Predicting Traffic Collision Occurrence and Severity

Keyur Thakkar
Arizona State University
Tempe, Arizona, USA
kthakk14@asu.edu

Ian McDonough
Arizona State University
Tempe, Arizona, USA
itmcdono@asu.edu

Shagufta Memon
Arizona State University
Tempe, Arizona, USA
smemon5@asu.edu

Simran Panchal
Arizona State University
Tempe, Arizona, USA
spanch12@asu.edu

ABSTRACT

In an era where traffic collisions continue to pose significant challenges to urban safety, leveraging advanced data mining techniques offers a promising avenue to enhance emergency response systems and mitigate collision impacts. This project, conducted as part of the Data Mining (CSE 572) course in Spring 2024, employs predictive modeling and comprehensive data analysis to predict the occurrence and severity of traffic collisions in Tempe. By integrating sophisticated algorithms and a rigorous data preprocessing approach, we aim to provide actionable insights that can guide emergency services to respond more effectively and implement preventive measures. Our methodology encompasses a meticulous selection of supervised learning models, enhanced through hyperparameter tuning and balanced by addressing data imbalances. The predictive framework developed not only showcases the potential of data mining in public safety applications but also serves as a critical step towards minimizing the economic, societal, and human toll of traffic accidents. This report encapsulates our journey through data collection, model evaluation, and the synthesis of findings into practical recommendations, reflecting a profound engagement with the intricacies of traffic safety and data science.

ACM Reference Format:
Keyur Thakkar, Ian McDonough, Shagufta Memon, and Simran Panchal. 2024. Predicting Traffic Collision Occurrence and Severity. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This report is presented as a culmination of the Data Mining (CSE 572) course undertaken in the Spring semester of 2024. At the core of this project is the urgent global challenge of traffic collisions—a multifaceted issue that continues to claim lives and impact societies despite advancements in technology and heightened safety awareness. The city of Tempe serves as the focal point for our investigation, reflecting a microcosm of the broader global struggle against the dangers posed by urban traffic systems and the unpredictable nature of human behavior on the road. The relentless pace of traffic accidents, coupled with their profound societal, economic, and emotional toll, underscores the imperative for innovative solutions. Our project seeks to bridge the gap between data science and public safety, leveraging predictive analytics to shed light on the patterns and predictors of traffic collisions. By harnessing a rich dataset sourced directly from the City of Tempe's official records, we embark on a data-driven exploration aimed at understanding the nuances of traffic collision occurrences and their severity. The intricacies of urban traffic dynamics, coupled with limited emergency response resources and the inherent unpredictability of accidents, present significant challenges. These obstacles necessitate a proactive and informed approach to emergency response and accident prevention. Our methodology, centered around a comprehensive data mining pipeline, is designed to distill actionable insights from complex datasets. Through predictive modeling and rigorous data analysis, we aspire to equip authorities with the knowledge needed to implement effective preventative measures and enhance the efficiency of emergency responses. This report outlines our approach, from the initial stages of data preprocessing and model selection to the in-depth analysis and interpretation of our findings. It reflects a concerted effort to advance traffic safety in Tempe, contributing to the creation of a safer, more resilient urban environment. The implications of our work extend beyond the immediate context, offering a blueprint for leveraging data mining techniques in the service of public safety and community welfare.

2 RELATED WORK

In the realm of traffic safety, the application of machine learning (ML) techniques to predict traffic collision occurrence and severity has gained significant attention. This part of the report synthesizes findings from a series of pertinent studies, shedding light on various methodologies, strengths, and limitations of current approaches in traffic accident analysis and prediction. These studies collectively underscore the potential of ML to enhance predictive accuracy, improve emergency response strategies, and ultimately contribute to safer roadways.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.1 Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers

A pivotal study by Madushani et al. (2023) [4] presents an innovative approach to evaluating expressway traffic crash severity, integrating logistic regression with advanced machine learning classifiers, including XGBoost, Random Forest, Decision Tree, and KNN. This research is especially relevant to our project as it aligns with our objective of leveraging machine learning techniques to predict traffic collision occurrence and severity. Madushani and colleagues distinguish their methodology by utilizing explainable AI methods, such as SHAP and LIME, to interpret the influence of various factors on crash severity. Their findings underscore the superior predictive capabilities of machine learning models over traditional logistic regression, particularly highlighting XGBoost's exceptional performance.

This study's strengths lie in its effective combination of logistic regression with machine learning for enhanced predictive accuracy and its innovative use of explainable AI techniques to provide insights into the underlying factors influencing crash severity. The transparency and interpretability of the models through SHAP and LIME analyses offer a valuable framework for our project, especially in understanding and communicating the impact of different variables on traffic collision outcomes.

However, Madushani et al.'s research also presents limitations, notably its focus on expressway conditions, which may not fully encapsulate the urban traffic environments central to our study. Furthermore, the exclusion of certain environmental characteristics and driver behavior aspects suggests that their model may benefit from incorporating a wider array of influencing factors for a more comprehensive analysis.

Despite these limitations, Madushani et al.'s work contributes significantly to the body of knowledge on applying machine learning to traffic safety analysis. Their methodology and findings provide a solid foundation for our project, particularly in advocating for the use of advanced machine learning techniques and explainable AI to enhance predictive accuracy and model interpretability. By building upon their approach and extending the analysis to include a broader range of conditions and factors relevant to urban traffic systems, our project aims to further the application of machine learning in improving traffic collision prediction and prevention strategies.

2.2 Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents

Obasi and Benson (2023) [5] conducted a comprehensive study employing machine learning models to forecast the severity of traffic accidents based on contributing factors, analyzing ten years of UK traffic accident data from 2005 to 2014. Their methodology leveraged Naive Bayes, Random Forest, Logistic Regression, and Artificial Neural Networks, highlighting the predictive strength of machine learning over traditional statistical models. Their work particularly noted the superior accuracy of Random Forest and

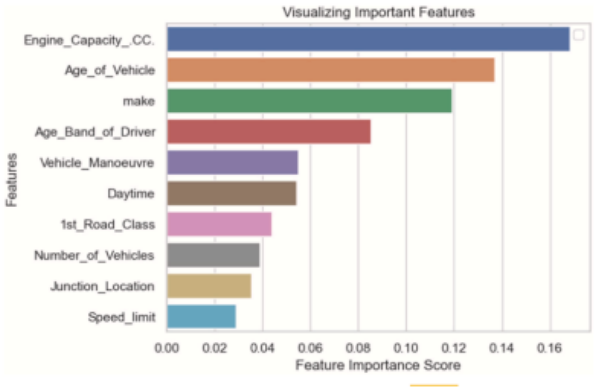

Figure 1: Feature Importance Analysis

	Predicted:Non-Serious	Predicted:Serious
True:Non-Serious	T:Non-Serious	F:Serious
True:Serious	F:Non-Serious	T:Serious

Table 1: Confusion Matrix for Binary Classification

Logistic Regression models, which achieved an 87% prediction accuracy, significantly outperforming Naive Bayes and Artificial Neural Networks.

A notable contribution from their research includes the use of Random Forest-based feature importance analysis to identify critical variables influencing traffic accident severity prediction. The analysis underscored Engine Capacity, Age of the vehicle, make of the vehicle, Age of the driver, vehicle manoeuvre, daytime, and 1st road class as pivotal factors. Figure 1 in the paper visually represents the variable importance plot generated using Random Forest Feature Selection, illustrating the significant impact of variables like Engine Capacity and Age of the vehicle on the prediction of traffic accident severity. This figure underscores the utility of machine learning in identifying and prioritizing factors for accident severity prediction. Model Evaluation Metrics: Their study outlines essential model evaluation metrics derived from a confusion matrix, including accuracy, precision, recall, and F1-score, providing a structured framework for assessing machine learning models' performance.

The confusion matrix format they employed (Table 1) offers a binary classification of predicted versus actual outcomes, serving as a fundamental tool in evaluating predictive accuracy.

Predictive Performance Results: The predictive results presented in their study offer a comprehensive comparison of the tested models. The Random Forest model notably outperformed others, achieving the highest weighted average recall, precision, and F-score, substantiating its effectiveness in traffic accident severity prediction.

Their research contributes significantly to understanding traffic accident severity prediction, emphasizing the effectiveness of machine learning models like Random Forest. By identifying key variables and employing robust evaluation metrics, their study provides valuable insights for enhancing traffic safety strategies and improving predictive accuracy in accident severity assessment.

2.3 Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction

Santos et al. (2021) [6] conducted a study leveraging machine learning techniques to analyze traffic accident data and forecast accident hotspots in Portugal's Setúbal district. Their research identified key factors such as pedestrian accidents and motorcycle involvement as significant contributors to predicting accident severity. The team developed a predictive model utilizing the random forest algorithm, achieving a commendable 73% accuracy rate in anticipating accidents, highlighting the potential of machine learning in enhancing road safety.

One notable strength of the study is its successful identification of influential factors in accident severity through machine learning algorithms. The integration of historical weather data with accident records further improved prediction accuracy, showcasing a holistic approach to analyzing accident patterns. Additionally, the use of clustering techniques facilitated rule generation, aiding in the comprehensive analysis of accident trends and patterns within the dataset.

However, the study also faces certain limitations. The relatively low sensitivity (0.08) in the predictive model indicates room for improvement in accurately forecasting accidents. Moreover, the restricted scope of data covering the years 2016 to 2019 may not capture recent trends and changes in traffic patterns, necessitating ongoing data collection and analysis. Furthermore, while clustering techniques assist in rule generation, the generated rules should undergo rigorous testing for generalization to the entire dataset and real-world applicability, ensuring the reliability and robustness of the predictive model.

2.4 Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh

M. F. Labib and colleagues' (2019) [2] research delves into a thorough analysis of traffic accidents in Bangladesh, employing machine learning methodologies like Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and AdaBoost. The study's primary objective is to classify accident severity into categories such as Fatal, Grievous, Simple Injury, and Motor Collision, with AdaBoost demonstrating high accuracy in this task. The research offers valuable insights into critical factors influencing accidents, including vehicle types, road conditions, junction types, and the time of occurrence, providing a nuanced understanding of road safety dynamics.

One notable strength of the study lies in its utilization of advanced machine learning algorithms, which enables a more precise and insightful analysis compared to traditional statistical methods. Furthermore, the research provides actionable recommendations, such as implementing a recommender system to predict accidents and alert road users, potentially mitigating accident rates. The comprehensive examination of various factors impacting accident severity, including vehicle types, road conditions, and time of day, contributes to a holistic approach to road safety analysis, enhancing the study's applicability and relevance.

However, the study acknowledges certain limitations that warrant consideration. The reliance on historical data up to 2015 may restrict the generalizability of the findings to current traffic conditions, highlighting the importance of utilizing updated datasets for more accurate predictions and recommendations. Additionally, while achieving high accuracy with AdaBoost, the paper acknowledges the need for further refinement of the approach, particularly in terms of sensitivity and false positive rate, to enhance its predictive capabilities. Furthermore, while the study primarily focuses on categorizing accident severity, exploring additional factors such as weather conditions, driver behavior, and road infrastructure could enrich the analysis for a more comprehensive understanding of accident rates and patterns.

2.5 An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors

Zhengjing Ma, Gang Mei, and Salvatore Cuomo (2021) [5] present an innovative analytic framework for predicting traffic accident injury severity based on contributing factors. The framework integrates CatBoost and Shapley value for feature analysis, k-means clustering for spatial pattern capturing, and SSAE deep learning for injury severity prediction. The study's key contributions include providing insights into the importance of contributing factors in injury severity prediction through feature analysis and demonstrating the effectiveness of the proposed framework, particularly the SSAE model, in outperforming baseline models.

The utilization of CatBoost and Shapley value for feature analysis stands out as a strength of the study, offering valuable insights into the significance of different factors in predicting injury severity levels. Additionally, the integration of k-means clustering helps capture spatial patterns in accident data, which enhances the modeling process and improves the accuracy of predictions. The superior performance of the SSAE-based deep learning model further solidifies the robustness of the proposed framework, showcasing its potential in accurately predicting injury severity.

However, the study acknowledges several limitations that require consideration. The lack of detailed information on vulnerable road users (VRUs) limits the applicability of the framework for VRU safety analysis. Additionally, the limited discussion on the selection process and parameters of machine learning approaches may hinder reproducibility and generalizability. Furthermore, the study does not address potential biases or uncertainties in the data, which could impact the accuracy of injury severity predictions and warrant further investigation for a more comprehensive analysis.

2.6 Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities

Yubian Wang and Wei Zhang (2017) [7] conducted a study to evaluate the impacts of roadway and environmental factors on traffic crash severities using logistic regression modeling. Their analysis identified several key factors influencing crash severity, including road function class, crash location, road alignment, light conditions, road surface condition, and speed limit. The study found that higher crash severity is often associated with rural roadways, major arterials, non-intersection locations, curved road sections, dark conditions without street lights, dry road surfaces, and high speed limits.

One of the study's strengths lies in its use of a large dataset from the SHRP2 study, allowing for comprehensive insights into traffic crash severities. By employing logistic regression modeling, a robust statistical method, the researchers were able to quantify the impacts of various factors on crash severity with a high degree of reliability. The study also provides clear recommendations for traffic engineers and planners, highlighting the importance of prioritizing certain roadway types and environmental conditions to mitigate crash severities effectively.

However, the study has several limitations that need to be considered. Firstly, the reliance on crash data from Florida may limit the generalizability of the findings to nationwide crash patterns and severities. Secondly, the focus on infrastructure-related factors might overlook the complex interactions with human and vehicle-related factors, which also play a crucial role in crash severities. Lastly, the recommendations proposed in the study may require further validation and testing in real-world scenarios beyond the scope of the initial analysis to ensure their practical effectiveness and feasibility.

2.7 A study on traffic crash severity prediction using machine learning algorithms

Türker and Gündüz (2023) [1] conducted a study focused on predicting traffic accident severity using various machine learning algorithms. They employed Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes algorithms with real-time data. Notably, the Decision Tree algorithm exhibited outstanding performance, achieving a remarkable accuracy rate of 99.6%. The study highlights the crucial role of advanced computational models in accurately predicting accident severity and implementing effective preventive measures to reduce traffic accidents.

One of the notable strengths of the study is its comprehensive analysis using a wide range of machine learning algorithms. The high accuracy rates achieved in predicting accident severity, especially with the Decision Tree algorithm, underscore the efficacy of these advanced predictive methods. The emphasis on the significance of employing such models for targeted interventions aimed at reducing traffic accidents is a significant contribution of this research.

However, the study also has some limitations. It lacks detailed discussion on the specific features or variables that contributed most significantly to the accuracy of the predictive models. Additionally, there's a gap in the analysis regarding the impact of different hyperparameters or model tuning techniques on the performance of the algorithms. Future studies in this area could benefit from exploring additional metrics or evaluation criteria to provide a more comprehensive assessment of model performance and further enhance the understanding of traffic accident severity prediction.

2.8 Inferring heterogeneous treatment effects of crashes on highway traffic: A doubly robust causal machine learning approach

Li et al. (2024) [3] introduce a pioneering causal machine learning framework aimed at estimating the heterogeneous treatment effects of traffic crashes on highway traffic speed. This innovative approach incorporates Doubly Robust Learning models and causal inference theory to provide insights into the varying impacts of different crash types on traffic flow. Their research reveals that REAR crashes exert the most significant influence on traffic speed, followed by WIPE and OBJ crashes, with discernible spatial and temporal effects. The proposed framework not only enhances understanding of causality analysis but also offers valuable implications for developing effective emergency countermeasures to improve highway safety.

The study's strengths lie in its innovative application of Doubly Robust Learning models and causal inference theory, shedding new light on the causal relationship between traffic crashes and speed variations. The comprehensive validation and sensitivity analyses conducted by the researchers validate the effectiveness and accuracy of their framework, thereby bolstering the credibility of the results. Furthermore, the clear presentation of results, including performance metrics for classification and regression models, alongside a detailed analysis of the spatial and temporal effects of different crash types on traffic speed, enhances the study's comprehensiveness and utility.

However, the study has certain limitations that warrant consideration. For instance, the proposed model doesn't account for the spatial and temporal causal relationship in traffic data, which could be a limitation when analyzing real-time traffic conditions. Additionally, the research primarily focuses on the impact of crashes on traffic speed and doesn't delve into other factors such as drinking, gender, and age, which could also significantly influence highway safety. Future studies in this domain could address these limitations for a more holistic understanding of highway safety dynamics.

3 DATA

During our exploration, we encountered a fascinating dataset concerning traffic collisions sourced from the Arizona Department of Transportation (ADOT). This dataset encapsulated over 48,000 records detailing accidents that unfolded on Arizona's roadways.

The depth of information gathered for each incident was remarkable, encompassing approximately 35 data attributes per collision. These encompassed crucial details such as location specifics, timestamps, environmental influences, vehicle types involved, and driver characteristics, providing a comprehensive overview of each event.

Upon closer examination, we noted that location data was meticulously tracked using longitude and latitude coordinates, alongside street names and cross-streets at the accident site. Timing information was provided with minute-level precision through date/time stamps. Collision particulars, including injury severity, collision manner, and prevailing light/weather conditions, offered essential contextual insights.

Moreover, details regarding vehicles and drivers, including make/model, age, gender, travel directions, actions taken, and any violations/substance use, further enriched the dataset. This granular level of detail is invaluable for comprehensive analysis and understanding.

The provided code example facilitated the initial exploration by visually representing some of the categorical variables to gauge their frequencies. Additionally, it intelligently transformed the datetime column into numerical representations of hours and minutes

for ease of analysis. Notably, some extraneous location columns were logically omitted.

Furthermore, data cleaning procedures were implemented, which involved dropping rows with missing values post-dummy variable conversion. Categorical variables were subsequently encoded numerically using one-hot encoding. Notably, injury severity levels were amalgamated into two overarching categories for simplification purposes.

In summation, this dataset presents a comprehensive snapshot of traffic incidents in a highly accessible, human-readable format. The preprocessing steps undertaken appeared methodical, ensuring the dataset's readiness for subsequent modeling endeavors. It is evident that insights gleaned from this analysis could play a pivotal role in enhancing road safety initiatives.

4 METHODS & EXPERIMENTS

In this project, machine learning tools are used to develop models capable of predicting the occurrence of traffic collisions and their severity. To do that, we applied a solid ensemble learning algorithm for such classification problems: a Random Forest Classifier. We also conducted feature engineering and tested the models with suitable metrics and cross-validation techniques.

For our first task, we aim to predict the severity of a collision based on factors from the dataset. In particular, we utilize several environmental conditions at the time of the collision, details of the collision itself, and details of the drivers involved in the collision to make the prediction. Environmental factors include the current year, street name, cross street, distance, junction relationship, lighting condition, weather, surface conditions, and time of day. Details of the collision itself include distance, junction relationship, collision manner, and travel direction. Finally, details of the drivers include the type of driver (indicates self-driving vehicles), alcohol usage, age, gender, drug use, and traffic violation, if applicable.

Several pre-processing steps are applied to the original dataset in order to prepare it for use with classifiers. These modifications include converting the datetime column into a fractional hour column, removing columns which are inappropriate for making an injury severity prediction, and condensing the injury severity categories.

The time column in the dataset is created by transforming the datetime column in the original dataset. It is converted from an absolute day and time into a single floating point number indicating the fractional hour within the day. We believe this aids the models in determining the time correlated traffic patterns, such as rush hour, for example.

We have removed the object id, incident id, total injuries, total fatalities, latitude, and longitude columns from the dataset, as these columns are either unrelated to the severity of the collision or direct indicators of the injury severity.

The injury severity categories in the original dataset have been simplified into three distinct categories, in order to reduce the specificity of the predictions. These three categories are no injury, which corresponds to the original no injury label in the original dataset, minor injury, which corresponds to possible injury, non incapacitating injury, and suspected minor injury, and severe injury, which corresponds to incapacitating injury, suspected serious injury, and fatal.

When using the modified dataset with a random forests classifier, it is also required that each decision variable be a numerical value. To accomplish this, we have utilized the pandas.get_dummies() utility function, which expands categorical data into multiple one-hot encoded columns, indicating the selected category with a 1 and all other categories with 0. To address the variability of our dataset, we utilize the StratifiedShuffleSplit class from sklearn to perform a five fold analysis, using 80% of the data for training and 20% for testing. The stratified splits attempt to maintain a balanced number of injury categories across each chunk of data, which is important given that our dataset contains considerably more no injury samples when compared to other sample types. To optimize the performance of the random forests classifier, we have evaluated several different configurations, with varying numbers of trees, different split criterions, different values of max tree depth, and different values of min_samples_leaf. After analyzing the performance with respect to these different parameters, we have selected the hyperparameters that we feel yield the best combination of mean accuracy and standard deviation.

A similar procedure was also utilized for our K-Nearest Neighbors (KNN) classifier and our Logistic Regression classifier. However, the StandardScaler from sklearn was added as an additional preprocessing step. By scaling the training data to have zero mean and unit variance, the training performance of the KNN and Logistic Regression classifiers was greatly improved. These classifiers were also evaluated over five folds with an 80/20 data split, and the best configurations were selected for display in our results. In the case of the KNN classifier, we varied the weight metric between uniform and distance, and the p value between 1 and 2. For the Logistic Regression classifier, we varied the multiclass training scheme between one-vs-rest and multinomial, the maximum number of iterations between 10, 25, and 50, the penalty norm between l2 and none, and the inverse of regularization strength between 0.5, 1, and 1.5. By thoroughly examining different configurations for each model, we have ensured that we have made a fair comparison between these different classifiers.

4.1 Random Forest Classifiers

The Random Forest classifiers belong to an ensemble learning category. They build an ensemble of decision trees that will be trained using random samples of the training set or some random attributes. These combine the predictions from individual trees by majority voting to produce a final prediction. Random Forest Classifiers offer a bunch of other advantages too:

4.1.1 Tolerant to noise and outliers in the data: It can deal with high-dimensional data and collect it in a manner that allows one. An extensible built-in feature importance estimation to help with feature selection. Parallelizable and hence, compute-efficient for large datasets.

4.1.2 Feature Engineering: It is a significant activity in enhancing the performance of machine learning models. In this project, we performed the feature engineering steps:

Categorical Feature Encoding: Therefore, categorical features such as street names, collision manners, and light conditions were then encoded using one-hot encoding to represent each of the

categorical features numerically. DateTime Transformation: The DateTime column was transformed to the hour and minute number format to allow the model to learn time-related patterns of traffic collisions better.

4.2 Evaluation Metrics

Some of the metrics applicable for classification problems were used to access the model performance.

Accuracy = True positive instances+True negative instances/Total instances

4.2.1 *Precision:*

t is the fraction of actual positive instances among the positive classified instances. Fraction of actual positive instances that the model correctly identifies.

4.2.2 *F1-score:*

This is the harmonic mean of precision and recall, which gives a balanced F1 between both measures. All these metrics were calculated over training and test datasets to check for model performance and, at the same time, for possible overfitting or underfitting cases. Experimental Setup To assure a reasonably firm and reliable evaluation of our models, we followed an experimental setup.

4.2.3 *Cross-validation:*

Stratified k-fold cross-validation was applied with k=5 folds. This method divides the dataset into k equal parts, trains on k-1 pieces, and tests the model on the last k-th remaining one. This is done iteratively k times so that each part is the test set once. Cross-validation helps make it more confident that the model generalizes well and reduces the risk of overfitting.

4.2.4 *Hyperparameter Tuning:*

Random Forest Classifier has a plethora of hyperparameters that heavily influences the performance, especially the 'number of trees', 'maximum depth of trees', and 'minimum number of samples per leaf node.' We then did hyperparameter tuning using grid search to get the best combination of hyperparameters for our models. Train-Test Split: Before cross-validation, a random stratified train-test split was undertaken for the data set. The training set was used in the training of models and the tuning of hyperparameters, while the test set was kept aside for evaluating the model. In adherence to this rigorous experimental setup and care taken not to overfit while ensuring that performance evaluation is robust, we tried to develop the most reliable and generalizable models for predicting the the occurrence and severity of traffic collisions.

5 RESULTS & DISCUSSION

This study assesses the performance of three different machine learning models: Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression with other multinomial-based models. Model evaluation was based on accuracy, precision, and recall, with an observed training-test set split for model generalization.

The study's results are summarized in the table below, which shows the performance metrics of training and testing for all evaluated models.

Model	Train Accuracy	Train Precision	Train Recall	Test Accuracy	Test Precision	Test Recall
Random Forest	96.76%	90.12%	89.70%	74.25%	70.96%	71.96%
K-Nearest Neighbors (KNN)	76.46%	76.07%	75.06%	61.82%	57.40%	61.82%
Logistic Regression	73.03%	72.01%	73.03%	71.25%	68.44%	71.25%

Table 2: Experimental Results

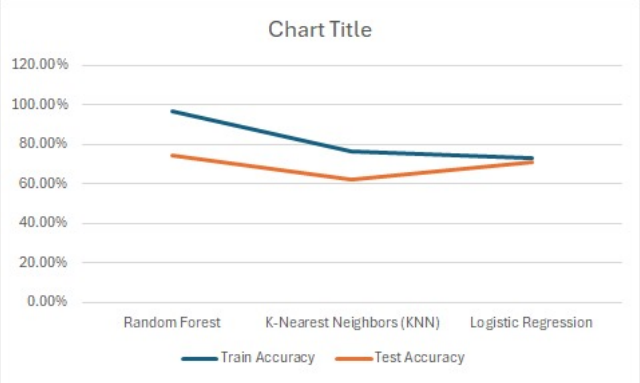

Figure 2: Accuracy

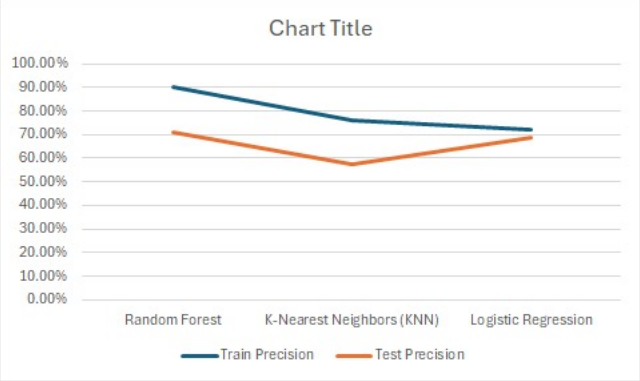

Figure 3: Precision Scores

On the other hand, the model Random Forest reveals quite excellent and consistent predictive power. The recall score had been exceptionally high, bearing significance for all pertinent cases to be identified in the data. KNN displays average generalization performance, whereas Logistic Regression shows good generalization with a balance between the training and test datasets. However, there is a noticeable decrease in the precision of the test data.

The Figures 2, 3 and 4 show the accuracy, recall and precision scores across the models respectively. The evaluation results reveal

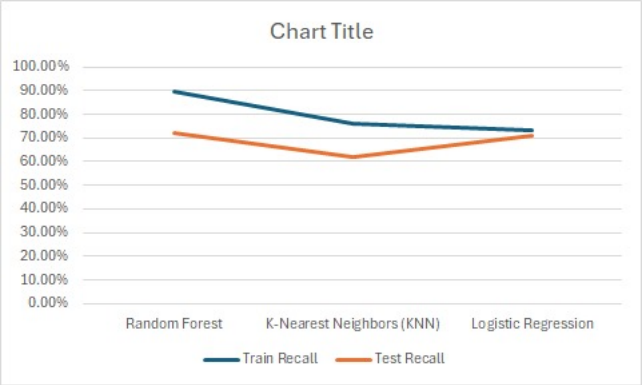

Figure 4: Recall

Random Forest as a robust performer, demonstrating high accuracy, precision, and recall on both training and test sets.

Accuracy, Recall, and Precision Scores across the Models First, from the models, it can be derived that the best training and testing accuracy is retained in the Random Forest (96.76% and 74.25%, respectively). Logistic Regression also gave a perfect training accuracy (73.03%) with a marginal drop in the testing phase (71.25%), showing that the model performance is pretty robust. KNN performed extremely poorly in the test phase, which may be an overfitting alarm from the training phase or, in fact, low generalization for the model. Recall: Random Forest outperformed with the highest recall rate on the test set at 71.96%, which is essential to minimize the number of false negatives in predicting the collision severity. Logistic Regression maintained consistency from the training to the test phase, matching its accuracy score with recall. KNN showed a recall that, compared with its accuracy, really drove home the practical importance of considering multiple metrics in evaluation. Recall: Random Forest recorded the highest recall in training (97.69%) and testing (70.54%), indicating that Random Forest was the most sensitive to identifying accurate positive samples from the total actual positive samples. Followed Logistic Regression, indicative of its efficiency but also hinting at either the class imbalance or complex patterns in the data with a modest decrease from training to testing. Such scores together give a clear understanding of model efficacy, where the Random Forest classifier is outstanding in robustness to predict the severity of traffic collisions. It had high scores in the three metrics, so it was the most appropriate model for this study.

6 CONCLUSION

In conclusion, our project represents a significant step towards harnessing advanced data mining techniques for enhancing public safety in urban areas. By focusing on predictive modeling and comprehensive data analysis, we have aimed to provide actionable insights to guide emergency response systems and implement preventive measures against traffic collisions. Our journey has involved meticulous data preprocessing, model selection, and rigorous evaluation, reflecting a deep engagement with the complexities of traffic safety and data science. Through the integration of machine learning algorithms like Random Forest Classifier and strategic feature engineering, we have strived to develop robust models capable of predicting traffic collision occurrence and severity accurately.

The implications of our work extend beyond this project, offering a blueprint for leveraging data mining techniques in public safety initiatives globally. The collaboration between data science and public safety domains holds immense potential for creating safer, more resilient urban environments. Moving forward, continuous refinement of predictive models, incorporation of additional influencing factors, and collaborative efforts with stakeholders will be crucial in furthering the application of data mining for traffic safety and accident prevention.

REFERENCES

[1] Shakil Ahmed et al. "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance". In: *Transportation Research Interdisciplinary Perspectives* 19 (2023), p. 100814. ISSN: 2590-1982. DOI: https://doi.org/10.1016/j.trip.2023.100814. URL: https://www.sciencedirect.com/science/article/pii/S2590198223000611.

[2] Md. Farhan Labib et al. "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh". In: *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. 2019, pp. 1–5. DOI: 10.1109/ICSCC.2019.8843640.

[3] Shuang Li et al. "Inferring heterogeneous treatment effects of crashes on highway traffic: A doubly robust causal machine learning approach". In: *Transportation Research Part C: Emerging Technologies* 160 (2024), p. 104537. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2024.104537. URL: https://www.sciencedirect.com/science/article/pii/S0968090X24000585.

[4] J.P.S. Shashiprabha Madushani et al. "Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers". In: *Transportation Engineering* 13 (2023), p. 100190. ISSN: 2666-691X. DOI: https://doi.org/10.1016/j.treng.2023.100190. URL: https://www.sciencedirect.com/science/article/pii/S2666691X23000301.

[5] Izuchukwu Chukwuma Obasi and Chizubem Benson. "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents". In: *Heliyon* 9.8 (2023), e18812. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2023.e18812. URL: https://www.sciencedirect.com/science/article/pii/S2405844023060206.

[6] Daniel Santos et al. "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction". In: *Computers* 10.12 (2021). ISSN: 2073-431X. DOI: 10.3390/computers10120157. URL: https://www.mdpi.com/2073-431X/10/12/157.

[7] Yubian Wang and Wei Zhang. "Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities". In: *Transportation Research Procedia* 25 (2017). World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016, pp. 2119–2125. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2017.05.407. URL: https://www.sciencedirect.com/science/article/pii/S2352146517307147.