Predictive Modelling of Traffic Accidents in Bogota, Colombia: Uncovering Key Contributing Factors

Sebastián Castellanos, Alejandra Baena and Juan Camilo Ramírez

Universidad Antonio Nariño, Bogota, Colombia

Abstract

Traffic accidents pose a significant threat to public safety, making their prediction and understanding of contributing factors crucial for effective preventive measures. This study focuses on leveraging historical accident data from Bogotá, Colombia, to design and evaluate machine learning models for traffic accident prediction and take the initial steps toward the identification of the most influential factors associated with each accident. The main objective of this research is to develop accurate machine learning models that can effectively predict traffic accidents in Bogotá and that can later be used in the identification of the key contributing factors leading to these incidents. By achieving this objective, it will be possible to enhance road safety and devise targeted interventions to reduce the occurrence of accidents in the city. A comprehensive dataset comprising historical traffic accident records in Bogotá was collected and preprocessed for analysis. Various machine learning algorithms, including decision trees, random forests, and neural networks, were applied to develop predictive models. The models were trained and evaluated using the F1-score as well as the area under the ROC curve. The experimental results demonstrate the effectiveness of machine learning models in predicting traffic accidents in Bogotá. The best-performing models achieved performance scores over 0.80, both for the F1 metric and the area under the ROC curve, outperforming traditional statistical methods. These preliminary results are novel in the use of a more comprehensive and updated dataset of accidents in Bogotá and are envisaged to be extended with further analyses in order critical factors that strongly influence accident occurrence.

Keywords

Traffic accidents, Machine Learning, Prediction Models, Contributing Factors, Road Safety

1. Introduction

The increase in road traffic accidents poses a significant challenge to urban areas worldwide, affecting public safety, transportation efficiency, and overall societal well-being. The development of effective strategies for accident prevention requires a deep understanding of the contributing factors as well as an ability to accurately predict accident occurrences [1, 2, 3]. In recent years, machine learning techniques have shown great potential in this domain by leveraging historical accident data to identify patterns and extract valuable insights. The prediction of traffic accidents using machine learning techniques has garnered significant attention globally due to its potential to enhance road safety and inform effective accident prevention strategies. Numerous studies have explored this area across various regions, aiming to develop accurate

models capable of forecasting accident occurrences [4, 5, 6]. Concurrently, a few investigations have examined similar aspects within the context of Bogotá, Colombia [7, 8, 9, 10, 11]. Despite these efforts, a notable gap persists in the current body of research: the absence of studies leveraging the most recent and updated historical traffic accident data provided by the local government of Bogotá.

Bogotá, the capital city of Colombia, is home to a densely populated urban environment characterized by complex traffic dynamics and diverse transportation modes. Understanding the factors that contribute to traffic accidents in Bogotá is crucial for designing targeted interventions and improving road safety initiatives. By harnessing the power of machine learning algorithms, investigation in this line of research aims to develop accurate prediction models that can assist policymakers, city planners, and law enforcement agencies in making informed decisions and allocating resources effectively. Within the local context of Bogotá, previous studies have made strides toward enhancing the understanding of traffic accident patterns and risk factors. However, the majority of these investigations relied on historical data that may no longer accurately reflect the evolving dynamics of the city's traffic landscape [7, 8, 9, 10, 11]. Notably, the local government of Bogotá has recently made available an updated repository of historical traffic accident data, rendering previous analyses outdated and prompting the need for new insights drawn from this comprehensive and up-to-date dataset.

The present study builds upon a rich dataset comprising historical accident records collected over the past decade in Bogotá. The dataset encompasses a wide range of variables, both temporal and geographical, and accident severity. By systematically analyzing this comprehensive dataset, we strive to identify the key factors influencing accident occurrences and evaluate the performance of various machine learning models in predicting accidents with high precision and recall rates.

In conclusion, this article contributes to the growing body of research on using machine learning for traffic accident prediction and factor identification. By applying advanced machine learning techniques to historical accident data in Bogotá, we aim to enhance our understanding of the factors influencing accident occurrences and provide valuable insights for policymakers and stakeholders. The results of this study have the potential to inform evidence-based strategies for improving road safety, reducing accident rates, and creating a safer transportation environment in Bogotá and beyond.

2. Related work

Several studies have been conducted worldwide to leverage machine learning techniques for traffic accident prediction and the identification of contributing factors. In this section, we present a review of relevant literature, focusing on similar research efforts and their findings in the context of traffic accident analysis.

[7] consolidated a comprehensive dataset of motorcycle accidents in Colombia in the 2013-2018 period using various government sources. This dataset, including variables such as road and weather conditions surrounding each incident, is proposed as the basis for future investigations using predictive models in order to examine the main causes of traffic accidents involving motorcycles. The same authors use this information in order to investigate motorcycle-related

accidents in Cartagena, Colombia, in order to identify areas within the city where the most incidents occur, using a Bayes' empirical approach, as well as contributing factors, such as the number of intersections used by motorcyclists, which can be then used in order to implement countermeasures [8].

Following this line of research, [9] address the limited availability of sufficient historical data regarding traffic accidents in medium-size cities, such as Popayán, by employing complementary data collection techniques, such as naturalistic driving, *i.e.*, the continuous recording of driving information in real-time. Prediction models trained on these data were found to exhibit high-performance metrics and were used to identify regions within the urban area of the city where accidents are concentrated. [10] conducted a similar study with accident data from 2016 in Bogotá using multilayer perceptrons and naive Bayes models, finding that the former exhibit the best performance and that the most contributing factor to traffic accidents is drivers' behavior. Finally, taking a different direction, [11] proposed the use of social media and meteorological information as data sources resulting in high-performance models.

While these studies above have contributed significantly to the field of traffic accident prediction and factor identification, there is a limited number of studies focusing specifically on Bogotá, Colombia. The present study aims to address this research gap by utilizing a comprehensive dataset of historical accident records from Bogotá and applying a range of machine learning models to predict accident occurrences accurately. Furthermore, our study seeks to identify the most influential factors contributing to traffic accidents in the unique urban context of Bogotá, providing valuable insights for policymakers, transportation authorities, and urban planners in their efforts to improve road safety and reduce accident rates.

3. Methodology

Historical traffic accident data for Bogotá, Colombia, spanning the years 2015 to 2022, was sourced from three distinct repositories managed by the local government. These repositories contained detailed information on the vehicles involved, injured individuals, and fatalities associated with each accident. The data, provided in CSV format, was merged into a consolidated dataset based on a unique accident identification number shared across all three repositories. The integrated dataset comprised over 220,000 instances, each encompassing the accident's temporal attributes (time, year, month, weekday), geographical coordinates (latitude, longitude), accident severity (vehicle damage, injured, death), and the categorical class attribute "Type" (crash, runover, fall from a vehicle). The dimensionality of the integrated dataset was streamlined through the implementation of Principal Component Analysis (PCA). By selecting a subset of factors that collectively explain no less than 95% of the variance, PCA is used to distill the essential features while curtailing redundancy and noise.

Three distinct Machine Learning models were selected for prediction: multilayer neural networks, random forests, and decision trees. These models were chosen due to their capacity to handle complex datasets and demonstrate proficiency in predictive tasks. Prior to training the models, data preprocessing was carried out. This encompassed handling missing values, encoding categorical variables, and normalizing numerical features to ensure consistency and optimal model performance.

Model	F1-Score	Area Under ROC Curve
Multilayer Neural Network	0.83	0.87
Random Forest	0.81	0.84
Support Vector Machine	0.84	0.86

Table 1
Performance metrics of the predictive models trained on historical traffic accident data in Bogotá between 2015 and 2022.

To robustly assess the models' performance, a 10-fold cross-validation approach was employed. The dataset was divided into ten subsets, with each model trained and evaluated ten times, using a different subset as the validation set in each iteration while the rest were used for training. The performance of the prediction models was evaluated using two key metrics: the F1 score and the area under the Receiver Operating Characteristic (ROC) curve. The F1 score offers a balanced assessment of precision and recall, particularly relevant for imbalanced datasets like this. The ROC curve and its associated area provide insights into the model's ability to discriminate between classes. Following the cross-validation process, the three models were compared based on their F1 scores and ROC curve areas. This comparison aimed to identify the model that demonstrated the most robust and accurate performance in predicting traffic accidents and classifying their types.

The study strictly adhered to ethical guidelines and data privacy regulations. The utilized data was obtained from publicly available sources and did not contain any personally identifiable information.

4. Results

All three prediction models exhibited a commendable level of performance, as evidenced by F1-scores and AUC-ROC scores exceeding 0.80. This underscored their efficacy in effectively predicting traffic accidents based on historical data. However, closer scrutiny of the results revealed nuanced distinctions between the models. The multilayer neural network model notably outperformed the others in terms of AUC-ROC score. This outcome signified its superior ability to differentiate between distinct accident types. Conversely, the support vector machine model demonstrated a slight superiority in terms of F1-score, reflecting its capacity to achieve a harmonious balance between precision and recall.

The tabulated comparison is shown in Table 1 succinctly outlines the contrasting performance of the three models in relation to both the F1-score and the area under the ROC curve. The analysis reaffirms the multilayer neural network's superior AUC-ROC score, while the support vector machine model's marginally elevated F1-score showcases its prowess in achieving precision-recall equilibrium.

5. Conclusions

This study addresses a significant gap in the field of traffic accident prediction by leveraging advanced Machine Learning techniques to analyze the most recent and integrated historical accident data provided by the local government of Bogotá. While previous research has explored computational methods for investigating traffic accidents, our work stands out as the first to employ an innovative approach that integrates multiple updated data repositories managed by the city's government.

The outcomes of this study underscore the potential of machine learning methods in predicting traffic accidents and shedding light on their underlying dynamics. By demonstrating the viability of these approaches, we offer compelling evidence that such predictive models can provide crucial insights to enhance the local government's ability to comprehend and address this pressing issue. Our findings not only confirm the feasibility of utilizing machine learning in the realm of traffic safety but also highlight the potential of these models to support the design and implementation of preventive measures aimed at curbing this phenomenon.

It is important to note that while our study presents promising preliminary results, there are avenues for further exploration and refinement. Our current findings serve as a foundational platform for future investigations, where a more comprehensive exploration of the most influential factors contributing to traffic accidents can be undertaken. By integrating additional datasets managed by the local government, such as weather and vehicle conditions, we anticipate that these computational methods will provide even deeper insights into the complex interactions surrounding traffic accidents.

In conclusion, this study marks a significant step forward in the domain of traffic accident prediction and prevention in Bogotá. The pioneering integration of diverse and updated datasets, coupled with the application of machine learning models, has illuminated the potential to tackle this challenge in a novel and effective manner. As we move forward, the insights gleaned from this study will not only inform targeted interventions but also inspire continued research to comprehensively understand and mitigate the causes and consequences of traffic accidents.

Acknowledgments

References

[1] F.-R. Chang, H.-L. Huang, D. C. Schwebel, A. H. S. Chan, G.-Q. Hu, Global road traffic injury statistics: Challenges, mechanisms and solutions, Chinese journal of traumatology 23 (2020) 216–218.

[2] A. A. Mohammed, K. Ambak, A. M. Mosa, D. Syamsunur, A review of traffic accidents and related practices worldwide, The Open Transportation Journal 13 (2019).

---

[1]Universidad Antonio Nariño (https://www.uan.edu.co/).

[3] L. Wanumen, J. Moreno, H. Florez, Mobile based approach for accident reporting, in: Technology Trends: 4th International Conference, CITT 2018, Babahoyo, Ecuador, August 29–31, 2018, Revised Selected Papers 4, Springer, 2019, pp. 302–311.

[4] T. Bokaba, W. Doorsamy, B. S. Paul, Comparative study of machine learning classifiers for modelling road traffic accidents, Applied Sciences 12 (2022) 828.

[5] B. K. Mohanta, D. Jena, N. Mohapatra, S. Ramasubbareddy, B. S. Rawal, Machine learning based accident prediction in secure iot enable transportation system, Journal of Intelligent & Fuzzy Systems 42 (2022) 713–725.

[6] J. Garcia, W. Arias-Rojas, G. Hernandez, On the feasibility of real-time prediction of driver's traffic accident risk with auto ml using demographics and brain concentration levels information, 2022, pp. 1–5.

[7] H. Ospina-Mateus, S. B. Garcia, L. Q. Jiménez, K. Salas-Navarro, Dataset of traffic accidents in motorcyclists in bogotá, colombia, Data in brief 43 (2022) 108461.

[8] H. Ospina-Mateus, L. A. Q. Jiménez, F. J. Lopez-Valdes, S. S. Sana, Prediction of motorcyclist traffic crashes in cartagena (colombia): development of a safety performance function, RAIRO-Operations Research 55 (2021) 1257–1278.

[9] S. F. Yepes-Chamorro, J. J. Paredes-Rosero, R. Salazar-Cabrera, Álvaro de la Cruz, J. M. Madrid-Molina, Design, development and validation of an intelligent collision risk detection system to improve transportation safety: the case of the city of popayán, colombia, Sustainability 14 (2022) 10087.

[10] H. Vélez-Sánchez, H. Saavedra-Angulo, Use of data mining for root cause analysis of traffic accidents in colombia, 2021, pp. 674–688.

[11] C. Gutierrez-Osorio, F. A. González, C. A. Pedraza, Deep learning ensemble model for the prediction of traffic accidents using social media data, Computers 11 (2022) 126.