Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Machine Learning-Based System for Predictive Management of Road Accidents in Metropolitan Lima

Angel Portal, Marlonn Sandoval, Pedro Castañeda

Facultad de Ingeniería de Sistemas de Información, Universidad Peruana de Ciencias Aplicadas, Perú
Email: u20201B307@upc.edu.pe

This work presents a machine learning-based predictive system for managing traffic accidents in Metropolitan Lima. Given the high incidence of accidents in this region, the system employs a Random Forest model trained with historical accident data, weather variables, and road infrastructure characteristics. The methodology includes data collection and preprocessing, feature extraction, and deployment in a cloud-based architecture that enables real-time information processing. The results demonstrate notable performance in terms of accuracy and balance in incident classification, with high effectiveness in identifying high-risk areas. This system not only allows for precise predictions but also has a scalable and adaptable design that facilitates implementation in urban environments with similar characteristics, significantly contributing to accident reduction and improved traffic management in Metropolitan Lima.

1. Introduction

Traffic accidents in Metropolitan Lima represent a critical challenge for road safety, public health, and the local economy. Between 2017 and 2022, more than 50% of accidents nationwide occurred in the Peruvian capital, with a total of 5,449 accidents recorded in 2022, resulting in 930 fatalities and 7,817 injuries. The main identified causes include reckless driving, speeding, and driving under the influence, compounded by poor road infrastructure and insufficient traffic education [1]. These incidents not only compromise the safety of citizens but also contribute to traffic congestion, deteriorate the quality of public transportation, and impose significant economic and social costs on the city.

Globally and locally, various technological solutions have been implemented to address this problem. Internationally, intelligent traffic systems based on Big Data and machine learning algorithms have been prominent. For example, Yang Yang in China used statistical and machine learning methods to identify accident risks in Beijing [2], while Zhixiong in South Korea applied deep neural networks (DNN) to improve accident prediction accuracy [3]. In Peru, studies have identified contributing factors to traffic accidents [4], although without proposing applied technological solutions.

However, many of these solutions have limitations, such as reliance on specific technological infrastructures and lack of integration of multidimensional data. Approaches focusing solely on traffic or weather data [5] often omit crucial factors such as driver behavior and the socioeconomic characteristics of affected areas. This lack of comprehensiveness limits the adaptability and scalability of existing systems, especially in complex urban settings like Metropolitan Lima.

In response to these deficiencies, this paper proposes an innovative approach to traffic accident prediction in Metropolitan Lima by integrating multidimensional data and using machine learning algorithms. The proposed system incorporates both traditional variables and less conventional factors, allowing for more accurate and real-time predictions. Additionally, its adaptable and scalable design facilitates implementation in various urban contexts with similar characteristics, effectively contributing to accident reduction and improved traffic management.

The following sections detail the development and implementation of the proposed predictive system. Section 2 reviews related works relevant to the project. Section 3 covers the system design, including architecture, the machine learning model, dataset description, indicators used, and developed interfaces. Section 4 presents the results obtained from the conducted experiments. Section 5 discusses the impact and implications of the findings. Finally, Section 6 concludes with a reflection on the system's benefits and suggestions for potential future improvements.

2. Related works

In the context of predictive systems for traffic accidents, multiple studies have addressed issues related to data scarcity and improved accuracy in complex scenarios. Jin and Noh implemented a deep learning-based system to predict accidents in dense urban areas, where the high proportion of missing data poses a considerable challenge. Using historical mobility and driving behavior data, they achieved an accuracy of 94% and an F1 score of 0.72 [6]. This approach highlights the importance of integrating diverse data to improve predictions in high-density traffic environments. Baykal, Eriskin, and Terzi focused their study on the use of big data and machine learning to predict accident severity on highways, where high speeds and traffic conditions exacerbate the consequences. They used techniques such as XGBoost and Bayesian networks to model traffic characteristics, achieving a 30% improvement in prediction accuracy [7].

Abohassan, El-Basyouny, and Kwon analyzed the impact of adverse weather events, such as snowstorms, on pavement friction, showing that when it fell below 0.35 [8], the risk of collision increased considerably. Yang et al. combined statistical and machine learning models to anticipate accidents on highways in different areas, demonstrating that variables such as traffic volume and its standard deviation correlated with accident risk [9], emphasizing the need for area-specific approaches. Bokaba et al. compared machine learning classifiers in Gauteng, South Africa, and found that the random forest, combined with multiple imputations, offered the best performance with an accuracy of 97% [10].

However, long-term scalability was not addressed. Koo, Baek, and Chung employed an ensemble model based on weight feedback and MDG harmony, which achieved an accuracy of 96.86% [11], but presented limitations in environments with inconsistent data. Santos et al. investigated machine learning techniques in Setúbal, Portugal, between 2016 and 2019, using Random Forest to predict critical accident areas, although the generalization of the results was limited to that region [12]. Cao et al. proposed the MSGSGCN model to improve traffic speed prediction accuracy on irregular road networks, with a 4% increase in accuracy compared to other advanced models [13]. Lu et al. explored automatic incident detection using the STVDAE model, achieving a 26.3% improvement in detection accuracy and a processing speed eight times faster [14].

Finally, Yuan et al. focused on real-time accident risk prediction with limited data on the I-5 freeway in Washington, using support vector machines (SVM), achieving an accuracy of 78.7% [15], although without sufficiently addressing the model's applicability in other environments with dynamic traffic conditions.

3. System Design

A. Architecture

The proposed system for traffic accident prediction in Lima, Peru, follows an integrated architecture based on cloud services, real-time data processing, and frontend technologies accessible to the end user. The architecture design includes the following components:

Frontend: The frontend is a web application that allows users, including private entities and the general public, to access the platform from their browsers. Users can view prediction results for accident risk zones, generate reports, and query previous incidents. Access to the services is provided through an API that connects the frontend with the other system components.

AWS Services (Amazon Web Services): The platform uses AWS Lambda for data processing and Amazon SageMaker for training and executing the predictive model. These services allow the infrastructure to scale without the need to manage servers, facilitating the implementation of agile, high-performance solutions. Integration with other AWS services, such as Amazon S3 for data storage and API Gateway for component communication, enables the system to process large volumes of data efficiently.

Big Data: The Big Data component in this system focuses on adopting advanced tools to process large volumes of data from various sources related to traffic accidents. For this purpose, Apache Spark ecosystem tools are used, allowing efficient real-time and batch data processing, facilitating massive analysis.

Predictive Model: The predictive model is trained with historical accident data, weather conditions, and road infrastructure characteristics in Lima. This model is implemented in Amazon SageMaker, where machine learning algorithms are executed to predict possible traffic accidents. The architecture allows real-time information processing, enabling proactive alert generation.

Data Sources: The data comes from various sources, including public datasets such as the records from the National Road Safety Observatory (ONSV), data from the Ministry of Transport and Communications (MTC), and historical weather data. These data are stored and managed through cloud database services, allowing quick and secure access.

The architecture diagram (Fig. 1) provides a clear view of the system components and how they are interconnected

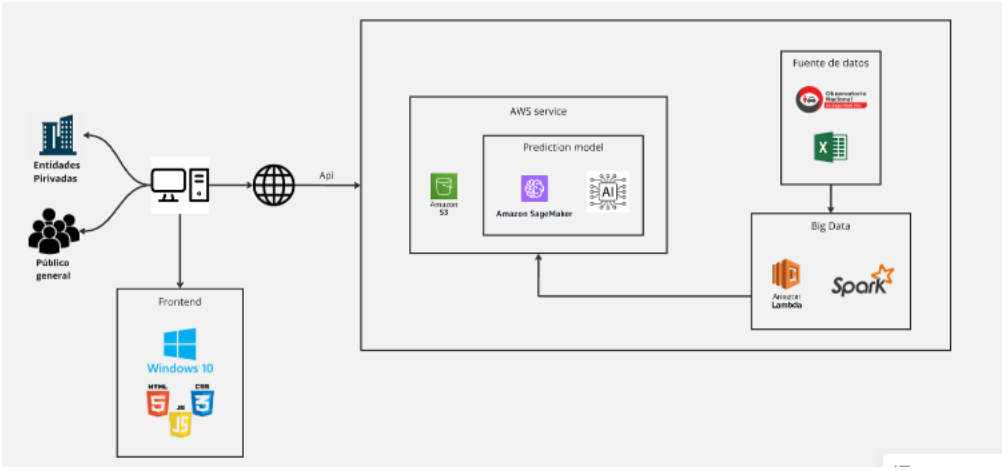

Fig. 1. Integrated Architecture Diagram

B. Methodology

1. Dataset

For the development of the predictive model for traffic accidents, various datasets have been used, which include key information to identify patterns and predict accidents in Lima. Among them, the ONSV dataset (2008-2023) records fatal accidents, detailing location, type of accident, and consequences. Additionally, data on people and vehicles involved in fatal incidents (2021-2023), the state of road infrastructure according to the MTC (2022), and historical weather data from Lima (2000-2014) were included to analyze the influence of climate. The conditions under which the incident occurred were also added, such as whether it was a holiday or in which time slot the accident happened.

The data is divided into three sets: training (80%), validation (10%), and testing (10%). The training set is used to teach the model, while the validation set adjusts its performance and prevents overfitting. Finally, the testing set evaluates the model's accuracy on unseen data. The cross-validation approach ensures robust evaluation and improves the model's predictive capability.

2. Model

The developed predictive system uses a machine learning model based on Random Forest to identify traffic accident patterns in Metropolitan Lima. This workflow is organized into several stages: data collection, preprocessing, feature extraction, model training, and prediction generation. The tool integrates different data sources, including accident history and climate variables, and uses cloud infrastructure (AWS) to manage data storage and processing.

Data cleaning was conducted by removing missing values and anomalies. Additionally, data augmentation was considered, evaluating whether the data could be enhanced using techniques like web scraping. A strict separation was applied between the training, validation, and test sets to ensure no examples were shared among these sets, preserving model integrity.

In the feature extraction stage, the most relevant variables in the dataset were identified. A key aspect of this phase was converting categorical variables to numerical ones, which was necessary for the machine learning model to process them properly. Features such as road type, weather, and time of day were transformed using techniques like one-hot encoding or ordinal encoding as appropriate. These numerical features were used to train the Random Forest model, improving accident prediction accuracy.

3. Training

The traffic accident predictive model was trained using the Random Forest algorithm, leveraging a diversified dataset with historical accident information, meteorological variables, and demographic data. To optimize performance, key hyperparameters were adjusted, such as the number of trees (n_estimators=100), maximum depth (max_depth=10), and a strategic feature selection using the Gini feature selection function. Data partitioning was done with 80% for training and 20% for testing, ensuring an adequate representation of input data in each set.

4. Evaluation and Statistical Analysis

To evaluate the model's performance, several classification metrics were used, such as accuracy, sensitivity, and specificity, focusing on achieving a balance between accurate accident detection and minimizing false alarms. Additionally, a 5-fold stratified cross-validation was conducted to ensure model stability. The results showed an average accuracy of 99.5% with a high F1 score in both classes of the target variable. The confusion matrix revealed a low false positive rate, thereby validating the model's effectiveness in traffic accident prediction contexts in Metropolitan Lima.

TABLA. I. METRICS

#	Metric	Description	Formula
1	Accuracy	Proportion of correct predictions out of all predictions. Evaluates the model's overall accuracy.	$\frac{TP + TN}{TP + TN + FP + FN}$
2	Recall	Proportion of true positives out of all actual positive cases. Important for accident detection.	$\frac{TP}{TP + FN}$
3	F1-Score	Harmonic mean of precision and recall, useful in imbalanced class scenarios.	$2 * \frac{Precision * Recall}{Precision + Recall}$
4	Specificity	Proportion of true negatives out of all actual negative cases. Evaluates the ability to avoid false positives.	$\frac{TN}{TN + FN}$

4. Results

The Random Forest model achieved perfect accuracy in the test set, obtaining values of 1.00 for precision, recall, and F1 score in both classes, which is remarkable in terms of balance and model robustness on the evaluated dataset. This performance is detailed in the following classification report:

Class 0 (No accident): The model achieved a precision, recall, and F1 score of 1.00 in the majority class of "No accidents," with 8,972 instances correctly classified and no errors.

Class 1 (Accident): In the "Accident" class, with fewer instances, the model maintained perfect precision and F1 score of 1.00, with only two false negatives. This indicates a minimal error rate in accident detection.

The confusion matrix confirms these results, showing an almost zero error rate in both classes, as seen in Fig. 2.

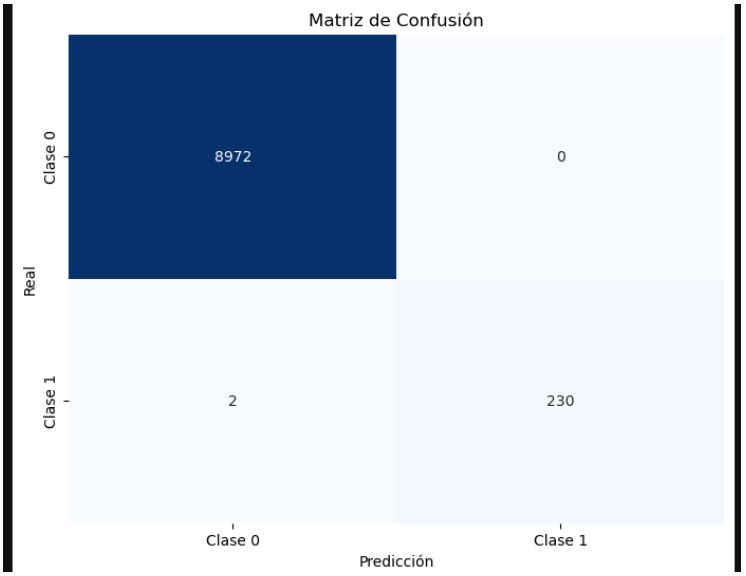

Fig. 2. Confusion Matrix

Additionally, feature importance in the model was evaluated, revealing that specific factors have a predominant influence on accident prediction. In the feature importance chart, categorical variables related to "Time of Day" were the most significant, highlighting:

Time Category (Night): As the most influential feature, it suggests a higher likelihood of accidents occurring at night.

Time Category (Morning) and Time Category (Afternoon): These were also relevant, indicating that time periods significantly impact accident occurrence.

Wind Direction (winddir) and Humidity (humidity): These meteorological variables showed a strong correlation, which could indicate adverse atmospheric conditions that increase risk.

Wind Gusts (windgust), Sea Level Pressure (sealevelpressure), and Moon Phase (moonphase): These factors also influence the model, suggesting that atmospheric conditions and visibility may indirectly affect accident occurrence.

According to Fig. 3, this analysis suggests that both time and atmospheric conditions are determinants in accident risk in Metropolitan Lima, particularly at night.

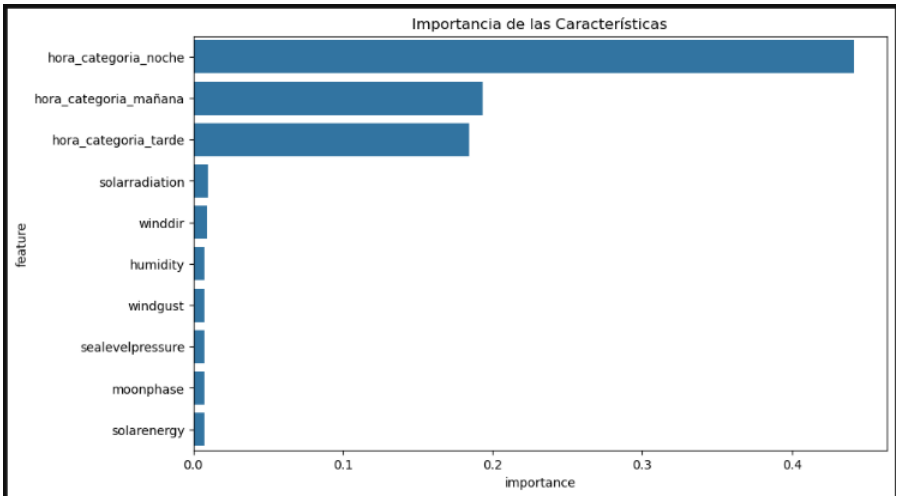

Fig. 3. Most Important Features

5. Discussion

The results reflect a high accuracy and robustness of the model for predicting accidents in Metropolitan Lima, especially in the districts included in the dataset. The evaluation metrics show excellent classification ability in both groups (accident and no accident), suggesting that the model is reliable within the current data context. However, certain aspects need to be strengthened to improve the model's applicability and generalization in a complex urban environment.

One critical issue observed is the geographical and temporal limitation of the dataset, which includes only certain districts of Lima and does not fully cover the daily variability across the entire city. This limitation increases the likelihood of prediction matches but does not guarantee that the model can effectively scale to a complete mapping of Metropolitan Lima. Incorporating additional data that more accurately represent the city’s overall dynamics, such as real-time traffic information and district-specific climate data, would enable a more granular and adaptable analysis.

Moreover, the obtained feature importance suggests that time and weather conditions significantly influence accident prediction, which aligns with previous studies identifying nighttime traffic and adverse conditions as risk factors. To reinforce this conclusion, it would be beneficial to expand the dataset and explore incorporating additional variables, such as vehicle type and specific characteristics of road infrastructure in each area, which would enhance the model's accuracy and usefulness in different contexts.

6. Conclusions

This work demonstrates the potential of the Random Forest model to predict traffic accidents in selected areas of Metropolitan Lima, achieving outstanding performance in terms of classification accuracy and balance. The model's structure and the data used provide a solid foundation for creating an accident monitoring system, particularly in an urban environment with high traffic density and variable road conditions.

However, the model's applicability would greatly benefit from a more extensive and representative dataset covering the entire city. Including district-level data and different time slots would optimize its accuracy and adaptability. As a future direction, it is recommended to work on developing a real-time data collection system and expanding the coverage of demographic, climate, and traffic information. This would enable a more comprehensive mapping of Metropolitan Lima and enhance the model's capability to provide accurate predictions in potential risk areas.

Acknowledgement

The authors are grateful to the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas for the support provided for this research work through the economic incentive.

References

1. Defensoría del Pueblo, Informe de Adjuntía N° 022-2022-DP/AMASPPI: La necesidad de contar con una Agencia Nacional de Seguridad Vial en el Perú. Defensoría del Pueblo, 2022.
2. H. Yang, X. Zhao, S. Luan, and S. Chai, "A traffic dynamic operation risk assessment method using driving behaviors and traffic flow Data: An empirical analysis," Expert Syst. Appl., vol. 249. Expert Syst. Appl., p. 123619.
3. Z. Jin and B. Noh, "From Prediction to Prevention: Leveraging Deep Learning in Traffic Accident Prediction Systems," Electronics. Electronics, Oct. 19, 2023.
4. J. Chipana Miranda, Factores que influyen en los accidentes de tránsito ocasionados por el transporte público terrestre en Villa El Salvador, 2021, Tesis de pregrado, Universidad Autónoma del Perú, 2023. Repositorio de la Universidad Autónoma del Perú.
5. Y. Yang, K. Wang, Z.- zhou Yuan, and D. Liu, "Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction," Journal of Advanced Transportation. Journal of Advanced Transportation, Apr. 30, 2022..
6. Z. Jin and B. Noh, "From Prediction to Prevention: Leveraging Deep Learning in Traffic Accident Prediction Systems," Electronics. Electronics, Oct. 19, 2023.
7. T. Baykal, F. Ergezer, E. Eriskin, and S. Terzi, "Accident Severity Prediction in Big Data Using Auto-Machine Learning," Scientia Iranica. Scientia Iranica, Feb. 22, 2023.
8. A. Abohassan, K. El-Basyouny, and T. Kwon, "Effects of Inclement Weather Events on Road Surface Conditions and Traffic Safety: An Event-Based Empirical Analysis Framework," Transportation Research Record, vol. 2676. Transportation Research Record, pp. 51–62, Apr. 29, 2022.
9. Y. Yang, K. He, Y. Wang, Z.- zhou Yuan, Y. Yin, and M. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," Physica A: Statistical Mechanics and its Applications. Physica A: Statistical Mechanics and its Applications, Feb. 01, 2022.

10. T. Bokaba, W. Doorsamy, and B. Paul, "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents," Applied Sciences. Applied Sciences, Jan. 14, 2022.
11. B.-K. Koo, J.-W. Baek, and K. Chung, "Weight Feedback-Based Harmonic MDG-Ensemble Model for Prediction of Traffic Accident Severity," Applied Sciences, vol. 11. Applied Sciences, p. 5072, May 30, 2021.
12. D. Santos, J. Saias, P. Quaresma, and V. Nogueira, "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction," Comput., vol. 10. Comput., p. 157, Nov. 24, 2021.
13. C. Cao, Y. Bao, Q. Shi, and Q. Shen, "Dynamic Spatiotemporal Correlation Graph Convolutional Network for Traffic Speed Prediction," Symmetry, vol. 16. Symmetry, p. 308, Mar. 05, 2024.
14. Y. Lu, Q. Lin, H. Chi, and J.-Y. Chen, "Automatic incident detection using edge-cloud collaboration based deep learning scheme for intelligent transportation systems," Applied Intelligence, vol. 53. Applied Intelligence, pp. 24864–24875, Jul. 29, 2023.
15. Z.- zhou Yuan, K. He, and Y. Yang, "A Roadway Safety Sustainable Approach: Modeling for Real-Time Traffic Crash with Limited Data and Its Reliability Verification," Journal of Advanced Transportation. Journal of Advanced Transportation, Jan. 15, 2022.