestros de tránsito.

La limitación más importante que se ha presentado en el TT, durante el desarrollo de esta fase, fue el trabajar solo con una base de datos que contenía información referente a solo un año en específico, debido a que no fue posible conocer como fue la ocurrencia de siniestros de tránsito en años atrás, esto con la posibilidad de haber extrapolado y mejorado los resultados actuales.

Objetivo 2: Aplicar técnicas de minería de datos a la base de datos obtenida para determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

Durante la presente etapa se aplicó los algoritmos predictivos AD CHAID, AD CHAID Exhaustivo, AD CRT, RN Perceptrón Multicapa, RN de Base de Función Radial, Naive Bayes y BayesNet, mediante las herramientas SPSS Statistics y Weka; si bien actualmente existen muchos algoritmos que se pueden aplicar al proceso de minería de datos, considerando que la elección de los mismos depende del problema en estudio y de las variables que contenga el conjunto de datos, se optó por los antes mencionados debido a la literatura encontrada en la que mencionan que los algoritmos de minería de datos más utilizados en el campo de la seguridad vial, analizando siniestros de tránsito son los Árboles de Decisión, las Redes Neuronales y las Redes Bayesianas.

Lo mismo sucede al comparar los resultados con trabajos relacionados [27] [28] [33] [36] [38], en donde se realizó una aplicación bastante similar de los algoritmos predictivos antes mencionados, mediante Weka, R y RapidMiner, acotando que no en todos los estudios aplicaron específicamente las variantes de árboles de decisión como CHAID, CHAID Exhaustivo o CRT, en unos trabajos aplican el algoritmo Random Forest y en otros el algoritmo J 48 y C4.5, resaltando que en todas las aplicaciones se configuraban los criterios más óptimos con el fin de identificar patrones y relaciones entre las variables para obtener resultados concretos, muy útiles para ayudar a la toma de decisiones; dentro de la implementación del TT se debe destacar, que todo el proceso de la aplicación de los algoritmos predictivos, tuvo el propósito de determinar los factores más influyentes para la ocurrencia de

siniestros de tránsito en Ecuador en el año 2020 con la posible finalidad de ayudar a la toma de decisiones de las instituciones encargadas de la seguridad vial en el país, para poder solventar el problema de salud pública en el que se ha convertido la ocurrencia de los siniestros de tránsito, con el fin de disminuir el impacto devastador que afecta a las personas y más importante el poder mejorar la calidad de vida de la sociedad.

Además, se expone que la ejecución de los algoritmos predictivos en las herramientas SPSS Statistics y Weka resultó más rápida y sencilla que al hacerlo en un lenguaje de programación, debido a que se la puede realizar mediante la interfaz gráfica que ofrecen estas herramientas.

La limitación que se encontró durante el desarrollo de esta fase, fue que, a pesar de que se han utilizado técnicas predictivas para obtener información importante y predominante dentro de registros de siniestros de tránsito es posible aplicar al conjunto de datos técnicas descriptivas, las cuales brindarán un enfoque a futuro y al mismo tiempo, servirán como sustento a los patrones que se han determinado y de esta manera extrapolar los resultados desde otro enfoque.

Objetivo 3: Interpretar y presentar los resultados obtenidos sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador.

Realmente pocos son los estudios que muestran de manera detallada la etapa de evaluación de los algoritmos de minería de datos aplicados. Es por ello, que en el presente TT se realizó el proceso de evaluación, mediante la comparación de las métricas de rendimiento basadas en la matriz de confusión generada por cada algoritmo aplicado, estas métricas fueron el porcentaje global de instancias clasificadas correctamente y el porcentaje de precisión global especificado para cada categoría de la variable objetivo, se eligió este proceso por que a través de este se compararon todos los algoritmos aplicados, mediante la utilización de tablas y gráficos que muestran los porcentajes de rendimiento para cada uno de ellos, con el fin de elegir el algoritmo que presentó mejores resultados. De la misma forma, este proceso es utilizado en el trabajo [31], en el cual compararon todos los modelos aplicados con sus diferentes algoritmos y técnicas en una tabla que mostraba valores de métricas con respecto a su rendimiento, teniendo en cuenta las mismas métricas en todos los modelos, para proporcionar una comparación de porcentajes entre ellos.

Los resultados plasmados mediante los gráficos presentados fueron realizados en base al algoritmo CHAID Exhaustivo, considerando que este obtuvo los porcentajes

más altos, específicamente 58,38% de clasificación correcta y 44,60% de precisión. Al contrastar los resultados antes mencionados con el estudio relacionado [31], estos difieren, debido a que en este estudio un algoritmo de Red Bayesiana obtuvo mejores resultados con respecto a métricas de clasificación correcta y precisión, dejando en segundo lugar al algoritmo CHAID Exhaustivo.

En el presente TT, a través de las reglas generadas por el algoritmo CHAID Exhaustivo, se determinó que la causa probable con más probabilidad de ocurrencia con un 95,20% fue conducir en sentido contrario a la vía normal de circulación, mientras que la causa probable de conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distracto) fue la que presentó la menor probabilidad de ocurrencia con un 20,50%, destacando que estas dos causas probables fueron categorizadas dentro del Factor Humano, añadiendo que este factor fue el más influyente en la ocurrencia de siniestros de tránsito, al igual que el estudio [27] en donde determinan de la misma forma, que el Factor Humano tiene un mayor efecto de ocurrencia de siniestros de tránsito. Por el contrario, el estudio [8] difiere con los resultados, al presentar a la causa relacionada a conducir con falta de atención a las condiciones de tránsito, como la que obtuvo un mayor nivel de soporte y confianza, esto a través de la aplicación del algoritmo árbol de decisión C4.5. con una precisión del 58,00%, pero concuerda en haberla categorizado dentro del Factor Humano, y que este también fue el más influyente.

Es importante mencionar que la definición de reglas a partir de la minería de datos es más efectiva que analizar información con un simple análisis exploratorio de los datos, debido a esto, fue necesario realizar una diferenciación a través de la presentación de los resultados mediante perfiles de conductores y peatones, estructurando cada regla encontrada con respecto a cada clase de siniestro de tránsito en tablas, con el fin de una presentación más técnica.

7.2. Valoración técnica, económica, ambiental y social

Valoración técnica

El presente TT se valora técnicamente mediante la integración de múltiples herramientas tecnológicas que trabajan conjuntamente con el objetivo de determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador. Herramientas como SPSS Statistics y Weka fueron primordiales ya que permitieron la aplicación de los algoritmos de clasificación de AD CHAID, AD CHAID

Exhaustivo, AD CRT, RN Perceptrón Multicapa, RN de Base de Función Radial, Naive Bayes y BayesNet, además se integran Openrefine y RStudio utilizadas para la limpieza y depuración del conjunto de datos, dando como resultado un conjunto de datos estandarizado y depurado lo que aportó a mejorar el desempeño de los algoritmos aplicados.

Valoración económica

Para el desarrollo y cumplimiento del presente TT, fueron necesarios ciertos recursos económicos, los cuales se detallan en el siguiente presupuesto establecido a través de ciertas categorías tales como Recursos Humanos (ver TABLA L), Recursos Técnicos y Tecnológicos (ver TABLA LI), y por último de Servicios (ver TABLA LII).

TABLA L
RECURSOS PARA TALENTO HUMANO

Talento Humano			
Responsable	Número de horas	Costo por hora	Costo total
Tesista	400	\$4.00	\$1,600.00
Director	20	\$10.47	\$209.40
TOTAL			\$1,809.40

TABLA LI
RECURSOS TÉCNICOS Y TECNOLÓGICOS

Recursos Técnicos y Tecnológicos	
Recursos de Software	
Nombre	Costo total
LibreOffice	\$0.00
Mendeley Desktop	\$0.00
Google Chrome	\$0.00
OpenRefine	\$0.00
RStudio	\$0.00
IBM SPSS Statistics	\$0.00
Weka	\$0.00
Firma electrónica	\$27.45
SUBTOTAL	\$ 27.45
Recursos de Hardware	

Nombre	Cantidad	Costo Unitario	Costo total
Laptop	1	\$1,700.00	\$1,700.00
SUBTOTAL			\$1,700.00
TOTAL			\$1,727.45

TABLA LII
RECURSOS PARA SERVICIOS

Servicios			
Nombre	Meses	Costo unitario	Costo total
Internet	5	\$25.00	\$125.00
Transporte	5	\$10.00	\$50.00
TOTAL			\$175.00

Conforme a todos los recursos económicos anteriormente presentados, se genera el Presupuesto General (ver TABLA LIII), en el que se presenta la sumatoria de cada uno de los recursos categorizados, mostrando el valor total de los gastos que llevó ejecutar el presente TT.

TABLA LIII
TOTALIDAD DE LOS RECURSOS ECÓMICOS

Presupuesto General	
Recursos	Costo total
Talento Humano	\$1,809.40
Técnicos y Tecnológicos	\$1,727.45
Servicios	\$175.00
TOTAL	\$3,711.85

Valoración ambiental

Con respecto al ámbito ambiental, el presente TT se ejecutó en su totalidad con recursos digitales y tecnológicos, los cuales no tienen un mayor impacto al entorno ambiental, además se tuvo un bajo consumo de recursos materiales o de otros elementos que puedan llegar a dañar al planeta.

Valoración social

El presente TT tiene una fuerte valoración social, ya que gracias a este se determinó cuáles son los factores más influyentes para la ocurrencia de siniestros de tránsito

en Ecuador, lo que acorde a varias entrevistas como, por ejemplo, la realizada al Cnel. Paul Aguilar que afirma que la mayoría de siniestros de tránsito ocurren debido a causas implicadas dentro del factor humano. Es por esta razón por lo cual, con los resultados obtenidos a través del desarrollo del presente TT se está dando una fuente confiable para que las autoridades de instituciones competentes en esta área, tomen medidas a su criterio que conlleven a generar concientización en la población con respecto a la seguridad vial y además les permita implementar estrategias para reducir la ocurrencia de siniestros de tránsito (ver Anexo 9).

8. CONCLUSIONES

De acuerdo al trabajo de titulación realizado, se puede concluir lo siguiente:

- A través de la aplicación de minería de datos, utilizando técnicas predictivas, como son los Árboles de Decisión, Redes Neuronales y Redes Bayesianas, fue posible determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador en el año 2020, dando como resultado que el Factor Humano es el factor más influyente con una probabilidad de ocurrencia del 69,64%, implicando a seis causas probables principales que son: conducir en sentido contrario a la vía normal de circulación, no respetar las señales reglamentarias de tránsito, no guardar la distancia lateral mínima de seguridad entre vehículos, no respetar las señales manuales del agente de tránsito, el no mantener la distancia prudencial con respecto al vehículo que le antecede y por último conducir desatento a las condiciones de tránsito, esto en concordancia con las entrevistas realizadas a autoridades de la Unidad de Control Operativa de Tránsito (UCOT) del GAD Municipal de Loja, con el fin de sustentar el presente TT, en donde mencionan que el Factor Humano, es el más influyente para que en la ocurrencia de siniestros de tránsito, esto debido a la imprudencia, la impericia y la inobservancia de las leyes.
- Los datos obtenidos sobre siniestros de tránsito en Ecuador en el año 2020 a través de la página oficial de la ANT, fueron el principal insumo y tuvieron un aporte muy significativo al desarrollo del presente TT, ya que son datos que fueron recolectados por entes de control gubernamentales de las 24 provincias del país, lo que permitió que la aplicación de la minería de datos presente resultados satisfactorios referentes al objeto de estudio.
- Comparando los resultados de cada algoritmo predictivo, se concluye que el algoritmo con mejores resultados de rendimiento con respecto a porcentajes de clasificación correcta de las instancias y de precisión con valores de 58,38% y 44,60% respectivamente es el árbol de decisión CHAID Exhaustivo, el cual permitió determinar cuáles fueron los factores más influyentes para que ocurran siniestros de tránsito en Ecuador en el año 2020, mostrando su contexto de ocurrencia y además permitió la creación de perfiles de conductores y peatones involucrados en cada clase de siniestro de tránsito.
- La categorización de los tres factores fue realizada en relación a las categorías presentes en la variable “CAUSA_PROBABLE”, estas fueron diferenciadas de

acuerdo al factor humano, factor vehículo y factor entorno, esto con el propósito de tratar más concretamente a las causas probables, para poder cumplir con el objetivo de identificar patrones utilizados para la toma de decisiones, con respecto en la educación de los conductores y peatones para prevenir la ocurrencia de siniestros de tránsito.

9. RECOMENDACIONES

En base al Trabajo de Titulación realizado, se puede dar las siguientes recomendaciones:

- Al obtener el conjunto de datos se debe considerar la evaluación de cada de las variables contenidas en este, esto debido a que pueden existir datos erróneos o redundantes, además que permite conocer la manera en que está estructurada la información
- Usar OpenRefine para la limpieza o depuración de la información, ya que permite reemplazar o eliminar valores, estandarizar el conjunto de datos y optimizar la calidad del mismo de manera más rápida y sencilla.
- Para trabajos similares usar los diferentes tipos de algoritmos de árboles de decisión encaminados a encontrar información única y relevante, ya que son muy beneficiosos o eficientes para encontrar patrones o relaciones fiables entre las variables ya que proporcionan diversos criterios de configuración para obtener los resultados más óptimos.
- Para realizar una mejor evaluación de resultados, se sugiere utilizar varias métricas de rendimiento en relación a la matriz de confusión, como el porcentaje de instancias clasificadas correctamente y el porcentaje de precisión ya que en algunos casos una no es suficiente.
- Para la presentación de resultados se recomienda hacerlo de la manera más sencilla y simplificada posible dada la complejidad del problema esto con el fin de que puedan ser entendidos con facilidad.
- A las instituciones que velan por la seguridad vial se recomienda estandarizar la forma en la que se almacenan los datos sobre los registros de los siniestros de tránsito ocurridos en Ecuador.
- Se recomienda el uso de la herramienta SPSS Statistics para la aplicación de minería de datos, debido a su simple funcionamiento, fácil manejo y por qué mejora tanto el rendimiento como el despliegue de los resultados.
- Para cargar, filtrar, manipular datos y capacitar e implementar modelos de minería de datos se recomienda utilizar sistemas de computación de alto rendimiento, con nodos de computación de 32 GB de RAM, además, que cuente con CPUs de 12 núcleos a 2,60 GHz, esto con la finalidad de que todos los procesos se ejecuten de una manera más rápida y eficaz.

Trabajos futuros

- Para futuros trabajos se pueden considerar nuevas variables o factores, de manera que esto permita obtener un mayor grado de precisión y confiabilidad de los modelos generados.
- Con referencia al trabajo ya realizado en el presente TT, se puede aplicar otros algoritmos predictivos, al igual que descriptivos, para mejorar los resultados obtenidos, además de que se generan nuevos resultados para ser comparados con los actuales, además que se amplía el rango de soluciones y métricas al presente proyecto.
- Implementar una solución de software capaz de automatizar el proceso de determinar los factores más influyentes para que ocurran siniestros de tránsito en Ecuador, esto con relación a las variables principales involucradas en la ocurrencia de cada clase de siniestro para que de esta manera una persona pueda saber la probabilidad de estar involucrado en un siniestro de tránsito.

10. BIBLIOGRAFÍA

- [1] Organización Mundial de la Salud, "Informe sobre la situación mundial de la seguridad vial 2018," Suiza, 2018.
- [2] T. INEC, "Anuario de Estadísticas de Transporte 2019," Ecuador, Dec. 2020.
- [3] M. M. Rodriguez Hassiger, "APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN ACCIDENTES DE TRÁFICO," Universidad Politécnica de Valencia, 2014.
- [4] M. Beatriz Beltrán Martínez, "MINERÍA DE DATOS," México.
- [5] M. I. Á. Larrieta and A. M. Santillán Gómez, "Minería de datos: Concepto, características, estructura y aplicaciones."
- [6] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques. Third Edition*. San Francisco: CA: Morgan Kaufmann publishers., 2014.
- [7] M. Servente, "Algoritmos TDIDT aplicados a la Minería de Datos Inteligente," 2002.
- [8] A. Pumares, "Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador," Universidad Tecnológica Israel, Quito, 2019.
- [9] G. López Maldonado, "Análisis de la severidad de los accidentes de tráfico utilizando Técnicas de Minería de Datos," Universidad de Granada, 2013.
- [10] J. A. Lara Torralbo, "Marco de Descubrimiento de Conocimiento para Datos Estructuralmente Complejos con Énfasis en el Análisis de Eventos en Series Temporales," UNIVERSIDAD POLITÉCNICA DE MADRID, 2017.
- [11] D. Hasbleidy, C. Díaz, D. Felipe, and S. Vargas, "Análisis de accidentalidad vehicular usando técnicas de minería de datos," Universidad Distrital Francisco José de Caldas, 2019.
- [12] M. Beatriz Beltrán Martínez, "MINERÍA DE DATOS."
- [13] C. Espino Timón and X. Martínez Fontes, "Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso," Universitat Oberta de Catalunya, 2017.
- [14] S. L. González-Ruiz, I. Gómez-Gallego, J. L. Pastrana-Brincones, and A. Hernández-Mendo, "Algoritmos de clasificación y redes neuronales en la observación automatizada de registros Classification algorithms and neural networks in automated observation records," *Cuad. Psicol. del Deport.*, vol. 15, no. 1, pp. 31–40, 2015, doi: 10.4321/S1578-84232015000100003.
- [15] D. Delgado Castillo, R. Martín Pérez, L. Hernández Pérez, R. Orozco Morález, and J. Lorenzo Ginori, "Algoritmos de aprendizaje automático para la clasificación

de neuronas piramidales afectadas por el envejecimiento Machine learning algorithms for classification of pyramidal neurons affected by aging," vol. 8, 2016, Accessed: Sep. 03, 2021. [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18592016000300008.

- [16] A. N. Girón Gómez and L. S. Patarroyo Chaparro, "‘CSISNE’ Plugin para la Clasificación Supervisada de Imágenes Satelitales Mediante el Uso del Algoritmo Perceptrón Multicapa Basados en Redes Neuronales. - hdl:11349/5182," Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, 2016.
- [17] R. Rivera Camacho, R. Barrón Fernández, and A. G. Arenas, "Modelado y propagación de valores de sentimiento en relaciones de usuario Modelling and Propagation of Sentiment Values in Relations between Users," *Res. Comput. Sci.*, vol. 107, pp. 9–17, 2015.
- [18] I. Rufaza Esteve, "DEPURACIÓN DE UNA BASE DE DATOS," Universidad Complutense de Madrid, 2018.
- [19] S. Toscano Vizcaíno, "Derecho Ecuador - Qué es un accidente de tránsito," Nov. 2005.
- [20] N. W. Zamani and S. S. M. Khairi, "A comparative study on data mining techniques for rainfall prediction in Subang," in *AIP Conference Proceedings*, 2018, vol. 2013.
- [21] L. E. Sucar, "Redes Bayesianas," México.
- [22] "OpenRefine." <https://openrefine.org/> (accessed Jul. 11, 2021).
- [23] P. Larsson, "Evaluation of Open Source Data Cleaning Tools : Open Refine and Data Wrangler," 2013.
- [24] "RStudio ." <https://www.rstudio.com/> (accessed Jul. 11, 2021).
- [25] IBM, "SPSS Software | IBM," 2021. <https://www.ibm.com/analytics/spss-statistics-software> (accessed Jul. 11, 2021).
- [26] C. A. Escobar Canizales, S. A. Rodríguez Rubiano, and J. C. Vega Figueroa, "Modelo Big Data, aplicando análisis de datos y algoritmos predictivos, basado en la inteligencia computacional, para predecir la probabilidad de los accidentes de tránsito en la ciudad de Medellín," Universidad Cooperativa de Colombia, 2020.
- [27] B. Atnafu and G. Kaur, "Analysis and Predict the Nature of Road Traffic Accident Using Data Mining Techniques in Maharashtra, India," *Gagandeep Kaur Int. J. Eng. Technol. Sci. Res. IJETSR www.ijetsr.com ISSN*, vol. 4, no. 1, pp. 2394–3386, 2017.

- [28] H. Ospina-Mateus and L. A. Quintana Jiménez, "Predicción de accidentes viales en Cartagena, Colombia, con árboles de decisión y reglas de asociación," *Econ. Región*, vol. 13, no. 2, pp. 83–115, 2019.
- [29] K. P. P, "Road Accident Prediction Using Data Mining Techniques," vol. 3, no. 4, pp. 492–495, 2019.
- [30] R. Gomes and S. Barcellos, "BRAZILIAN FEDERAL ROADS: IDENTIFYING PATTERNS IN TRAFFIC ACCIDENTS USING DATA MINING TECHNIQUES WITH APRIORI ALGORITHM," no. December, 2020.
- [31] 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, 2020, doi: 10.1016/j.isatra.2020.06.018.
- [32] C. Horn, "Traffic Accidents Prediction Using Ensemble Machine Learning Approach MSc Research Project Monisha Lakshme Gowda School of Computing National College of Ireland."
- [33] Z. Yuan, X. Zhou, T. Yang, J. Tamerius, and R. Mantilla, "Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study," *Urban Comput.*, pp. 1–9, 2017.
- [34] A. Irfan, R. Al Rasyid, and S. Handayani, "Data mining applied for accident prediction model in Indonesia toll road," *AIP Conf. Proc.*, vol. 1977, no. June, 2018.
- [35] S. Hussain, L. J. Muhammad, F. S. Ishaq, A. Yakubu, and I. A. Mohammed, *Performance evaluation of various data mining algorithms on road traffic accident dataset*, vol. 106. Springer Singapore, 2019.
- [36] A. Makkar, H. S. Gill, M. T. Scholar, and C. Science, "A Radical Approach to Forecast the Road Accident Using Data Mining Technique," vol. 2, no. 8, 2017.
- [37] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, vol. 2018-Novem, pp. 3346–3351, 2018.
- [38] R. E. Almamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," *2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2019 - Proc.*, pp. 272–276, 2019.
- [39] N. López and I. Sandoval, "Métodos y técnicas de investigación cuantitativa y cualitativa," México, 2017.
- [40] Y. Torres Quezada, "Entrevista al Cnel. Paul Aguilar - Agt. Patricio Benítez,"

- Google Drive*, 2021. <https://drive.google.com/file/d/1IkTc3ifvY1S1BQY6iQE0xI7-Ds1f-Pmm/view?usp=sharing> (accessed Aug. 31, 2021).
- [41] J. L. Abreu, "El Método de la Investigación," *Daena Int. J. Good Conscienc.*, vol. 9, no. 3, pp. 195–204, 2014.
 - [42] F. E. V. Valdés Medina, M. del C. Hernández Silva, M. T. Martínez Contreras, and J. Y. Gomora Miranda, "ANTOLOGÍA DE METODOLOGÍA DE LA INVESTIGACIÓN," México, 2019.
 - [43] C. Espinoza Montes, *Metodología de la Investigación Tecnológica*. 2010.
 - [44] C. Espinoza Montes, *Metodología de investigación tecnológica Pensando en sistemas*, Primera Edición. Perú: Imagen Grafica SAC, 2010.
 - [45] ANT, "Siniestros de Tránsito," 2020.
 - [46] ANT, "Reporte Nacional de Siniestros de Tránsito," Ecuador, 2020.
 - [47] "REGLAMENTO A LEY DE TRANSPORTE TERRESTRE TRANSITO Y SEGURIDAD VIAL," Ecuador, Jun. 2012.
 - [48] Agencia Nacional de Tránsito, "EMISIÓN DE PERMISO DE CONDUCCIÓN PARA MENOR ADULTO (LICENCIA NO PROFESIONAL)," Ecuador. Accessed: Aug. 02, 2021. [Online]. Available: <https://www.gob.ec/ant/tramites/emision-permiso-conduccion-menor-adulto-licencia-no-profesional>.
 - [49] "Criterios para CHAID - Documentación de IBM." <https://www.ibm.com/docs/es/spss-statistics/26.0.0?topic=criteria-chaid> (accessed Aug. 03, 2021).
 - [50] "Criterios para CRT - Documentación de IBM." <https://www.ibm.com/docs/es/spss-statistics/26.0.0?topic=criteria-crt> (accessed Aug. 04, 2021).
 - [51] C. L. Corso, "Aplicación de algoritmos de clasificación supervisada usando Weka."
 - [52] N. V. Constante Tipán, "Accidentes de Tránsito producidos por Imprudencia y Negligencia de Conductores y Peatones en la Avenida Simón Bolívar del DMQ, Año 2016," UNIVERSIDAD CENTRAL DEL ECUADOR, Quito, 2017.
 - [53] L. A. Pulgarin Crespo, "Análisis De Los Accidentes De Tránsito En La Ciudad De Cuenca Para Los Años 2010 - 2011 - 2012," p. 82, 2014, [Online]. Available: <http://dspace.ucuenca.edu.ec/handle/123456789/19861>.
 - [54] E. F. Calle Reinoso and L. T. Sarabia Paucay, "DESARROLLO DE UNA BASE DE DATOS PARA EVALUAR LA PERCEPCIÓN DE LA SEGURIDAD VIAL EN ECUADOR," Universidad Politécnica Salesiana, Cuenca, 2020.

- [55] D. X. Román Matamoros, "INTEGRACIÓN DE UN PROGRAMA DE SEGURIDAD VIAL AL MODELO ECUADOR," Universidad San Francisco de Quito-Ecuador, Universidad de Huelva-España, Quito, 2015.
- [56] U. Aftab and G. F. Siddiqui, "Big Data Augmentation with Data Warehouse: A Survey," *2018 IEEE Int. Conf. Big Data (Big Data)*, pp. 2785–2794, Dec. 2018, doi: 10.1109 / BigData.2018.8622206.

11. ANEXOS

Anexo 1

Entrevistas realizadas con la finalidad de justificar y sustentar la realización del presente TT, tanto a nivel académico como social. Como apoyo de la información captada en las entrevistas, se anexa las grabaciones de las mismas, en el siguiente enlace:
<https://drive.google.com/file/d/1IkTc3ifvY1S1BQY6iQE0xI7-Ds1f-Pmm/view?usp=sharing>

A continuación, se redacta puntualmente las respuestas por sus actores

Entrevista 1

Cargo: Director Estratégico Municipal del Cuerpo de Agentes Civiles de Tránsito de la Unidad de Control Operativa de Tránsito (UCOT) del GAD Municipal de Loja

Nombre: Abg. Cnel. Paul Mauricio Aguilar Sotomayor

Descripción:

1. ¿Considera a la ocurrencia de siniestros de tránsito como un problema social?

Si, por que dentro de la movilidad existen los actores y el actor fundamental son los peatones, los conductores y los pasajeros, desde este punto de vista constituye una situación social y una problemática social que necesita de la cultura y de la educación dentro del ámbito de la movilidad.

2. ¿Qué opinión tiene usted acerca de la ocurrencia de siniestros de tránsito en Ecuador?

Que deberíamos clasificarlos por las diferentes formas en la que se suscitan los accidentes de tránsito, en muchas de ellas, si bien es cierto respetamos lo que estipula tanto el código orgánico integral penal en lo que respecta a las contravenciones de tránsito y también la ley de tránsito, no comarto en muchas de ellas, por ejemplo, cuando habla que todos los accidentes de tránsito son culposos, si tomamos en consideración que una persona que sale a ingerir bebidas alcohólicas y conduce su vehículo bajo los efectos del alcohol deja de ser culposo y se convierte en dolo, lo que esta persona ocasiona, pero sin embargo los han determinado así como culposos y así está estipulado en la ley y eso es lo que respetamos nosotros al momento de poner las sanciones y de elevar los partes correspondientes para que la fiscalía y los juzgados tomen acciones.

3. ¿Usted cuáles cree que serían los factores más influyentes para que ocurran siniestros de tránsito en Ecuador?

Bueno, básicamente tomando en consideración el análisis local dentro de los tres factores que existen le daríamos al factor humano como el más influyente para que ocurran siniestros de tránsito, tomando en consideración que si hacemos las estadísticas solo del cantón Loja el 98% es factor Humano y ese 2% restante se divide en lo que es el factor ambiental y el factor de fallas técnica mecánicas y este factor humano se distribuye en diferentes situaciones como tal, lo estipula incluso la ley que son la imprudencia, la impericia y la inobservancia de las leyes.

4. Realizar un estudio sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador ¿le aportaría de alguna manera?

Siempre todo estudio permite hacer estrategias y con las estrategias se estipula acciones para reducir los accidentes de tránsito, desde esa perspectiva toda acción que permita implementar estrategias dentro de los entes de control es productiva para reducir los accidentes de tránsito.

Entrevista 2

Cargo: Agente Civil de Tránsito de la Unidad de Control Operativa de Tránsito (UCOT) del GAD Municipal de Loja

Nombre: Sr. Patricio Bolívar Benítez Lanche

Descripción:

1. ¿Considera a la ocurrencia de siniestros de tránsito como un problema social?

Bueno los accidentes de tránsito, si es un problema, ciertamente por el problema social, por lo que implica a un conductor, un peatón, hasta el mismo ocupante de un vehículo, como pasajero, entonces, un problema muy conocido es que no hay mucha educación respecto a este tema vial, entonces pienso que si es un problema social

2. ¿Qué opinión tiene usted acerca de la ocurrencia de siniestros de tránsito en Ecuador?

Bueno, en el país prácticamente en el país la accidentología vial es considerada como una de las causas más frecuentes para vulnerar se podría decir a las personas, como lo manifesté anteriormente, conductores, pasajeros, o hasta la misma parte pública y

privada es afectada por este motivo, entonces considero que los accidentes de tránsito es una de las causas más comunes en lo que es el país.

3. ¿Usted cuáles cree que serían los factores más influyentes para que ocurran siniestros de tránsito en Ecuador?

Bueno, hay que ser sinceros, el factor humano, sería uno de los factores más influyentes y aunque tenemos al factor vial o también conocido como ambiental y el factor vehículo, consideremos que la mayoría de siniestros de tránsito se producen por que el conductor se encuentra con halitosis alcohólica, exceso de velocidad, mal usos del teléfono celular mientras conduce, estos factores prácticamente implican dentro del factor humano, entonces se podría considerar este como uno de los factores que más influye al momento de ocasionarse un siniestro de tránsito, mientras que sin embargo con respecto a los vehículos son muy pocas las veces que ocurren muchos accidentes de tránsito con respecto a digamos sean fallas técnicas mecánicas del vehículo o tal vez por el factor ambiental, ya sean por accidentes atípicos, como deslaves o ese factor, pero bueno más se considera al factor humano.

4. Realizar un estudio sobre los factores más influyentes para que ocurran siniestros de tránsito en Ecuador ¿le aportaría de alguna manera?

Por supuesto, prácticamente si es que no hubiera algún tipo de información, nosotros desconoceríamos el sector donde más hay ocurrencia de siniestros de tránsito, por lo general la ciudadanía siempre se basa en lo que es, la información en la que uno como entidad realiza y el trabajo con el que labora, siempre se va a hacer esto indispensable para que la gente mismo tenga seguridad, se podría decir al momento de trasladarse a cierto lugar, entonces, por lo general sería que, bueno con un estudio adecuado y si se pudiera implementar de forma más óptima, sería un beneficio a cualquier institución pública en este caso.

El respaldo formal de las entrevistas realizadas se especifica en la siguiente página adjunta.

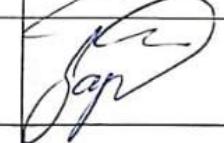


Universidad Nacional de Loja

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

Carrera de Ingeniería en Sistemas

Entrevista realizada por: Yulissa Stefania Torres Quezada

Nombres y Apellidos	Cédula	Institución/Empresa	Actividad en la Institución/Empresa	Firma
Patricia Bolívar Benítez Lancha	1105665044	GAD Municipal de Loja	Agente Civil de Tránsito	
Rebel Moncayo Aguirre Estévez	11029350746	UCOT - Municipio de Loja	Director.	

Anexo 2

1	MES	AÑO	CÓDIGO	ENTE DE CÓPROVINCIA	ZONA PLANDIA	FECHA	HORA	PERIODO	FERIADO	CÓDIGO CCAUSA PRÓCLASE FINAZONA	LATITUD (Y)	LONGITUD	DIRECCIÓN CANTÓN	PAR	
2	ENERO	2020	CTE178112CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	01:10:00	01	SI	C14	CONDUCIR ATROPELLC URBANA	-2,165205	-79,843158	AV NICOLÁS DURAN ELO
3	ENERO	2020	CTE178115CCTE	MANABÍ	ZONA 4	miércoles	01/01/2020	05:15:00	05	SI	C23	NO RESPET/CHOQUE L RURAL	-0,963261	-80,439908	RED VIAL E ROCAFUER ROC
4	ENERO	2020	CTE178116CCTE	LOS RÍOS	ZONA 5	miércoles	01/01/2020	03:10:00	03	SI	C11	NO MANTE CHOQUE P RURAL	-0,966081	-79,422838	VA TRANSV QUEVEDO LA E
5	ENERO	2020	CTE178118CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	03:10:00	03	SI	C14	CONDUCIR ATROPELLC RURAL	-1,876616	-80,177966	ISIDRO AY CISIDRO AY CISID
6	ENERO	2020	CTE178121CCTE	SANTO DO	ZONA 4	miércoles	01/01/2020	07:30:00	07	SI	C18	CONDUCIR CHOQUE F RURAL	-0,269831	-79,093383	RED VIAL E SANTO DO S. D
7	ENERO	2020	CTE178122CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	04:30:00	04	SI	C14	CONDUCIR PÉRDIDA D RURAL	-2,236333	-80,113972	RED VIAL E GUAYAQU JUA
8	ENERO	2020	CTE178123CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	04:00:00	04	SI	C18	CONDUCIR CHOQUE F RURAL	-1,202031	-79,757951	CANTON E BALZAR
9	ENERO	2020	CTE178124CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	07:00:00	07	SI	C12	NO GUARD ROZAM/EN RURAL	-2,171958	-79,540744	VIA E488 V MILAGRO
10	ENERO	2020	CTE178126CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	07:00:00	07	SI	C14	CONDUCIR ESTRELLAM URBANA	-2,157301	-79,829626	DURAN AV DURAN ELO
11	ENERO	2020	CTE178129CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	09:00:00	09	SI	C14	CONDUCIR ATROPELLC URBANA	-1,365001	-79,905001	CANTON B BALZAR BAL
12	ENERO	2020	CTE178130CCTE	SANTO DO	ZONA 4	miércoles	01/01/2020	04:00:00	04	SI	C14	CONDUCIR PÉRDIDA D RURAL	-0,244205	-79,291511	RED VIAL E SANTO DO S. D
13	ENERO	2020	CTE178131CCTE	EL ORO	ZONA 7	miércoles	01/01/2020	07:20:00	07	SI	C14	CONDUCIR PÉRDIDA D RURAL	-3,427972	-79,958583	RED VIAL E SANTA ROSSAN
14	ENERO	2020	CTE178132CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	10:30:00	10	SI	C18	CONDUCIR CHOQUE F RURAL	-1,970427	-79,596451	JUJAN VIA ALFREDO BAQU
15	ENERO	2020	CTE178133CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	12:30:00	12	SI	C17	BAJARSE O CAÍDA DE RURAL	-1,113319	-79,677038	CANTON E EL EMPALME
16	ENERO	2020	CTE178134CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	02:01:00	02	SI	C14	CONDUCIR ARROLLAM URBANA	-2,191497	-79,768172	DORAN VI DURAN ELO
17	ENERO	2020	CTE178135CCTE	AZUAY	ZONA 6	miércoles	01/01/2020	03:40:00	03	SI	C14	CONDUCIR PÉRDIDA D RURAL	-3,039055	-79,742111	VA ESTATAL CAMILO PC CAN
18	ENERO	2020	CTE178137CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	13:40:00	13	SI	C11	NO MANTE CHOQUE P URBANA	-2,07622	-79,940469	GUAYAQU GUAYAQUIL
19	ENERO	2020	CTE178138CCTE	LOS RÍOS	ZONA 5	miércoles	01/01/2020	08:10:00	08	SI	C14	CONDUCIR PÉRDIDA D RURAL	-1,077419	-79,601213	VA TRANSV QUEVEDO GU
20	ENERO	2020	CTE178139CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	01:30:00	01	SI	C14	CONDUCIR ESTRELLAM RURAL	-2,256305	-79,594244	RED VIAL E EL TRIUNFO
21	ENERO	2020	CTE178140CCTE	SANTO DO	ZONA 4	miércoles	01/01/2020	14:00:00	14	SI	C11	NO MANTE CHOQUE P RURAL	-0,060047	-79,564333	RED VIAL E LA CONCORDIA
22	ENERO	2020	CTE178142CCTE	SANTA ELEN	ZONA 5	miércoles	01/01/2020	09:40:00	09	SI	C14	CONDUCIR PÉRDIDA D RURAL	-2,234333	-80,825583	CANTN SAN SANTA ELENA
23	ENERO	2020	CTE178143CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	17:20:00	17	SI	C14	CONDUCIR ATROPELLC URBANA	-2,553631	-79,679955	EL TRIUNFO EL TRIUNFO EL TI
24	ENERO	2020	CTE178144CCTE	SANTA ELEN	ZONA 5	miércoles	01/01/2020	09:45:00	09	SI	C18	CONDUCIR CHOQUE F RURAL	-2,008694	-80,641083	CANTN SAN SANTA ELEN COL
25	ENERO	2020	CTE178145CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	03:40:00	03	SI	C14	CONDUCIR ATROPELLC URBANA	-2,329008	-79,397872	AV JUAN N EL TRIUNFO
26	ENERO	2020	CTE178146CCTE	SANTA ELEN	ZONA 5	miércoles	01/01/2020	09:00:00	09	SI	C11	NO MANTE CHOQUE P URBANA	-2,215666	-80,949777	CANTN SASALINAS CAR
27	ENERO	2020	CTE178147CCTE	AZUAY	ZONA 6	miércoles	01/01/2020	15:08:00	15	SI	C18	CONDUCIR CHOQUE F RURAL	-3,115277	-79,110001	RED ESTATAGIRON GIRON
28	ENERO	2020	CTE178148CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	19:00:00	19	SI	C14	CONDUCIR ATROPELLC URBANA	-2,130833	-79,592777	MILAGRO (MILAGRO
29	ENERO	2020	CTE178149CCTE	GUAYAS	ZONA 8	miércoles	01/01/2020	17:15:00	17	SI	C14	CONDUCIR PÉRDIDA D RURAL	-2,195651	-80,055556	GUAYAQU GUAYAQU CHO
30	ENERO	2020	CTE178151CCTE	MANABÍ	ZONA 4	miércoles	01/01/2020	14:15:00	14	SI	C23	NO RESPET/CHOQUE L RURAL	-0,954445	-80,777861	RED VIAL E MANTA MA
31	ENERO	2020	CTE178153CCTE	GUAYAS	ZONA 5	miércoles	01/01/2020	19:45:00	19	SI	C18	CONDUCIR CHOQUE F RURAL	-1,039913	-79,637985	CANTON E EL EMPALME
32	ENERO	2020	CTE178155CCTE	LOS RÍOS	ZONA 5	miércoles	01/01/2020	21:10:00	21	SI	C14	CONDUCIR ATROPELLC RURAL	-1,382677	-79,435508	RFD VIAL E VENTANAS VEN

Fig. 56 Conjunto de datos obtenido, aún sin procesar

Anexo 3

TABLA LIV
DICCCIONARIO DE DATOS DEL CONJUNTO DE DATOS INICIAL

DICCCIONARIO DE DATOS		
Nombre del Conjunto de Datos:	dataset_inicial	
Descripción del Conjunto de Datos:	Contiene información referente a: siniestros de tránsito, en cuanto a causas del siniestro, mes, provincia, número de fallecidos, lesionados, etc., proporcionada por la Agencia Nacional de Tránsito (ANT), referente al 2020	
URL del Conjunto de Datos:	https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/dataset_inicial.xlsx	
Nombre del Campo (Encabezado Columna)	Descripción del Campo	Categoría
MES	Mes de ocurrencia del siniestro	
AÑO	Año de ocurrencia del siniestro	
CÓDIGO	Código del siniestro	
ENTE DE CONTROL	Ente de control que registro el siniestro	
PROVINCIA	Provincia de ocurrencia del siniestro	
ZONA PLANIFICACIÓN	Zona de planificación de ocurrencia del siniestro	
DIA	Día de ocurrencia del siniestro	
FECHA	Fecha de ocurrencia del siniestro	
HORA	Hora de ocurrencia del siniestro	
PERIODO	Código para la hora de ocurrencia del siniestro	00=00:00 A 00:59 01=01:00 A 01:59 02=02:00 A 02:59 03=03:00 A 03:59 04=04:00 A 04:59

		05=05:00 A 05:59 06=06:00 A 06:59 07=07:00 A 07:59 08=08:00 A 08:59 09=09:00 A 09:59 10=10:00 A 10:59 11=11:00 A 11:59 12=12:00 A 12:59 13=13:00 A 13:59 14=14:00 A 14:59 15=15:00 A 15:59 16=16:00 A 16:59 17=17:00 A 17:59 18=18:00 A 18:59 19=19:00 A 19:59 20=20:00 A 20:59 21=21:00 A 21:59 22=22:00 A 22:59 23=23:00 A 23:59
FERIADO	Feriado en el día de la ocurrencia del siniestro	
CÓDIGO CAUSA	Código para la causa probable del siniestro	C01=CASO FORTUITO O FUERZA MAYOR (EXPLOSIÓN DE NEUMÁTICO NUEVO, DERRUMBE, INUNDACIÓN, CAÍDA DE PUENTE, ÁRBOL, PRESENCIA INTEMPESTIVA E IMPREVISTA DE SEMOVIENTES EN LA VÍA, ETC.). C02=PRESENCIA DE AGENTES EXTERNOS EN LA VÍA (AGUA, ACEITE, PIEDRA, LASTRE, ESCOMBROS, MADEROS, ETC.). C03=CONDUCIR EN ESTADO DE SOMNOLENCIA O MALAS CONDICIONES FÍSICAS (SUEÑO, CANSANCIO Y FATIGA).

	<p>C04=DAÑOS MECÁNICOS PREVISIBLES.</p> <p>C05=FALLA MECÁNICA EN LOS SISTEMAS Y/O NEUMÁTICOS (SISTEMA DE FRENIOS, DIRECCIÓN, ELÉCTRÓNICO O MECÁNICO).</p> <p>C06=CONDUCE BAJO LA INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS.</p> <p>C07=PEATÓN TRANSITA BAJO INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS.</p> <p>C08=PESO Y VOLUMEN - NO CUMPLIR CON LAS NORMAS DE SEGURIDAD NECESARIAS AL TRANSPORTAR CARGAS.</p> <p>C09=CONDUCIR VEHÍCULO SUPERANDO LOS LÍMITES MÁXIMOS DE VELOCIDAD.</p> <p>C10=CONDICIONES AMBIENTALES Y/O ATMOSFÉRICAS (NIEBLA, NEBLINA, GRANIZO, LLUVIA).</p> <p>C11=NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO AL VEHÍCULO QUE LE ANTECEDE.</p> <p>C12=NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDAD ENTRE VEHÍCULOS.</p> <p>C14=CONDUCIR DESATENTO A LAS CONDICIONES DE TRÁNSITO (CELULAR, PANTALLAS DE VIDEO, COMIDA, MAQUILLAJE O CUALQUIER OTRO ELEMENTO DISTRACTOR).</p>
--	---

	<p>C15=DEJAR O RECOGER PASAJEROS EN LUGARES NO PERMITIDOS.</p> <p>C16=NO TRANSITAR POR LAS ACERAS O ZONAS DE SEGURIDAD DESTINADAS PARA EL EFECTO.</p> <p>C17=BAJARSE O SUBIRSE DE VEHÍCULOS EN MOVIMIENTO SIN TOMAR LAS PRECAUCIONES DEBIDAS.</p> <p>C18=CONDUCIR EN SENTIDO CONTRARIO A LA VÍA NORMAL DE CIRCULACIÓN.</p> <p>C19=REALIZAR CAMBIO BRUSCO O INDEBIDO DE CARRIL.</p> <p>C20=MAL ESTACIONADO - EL CONDUCTOR QUE DETENGA O ESTACIONE VEHÍCULOS EN SITIOS O ZONAS QUE ENTRAÑEN PELIGRO, TALES COMO ZONA DE SEGURIDAD, CURVAS, PUENTES, TÚNELES, PENDIENTES.</p> <p>C21=MALAS CONDICIONES DE LA VÍA Y/O CONFIGURACIÓN. (ILUMINACIÓN Y DISEÑO).</p> <p>C22=ADELANTAR O REBASAR A OTRO VEHÍCULO EN MOVIMIENTO EN ZONAS O SITIOS PELIGROSOS TALES COMO: CURVAS, PUENTES, TÚNELES, PENDIENTES, ETC.</p> <p>C23=NO RESPETAR LAS SEÑALES REGLAMENTARIAS DE TRÁNSITO. (PARE, CEDA EL PASO, LUZ ROJA DEL SEMAFORO, ETC).</p> <p>C24=NO RESPETAR LAS SEÑALES MANUALES DEL AGENTE DE TRÁNSITO.</p>
--	--

		C25=NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO A VEHÍCULOS. C26=NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO AL PEATÓN. C27=PEATÓN QUE CRUZA LA CALZADA SIN RESPETAR LA SEÑALIZACIÓN EXISTENTE (SEMÁFOROS O SEÑALES MANUALES).
CAUSA PROBABLE	Causa probable del siniestro	
CLASE FINAL	Clase del siniestro	
ZONA	Zona de ocurrencia del siniestro	
LATITUD (Y)	Latitud de ocurrencia del siniestro	
LONGITUD (X)	Longitud de ocurrencia del siniestro	
DIRECCIÓN	Dirección de ocurrencia del siniestro	
CANTÓN	Cantón de ocurrencia del siniestro	
PARROQUIA	Parroquia de ocurrencia del siniestro	
[TIPO DE VEHÍCULO 1 – TIPO DE VEHÍCULO 10]	Tipo de vehículo involucrado en el siniestro	
[SERVICIO 1 – SERVICIO 10]	Servicio del vehículo involucrado en el siniestro	
AUTOMÓVIL	Tipo de vehículo involucrado en el siniestro	
BICICLETA	Tipo de vehículo involucrado en el siniestro	
BUS	Tipo de vehículo involucrado en el siniestro	
CAMIÓN	Tipo de vehículo involucrado en el siniestro	
CAMIONETA	Tipo de vehículo involucrado en el siniestro	
EMERGENCIA	Tipo de vehículo involucrado en el siniestro	
ESPECIAL	Tipo de vehículo involucrado en el siniestro	
FURGONETA	Tipo de vehículo involucrado en el siniestro	
MOTOCICLETA	Tipo de vehículo involucrado en el siniestro	

NO IDENTIFICADO	Tipo de vehículo no identificado	
VEHÍCULO DEPORTIVO UTILITARIO	Tipo de vehículo involucrado en el siniestro	
SUMA DE VEHÍCULOS	Suma de los vehículos involucrado en el siniestro	
NUM_FALLECIDO	Número de lesionados en el siniestro	
NUM_LESIONADO	Número de fallecidos en el siniestro	
[TIPO DE IDENTIFICACIÓN 1 – TIPO DE IDENTIFICACIÓN 52]	Tipo de identificación de la víctima	
[EDAD 1 – EDAD 52]	Edad de la víctima	
[SEXO 1 – SEXO 52]	Sexo de la víctima	
[CONDICIÓN 1 – CONDICIÓN 52]	Condición de la víctima	
[PARTICIPANTE 1 – PARTICIPANTE 52]	Tipo de participante en el siniestro	
[USO DE CASCO 1 – USO DE CASCO 52]	Uso de casco en el siniestro	
[USO DE CINTURÓN DE SEGURIDAD 1 - USO DE CINTURÓN DE SEGURIDAD 52]	Uso de cinturón de seguridad en el siniestro	

Fuente: Agencia Nacional de Tránsito – ANT

Elaborado: Yulissa Stefania Torres Quezada

Anexo 4

TABLA LV
DICCCIONARIO DE DATOS DEL CONJUNTO DE DATOS UTILIZADO

DICCCIONARIO DE DATOS		
Nombre del Conjunto de Datos:	dataset_dep.csv	
Descripción del Conjunto de Datos:	Contiene información referente a: siniestros de tránsito, en cuanto a causas del siniestro, provincia, día, hora clase de siniestro, etc., proporcionada por la Agencia Nacional de Tránsito (ANT), referente al 2020	
URL del Conjunto de Datos:	https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT/blob/main/datos/dataset_dep.csv	
Nombre del Campo (Encabezado Columna)	Descripción del Campo	Categoría
PROVINCIA	Provincia de ocurrencia del siniestro	AZUAY BOLÍVAR CAÑAR CARCHI COTOPAXI CHIMBORAZO EL ORO ESMERALDAS GUAYAS IMBABURA LOJA LOS RÍOS MANABÍ MORONA SANTIAGO NAPO PASTAZA PICHINCHA

		TUNGURAHUA ZAMORA CHINCHIPE GALÁPAGOS SUCUMBÍOS ORELLANA SANTO DOMINGO DE LOS TSÁCHILAS SANTA ELENA
DIA	Día de ocurrencia del siniestro	LUNES MARTES MIERCOLES JUEVES VIERNES SÁBADO DOMINGO
HORA	Hora de ocurrencia del siniestro	P00=00:00:00 A 00:59:00 P01=01:00:00 A 01:59:00 P02=02:00:00 A 02:59:00 P03=03:00:00 A 03:59:00 P04=04:00:00 A 04:59:00 P05=05:00:00 A 05:59:00 P06=06:00:00 A 06:59:00 P07=07:00:00 A 07:59:00 P08=08:00:00 A 08:59:00 P09=09:00:00 A 09:59:00 P10=10:00:00 A 10:59:00 P11=11:00:00 A 11:59:00 P12=12:00:00 A 12:59:00 P13=13:00:00 A 13:59:00 P14=14:00:00 A 14:59:00 P15=15:00:00 A 15:59:00 P16=16:00:00 A 16:59:00 P17=17:00:00 A 17:59:00 P18=18:00:00 A 18:59:00 P19=19:00:00 A 19:59:00

		P20=20:00:00 A 20:59:00 P21=21:00:00 A 21:59:00 P22=22:00:00 A 22:59:00 P23=23:00:00 A 23:59:00
FERIADO	Feriado en el día de la ocurrencia del siniestro	SI NO
CAUSA PROBABLE	Código para la causa probable del siniestro	CP01=CASO FORTUITO O FUERZA MAYOR (EXPLOSIÓN DE NEUMÁTICO NUEVO, DERRUMBE, INUNDACIÓN, CAÍDA DE PUENTE, ÁRBOL, PRESENCIA INTEMPESTIVA E IMPREVISTA DE SEMOVIENTES EN LA VÍA, ETC.). CP02=PRESENCIA DE AGENTES EXTERNOS EN LA VÍA (AGUA, ACEITE, PIEDRA, LASTRE, ESCOMBROS, MADEROS, ETC.). CP03=CONDUCIR EN ESTADO DE SOMNOLENCIA O MALAS CONDICIONES FÍSICAS (SUENO, CANSANCIO Y FATIGA). CP04=DAÑOS MECÁNICOS PREVISIBLES. CP05=FALLA MECÁNICA EN LOS SISTEMAS Y/O NEUMÁTICOS (SISTEMA DE FRENIOS, DIRECCIÓN, ELECTRÓNICO O MECÁNICO). CP06=CONDUCE BAJO LA INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS. CP07=PEATÓN TRANSITA BAJO INFLUENCIA DE ALCOHOL, SUSTANCIAS ESTUPEFACIENTES O PSICOTRÓPICAS Y/O MEDICAMENTOS. CP08=PESO Y VOLUMEN - NO CUMPLIR CON LAS NORMAS DE SEGURIDAD NECESARIAS AL TRANSPORTAR CARGAS. CP09=CONDUCIR VEHÍCULO SUPERANDO LOS LIMITES MÁXIMOS DE VELOCIDAD. CP10=CONDICIONES AMBIENTALES Y/O ATMOSFÉRICAS (NIEBLA, NEBLINA, GRANIZO,

	<p>LLUVIA).</p> <p>CP011=NO MANTENER LA DISTANCIA PRUDENCIAL CON RESPECTO AL VEHÍCULO QUE LE ANTECEDE.</p> <p>CP12=NO GUARDAR LA DISTANCIA LATERAL MÍNIMA DE SEGURIDAD ENTRE VEHÍCULOS.</p> <p>CP13=CONDUCIR DESATENTO A LAS CONDICIONES DE TRÁNSITO (CELULAR, PANTALLAS DE VIDEO, COMIDA, MAQUILLAJE O CUALQUIER OTRO ELEMENTO DISTRACTOR).</p> <p>CP14=DEJAR O RECOGER PASAJEROS EN LUGARES NO PERMITIDOS.</p> <p>CP15=NO TRANSITAR POR LAS ACERAS O ZONAS DE SEGURIDAD DESTINADAS PARA EL EFECTO.</p> <p>CP16=BAJARSE O SUBIRSE DE VEHÍCULOS EN MOVIMIENTO SIN TOMAR LAS PRECAUCIONES DEBIDAS.</p> <p>CP17=CONDUCIR EN SENTIDO CONTRARIO A LA VÍA NORMAL DE CIRCULACIÓN.</p> <p>CP18=REALIZAR CAMBIO BRUSCO O INDEBIDO DE CARRIL.</p> <p>CP19=MAL ESTACIONADO - EL CONDUCTOR QUE DETENGA O ESTACIONE VEHÍCULOS EN SITIOS O ZONAS QUE ENTRAÑEN PELIGRO, TALES COMO ZONA DE SEGURIDAD, CURVAS, PUENTES, TÚNELES, PENDIENTES.</p> <p>CP20=MALAS CONDICIONES DE LA VÍA Y/O CONFIGURACIÓN. (ILUMINACIÓN Y DISEÑO).</p> <p>CP21=ADELANTAR O REBASAR A OTRO VEHÍCULO EN MOVIMIENTO EN ZONAS O SITIOS PELIGROSOS TALES COMO: CURVAS, PUENTES, TÚNELES, PENDIENTES, ETC.</p> <p>CP22=NO RESPETAR LAS SEÑALES</p>
--	--

		REGLAMENTARIAS DE TRÁNSITO. (PARE, CEDA EL PASO, LUZ ROJA DEL SEMÁFORO, ETC). CP23=NO RESPETAR LAS SEÑALES MANUALES DEL AGENTE DE TRÁNSITO. CP24=NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO A VEHÍCULOS. CP25=NO CEDER EL DERECHO DE VÍA O PREFERENCIA DE PASO AL PEATÓN. CP26=PEATÓN QUE CRUZA LA CALZADA SIN RESPETAR LA SEÑALIZACIÓN EXISTENTE (SEMÁFOROS O SEÑALES MANUALES).
CLASE FINAL	Código para la Clase del siniestro	ARROLLAMIENTOS ATROPELLOS CAÍDA DE PASAJERO CHOQUE FRONTAL CHOQUE LATERAL CHOQUE POSTERIOR COLISIÓN ESTRELLAMIENTOS PÉRDIDA DE CARRIL PÉRDIDA DE PISTA ROZAMIENTOS VOLCAMIENTOS OTROS
ZONA	Zona de ocurrencia del siniestro	URBANA RURAL
TIPO DE VEHÍCULO 1	Tipo de vehículo involucrado en el siniestro	AUTOMÓVIL MOTOCICLETA NO IDENTIFICADO CAMIONETA VEHÍCULO DEPORTIVO UTILITARIO BUS BICICLETA FURGONETA

		ESPECIAL EMERGENCIAS
SERVICIO 1	Servicio del vehículo involucrado en el siniestro	PARTICULAR COMERCIAL PÚBLICO ESTADO CUENTA PROPIA GOBIERNOS SECCIONALES
EDAD 1	Edad de la víctima	
SEXO 1	Sexo de la víctima	HOMBRE MUJER NO IDENTIFICADO
CONDICIÓN 1	Condición de la víctima	ILESO LESIONADO FALLECIDO NO IDENTIFICADO
PARTICIPANTE 1	Tipo de participante en el siniestro	CONDUCTOR PASAJERO PEATÓN

Fuente: Agencia Nacional de Tránsito – ANT

Elaborado: Yulissa Stefania Torres Quezada

Anexo 5

Análisis Exploratorio de los datos

En la Fig. 57 se observa que la mayor cantidad de siniestros ocurren en la provincia de Guayas, con un total de 6364 siniestros registros, correspondiente al 37,57%, mientras que la provincia que registra menos siniestros, es Galápagos, con un total de 0 siniestros registrados.

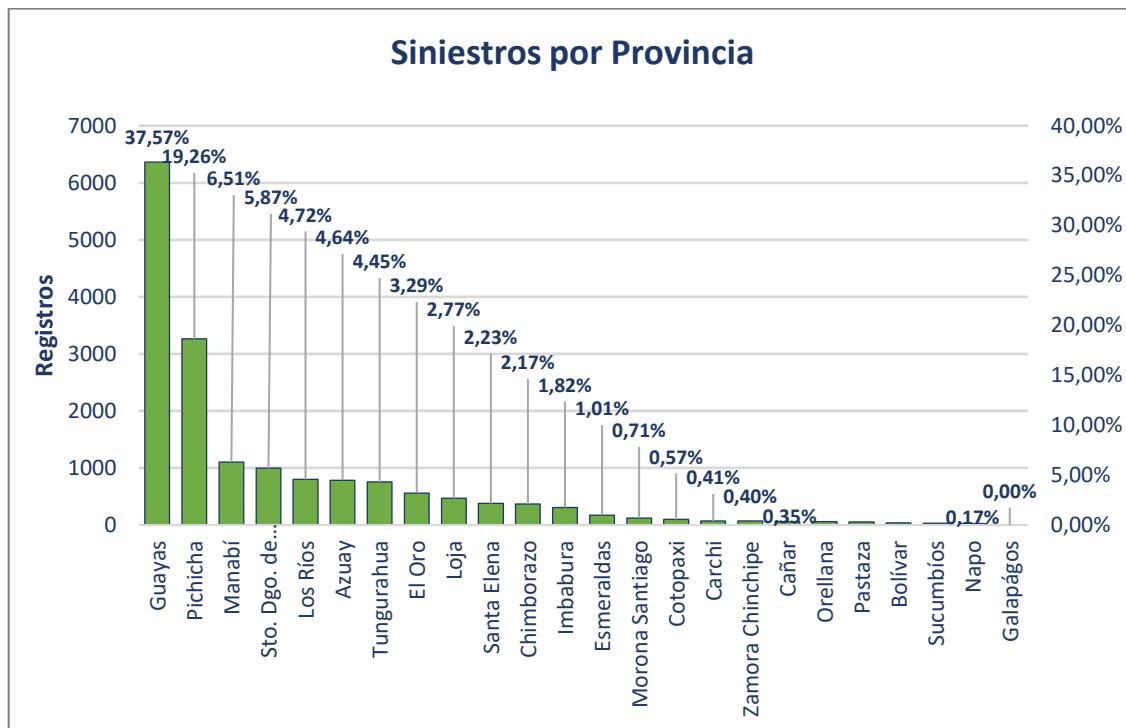


Fig. 57 Índice de Siniestralidad por Provincia

Dentro de los días con más siniestralidad se destacan los días Sábado, Domingo y Viernes, con un total de siniestros registrados de 3138, 2966 y 2437 respectivamente, que corresponden al 18,52%, 17,51% y 14,39%, tal como se puede ver en la Fig. 58. Por el contrario, el día con menor índice de siniestralidad, fue el día Martes con 1978 siniestros registrados, que corresponde al 11,68% del total.

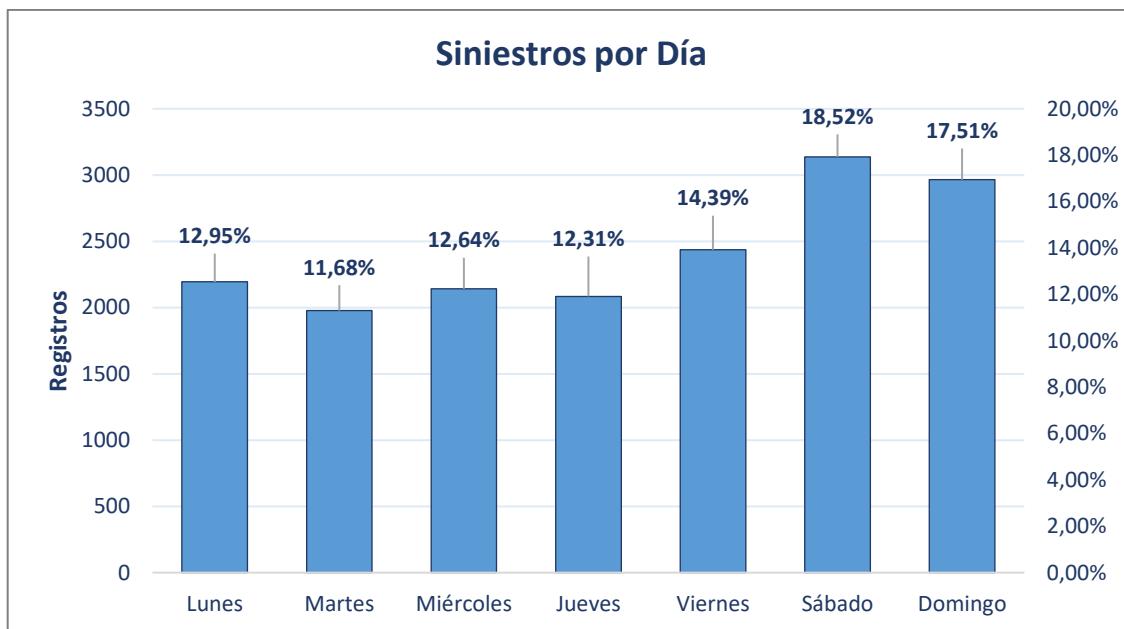


Fig. 58 Índice de Siniestralidad por Día

Las horas con mayor ocurrencia de siniestros de tránsito según la Fig. 59 es la del periodo entre las 19:00:00 horas a las 19:59:00 horas, seguido del periodo entre las 20:00:00 horas a las 20:59:00 horas, con un total de 1146 y 941 registros de siniestros respectivamente, que corresponden al 6,77% y 5,55% del total de los siniestros registrados, mientras que el periodo entre las 03:00:00 horas a las 03:59:00 horas, es el periodo con menor ocurrencia de siniestros con un total de 321 registros, correspondiente al 1,89% del total de siniestros de tránsito registrados.



Fig. 59 Siniestros por Hora

En la Fig. 60 se observa que los vehículos más involucrados en los siniestros fueron los automóviles con un 31,63%, seguidos de las motocicletas con un 21,50% del total de siniestros registrados, mientras que el vehículo menos involucrado en siniestros fue el de emergencias con un 0,06% del total de siniestros registrados.

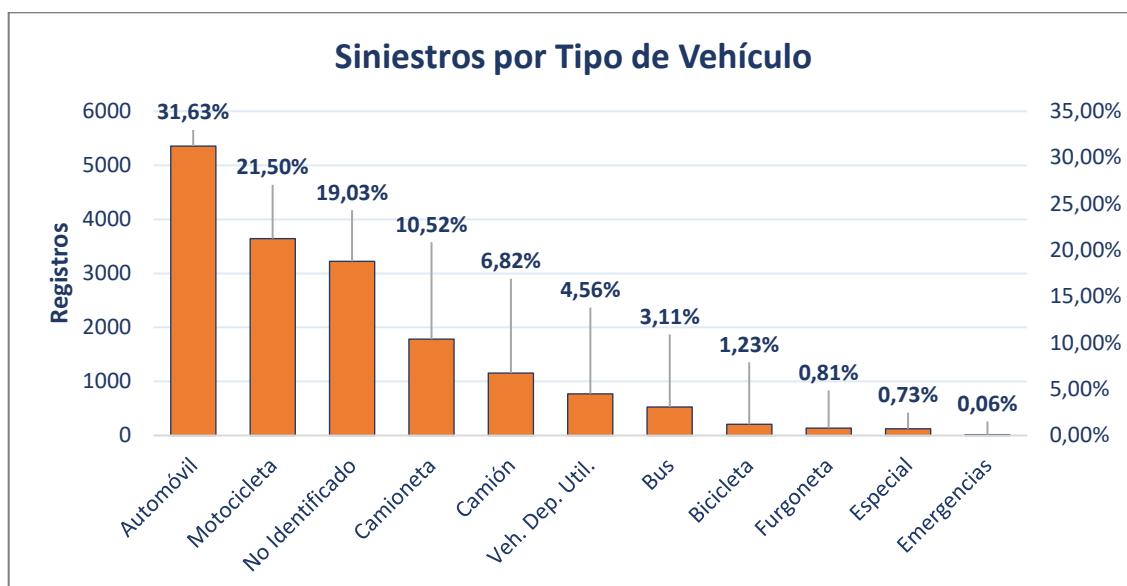


Fig. 60 Índice de Siniestralidad en Vehículos

El servicio de los vehículos que causaron más siniestros corresponde al servicio Particular con un total de 14904 siniestros registrados, correspondiente al 87,98%, tal como se puede observar en la Fig. 61, en cambio el servicio de Gobiernos Seccionales con un 0,03% del total de siniestros registrados, representa el servicio del vehículo involucrado que causa menos siniestros de tránsito.

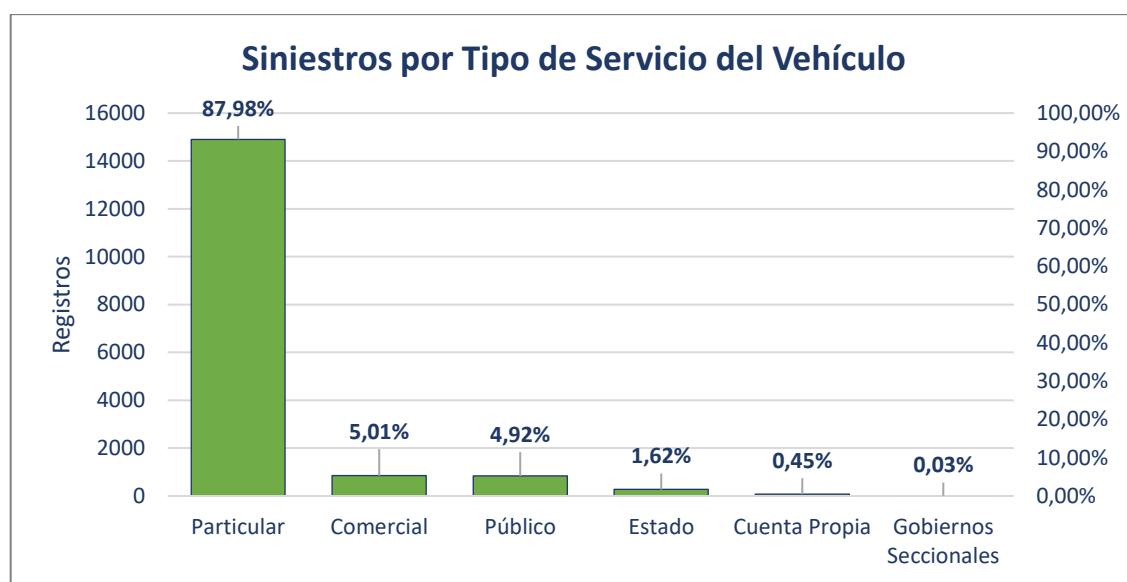


Fig. 61 Siniestros por Tipo de Servicio del Vehículo

La zona con mayor ocurrencia de siniestros de tránsito, es la Zona Urbana con 11182 siniestros registrados, correspondiente al 66,00% del total, como se puede ver en la Fig. 62, mientras que en la Zona Rural se registraron 5758 siniestros correspondientes al 34%.



Fig. 62 Siniestros por Zona

Dentro de la Fig. 63, se observa que la mayor cantidad de siniestros de tránsito son producidos por los conductores con un total de 14717 siniestros registrados, correspondiente al 86,88%, mientras que con un 5,47% los peatones son los participantes que menos siniestros producen.



Fig. 63 Índice de Siniestralidad por Participante

En la Fig. 64, se observa que la mayoría de siniestros de tránsito son ocasionados por hombres con un total del 55,59%, mientras que las mujeres ocasionan un 9,56% del total de siniestros registrados.

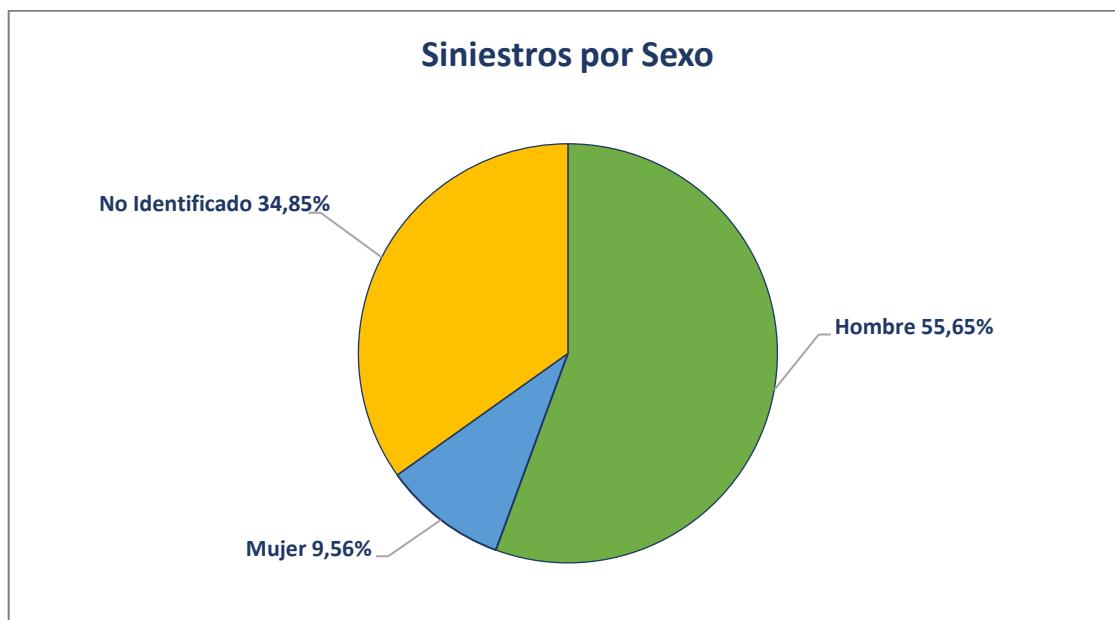


Fig. 64 Siniestros por Sexo

La Fig. 65 muestra que la principal causa según los datos de la ANT, fue la de conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractivo), con un total de 5151 siniestros registrados, correspondientes al 30,41%, en cambio la causa que menos siniestros de tránsito produce es la causa de no respetar las señales manuales del agente de tránsito, con un total del 0,01% de los siniestros registrados.

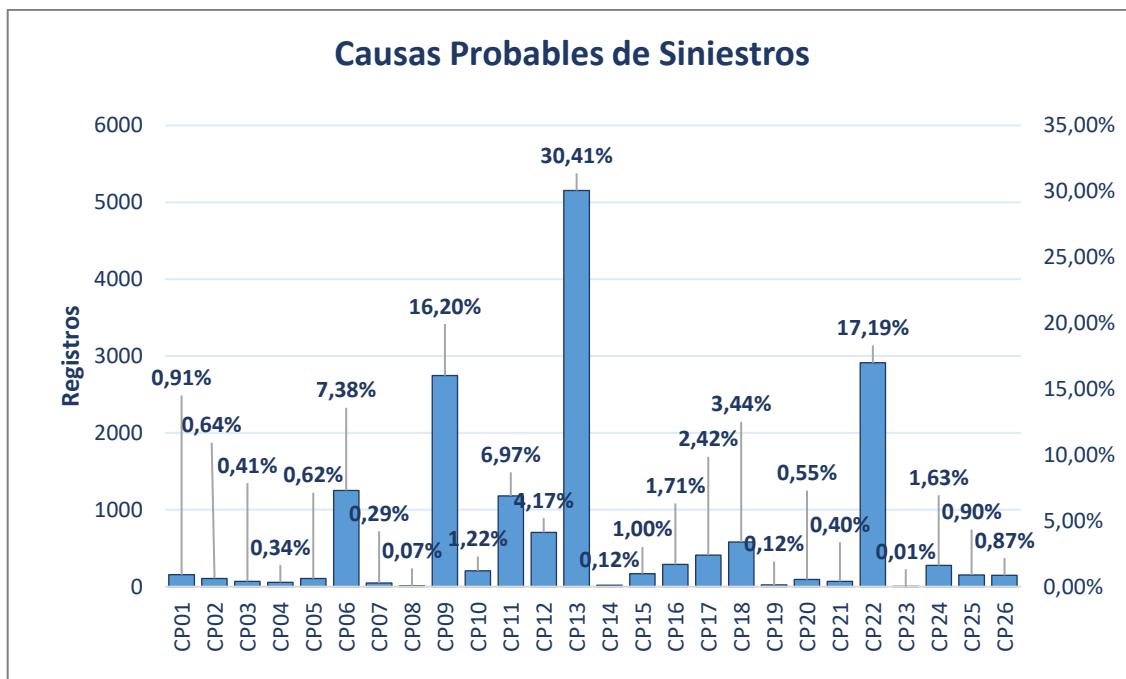


Fig. 65 Principales Causas Probables de Siniestros

La Fig. 66, describe las clases de los siniestros de tránsito, en la cual se observa que la clase de siniestro con más ocurrencia es el de Choque Lateral, con un total de 4859 registros correspondiente al 28,64%, mientras que la clase de siniestro con menor ocurrencia es el de Volcamientos, con 245 siniestros registrados correspondientes al 1,44% del total de los siniestros registrados.

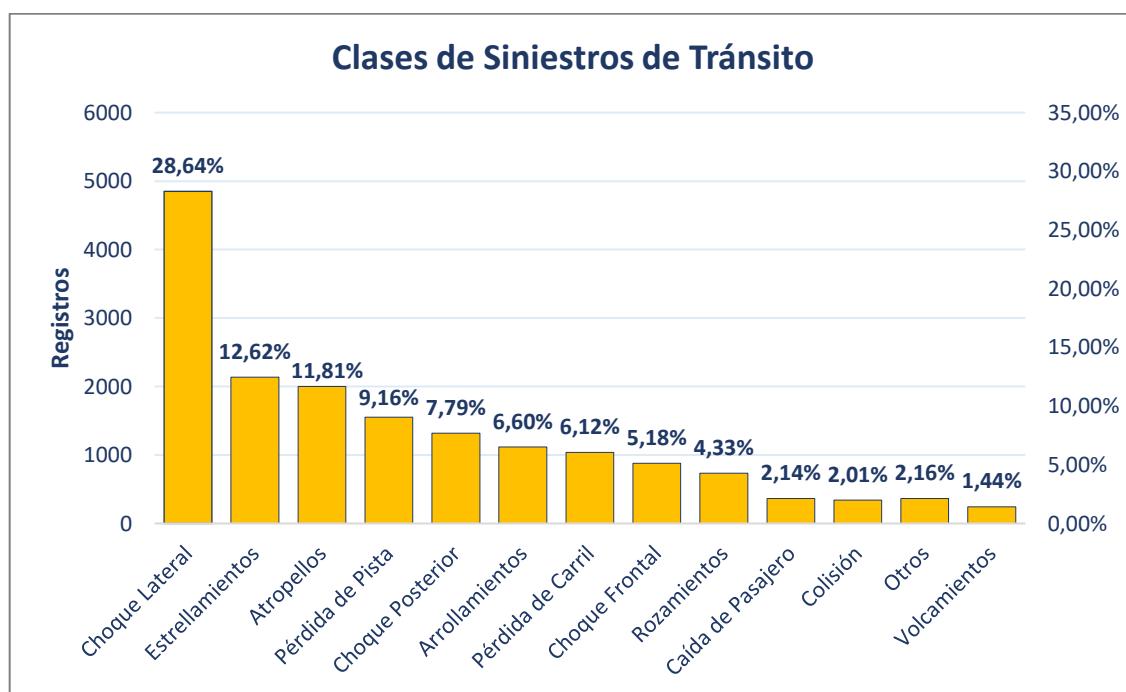


Fig. 66 Clases de Siniestros

Por último, la Fig. 67, muestra que en la mayoría de los siniestros registrados las personas involucradas resultaron ilesas con un total del 36,04%, seguido de un menor número de personas lesionadas correspondiente al 28,80% del total y una pequeña cantidad de fallecidos con un total del 4,80% de siniestros registrados.

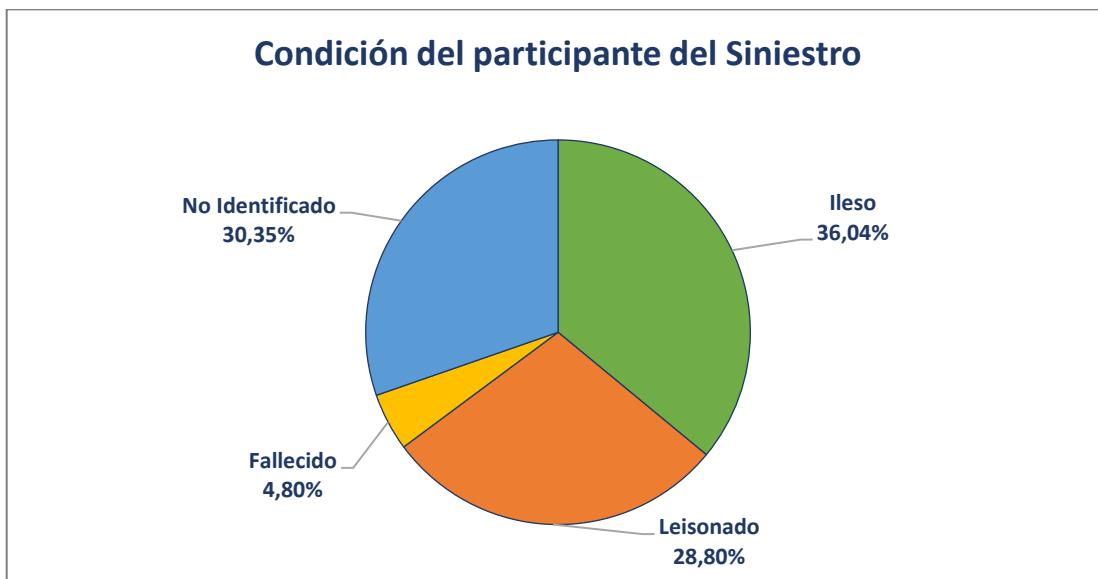


Fig. 67 Condición del Participante del Siniestro

Anexo 6

TABLA LVI
RESULTADOS DEL PROCESO DE EXPERIMENTACIÓN

Observado	CP01	CP02	CP03	CP04	CP05	CP06	CP07	CP09	CP10	CP11	CP12	CP13	CP15	CP17	CP18	CP19	CP20	CP21	CP22	CP24	Porcentaje correcto
CP01	0	0	0	0	0	2	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0,0%
CP02	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0,0%
CP03	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0,0%
CP04	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0,0%
CP05	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0,0%
CP06	0	0	0	0	0	49	0	0	0	0	0	20	0	38	0	0	0	0	0	0	45,8%
CP07	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0,0%
CP09	0	0	0	0	0	9	0	0	0	0	0	24	0	4	0	0	0	0	0	0	0,0%
CP10	0	0	0	0	0	7	0	0	0	0	0	6	0	1	0	0	0	0	0	0	0,0%
CP11	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0,0%
CP12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0,0%
CP13	0	0	0	0	0	37	0	0	0	0	0	134	0	35	0	0	0	0	0	0	65,0%
CP15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0,0%
CP17	0	0	0	0	0	20	0	0	0	0	0	25	0	317	0	0	0	0	0	0	87,6%
CP18	0	0	0	0	0	5	0	0	0	0	0	6	0	9	0	0	0	0	0	0	0,0%
CP19	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0,0%
CP20	0	0	0	0	0	1	0	0	0	0	0	2	0	8	0	0	0	0	0	0	0,0%
CP21	0	0	0	0	0	2	0	0	0	0	0	5	0	25	0	0	0	0	0	0	0,0%
CP22	0	0	0	0	0	21	0	0	0	0	0	13	0	12	0	0	0	0	0	0	0,0%
CP24	0	0	0	0	0	7	0	0	0	0	0	1	0	7	0	0	0	0	0	0	0,0%
Porcentaje global	0,0%	0,0%	0,0%	0,0%	0,0%	18,6%	0,0%	0,0%	0,0%	0,0%	0,0%	28,6%	0,0%	52,8%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	57,0%

Anexo 7

TABLA LVII
MATRIZ DE CONFUSIÓN DEL ALGORITMO CHAID

Observado	ARROLLAMIENTOS	ATROPELLOS	CAIDA DE PASAJERO	CHOQUE FRONTAL	CHOQUE LATERAL	CHOQUE POSTERIOR	COLISION	ESTRELLAMIENTOS	OTROS	PERDIDA DE CARRIL	PERDIDA DE PISTA	ROZAMIENTOS	VOLCAMIENTOS	Porcentaje correcto
ARROLLAMIENTOS	557	184	1	1	138	0	0	40	2	7	188	0	0	49,8%
ATROPELLOS	116	1214	4	9	200	3	0	107	27	26	291	4	0	60,7%
CAIDA DE PASAJERO	6	37	287	0	12	1	0	6	3	5	6	0	0	79,1%
CHOQUE FRONTAL	27	13	6	394	332	2	0	52	7	22	21	1	0	44,9%
CHOQUE LATERAL	73	52	12	28	4177	191	0	210	9	29	52	18	0	86,1%
CHOQUE POSTERIOR	16	34	0	4	327	806	0	65	6	11	45	6	0	61,1%
COLISION	13	31	1	0	87	145	0	49	0	2	13	0	0	0,0%
ESTRELLAMIENTOS	182	181	9	15	470	5	0	873	10	43	344	5	0	40,9%
OTROS	10	19	10	4	118	6	0	107	67	3	9	13	0	18,3%
PERDIDA DE CARRIL	54	72	3	8	266	1	0	209	13	127	283	1	0	12,2%
PERDIDA DE PISTA	97	249	5	6	231	8	0	73	5	40	834	3	0	53,8%
ROZAMIENTOS	6	8	1	8	172	12	0	13	1	3	8	502	0	68,4%
VOLCAMIENTOS	30	26	4	1	65	1	0	49	5	6	55	2	0	0,0%
Porcentaje global	7,0%	12,5%	2,0%	2,8%	38,9%	7,0%	0,0%	10,9%	0,9%	1,9%	12,7%	3,3%	0,0%	58,1%

TABLA LVIII
MATRIZ DE CONFUSIÓN DEL ALGORITMO CHAID EXHAUSTIVO

Observado	ARROLLAMIENTOS	ATROPELLOS	CAIDA DE PASAJERO	CHOQUE FRONTAL	CHOQUE LATERAL	CHOQUE POSTERIOR	COLISION	ESTRELLAMIENTOS	OTROS	PERDIDA DE CARRIL	PERDIDA DE PISTA	ROZAMIENTOS	VOLCAMIENTOS	Porcentaje correcto
ARROLLAMIENTOS	557	202	1	1	139	0	0	46	2	4	166	0	0	49,8%
ATROPELLOS	116	1260	4	9	182	3	0	115	27	26	255	4	0	63,0%
CAIDA DE PASAJERO	6	37	287	0	12	1	0	4	3	3	10	0	0	79,1%
CHOQUE FRONTAL	27	16	6	394	302	2	0	53	7	29	40	1	0	44,9%
CHOQUE LATERAL	73	51	12	28	4142	191	0	225	9	44	58	18	0	85,4%
CHOQUE POSTERIOR	16	43	0	4	323	806	0	61	6	12	43	6	0	61,1%
COLISION	13	35	1	0	90	145	0	47	0	4	6	0	0	0,0%
ESTRELLAMIENTOS	182	181	9	15	440	5	0	893	10	72	325	5	0	41,8%
OTROS	10	14	10	4	113	6	0	107	67	10	12	13	0	18,3%
PERDIDA DE CARRIL	54	77	3	8	240	1	0	204	13	148	288	1	0	14,3%
PERDIDA DE PISTA	97	248	5	6	219	8	0	87	5	39	834	3	0	53,8%
ROZAMIENTOS	6	7	1	8	173	12	0	13	1	2	9	502	0	68,4%
VOLCAMIENTOS	30	22	4	1	47	1	0	56	5	18	58	2	0	0,0%
Porcentaje global	7,0%	12,9%	2,0%	2,8%	37,9%	7,0%	0,0%	11,3%	0,9%	2,4%	12,4%	3,3%	0,0%	58,4%

TABLA LIX
MATRIZ DE CONFUSIÓN DEL ALGORITMO CRT

Observado	ARROLLAMIENTOS	ATROPELLOS	CAIDA DE PASAJERO	CHOQUE FRONTAL	CHOQUE LATERAL	CHOQUE POSTERIOR	COLISION	ESTRELLAMIENTOS	OTROS	PERDIDA DE CARRIL	PERDIDA DE PISTA	ROZAMIENTOS	VOLCAMIENTOS	Porcentaje correcto
ARROLLAMIENTOS	0	122	0	0	3	0	0	993	0	0	0	0	0	0,0%
ATROPELLOS	0	677	0	8	65	3	0	1244	0	0	0	4	0	33,8%
CAIDA DE PASAJERO	0	0	0	0	4	1	0	358	0	0	0	0	0	0,0%
CHOQUE FRONTAL	0	2	0	394	81	2	0	397	0	0	0	1	0	44,9%
CHOQUE LATERAL	0	6	0	28	3331	191	0	1277	0	0	0	18	0	68,7%
CHOQUE POSTERIOR	0	2	0	4	50	806	0	452	0	0	0	6	0	61,1%
COLISION	0	0	0	0	14	145	0	182	0	0	0	0	0	0,0%
ESTRELLAMIENTOS	0	22	0	14	122	5	0	1969	0	0	0	5	0	92,1%
OTROS	0	18	0	4	88	6	0	237	0	0	0	13	0	0,0%
PERDIDA DE CARRIL	0	6	0	8	40	1	0	981	0	0	0	1	0	0,0%
PERDIDA DE PISTA	0	7	0	6	34	8	0	1493	0	0	0	3	0	0,0%
ROZAMIENTOS	0	2	0	8	85	12	0	125	0	0	0	502	0	68,4%
VOLCAMIENTOS	0	1	0	1	7	1	0	232	0	0	0	2	0	0,0%
Porcentaje global	0,0%	5,1%	0,0%	2,8%	23,2%	7,0%	0,0%	58,7%	0,0%	0,0%	0,0%	3,3%	0,0%	45,3%

TABLA LX
MATRIZ DE CONFUSIÓN DEL ALGORITMO PERCEPTRÓN MULTICAPA

Observado	ARROLLAMIENTOS	ATROPELLOS	CAIDA DE PASAJERO	CHOQUE FRONTAL	CHOQUE LATERAL	CHOQUE POSTERIOR	COLISION	ESTRELLAMIENTOS	OTROS	PERDIDA DE CARRIL	PERDIDA DE PISTA	ROZAMIENTOS	VOLCAMIENTOS	Porcentaje correcto
ARROLLAMIENTOS	114	40	1	0	21	1	0	18	0	3	32	0	0	49,6%
ATROPELLOS	37	247	1	2	20	0	0	12	6	2	54	0	0	64,8%
CAIDA DE PASAJERO	1	3	54	3	4	0	0	2	1	2	2	0	0	75,0%
CHOQUE FRONTAL	8	4	0	112	45	2	0	8	2	7	13	2	0	55,2%
CHOQUE LATERAL	21	11	1	17	806	37	0	55	1	6	22	19	0	80,9%
CHOQUE POSTERIOR	3	7	0	5	49	149	0	22	1	3	11	5	0	58,4%
COLISION	4	6	0	0	9	31	0	9	0	0	7	0	0	0,0%
ESTRELLAMIENTOS	41	18	0	10	70	2	0	207	2	9	85	1	0	46,5%
OTROS	2	7	2	2	28	1	0	18	11	1	4	3	0	13,9%
PERDIDA DE CARRIL	12	18	1	6	28	1	0	56	0	23	80	1	0	10,2%
PERDIDA DE PISTA	31	50	0	9	25	3	0	40	0	6	155	1	0	48,4%
ROZAMIENTOS	2	4	0	3	26	2	0	2	0	0	3	116	0	73,4%
VOLCAMIENTOS	7	2	1	3	5	1	0	9	1	6	12	0	0	0,0%
Porcentaje global	8,1%	12,0%	1,8%	4,9%	32,7%	6,6%	0,0%	13,2%	0,7%	2,0%	13,8%	4,3%	0,0%	57,3%

TABLA LXI
MATRIZ DE CONFUSIÓN DEL ALGORITMO FUNCIÓN DE BASE RADIAL

Observado	ARROLLAMIENTOS	ATROPELLOS	CAIDA DE PASAJERO	CHOQUE FRONTAL	CHOQUE LATERAL	CHOQUE POSTERIOR	COLISION	ESTRELLAMIENTOS	OTROS	PERDIDA DE CARRIL	PERDIDA DE PISTA	ROZAMIENTOS	VOLCAMIENTOS	Porcentaje correcto
ARROLLAMIENTOS	0	48	0	0	110	5	0	24	0	0	43	0	0	0,0%
ATROPELLOS	0	259	0	0	60	12	0	10	0	0	42	0	0	67,6%
CAIDA DE PASAJERO	0	13	0	0	22	37	0	0	0	0	3	0	0	0,0%
CHOQUE FRONTAL	0	10	0	0	99	25	0	5	0	0	22	0	0	0,0%
CHOQUE LATERAL	0	20	0	0	862	39	0	25	0	0	38	0	0	87,6%
CHOQUE POSTERIOR	0	7	0	0	109	110	0	16	0	0	24	0	0	41,4%
COLISION	0	2	0	0	33	19	0	4	0	0	5	0	0	0,0%
ESTRELLAMIENTOS	0	57	0	0	177	9	0	104	0	0	80	0	0	24,4%
OTROS	0	10	0	0	34	11	0	19	0	0	4	0	0	0,0%
PERDIDA DE CARRIL	0	38	0	0	104	3	0	23	0	0	62	0	0	0,0%
PERDIDA DE PISTA	0	91	0	0	72	3	0	3	0	0	140	0	0	45,3%
ROZAMIENTOS	0	4	0	0	108	39	0	0	0	0	6	0	0	0,0%
VOLCAMIENTOS	0	4	0	0	16	1	0	1	0	0	15	0	0	0,0%
Porcentaje global	0,0%	16,6%	0,0%	0,0%	53,1%	9,2%	0,0%	6,9%	0,0%	0,0%	14,2%	0,0%	0,0%	43,4%

TABLA LXII
MATRIZ DE CONFUSIÓN DEL ALGORITMO NAIVE BAYES

Observado	ATROPELOS	CHOQUE LATERAL	CHOQUE POSTERIOR	CHOQUE FRONTAL	PERDIDA DE CARRIL	ROZAMIENTOS	ESTRELLAMIENTOS	CAIDA DE PASAJERO	ARROLLAMIENTOS	VOLCAMIENTOS	COLISION	PERDIDA DE PISTA	OTROS	Porcentaje correcto
ATROPELOS	1208	177	4	9	19	3	108	16	62	4	22	350	19	63,0%
CHOQUE LATERAL	60	3776	176	58	103	70	310	9	115	22	31	107	14	68,1%
CHOQUE POSTERIOR	28	221	746	20	29	9	118	3	22	14	60	47	3	68,7%
CHOQUE FRONTAL	19	181	3	423	48	1	78	3	43	20	5	49	4	70,8%
PERDIDA DE CARRIL	62	142	2	15	253	2	238	2	60	17	5	227	12	32,2%
ROZAMIENTOS	12	121	14	10	5	511	23	0	9	0	3	21	5	81,7%
ESTRELLAMIENTOS	95	336	7	22	112	6	1020	7	104	21	27	351	29	41,5%
CAIDA DE PASAJERO	15	17	1	0	6	0	0	289	8	4	7	7	9	82,1%
ARROLLAMIENTOS	199	193	0	3	32	0	143	5	301	9	1	228	4	34,2%
VOLCAMIENTOS	14	23	2	7	31	3	48	1	24	28	2	59	2	17,1%
COLISION	8	68	117	0	4	2	70	0	11	2	41	17	1	19,4%
PERDIDA DE PISTA	153	184	9	26	124	4	208	8	109	21	2	703	0	32,2%
OTROS	40	95	4	4	17	14	85	9	10	1	4	19	64	38,6%

TABLA LXIII
MATRIZ DE CONFUSIÓN DEL ALGORITMO BAYESNET

Observado	ATROPELOS	CHOQUE LATERAL	CHOQUE POSTERIOR	CHOQUE FRONTAL	PERDIDA DE CARRIL	ROZAMIENTOS	ESTRELLAMIENTOS	CAIDA DE PASAJERO	ARROLLAMIENTOS	VOLCAMIENTOS	COLISION	PERDIDA DE PISTA	OTROS	Porcentaje correcto
ATROPELOS	1236	175	4	7	25	4	106	17	63	4	22	318	20	61,9%
CHOQUE LATERAL	68	3792	175	56	112	61	302	10	107	24	27	103	14	67,8%
CHOQUE POSTERIOR	28	235	742	19	30	8	112	2	18	14	62	46	4	68,4%
CHOQUE FRONTAL	22	186	3	423	50	1	79	3	37	20	3	46	4	70,6%
PERDIDA DE CARRIL	62	149	2	10	263	2	239	1	59	18	5	217	10	30,9%
ROZAMIENTOS	12	124	3	13	5	513	22	0	9	1	2	15	5	82,4%
ESTRELLAMIENTOS	107	334	8	27	123	6	1009	8	94	22	23	348	28	41,5%
CAIDA DE PASAJERO	16	16	1	0	7	0	1	291	9	4	6	5	7	82,2%
ARROLLAMIENTOS	210	198	0	6	31	0	136	5	292	9	1	227	3	34,3%
VOLCAMIENTOS	14	21	2	6	33	2	47	1	25	30	2	59	2	17,2%
COLISION	10	68	117	0	3	2	73	0	10	2	39	16	1	19,2%
PERDIDA DE PISTA	165	187	9	26	141	4	209	7	93	21	3	686	0	32,3%
OTROS	38	92	4	4	19	14	81	9	10	1	5	17	72	41,3%

Anexo 8

TABLA LXIV
CATEGORIZACIÓN DE LAS CAUSAS PROBABLES DE LOS SINIESTROS DE
TRÁNSITO

Causa Probable	Descripción	Categoría
CP03	Conducir en estado de somnolencia o malas condiciones físicas (sueño, cansancio y fatiga).	Factor Humano
CP06	Conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos.	
CP08	Peso y volumen - no cumplir con las normas de seguridad necesarias al transportar cargas.	
CP09	Conducir vehículo superando los límites máximos de velocidad.	
CP11	No mantener la distancia prudencial con respecto al vehículo que le antecede.	
CP12	No guardar la distancia lateral mínima de seguridad entre vehículos.	
CP13	Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).	
CP14	Dejar o recoger pasajeros en lugares no permitidos.	
CP17	Conducir en sentido contrario a la vía normal de circulación.	
CP18	Realizar cambio brusco o indebido de carril.	
CP19	Mal estacionado - el conductor que detenga o estacione vehículos en sitios o zonas que entrañen peligro, tales como zona de seguridad, curvas, puentes, túneles, pendientes.	
CP21	Adelantar o rebasar a otro vehículo en movimiento en zonas o sitios peligrosos tales como: curvas, puentes, túneles, pendientes, etc.	

CP22	No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).	
CP23	No respetar las señales manuales del agente de tránsito.	
CP24	No ceder el derecho de vía o preferencia de paso a vehículos.	
CP25	No ceder el derecho de vía o preferencia de paso al peatón.	
CP04	Daños mecánicos previsibles.	Factor Vehículo
CP05	Falla mecánica en los sistemas y/o neumáticos (sistema de frenos, dirección, electrónico o mecánico).	
CP01	Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.).	Factor Entorno
CP02	Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).	
CP07	Peatón transita bajo influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos.	
CP10	Condiciones ambientales y/o atmosféricas (niebla, neblina, granizo, lluvia).	
CP15	No transitar por las aceras o zonas de seguridad destinadas para el efecto.	
CP16	Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.	
CP20	Malas condiciones de la vía y/o configuración. (iluminación y diseño).	
CP26	Peatón que cruza la calzada sin respetar la señalización existente (semáforos o señales manuales).	

Anexo 9

Informe dirigido al Abg. Cnel. Paul Aguilar Sotomayor

Loja, 17 de septiembre de 2021

Abg. Cnel. Paul Mauricio Aguilar Sotomayor

Director Estratégico Municipal del Cuerpo de Agentes Civiles de Tránsito de la Unidad de Control Operativa de Tránsito (UCOT) del GAD Municipal de Loja

De mis consideraciones.

Reciba un cordial saludo y a la vez deseándole toda clase de éxitos en las funciones a su cargo.

La presente tiene la finalidad de poner a su conocimiento y a la institución a la cual usted representa, los resultados de un estudio¹ que realicé en la Carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja, titulado "Minería de datos para determinar los factores más influyentes en la ocurrencia de Siniestros de Tránsito en Ecuador en el año 2020" en el que se analizó 16940 registros de siniestros de tránsito ocurridos en nuestro país, en donde se determinó que con un porcentaje de probabilidad de ocurrencia del 69,65% el factor mas influyente es el factor humano. Por esta razón pongo a vuestra disposición estos resultados para que los consideren como una fuente de información que lleve a implementar acciones y estrategias en beneficio de la comunidad.

Esperando que el presente informe sea de utilidad para la institución, le expreso mis sentimientos de gratitud.

Atentamente,



Yulissa Stefania Torres Quezada
Estudiante de la Universidad Nacional de Loja
C.I 1106035536
Email: yulissa.torres@unl.edu.ec

¹ https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT

RECIBIDO 20/SEP/2021



Informe dirigido al Econ. Santos Aurelio Araujo

Macará, 17 de septiembre de 2021

Econ. Santos Aurelio Araujo

Director de la Unidad de Tránsito del GAD Municipal de Macará

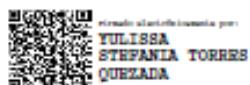
De mis consideraciones.

Reciba un cordial saludo y a la vez deseándole toda clase de éxitos en las funciones a su cargo.

La presente tiene la finalidad de poner a su conocimiento y a la institución a la cual usted representa, los resultados de un estudio¹ que realicé en la Carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja, titulado "Minería de datos para determinar los factores más influyentes en la ocurrencia de Siniestros de Tránsito en Ecuador en el año 2020" en el que se analizó 16940 registros de siniestros de tránsito ocurridos en nuestro país, en donde se determinó que con un porcentaje de probabilidad de ocurrencia del 69,65% el factor más influyente es el factor humano. Por esta razón pongo a vuestra disposición estos resultados para que los consideren como una fuente de información que lleve a implementar acciones y estrategias en beneficio de la comunidad.

Esperando que el presente informe sea de utilidad para la institución, le expreso mis sentimientos de gratitud.

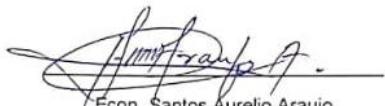
Atentamente,



Yulissa Stefania Torres Quezada
Estudiante de la Universidad Nacional de Loja
C.I 1106035536
Email: yulissa.torres@unl.edu.ec

¹ https://github.com/yulissatq/Factores_Influyentes_Siniestros_Transito_TT

RECIBIDO 17/SEP/2021



Econ. Santos Aurelio Araujo
Director de la Unidad de Tránsito del
GAD Municipal de Macará