e-ISSN: 1390-5902
CEDAMAZ, Vol. 11, No. 02, pp. 124–132, julio–diciembre 2021
DOI: 10.54753/cedamaz.v11i2.1181

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

Data mining to determine the most influential factors in the occurrence of traffic accidents in Ecuador in the year 2020

Yulissa Torres-Quezada[1,*]

[1] *Carrera de Ingeniería en Sistema/Computación, Universidad Nacional de Loja, Loja, Ecuador*
[*] *Autor para correspondencia: yulissa.torres@unl.edu.ec*

Fecha de recepción del manuscrito: 10/11/2021 Fecha de aceptación del manuscrito: 27/11/2021 Fecha de publicación: 24/12/2021

Resumen—Actualmente, 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 fundamental e importante plantear un estudio que permita determinar cuáles son los factores que ocasionan la ocurrencia de los siniestros de tránsito. En este trabajo de investigación se aplica 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 empleando cinco fases de la metodología Knowledge Discovery in Databases (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, estas, utilizadas para el descubrimiento de patrones ocultos en el conjunto de datos, el cual fue recolectado por la Agencia Nacional de Tránsito (ANT) y tiene un total de 418 variables y 16972 registros de eventos registrados sobre siniestros de tránsito en Ecuador. Se aplicaron siete técnicas de minería de datos, tales como: CHAID, CHAID Exhaustivo, CRT, Perceptrón Multicapa, Función de Base Radial, Naive Bayes y BayesNet. El algoritmo CHAID Exhaustivo fue el que obtuvo los mejores resultados 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.

Abstract—Currently, 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 essential and important to propose a study to determine what are the factors that cause the occurrence of traffic accidents. In this research work, data mining is applied to determine the most influential factors in the occurrence of traffic accidents in Ecuador in the year 2020, this was carried out using five phases of the Knowledge Discovery in Databases (KDD) methodology consisting of: information search, data collection, database cleansing, application of data mining techniques and interpretation and presentation of results, these, used for the discovery of hidden patterns in the dataset, which was collected by the National Traffic Agency (ANT) and has a total of 418 variables and 16972 records of events recorded on traffic crashes in Ecuador. Seven data mining techniques were applied, such as: CHAID, Exhaustive CHAID, CRT, Multilayer Perceptron, Radial Basis Function, Naive Bayes and BayesNet. The Exhaustive CHAID algorithm was the one that obtained the best results with which the most important patterns in the data were identified and the possible associations between the collected variables 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.

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 (Organización Mundial de la Salud, 2018). Los siniestros de tránsito 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 de tránsito por cada mil vehículos que circulan en el país, esto lo ubica en segundo país en Latinoamérica con mayor índice de ocurrencia de siniestros de tránsito (INEC, 2020).

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 (Rodriguez Hassiger, 2014).

El objetivo de este trabajo fue 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. Este análisis de datos fue realizado únicamente para siniestros de tránsito ocurridos en Ecuador en el año 2020. El presente estudio está conformado por varias secciones, entre ellas la sección de Materiales y Métodos en donde se definen las fases de la metodología KDD. Posteriormente, se encuentra la sección de Resultados estructurada a partir de los hallazgos basados en cada fase de la metodología utilizada. En la sección de Discusión se analiza, explica y contrasta los resultados que se obtuvieron y finalmente en la sección de Conclusiones se plasman las deducciones alcanzadas a partir de las experiencias obtenidas durante el proceso de cumplimiento de la metodología KDD.

MATERIALES Y MÉTODOS

Para realizar el presente trabajo de minería de datos 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 presentación e interpretación de los resultados. En la Figura 1 se presenta el propósito de cada fase antes mencionada. Se seleccionaron las herramientas OpenRefine y RStudio para llevar a cabo la fase de depuración de la base de datos mientras que las herramientas SPSS Statistics y Weka fueron utilizadas para cumplir con las fases de aplicación de técnicas de minería de datos e interpretación y presentación de los datos.

En la fase de búsqueda de información 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

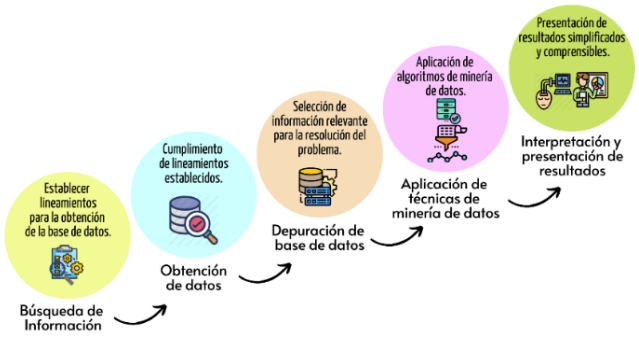


Fig. 1: Metodología utilizada en la investigación

tránsito. Al final se establecieron lineamientos que facilitaron la obtención de bases de datos confiables. Para el desarrollo del presente trabajo, fue necesaria 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 (ANT, 2020b).

Dentro de la fase de depuración de la base de datos, 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.

Más adelante, en la fase correspondiente a la minería de datos, se aplicaron siete algoritmos de clasificación al conjunto de datos depurado, los algoritmos seleccionados de acuerdo a la literatura encontrada fueron los CHAID (Chi-squared Automatic Interaction Detection), CHAID Exhaustivo, CRT (Claiming Rule Team), Perceptrón Multicapa, Función de Base Radial, Naive Bayes y 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.

Finalmente, en la fase de interpretación y presentación de resultados 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 porcen-

e-ISSN: 1390-5902
CEDAMAZ, Vol. 11, No. 02, pp. 124–132, julio–diciembre 2021
DOI: 10.54753/cedamaz.v11i2.1181

tajes presentados y en base a este se generaron reportes con las reglas de asociación de patrones y gráficos 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.

RESULTADOS

Fase I: Búsqueda de información

En esta fase se realizó una investigación acerca de instituciones que proporcionen bases de datos, las cuales contengan registros de sucesos de siniestros de tránsito ocurridos en Ecuador, una vez identificado dicho repositorio se procedió a analizar trabajos relacionados con el presente, como (ANT, 2020a), (Gomes Barcellos, 2020), (Pumares, 2019) y (Ospina-Mateus Quintana Jiménez, 2019) , con la finalidad de obtener referencias para establecer los lineamientos para la selección del conjunto de datos más factible. Posterior a haber realizado dicha búsqueda fueron establecidos los lineamientos que se proponen en la Tabla 1.

Tabla 1: 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

Ya establecidos los criterios para la selección de las bases de datos mostrados en la Tabla 1, se inició el proceso para la obtención de las mismas. Se trabajó con el conjunto de datos alojado en la página oficial de la (ANT, 2020b), en el cual 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, 2020b) y 16972 registros de eventos ocurridos sobre siniestros de tránsito en Ecuador. Considerando lo establecido en la Tabla 1, 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 fases.

Fase III: Depuración de la base de datos

En esta fase, se procedió a la evaluación de la misma, de esta manera se determinó que variables del conjunto de datos son las más relevantes y útiles para el proceso de aplicación de las técnicas de minería de datos, una vez evaluado el conjunto de datos se identificó trece variables relevantes, estas variables seleccionadas con referencia a las investigaciones de (Pumares, 2019), (Ospina-Mateus Quintana Jiménez, 2019) y (Gomes Barcellos, 2020) que tratan de la utilización de minería de datos para el análisis de los siniestros de tránsito.

Luego se realizó la limpieza a la base de datos, la herramienta usada para la eliminación de la información no útil fue OpenRefine , a través del uso de la función Transformaciones Comunes y Transformar, se realizó la estandarización del conjunto de datos, aplicando comandos para renombrar las variables, al igual para convertir los registros de las variables a mayúsculas y reemplazar un valor por otro, de esta manera se eliminó las tildes y se reemplazó la letra “Ñ”, por la letra “N”.

Una vez estandarizado el conjunto de datos, a través del software RStudio se procedió a eliminar información inconsistente presente 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 la ocurrencia de los siniestros de tránsito, la eliminación de esta información se llevó a cabo en base al Art. 125 del Reglamento a Ley De Transporte Terrestre Tránsito y Seguridad Vial, con esto se eliminaron registros controlando que estos almacenen datos de los participantes que sean conductores de los tipos de vehículos, ya sean automóviles y camionetas, que contaban con un servicio particular y de cuenta propia, que tengan una edad menor a 16 años y además controlando que para los demás tipos de vehículos involucrados y su tipo de servicio, diferentes a los antes mencionados, se eliminen los que cuenten con una edad menor a 18 años, luego de la aplicación de este control el número de registros se redujo de 16972 a 16940.

Fase IV: Aplicación de técnicas de minería de datos

Para la aplicación de los algoritmos de minería de datos fueron utilizadas las herramientas SPSS Statistics y Weka, los algoritmos seleccionados de acuerdo a literatura encontrada en los cuales según (Rodriguez Hassiger, 2014) y (López Maldonado, 2013), 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.

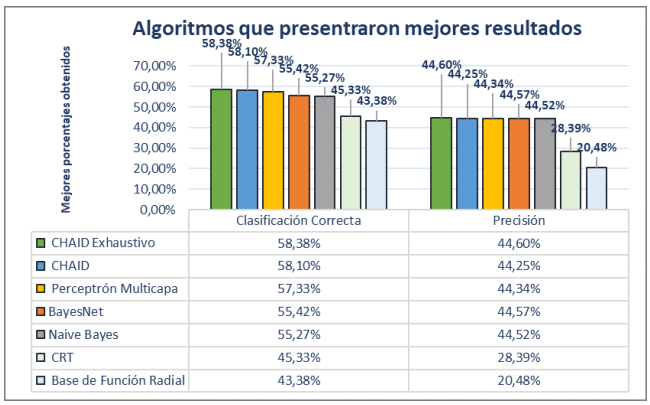
Es por lo antes mencionado que se aplicaron tres tipos de algoritmos de árboles de decisión: CHAID, CHAID Exhaustivo y CRT; dos tipos de algoritmos de redes neuronales: Perceptrón multicapa y de Función de Base Radial; y por último los algoritmos de Redes Bayesianas: Naive Bayes y BayesNet, al aplicar dichos algoritmos se configuró la variable “CLASE_FINAL” como variable objetivo esto en relación al objeto de estudio del presente trabajo. Mediante esta aplicación se realizó la clasificación del conjunto de datos sobre siniestros de tránsito registrados en Ecuador en el año 2020.

Fase V: Interpretación y presentación de resultados

En esta fase primero se realizó el análisis de los resultados obtenidos después de la aplicación 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 estas métricas fueron el porcentaje global de instancias clasificadas correctamente y el porcentaje de precisión global especificado para cada categoría de la variable objetivo.

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 Figura 2.



	Clasificación Correcta	Precisión
CHAID Exhaustivo	58,38%	44,60%
CHAID	58,10%	44,25%
Perceptrón Multicapa	57,33%	44,34%
BayesNet	55,42%	44,57%
Naive Bayes	55,27%	44,52%
CRT	45,33%	28,39%
Base de Función Radial	43,38%	20,48%

Fig. 2: Mejores resultados de clasificación correcta y precisión de acuerdo a cada algoritmo

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. El árbol generado cuenta con una profundidad máxima de 5 nodos, con 216 nodos de los cuales 148 son nodos terminales. Se visualiza parte del árbol obtenido en la Figura 3.

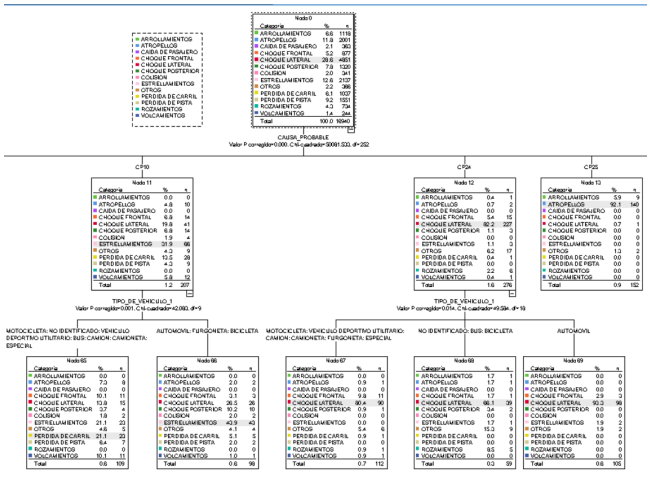


Fig. 3: Árbol obtenido con el algoritmo CHAID Exhaustivo.

A través del árbol de decisión generado del algoritmo CHAID Exhaustivo 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, mostrados a continuación:

La Figura 4 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.

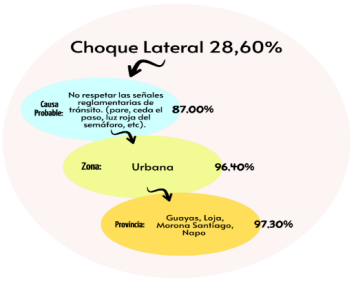


Fig. 4: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral

Tal como se muestra en la Figura 5, 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. 5: 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 Figura 6.

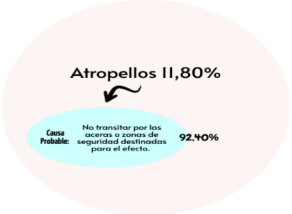


Fig. 6: Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Estrellamientos

Como se muestra en la Figura 7, 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 %.

En la Figura 8 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

e-ISSN: 1390-5902
CEDAMAZ, Vol. 11, No. 02, pp. 124–132, julio–diciembre 2021
DOI: 10.54753/cedamaz.v11i2.1181

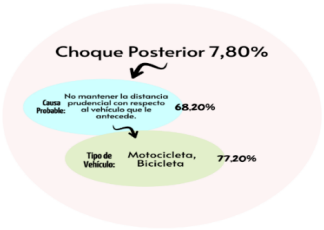


Fig. 7: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Posterior

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.

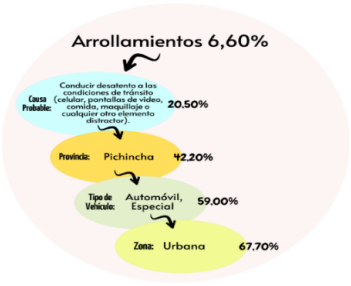


Fig. 8: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Arrollamientos

Para que ocurran siniestros de tránsito de clase pérdida de carril, de acuerdo a la Figura 9, 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 %.

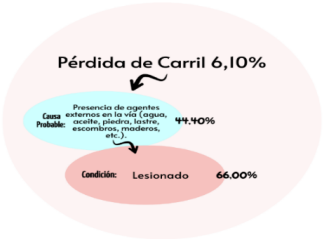


Fig. 9: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Lateral

En la mayoría de siniestros de tránsito de clase choque frontal como se muestra en la Figura 10, 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.

De acuerdo a la Figura 11, 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

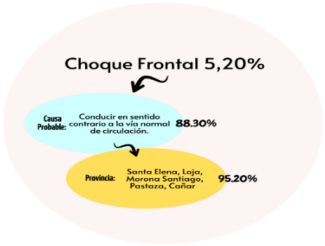


Fig. 10: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Choque Frontal

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 periodos de tiempo de las horas mostradas en la Tabla 2 con una probabilidad del 100 % de ocurrencia.

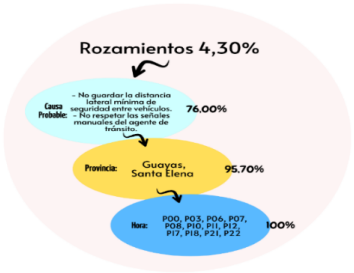


Fig. 11: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Rozamientos

Tabla 2: 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 Figura 12 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 %.

De acuerdo a lo expuesto en la Figura 13, 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 ocurran en los periodos de

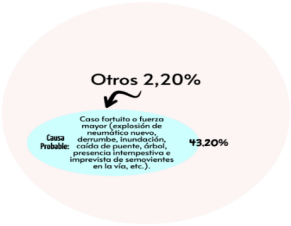


Fig. 12: Principal variable involucrada en la ocurrencia del siniestro de tránsito de clase Otros

tiempo de las horas mostradas en la Tabla 3 con una probabilidad del 100 % de ocurrencia.

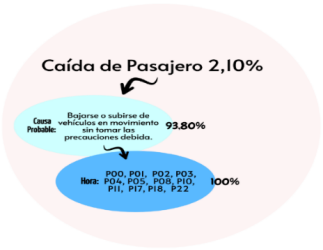


Fig. 13: Principales variables involucradas en la ocurrencia del siniestro de tránsito de clase Caída de Pasajero

Tabla 3: 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

Según los gráficos presentados 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 categorizada para el año 2020 las causas probables de los siniestros de tránsito tal como se muestra en la Tabla 4.

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 distractor), 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 Figura 14.

Tabla 4: Causas Probables de siniestros de tránsito en el Ecuador

N°	Cód.	Causa probable	Prob. de Ocu.
1	CP01	Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación,caída de puente, árbol,presencia intempestiva e imprevista de semovientesen la vía, etc.).	43,20 %
2	CP02	Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.).	44,40 %
3	CP05	Falla mecánica en los sistemas y/o neumáticos(sistema de frenos, dirección, electrónico o mecánico).	51,40 %
4	CP11	No mantener la distancia prudencial con respecto al vehículo que le antecede.	69,20 %
5	CP12	No guardar la distancia lateral mínima de seguridad entre vehículos.	76,00 %
6	CP13	Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).	20,50 %
7	CP15	No transitar por las aceras o zonas de seguridad destinadas para el efecto.	92,40 %
8	CP16	Bajarse o subirse de vehículos en movimiento sin tomar las precauciones debida.	93,80 %
9	CP17	Conducir en sentido contrario a la vía normal de circulación.	95,20 %
10	CP22	No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc).	87,00 %
11	CP23	No respetar las señales manuales del agente de tránsito.	76,00 %

Para cumplir con el objeto de estudio las causas probables mostradas en la Figura 14, se las categorizó en tres factores que son: Factor Humano, Factor Vehículo y Factor Entorno, tomando como referencia los trabajos realizados por (Constante Tipán, 2017), (Calle Reinoso Sarabia Paucay, 2020) y (Román Matamoros, 2015); 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 como se muestra en la Figura 15.

Como muestra la Figura 15, al categorizar las once causas

e-ISSN: 1390-5902
CEDAMAZ, Vol. 11, No. 02, pp. 124–132, julio–diciembre 2021
DOI: 10.54753/cedamaz.v11i2.1181

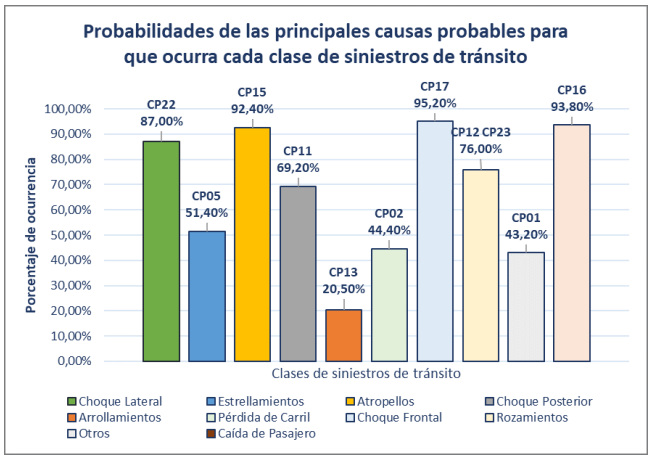


Fig. 14: Resultados de las probabilidades de ocurrencia de las causas probables cada clase de siniestro de tránsito

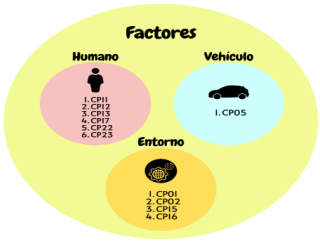


Fig. 15: Categorización en factores a las causas probables de los siniestros de tránsito

probables mostradas en la Figura 14, 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.

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 Figura 16.

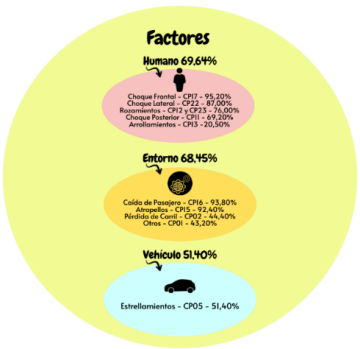


Fig. 16: Resultados de las probabilidades de ocurrencia en relación a los tipos de factores

La Figura 16 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.

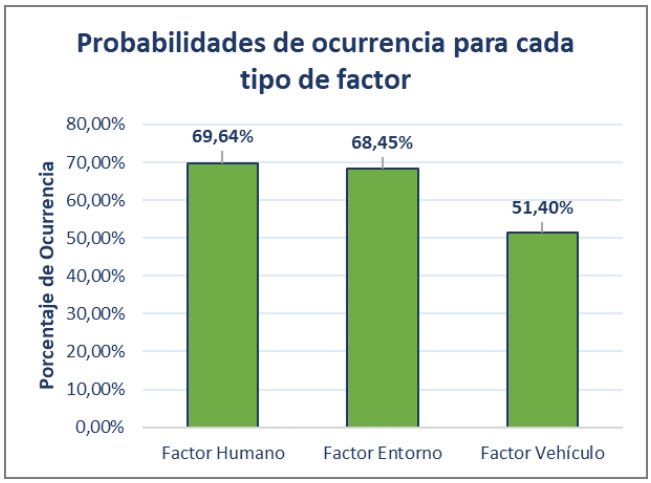


Fig. 17: Factores influyentes para la ocurrencia de siniestros de tránsito en Ecuador

En la Figura 17 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.

Finalmente, en el siguiente hipervínculo se comparte el conjunto de datos referente a los siniestros de tránsito ocurridos en Ecuador en el año 2020 y el tratamiento de los mismos con respecto a la ejecución de la metodología KDD para el cumplimiento del objetivo del presente trabajo.

DISCUSIÓN

Para el cumplimiento de la primera fase de la metodología KDD, se generó un protocolo de búsqueda para identificar las bases de datos, esto a través del establecimiento de lineamientos relacionados para obtener información relevante sobre siniestros de tránsito en Ecuador en el año 2020. Este protocolo permitió una búsqueda sistemática de la información, priorizando las bases de datos a utilizar en cada momento, dando como resultado la obtención de una base de datos consolidada con información confiable, precisa y actualizada que permitió el cumplimiento del objeto de estudio. Por el contrario, el estudio realizado por (Pumares, 2019) ejecutó 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 adecue a la estructura que tiene el data warehouse para su almacenamiento, este tipo de proceso fue llevado a cabo mediante un sistema conocido como sistema ETL .Por tanto, se resalta que la generación de un protocolo de búsqueda se considera una contribución muy importante por parte del presente artículo.

La fase de depuración de la base de datos obtenida se realizó con el objetivo de filtrar aquellos datos que no fueron rele-

vantes para el análisis posterior, inicialmente se filtraron atributos mediante el análisis de las variables, como resultado fueron seleccionadas trece variables relevantes, de igual manera se realizó un filtrado de registros con el fin de eliminar registros almacenados que afectan al proceso de descubrimiento del conocimiento. En el trabajo relacionado realizado por (Gomes Barcellos, 2020), ejecutan un proceso similar al aplicado en el presente artículo con respecto a la depuración del conjunto de datos utilizado, en el cual realizan 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.

En la fase de minería de datos se aplicó los algoritmos de árboles de decisión CHAID, CHAID Exhaustivo y CRT; redes neuronales Perceptrón Multicapa y de Base de Función Radial; y redes bayesianas Naive Bayes y BayesNet, mediante las herramientas SPSS Statistics y Weka; si bien actualmente existen muchos algoritmos de minería de datos, se optó por los antes mencionados debido a la literatura encontrada en la que mencionan que los algoritmos más utilizadas en el campo de la seguridad vial, analizando siniestros de tránsito son los Árboles de Decisión, las Redes Neuronales y las Redes Bayesianas.

En los trabajos relacionados realizados por Atnafu Kaur (2017), Ospina-Mateus Quintana Jiménez (2019), Yuan et al. (2017), Makkar et al. (2017), Almamlook et al. (2019), se realiza una aplicación de los algoritmos antes mencionados bastante similar mediante Weka, R y Rapid Miner, pero no todos aplican 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.; dentro del desarrollo del presente trabajo se debe destacar, que la aplicación de los algoritmos tuvo el propósito de determinar los factores más influyentes para la ocurrencia de siniestros de tránsito en Ecuador en el año 2020. Se debe exponer además que la ejecución de los algoritmos en las herramientas SPSS Statistics y Weka resulta más sencilla que al hacerlo en un lenguaje de programación, debido a que se la puede realizar mediante la interfaz gráfica que ofrecen estas herramientas.

Con respecto a la fase de interpretación de resultados, primero se realizó la evaluación de cuales fueron los algoritmos con mejores resultados, para que a través del mejor se presenten los mismos de manera simplificada y comprensibles, resaltando que 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 artículo se realizó el proceso de evaluación mediante la comparación de las métricas de rendimiento, esta 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 mejor algoritmo.

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. El trabajo relacionado de AlKheder et al. (2020), realiza el mismo proceso de evaluación de los algoritmos aplicados con el fin de decidir que algoritmo fue mejor en la predicción de las variables independientes, esto ayudó a entender que algoritmo funcionó con más precisión con los datos utilizados para la predicción de la muestra para cada factor, en este estudio el algoritmo de Red Bayesiana tuvo la mayor precisión, seguido de algoritmo CHAID Exhaustivo, dejando al final al algoritmo Máquina de Vectores de Apoyo (SVM) debido a que tuvo la menor precisión en comparación con los otros algoritmos.

A través del 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 distractor) 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, por el contrario, con el estudio de Pumares (2019) el cual presenta 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 %.

CONCLUSIONES

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 trabajo, 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 de clasificación, 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 el contexto de ocurrencia de 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 desarrollar el presente trabajo.

A través de la aplicación de minería de datos, 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 cinco causas probables principales que son: conducir en sentido contrario a la vía normal de circulación, no respetar las señales reglamentarias de tránsito (pare, ceda el paso, luz roja del semáforo, etc.), no guardar la distancia lateral

e-ISSN: 1390-5902
CEDAMAZ, Vol. 11, No. 02, pp. 124–132, julio–diciembre 2021
DOI: 10.54753/cedamaz.v11i2.1181

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 (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).

AGRADECIMIENTOS

A la Universidad Nacional de Loja y a los distinguidos docentes de la Carrera de Ingeniería en Sistemas por los conocimientos brindados y de manera especial al Ing. Oscar Cumbicus Pineda, por su guía durante esta investigación.

CONTRIBUCIONES DE LOS AUTORES

Conceptualización, YTQ; metodología, YTQ; análisis formal, YTQ; investigación, YTQ; recursos, YTQ; curación de datos, YTQ; redacción — preparación del borrador original, YTQ; redacción — revisión y edición, YTQ. Todos los autores han leído y aceptado la versión publicada del manuscrito. Yulissa Torres-Quezada: YTQ.

FINANCIAMIENTO

El presente estudio tuvo un financiamiento de procedencia propia por parte del autor.

REFERENCIAS

Aftab, U., y Siddiqui, G. F. (2018). Big Data Augmentation with Data Warehouse: A Survey. 2018 IEEE International Conference on Big Data (Big Data), 2785–2794. https://doi.org/10.1109 / BigData.2018.8622206

AlKheder, S., AlRukaibi, F., y Aiash, A. (2020). Risk analysis of traffic accidents' severities: An application of three data mining models. ISA Transactions, 106, 213–220. https://doi.org/10.1016/j.isatra.2020.06.018

Almamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., y Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings, 272–276.

ANT. (2020a). Reporte Nacional de Siniestros de Tránsito.

ANT. (2020b). Siniestros de Tránsito.

Atnafu, B., y Kaur, G. (2017). Analysis and Predict the Nature of Road Traffic Accident Using Data Mining Techniques in Maharashtra, India. Gagandeep Kaur International Journal of Engineering Technology Science and Research IJETSR Www.Ijetsr.Com ISSN, 4(1), 2394–3386.

Calle Reinoso, E. F., Sarabia Paucay, L. T. (2020). Desarrollo de una base de datos para evaluar la percepcion de la seguridad vial en Ecuador. Universidad Politecnica Salesiana.

Constante Tipán, N. V. (2017). 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.

Gomes, R., Barcellos, S. (2020). Brazilian federal roads: identifying patterns in traffic accidents using data mining techniques with apriori algorithm. December..

IBM. (2021). SPSS Software | IBM. https://www.ibm.com/analytics/spss-statistics-software
INEC, T. (2020). Anuario de Estadísticas de Transporte 2019.

López Maldonado, G. (2013). Análisis de la severidad de los accidentes de tráfico utilizando Técnicas de Minería de Datos. Universidad de Granada.

Makkar, A., Gill, H. S., Scholar, M. T., y Science, C. (2017). A Radical Approach to Forecast the Road Accident Using Data Mining Technique. 2(8).

Organización Mundial de la Salud. (2018). Informe sobre la situación mundial de la seguridad vial 2018.

Ospina-Mateus, H., y Quintana Jiménez, L. A. (2019). Predicción de accidentes viales en Cartagena, Colombia, con árboles de decisión y reglas de asociación. Economía Región, 13(2), 83–115.

Pumares, A. (2019). Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador. Universidad Tecnológica Israel.

Rodriguez Hassiger, M. M. (2014). Aplicacion de tecnicas de mineria de datos en accidentes de trafico. Universidad Politecnica de Valencia.

Roman Matamoros, D. X. (2015). Integracion de un programa de seguridad vial al modelo Ecuador [Universidad San Francisco de Quito-Ecuador, Universidad de Huelva-Espana]

Yuan, Z., Zhou, X., Yang, T., Tamerius, J., y Mantilla, R. (2017). Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. Urban Computing, 1–9.