iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78253

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

RTAverse: A Machine Learning-Based Analysis and Forecasting of Road Traffic Accidents in Angeles City

Raven Y. Butial[1], Ivan T. David[2], Kyla Marie P. De Leon[3], Rianne Louisa R. Magno[4], Melissa M. Pantig[5]
College of Computer Studies Angeles University Foundation, Angeles City, Pampanga, Philippines

Abstract: ***Road traffic accidents (RTAs) in Angeles City remain a persistent public-safety concern, yet operational planning is often reactive because risk signals are not transformed into actionable, forward-looking evidence. This study presents RTAverse, a private web-based decision-support system that operationalizes official accident records into spatiotemporal forecasting and hotspot-level risk mapping for authorized stakeholders (e.g., LGU and traffic enforcement units). Following CRISP-DM, RTA records from Camp Tomas J. Pepito (2015–2024) were consolidated and preprocessed through automated cleaning (canonicalized headers, spatiotemporal deduplication, and barangay-name normalization), temporal transformation (datetime parsing and engineered season/time attributes), and feature reduction for modeling. Seven learning algorithms (Decision Tree, Random Forest, AdaBoost, XGBoost, k-NN, Naive Bayes, and SVM) were screened using error-based forecasting metrics; Random Forest and XGBoost achieved the lowest initial errors (MAE ≈ 0.22–0.25). Under sequential time-series evaluation, XGBoost produced the most consistent performance, and a Poisson-objective XGBoost achieved cross-validation MAE scores of 0.39, 0.19, 0.18, and 0.13 (overall MAE = 0.22), reflecting improved suitability for count-based outcomes. A hybrid variant strengthened spatial utility by integrating time-of-day clustering into hotspot forecasts, yielding absolute error 0–1 for 20 of 26 hotspots in the final period. Feature importance analysis indicated late-night time clusters as the strongest predictors (nearly 70% of the importance score), followed by rolling temporal trends. Expert evaluation using ISO/IEC 25010 and TAM affirmed the dashboard's usability and perceived usefulness. Overall, RTAverse demonstrates how privacy-preserving, localized accident data can be modeled as an evolving urban risk system and translated into practical forecasts that support preventive traffic-safety planning in Angeles City.***

Keywords: Road traffic accidents; forecasting; XGBoost; hotspot mapping; CRISP-DM; decision-support dashboard

I. INTRODUCTION

Road traffic accidents (RTAs) are a persistent public-safety problem that produce injuries, fatalities, and economic losses while placing continuous pressure on local governance and enforcement units. In urban settings, accident risk is rarely uniform; it emerges from interacting factors such as mobility patterns, road design, land-use activity, enforcement routines, and time-dependent human behavior. These interactions can produce non-linear changes over time and localized concentrations of risk, where a small set of locations and time windows repeatedly accounts for a disproportionate share of incidents. In Angeles City, the availability of historical accident records provides an opportunity to move beyond after-the-fact reporting and toward risk-aware planning grounded in evidence. Conventional approaches in local settings often emphasize post-incident documentation and reactive deployment of resources. While these practices are necessary for accountability, they do not directly support prevention when risk is evolving across weeks, seasons, and hotspots. Forecasting systems that anticipate when and where accidents are likely to occur can enable preemptive strategies such as targeted enforcement, time-specific advisories, and location-focused interventions. However, implementing forecasting for local decision-making requires more than model selection; it requires a reproducible data pipeline, evaluation designs that respect temporal order, and outputs that translate into actionable signals at the hotspot level.

This study introduces RTAverse, a machine learning–based forecasting and visualization system designed specifically for Angeles City. RTAverse is intended for authorized administrators (e.g., LGU and police traffic management officers) and includes an integrated preprocessing and retraining workflow that standardizes new uploads and refreshes forecasting outputs as records are updated. The study is guided by the following objectives: (1) compare multiple machine learning models and identify the most accurate and reliable forecasting approach for the local dataset; (2) forecast high-risk accident conditions across time and location; (3) develop a dashboard that visualizes accident trends and predicted hotspot risks; (4) identify and rank key contributing factors; and (5) assess system quality and acceptability using ISO/IEC 25010 and the Technology Acceptance Model (TAM).

The primary contributions of this work are: (a) a consolidated and cleaned accident dataset for Angeles City (2015–2024) prepared for temporal and spatial forecasting; (b) an empirical comparison of candidate models and forecasting configurations, including count-aware XGBoost with a Poisson objective and a hybrid variant integrating temporal clustering; and (c) a private operational dashboard that presents forecasts and hotspot risk classifications to support data-driven decision-making. Overall, the study positions localized accident forecasting as an operational way to understand and manage a changing urban risk system, where spatiotemporal patterns and hotspot dynamics can be monitored and acted upon.

A. Related Work

Recent studies demonstrate the use of machine learning to analyze and predict road safety outcomes across spatial and temporal dimensions. GIS-integrated approaches have been applied to identify and forecast accident hotspots and generate actionable mapping outputs (Agoylo, 2024; Amorim et al., 2023). Other work combines predictive modeling with spatial network analysis to support safer routing and resource allocation (Berhanu et al., 2024). Temporal pattern discovery is commonly supported through time-series visualization and supervised learning, with studies often reporting distinct high-risk periods at night and on weekends, motivating the inclusion of time-of-day features (Ackaah et al., 2020). Reviews also emphasize that temporal signals may be underutilized without appropriate feature engineering and evaluation designs (Behboudi et al., 2024; Silva et al., 2020). Ensemble methods such as Random Forest and boosting-based models are frequently favored for their robustness on heterogeneous crash datasets, while clustering is sometimes used to support interpretable risk groupings (Assi et al., 2020). In the Philippine context, similar accident analytics efforts commonly focus on Metro Manila, leaving a geographic gap for other urban centers. This study addresses that gap by developing a localized forecasting and visualization system for Angeles City using official records and an implementation-oriented pipeline that supports periodic retraining and secure administrative access.

II. METHODS

A. Research Design and CRISP-DM Framework

This study developed RTAverse, a privacy-preserving, machine learning-based dashboard for analyzing and forecasting road traffic accidents in Angeles City. Using official accident records and an automated preprocessing and retraining workflow, the system supports both temporal forecasts and hotspot risk visualization for authorized stakeholders.

Model comparisons showed that XGBoost is the most suitable algorithm for integration into the forecasting system, achieving the lowest errors and highest goodness-of-fit in sequential evaluation. A Poisson-objective configuration improved count-data modeling, while a hybrid variant enhanced hotspot-level utility by incorporating time-of-day clustering signals.

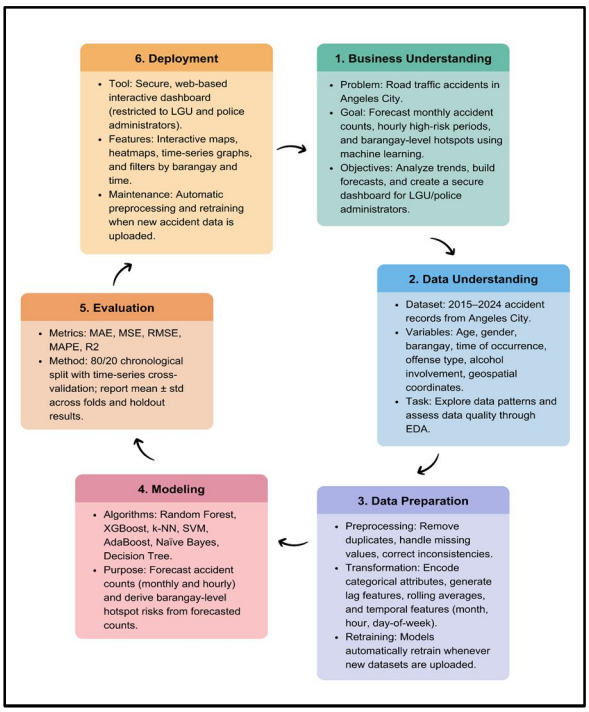


Figure 1. CRISP-DM framework used to guide the RTAverse development lifecycle.

B. Data Source and Ethics

The primary dataset was obtained from the Camp Tomas J. Pepito database maintained by the City Investigation and Detective Management Unit in Angeles City. The records span 2015-2024 and include incident-level attributes capturing temporal, spatial, demographic, and contextual information.

Because incident records are sensitive, RTAverse was designed as a private dashboard for authorized administrators only. Data handling emphasized confidentiality, access control, and secure processing, and the system is not intended for public access.

The study also noted that bias may arise from incomplete or inconsistently reported attributes (e.g., missing timestamps or coordinates). Automated preprocessing and standardized encoding were applied to reduce distortions in downstream forecasts.

C. Data Description

The source data were provided as multiple Excel workbooks with standardized column structures, including accident records and related vehicle and party information. For 2015-2024, the accident workbook contained 4,903 rows; related workbooks contained 8,131 and 4,913 rows for other involved parties and vehicle data, respectively. An additional workbook covering January-August 2025 was also provided for system updating.

Key variables used in analysis included age, gender, barangay, police station identifier, suspect and victim counts, date and time committed, offense type under Article 365 of the Revised Penal Code, alcohol involvement indicator, latitude, longitude, and an engineered season attribute based on PAGASA definitions (rainy: June-November; dry: December-May).

D. Data Preprocessing

Preprocessing was implemented as an automated pipeline to support periodic retraining when new records are uploaded. Cleaning removed spreadsheet artifacts, standardized column headers, normalized categorical inputs, and corrected barangay name inconsistencies.

Missing value placeholders were converted to proper null entries. Records with missing coordinates were dropped to preserve spatial accuracy in hotspot mapping.

Deduplication grouped incidents by spatiotemporal keys (date, time, and coordinates) and consolidated multiple rows describing the same incident into single entries. Date and time fields were converted into typed formats and expanded into temporal attributes (year, month, weekday).

E. Feature Engineering and Hotspot Risk Categorization

The modeling pipeline incorporated temporal features, lag features (prior months), and rolling statistics (e.g., a three-month rolling mean) to capture sequential dependencies in accident counts.

For spatial utility, incidents were aggregated into hotspot clusters based on geolocation and forecast outputs were translated into operational risk categories using percentile-based thresholds.

Risk category	Threshold (percentile basis)	Map color
No Risk	-	Grey
Low	0 > and <= 50th percentile (median)	Green
Moderate	> 50th and <= 75th percentile	Orange
High	> 75th percentile	Red

Table 1. Percentile-based hotspot risk categorization used for map visualization in RTAverse.

F. Modeling and Evaluation

Seven algorithms were compared: Decision Tree, Random Forest, AdaBoost, XGBoost, k-NN, Naive Bayes, and SVM. Models used consistent feature inputs for fair comparison.

Baseline screening used an 80/20 training-test split and MAE. Sequential forecasting evaluation used MAE, MSE, RMSE, MAPE, and R-squared with time-series cross-validation to preserve chronological order.

An XGBoost configuration with a Poisson regression objective was evaluated for count-based outcomes, and a hybrid variant integrated time-of-day clustering to improve hotspot-level interpretability.

G. RTAverse Dashboard

RTAverse is a web-based dashboard that supports secure administrative access, interactive filtering, and visual analytics. Outputs include time-based charts, location summaries, and a map highlighting forecasted hotspot risks.

The dashboard provides a dataset upload workflow that triggers automated preprocessing and model retraining, enabling the system to remain current as additional records are provided.

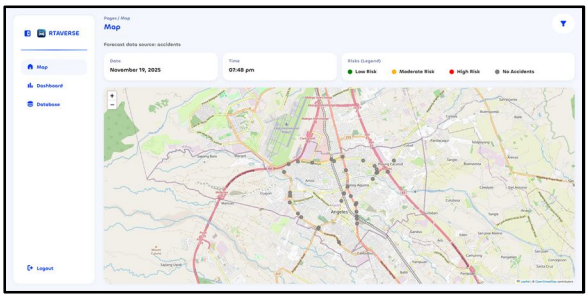


Figure 2. RTAverse - Map/Hotspot

Figure 3. RTAverse - Dashboard

III. RESULTS AND DISCUSSION

A. Baseline Algorithm Comparison

Initial baseline screening using MAE clearly separated high-performing ensemble methods from weaker baselines. Random Forest and XGBoost achieved the lowest MAE values (approximately 0.22–0.25), indicating strong predictive capacity on the prepared feature set. The Decision Tree model produced moderate performance (MAE ≈ 0.35), while k-NN and Naive Bayes showed higher errors (≈ 0.45–0.55).

AdaBoost exceeded 1.0 MAE, and SVM exceeded 1.6 MAE, making them less suitable for deployment under the current data configuration. These results support prioritizing ensemble learners—particularly boosting and bagging methods—when forecasting accident counts from heterogeneous, engineered temporal features.

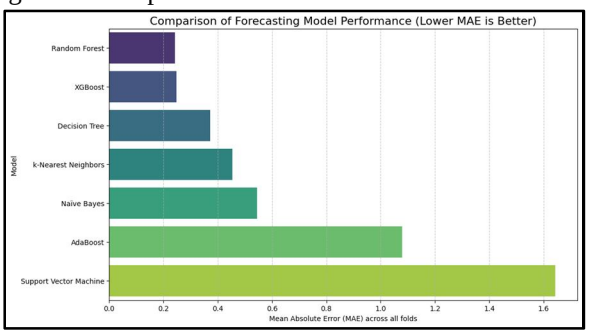


Figure 4. Baseline model comparison using MAE during initial training (Part 1).

B. Time-Series Forecasting Evaluation

To evaluate performance under realistic forecasting conditions, sequential time-series evaluation was conducted using MAE, MSE, RMSE, MAPE, and R-squared. XGBoost consistently achieved the lowest error values and the highest R-squared, with Random Forest following closely. Focused evaluation among top candidates confirmed XGBoost as the strongest performer once temporal structure was explicitly represented through engineered features such as time clusters and lagged months. This result indicates that modeling the temporal organization of crash events improves forecasting stability and helps the model generalize across shifting weekly patterns.

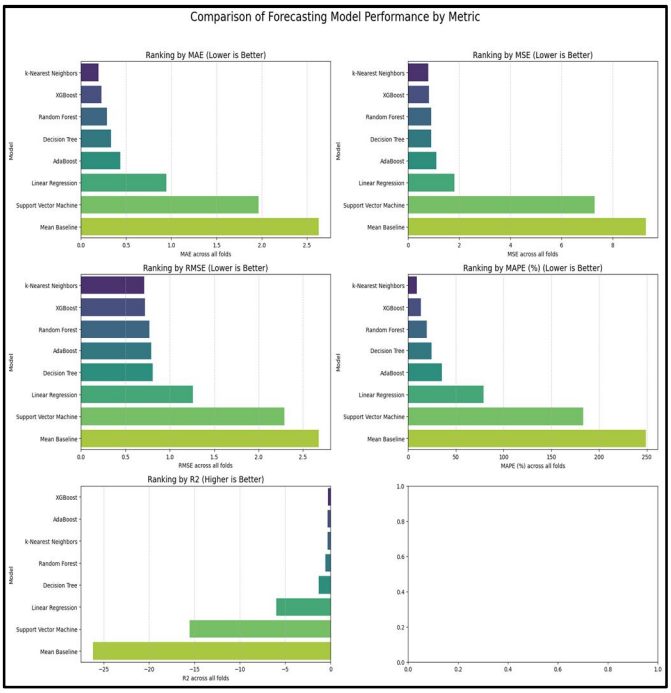


Figure 5. Focused comparison among top-performing models using multiple metrics (Part 3).

C. Random Forest Weekly Forecasting

A Random Forest configuration trained on weekly accident counts achieved MAE = 1.60 and RMSE = 1.95. Errors remained low for low- to moderate-frequency hotspots, but the model systematically underpredicted in high-frequency clusters. This behavior is consistent with smoothing effects observed in ensemble averaging, where extreme values are pulled toward the mean. Operationally, this underprediction matters because extreme-count hotspots often correspond to locations where preventive action is most urgent and where forecast bias can reduce situational preparedness.

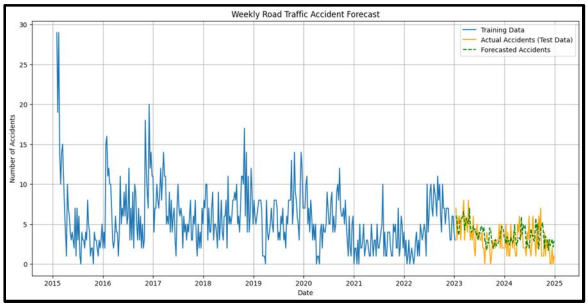


Figure 6. Weekly road traffic accident forecast using Random Forest.

D. XGBoost Forecasting and Hybrid Variant

A count-aware XGBoost configuration using a Poisson regression objective (three lag features, a three-month rolling mean, and seasonal variables) produced cross-validation MAE scores of 0.39, 0.19, 0.18, and 0.13, with an overall MAE = 0.22. While some underestimation persisted in extremely high-count hotspots (e.g., 115 actual incidents forecasted as 67), the Poisson objective improved alignment with count-based outcomes and reduced systematic bias compared with non-count-aware alternatives. A hybrid variant further integrated time-of-day cluster aggregates alongside lag and seasonal features. Although its four-fold cross-validation MAE values (1.17, 1.13, 0.32, 0.22; overall MAE = 0.71) were higher in aggregate, it demonstrated strong near-term spatial utility: in the final period, 20 of 26 hotspots achieved absolute error 0–1, supporting actionable hotspot-level forecasting where decision-making is location-focused. This highlights an important operational trade-off: models optimized for global error may differ from models optimized for hotspot-level interpretability and deployment readiness.

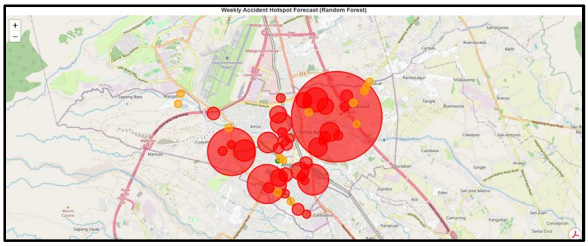


Figure 7. Weekly accident hotspot forecast map (example visualization).

E. Key Factors Influencing Accident Forecasts

Feature importance analysis indicates that time-based variables and rolling temporal trends are the dominant predictors of accident occurrence. TIME_CLUSTER_Midnight contributed nearly 70% of the importance score, followed by the three-month rolling mean. Other time clusters contributed meaningful additional signal, while lag features and seasonal indicators provided smaller but relevant contributions. These results suggest that late-night periods represent a consistent risk regime in the dataset, and that recent trend behavior (rolling means) captures evolving system conditions that influence near-term forecasting.

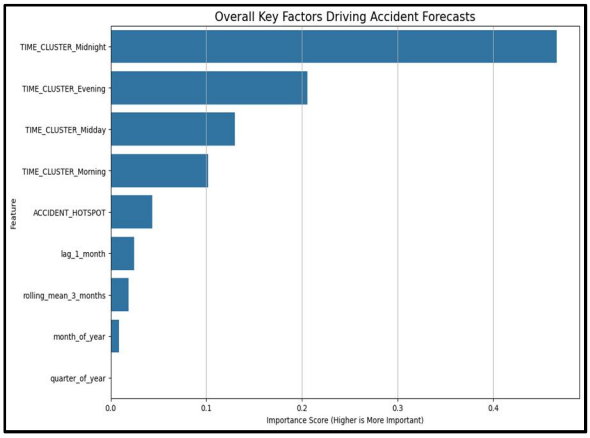


Figure 8. Feature importance summary for key predictors in the forecasting model.

F. System Quality and Acceptability

Expert assessment using ISO/IEC 25010 and TAM indicated that the dashboard interface and workflows were straightforward and that the visualizations were readable and effective in communicating actionable insights. The evaluation supports strong perceived ease of use and perceived usefulness for authorized stakeholders, indicating readiness for operational use in local safety planning contexts.

Across experiments, XGBoost and Random Forest outperformed simpler baselines, with XGBoost providing the most consistent results under sequential evaluation. Random Forest underprediction in high-frequency hotspots highlights the operational importance of modeling extreme counts, while the Poisson objective improved XGBoost's suitability for count data. The hybrid XGBoost variant strengthened hotspot-level interpretability and short-horizon spatial utility, illustrating that decision-support deployments may prioritize hotspot accuracy alongside aggregate error metrics. The cleaned historical dataset contained 2,780 accident records and was sufficient to validate core forecasting and dashboard functions; however, identified gaps (e.g., limited pedestrian and self-accident records) motivate future data enrichment and broader contextual integration.

The concentration of incidents within specific time clusters and locations suggests the presence of stable and transitional risk regimes within the urban traffic system. Future work may model these regimes using dynamical systems or network-based approaches to better characterize tipping behaviors and structural transitions in urban accident risk.

IV. CONCLUSION AND RECOMMENDATIONS

RTAverse demonstrates how official accident records can be transformed into a privacy-preserving forecasting and visualization dashboard for Angeles City that supports authorized stakeholders in identifying temporal patterns and actionable hotspots. Across experiments, XGBoost and Random Forest consistently outperformed simpler baselines. XGBoost produced the most stable results under sequential evaluation, and a Poisson-objective configuration aligned better with count-based crash data, achieving an overall MAE of 0.22. While Random Forest tended to underpredict in high-frequency areas, the hybrid XGBoost variant improved hotspot-level interpretability and near-term spatial utility, supporting operational usebeyond aggregate accuracy. The cleaned dataset contained 2,780 historical records and was sufficient for model validation, but remaining gaps motivate further enrichment.

A. Recommendations

Future work should (1) expand and standardize data coverage (especially for underrepresented cases and incomplete fields), (2) refine hotspot calibration to better capture extreme/high-frequency clusters, (3) integrate additional contextual factors (e.g., traffic exposure, road attributes, weather or event indicators where available), (4) automate ingestion and periodic retraining to keep forecasts current, and (5) conduct broader operational evaluations to measure decision impact in real deployments.

B. Data Availability

The dataset used in this study was obtained from official local government documentation and is not publicly available due to confidentiality and data privacy restrictions.

C. Acknowledgement

The researchers are extremely grateful to all the people who have been part of this study in one way or another

REFERENCES

[1] Ackaah, W., Apuseyine, B. A., & Afukaar, F. K. (2020). Road traffic crashes at night-time: Characteristics and risk factors. International Journal of Injury Control and Safety Promotion, 27(3), 392-399. https://doi.org/10.1080/17457300.2020.1785508

[2] Agoylo, J. C. (2024). GIS-based traffic accident hotspot prediction using machine learning. International Journal of Advanced Research in Computer Science, 15(2), 45–53. https://doi.org/10.22541/au.173347433.37543456/v1

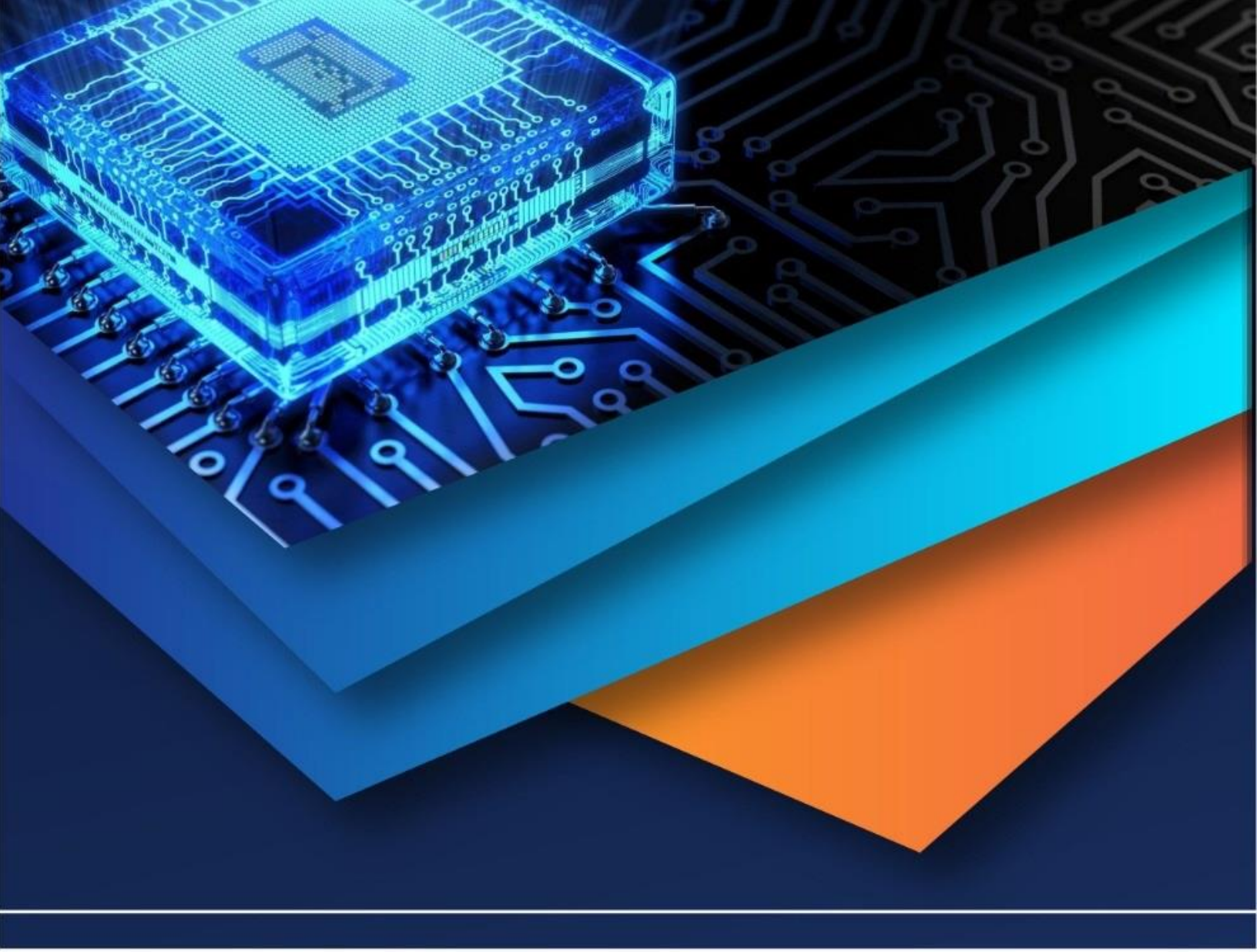
[3] Al-Mistarehi, B. W., Alomari, A. H., Imam, R., & Mashaqba, M. (2022). Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS. Frontiers in Built Environment, 8. https://doi.org/10.3389/fbuil.2022.860805

[4] Amorim, B. D. S. P., Firmino, A. A., Baptista, C. D. S., Júnior, G. B., Paiva, A. C. D., & Júnior, F. E. D. A. (2023). A machine learning approach for classifying road accident hotspots. ISPRS International Journal of Geo-Information, 12(6), 227. https://doi.org/10.3390/ijgi12060227

[5] Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. International journal of environmental research and public health, 17(15), 5497. https://doi.org/10.3390/ijerph17155497

[6] Berhanu, Y., Schröder, D., Wodajo, B. T., & Alemayehu, E. (2024). Machine Learning for Predictions of Road Traffic Accidents and Spatial Network Analysis for Safe Routing on Accident and Congestion-Prone Road Networks. Results in Engineering, 23, 102737. https://doi.org/10.1016/j.rineng.2024.102737

[7] Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques.
[8] arXiv preprint arXiv:2406.13968. https://doi.org/10.48550/arXiv.2406.13968
[9] Silva, P. B., Andrade, M., & Ferreira, S. (2020). Machine learning applied to road safety modeling: A systematic literature review. Journal of traffic and transportation engineering (English edition), 7(6), 775-790. https://doi.org/10.1016/j.jtte.2020.07.004
[10] Datu, N. H. (2023, March). Road traffic accidents analysis using association rule mining and descriptive analytics. In AIP Conference Proceedings (Vol. 2508, No. 1). AIP Publishing. https://doi.org/10.1063/5.0117371
[11] Dong, C., & Chang, N. (2023). Overview of the identification of traffic accident-prone locations driven by big data. Digital Transportation and Safety, 2(1), 67-76. https://doi.org/10.48130/DTS-2023-0006
[12] Pitarque, A., & Guillen, M. (2022). Interpolation of quantile regression to estimate driver's risk of traffic accident based on excess speed. Risks, 10(1), 19. https://doi.org/10.3390/risks10010019
[13] Quistberg, D. A., Hessel, P., Rodriguez, D. A., Sarmiento, O. L., Bilal, U., Caiaffa, W. T., ... & Roux, A. V. D. (2022). Urban landscape and street-design factors associated with road-traffic mortality in Latin America between 2010 and 2016 (SALURBAL): an ecological study. The Lancet Planetary Health, 6(2), e122-e131. https://doi.org/10.1016/S2542-5196(21)00323-5
[14] Dorado, D., & Aviles, J. (2024, July). Machine Learning Regression Model Development and Data Visualization of Road Accident in Urdaneta City, Pangasinan, Philippines. In Proceedings of the 2024 6th Asia Conference on Machine Learning and Computing (pp. 27-32). https://doi.org/10.1145/3690771.3690785
[15] Libnao, M., Misula, M., Andres, C., Mariñas, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve bayes algorithm. Procedia Computer Science, 227, 316-325. https://doi.org/10.1016/j.procs.2023.10.530



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY