UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE INGENIERÍA CIVIL, DE SISTEMAS Y DE ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

Modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

Presentado por:

Steven Edu Herrera Chirinos

Asesorado por:

Msc. Ing. Maria de los Angeles Guzman Valle

LAMBAYEQUE – PERÚ

2024

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE INGENIERÍA CIVIL, DE SISTEMAS Y DE ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

Modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos

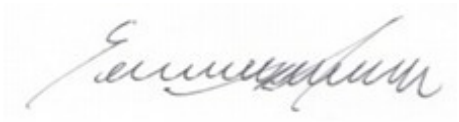
PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

APROBADO POR LOS MIEMBROS DEL JURADO:

Dr. Ing. Ernesto Karlo Celi Arévalo

PRESIDENTE



Dr. Ing. Edward Ronal Haro Maldonado

SECRETARIO

Dr. Ing. Juan Elías Villegas Cubas

VOCAL

LAMBAYEQUE – PERÚ

2024

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE INGENIERÍA CIVIL, DE SISTEMAS Y DE ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

Modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

Steven Edu Herrera Chirinos

Msc. Ing. Maria de los Angeles Guzman Valle

ASESOR

LAMBAYEQUE – PERÚ

2024

DEDICATORIA

A mi madre, por darme su apoyo incondicional, ese apoyo, esa palabra de aliento que siempre necesito, para no tirar la toalla, a ella que me demuestra que hasta en los momentos más críticos siempre se puede salir airoso.

AGRADECIMIENTO

Gracias a Dios por ser esa esperanza que siempre necesitamos para seguir adelante, mis padres, mi abuela, familia y amigos que confían en mí y me brindan su apoyo en todo momento. Gracias a mi universidad y a mis profesores por formarme con sus conocimientos y consejos en mi carrera profesional. Gracias a cada uno de ustedes por ser parte, para que este informe de tesis esté terminado.

RESUMEN

En Perú, la gestión eficiente del tráfico en peajes es crucial para mejorar la seguridad vial. La minería de datos y modelos predictivos son herramientas esenciales para entender patrones del transporte y tomar decisiones estratégicas que optimicen las operaciones de peaje y el flujo vehicular, debido a esta problemática, se propone implementar un modelo predictivo basado en minería de datos que permita predecir patrones de tráfico en unidades de peaje peruanos, con una metodología con enfoque cuantitativo, de tipo aplicada y con un diseño no experimental, con una población de 40 177 registros y 10 atributos, proveniente de Open Data, OSITRAN, como técnica se utilizaron los árboles de decisión de tipo regresión, con la finalidad de realizar la predicción de flujo vehicular en los peajes, optando por utilizar el 75% de los datos con la finalidad de realizar el entrenamiento del algoritmo, y utilizando el 25% restante como control, dando como resultado de la evaluación del modelo predictivo, muestra un 84% de en la precisión de la predicción, en conclusión, es posible realizar la implementación de la metodología, porque las pruebas de software garantizan el correcto funcionamiento del algoritmo de predicción.

Palabras clave: Minería de datos, Predicción, Tráfico, Metodología CRIPS-DM, algoritmo de árbol de decisiones.

ABSTRACT

In Peru, efficient traffic management at tolls is crucial to improve road safety. Data mining and predictive models are essential tools to understand transportation patterns and make strategic decisions that optimize toll operations and vehicle flow. Due to this problem, it is proposed to implement a predictive model based on data mining that allows predicting patterns. of traffic in Peruvian toll units, with a methodology with a quantitative approach, of an applied type and with a non-experimental design, with a population of 40,177 records and 10 attributes, coming from Open Data, OSITRAN, as a technique the trees of regression type decision, with the purpose of predicting vehicular flow at tolls, choosing to use 75% of the data for the purpose of training the algorithm, and using the remaining 25% as a control, resulting in From the evaluation of the predictive model, it shows an 84% accuracy of the prediction. In conclusion, it is possible to implement the methodology, because the software tests guarantee the correct functioning of the prediction algorithm.

Keywords: Data mining, Prediction, Traffic, CRIPS-DM Methodology, decision tree algorithm.

INDICE DE CONTENIDO

INDICE DE TABLAS

TABLA DE FIGURAS

INTRODUCCIÓN

En la actualidad, la gestión eficiente del tráfico en las unidades de peaje es un desafío clave mejorar el tráfico y la seguridad vial en Perú. El análisis y comprensión de los patrones del transporte por carretera juega un papel fundamental a la hora de tomar decisiones estratégicas y aplicar medidas efectivas para optimizar las operaciones de peaje. En este contexto, la minería de datos y los modelos predictivos han surgido como herramientas poderosas para el estudio y la predicción de los flujos de tráfico en estas áreas (Andina, 2021).

Este estudio se centra en aplicar un modelo de predicción basado en minería de datos para analizar y pronosticar patrones de tráfico en las unidades de cobro de peaje peruanos. La minería de datos se refiere a la extracción de información útil y conocimientos relacionados a partir de grandes conjuntos de datos. Al aplicar técnicas de minería de datos a los datos de tráfico recopilados en las cabinas de peaje, es posible identificar patrones, correlaciones y tendencias ocultas que pueden ayudar a comprender y predecir el comportamiento del tráfico.

El principal objetivo de este modelo predictivo es proporcionar a las autoridades de tráfico y a los gestores de peajes información valiosa que les permita tomar decisiones más informadas y eficaces. Al predecir los patrones de tráfico, se pueden anticipar y mitigar los problemas de congestión, establecer estrategias de peaje dinámicas, mejorar la planificación de infraestructuras y brindar una experiencia de conducción más fluida y segura para los usuarios de las carreteras.

En este estudio, se emplearán diferentes técnicas de data mining, como el análisis de series temporales, la clasificación y la regresión, con el fin de construir un modelo predictivo robusto y preciso. Además, se utilizarán datos históricos de tráfico de unidades de peaje peruanos para entrenar y validar el modelo, garantizando su capacidad para capturar y predecir los patrones de tráfico de manera efectiva.

En resumen, el objetivo principal de este trabajo es desarrollar un modelo de predicción basado en minería de datos para analizar y pronosticar patrones de tráfico en las unidades de cobro de peaje peruanos. Se espera que el modelo proporcione información valiosa y respalde la toma de decisiones estratégicas para mejorar la gestión del tráfico, reducir la congestión vial y promover una movilidad más eficiente en el Perú.

CAPÍTULO I: GENERALIDADES

1.1. Planteamiento del problema

1.1.1. Síntesis de la problemática

Hoy en día, las ciudades modernas de todo el mundo se caracterizan por altos índices de movimiento de personas y bienes, una dinámica ligada a los aspectos físicos de las actividades poblacionales y su distribución territorial. Cuanto más se desarrolla una ciudad, más fomenta los viajes más largos (El Comercio, 2019).

Según la consultora de transporte "INRIX Global Traffic Scorecard 2018", que realiza la investigación más profunda de su tipo en el mundo, que analiza y elabora un ranking del impacto de la congestión del tráfico en todo el mundo, estiman que el tiempo un coche queda atrapado en atascos al desplazarse de un lugar a otro, es de 80 horas al año (BBCM, 2017).

El tráfico vehicular en el Perú es un desafío creciente que afecta a las principales ciudades del país. A medida que la urbanización y la población continúan en aumento, la congestión vial se ha convertido en un problema significativo que afecta a la movilidad, la economía y la calidad de vida de los peruanos, el aumento en la posesión de vehículos particulares, la falta de inversión en infraestructura vial adecuada y la ausencia de un transporte público eficiente son solo algunos de los factores que contribuyen a la congestión del tráfico en las principales ciudades del Perú. La movilidad urbana se ha vuelto un desafío complejo que requiere una planificación cuidadosa, la adopción de nuevas tecnologías y cambios en la mentalidad de la sociedad para fomentar una movilidad sostenible.

En Perú, el índice nacional de tráfico de vehículos de peaje aumentó un 19,6% en abril de este año respecto al mismo mes de 2020, informó el Instituto Nacional de Estadística e Informática (INEI). Además, en este resultado influyó el aumento del tráfico de vehículos pesado y ligeros, del 84,7% y 141,8% respectivamente (Andina, 2021).

Por lo cual está presente investigación, se centrará en responder al problema ¿De qué manera un modelo predictivo basado en técnicas de minería de datos permite predecir patrones de tráfico en las unidades de peaje peruanos?

En relación a lo mencionado anteriormente, se tomarán en cuenta los factores centrándose principalmente en el flujo de tráfico, es decir que a través de utilización de un

método predictivo basado de minería de datos se pretende establecer un comportamiento en relación a los patrones de tráfico.

1.1.2. Formulación del problema

Problema general

¿De qué manera un modelo predictivo basado en técnicas de minería de datos permite predecir patrones de tráfico en las unidades de peaje peruanos?

Problema especifico

- ¿Se podrá obtener y almacenar datos históricos detallados de tráfico en las unidades de peaje peruanos?
- ¿Se podrá Identificar las variables más influyentes en los patrones de tráfico, prepararlos para su uso en el modelo?
- ¿Cómo diseñar y construir modelos de minería de datos, en árboles de decisión, y entrenarlos utilizando datos históricos para predecir patrones de tráfico futuros?
- ¿Se podrá evaluar el rendimiento de los modelos mediante métricas de precisión y error, y refinarlos según sea necesario para mejorar su capacidad predictiva en las unidades de peaje peruanos?

1.2. Objetivos

1.2.1. Objetivo general

Implementar un modelo predictivo basado en minería de datos que permita predecir patrones de tráfico en unidades de peaje peruanos.

1.2.2. Objetivos específicos

- Obtener y almacenar datos históricos detallados de tráfico en las unidades de peaje peruanos.
- Identificar las variables más influyentes en los patrones de tráfico, prepararlos para su uso en el modelo.
- Diseñar y construir modelos de minería de datos, en árboles de decisión, y entrenarlos utilizando datos históricos para predecir patrones de tráfico futuros.
- Evaluar el rendimiento de los modelos mediante métricas de precisión y error, y refinarlos según sea necesario para mejorar su capacidad predictiva en las unidades de peaje peruanos.

1.3. Justificación de la investigación

1.3.1. Práctico

Desde el punto de vista práctico esta investigación, parte de la necesidad de identificar patrones de tráfico vehicular en peajes peruanos, debido a las consecuencias de los altos índices de tráfico como la contaminación ambiental, los efectos negativos en la salud y disminución de la productividad laboral, es por eso que se plantea un modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos, aplicando la teoría de las diferentes técnicas para la minería de datos.

1.3.2. Social

Desde la perspectiva social, esta investigación propone un modelo predictivo basado en minería de datos para predecir patrones de tráfico en unidades de peaje peruanos, esto indudablemente favorece a la población, brindando una serie de factos recurrentes de tráfico, para tomar medidas de precaución y evitar la congestión vehicular.

CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes

Internacionales

Yan, Shen (2022) en su artículo científico titulado "Road Accident Severity Prediction Based on Random Forest" con traducción al español "Predicción de la gravedad de los accidentes de tráfico basada en Random Forest". Predecir la gravedad de los accidentes de tráfico es importante para la gestión de accidentes. El estudio propone un modelo híbrido que integra bosque aleatorio (RF) y optimización bayesiana (BO), para predecir la gravedad de los accidentes de tráfico en vías urbanas. En el modelo propuesto BO-RF, RF se utiliza como modelo de predicción básico y BO se utiliza para ajustar los parámetros de RF. Los resultados experimentales muestran que BO-RF no sólo tiene una buena precisión de predicción, sino que también predice más que los modelos tradicionales de aprendizaje automático, sino que también proporcionan resultados explicables. Al predecir con precisión la gravedad de los accidentes de tránsito, los controladores de tránsito pueden tomar medidas oportunas para reducir los efectos secundarios de los accidentes, como brindar asistencia médica oportuna a las víctimas de lesiones en accidentes de tránsito, reduciendo así el número de víctimas. Además, los principales factores que influyen en la gravedad de los accidentes de tráfico pueden determinarse por la importancia relativa del

modelo propuesto. La forma en que influyen en la gravedad de los accidentes de tráfico se puede estudiar mediante un gráfico de dependencia parcial. Los resultados proporcionan sugerencias básicas para implementar medidas para reducir la gravedad de las consecuencias de los accidentes y mejorar la seguridad vial.

Torres (2021) en su investigación titulada "Data mining to determine the most influential factors in the occurrence of traffic accidents in Ecuador in the year 2020" con traducción al español "Minería de datos para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador en el año 2020", mencionó que actualmente la incidencia de accidentes de tránsito es un problema de salud pública a nivel nacional y regional, causando pérdidas de vidas y aumentando día a día en el mundo, por ello, la propuesta de un estudio para determinar las variables que causan los accidentes de tránsito es fundamental e importante. . En este estudio se utilizaron técnicas de minería de datos para identificar patrones que influyen en la ocurrencia de accidentes viales en Ecuador durante el año 2020. Para ello, se utilizaron cinco etapas de la metodología descubrimiento de conocimiento en bases de datos (KDD): veces búsquedas de información, recolección de datos, base de datos limpieza, aplicación de técnicas de minería de datos, interpretación y presentación de los resultados, utilizados para identificar patrones ocultos en el conjunto de datos, perteneciente a la Agencia Nacional de Transporte (ANT) con 418 variables, así como 16.972 registros de accidentes viales en el Ecuador. Se aplicaron siete técnicas de minería de datos, tales como: CHAID, Full CHAID, CRT, Perceptrón multicapa, Función de base radial, Naive Bayes y BayesNet. Los mejores resultados se obtienen utilizando el algoritmo CHAID integral, que identifica los patrones más importantes en los datos y evalúa posibles correlaciones entre las variables recopiladas. Finalmente, con una probabilidad del 69,64%, el factor humano es considerado el componente más influyente.

Jafari et al. (2022) En su artículo titulado "Designing the Controller-Based Urban Traffic Evaluation and Prediction Using Model Predictive Approach" con traducción al español "Diseño de la evaluación y predicción del tráfico urbano basado en el controlador utilizando el enfoque predictivo del modelo" nos menciona que a medida que la sociedad crece, la población urbanizada prolifera y la urbanización se acelera. Los crecientes problemas de tráfico afectan el proceso normal de la ciudad. El sistema de transporte urbano es vital para el funcionamiento efectivo de cualquier ciudad. La ciencia y la tecnología son elementos críticos para mejorar el rendimiento del tráfico en las zonas urbanas. En este trabajo se propone una nueva estrategia de control, basada en la selección el tipo de semáforo

y la duración de la fase verde para lograr un equilibrio óptimo en las intersecciones. Este equilibrio debe ser adaptable al comportamiento fijo del tiempo y la aleatoriedad en una situación de tránsito; el objetivo del método propuesto es reducir el volumen de tráfico en el transporte, la demora promedio de cada vehículo y controlar el choque de automóviles. Debido a la distribución del tráfico urbano y la red de transporte urbano entre los métodos inteligentes para el control del tráfico, el sistema multifactor ha diseñado como un modelo adecuado, inteligente, emergente y exitoso. El control del tráfico en las intersecciones se verifica a través de la temporización adecuada de los semáforos modelada en sistemas multifactoriales. Su capacidad para resolver problemas complejos del mundo real ha convertido a los sistemas multiagente en un campo de inteligencia artificial distribuida que está ganando popularidad rápidamente. El método propuesto investigó explícitamente en la intersección a través de un tiempo de semáforo apropiado mediante el muestreo de un sistema multiagente. Consta de muchas intersecciones, y cada una de ellas se considera un agente independiente que comparte información entre sí. La estabilidad de cada agente se prueba por separado. Una de las características más destacadas del método propuesto para la programación de semáforos es que no hay límite para el número de intersecciones y la distancia entre las intersecciones. En este artículo, se propuso un método de control predictivo del modelo para la estabilidad de cada intersección; los resultados de la simulación demuestran que el controlador de modelo predictivo en este sistema predictivo de modelo multifactorial es más valioso que la programación en el método de tiempo fijo. Reduce la longitud de las colas de vehículos.

Xu et al. (2021), en su artículo científico mencionan que la extracción adecuada de patrones de movilidad individual con fuentes de datos de alta resolución, como la que se extrae de las aplicaciones para teléfonos inteligentes, ofrece importantes oportunidades. Las oportunidades potenciales que no ofrecen los registros detallados de llamadas (CDR), que ofrecen resoluciones trianguladas desde antenas, son opciones de ruta, detección de modos de viaje y encuentros cercanos. Hoy en día, no existe un conjunto de datos estándar y de grandes escalas recopiladas durante largos períodos que nos permita caracterizarlos. En este trabajo, analiza de cerca el uso de datos de aplicaciones de teléfonos inteligentes, también conocidos como datos de servicios basados en la ubicación (LBS), para extraer y comprender el comportamiento de elección de rutas de los vehículos. Tomando como ejemplo el área metropolitana de Dallas-Fort Worth, primero se extrae los viajes vehiculares con reglas simples y reconstruimos la matriz origen-destino acoplando los viajes vehiculares extraídos

de los usuarios activos de LBS y los datos del censo de los Estados Unidos. Luego presenta un método para derivar las rutas comúnmente utilizadas por las personas a partir de las trazas LBS con intervalos de frecuencia de muestreo variables. Además, inspecciona la relación entre el número de rutas y las características del viaje, incluida la hora de salida, la duración del viaje y el tiempo de viaje. Específicamente, considera el índice de tiempo de viaje y el índice de amortiguamiento para los usuarios de LBS que toman diferentes números de rutas. Los resultados empíricos demuestran que durante las horas pico, los viajeros tienden a reducir el impacto de la congestión del tráfico tomando rutas alternativas. En general, el marco de análisis de datos propuesto es rentable para tratar los datos escasos generados por el uso de teléfonos inteligentes para informar el comportamiento de enrutamiento. El potencial en la práctica es informar las estrategias de gestión de la demanda, dirigiéndose a usuarios individuales mientras se generan estimaciones a gran escala de la mitigación de la congestión.

ULLAH et al. (2019) en su trabajo de investigación titulado "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector" con su traducción en español "Un modelo de predicción de abandono utilizando Random Forest: Análisis de técnicas de aprendizaje automático para predicción de abandono e identificación de factores en el sector de las telecomunicaciones" nos esboza como principal problemática que en este caso está relacionada con el sector de telecomunicaciones las cuales abarcan el entorno de la migración de clientes entre proveedores, el tipo de servicio con el que cuenta y el grado de vinculación del servicio (pospago y prepago). El contar con un servicio prepago los hace más próximos a desertar del servicio o cambiar de proveedor. Además que se pierdan clientes fieles puede repercutir no solo en sus ingresos, sino en su marca ya que puede traer a una baja de clientes potenciales, ya que en algunos casos estos suelen ser familiares, amigos o conocidos, por lo cual se amerita que la predicción de rotación es esencial en dicho sector, ya que permite la toma de decisiones entorno a la retención y fidelización de clientes preciados, además que permita desarrollar sus CRM, ya que estos requieren entender los motivos que provocan perder clientes, y también reconocer las pautas de conducta de los clientes existentes. Por lo tanto, se planteó un modelo anticipatorio de abandono que emplea múltiples técnicas de algoritmos de Machine Learning (Aprendizaje automático), además de que el modelo es evaluado empleando métricas de restablecimiento de información y la exactitud del modelo se realiza aplicando frecuencias FP, TP, Medida f, Precisión, Área

ROC y Recuperación. Por consiguiente, el desarrollo de este estudio se centra en identificar una técnica de Machine learning y minería de datos que permita determinar un modelo predictivo de deserción de clientes que permita establecer factores desertivos y aportar estrategias de retención. Para la selección de atributos se emplean técnicas de selección como el filtro correlacional de catalogalización de atributos y el descubrimiento de información. Además, se emplea técnicas de Machine Learning para clasificar en dos tipos de clientes de sector de telecomunicaciones: desertores y no desertores. Siendo predominante la aplicación del algoritmo Dom Forest ya que produce el mayor grado de precisión con respecto a otros, también se estableció un perfil de cliente basado en su comportamiento y usando el agrupamiento K-medias para finalmente permitir agruparlos en tres tipos: bajo, medios y de riesgo. De forma más sintética este modelo se divide en los pasos de preprocesamiento de datos (subprocesos de eliminación de ruido y selección de características); clasificación y predicción del cliente(se utilizan algoritmos catiónicos, incluyendo árbol de decisión, bosques aleatorios[RF], árbol aleatorio con aprobación cruzada con un ciclo de diez, J48, Decision Stump, Decision Stump más AdaboostMl, bagging más árbol aleatorio, Naive Bayes[NB], IBK, Regresión logística[LR], Perceptron[MLP] y LWL, además de emplearse Weka para la simulación); la creación de perfiles de clientes(se emplea técnicas de agrupamiento de K-means); el estudio de clústeres(en función a la herencia de comportamiento de los clientes, obtenido de los datos evaluados) y la recomendación de formas de retención para los diversos tipos de clientes desertores. Se obtuvo como resultado que el bosque aleatorio tuvo una mayor eficiencia en la clasificación correcta con un 88,63%, mientras que J48 con un 88.58%, Árbol aleatorio con un 84.34%, Decisión Stump con un 70.98%, Decisión Stump y AdaboostMl con un 83.95%, bagging más árbol aleatorio un 88.61%, el perceptrón con un 82.04%. Además, Rmdon Forest demoro 108,48 segundos en ejecutar la precisión, pero obtuvo el mayor porcentaje, además de tener una tasa TP y FP es alta. Por último, se obtienen cuatro factores (OFFNET_CALLS, ONNET_CALLS, TOTAL_CALLS y OFFNET_MINS; TOTAL_CALLS_REV; TOTAL_CALLS, REVENUE_SMS, RECHRG_TOTAL_LOAD, TOTAL_MINS y FREE_MINS; un factor cuatro) que permiten perfilar a los clientes usando K-medias el arriesgado, medio y el bajo. Para finalmente obtener como conclusión más relevante del estudio que el algoritmo J48 y Aletorio Forest proporcionan un resultado sobresaliente de la medida F con 88%.

Zhang et al. (2022) en su trabajo de investigación se propone un método de reconocimiento preciso del patrón de flujo de tránsito aéreo basado en datos históricos de

trayectorias de vuelo para comprender la distribución espacial del flujo de tránsito aéreo y mejorar el aprovechamiento del espacio aéreo dentro de la zona terminal. Debido a la alta dimensionalidad de los datos de trayectoria de vuelo, establecemos un modelo de similitud de trayectoria basado en la distancia geodésica y utilizamos un algoritmo de agrupación espectral mejorado para clasificar los datos de muestra de trayectoria de vuelo. Se presenta un algoritmo mejorado basado en el árbol de tensión mínima para extraer la similitud del esqueleto entre las vías y obtener el modelo del flujo de tráfico predominante. Los resultados experimentales muestran que el método puede dividir con precisión 1070 rutas de vuelo en 5 categorías y extraer 7 flujos de tráfico predominantes, mostrando una gran solidez frente a trayectorias anormales y ruido. En conclusión, en el artículo se establece un modelo de cálculo de similitud basado en la distancia geodésica para reflejar las características espaciales de los datos de trayectoria de vuelo de alta dimensión. Las muestras de la trayectoria de vuelo en el área terminal se agruparon utilizando el algoritmo de agrupamiento espectral. Los hallazgos derivados del análisis de caso evidenciaron que se obtuvieron resultados de agrupamiento satisfactorios a partir del modelo de cálculo de similitud basado en la distancia geodésica. El árbol de tensión mínima se utilizó para extraer las trayectorias de vuelo clave del grupo y el esqueleto de similitud. Las trayectorias de vuelo clave exhibieron una gran robustez frente a la interferencia del ruido. Este estudio proporciona una base técnica para una mayor optimización del sector del espacio aéreo terminal y las rutas de entrada y salida para adaptarse al flujo de tráfico. Lo estructural las características de la trayectoria de vuelo se determinaron mediante procedimientos estándar. Sin embargo, debido a las diferencias en el desempeño de diferentes aeronaves, la implementación de los procedimientos estándar puede ser diferente.

Nacionales

Tarazona (2016) En su investigación titulada “identification of main factors and variables describing the quantity and distribution of fatal vehicular accidents in metropolitan city of Lima using data minig techiques: ramdon forrest, boosting, decision tres”, El principal objetivo de este estudio fue identificar los factores primordiales que influyen en los accidentes de tránsito mediante el uso de técnicas de Minería de Datos, como Random Forest, Boosting y Árbol de decisiones CART. La investigación se realizó de manera transversal y causal, utilizando los registros de accidentes de tránsito en Lima Metropolitana del año 2014, que estaban disponibles en el sistema de denuncias policiales de comisarías cercanas. Después de analizar los resultados, se pudo observar que los principales factores

que influyen en los accidentes de tránsito en Lima Metropolitana, identificados a través de las técnicas de Minería de Datos mencionadas, incluyen: la ocurrencia del accidente en una avenida o carretera, el incumplimiento de las señales de tránsito por parte del conductor como causa, la presencia de vehículos tipo combi, la invasión de carril y la presencia de vehículos como motocar o mototaxi. Los modelos utilizados mostraron un índice Gini superior al 50% en todos los casos y superior al 70% en el modelo Boosting, además de una sensibilidad superior al 55% en todos los modelos y una especificidad superior al 60% en todos los casos. Los modelos generados también presentaron tasas de error inferiores al 9% en el modelo Boosting, inferiores al 15% en el modelo Random Forest e inferiores al 25% en el modelo Árbol de decisión CART, según se determinó mediante validación cruzada k-fold con k=10. Además, las variables más relevantes identificadas incluyen el tipo de vía donde ocurrió la colisión (carretera), el tipo de vehículo mayor involucrado (camión-combi) y la falta de respeto del conductor a las señales de tráfico, entre otras.

2.2. Bases teóricas

Metodología CRISP-DM

Son las siglas de Cross-Industry Standard Process for Data, siendo una metodología óptima para la realización de tareas de minería de datos, como metodología, tiene seis fases para la realización de un proyecto, así como una explicación de la relación entre las tareas, asimismo, como modelo de proceso, brinda un resumen del proceso para minería de datos **Fuente especificada no válida.**.

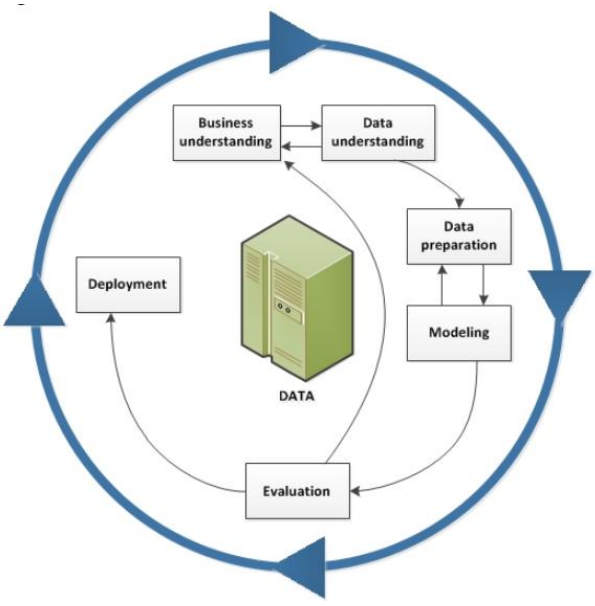


Figura 1. Ciclo de vida de minería de datos.

Comprensión del negocio

El conocer los negocios es de vital importancia, esto porque brinda diversas ventajas, en diferentes negocios, esto es el inicio de la Inteligencia de Negocios (Business Intelligence, BI), esto permite tomar decisiones estratégicas, evitar sobrecostos, etc. **Fuente especificada no válida.**.

Es fundamental invertir tiempo en comprender las expectativas de la organización sobre el uso de la minería de datos. Busque la participación activa de un amplio grupo de personas en estas conversaciones y registre detalladamente los resultados obtenidos. La última etapa del proceso CRISP-DM se centra en la elaboración de un plan de proyecto basado en la información recopilada en esta documentación **Fuente especificada no válida.**.

Determinación de los objetivos comerciales

Es el primer paso para determinar los objetivos comerciales, debiendo ser una tarea minuciosa, debido a que esta información permite la reducción de riesgos mediante la clarificación de problemas, objetivos y recursos, teniendo 3 fases **Fuente especificada no válida.**:

- Recopilación de la información actual.
- Registro de los objetivos específicos comerciales.
- Criterios utilizados para determinar rendimientos en el proceso de minería de datos.

Evaluación de la situación

Evalúa la situación actual **Fuente especificada no válida.**:

- ¿Qué datos están disponibles?
- ¿Se dispone del personal necesario?
- ¿Cuáles son los factores de riesgo principales?
- ¿Existen planes de consistencia para cada factor?

Determinación de los objetivos de minería de datos

Continúa luego de la determinación de los objetivos comerciales, consiste en **Fuente especificada no válida.**:

- Identificar clientes.
- Crear un modelo con los datos disponibles.

- Asignar un rango.

Producción de un plan de proyecto

Toma en conjunto los objetivos comerciales y de minería de datos para elaborar un plan, consiste en **Fuente especificada no válida.**:

- Escribir un plan de proyecto.
- Realizar un piloto.
- Evalúa las técnicas y herramientas.

Compresión de datos

Estudia de forma cercana los datos disponibles para la minería de datos, esencial para la prevención de inconvenientes para la preparación de datos, implica explorar y acceder, mediante tablas y/o gráficos, además, permite determinar la calidad de los datos, así como una descripción de los resultados **Fuente especificada no válida.**.

Recopilación de datos iniciales

Se recopilan datos de orígenes variados, como **Fuente especificada no válida.**:

- Datos existentes
- Datos adquiridos
- Datos adicionales

Descripción de los datos

Existen diversas formas para la descripción de datos, principalmente centradas en la cantidad y claridad de los mismos, con las características de **Fuente especificada no válida.**.

- **Cantidad de datos.** Debe existir un equilibrio en los datos, una mayor cantidad de datos da resultados más precisos, sin embargo, conlleva un mayo tiempo de procesamiento, considerar los atributos al describir los datos.
- **Tipos de valores.** Pueden ser numéricos, categóricos o Booleano
- **Esquemas de codificación.** Características como tipo de producto o género.

Registros de datos

Acción que implica almacenar algún contenido o información en un documento específico. Por otra parte, los datos representan información que facilita el acceso al conocimiento **Fuente especificada no válida.**.

Atributos de datos

Principalmente, existen 3 categorías par la selección de atributos, métodos de filtrado, métodos de envoltura y métodos integrados **Fuente especificada no válida.**

Exploración de datos

Realizado mediante tablas, gráficos y diversas herramientas, ayuda a describir los objetivos de minería de datos, así como formular hipótesis y dar forma a las tareas **Fuente especificada no válida.**.

Verificación de la calidad de los datos

En general, los datos no son perfectos, estos tienen errores en la codificación, valores perdido y demás, para evitar esto, se realiza el análisis de la calidad de los datos recolectados previos al modelado, para realizar la verificación de estos datos se buscan **Fuente especificada no válida.**.

- **Datos perdidos.** Vacíos o sin respuesta.
- **Errores de datos.** Tipográficos al introducir datos.
- **Errores de medición.** Incluye datos erróneos, basados en mediciones incorrectas.
- **Codificación incoherente.** Incluyen unidades no estándar de medición o valores incoherentes.
- **Metrados erróneos.** Errores entre el significado o definición de un campo.

Preparación de datos

Es el aspecto más relevante y el más tiempo necesita para la minería de datos, conlleva entre el 50% y 70% del tiempo, suele indicar **Fuente especificada no válida.**:

- Fusionar conjuntos y/o registros.
- Seleccionar muestras de subconjuntos.
- Agregar registros.
- Derivar nuevos recursos.
- Eliminar o sustituir datos perdidos.
- Dividir en conjuntos de prueba y entrenamiento.

Selección de datos

Esta etapa selecciona los datos necesarios relevantes para cada objetivo, existiendo dos formas para la selección **Fuente especificada no válida.**:

- **Por filas.** Cuentas, productos o clientes.
- **Por columnas.** Por características.

Limpieza de datos

Observa de manera minuciosa los problemas presentes en los datos, utiliza el informe de calidad de datos, ya que este caracteriza los problemas concretos de los datos **Fuente especificada no válida.**.

Construcción de nuevos datos

Frecuentemente es necesario construir nuevos datos, existiendo dos formas de construir datos nuevos **Fuente especificada no válida.**:

- Derivar atributos.
- Generar registros.

Integración de datos

Compuesto or diversos orígenes para el mismo conjunto, en caso que los conjuntos compartan el mismo identificador, pueden fusionarse, existiendo dos métodos para la integración de datos **Fuente especificada no válida.**.

- **Fisión de datos.** Unión de datos con registros similares, pero con atributos diferentes.
- **Adición de datos.** Unión de datos con atributos similares, pero con registros diferentes.

Formato de datos

El paso previo a la modelación, comprueba las técnicas requeridas, así como clasificación de datos, para un algoritmo de secuencia, es necesario que los datos sean clasificados previo a la ejecuciuón del modelado, ahorra tiempo utilizando un nodo Ordenar, considerando lo siguiente **Fuente especificada no válida.**.

- ¿Qué modelos se pueden usar?
- ¿Requieren un formato o clasificación concreta?

Modelado

Contempla todos los pasos anteriores, utilizando herramientas analíticas y se empiezan a ver resultados de los planteado en la comprensión del negocio, el modelado ejecuta diversas interacciones, por lo general, los analistas ejecutan diversos modelados utilizando parámetros predeterminados, a su vez, ajustan parámetros y regresan al paso anterior para manipular las bases para el modelado, existiendo una variedad de método para resolver un problema en concreto **Fuente especificada no válida.**.

Selección de técnicas de modelado

Basado en **Fuente especificada no válida.**.:

- Datos disponibles
- Objetivos de minería de datos
- Requisitos específicos del modelado

Generación de un diseño de comprobación

Previo a la generación del modelado se debe comprobar los resultados del modelo, siendo un proceso iterativo, comprueba los resultados de varios modelos, siendo 2 verificaciones **Fuente especificada no válida.**:

- **Criterios de bondad.** puede ser medido de formas diverss, para C5.0 y C&R Tree, la medición calcula la tasa de error.
- **Definición de los datos.**

Generación de los modelos

En este punto se generan los modelados considerados, los analistas generan diversos modelos para comparar resultados previo al despliegue, al final del modelado, se dispone de tres tipos de información, la cual se utilizará en la toma de decisiones de minería de datos **Fuente especificada no válida.**.

- Configuración de parámetros
- Modelos reales producidos
- Descripción de resultados

Evaluación del modelado

Considera el conjunto de modelos iniciales, para identificar al más preciso o eficaz, también se puede evaluar desde la percepción de la organización **Fuente especificada no válida.**.

Variables categóricas

Son atributos que representan categorías discretas o etiquetas, en lugar de valores numéricos. Se utilizan para clasificar datos en grupos específicos, facilitando el análisis y modelado en diversos sistemas, como el procesamiento de información y la toma de decisiones automatizada.

Concesión

Proceso donde la Administración Pública o empresas ceden a entidades privadas el derecho de operar bienes y servicios para mejorar su eficiencia. La concesión, gestionada por la empresa ganadora mediante presupuesto público, tiene un plazo fijo establecido por la Administración, otorgado a través de métodos como la concesión directa o el concurso público **Fuente especificada no válida.**.

Unidad de peaje

Permite financiar los gastos de inversión, funcionamiento y mantenimiento que asuma el concesionario. Se podría considerar la posibilidad de establecer nuevos puntos de peaje, siempre y cuando se construyan y operen según lo acordado entre las partes. Esto podría surgir durante la ejecución de obras adicionales o la aplicación de ajustes contemplados en el contrato de concesión. Además, se contemplarían otros casos específicos detallados en dicho acuerdo. **Fuente especificada no válida.**.

Python

Lenguaje de programación de alto nivel, interpretado y versátil, creado por Guido van Rossum en 1991. Su sintaxis clara y legible, junto con su capacidad para soportar múltiples paradigmas de programación, lo convierten en una opción popular entre desarrolladores novatos y experimentados. Es multiplataforma, cuenta con una amplia biblioteca estándar y tiene una activa comunidad de desarrolladores. Python se utiliza en diversas aplicaciones, desde desarrollo web hasta inteligencia artificial y aprendizaje automático. Su enfoque en la simplicidad y productividad lo ha convertido en un pilar esencial en la programación moderna**Fuente especificada no válida.**.

Evaluación

En esta etapa avanzada del proyecto de minería de datos, se verifica la efectividad técnica de los modelos desarrollados. Previo a la continuación, es esencial evaluar los resultados con los criterios comerciales establecidos al inicio. Los resultados comprenden modelos finales y descubrimientos derivados del proceso de minería de datos **Fuente especificada no válida.**.

Evaluación de resultados

Se registra la evaluación, así como el cumplimiento de los criterios, para los resultados de minería de datos **Fuente especificada no válida.**:

- ¿Los resultados que obtuvimos se explican de manera clara y fácil de presentar?
- ¿Hemos hecho descubrimientos especiales o muy relevantes que debamos destacar?
- ¿Podemos evaluar los modelos y hallazgos considerando su aplicabilidad a los objetivos?
- ¿Hasta qué punto estos resultados se alinean con los objetivos de nuestra organización?
- ¿Qué otras preguntas surgen a partir de los resultados?

Proceso de revisión

Se resumen las actividades y las decisiones tomadas en cada fase **Fuente especificada no válida.**.

- ¿Ha aportado esta etapa al valor de los resultados finales obtenidos?
- ¿Existen situaciones sin solución aparente, como modelos específicos que no generan resultados significativos?
- ¿Existen decisiones alternativas o estrategias que podrían implementarse en una fase específica de proyectos de ingeniería de sistemas?

Registre estas alternativas con el objetivo de aplicarlas en futuros proyectos de minería de datos de manera efectiva y exitosa.

Despliegue

La implementación implica aplicar nuevos conocimientos para mejorar la organización, ya sea integrando formalmente modelos como IBM® SPSS Modeler o ajustando la estrategia basándose en la minería de datos. La fase de despliegue de CRISP-

DM abarca la planificación y control de resultados, así como la finalización de tareas de presentación según las necesidades organizativas **Fuente especificada no válida.**.

Planificación del despliegue

En la implementación de resultados de minería de datos, se aconseja al profesional de ingeniería en sistemas peruano que, antes de compartir sus logros, elabore planes detallados para integrar modelos y comunicar descubrimientos. Además, debe considerar posibles problemas y contar con planes alternativos, demostrando una sólida planificación estratégica **Fuente especificada no válida.**.

Planificación del control y del mantenimiento

En un contexto de minería de datos aplicada a la ingeniería de sistemas en Perú, se destaca la importancia de evaluar y ajustar continuamente los modelos desplegados para prever compras en línea y mejorar la retención de clientes clave. La eficacia se asegura mediante la medición constante, y la documentación es esencial para evaluar proyectos futuros **Fuente especificada no válida.**.

Creación de un informe final

Aborda la comunicación de resultados a administradores técnicos y gestión. Al considerar receptores, se destaca la necesidad de informes adaptados a desarrolladores técnicos y gestores de ventas. Elementos clave incluyen descripción del problema, procedimiento de minería de datos, costos, desviaciones del plan original y recomendaciones para futuros proyectos **Fuente especificada no válida.**.

Revisión final del proyecto

En la fase final del método CRISP-DM, se realiza una entrevista con los involucrados en la minería de datos. Se exploran las impresiones generales, conocimientos adquiridos y evaluación de las partes exitosas y desafíos del proyecto. Posteriormente, se entrevistan a personas afectadas por los resultados obtenidos. La recopilación de estas impresiones, junto con observaciones personales, se sintetiza en un informe final que destaca los aprendizajes clave derivados de la experiencia de minería de datos en los almacenes correspondientes **Fuente especificada no válida.**.

Principios de tráfico vehicular

Definición

El efecto originado por la circulación de vehículos en una carretera, calle o autopista se denomina congestión vehicular, o simplemente tráfico. Es esencial tener información sobre las particularidades del tráfico que utilizará esa carretera o calle antes de emprender cualquier planificación geométrica de la vía (MTC, 2018).

La hipótesis del flujo de tráfico se basa en un procedimiento matemático que relaciona los elementos fundamentales del flujo de vehículos, como el caudal, la densidad y la velocidad. Las características y el comportamiento del tráfico pueden descubrirse mediante el estudio del flujo vehicular, así como la forma en que los coches circulan por las carreteras (MTC, 2018).

Es viable entender las particularidades y el comportamiento del tráfico, así como las necesidades fundamentales para la planificación, construcción y operación de carreteras, calles y sus infraestructuras relacionadas en el sistema de transporte, al examinar los aspectos del flujo de vehículos. El análisis del flujo vehicular implica estudiar cómo los automóviles se desplazan en diversos tipos de vías mediante la aplicación de principios de física y matemáticas, lo que permite evaluar la eficacia de la operación. La creación de modelos detallados y generales que relacionan variables como el volumen de tráfico, la velocidad, la densidad, el espacio entre vehículos y el tiempo de seguimiento representa uno de los logros más significativos del análisis del flujo vehicular. Estos modelos sentaron las bases para conceptos como la capacidad y los niveles de servicio, que se aplican a diversas características de las carreteras (MTC, 2018).

Variables de flujo vehicular

La frecuencia, es decir, la cantidad de vehículos que atraviesan un punto o segmento específico de un carril o carretera durante un intervalo de tiempo determinado es conocido como tasa de flujo (q), que se define como: n/t = q (MTC, 2018).

En dónde el Manual de D i s e ñ o Geométrico, DG (2018), menciona:

q = Cantidad de vehículos que transitan por un periodo de tiempo

n = Cantidad total de vehículos que transitan

t = Periodo determinado

Velocidad (v). La distancia que un vehículo viaja en una unidad de tiempo, expresada en kilómetros por hora (km/h), se conoce como velocidad de circulación.

La densidad o concentración (k). En un momento dado, k = n /d es el número de vehículos que ocupan una longitud determinada (d) de un sector de la carretera.

La ecuación vehicular fundamental, q vehículos/tiempo = v k, conecta estas tres variables.

Elementos del transito

Según Jaramillo (2017) La Ingeniería de Tráfico consta de tres componentes esenciales, los cuales son:

El Usuario: En el diseño, el análisis, el proyecto y el funcionamiento de un sistema de transporte vehicular, es fundamental tener en cuenta el comportamiento del usuario. Los peatones y los automovilistas son los actores fundamentales que requieren investigación para garantizar la disciplina y la seguridad en las vías, y el usuario se encuentra estrechamente relacionado con ellos.

El Vehículo: El peso, las dimensiones y las características operativas de un vehículo hipotético se utilizan para construir los parámetros establecerán las directrices para la planificación geométrica de las carreteras, calles e intersecciones, considerando su capacidad para acomodar vehículos de diferentes dimensiones, los cuales se pueden clasificar en dos categorías: vehículos de dimensiones reducidas y vehículos de mayores dimensiones.

La Vía o Vialidad: La calzada o vía por la que se desplazan los vehículos El tercer componente esencial en el sistema de transporte es la red vial, que consiste en una infraestructura diseñada y preparada específicamente para facilitar el flujo ininterrumpido de vehículos con altos estándares de seguridad y comodidad, abarcando tanto el espacio como el tiempo. La calidad de vida de un país está intrínsecamente relacionada con la calidad de su sistema de carreteras, y viceversa.

Congestión vehicular

Causas del congestionamiento

La congestión en las áreas urbanas se deriva principalmente de la dinámica del transporte en la ciudad, que incluye la cantidad significativa de automóviles, la condición de las carreteras, el comportamiento de los conductores y los desafíos en la gestión del transporte público (Márquez, 2015).

Además, Bull (2003) señala que es fundamental recordar que la congestión del tráfico debe ser abordada para que estas intersecciones funcionen bien, y que se puede lograr una

mejora en el sistema de transporte teniendo un mejor plan de mejora y dando suficiente mantenimiento a las avenidas para conseguir mejoras, esto ayudará al usuario tanto en el aspecto de conductor como los peatones que transita la avenida, así como el sistema también aumentará su actitud a la hora de conducir su método de transporte preferido, además de proporcionar un mejor servicio para las tareas que desee.

Efectos negativos

Según Jaramillo (2017) algunos aspectos negativos a mencionar del congestionamiento de tráfico son:

El tiempo de los conductores y pasajeros se desperdicia, lo que representa un "costo de oportunidad". Además, tiene un impacto negativo en la economía regional, ya que la mayoría de las personas no están realizando actividades productivas durante este tiempo. Los retrasos pueden hacer que se llegue tarde al trabajo, a las reuniones y a la escuela, con la consiguiente pérdida de negocios, medidas de disciplina u otras pérdidas en el ámbito personal.

Los conductores dedican más tiempo a viajar y menos a actividades productivas, debido a su incapacidad para prever con precisión el tiempo de viaje.

El incremento en la velocidad, la aceleración y las frenadas resulta en un mayor consumo de combustible, mayor contaminación del aire y emisión de dióxido de carbono, lo cual puede contribuir al fenómeno del calentamiento global. El uso de combustible ha aumentado.

En el caso de ocurrencia de alguna emergencia; Si el tráfico está atascado, los vehículos de emergencia pueden ser incapaces de llegar a sus destinos de manera oportuna.

Minería de Datos

En el pasado, se exploraban conexiones, tendencias, desviaciones, comportamientos excepcionales, patrones y trayectorias ocultas con el fin de respaldar los procesos de toma de decisiones con un mayor nivel de conocimiento. En ese entonces, la Minería de Datos se situaba en el escalón más avanzado de la evolución de los procedimientos tecnológicos para analizar datos. (Beltrán, 2003).

La denominación "Minería de Datos" (Data Mining) se originó a partir de la semejanza entre una montaña y la vasta cantidad de datos resguardados en cualquier organización. Al igual que los diamantes de alto valor que permanecen ocultos entre las

rocas y el suelo dentro de una montaña, estos datos eran descubiertos y aprovechados mediante prácticas mineras (Beltrán, 2003).

Se describía como un conjunto de técnicas que automatizaban la identificación de patrones importantes o como el proceso que habilitaba la conversión de información en conocimiento valioso para la empresa, al revelar y medir conexiones dentro de una amplia base de datos (Timarán et al., 2016)

Técnicas de Minería de Datos

Las metodologías posibilitan la adquisición de patrones o estructuras en relación de unas bases de datos o big data con la que cuentan. De forma general, las técnicas predictivas permiten realizar predicciones futuras de un comportamiento, mientras que las técnicas descriptivas permiten describir un comportamiento para un mejor entendimiento. Por consiguiente, se esquematiza gráficamente la categorización de las metodologías de minería de datos junto con sus correspondientes algoritmos (Espino, 2017).

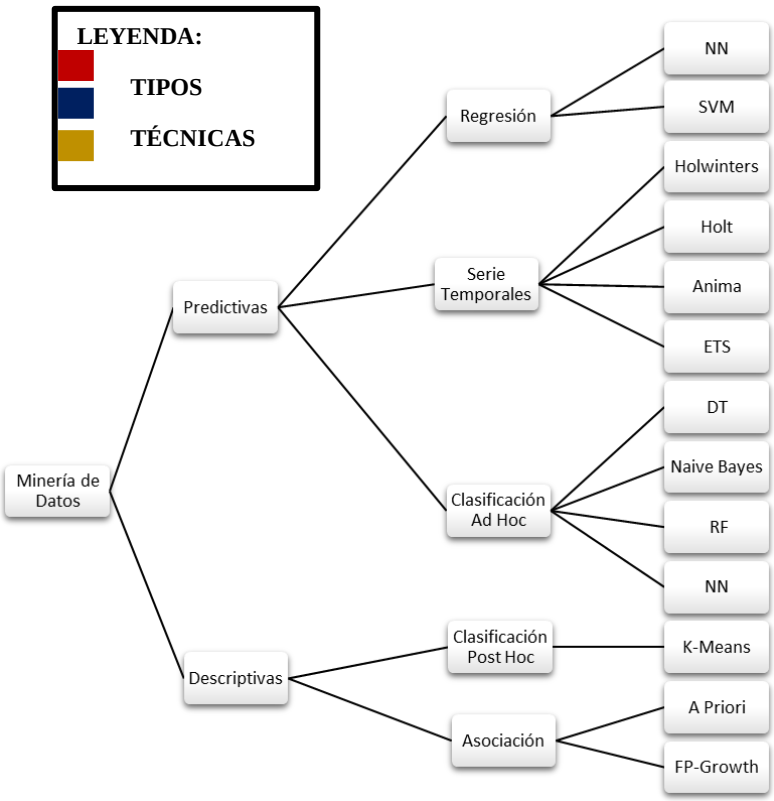


Figura 2. *Técnicas de minería de datos (Pérez & Santín, 2008).*

Nota: (Pérez & Santín, 2008)

Se presenta un resumen gráfico de las metodologías de minería de datos y sus principales submetodologías o algoritmos:

Tabla 1.

Técnicas de Minería de datos y sus principales subtécnicas o algoritmos.

TIPO	CARACTERISTICAS	SUBTÉCNICAS O ALGORITMOS		
Predictivo	Interpolación y Predicción Secuencial	Datos continuos	Regresión Lineal	- Regresión lineal global (clásica). - Regresión lineal ponderada localmente
			Regresión no lineal	- Logarítmica. - Pick & mix. - Entre otras.
		Datos discretos	No hay técnicas específicas: se suelen utilizar técnicas de algoritmos genéticos o algoritmos de enumeración refinados.	
	Aprendizaje supervisado	- K –NN (Nearest Neighbor). - K –means (competitive learning). - Perceptron Learning. - Multilayer ANN methods (e. g. backpropagation). - Radial Basis Functions. - Decision Tree Leaming (e. g. 103, C4.5, CART). - Bayes Classifiers. - Center Splitting Methods. - Rules (CN2).		

		- Pseudo-relational: Supercharging, Pick-and-Mix. - Relational: ILP, IFLP, SCIL - Entre otros
Descriptivo	Exploratorio	- Estudios correlacionales. - Dependencias. - Detección datos anómalos. - Análisis de dispersión.
	Aprendizaje no supervisado (Segmentación)	- K –means (competitive learning). - Redesneuronalesde Kohonen - EM (Estimated Means) (Dempster 1977). - Cobweb (Fisher 1987). - AUTOCLASS

Nota: (Gutierrez & Molina, 2016).

Modelos empleados en el estudio

Figura 3. Algoritmos a evaluar en el estudio (Shewan, 2021)

Nota: (Shewan, 2021)

En este estudio se utilizan los algoritmos para la evaluación y posteriormente el desarrollo del modelo predictivo, los cuales son: SVM**, RF, K-NN, DL** (estas técnicas se

escogieron por su efectividad en otros sectores industriales al predecir y clasificar la deserción de clientes.

SVM (SOPORTE DE MÁQUINAS VECTORIALES)

El algoritmo de machine learning supervisado de Soporte de Máquinas vectoriales (SVM), cuyo objetivo de es encontrar el híper separador óptimo plano que maximice el margen de los datos de entrenamiento dividiendo el espacio n-dimensional representación de los datos en dos regiones mediante un hiperplano. También tiene bases sólidas en la teoría del aprendizaje estadístico es aplicable con eficacia a una amplia gama de problemas de clasificación, tanto lineales como no lineales. Hay muchas funciones basadas en el kernel, como la función lineal del kernel, la poli normalizada kernel, función kernel polinomial, función de base radial (RBF) o kernel gaussiano e Hy-Perbolic Tangent (Sigmoid) Kernel Función sigmoidea que se puede implementar en SVM. SVM genera una etiqueta de clase, ya sea positiva o negativa para cada muestra de caso de binomio, la clasificación: para calcular métricas como la curva ROC, etc. También podemos encontrar la distancia entre del hiperplano que separan las clases. SVM tiene muchas ventajas como obtener el mejor resultado cuando se trata de la representación binaria, capaz de hacer frente a un número bajo de características (Cuevas et al., 2019).

RF (RANDOM FOREST)

Es una técnica predictiva que se basa en la combinación de árboles predictivos, donde cada árbol opera independientemente y de manera no correlacionada. En el método conocido como "Random Forest" o "Selvas Aleatorias", todos los clasificadores utilizados son árboles de decisión. Cada uno de estos modelos produce una predicción, y la predicción final se determina por mayoría de votos (Medina & Ñique, 2017).

K-NN (K Vecinos Próximos)

El método K-NN se basa en una técnica de clasificación supervisada que implica calcular la distancia entre un conjunto específico de muestras (K vecinos) y la muestra que se desea clasificar. Luego, se determina su afinidad con la clase que posee la mayoría de vecinos etiquetados, utilizando el criterio de la mínima distancia. Es relevante destacar que esta metodología es adecuada únicamente para datos numéricos y no es aplicable a clasificadores de texto (5). Para un conjunto de muestras dado x = {x1, x2,..., xn} en un espacio de características Rp, donde se utiliza una función de distancia d, se pueden observar dos aspectos importantes:

- Vecino más cercano: Esto implica identificar la muestra xl en el conjunto x que se encuentra más próxima a la muestra xk, donde l es menor o igual a k y k es diferente de 1.
- Rango r: Cuando se establece un umbral r específico y se tiene un punto xk, se excluyen aquellos puntos xl que cumplan con la condición $0 \leq d(xk, xl) = dkl \leq r$. (Haro et al., 2018).

DL (DEEP-LEARNING)

Este tipo de aprendizaje se refiere a un conjunto de técnicas que permiten a una máquina recibir datos en su forma original y, a través del uso de algoritmos de propósito general y múltiples capas no lineales, descubrir automáticamente las representaciones necesarias para tareas como la detección o clasificación. Los métodos basados en el aprendizaje profundo (DL) incluyen múltiples niveles de representación que se obtienen mediante la composición de módulos simples, pero no lineales. Estos módulos transforman gradualmente una representación desde un nivel inicial, que comienza con los datos sin procesar, hacia niveles superiores más abstractos. Al combinar suficientes de estas transformaciones, es posible aprender funciones altamente complejas.

En el contexto de tareas de clasificación, las capas superiores de representación realzan características importantes de la entrada mientras reducen las variaciones irrelevantes. Un aspecto crucial del DL es que estas capas no son diseñadas por ingenieros humanos, sino que se generan a partir de los datos mediante un proceso de aprendizaje generalizado (Molina & García, 2021).

Árbol de decisiones

El aprendizaje del árbol de decisiones sigue una estrategia de "divide y vencerás", utilizando un método de búsqueda codicioso para encontrar puntos de división óptimos en el árbol. Esta división se repite recursivamente de arriba a abajo hasta que la mayoría de los registros se clasifican en categorías específicas. Que todos los datos se agrupen en categorías homogéneas depende de la complejidad del árbol. Los árboles más pequeños tienden a acercarse a nodos de hojas puras, donde todos los datos pertenecen a una sola capa. Sin embargo, a medida que el árbol crece, mantener esta pureza se vuelve más difícil, lo que a menudo conduce a la fragmentación de los datos y un posible sobreajuste. Por lo tanto, los árboles de decisión tienden a preferir tamaños más pequeños, de acuerdo con el principio de análisis sintáctico de Occam, que establece por lo general, la simplicidad suele ser la mejor

opción en términos de explicación. Para mitigar la complejidad y evitar problemas de sobreajuste, se recurre a la poda, que es un procedimiento que consiste en eliminar ramificaciones que no aportan información relevante (IBM, IBM, 2023).

CAPÍTULO III: DISEÑO METODOLÓGICO

3.1. Tipo y diseño de la investigación

3.1.1. Tipo y Enfoque

La investigación de tipo aplicada, pretende solucionar problemas prácticos y específicos, mismos que suelen afectar a los usuarios que están implicados en dichos problemas (Hernández y otros, 2014).

Una investigación con enfoque cuantitativo, pretende medir unidades de magnitud, mediante un esquema deductivo y lógico para explicar los resultados obtenidos a partir de la investigación (Hernández y otros, 2014).

3.1.2. Diseño

Este estudio ha sido planificado con un diseño experimental: Cuasi experimental, (Puede ser no experimental, en caso no se cuente con el algoritmo) entendiéndose por ello que, aunque existe manipulación deliberada de variables, esta se realiza de forma parcial, es decir se manipula una o más variables independientes se manipulan para observar su influencia en las variables dependientes (Hernández y otros, 2014).

Además, en este tipo de investigación, los sujetos no son conformados o agrupados al azar, es decir dichas agrupaciones ya estuvieron formadas antes del experimento, es por eso que por consiguiente estas pertenecen a conjuntos intactos (la agrupación es independiente o desligada del experimento). (Hernández y otros, 2014).

Esta investigación tiene un alcance explicativo, entendiéndose por ello como la búsqueda de respuestas de una causante de eventos o acontecimientos sociales o físicos. Además, se enfoca en interpretar y justificar la razón por la que sucede un acontecimiento, las condiciones en la que se presenta y por cual razón se vinculan dos o más variables. (Hernández y otros, 2014)

3.2. Operacionalización de variables

3.2.1. Variable independiente

Método predictivo basado en minería de datos

3.2.2. Variable dependiente

Patrones de tráfico vehicular

Tabla 2.

Operacionalización de variables de investigación

Variable	Definición conceptual	Definición operacional	Dimensión	Indicadores	Escala
Modelado predictivo basado en minería de datos	Son modelos, técnicas y algoritmos que permiten predecir en base a una cantidad de datos, un factor o patrón de estudio.	Se realizará un modelado predictivo de patrones de tráficos, el cual utilizará árbol de decisiones y aprendizaje supervisado	Comprensión del negocio	Campo Tipo Descripción	De razón
			Comprensión de datos	Registros Atributos	Numérica Descriptiva
			Preparación de datos	Selección de datos	De razón
			Modelado	Variables categóricas Phyton	Concesión Entidad prestadora unidad de Peaje Programación
Patrones de tráfico	Características recurrentes en relación con los índices de trafico	Son extraído de OpenData - OSITRAN, y comparados con los resultados del modelado para la predicción de peajes	Evaluación	Porcentaje de aceptación	%
			Despliegue	Validez	%

Nota: Generación de contenido de autoría propia

3.3. Población y muestra

Población

Registro de indicadores mensuales de tráfico Open Data de OSITRAN 2022.

Muestra

La investigación se enfocará en el peaje del sistema de transporte eléctrico masivo de Lima y Callao, específicamente en la línea 1, Villa El Salvador, durante el período comprendido entre 2019 - 2021.

3.4. Técnicas e instrumentos de recolección de datos

3.4.1. Técnicas

Uso de Metodología para diseño de modelos predictivos

Se emplean las metodologías correspondientes para llevar a cabo la investigación. En este caso específico, se utiliza la metodología Crisp-DM para aplicar técnicas de minería de datos y construir los modelos necesarios.

3.4.2. Instrumentos

En el presente proyecto de investigación, se llevará a cabo la implementación del modelo de árbol de decisiones regresivo utilizando la plataforma de Google Colab. Para ello, se requiere la conexión estable a Internet y una cuenta de Google para acceder a la plataforma.

Además, se necesitará la instalación de las bibliotecas necesarias de Python, como Pandas, NumPy, Scikit-learn y Matplotlib, que se pueden instalar mediante los comandos "pip install" en la consola de Colab.

3.4.3. Equipos y herramientas

En el análisis de la información, se requerirá una laptop con características regulares como mínimo 4 GB de RAM, procesador de al menos 2.5 GHz y almacenamiento suficiente para obtención de la base de datos y otros archivos necesarios. Además, se utilizará Google Colab como herramienta de trabajo para la implementación del modelo de árbol de decisiones regresivo, lo que permitirá utilizar recursos de procesamiento adicionales en la nube y evitar la carga excesiva del equipo local

3.4.4. Materiales

En cuanto a los materiales necesarios, se utilizará la base de datos proporcionada por Open Data OSITRAN, que contiene información sobre las tarifas y los pagos de los peajes en las carreteras de la red supervisada por esta entidad. El conjunto de datos está disponible para su descarga en formato CSV desde el sitio web de Open Data OSITRAN y cargar en la plataforma de Colab para su procesamiento. Asimismo, se utilizará el código del modelo que se implementará en Python para realizar el análisis y la evaluación de los datos.

CAPÍTULO IV: METODOLOGÍA DESARROLLADA

Se utilizará la metodología CRISP-DM.

4.1. Obtener y almacenar datos históricos detallados de tráfico en las unidades de peaje peruanos.

Paso 1.Comprensión del negocio

El Organismo **Supervisor** de la Inversión en Infraestructura del Transporte Público (OSITRAN) es el ente encargado de monitorear y regular las actividades relacionadas con la infraestructura del transporte público en el Perú. Su objetivo fundamental es garantizar la operación eficiente, segura y sostenible de las concesiones de infraestructuras de transporte como carreteras, puertos, aeropuertos y ferrocarriles.

Los objetivos del OSITRAN se centran en varios aspectos clave:

Promoción de inversiones:

OSITRAN busca incentivar la inversión privada en infraestructura de transporte, contribuyendo al desarrollo económico del país mejorando la conectividad y movilidad de personas y bienes.

Protección de los derechos de los usuarios:

La unidad protege los derechos y las preocupaciones de los usuarios del servicio de transporte, garantizando calidad, seguridad y precios razonables.

Garantía competitiva:

OSITRAN promueve la competencia en el sector de infraestructura de transporte, evita prácticas anticompetitivas y garantiza que las concesiones se otorguen de manera transparente. Garantizar la sostenibilidad:

La unidad se preocupa para garantizar la sostenibilidad ambiental y social de los proyectos de infraestructura, es crucial cumplir con las regulaciones y reducir al mínimo los efectos adversos en el entorno.

Monitorear el cumplimiento del contrato:

OSITRAN verifica que las empresas franquiciadoras cumplan con los términos del contrato de franquicia, incluyendo la realización de inversiones y mantenimiento adecuados en infraestructura (OSITRAN, 2023).

El propósito del modelo propuesto está estrechamente vinculado con los objetivos básicos del proyecto. Este modelo está diseñado para analizar y obtener una comprensión más un análisis exhaustivo de los datos presentes en el conjunto de datos, con el objetivo de extraer conocimientos valiosos y patrones relevantes que respalden la toma de decisiones. Además, busca proporcionar pronósticos y recomendaciones que puedan ser útiles para lograr los objetivos generales del proyecto.

La viabilidad de la contribución de este modelo depende de la calidad y cantidad de datos existentes en el conjunto de datos. Con la gran cantidad de información disponible, el modelo tiene el potencial de producir resultados significativos que contribuyan al éxito del proyecto. Asimismo, el diseño y la implementación adecuados del modelo, así como la interpretación precisa de los resultados, son factores clave para garantizar la viabilidad y utilidad del modelo en el contexto de un proyecto.

En esta primera fase, el objetivo principal es analizar los objetivos y necesidades del proyecto, definir el contexto y elaborar un plan. Esto incluye evaluar la situación actual, leer el estudio, evaluar los costos y beneficios, encontrar planes de negocios que guíen el estudio y crear un plan de proyecto.

Antes de establecer los objetivos, es crucial comprender el contexto o la problemática que motiva la investigación. El tema central se refiere al crecimiento del tráfico vehicular desde 2019 hasta 2022 y su continuo aumento en el país. Por tanto, resulta fundamental evaluar el incremento en los años subsiguientes, con el propósito principal de adquirir un conocimiento que pueda aplicarse en diversas áreas de utilidad, como:

- Planificación de recursos:

Los operadores y administradores de peaje pueden utilizar los pronósticos del recuento de vehículos de las estaciones de peaje para asignar de manera óptima los recursos, incluidos el personal y el equipo. Por lo tanto, gestionan los flujos de tráfico esperados y ajustan sus operaciones para garantizar una experiencia fluida y eficiente para los conductores.

- Gestión del tráfico:

Predecir el número de vehículos en las plazas de peaje permite a los gestores de tráfico predecir posibles atascos y atascos cerca de las casetas de peaje. Esto les brinda la oportunidad de implementar medidas de gestión del tráfico, como ajustar las señales, abrir

carriles adicionales o crear desvíos para reducir los impactos negativos en el tráfico y garantizar un flujo de tráfico más fluido.

- Optimización de la eficiencia de los peajes:

Al predecir el número de vehículos en los centros de peaje, los operadores pueden optimizar los sistemas y la infraestructura de cobro de peaje para agilizar el proceso de cobro y disminuir los períodos de espera en los puestos de peaje. Esto puede implicar la adopción de tecnologías de cobro de peaje electrónico, como el cobro de peaje electrónico o el cobro de peaje automatizado, para aumentar la velocidad de desplazamiento de los vehículos y reducir la congestión.

- Estimación de ingresos:

La previsión del número de vehículos en los centros de peaje también influye en la estimación de los ingresos generados por las tarifas de peaje. El operador puede calcular con precisión los ingresos esperados si conoce la cantidad de vehículos que se espera que pasen por la cabina de peaje. Esto les permite planificar y gestionar los presupuestos asignados para mantener y mejorar la infraestructura vial de manera más efectiva.

Ahora bien, planteado la importancia de la realización del proyecto, es importante considerar los siguientes propósitos:

- ✓ Predecir el total de vehículos pagantes según las unidades de peaje peruanos
- ✓ Predecir el total de vehículos exonerados según las unidades de peaje peruanos
- ✓ El obtener estos objetivos, se podrá obtener una serie de resultados a lo necesitado para la investigación, lo cual es el saber a través de la minería de datos cuál será el número de vehículos según el tiempo establecido y las unidades de tarifas de peaje extraídas de la base de datos.

- Evaluación de la situación

Disponemos de una base de datos proporcionada por la entidad OSITRAN, cuya principal función es la supervisión y regulación de la infraestructura de transporte de uso público. Contar con los datos de esta entidad nos brinda ventajas en los aspectos siguientes:

- Protección de los derechos de los usuarios:

- ✓ La entidad reguladora OSITRAN Su misión se centra en proteger los derechos de los pasajeros que utilizan el transporte público en el territorio peruano.
- ✓ Esta información tiene la capacidad de anticipar las exigencias y requerimientos de los usuarios en relación a la infraestructura y prestaciones de transporte, facilitando

una planificación más efectiva y un incremento en la excelencia del servicio en los puntos de peaje.

- Mejora de la calidad y eficiencia de los servicios:

✓ OSITRAN trabaja para garantizar que los servicios de transporte público cumplan con estándares de calidad apropiados.

✓ Al analizar esta información, se pueden predecir posibles mejoras en la infraestructura de transporte y evaluar su impacto en la eficiencia del tráfico en las unidades de peaje, lo que se traduce en un flujo más rápido y eficiente de los vehículos.

- Arbitraje de conflictos:

✓ OSITRAN ofrece un mecanismo de arbitraje para resolver disputas o conflictos entre los usuarios y los operadores de transporte.

✓ Al anticipar posibles conflictos y utilizar esta información, se pueden diseñar estrategias que permitan resolverlos de manera eficiente, evitando retrasos innecesarios en los peajes.

- Inventario por utilizar: El inventario con el que se cuenta es con una laptop y la base de datos de la empresa reguladora.
- Costes y beneficios: como coste se tomará el presupuesto ya establecido por el investigador y será a cuenta propia. Como beneficio solo he de mencionar que no contaría con ningún aumento del recurso financiero del investigador, sino que ayudará a modo de contribución a próximas investigaciones que vayan por la misma rama.
- Determinación de los objetivos del negocio:

✓ Realizar el análisis de predicción sobre los patrones de tráfico en unidades de peaje peruanos.

✓ Recopilar datos precisos sobre la cantidad de vehículos que realizan pagos y están exentos de pago, con el fin de contribuir a la toma de decisiones relacionadas con los comportamientos de tráfico en los peajes de Perú.

Cabe resaltar que El nivel de precisión se considera como un indicador de éxito empresarial y este debe ser mayor al 80%.

Paso 2.Compresión de datos

- Descripción de los datos

La data se obtuvo por medio de la Open Data de OSITRAN, cuenta con un total de 40177 registros, 10 atributos y no se encuentra normalizada, tal como se aprecia en la tabla.

Tabla 3.

Descripción de los datos

Campo	Tipo	Descripción
Año	Int	Unidad de tiempo.
Mes	Int	Unidad de tiempo.
Concesión	Categórico	Entidad encargada de la prestación de un servicio de carreteras.
Entidad prestadora	Categórico	Nombre de la concesión.
Unidad de peaje	Categórico	Lugar donde se realiza el pago de peaje.
Sentido de cobro en unidad de peaje	Categórico	Ascendente/Descendente/Ambos
Total vehículos pagantes	Int	Número de vehículos que pagaron el peaje.
Total vehículos exonerados	Int	Número de vehículos que no pagaron el peaje por tener los siguientes vehículos: motos, bicicletas, bomberos, ambulancias, fuerzas militares, policía, INPEC, Defensa Civil.
Total vehículos evadidos	Int	Número de vehículos los cuales evadieron el peaje.
Total de vehículos exonerados eventos extraordinarios	Int	Número de vehículos que se encuentran exonerados por temas de revisión técnica.

Nota: Generación de contenido de autoría propia

- Exploración y verificación de la calidad de los datos

Como se aprecia en la figura, el 100% de los registros de los atributos no se encuentran nulos.

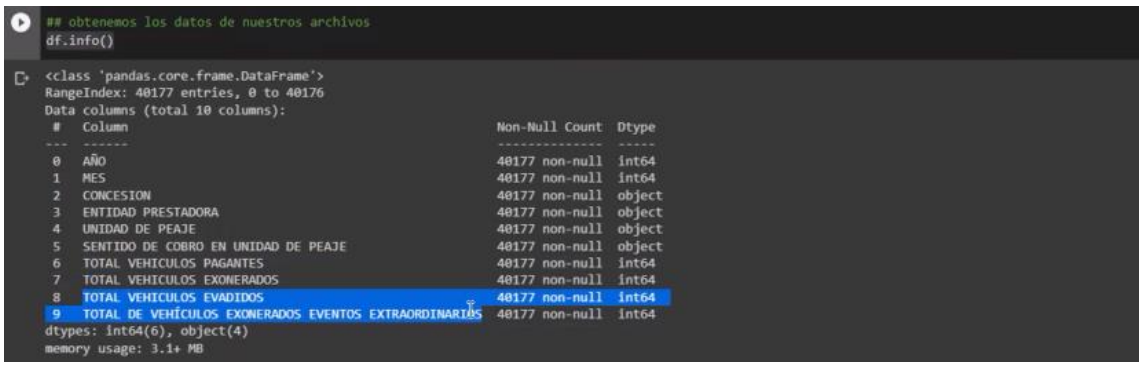


Figura 4. *Dataframe donde se leen los atributos*

Nota: Generación de contenido de autoría propia

No obstante, por parte de los atributos categóricos (concesión, entidad prestadora y unidad de peaje), es decir, los que servirán como variables de entrada para entrenar al algoritmo, ya que son importantes para la investigación, deben ser normalizados y separarse en tablas para obtener un fácil análisis de data para el árbol de decisiones regresivo.

A	B	C	D	E	F	G	H	I	J
AÑO	MES	CONCESION	ENTIDAD PRESTADORA	UNIDAD DE PEAJE	SENTIDO DE COBRO EN UNI	TOTAL VEHICULOS PAGANTE	TOTAL VEHICULOS EXONERAC	TOTAL VEHICULOS EVADIDOS	TOTAL DE VEHÍCULOS EXONERADOS EVENTOS EXTRAORDINARIOS
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP BAYOVAR	ASCENDENTE	677	9	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP CHICAMA	ASCENDENTE	3905	24	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP CHICAMA	DESCENDENTE	4795	26	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP MÓRROPE	ASCENDENTE	899	5	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP PACANGUILLA	ASCENDENTE	1882	11	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP PACANGUILLA	DESCENDENTE	2019	17	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP SULLANA	ASCENDENTE	2114	20	0	0
2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP SULLANA	DESCENDENTE	2948	24	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	CCASACANCHA	ASCENDENTE	827	8	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	CCASACANCHA	DESCENDENTE	814	9	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	PAMPA GALERAS	DESCENDENTE	263	11	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	PAMPAMARCA	ASCENDENTE	144	2	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	PAMPAMARCA	DESCENDENTE	216	3	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	PICHIRHUA	ASCENDENTE	381	13	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	PICHIRHUA	DESCENDENTE	458	14	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	SAN JUAN DE MARCONA	ASCENDENTE	668	2	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 1: IVA	SURVIAL S.A.	SAN JUAN DE MARCONA	DESCENDENTE	737	1	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 2 : UR	CONCESIONARIA INTEROCEANICA SUR - TRAMO 2 S.A.	QUINCEMIL	AMBOS SENTIDOS	452	0	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 2 : UR	CONCESIONARIA INTEROCEANICA SUR - TRAMO 2 S.A.	QUINCEMIL	ASCENDENTE	222	0	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 2 : UR	CONCESIONARIA INTEROCEANICA SUR - TRAMO 2 S.A.	QUINCEMIL	DESCENDENTE	230	0	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	ALERTA (PLANCHÓN)	AMBOS SENTIDOS	659	17	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	ALERTA (PLANCHÓN)	ASCENDENTE	357	8	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	ALERTA (PLANCHÓN)	DESCENDENTE	302	9	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	IÑAPARI (SAN LORENZO)	AMBOS SENTIDOS	227	10	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	IÑAPARI (SAN LORENZO)	ASCENDENTE	121	5	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	IÑAPARI (SAN LORENZO)	DESCENDENTE	106	5	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	UNIÓN PROGRESO	AMBOS SENTIDOS	1023	14	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	UNIÓN PROGRESO	ASCENDENTE	489	7	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 3: INA	CONCESIONARIA INTEROCEANICA SUR - TRAMO 3 S.A.	UNIÓN PROGRESO	DESCENDENTE	534	7	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	MACUSANI	ASCENDENTE	141	5	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	MACUSANI	DESCENDENTE	136	6	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	SAN ANTÓN	ASCENDENTE	492	13	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	SAN ANTÓN	DESCENDENTE	432	14	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	SAN GABÁN	ASCENDENTE	194	35	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 4: AZÁ	INTERSUR CONCESIONES S.A.	SAN GABÁN	DESCENDENTE	186	36	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 5: ILO	CONCESIONARIA VIAL DEL SUR S.A.	CARACOTO - ILLPA	ASCENDENTE	2102	42	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 5: ILO	CONCESIONARIA VIAL DEL SUR S.A.	CARACOTO - ILLPA	DESCENDENTE	2997	45	2	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 5: ILO	CONCESIONARIA VIAL DEL SUR S.A.	ILO	ASCENDENTE	1810	2	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 5: ILO	CONCESIONARIA VIAL DEL SUR S.A.	ILO	DESCENDENTE	1733	1	0	0
2019	1	CORREDOR VIAL INTEROCEÁNICO SUR PERÚ -BRASIL, TRAMO 5: ILO	CONCESIONARIA VIAL DEL SUR S.A.	MATARANI	ASCENDENTE	2129	0	0	0

Figura 5. *Atributos antes de la normalización*

Nota: Generación de contenido de autoría propia

En investigación realizada no se han creado las variables objetivas que servirán como predicción para los patrones de tráfico en unidades de peaje peruanos.

También, se han hecho gráficos estadísticos con los datos obtenidos.

En la figura 6, se puede apreciar el total por años sobre los vehículos exonerados, exonerados por eventos extraordinarios, evadidos y pagantes. En los años 2019 a 2022

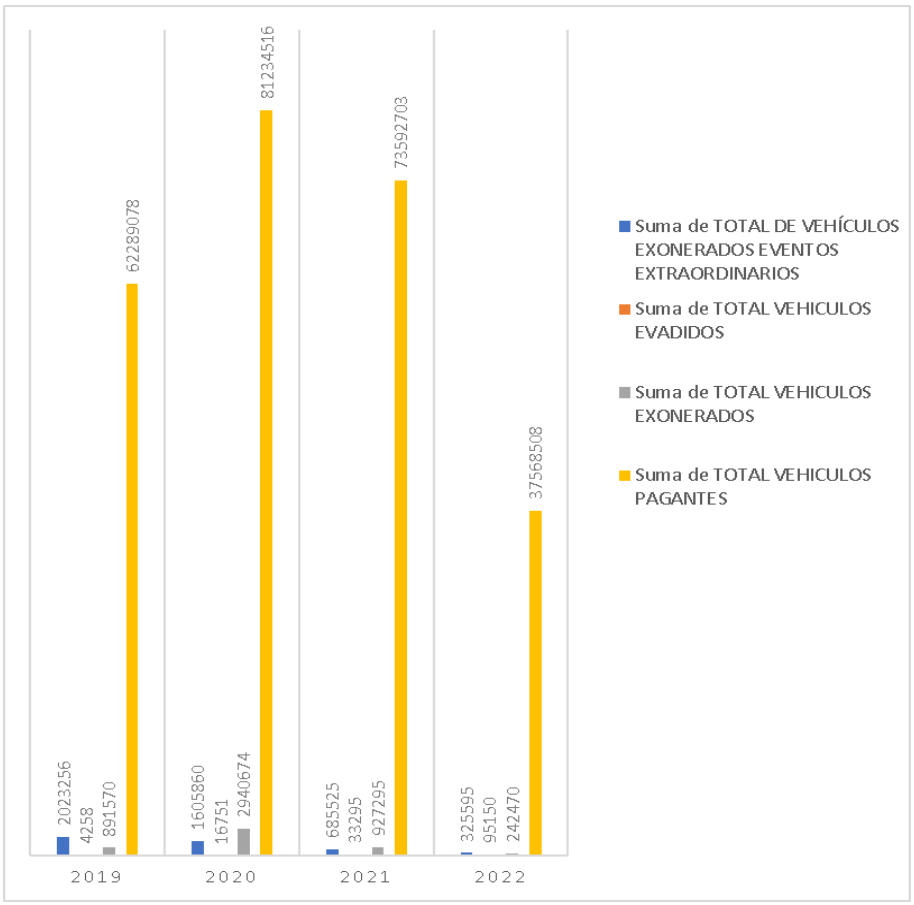


Figura 6. *Total, según la clasificación de los vehículos*

Nota: Generación de contenido de autoría propia

El gráfico 7, se muestra el total de veces que los vehículos han transitado en los concesionarios en el año 2019.

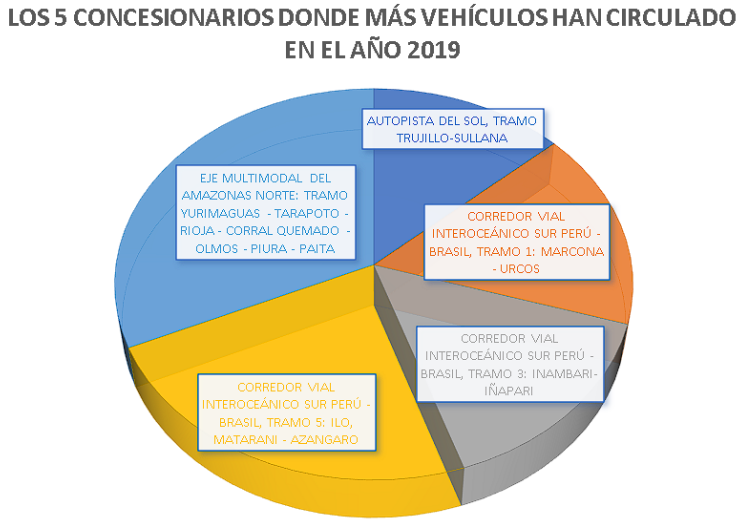


Figura 7. *Top 5 concesionarios de más circulación en el año 2022*

Nota: Generación de contenido de autoría propia

El gráfico 8, se muestra el total de veces que los vehículos han transitado en los concesionarios en el año 2020.

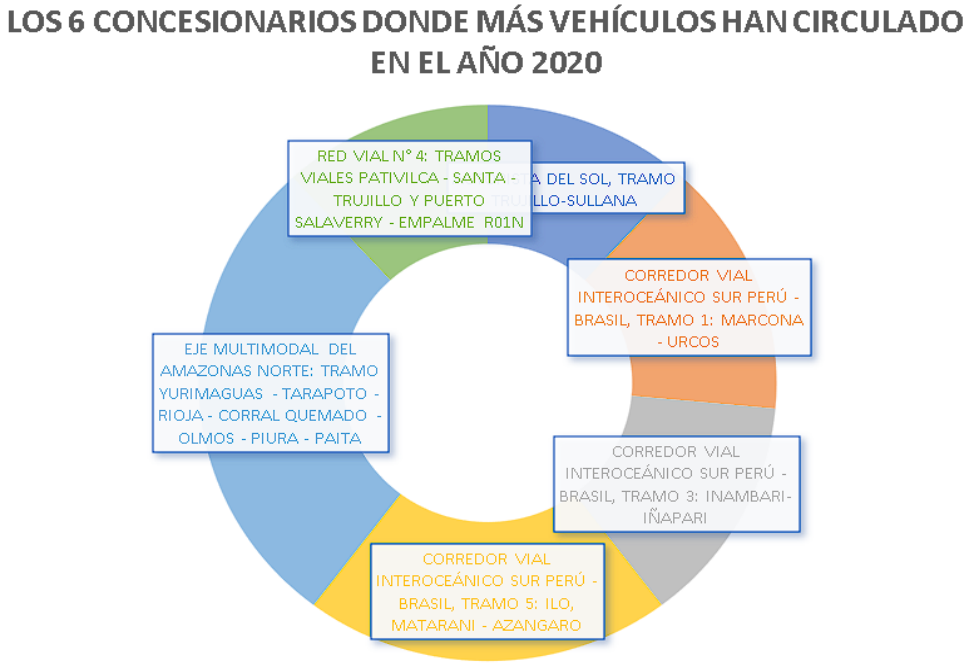


Figura 8. *Top 6 concesionarios de más circulación en el año 2022*

Nota: Generación de contenido de autoría propia

El gráfico 9, se muestra el total de veces que los vehículos han transitado en los concesionarios en el año 2021.

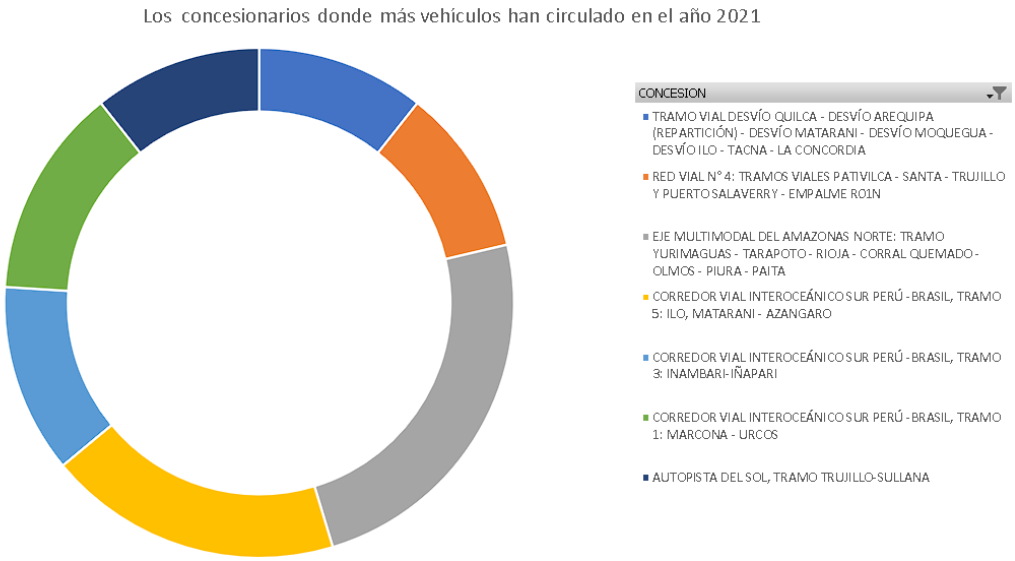


Figura 9. *Top 7 concesionarios de más circulación en el año 2022*

Nota: Generación de contenido de autoría propia

El gráfico 10, se muestra el total de veces que los vehículos han transitado en los concesionarios en el año 2021.

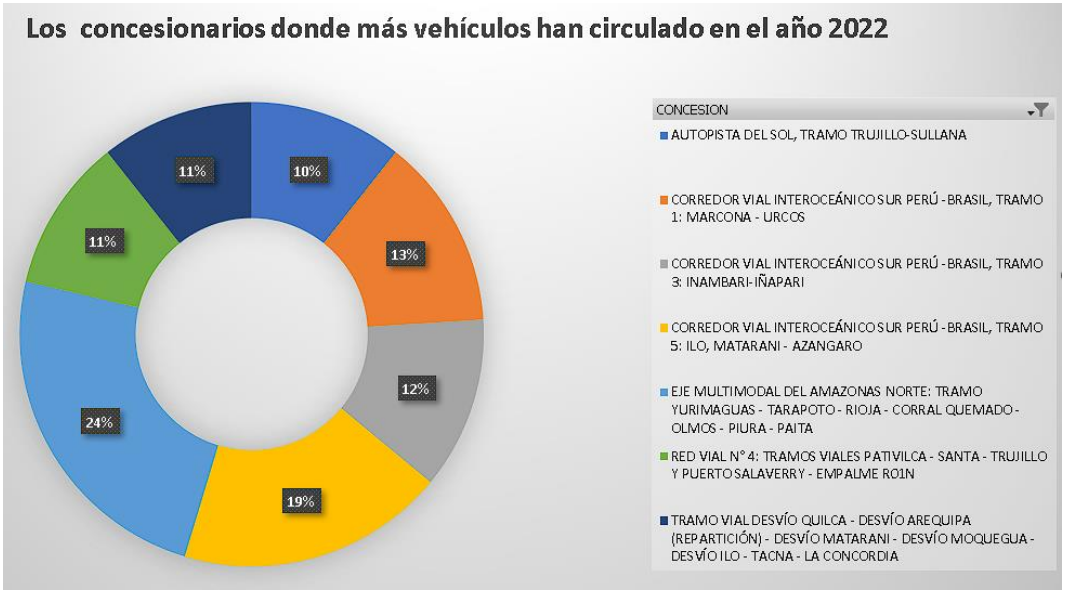


Figura 10. *Top 7 concesionarios de más circulación en el año 2022*

Nota: Generación de contenido de autoría propia.

El gráfico 11, se verifica el sentido de cobro por unidades entre los años 2019 a 2022.

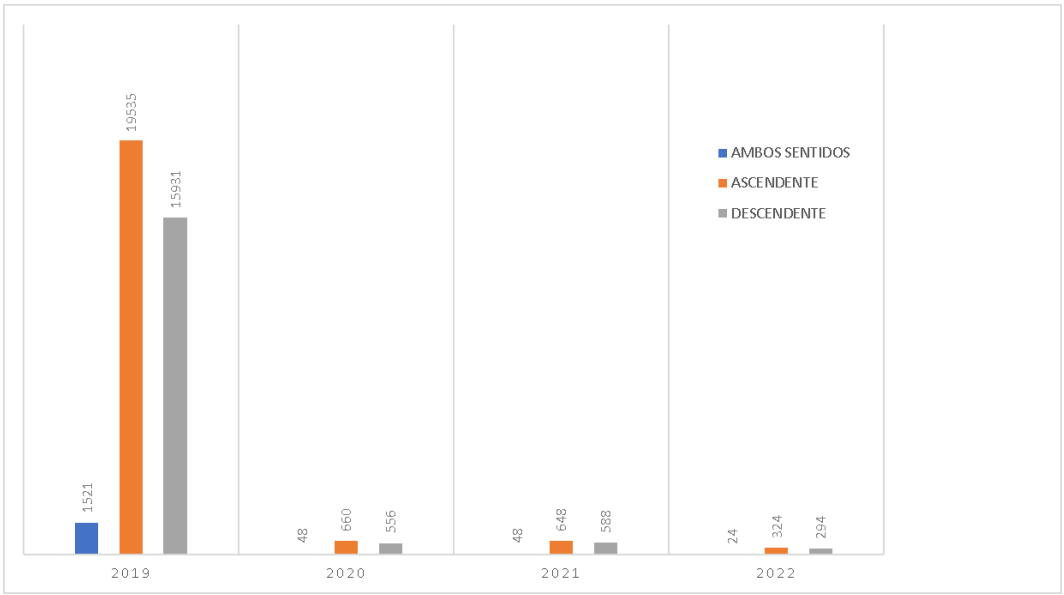


Figura 11. *Cantidad de sentido de cobro por unidad*

Nota: Generación de contenido de autoría propia

En el gráfico 12. muestra las unidades de peaje donde a los que más vehículos se han acercado en él.

Figura 12. *Unidades de peaje más visitadas en el 2022*

Nota: Generación de contenido de autoría propia

El gráfico 13, se muestra las unidades de peaje donde a los que más vehículos se han acercado en el año 2020.

Figura 13. *Unidades de peaje más visitadas en el 2022*

Nota: Generación de contenido de autoría propia

El gráfico 14, se muestra las unidades de peaje donde a los que más vehículos se han acercado en el año 2021.

Figura 14. *Unidades de peaje más visitadas en el 2022*

Nota: Generación de contenido de autoría propia

El gráfico 15, se muestra las unidades de peaje donde a los que más vehículos se han acercado en el año 2022.

Figura 15. *Unidades de peaje más visitadas en el 2022*

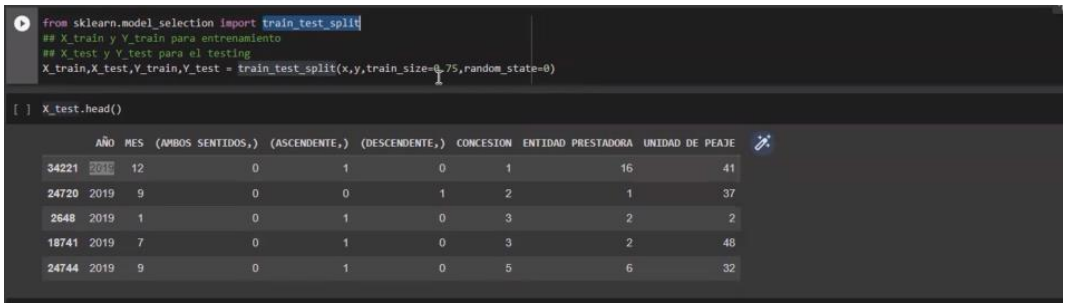
Nota: Generación de contenido de autoría propia

5. Identificar las variables más influyentes en los patrones de tráfico, prepararlos para su uso en el modelo.

Paso 3.Preparación de los datos.

- Selección de datos

De acuerdo a lo explicado en los apartados previos, es necesaria una normalización de la tabla para poder facilitar y tener una mejor carga de data a la hora de ser analizada por el algoritmo en cuestión, de manera que a través del uso de librerías de python, se obtuvo un cambio de nombre a los atributos, así como la respectiva normalización de la tabla.



```
from sklearn.model_selection import train_test_split
## X_train y Y_train para entrenamiento
## X_test y Y_test para el testing
X_train,X_test,Y_train,Y_test = train_test_split(x,y,train_size=0.75,random_state=0)
```

```
[ ] X_test.head()
```

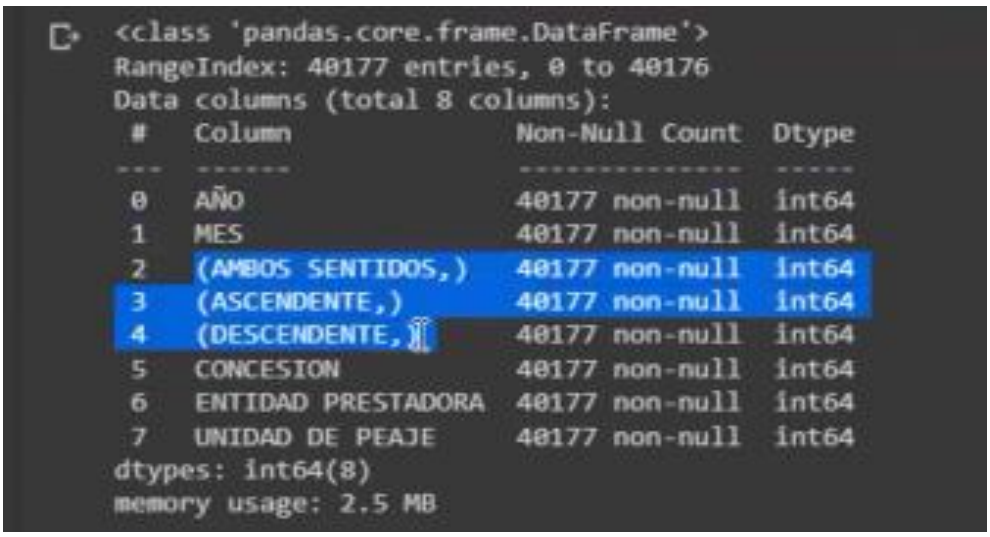
	AÑO	MES	(AMBOS SENTIDOS,)	(ASCENDENTE,)	(DESCENDENTE,)	CONCESION	ENTIDAD PRESTADORA	UNIDAD DE PEAJE
34221	2019	12	0	1	0	1	16	41
24720	2019	9	0	0	1	2	1	37
2648	2019	1	0	1	0	3	2	2
18741	2019	7	0	1	0	3	2	48
24744	2019	9	0	1	0	5	6	32

Figura 16. *Dataframe donde se muestra de manera normalizada*

Nota: Generación de contenido de autoría propia

A veces, cuando se trabaja con conjuntos de datos, es necesario seleccionar un subconjunto de datos porque es muy relevante para el análisis o la tarea en cuestión. Esto se hace para centrar la atención en los datos que proporcionan información importante y significativa para los objetivos del proyecto, evitando la inclusión de datos redundantes o menos importantes. Al tomar solo los 7 datos más relevantes de la tabla 10, se simplifica el análisis, se reduce el ruido en los resultados y los principales aspectos de interés pueden comprenderse de manera más precisa y efectiva. Esta selección estratégica de datos altamente la consideración de elementos significativos puede acelerar la toma de decisiones y ayudar a identificar patrones y tendencias significativos de manera más efectiva.

Se clasifican 3 variables (franquicia, proveedor y cobrador de peaje), gracias a la función LabelEncode, los atributos relacionados se recuperan convirtiendo los datos categóricos (las variables mencionadas se convierten a número).



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40177 entries, 0 to 40176
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   AÑO                 40177 non-null  int64
 1   MES                 40177 non-null  int64
 2   (AMBOS SENTIDOS,)   40177 non-null  int64
 3   (ASCENDENTE,)       40177 non-null  int64
 4   (DESCENDENTE,       40177 non-null  int64
 5   CONCESION           40177 non-null  int64
 6   ENTIDAD PRESTADORA  40177 non-null  int64
 7   UNIDAD DE PEAJE     40177 non-null  int64
dtypes: int64(8)
memory usage: 2.5 MB
```

Figura 17. *Dataframe de los atributos seleccionados*

Nota: Generación de contenido de autoría propia.

5.1. Diseñar y construir modelos de minería de datos, en árboles de decisión, y entrenarlos utilizando datos históricos para predecir patrones de tráfico futuros.

Paso 4.Modelado.

- Preparación de los datos:

Anteriormente, se hizo ya una selección de atributos y se verificó que no haya datos nulos dentro de la tabla de datos y la normalización de esta, ahora es necesario definir las variables categóricas usadas para la predicción de los datos.

Las variables categóricas elegidas son las siguientes:

- Concesión.
- Entidad prestadora.
- Unidad de peaje.

Por el hecho que estos 3 atributos son donde más relación tiene con los vehículos, son más que necesarios para la presente investigación.

Durante el proceso de entrenamiento del sistema, se generaron dos variables clave: "x" y "y". Estas variables representan la entrada y la variable a predecir, respectivamente. Esta separación permite dividir la tabla principal en dos grupos: el 75% de los datos se destina al entrenamiento del algoritmo, actuando como la variable de ingreso, mientras que el 25% restante se utiliza para las pruebas posteriores una vez que el algoritmo ha sido entrenado.

Para garantizar la efectividad del algoritmo, se dedicaron tres meses a la construcción y configuración del mismo. Durante este período de preparación, se llevaron a cabo diversas actividades, como el análisis de los datos, la selección y ajuste de parámetros, la elección de algoritmos adecuados y la validación del modelo resultante.

Esta fase de preparación es fundamental para lograr un algoritmo preciso y confiable. Se realizan pruebas exhaustivas, se realizan ajustes y se optimizan los parámetros para obtener resultados más precisos en la etapa de predicción. Además, se implementan técnicas de validación cruzada y se realizan análisis de rendimiento con el fin de medir la calidad del modelo y asegurar su confiabilidad.

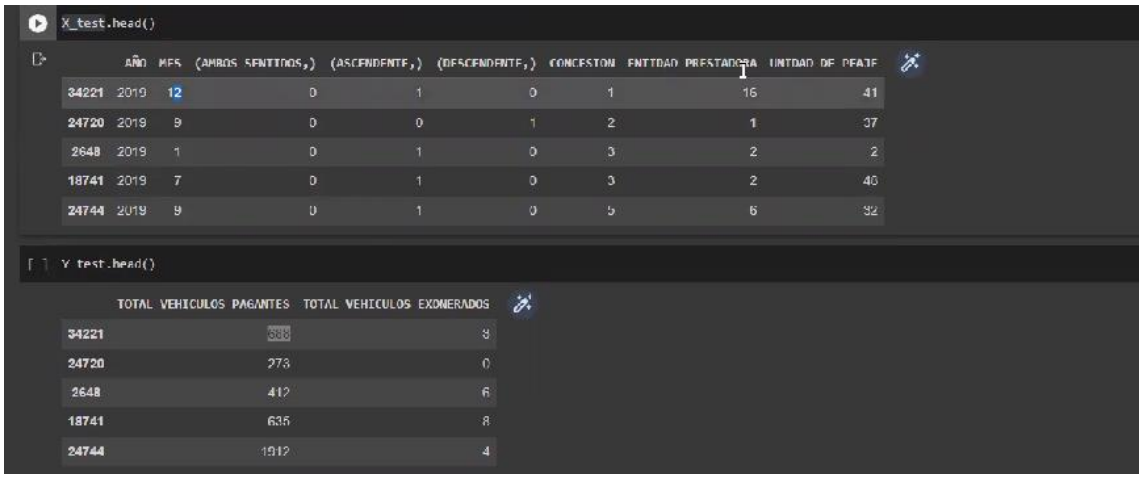


Figura 18. *Dataframes de la variable de entrada y a predecir*

Nota: Generación de contenido de autoría propia

A continuación, se explicará el paso a paso utilizado para la construcción del algoritmo usando la técnica de árbol de decisiones regresivo:

Paso 1: En el inicio de un proyecto de programación en Python, es común importar una serie de librerías para aprovechar las funcionalidades que ofrecen. En este contexto, tres librerías fundamentales son io, pandas y Google Colab, cada una con su propio propósito y beneficio.

En conjunto, al importar las librerías io, pandas y Google Colab, se obtiene un conjunto de herramientas poderosas para el manejo de recursos del hardware, el análisis y manipulación de datos, y el desarrollo colaborativo en la nube. Esta combinación de librerías permite trabajar de manera eficiente y efectiva en proyectos de programación, proporcionando las capacidades necesarias para abordar desafíos complejos y obtener resultados precisos y significativos en el análisis de datos.

Figura 19. *Librerías a utilizar*

Nota: Generación de contenido de autoría propia

Paso 2: Una de las funciones clave para trabajar con archivos de datos, en particular aquellos en formato Excel, es la función de carga que permite importar el archivo con los datos en cuestión. Para realizar esta tarea, se utilizó el comando upload () de Python, el cual proporciona una manera sencilla y eficiente de cargar archivos desde el hardware local al entorno de programación. Al ejecutar el comando upload (), se habilita un cuadro de diálogo

que permite seleccionar y cargar el archivo deseado. Esta funcionalidad resulta especialmente útil cuando se requiere acceder a conjuntos de datos almacenados localmente y se desea utilizarlos en el contexto del proyecto en desarrollo. Una vez que el archivo ha sido cargado exitosamente, se obtiene acceso a los datos contenidos en él, lo que permite su posterior procesamiento y análisis utilizando las capacidades de las librerías y herramientas disponibles. La función de carga de archivos a través del comando upload () de Python, por lo tanto, representa una poderosa funcionalidad para importar datos de manera rápida y conveniente, facilitando así el flujo de trabajo y la gestión de datos en la ejecución de proyectos de análisis de datos y programación en términos generales.

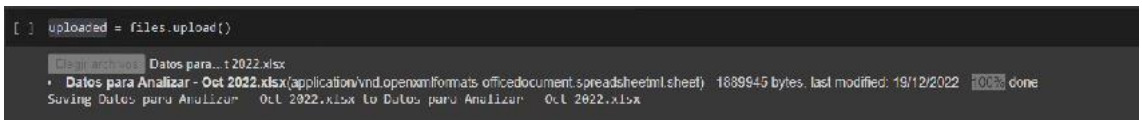


Figura 20. *Uso del comando upload*

Nota: Generación de contenido de autoría propia

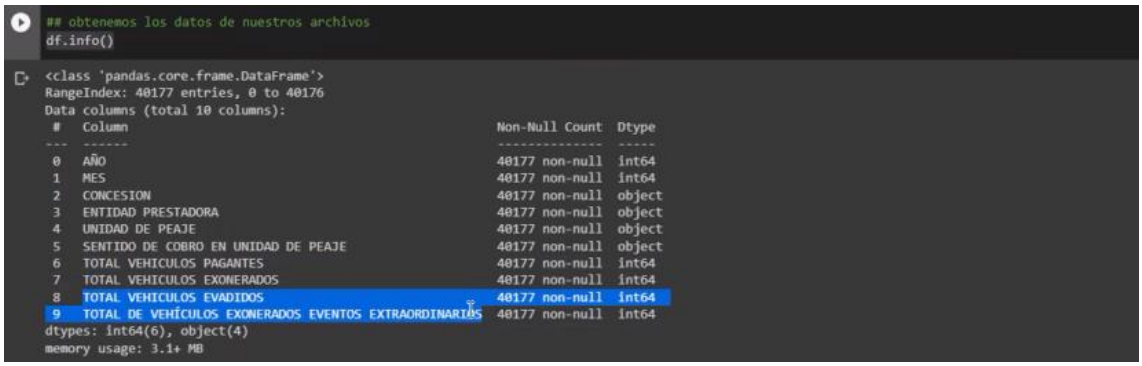
Paso 3: Una vez que el archivo ha sido cargado utilizando el comando upload () de Python, se procede a la creación de una variable denominada df, la cual hace referencia a un DataFrame. Este DataFrame se utiliza para leer y almacenar los datos cargados anteriormente, lo que permite un fácil acceso y manipulación de la información contenida en el archivo. Al asignar los datos a la variable df, se crea una estructura tabular organizada en filas y columnas que refleja la estructura del conjunto de datos original. Esta estructura tabular ofrece una serie de funcionalidades y métodos propios de los DataFrames, como la capacidad de filtrar y transformar los datos, realizar cálculos estadísticos, visualizar el contenido de las columnas y realizar operaciones de manipulación, como la modificación de columnas o la creación de nuevas variables derivadas. Al utilizar el DataFrame df, se simplifica la posibilidad de acceder y manejar los datos que se han cargado, lo que posibilita realizar análisis y procesamientos posteriores de manera más eficaz y organizada.

Figura 21 . *Dataframes de la variable de entrada y a predecir*

Nota: Generación de contenido de autoría propia

Paso 4: Al cargar la data, se pueden observar diversos atributos que proporcionan información detallada sobre los registros. Estos atributos incluyen el año (int), mes (int), concesión (categórico), entidad prestadora (categórico), unidad de peaje (categórico),

sentido de cobro en unidad de peaje (categórico), total de vehículos pagantes (int), total de vehículos exonerados (int), total de vehículos evadidos (int) y total de vehículos exonerados por eventos extraordinarios (int). El año y mes indican la temporalidad de los registros, mientras que la concesión y entidad prestadora se refieren a las entidades responsables del servicio de peaje. La unidad de peaje identifica ubicaciones específicas, y el sentido de cobro indica la dirección del tráfico. Los totales de vehículos pagantes, exonerados, evadidos y exonerados por eventos extraordinarios reflejan la cantidad de vehículos en cada categoría. Estos atributos proporcionan información relevante para analizar patrones, tendencias y cumplimiento de pagos en relación con la infraestructura de transporte y el flujo de vehículos.



```
## obtenemos los datos de nuestros archivos
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40177 entries, 0 to 40176
Data columns (total 10 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   AÑO                                     40177 non-null  int64
 1   MES                                     40177 non-null  int64
 2   CONCESION                               40177 non-null  object
 3   ENTIDAD PRESTADORA                      40177 non-null  object
 4   UNIDAD DE PEAJE                         40177 non-null  object
 5   SENTIDO DE COBRO EN UNIDAD DE PEAJE     40177 non-null  object
 6   TOTAL VEHICULOS PAGANTES                40177 non-null  int64
 7   TOTAL VEHICULOS EXONERADOS              40177 non-null  int64
 8   TOTAL VEHICULOS EVADIDOS                40177 non-null  int64
 9   TOTAL DE VEHÍCULOS EXONERADOS EVENTOS EXTRAORDINARIOS  40177 non-null  int64
dtypes: int64(6), object(4)
memory usage: 3.1+ MB
```

Figura 22. *Dataframes de la tabla otorgada por OSITRAN*

Nota: Generación de contenido de autoría propia

Paso 5: Después de cargar el archivo de datos, se utiliza la función head () para mostrar una vista previa de los datos contenidos en los atributos mencionados. Esta función permite visualizar las primeras filas del conjunto de datos cargado, lo que nos ayuda a confirmar que el archivo se ha cargado correctamente y que los atributos están siendo interpretados adecuadamente. Al utilizar la función head (), se proporciona un resumen tabular que exhibe una muestra representativa de los primeros registros del conjunto de información. Esta vista previa nos permite verificar la integridad de los datos y asegurarnos de que se hayan leído correctamente todas las columnas y valores. Al inspeccionar estos datos iniciales, podemos identificar posibles errores de formato, ausencia de información o discrepancias en los datos. Además, esta visualización preliminar nos brinda una idea conceptualmente, se obtiene una visión general de cómo están estructurados y qué contienen los datos que han sido cargados, lo que simplifica la exploración y análisis subsiguientes.

```
df.head()
```

	AÑO	MES	CONCESION	ENTIDAD PRESTADORA	UNIDAD DE PEAJE	SENTIDO DE COBRO EN UNIDAD DE PEAJE	TOTAL VEHICULOS PAGANTES	TOTAL VEHICULOS EXONERADOS	TOTAL VEHICULOS EVADIDOS	TOTAL DE VEHÍCULOS EXONERADOS EVENTOS EXTRAORDINARIOS
0	2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP BAYOVAR	ASCENDENTE	677	9	0	0
1	2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP CHICAMA	ASCENDENTE	3985	24	0	0
2	2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP CHICAMA	DESCENDENTE	4795	26	0	0
3	2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP MÓRROPE	ASCENDENTE	899	5	0	0
4	2019	1	AUTOPISTA DEL SOL, TRAMO TRUJILLO-SULLANA	CONCESIONARIA VIAL DEL SOL S.A.	UP PACANGUILLA	ASCENDENTE	1682	11	0	0

Figura 23. *Uso de la función head ()*

Nota: Generación de contenido de autoría propia

Paso 6: Dado que la columna "sentido de cobro de peaje" contiene datos categóricos, es necesario realizar una transformación para convertirlos en datos binarios. Para lograr esto, se utiliza la función OneHotEncoder (), la cual asigna una representación binaria a cada categoría presente en la columna. Una vez aplicada esta transformación, la columna "sentido de cobro" se reemplaza por tres nuevas columnas: "ascendente", "descendente" y "ambos sentidos". Estas columnas resultantes serán de utilidad como datos de entrada para la tarea de predicción.

La columna "ascendente" contendrá valores binarios que indican si el sentido de cobro es ascendente en la unidad de peaje correspondiente. De manera similar, la columna "descendente" representará si el sentido de cobro es descendente, y la columna "ambos sentidos" indicará si el cobro se realiza en ambos sentidos de la unidad de peaje.

Esta transformación a datos binarios permite representar de manera más adecuada la información categórica del sentido de cobro, lo cual facilita su uso como entrada en modelos de predicción. Al convertirlo en variables binarias, se captura la relación entre las diferentes categorías y se brinda la posibilidad de utilizar esta información en algoritmos de aprendizaje automático para predecir resultados basados en los valores de estas nuevas columnas. Por lo tanto, esta conversión es esencial para habilitar un examen más exhaustivo y exacto de la información vinculada al método de recaudación de peajes.

```
from sklearn.preprocessing import OneHotEncoder


categorical_cols = ['SENTIDO DE COBRO EN UNIDAD DE PEAJE']
#categorical_cols = ['MES']
#categorical_cols = ['CONCESION']
#categorical_cols = ['ENTIDAD PRESTADORA']
##categorical_cols = ['UNIDAD DE PEAJE']


codificador = OneHotEncoder()

codificacion = codificador.fit_transform(df[["SENTIDO DE COBRO EN UNIDAD DE PEAJE"]])
#codificacion = codificador.fit_transform(df[["MES"]])
#codificacion = codificador.fit_transform(df[["CONCESION"]])
#codificacion = codificador.fit_transform(df[["ENTIDAD PRESTADORA"]])
##codificacion = codificador.fit_transform(df[["UNIDAD DE PEAJE"]])

#print(type(codificacion))
#print(codificacion)
#print(codificacion.toarray())

nuevas_cols = pd.DataFrame(codificacion.toarray(),columns=codificador.categories_,index=df.index)
#print(nuevas_cols)

data_other_cols = df.drop(columns=categorical_cols)

print (df.columns[df.columns.duplicated(keep=False)])
print (nuevas_cols.columns[nuevas_cols.columns.duplicated(keep=False)])

df = pd.concat([data_other_cols,nuevas_cols],axis=1)
df.head()
#df = df + nuevas_cols
#df
```

Figura 24. *Uso de OnehotEncoder*

Nota: Generación de contenido de autoría propia

Paso 7: Para categorizar las tres variables mencionadas (concesión, entidad prestadora y unidad de peaje), se utiliza la función LabelEncoder, la cual asigna un valor numérico único a cada categoría presente en las variables. Esta transformación de datos categóricos a numéricos permite que las variables sean utilizadas en modelos de aprendizaje automático, ya que muchos algoritmos requieren que los datos de entrada sean numéricos. Al aplicar el LabelEncoder, se recupera la información original que estaba siendo afectada por la transformación de los datos categóricos, ya que ahora cada categoría se representa con un valor numérico específico. Por ejemplo, a cada concesión se le asigna un valor único, al igual que cada entidad prestadora y unidad de peaje. Esto permite mantener la relación entre las categorías originales y los valores numéricos asignados, lo que es fundamental para un análisis preciso de los datos. Al recuperar los datos afectados por la transformación, se asegura que la información original no se pierda y se mantiene la capacidad de interpretar y comprender las categorías en su contexto original.

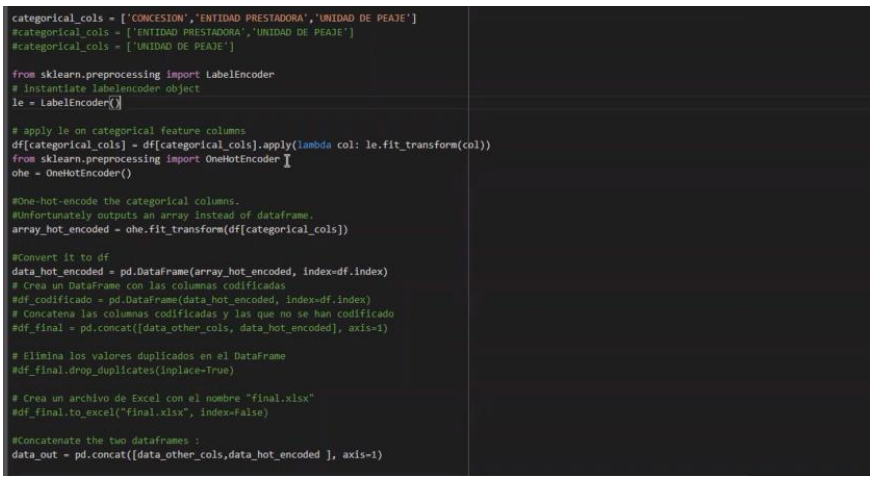


Figura 25. *Normalización del dataframe*

Nota: Generación de contenido de autoría propia

Paso 8: La utilización de la biblioteca NumPy nos permite llevar a cabo una transformación significativa en los valores de los vehículos pagantes y no pagantes. En particular, hemos convertido estos valores decimales en números enteros para lograr una representación más precisa y adecuada de los datos. Al aplicar esta transformación, se eliminan los decimales y se obtienen valores enteros que reflejan con mayor exactitud la cantidad de vehículos involucrados. Esta conversión resulta especialmente útil en casos

donde se requiere trabajar con valores enteros, como en análisis estadísticos o cálculos matemáticos. Al utilizar NumPy para esta transformación, nos aseguramos de obtener resultados consistentes y confiables, lo que a su vez contribuye a un análisis más preciso y coherente de los datos de vehículos pagantes y no pagantes.

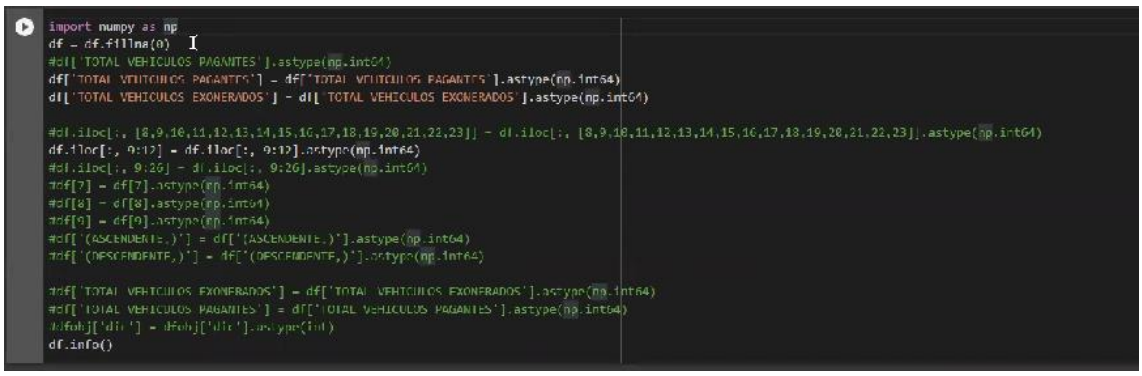


Figura 26. *Uso de biblioteca Numpy*

Nota: Generación de contenido de autoría propia

Paso 9: Durante el proceso de preparación de los datos, se han creado dos variables fundamentales: "x" y "y". Estas variables desempeñan un papel crucial en el análisis y la predicción de los datos. En primer lugar, se ha asignado la columna correspondiente a "x" para representar los datos de ingreso. Esta variable captura las características relevantes que se utilizarán como entrada en el modelo de predicción. Por otro lado, la variable "y" se ha asignado a la columna que contiene los datos a predecir. Esta variable representa el resultado objetivo que se busca estimar a partir de los datos de entrada. Al establecer y asignar correctamente estas variables, se establece una estructura clara y organizada para el análisis de los datos. Esto facilita el posterior procesamiento y modelado de los datos, así como la evaluación de la precisión del modelo de predicción.

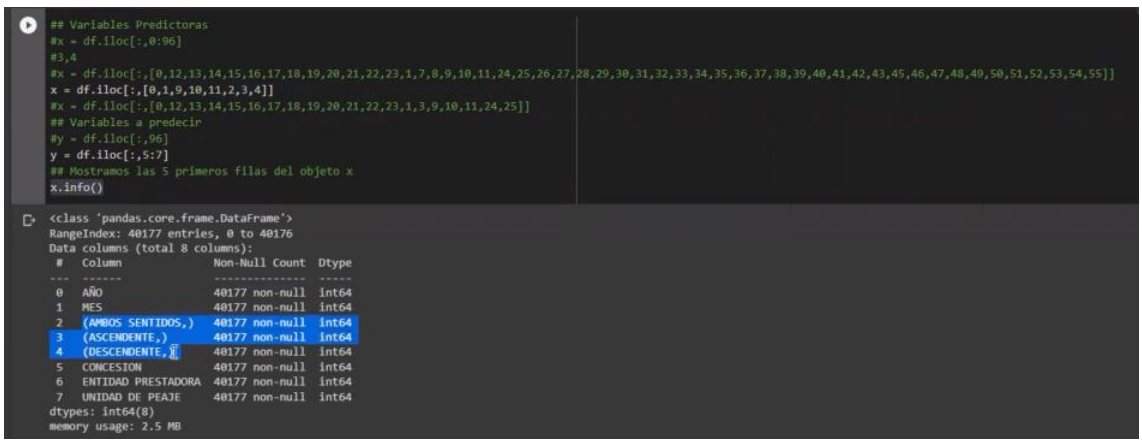
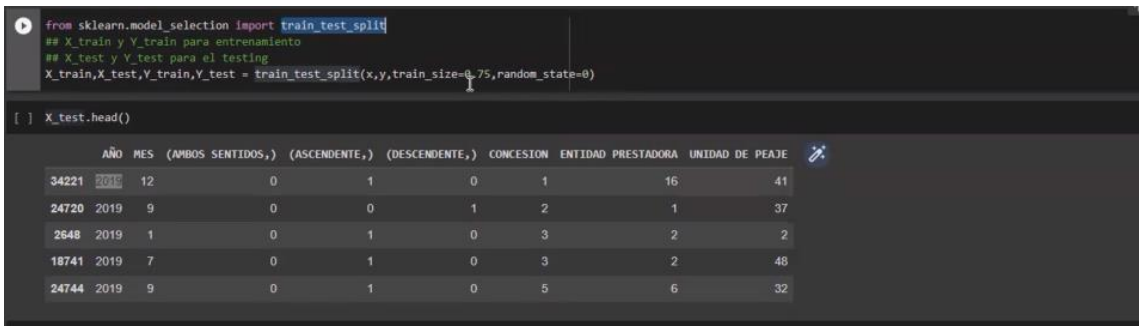


Figura 27. *Dataframe de la tabla a entrenar*

Nota: Generación de contenido de autoría propia

Paso 10: La inclusión de la biblioteca sklearn nos brinda acceso a una característica esencial conocida como train_test_split, la cual cumple un papel fundamental en la separación de datos para el entrenamiento y la evaluación de un modelo. Al importar esta función, podemos particionar de manera aleatoria un conjunto de datos en dos grupos distintos: uno destinado al entrenamiento del modelo y el otro para llevar a cabo pruebas y evaluar su rendimiento. En este escenario específico, se ha establecido una división del 75% para el grupo de entrenamiento y el 25% restante para las pruebas. Esta distribución aleatoria garantiza una representación equitativa y al azar de los datos en ambos conjuntos, lo cual es esencial para un entrenamiento y evaluación confiables del modelo. Al realizar esta división de datos de esta manera, podemos emplear el conjunto de entrenamiento para calibrar y mejorar el modelo, mientras que el conjunto de pruebas se utiliza para evaluar su capacidad predictiva y verificar su precisión. Esto nos permite obtener una evaluación realista del rendimiento del modelo en datos nuevos y no previamente vistos.



```
from sklearn.model_selection import train_test_split
## X_train y Y_train para entrenamiento
## X_test y Y_test para el testing
X_train,X_test,Y_train,Y_test = train_test_split(x,y,train_size=0.75,random_state=0)
```

```
[ ] X_test.head()
```

	AÑO	MES	(AMBOS SENTIDOS,)	(ASCENDENTE,)	(DESCENDENTE,)	CONCESION	ENTIDAD PRESTADORA	UNIDAD DE PEAJE
34221	2019	12	0	1	0	1	16	41
24720	2019	9	0	0	1	2	1	37
2648	2019	1	0	1	0	3	2	2
18741	2019	7	0	1	0	3	2	48
24744	2019	9	0	1	0	5	6	32

Figura 28. *Dataframe de la variable x*

Nota: Generación de contenido de autoría propia

Paso 11: Para asegurarnos de que los datos de prueba sean representativos y estén correctamente cargados, se muestra un vistazo a las primeras cinco filas de cada una de las variables. Esta visualización nos permite examinar de manera rápida y eficiente los valores y patrones presentes en los datos de prueba. Al observar estas filas iniciales, podemos identificar posibles discrepancias, errores o tendencias que podrían requerir una mayor atención durante el análisis. Además, esta vista preliminar nos brinda una idea general de cómo se distribuyen los datos y nos permite realizar una evaluación inicial de su calidad y coherencia. Al tener acceso a esta información, podemos tomar decisiones informadas sobre el preprocesamiento, la selección de características o cualquier otro ajuste necesario para mejorar la calidad y la eficacia de nuestro modelo de predicción.

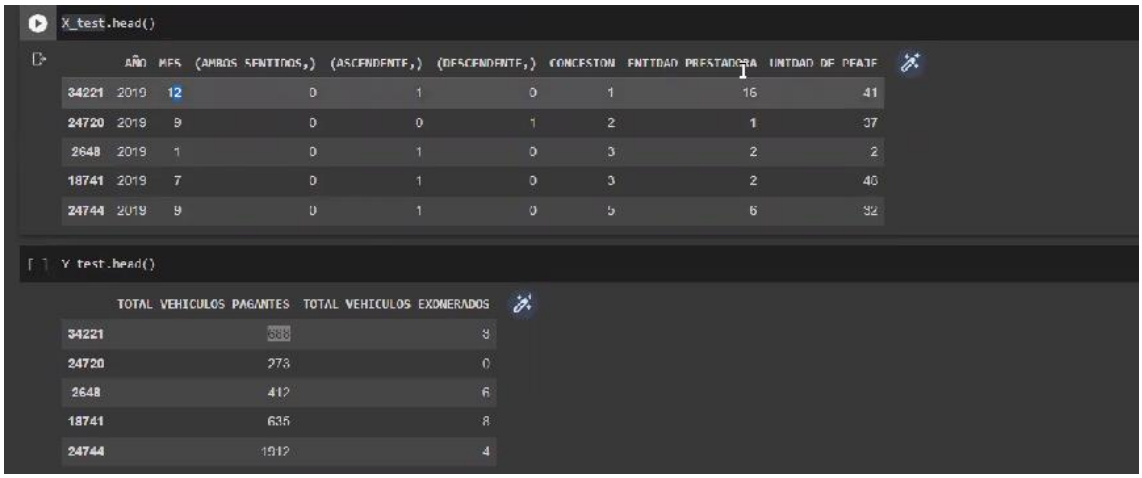


Figura 29. *Dataframes de la tabla normalizada)*

Nota: Generación de contenido de autoría propia

Paso 12: En el proceso de entrenamiento del modelo, hemos importado la funcionalidad de DecisionTreeRegressor, el cual es un modelo de regresión basado en árboles de decisión. Utilizando esta funcionalidad, hemos aplicado la función fit para entrenar el modelo utilizando nuestros datos de entrenamiento. Durante este proceso de entrenamiento, se han asignado 28 nodos al árbol de decisión, lo cual se considera una cantidad apropiada para obtener una precisión aceptable en nuestras predicciones. Una vez que el modelo ha sido entrenado, hemos utilizado la función r2_score para evaluar su rendimiento y obtener el porcentaje de aceptación del algoritmo, que en este caso ha sido del 84%. Este valor nos indica la proporción de la varianza de los datos que puede ser explicada por nuestro modelo, lo cual es un indicativo de la precisión y la calidad de las predicciones realizadas.

Figura 30. *Función DecisionTreeRegresor*

Nota: Generación de contenido de autoría propia

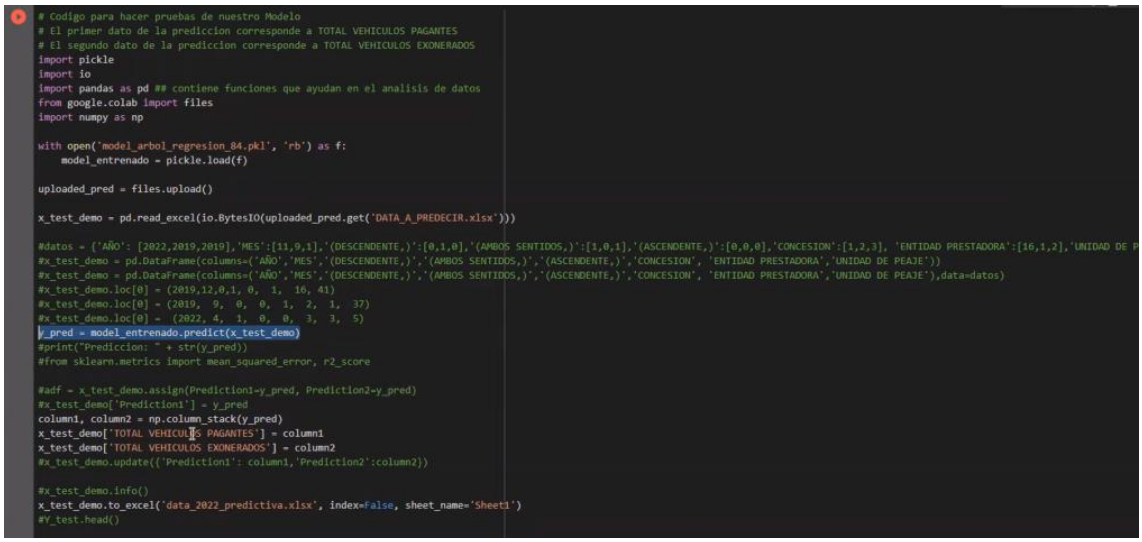
Paso 13: Después de entrenar el modelo y obtener un resultado satisfactorio, procedemos a realizar la predicción de la variable "x". Para ello, utilizamos la funcionalidad predict del modelo entrenado, lo que nos proporciona una matriz de predicciones. Esta matriz

se presenta dentro de otra matriz más grande, donde cada fila corresponde a una observación y cada columna representa una característica o variable. De esta manera, podemos asociar cada dato de la predicción con su respectiva columna y tener una visión clara de cómo se relacionan los valores predichos con las características específicas que hemos considerado en nuestro modelo. Esta estructura de matriz dentro de matriz nos permite organizar y visualizar de manera ordenada los resultados de la predicción, lo que facilita el análisis y la interpretación posterior. Al identificar qué dato pertenece a cada columna, podemos comprender mejor cómo el modelo ha utilizado las características de entrada para realizar las predicciones y evaluar su coherencia con respecto a los datos de prueba.

Figura 31. *Resultado de la predicción*

Nota: Generación de contenido de autoría propia

Paso 14: Para utilizar el modelo previamente entrenado, hacemos uso del bloque "with open". Dentro de este bloque, cargamos el modelo mediante la función "load" para asegurarnos de que se haya guardado correctamente. A continuación, utilizamos la función "read_excel" de la librería Pandas para leer el archivo de datos que deseamos predecir. Al cargar el archivo, obtenemos los datos necesarios para realizar las predicciones, en este caso, los valores correspondientes a los vehículos pagantes y no pagantes. Estos datos se extraen de manera adecuada y se preparan para ser procesados por el modelo previamente entrenado. Una vez que hemos realizado esta preparación, podemos proceder a utilizar el modelo para realizar las predicciones sobre estos datos. Esto nos permitirá obtener resultados precisos y relevantes que nos ayudarán a tomar decisiones informadas en base a las características de los vehículos en términos de su estado de pago.



```
# Codigo para hacer pruebas de nuestro Modelo
# El primer dato de la prediccion corresponde a TOTAL VEHICULOS PAGANTES
# El segundo dato de la prediccion corresponde a TOTAL VEHICULOS EXONERADOS
import pickle
import io
import pandas as pd ## contiene funciones que ayudan en el analisis de datos
from google.colab import files
import numpy as np

with open('model_arbol_regresion_84.pkl', 'rb') as f:
    model_entrenado = pickle.load(f)

uploaded_pred = files.upload()

x_test_demo = pd.read_excel(io.BytesIO(uploaded_pred.get('DATA_A_PREDECIR.xlsx')))

#datos = {'AÑO': [2022,2019,2019],'MES':[11,9,1],'(DESCENDENTE,)':[0,1,0],'(AMBOS SENTIDOS,)':[1,0,1],'(ASCENDENTE,)':[0,0,0],'CONCESION':[1,2,3], 'ENTIDAD PRESTADORA':[16,1,2],'UNIDAD DE P
#x_test_demo = pd.DataFrame(columns=('AÑO','MES','(DESCENDENTE,)','(AMBOS SENTIDOS,)','(ASCENDENTE,)','CONCESION', 'ENTIDAD PRESTADORA','UNIDAD DE PEAJE'))
#x_test_demo = pd.DataFrame(columns=('AÑO','MES','(DESCENDENTE,)','(AMBOS SENTIDOS,)','(ASCENDENTE,)','CONCESION', 'ENTIDAD PRESTADORA','UNIDAD DE PEAJE'),data=datos)
#x_test_demo.loc[0] = (2019,12,0,1, 0,  1,  16, 41)
#x_test_demo.loc[0] = (2019,  9,  0,  0,  1,  2,  1,  37)
#x_test_demo.loc[0] =  (2022,  4,  1,  0,  0,  3,  3,  5)
y_pred = model_entrenado.predict(x_test_demo)
#print("Prediccion: " + str(y_pred))
#from sklearn.metrics import mean_squared_error, r2_score

#adf = x_test_demo.assign(Prediction1=y_pred, Prediction2=y_pred)
#x_test_demo['Prediction1'] = y_pred
column1, column2 = np.column_stack(y_pred)
x_test_demo['TOTAL VEHICULOS PAGANTES'] = column1
x_test_demo['TOTAL VEHICULOS EXONERADOS'] = column2
#x_test_demo.update({'Prediction1': column1,'Prediction2':column2})

#x_test_demo.info()
x_test_demo.to_excel('data_2022_predictiva.xlsx', index=False, sheet_name='Sheet1')
#Y_test.head()
```

Figura 32. *Entrenamiento del algoritmo*

Nota: Generación de contenido de autoría propia

Paso 15: Una vez que el modelo predictivo ha sido ejecutado exitosamente, se realiza una actualización del ícono de la carpeta para indicar que el proceso de predicción ha finalizado. Esta actualización visual es útil para reconocer de manera rápida y sencilla que el archivo de salida está listo para ser descargado. En este caso, el archivo se denomina "dato_2022_predictiva". Al descargar este archivo, los usuarios podrán acceder a los datos de predicción de las unidades vehiculares pagantes y no pagantes. Esta información les permitirá visualizar y analizar en detalle las predicciones realizadas por el modelo en relación con el estado de pago de los vehículos. Al tener acceso a estas predicciones, los usuarios podrán tomar decisiones fundamentadas y estratégicas basadas en los patrones y tendencias identificados en los datos predictivos. En resumen, una vez que el modelo predictivo ha sido ejecutado, se actualiza el ícono de la carpeta para indicar la finalización del proceso y se habilita la descarga del archivo "dato_2022_predictiva", donde los usuarios podrán visualizar y analizar las predicciones de las unidades vehiculares pagantes y no pagantes.

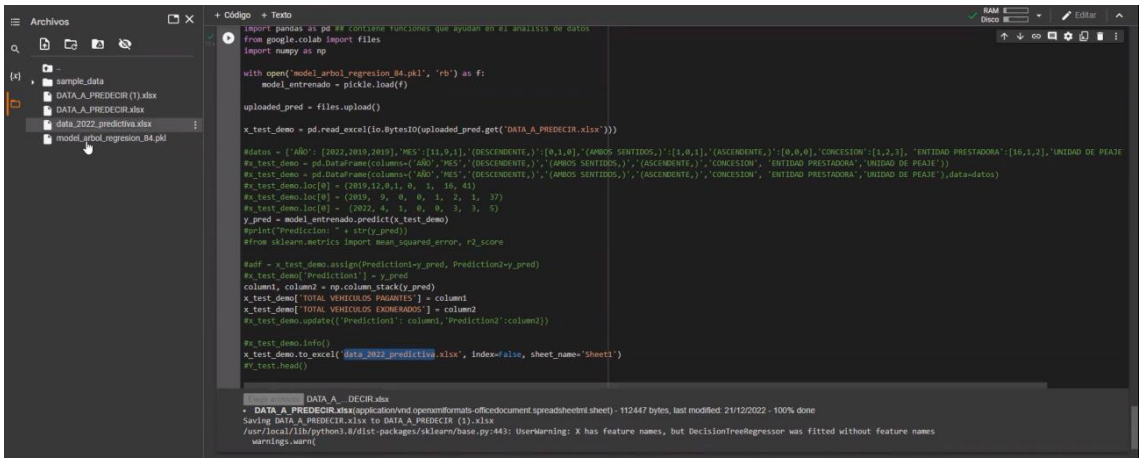


Figura 33. *Envío de los resultados del algoritmo*

Nota: Generación de contenido de autoría propia

Paso 16: Después de completar el proceso de predicción, se presenta la visualización de los datos de las columnas de predicción. Es importante destacar que las demás columnas del conjunto de datos se presentan en forma numérica, debido a que se han categorizado previamente. Esta categorización se realiza con el propósito de facilitar la interpretación y el análisis de los datos.

AÑO	MES	('AMBOS SENTIDOS',)	('ASCENDENTE',)	('DESCENDENTE',)	CONCESION	ENTIDAD PRESTADORA	UNIDAD DE PEAJE	TOTAL VEHICULOS PAGANTES	TOTAL VEHICULOS EXONERADOS
2022	11	0	11	0	0	5	49	47030	243
2022	11	0	11	0	0	5	50	175739	762
2022	11	0	0	11	0	5	50	175965	738
2022	11	0	11	0	0	5	511	136511	526
2022	11	0	11	0	0	5	52	97978	126
2022	11	0	0	11	0	5	52	98896	93
2022	11	0	11	0	0	5	53	136511	526
2022	11	0	0	11	0	5	53	135957	489
2022	11	0	11	0	11	116	6	349679	2363
2022	11	0	0	11	11	116	6	349679	2363
2022	11	0	0	11	11	116	29	207521	1529
2022	11	0	11	0	11	116	30	207521	1529
2022	11	0	0	11	11	116	30	207521	1529
2022	11	0	11	0	11	116	34	207521	1529
2022	11	0	0	11	11	116	34	207521	1529
2022	11	0	11	0	11	116	411	207521	1529
2022	11	0	0	11	11	116	411	207521	1529
2022	11	11	0	0	2	11	37	15096	224
2022	11	0	11	0	2	11	37	25418	259
2022	11	0	0	11	2	11	37	22933	138
2022	11	11	0	0	3	2	2	31044	442
2022	11	0	11	0	3	2	2	13389	253
2022	11	0	0	11	3	2	2	13332	253
2022	11	11	0	0	3	2	117	47593	424
2022	11	0	11	0	3	2	117	23765	214
2022	11	0	0	11	3	2	117	23765	214
2022	11	11	0	0	3	2	48	47593	424
2022	11	0	11	0	3	2	48	23765	214
2022	11	0	0	11	3	2	48	23765	214
2022	11	0	11	0	4	113	20	14476	138
2022	11	0	0	11	4	113	20	13913	135
2022	11	0	11	0	4	113	39	23146	124
2022	11	0	0	11	4	113	39	24945	257
2022	11	0	11	0	4	113	40	21485	68
2022	11	0	0	11	4	113	40	11801	648
2022	11	0	11	0	5	6	4	137337	1102
2022	11	0	0	11	5	6	4	135805	1104
2022	11	0	11	0	5	6	116	142900	1166
2022	11	0	0	11	5	6	116	128560	964
2022	11	0	11	0	5	6	211	142900	1166

Figura 34. *Dataframes de la variable de entrada y las predichas*

Nota: Generación de contenido de autoría propia

5.2. Evaluar el rendimiento de los modelos mediante métricas de precisión y error, y refinarlos según sea necesario para mejorar su capacidad predictiva en las unidades de peaje peruanos.

Paso 5.Evaluación.

Al realizar la verificación de la precisión del algoritmo utilizado, se ha determinado que el mismo presenta un porcentaje de aceptación del 84%. Este resultado confirma que la fiabilidad de los resultados se encuentra dentro de los estándares aceptados. Dicha medida de precisión representa un indicador sólido que respalda la confianza en la validez de los resultados obtenidos mediante el algoritmo utilizado. Un porcentaje de aceptación del 84% indica que la mayoría de las predicciones realizadas por el algoritmo concuerdan con los resultados reales dentro de un margen razonable de error. Esta información es crucial para asegurar que los resultados generados por el algoritmo son confiables y se pueden utilizar con confianza en la toma de decisiones o en el análisis de datos. Asimismo, es importante destacar que la verificación de la precisión del algoritmo es un proceso fundamental para garantizar que los resultados obtenidos sean consistentes y confiables, y permitir que los usuarios tengan una visión clara sobre la calidad de los resultados generados. Esta alta precisión del 84% indica que el algoritmo utilizado es capaz de realizar predicciones con un nivel significativo de exactitud, lo cual es fundamental para respaldar la toma de decisiones informadas y eficientes. Además, este porcentaje de aceptación demuestra que el algoritmo ha sido entrenado adecuadamente y ha aprendido de los datos disponibles, lo que permite obtener resultados coherentes y confiables en diferentes escenarios.

Figura 35. *Grado de aceptación del algoritmo*

Nota: Generación de contenido de autoría propia.

5.3. Implementar un modelo predictivo basado en minería de datos que permita predecir patrones de tráfico en unidades de peaje peruanos.

Paso 6.Despliegue.

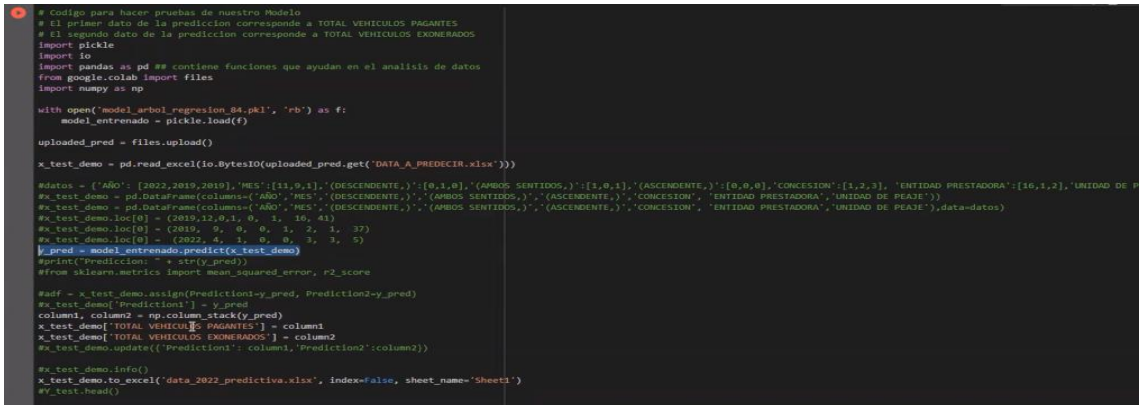
Una vez concluido el desarrollo del sistema, se procede a mostrar los resultados obtenidos a través de la implementación del algoritmo de investigación. Este proceso implica una serie de pasos fundamentales que aseguran la correcta ejecución y validación de los datos. En primer lugar, se realizan las tareas de carga de las librerías pertinentes, lo que permite disponer de las herramientas necesarias para llevar a cabo las siguientes etapas del análisis. Posteriormente, se lleva a cabo un proceso de limpieza de los datos, con el fin de eliminar posibles anomalías o inconsistencias que podrían afectar la calidad de los resultados. Una vez que los datos han sido depurados y están en condiciones óptimas, se procede a entrenar el modelo de árbol de decisiones, aplicando técnicas y ajustes adecuados para lograr un alto grado de precisión en las predicciones. Este entrenamiento es esencial para que el modelo sea capaz de analizar y clasificar los datos de manera eficiente y proporcionar resultados confiables. Luego de completar el proceso de entrenamiento, se lleva a cabo una verificación exhaustiva para evaluar el rendimiento del modelo y asegurar que cumpla con los estándares de precisión requeridos. Para facilitar la interpretación de los resultados, se imprimen los datos obtenidos en el algoritmo y se presentan en una representación gráfica clara y concisa, como se muestra en la figura 36. Esta figura visualiza de manera efectiva los resultados generados, permitiendo una fácil comprensión y análisis de los datos.

Figura 36. *Datos obtenidos del entrenamiento de la variable x*

Nota: Generación de contenido de autoría propia

En la figura 37, Al realizar el proceso de lectura del archivo Excel obtenido para realizar las predicciones, se siguen una serie de pasos específicos. En primer lugar, se utiliza la funcionalidad de lectura proporcionada por la librería pandas para cargar el archivo en un objeto de tipo DataFrame. Posteriormente, se aplican las transformaciones y preparaciones necesarias en los datos, como la codificación de variables categóricas, la limpieza de valores

nulos y la normalización de los datos si es necesario. Una vez completada esta etapa de preparación de los datos, se procede a generar las predicciones utilizando el modelo previamente entrenado. Estas predicciones se almacenan en nuevas variables dentro del DataFrame, correspondientes a las columnas de interés. Por último, se guarda el DataFrame modificado en un nuevo archivo Excel, utilizando la funcionalidad de escritura de pandas. El nombre asignado al archivo resultante es "data_2022_predictiva.xlsx", que contiene los datos predichos para el año 2022. Este archivo proporciona una visión detallada y organizada de los resultados obtenidos, lo que facilita su posterior análisis y evaluación.



```
# Codigo para hacer pruebas de nuestro Modelo
# El primer dato de la prediccion corresponde a TOTAL VEHICULOS PAGANTES
# El segundo dato de la prediccion corresponde a TOTAL VEHICULOS EXONERADOS
import pickle
import io
import pandas as pd ## contiene funciones que ayudan en el analisis de datos
from google.colab import files
import numpy as np

with open('model_arbol_regresion_84.pkl', 'rb') as f:
    model_entrenado = pickle.load(f)

uploaded_pred = files.upload()

x_test_demo = pd.read_excel(io.BytesIO(uploaded_pred.get('DATA_A_PREDECIR.xlsx')))

#datos = {'AÑO': [2022,2019,2019],'MES':[11,9,1],'(DESCENDENTE,)':[0,1,0],'(AMBOS SENTIDOS,)':[1,0,1],'(ASCENDENTE,)':[0,0,0],'CONCESION':[1,2,3], 'ENTIDAD PRESTADORA':[16,1,2],'UNIDAD DE P
#x_test_demo = pd.DataFrame(columns={'AÑO','MES','(DESCENDENTE,)','(AMBOS SENTIDOS,)','(ASCENDENTE,)','CONCESION', 'ENTIDAD PRESTADORA','UNIDAD DE PEAJE'})
#x_test_demo = pd.DataFrame(columns={'AÑO','MES','(DESCENDENTE,)','(AMBOS SENTIDOS,)','(ASCENDENTE,)','CONCESION', 'ENTIDAD PRESTADORA','UNIDAD DE PEAJE'),data=datos)
#x_test_demo.loc[0] = (2019,12,0,1, 0,  1,  16, 41)
#x_test_demo.loc[0] = (2019,  9,  0,  0,  1,  2,  1,  37)
#x_test_demo.loc[0] =  (2022, 4,  1,  0,  9,  3,  3,  5)
y_pred = model_entrenado.predict(x_test_demo)
#print("Prediccion: " + str(y_pred))
#from sklearn.metrics import mean_squared_error, r2_score

#adf = x_test_demo.assign(Prediction1=y_pred, Prediction2=y_pred)
#x_test_demo['Prediction1'] = y_pred
column1, column2 = np.column_stack(y_pred)
x_test_demo['TOTAL VEHICULOS PAGANTES'] = column1
x_test_demo['TOTAL VEHICULOS EXONERADOS'] = column2
#x_test_demo.update({'Prediction1': column1,'Prediction2':column2})

#x_test_demo.info()
x_test_demo.to_excel('data_2022_predictiva.xlsx', index=False, sheet_name='Sheet1')
#y_test.head()
```

Figura 37. *Proceso de envío de resultados*

Nota: Generación de contenido de autoría propia

CONCLUSIONES

1. Mediante la presente investigación, se implementó un modelo de algoritmo predictivo, donde se empleó 40177 registros (Registro de vehículos, Open Data de OSITRAN 2022), para el entrenamiento y pruebas necesarias, y se obtuvo una precisión de 84%; por tanto, el algoritmo permite predecir patrones de tráfico, con un porcentaje de condición de aceptable, para el uso correspondiente.
2. Como primera conclusión específica, se afirma que se logró obtener datos históricos, y consistentes de las unidades de peaje peruanos.
3. Como segunda conclusión especifica, se logró identificar 8 variables, de los cuales, tres fueron variables categórico; los mismos que fueron de gran importancia para el modelo predictivo a construir.
4. Como tercera conclusión, se afirma que, mediante el uso de las librerías como IO, Pandas, sklearn y las funciones como el head ( ), OneHotEncoder ( ), LabelEncoder (), DecisionTreeRegressor ( ) función predict (). Con un algoritmo de árbol de decisión que consta de 28 nodos. Se logró construir un modelo de algoritmo predictivo
5. Como cuarta y última conclusión, de afirma que, mediante la evaluación, se concluye que el modelo de algoritmo construido, tiene una precisión del 84%, por tanto, es considerado aceptable y tiene un buen rendimiento.

RECOMENDACIONES

1. Continuar con la obtención y el almacenamiento sistemático de datos históricos relacionados con el tráfico en las estaciones de peaje de Perú, enfocándose en aspectos como las horas pico, los días de la semana y eventos especiales. Este proceso debe ser continuo y riguroso, respaldado por tecnología avanzada de recopilación de datos y sistemas de almacenamiento seguros. Además, se sugiere que se explore la posibilidad de ampliar la recopilación de datos para incluir otros factores relevantes, como condiciones meteorológicas y eventos de construcción, para obtener una visión más completa y precisa de los patrones de tráfico. Estos datos serán esenciales para futuros análisis y la implementación de modelos predictivos, lo que permitirá una gestión más efectiva del tráfico y una toma de decisiones basada en datos en el sistema de peaje peruano.
2. Llevar a cabo un exhaustivo proceso de identificación y selección de las variables que ejercen una influencia significativa en los patrones de tráfico en las unidades de peaje. Esto implica considerar factores como las condiciones climáticas, festivos locales y datos económicos que puedan tener un impacto en el flujo vehicular. Una vez identificadas estas variables, es fundamental prepararlas adecuadamente para su integración en el modelo predictivo. Esto incluye la recopilación precisa y continua de estos datos, su limpieza y transformación, y la creación de un conjunto de datos sólido y coherente que sirva como entrada para el modelo. Además, se recomienda mantener un monitoreo constante de estas variables a lo largo del tiempo para asegurarse de que el modelo esté actualizado y sea capaz de adaptarse a posibles cambios en las condiciones que afectan el tráfico en las estaciones de peaje peruanos.
3. Llevar a cabo el diseño y construcción de modelos de minería de datos, como regresiones, redes neuronales o árboles de decisión, con el propósito de predecir patrones de tráfico futuros en las estaciones de peaje peruanos. Para ello, es esencial contar con un conjunto de datos históricos de alta calidad y actualizados que sirvan como base de entrenamiento para estos modelos. Además, se sugiere realizar un proceso de evaluación riguroso de los modelos para determinar su precisión y rendimiento antes de implementarlos en la toma de decisiones en tiempo real. Este enfoque de modelado predictivo puede proporcionar valiosas perspectivas sobre el tráfico en las estaciones de peaje y contribuir a una gestión más eficiente de las operaciones.

4. Realizar una evaluación exhaustiva del rendimiento de los modelos utilizando métricas de precisión y error específicas para la predicción de patrones de tráfico en las estaciones de peaje peruanos. Esta evaluación periódica permitirá identificar áreas de mejora y ajustar los modelos en consecuencia. Es importante mantener un proceso de refinamiento continuo para mejorar la capacidad predictiva de los modelos a medida que se obtengan más datos y se enfrenten a nuevas condiciones de tráfico. Esta práctica garantizará que los modelos sean efectivos y confiables en la toma de decisiones relacionadas con la gestión del tráfico en tiempo real

BIBLIOGRAFÍA

Andina. (2021). *Flujo vehicular por peajes a nivel nacional creció 19.6% en agosto 2021.* https://andina.pe/agencia/noticia-flujo-vehicular-peajes-a-nivel-nacional-crecio-196-agosto-2021-866399.aspx

Baena. (2017). *Metodología de la investigación.* http://www.biblioteca.cij.gob.mx/Archivos/Materiales_de_consulta/Drogas_de_Abuso/Articulos/metodologia%20de%20la%20investigacion.pdf

BBC. (2017). *¿Cuáles son las ciudades con mejor y peor transporte público en América Latina?* https://www.bbc.com/mundo/noticias-america-latina-38927134

Beltrán, B. (2003). *Minería de datos.* Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la ComputacióN. http://bbeltran.cs.buap.mx/NotasMD.pdf

BishopF.R.Eng., C. M. (2006). *Pattern Recognitionand Machine Learning.* USA: M. Jordan; J. Kleinberg; B. Scho¨lkopf. http://research.microsoft.com/ cmbishop

Brachman, R., & Anand, T. (1996). *The Process of Knowledge Discovery in Databases: A Human-Centered Approach.*

Bull, A. (2003). *Congestión de tránsito: El problema y cómo enfrentarlo.* Chile: Cepal. https://repositorio.cepal.org/bitstream/handle/11362/27813/S0301049_es.pdf

Comercio. (2019). *El transporte público era el quinto problema que más afecta a ciudadanos: ahora es el segundo.* https://elcomercio.pe/lima/el-transporte-publico-paso-a-ser-el-segundo-problema-que-mas-afecta-a-ciudadanos-lima-como-vamos-noticia/

Cuevas, V., Alvares, S., Azcona, M., & Rodríguez, A. (2019). Capacidad predictiva de las Máquinas de Soporte Vectorial. Una aplicación en la planificación financiera. *Revista Cubana de Ciencias Informáticas, 13*(3), 59-75. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992019000300059&lng=es&tlng=es

Espino, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.* [Tesis de fin de grado, Universitat Oberta de Catalunya]. https://openaccess.uoc.edu/bitstream/10609/59565/6/caresptimTFG0117mem%C3%B2ria.pdf

Fayyad, U. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Computer Society, 11*(5), 20-25.

Gutierrez, J., & Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare, 3*(2), 33–51. https://doi.org/10.21158/23823399.v3.n2.2015.1440

Haro, S., Zúñiga, L., Meneses, A., Vera, L., & Escudero, A. (2018). Métodos de clasificación en minería de datos meteorológicos. *Perfiles: Revista científicas*, 107-113. http://dspace.espoch.edu.ec/handle/123456789/9395

Hernández, R., Fernández, C., & Baptista, L. (2014). *Metodología de la investigación - Sexta Edición - UCA.* https://repository.usta.edu.co/bitstream/handle/11634/44171/metodolog%C3%ADa%20%202014%20siampieri.pdf

IBM. (24 de 09 de 2023). *IBM*. IBM: https://www.ibm.com/es-es/topics/decision-trees

IBM. (02 de 10 de 2023). *IBM*. IBM: https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model

Jafari, S., Shahbazi, Z., & Byun, Y. (2022). Designing the Controller-Based Urban Traffic Evaluation and Prediction Using Model Predictive Approach. *Applied Sciences, 12*(4). https://doi.org/10.3390/app12041992

Jaramillo, J. (2017). *Mejoramiento de la circulación del flujo vehicular en la intersección de los jirones Orellana y Alfonso Ugarte de la ciudad de Tarapoto, distrito de Tarapoto, provincia y región San Martín.* [Tesis de pregrad, Universidad Nacional de San Martín]. https://repositorio.unsm.edu.pe/bitstream/11458/2715/1/CIVIL%20-%20Janneth%20Jaramillo%20Delgado.pdf

Márquez, B. &. (2015). *LA CONGESTIÓN VEHICULAR EN LA CIUDAD DE PIURA.* https://www.unp.edu.pe/libros/librolacongestionvehicular.pdf

Medina, R., & Ñique, C. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases, 10*, 165-189. https://doi.org/10.26439/interfases2017.n10.1775

Molina, J., & García, J. (2021). *Técnicas de Minería de Datos basadas en Aprendizaje Automático.* Universidad Nacional Agraria La Molina. https://santiagozapatakdd.files.wordpress.com/2011/03/curso-kdd-full-cap-3.pdf

MTC. (2018). *Manual de Carreteras: Diseño Geométrico.* Lima: Ministerio de Transportes y Comunicaciones. https://portal.mtc.gob.pe/transportes/caminos/normas_carreteras/documentos/manuales/Manual.de.Carreteras.DG-2018.pdf

Ordoñes, Y., & Grass, H. (2011). HERMINWEB: Herramienta de Minería de Uso de la Web Aplicado a los Registros del Proxy. https://www.researchgate.net/figure/Fases-del-proceso-KDD-segun-la-metodologia-CRISP-DM-Primeramente-se-debe-estudiar-el_fig1_233426470

OSITRAN. (29 de 09 de 2023). *OSITRAN*. OSITRAN: https://www.ositran.gob.pe/anterior/

Pérez, C., & Santín, D. (2008). *Minería de datos: técnicas y herramientas.* Madrid: : Ediciones Paraninfo, S.A.

Shen, Y. y. (2022). *" Traffic Accident Severity Prediction Based on Random Forest".* https://www.mdpi.com/2071-1050/14/3/1729

Shewan, D. (2021). *10 Companies Using Machine Learning in Cool Ways.* WordStream. https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Tarazona. (2016). *https://acreditacion.uni.edu.pe/wp-content/uploads/2017/06/Identification-of-Main-Factors-and-Variables-Describing-the-Quantity-and-Distribution-of-Fatal-Vehicular-Accidents-in-Metropolitan-City-of-Lima-Using-data-Mining-Techniques-Random-Forest.pdf.* https://acreditacion.uni.edu.pe/wp-content/uploads/2017/06/Identification-of-Main-Factors-and-Variables-Describing-the-Quantity-and-Distribution-of-Fatal-Vehicular-Accidents-in-Metropolitan-City-of-Lima-Using-data-Mining-Techniques-Random-Forest.pdf

Timarán, S., Hernández, I., Caicedo, S., Hidalgo, A., & Alvarado, J. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesiona*, 63-86. https://doi.org/10.16925/9789587600490

Torres. (2022). *Traffic Accident Severity Prediction Based on Random Forest.* https://revistas.unl.edu.ec/index.php/cedamaz/article/view/1181/851

ULLAH, y. o. (2019). *A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector.* https://www.researchgate.net/publication/332928038_A_Churn_Prediction_Model_Using_Random_Forest_Analysis_of_Machine_Learning_Techniques_for_Churn_Prediction_and_Factor_Identification_in_Telecom_Sector

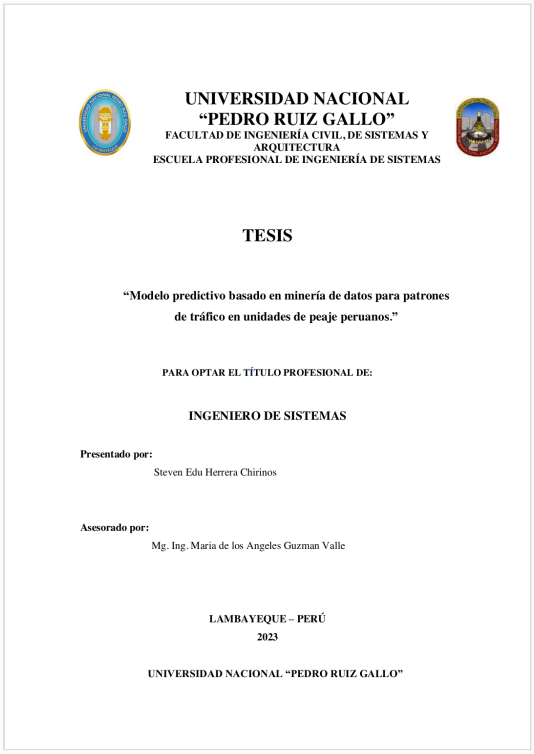
Xu, C. &. (2022). *Understanding vehicular routing behavior with location-based service data.* https://dspace.mit.edu/handle/1721.1/132006

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	Steven Herrera
Título del ejercicio:	Turni 2023
Título de la entrega:	Turnitin informe final de tesis
Nombre del archivo:	Informe_final_Steven_Herrera_3.docx
Tamaño del archivo:	3.63M
Total páginas:	72
Total de palabras:	15,580
Total de caracteres:	86,454
Fecha de entrega:	09-ago.-2023 07:55p. m. (UTC-0500)
Identificador de la entre...	2143722544



UNIVERSIDAD NACIONAL
"PEDRO RUIZ GALLO"
FACULTAD DE INGENIERÍA CIVIL, DE SISTEMAS Y ARQUITECTURA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

"Modelo predictivo basado en minería de datos para patrones de tráfico en unidades de peaje peruanos."

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

Presentado por:
Steven Edu Herrera Chirinos

Asesorado por:
Mg. Ing. Maria de los Angeles Guzman Valle

LAMBAYEQUE – PERÚ
2023

UNIVERSIDAD NACIONAL "PEDRO RUIZ GALLO"

Derechos de autor 2023 Turnitin. Todos los derechos reservados.

Turnitin informe final de tesis

INFORME DE ORIGINALIDAD

19%	19%	3%	%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	bbeltran.cs.buap.mx Fuente de Internet	3%
2	hdl.handle.net Fuente de Internet	2%
3	repositorio.espe.edu.ec Fuente de Internet	2%
4	repositorio.ucv.edu.pe Fuente de Internet	1%
5	repositorio.udd.cl Fuente de Internet	1%
6	acreditacion.uni.edu.pe Fuente de Internet	1%
7	dspace.espoch.edu.ec Fuente de Internet	1%
8	www.gsi.dit.upm.es Fuente de Internet	1%
9	docplayer.es Fuente de Internet	1%

10	revistas.unl.edu.ec Fuente de Internet	1 %
11	repositorio.unprg.edu.pe Fuente de Internet	<1 %
12	repositorio.unap.edu.pe Fuente de Internet	<1 %
13	issuu.com Fuente de Internet	<1 %
14	dspace.unl.edu.ec Fuente de Internet	<1 %
15	andina.pe Fuente de Internet	<1 %
16	www.scribd.com Fuente de Internet	<1 %
17	repositorio.uss.edu.pe Fuente de Internet	<1 %
18	www.slideshare.net Fuente de Internet	<1 %
19	www.coursehero.com Fuente de Internet	<1 %
20	nanopdf.com Fuente de Internet	<1 %
21	repositorio.unas.edu.pe Fuente de Internet	<1 %

22	pt.scribd.com Fuente de Internet	<1 %
23	repositorio.untels.edu.pe Fuente de Internet	<1 %
24	oa.upm.es Fuente de Internet	<1 %
25	prezi.com Fuente de Internet	<1 %
26	es.scribd.com Fuente de Internet	<1 %
27	www.cacic2016.unsl.edu.ar Fuente de Internet	<1 %
28	core.ac.uk Fuente de Internet	<1 %
29	pricila.senacyt.gob.pa Fuente de Internet	<1 %
30	repositorio.autonoma.edu.pe Fuente de Internet	<1 %
31	futur.upc.edu Fuente de Internet	<1 %
32	dspace.uazuay.edu.ec Fuente de Internet	<1 %
33	repositorio.unfv.edu.pe Fuente de Internet	<1 %

34	www.jove.com Fuente de Internet	<1 %
35	www.dspace.uce.edu.ec:8080 Fuente de Internet	<1 %
36	sired.udenar.edu.co Fuente de Internet	<1 %
37	repositorio.unam.edu.pe Fuente de Internet	<1 %

Excluir citas	Activo	Excluir coincidencias	< 15 words
Excluir bibliografía	Activo		

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE INGENIERÍA CIVIL, DE SISTEMAS Y ARQUITECTURA
UNIDAD DE INVESTIGACIÓN

"Año del Fortalecimiento de la Soberanía Nacional".

CONSTANCIA DE APROBACION DE ORIGINALIDAD DE TESIS

Según Res. N° 659-2020-R

Yo, Mg. Ing. María de los Ángeles Guzmán Valle**, asesora de tesis del bachiller**:

Bach. STEVEN EDU HERRERA CHIRINOS

TITULADA:

"MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA PREDECIR PATRONES DE TRÁFICO EN UNIDADES DE PEAJE PERUANOS"

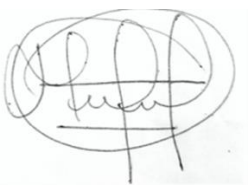
Luego de la revisión exhaustiva del documento constato que la misma tiene un índice de similitud de 19% verificable en el reporte de similitud del programa TURNITIN.

El suscrito analizó dicho reporte y concluyó que cada una de las coincidencias detectadas NO CONSTITUYEN PLAGIO. A mi leal saber y entender la tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad Nacional Pedro Ruiz Gallo.

Se expide la presente según lo dispuesto en la Resolución Nº 659-2020-R, de fecha 8 de setiembre de 2020 formativa para la obtención de Grados y Títulos de la UNPRG:

Lambayeque, 10 de agosto del 2023

ATENTAMENTE,



Mg. Ing. María de los Ángeles Guzmán Valle
DNI. 16730587

Se adjunta:
Recibo digital de Turnitin
Revisión de informe en Turnitin

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE INGENIERÍA CIVIL DE SISTEMAS Y DE ARQUITECTURA
DECANATO

ACTA DE SUSTENTACIÓN
N° 562-2024-FICSA-D

Siendo las 10:00 am del día 31 de enero del 2024, se reunieron los miembros de jurado de la Tesis titulada **"MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA PREDECIR PATRONES DE TRÁFICO EN UNIDADES DE PEAJE PERUANOS" con código N° IS_V_2022_007**, y designado por **Decreto Directoral N° 216-2023-UNPRG-FICSA-UI;** con la finalidad de Evaluar y Calificar la sustentación de la tesis profesional antes mencionada, conformado por los siguientes docentes:

DR. ING. ERNESTO KARLO CELI ARÉVALO — **PRESIDENTE**
DR. ING. EDWARD RONAL HARO MALDONADO — **SECRETARIO**
DR. ING. JUAN ELÍAS VILLEGAS CUBAS — **VOCAL**

Asesorado por MSC. ING. MARIA DE LOS ANGELES GUZMAN VALLE

El acto de sustentación fue autorizado por **OFICIO VIRTUAL N° 019-2024-UIFICSA**, la Tesis fue presentada y sustentada por el bachiller: **STEVEN EDU HERRERA CHIRINOS,** tuvo una duración de 60 minutos Después de la sustentación, y absueltas las preguntas y observaciones de los miembros del jurado; se procedió a la calificación respectiva:

	NUMERO	LETRAS	CALIFICATIVO
STEVEN EDU HERRERA CHIRINOS	17	DIECISIETE	BUENO

Por lo que quedan **APTOS** para obtener el Título Profesional de **INGENIERO DE SISTEMAS** de acuerdo con la Ley Universitaria 30220 y la normatividad vigente de la Facultad de Ingeniería Civil De Sistemas y de Arquitectura de la Universidad Nacional Pedro Ruiz Gallo.

Siendo las 11:30 AM ; del mismo día, se dio por concluido el presente acto académico, dándose conformidad al presente acto, con la firma de los miembros del jurado.

DR. ING. ERNESTO KARLO CELI ARÉVALO
PRESIDENTE

DR. ING. EDWARD RONAL HARO MALDONADO
SECRETARIO

DR. ING. JUAN ELÍAS VILLEGAS CUBAS
VOCAL

MSC. ING. MARIA DE LOS ANGELES GUZMAN VALLE
ASESOR

FACULTAD DE INGENIERÍA CIVIL, SISTEMAS Y ARQUITECTURA – U.N.P.R.G. – DECANATO

DR. ING. SERGIO BRAVO IDROGO
DECANO

SBI/