Spring 2024

Predictive Analytics of Road Traffic Incidents, A Machine Learning Approach

Maryam Essa Jaji
me3101@rit.edu

Predictive Analytics of Road Traffic Incidents, A Machine Learning Approach

by

Maryam Essa Haji

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
Spring 2024

Acknowledgments

First and foremost, I want to thank Allah, the Almighty, for providing me with the courage, knowledge, and grace to pursue and complete my academic journey.

My parents have been the most consistent support system in my life, and I am thankful to Allah for them. Their philanthropic deeds of kindness and moral support have given me the strength to complete this order.

In addition, I would like to thank Professor Sanjay Modak, the department chair, and my professors for their wise guidance and the pleasant environment that has allowed me to grow and learn with them. The perspective I've developed as a result of their constant pursuit of their students' success has enriched my life.

Another person to whom I am particularly grateful is my friends and his groups. During my time at college, we have been each other's largest supporter when it comes to school. We've fostered and encouraged each other.

Thank you, Dr. Esan Ullah Warriach, my project advisor, and matching guru, for his continued backing during difficult times. Dr. Ehsan's perseverance and expert advice turned what appeared to be an impossible challenge into a manageable one. During the most challenging periods, he viewed anything in me when I had no idea how much I could do in a short period. Thank you for your thoughtfulness and counsel.

I would also like to thank my friend, my coworker, who helped me understand my lessons and project. His advice and encouragement were always a source of strength for me. My coworker and friend merit a distinct word of gratitude for making the lessons and project simpler to solve by providing advice, tips, and suggestions.

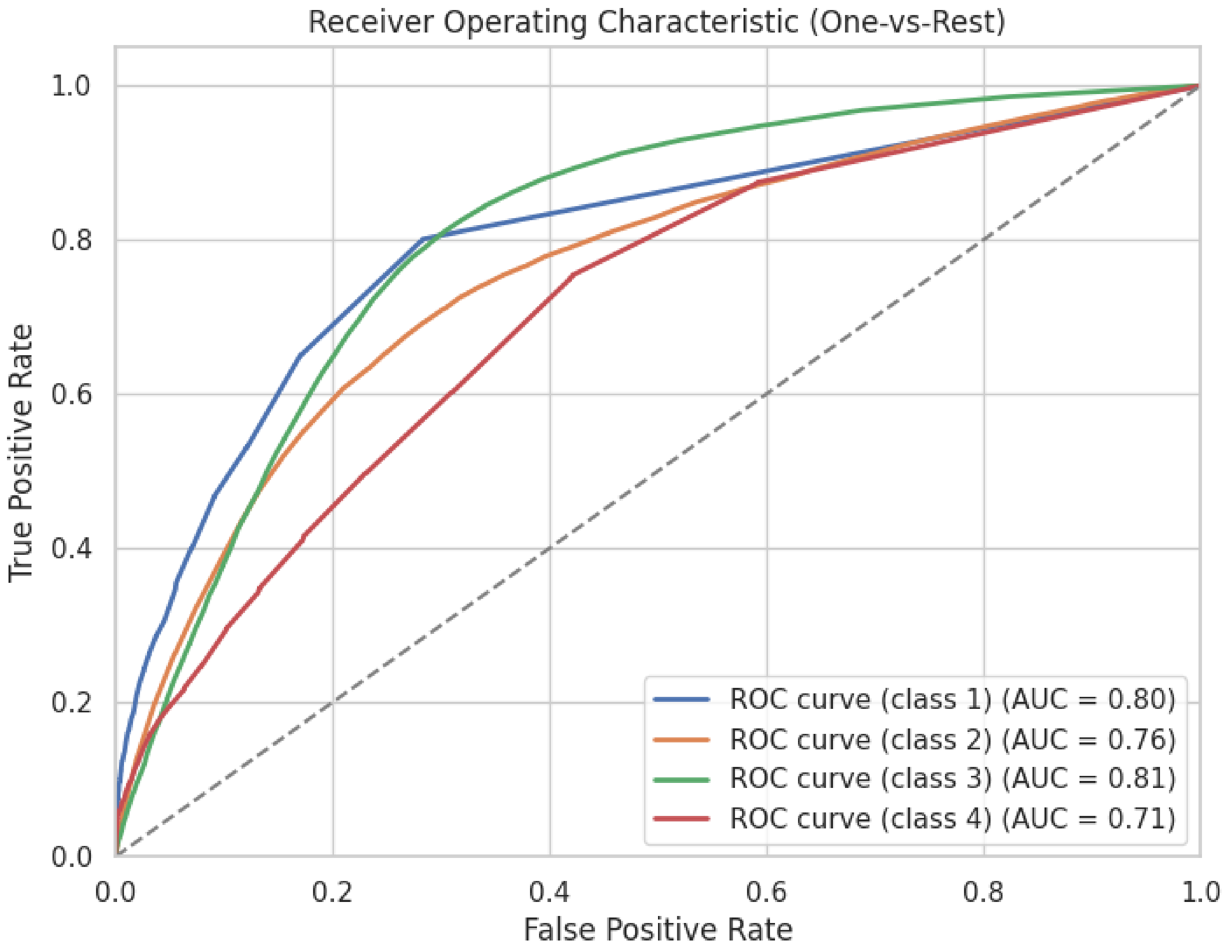
Abstract

Traffic accidents rank among the world's most serious concerns due to the high number of deaths, injuries, and fatalities as well as the enormous financial losses they cause every year. Road travel is a necessary component of modern civilization, but because of the rise in traffic accidents, it costs the world economy billions of dollars and over a million deaths annually. Road accidents can be caused by a variety of elements. It can be possible to take action to lessen the severity and extent of the effects if these elements are better recognized and predicted.

The goal of this project is to use machine learning techniques to forecast the severity of traffic incidents. Utilizing the US Accidents dataset sample from Kaggle, the project develops a Random Forest classifier prediction model, Decision Tree model, and K-Nearest Neighbor (KNN) model. To predict the severity of accidents, the model utilizes the use of several data, including time-related factors, road characteristics, and weather. Python programming language has been used to develop the predictive model. The project desires to improve public safety and minimize the effect of road traffic events by offering actionable data for emergency response teams and traffic management. The outcomes obtained have increased confidence that the use of advanced features contributes to improved traffic accident prediction. The random forest model has an accuracy of approx. 84%, the Decision Tree model has an accuracy of approx. 76% and K-Nearest Neighbor (KNN) model has an accuracy of approx. 83%. Hence, the random forest model performed well, but there is undoubtedly space for improvement especially when it comes to managing minority classes.

Key Words: Machine Learning, Data Analysis, Road Accident Data, Random Forests, Decision Tree, K-Nearest Neighbor (KNN), Traffic Accident Dataset, Traffic Accident Prediction

Table of Contents



List of Figures

Chapter 1

1.1 Introduction

Traffic accidents on roads are among the worst threats to human life. Forecasting possible accidents can help prevent them, lessen the harm they do, warn drivers of impending dangers, or enhance the emergency response system. If local authorities are informed in advance about which sections of the district's roadways are most likely to have an accident at different times of the day, officials may be able to reduce the amount of time it takes them to react (Santos , Saias, Quaresma & Nogueira, 2021).

The World Health Organization recently released statistics on fatalities, and they indicate a startling number of road accidents that occur annually around the world. 1.2 million people lost their lives in car crashes. Each year, 50 million people suffer injuries. There were over 13,7,000 injuries and 3,300 fatalities every day. 43 billion dollars in actual economic losses, in addition to the constant threat that road accidents represent to property safety and human life. Traffic-related incidents, such as collisions, gridlock, and dangers on the road, present serious threats to public safety and cause financial losses on a global scale. The World Health Organization (WHO) reports that traffic accidents are a major cause of mortality worldwide, especially for young adults between the ages of 15 and 29. Traffic accidents not only cause a human cost but also significant financial losses because of lost production, property damage, and medical bills. The forecasting of traffic accidents is one of the most important topics in traffic safety research. A number of factors, including road shape, traffic flow, driver characteristics, and road environment, significantly influence the risk of traffic accidents.

Numerous studies on the recognition of risky areas or regions, the assessment of accident injury severity, and the study of accident duration have all been conducted in an effort to anticipate accident frequencies and assess the components of traffic accidents. Numerous studies focus on the technical aspects of the accident. Other concerns include the clarity of the road and the weather. Weather conditions like as rain and fog can significantly raise the risk of accidents. Therefore, if you have a correct estimate of events and have knowledge of accident zones and contributing variables, taking action to decrease accidents will be easy. This calls for a thorough analysis of the incidents and the development of accident prediction models. This essay will discuss that an accident prediction model may be used to assess the risks associated with different accident situations (Vanitha & Swedha, 2023)

Traditional methods of managing traffic safety mostly focus on responding to incidents and analyzing accidents. These methods, however, frequently fall short of what is needed to avoid accidents and lessen their severity. By using past data and machine learning algorithms to predict the incidence and severity of traffic incidents, predictive analytics provides a proactive strategy. Authorities can improve emergency response plans, optimize traffic flow, and put preventative measures into place by identifying high-risk locations and periods.

1.2 Problem Statement

Road traffic accidents are a huge global problem that results in a large number of fatalities, severe injuries, and economic damage. A significant challenge in the management of traffic safety is the lack of reliable predictive technologies for predicting and reducing accidents. Emergency response teams and traffic management authorities find it difficult to deploy resources efficiently and react quickly to crises in the absence of proactive measures. Accurate predictive analytics solutions that can anticipate the severity of traffic events are therefore urgently required in order to improve public safety on highways by enabling safeguarding responses.

1.3 Project Goals

The project seeks to solve important issues about road accidents. The project's goals are as follows:

Finding important factors that are linked to accident severity is the main goal of factor identification. Through an analysis of accident data from years 2016 to 2023 the research aims to identify key factors that influence accident outcomes. These elements will help categorize the seriousness of incidents and offer guidance on the methods for reducing them.

Predictive Modelling: The research suggests a predictive model for upcoming traffic accidents alongside factor identification based on historical data. The machine learning models have been made using three machine learning techniques: random forests, decision tree, and K-nearest neighbor (KNN). The findings suggest that the random forests classifier model has a better potential in predicting accident areas.

1.4 Aims and Objectives

The project aims to develop machine learning models and choose the best one for predicting the predictive analytics of road traffic incidents. The primary objectives include:

1. The collection and preprocessing of the relevant data from the US Accidents dataset.
2. The selection of appropriate features and engineering new ones to enhance predictive accuracy.
3. Training of the Random Forest classifier, Decision Tree, and K-Nearest Neighbor (KNN) using the processed data to predict accident severity.
4. Evaluation of the performance of the models using metrics such as accuracy, precision, recall, and F1-score.
5. Developing effective prediction techniques, insights, and suggestions for traffic management and emergency response teams based on model predictions.

By achieving these objectives, the project seeks to contribute to the development of effective predictive analytics models which will enhance road safety and reduce the impact of road traffic incidents.

1.5 Research Methodology

1.5.1 Data Preprocessing

Data Collection

A Countrywide Traffic Accident Dataset sample from the years 2016 to 2023 on traffic events in the US has been downloaded from the Kaggle website. This dataset contains data on vehicle kinds, types of roads, weather, locations of accidents, and other relevant factors. This is a nationwide dataset of auto accidents that includes data from 49 US states. Using several APIs that offer stream traffic incident data, the accident data were gathered between February 2016 and March 2023 (Moosavi, Samavatian, Parthasarathy, & Ramnath, 2019).

Data Cleaning

The dataset's outliers, inconsistent values, and missing values have been taken into account. The downloaded data from the Kaggle website is accurate and ready for analysis.

Feature Engineering

Feature engineering is the process of integrating current features to produce new ones. I determine the characteristics of the road section the time of day etc.

1.5.2 Feature Selection

Feature selection includes cleaning the dataset by handling missing values, removing irrelevant columns, and encoding categorical variables.

Finding Relevant Features: Examine how each factor affects the severity of the accident. The selection approach is guided by methods like mutual information, and correlation analysis.

Dimensionality Reduction: Using techniques like feature significance ratings from machine learning models reduce the dimensionality of the dataset as needed.

1.5.3 Model Training Using Random Forest Classifier, Decision tree and K-Nearest Neighbor (KNN)

Random Forest (RF): Random Forest algorithm is preferred as it is resistant to overfitting, robust, and it is able to manage complex relationships.

Decision Tree: The decision tree approach makes use of conditional control statements to forecast the ultimate decision by building a tree-like graph of alternatives and potential outcomes.

K-Nearest Neighbor (KNN): The KNN algorithm attempts to categorize an observation using the k nearest observations in the feature space.

Training Set and Test Set: The dataset was divided into a testing set and a training set. Training of the Random Forest classifier, Decision Tree, and K-Nearest Neighbor (KNN) has been performed using the training data.

1.5.4 Evaluation of Model Performance

Analyzing the performance of trained models on the testing data using measures like F1-score, accuracy, precision, and recall. Use the testing set to evaluate the performance of the trained models.

Metrics are as follows:

Accuracy Score: It measures the overall correctness of predictions.

Classification Report: It provides precision, recall, F1-score, and support for each class.

Utilize the trained model to forecast the seriousness of upcoming traffic events and to offer useful information to traffic management and emergency response teams.

1.6 Limitations of the Study

The limitations of the study are as follows:

1. Model Assumptions: Our prediction approach is predicated on the idea that characteristics and accident severity will always be correlated in the same way. The prediction approach assumes a consistent correlation between characteristics and accident severity. However, real-world conditions change over time. The accuracy of the model can be impacted by modifications to traffic patterns, road conditions, or new legislation. If traffic patterns, road conditions, or legislation evolve then the model's accuracy may degrade. The model might fail to adapt to new scenarios, leading to incorrect predictions. Regularly update the model with fresh data to capture changing patterns. Consider dynamic feature engineering that adapts to evolving conditions.
2. Interpretability of the Model: Random Forest, Decision Tree and K-Nearest Neighbor (KNN) models are not as interpretable as simpler models, despite being strong. It's still difficult to fully understand how the model makes decisions. Random Forests, Decision Trees, and KNN models, while powerful, lack interpretability compared to simpler models. Stakeholders like traffic management professionals and policymakers may struggle to trust or understand the model's decisions. Lack of transparency hinders practical implementation. Use simpler models for interpretability when feasible. Provide feature important rankings to explain model decisions. Explore model-agnostic interpretability techniques.
3. External Factors: Things like driver behavior, road upkeep, and emergency response times are not taken into consideration in the project. The results of accidents are significantly influenced by these elements. Real-world accidents are influenced by these factors, which are not captured by the model. The model's predictions may not align with actual outcomes due to unaccounted external influences. Collect additional data on driver behavior e.g., aggressive driving, distractions, etc. Incorporate road maintenance schedules and emergency response times into the model. Collaborate with domain experts to identify relevant external features.

4. Generalization to Different Data Sets: The model’s performance may vary when applied to different geographical regions or diverse datasets. External validation of unseen data is essential to assess generalization. Collect data from multiple cities or regions to ensure broader applicability. Conduct cross-validation across various datasets to evaluate robustness.
5. Computational Resources: Limited computational resources may restrict model complexity and training time. More resources allow for deeper architectures, hyperparameter tuning, and ensemble methods. Leverage cloud computing or distributed systems for scalability. Optimize hyperparameters using grid search or Bayesian optimization.

Chapter 2 Literature Review

2.1 Literature Review

2.1.1 Traffic Incident Duration Analysis and Prediction:

Understanding and predicting the duration of traffic incidents is vital for managing congestion caused by accidents, breakdowns, or road closures. This review delves into the phases of incident duration from detection to recovery and highlighting the importance of prompt response and efficient clearance activities. Researchers leverage diverse data sources including incident reports and real-time traffic data to analyze factors influencing incident duration such as incident type, location, weather conditions, and emergency response time. Timely incident management is crucial for minimizing disruptions and improving transportation efficiency as emphasized by Li, Pereira, and Ben-Akiva (2018). Moving forward integrating real-time data and predictive models can enhance incident management strategies and mitigate traffic disruptions effectively. Overview of traffic incident duration analysis and prediction study reviews incident duration phases, data resources, and methods for influence factor analysis and prediction. It emphasizes the importance of timely incident management to reduce traffic disruptions (Li, Pereira, & Ben-Akiva, 2018).

2.1.2. Machine Learning Techniques for Road Traffic Automatic Incident Detection Systems:

In the realm of road traffic management, automatic incident detection systems powered by machine learning techniques play a pivotal role in ensuring safety and minimizing congestion. This review explores the application of machine learning in incident detection emphasizing real-time analysis and the importance of feature selection, quality training data, and model complexity. Despite progress, challenges such as data quality and model interpretability persist, urging researchers to seek hybrid approaches and embrace transfer learning for improved system robustness and effectiveness. Looking ahead, the integration of real-time model updates presents an exciting avenue for further advancement in enhancing transportation safety and efficiency through machine learning technologies. Machine Learning Techniques for Road Traffic Automatic Incident Detection Systems review explores machine learning techniques for incident detection, emphasizing real-time applications and system effectiveness (Hireche & Dennai, 2020)

2.1.3 Analysis of Road Accidents Prediction and Interpretation Using the KNN Classification Model:

Road accidents are a significant concern for public safety and transportation efficiency necessitating effective prediction and interpretation measures. This review scrutinizes studies on road accident prediction is particularly focusing on the K-nearest neighbors (KNN) classification model. The KNN model is known for its simplicity and efficacy in pattern recognition learns from labeled data during the training phase and predicts outcomes based on the majority class label among the nearest neighbors in feature space. Researchers explore various aspects of road accident prediction using KNN including feature selection, data preprocessing, and model evaluation. Despite its effectiveness, challenges such as scalability, hyperparameter tuning, and imbalanced data persist. Future research avenues include ensemble methods, feature engineering, and incorporating spatial-temporal considerations to enhance prediction accuracy. The study by Sahu, Maram, Gampala, and Daniya (2022) contributes valuable insights that are urging critical evaluation of existing research to ensure safer roadways. Analysis of Road Accidents Prediction and Interpretation Using the KNN Classification Model study focuses on road accident prediction using the K-nearest neighbors (KNN) classification model (Sahu, Maram, Gampala, & Daniya 2022).

2.1.4 Analytical Methods and Determinants of Frequency and Severity of Road Accidents:

Road traffic accidents (RTAs) persist as a pressing global issue causing countless fatalities and injuries annually. Despite ongoing safety initiatives, progress remains inadequate. Through a thorough 20-year systematic literature review (SLR), they examine analytical methodologies, contributing factors, and avenues for further investigation in RTA analysis. The review of 3888 papers published between January 2000 and June 2021 revealed four primary clusters each focusing on statistical analysis, machine learning, smart city technologies, or geographic information systems. Key findings highlight "Accident Analysis and Prevention" as the leading journal and Fred Mannering as a prominent RTA researcher. Analytical methods range from traditional negative binomial regression to emerging techniques like deep learning and convolutional neural networks offering promising avenues for RTA prediction. Exploring determinants of RTA frequency and severity reveals variables such as road type, weather conditions, and traffic density as crucial factors. Moreover, technological advancements, including 5G technology, Internet of Things (IoT), and Intelligent Transport Systems play pivotal roles in incident detection and response. This SLR not only addresses gaps in RTA analysis but also emphasizes the importance of computational algorithms and data visualization in shaping future research and developing intelligent systems for RTA prediction and prevention. Analytical Methods and Determinants of Frequency and Severity of Road Accidents research analyzes road accident frequency and severity determinants, providing insights into contributing factors (Ferreira-Vanegas, Vélez, & García-Llinás, 2022).

2.1.5 Road Accident Severity Prediction and Model Interpretation:

Understanding and predicting road accident severity is crucial for public safety and infrastructure preservation. Yassin and Pooja (2020) conducted a study utilizing a hybrid K-means and random forest (RF) approach for severity prediction. K-means clustering extracted hidden patterns from accident data while RF achieved an impressive accuracy of 99.86% in severity prediction. Key findings highlighted contributing factors such as driver experience, day of the week, light conditions, driver age, and vehicle service year for different severity levels. The study emphasizes the importance of accurate prediction models in developing effective road safety strategies for road transport agencies and insurance companies, ultimately contributing to safer roadways and better accident management. Road Accident Severity Prediction and Model Interpretation Using a Machine Learning Approach study identifies significant contributing factors for road accident severity, emphasizing the need for understanding primary causes (Yassin, & Pooja, 2020).

2.1.6 Traffic Incident Duration Prediction Using Hazard-Based Models

Predicting traffic incident durations is crucial for managing road transportation networks efficiently. Leveraging machine learning and real-time data, hazard-based models offer promising avenues for such predictions. The study, focusing on the M25 motorway in London, utilized dynamic predictions and interpretability through the Match-Net algorithm and Shapley values, respectively. Notably, time of day consistently influenced predictions, while time-series features played a stronger role at specific horizons. The dynamic hazard-based models outperformed static regression models, showcasing their effectiveness for incident duration prediction. By enhancing both prediction accuracy and interpretability, the findings contribute to better incident management and system-level optimization with broader applicability beyond traffic incidents. Traffic Incident Duration Prediction Using Hazard-Based Models study investigates hazard-based models for incident duration prediction considering factors such as incident type, location, and weather conditions Mehdizadeh, Cai, Hu, Alamdar Yazdi, Mohabbati-Kalejahi, Vinel, & Megahed, 2020).

2.1.7 Machine Learning Approaches for Traffic Incident Detection and Prediction:

Machine learning methods play a pivotal role in detecting and predicting traffic incidents are crucial for efficient transportation management. This review examines various algorithmic approaches including Random Forest, Support Vector Machines, Neural Networks, K-Nearest Neighbors, and Gradient Boosting, each with distinct strengths and limitations. Assessment methods involve metrics such as accuracy and precision while variables like traffic flow, weather conditions, and time of day contribute to incident prediction. Despite the consistent performance of Random Forest and Support Vector Machines, challenges persist, including imbalanced datasets and the interpretability of deep learning models. Bridging these gaps and exploring hybrid approaches could lead to safer and more efficient transportation systems. Machine Learning Approaches for Traffic Incident Detection and Prediction reviews various machine learning algorithms for incident detection and prediction. Based on the technique and algorithms used to estimate traffic flow, the retrieved material comprises the approaches, assessment methods, variables, datasets, and outcomes of each evaluated study (Razali, Shamsaimon, Ishak, Ramli, Amran, & Sukardi, 2021).

2.1.8 Traffic Accident Analysis Using SVM and MLP Models:

Traffic accidents pose significant challenges worldwide, impacting public safety and infrastructure. Sharma, Katiyar, and Kumar's (2016) study comparing Support Vector Machines (SVM) and Multilayer Perceptron (MLP) models in analyzing accident data reveals SVM's superiority in predicting accident severity, achieving 94% higher accuracy than MLP. The study emphasizes alcohol consumption and driving speed as critical variables influencing accident outcomes, highlighting the importance of accurate prediction models for effective accident prevention. Future research should focus on refining methodologies and exploring hybrid approaches to enhance prediction accuracy and contribute to safer roadways. SVM and MLP were used by Sharma to analyze data on traffic accidents on a small number of datasets. In addition, the authors included just two independent variables alcohol and speed as critical considerations. In the end, SVM with an RBF kernel produced results that were 94% more accurate than MLP. According to the survey, driving while intoxicated is the primary cause of accidents (Sharma, Katiyar, & Kumar, 2016).

2.1.9 Machine Learning Approaches for Motorcycle Crash Analysis:

Understanding and preventing motorcycle crashes is crucial for road safety, especially in countries like Ghana. A study by Wahab and Jiang (2020) assesses various machine learning classifiers to identify primary causes of motorcycle accidents. They analyze crash incidents in Ghana by employing models like Multilayer Perceptron, PART, and SimpleCART with SimpleCART demonstrating superior accuracy and interpretability. Factors such as driver experience, day of the week, light conditions, and vehicle service year emerge as significant predictors of crash severity. This research underscores the importance of accurate prediction models and highlights the need for further exploration of hybrid approaches to enhance road safety. Using MLP, PART, and SimpleCART, Wahab, and Jiang investigated crash incidents on a Ghanaian dataset with the goal of assessing classifiers and determining the primary causes of motorbike crashes. In order to determine the most significant variable influencing motorbike crashes in Ghana, the Authors employed InfoGainAttributeEval after comparing and analyzing datasets using Weka tools. Consequently, the simpleCART model outperformed other classification models in terms of accuracy (Wahab, & Jiang, 2020).

2.1.10 Characterizing Frequent Serious Accident Locations Using Data Mining:

Understanding frequent serious accident locations is crucial for improving road safety and reducing associated risks. Kumar and Toshniwal (2016) employ data mining techniques to analyze road accident data, focusing on K-means clustering and Association Rule Mining. They identify high-frequency accident areas, particularly noting intersections on highways as dangerous spots for various accident types, while two-wheeler accidents are prevalent in hilly regions. The study underscores the importance of analyzing accident data to prioritize preventive measures and highlights the need for further research to consider additional factors and broader datasets for a comprehensive understanding of accident patterns. In order to determine the frequently occurring serious accident locations and to retrieve hidden information, Kumar used data mining techniques such as kmeans and Association Rules. After eliminating accident locations with a frequency count of fewer than 20, 87 out of the 158 total locations were chosen (Kumar, & Toshniwal, 2016).

2.1.11 Analysis of Fatal Accidents Using Naive Bayes and Clustering Association Rule:

Understanding the root causes of fatal accidents is paramount for public safety, prompting Li, Shrestha, and Hu's (2017) study which employs data mining techniques to analyze fatal accidents in the United States. Their research utilizes the Naive Bayes algorithm for probabilistic classification and clustering association rules to uncover hidden patterns and factors associated with fatal accidents. By identifying human behavior and collusion as primary causes of mortality, the study underscores the importance of targeted prevention strategies and informed policy development. This analysis provides valuable insights for policymakers and safety agencies to mitigate the impact of fatal accidents and enhance road safety measures. The study used Nave Bayes and the Clustering Association rule to get statistics of fatal accidents in the United States. The study clarified and identified the forms of collusion and human beings as the primary causes of the mortality rate (Li, Shrestha, & Hu, 2017).

2.1.12 Identifying High-Risk Roadways Using Machine Learning Models

Understanding and preventing accidents on roadways is crucial for public safety, and a study by AlMamlook, Kwayu, Alkasisbeh, and Frefer (2019) employs machine learning models to identify high-risk roadways in Michigan. Utilizing algorithms such as AdaBoost, Naive Bayes, Logistic Regression, and Random Forest, the study aims to determine determinant variables associated with accident-prone areas. Performance metrics like F1-score, AUC, precision, recall, and ROC were used to rigorously evaluate model performance. While the study did not specify the best-performing model, it identified significant variables crucial for predicting high-risk roadways. This research contributes valuable insights for road safety strategies, enabling policymakers and transportation agencies to prioritize safety measures and enhance accident prevention efforts in Michigan and beyond. AdaBoost, Nave Bayes, Logistic Regression, and Random Forest were employed by AlMamlook to get determinant variables and to identify high-risk roadways for Michigan traffic agencies. Models were evaluated using performance metrics such as F1-score, AUC, precision, recall, and ROC (AlMamlook, Kwayu, Alkasisbeh, & Frefer, 2019).

2.1.13 Data Mining Techniques for Examining Traffic Incidents:

Examining traffic incidents through data mining techniques is essential for improving public safety and transportation efficiency. Tiwari, Kumar, and Kalitin (2017) conducted a study employing classification algorithms like Naive Bayes, Decision Tree, and Support Vector Machine, alongside clustering techniques such as K-modes and Self-Organizing Maps, to analyze causation class traffic incidents. Their findings highlight the effectiveness of clustering in identifying incident patterns accurately, contributing to targeted preventive measures and policy decisions. By critically evaluating existing research, we can refine methodologies and enhance accident prevention efforts, ultimately creating safer roadways and reducing incident severity. A data mining technique was used by Tiwari to examine causation class traffic incidents. The authors used algorithms for classification such as NB, DT, and SVM, as well as clustering such as K-modes and SOM. Better accuracy on the cluster dataset was thus shown than on the classification dataset (Tiwari, Kumar, & Kalitin, 2017).

2.1.14 Data Mining Methods for Analyzing Driver Responsibility

Analyzing driver responsibility in road accidents is crucial for improving accident prevention and safety measures. Regassa (2009) employed data mining techniques including Multilayer Perceptron and Decision Trees, utilizing the Weka tool to assess driver responsibility. By analyzing patterns and factors associated with driver behavior, the study aimed to determine the extent of driver responsibility in accidents. While the study didn't specify which model performed better, the use of data mining techniques provided a comprehensive understanding of driver responsibility. Through critical evaluation of existing research, we can refine methodologies and enhance road safety strategies by benefiting policymakers and transportation agencies in creating targeted interventions to reduce accidents caused by driver behavior. Research used data mining methods (MLP and Decision Trees) on the Weka tool, with a primary focus on driver responsibility (Regassa, 2009).

2.1.15 Data Mining Algorithms for Analyzing Driver and Vehicle Information:

Analyzing driver and vehicle information to understand accident severity is crucial for road safety measures. Getnet (2009) conducted a study utilizing data mining algorithms, including J48 and PART, with the Weka tool, to assess the impact of relevant variables on accident severity. By identifying significant risk factors associated with accident severity, the study provides insights for targeted safety interventions and policy decisions. Although the study did not specify which algorithm performed better, the use of data mining techniques allowed for a comprehensive understanding of accident severity determinants. Critically evaluating such research helps refine methodologies and enhance road safety strategies by benefiting policymakers and transportation agencies in creating safer roadways and reducing accident severity. Research used the J48 and PART data mining algorithms on driver and vehicle information that was thought to be a significant risk factor for the severity of accidents using the Weka tool (Getnet, 2009).

2.1.16 Investigating Incidents in Montreal Using Data-Driven Approaches:

Investigating incidents in Montreal using data-driven approaches is vital for enhancing public safety and urban planning. Hébert, Guédon, Glatard, and Jaumard (2019) conducted a study focusing on incident data related to Montreal, employing a balancing random forest algorithm to address class imbalance issues and ensure fair representation of incident types. By utilizing publicly available datasets such as the National Road Network Database, Historical Climate Dataset, and Montreal Vehicle Collisions, the study aimed to identify significant factors associated with incidents and their implications for urban planning. The findings provide valuable insights for policymakers and urban planners to design targeted interventions and prioritize safety measures, ultimately contributing to safer urban environments and reduced incidents in Montreal. A balancing random forest algorithm was employed by a Concordia University team to investigate the incidents that took place in Montreal. Three publicly available datasets were used to collect accident data. The National Road Network database, which included information on road segments, the Historical Climate Dataset, which included meteorological data, and Montreal Vehicle Collisions (Hébert, Guédon, Glatard, & Jaumard, 2019).

2.1.17 Machine Learning Approaches for Traffic Accident Severity Prediction:

Understanding and predicting traffic accident severity is crucial for public safety and economic stability, particularly in countries like Bangladesh. Siam, Hasan, Anik, Dev, Alita, Rahaman, and Rahman (2020) conducted a study employing machine learning techniques to analyze traffic accident data in Bangladesh. They collected data encompassing road characteristics, weather conditions, accident severity, and traffic incidents from 2015. By applying machine learning models such as Random Forest and Agglomerative Hierarchical Clustering, they predicted accident severity and identified significant predictor variables within homogeneous clusters of accidents. Their findings highlight the promising accuracy of the Random Forest model and provide valuable insights for prioritizing preventive measures and reducing accident severity, contributing to enhanced road safety strategies. To understand and anticipate the severity of the occurrences in Bangladesh, a team from North South University used a number of machine learning approaches. The data used include road data, weather conditions, accident severity, and traffic accidents from 2015. The authors identified the predictor variables for each cluster using random forest after extracting homogeneous clusters with the agglomerative hierarchical clustering technique (Siam, Hasan, Anik, Dev, Alita, Rahaman, & Rahman, 2020).

2.1.18 Factors Affecting Accident Severity Inside and Outside Urban Areas:

Understanding the factors influencing accident severity is essential for road safety and urban planning. Theofilatos, Graham, and Yannis (2012) investigated these factors both inside and outside urban areas in Greece. Analyzing disaggregated road accident data for 2008, they utilized binary logistic regression to estimate the probability of fatality/severe injury versus slight injury, identifying specific variables affecting accident severity. Within urban areas, factors such as young drivers, bicycles, intersections, and collisions with fixed objects played significant roles, while outside urban areas, weather conditions and types of collisions were influential. Their study contributes valuable insights for prioritizing targeted interventions in road safety strategies, benefiting policymakers and transportation agencies in Greece and beyond. Theofilatos analyzed both urban and rural variables. The study discovered that the severity of accidents was impacted by a variety of factors both inside and outside of cities. Bicycles, crossroads, young drivers, and accidents with objects were among the factors that influenced accident severity in metropolitan locations (Theofilatos, Graham, & Yannis, 2012).

2.1.19 Predicting Traffic Event Severity Using Machine Learning Approaches:

Understanding and predicting the severity of traffic events is crucial for mitigating their societal and economic impacts globally. Iranitalab and Khattak (2017) conducted a study utilizing machine learning methods to analyze accident data and identify influential factors associated with severity. By exploring techniques such as Random Forest, Support Vector Machines, Nearest Neighbor Classification, and Multinomial Logit, they found that Nearest Neighbor Classification outperformed other methods in predicting severe occurrences. These findings offer valuable insights for guiding safety measures and interventions, with Nearest Neighbor Classification showing potential for forecasting accident hotspots. Critically evaluating such research helps refine methodologies and enhance road safety strategies, enabling policymakers and transportation agencies to prioritize preventive actions and mitigate the impact of severe traffic events effectively. Iranitalab and Khattak investigated the use of RF analysis, support vector machines (SVM), nearest neighbor classification (NNC), and multinomial logit (MNL) to predict the severity of traffic events. The results show that NNC beats RF, SVM, and MNL in terms of overall prediction ability for more critical occurrences (Iranitalab, & Khattak, 2017).

2.1.20 Anticipating Traffic Accidents Using Machine Learning Techniques:

Understanding and anticipating traffic accidents is vital for minimizing casualties and economic losses worldwide. Lin, Wang, and Sadek (2015) conducted a study utilizing machine learning techniques to analyze accident data and identify influential factors associated with accident occurrence. Their investigation involved exploring methods like Random Forest, K-Nearest

Neighbor, and Bayesian Network for predicting accidents, with the top model demonstrating a 61% prediction rate despite a 38% false alert rate, showcasing its practical utility. The study emphasizes the importance of balancing trade-offs between sensitivity and specificity in real-world applications. By critically evaluating such research, we can refine methodologies and enhance road safety strategies, enabling policymakers and transportation agencies to prioritize preventive measures effectively and reduce accident severity. To anticipate traffic accidents, Lin looked into a number of machine learning techniques including random forest, K-nearest neighbor, and Bayesian network. The top model had a 38% false alert rate but could predict 61% of incidents (Lin, Wang, & Sadek, 2015).

2.1.21 Predicting Incidents Using CART Models:

Understanding and predicting road incidents is essential for public safety and infrastructure management. Chang and Chen (2005) conducted a study utilizing a CART model to analyze incident data and identify influential factors associated with incident occurrence. The CART model, a decision tree-based approach, showed moderate accuracy in predicting incidents, suggesting its potential utility despite not achieving high accuracy. While further research could explore methods to improve accuracy, the study contributes valuable insights for enhancing safety strategies and prioritizing preventive measures for policymakers and transportation agencies. In order to train and test a classifier that predicts incidents with a training and validation accuracy of 55%, Chang and Chen developed a CART model (Chang & Chen, 2005).

2.1.22 Regression Models for Forecasting Accident Frequency on Multi-Lane Highways:

Understanding and forecasting accident frequency on multi-lane highways is crucial for public safety and transportation efficiency. Caliendo, Guida, and Parisi (2007) conducted a study utilizing regression models to analyze accident data related to multi-lane highways. They explored three regression models Poisson Regression, Negative Binomial Regression, and Negative Multinomial Regression to predict accident frequency and assess their forecasting accuracy. The study's findings provide insights into the suitability of each regression model for predicting accident frequency by considering the dataset's characteristics and the trade-offs between accuracy and interpretability. By critically evaluating existing research, this study contributes to enhancing road safety strategies and guiding policymakers and transportation agencies in prioritizing preventive measures to reduce accident rates on multi-lane highways. To forecast the frequency of accidents on multi-lane highways, Caliendo employed the Poisson, negative binomial, and negative multinomial regression models (Caliendo, Guida, & Parisi, 2007).

2.1.23 Methods for Predictive Modeling in Various Contexts:

Predictive modeling is essential for understanding complex phenomena and improving outcomes across different domains. Silva, Andrade, and Ferreira (2020) conducted a study examining common methods like Artificial Neural Networks, Decision Trees, Support Vector Machines, Evolutionary Algorithms, and Closest Neighbor Classification for predictive modeling. They emphasize the importance of multivariate response models, which handle both regression and classification tasks by considering multiple dependent variables simultaneously. By critically evaluating these methods, the study highlights the need to choose appropriate techniques based on the problem context and data availability. Multivariate response models offer a holistic understanding of complex systems, capturing dependencies among variables and enhancing predictive accuracy. This research contributes to refining methodologies and improving decision-making processes across various domains, benefiting policymakers, researchers, and practitioners alike. The typical methods used for these objectives include artificial neural networks, decision trees, support vector machines, evolutionary algorithms, and closest neighbor classification, according to Silva. Multivariate response models are used in many different contexts due to their ability to handle both regression and classification issues (Silva, Andrade, & Ferreira, 2020).

2.1.24 Analyzing Traffic Accidents Using Real-Time Data:

Analyzing traffic accidents through real-time data offers valuable insights into accident probability and prevention strategies. Theofilatos (2017) utilized Bayesian logistic regression and Random Forest models to examine the relationship between traffic conditions and accident occurrence on metropolitan arterial highways. While the study provides significant contributions, gaps persist in data quality, spatial and temporal considerations, and model interpretability. Future research should focus on enhancing data collection and accuracy, investigating findings' generalizability across different contexts, and improving model transparency. Additionally, exploring causality versus correlation and incorporating human factors like distracted driving and fatigue are essential for a comprehensive understanding of accident risk. Overall, Theofilatos' work lays the groundwork for leveraging real-time data to enhance traffic safety, emphasizing the need for continued research to address remaining challenges and advance accident prevention strategies. Using real-time traffic data from metropolitan arterial highways, Theofilatos employed Bayesian logistic regression models and random forest models to examine the probability of traffic accidents (Theofilatos, 2017).

2.1.25 Machine Learning Approaches for Classification:

Machine learning techniques are widely utilized for classification tasks across various fields. The review integrates findings from studies by Theofilatos, Chen, & Antoniou (2019), which examined machine learning and deep learning algorithms for classification. These include k-Nearest Neighbors, Naive Bayes, Random Forest, Support Vector Machine, Classification Tree, Shallow Neural Network, and Deep Neural Network. While these studies offer valuable insights, gaps in data preprocessing and hyperparameter tuning are identified, along with the challenge of balancing interpretability and performance. Missing information pertains to the potential of transfer learning techniques and the effectiveness of ensemble approaches. Understanding these trade-offs and exploring novel techniques are crucial for informed decision-making in classification tasks. The author assessed a number of machine learning and deep learning approaches, such as kNN, naive Bayes, random forest, SVM, classification tree, shallow neural network, and deep neural network. It was discovered that the deep learning strategy yielded the best outcomes, with less sophisticated approaches like naive Bayes performing only marginally worse (Theofilatos, Chen, & Antoniou, 2019).

2.1.26 Long-Short Term Memory (LSTM) Models for Traffic Accident Risk Estimation:

Traffic accidents pose significant challenges globally, prompting researchers to explore innovative approaches for risk estimation and preventive measures. Ren et al. (2018) proposed an LSTM-based model to estimate traffic accident risk, leveraging the network's ability to capture temporal dependencies in sequential data. While offering valuable insights, gaps in data granularity, feature engineering, and interpretability are identified in existing literature. Questions remain regarding the model's transferability across different regions and its ability to provide uncertainty estimates for risk predictions. Addressing these challenges is crucial for advancing our understanding and improving safety measures in traffic accident risk estimation using LSTM models. A long-short term memory (LSTM) model was proposed by Ren as a means of estimating the risk of traffic accidents. Risk is defined as the total number of accidents in an area within a specific time period (Ren, Song, Wang, Hu, & Lei, 2018).

2.1.27 ConvLSTM Models for Traffic Accident Analysis:

Traffic accidents are a pressing global issue, prompting research into innovative methods for analysis and prevention. Yuan, Zhou, and Yang's (2018) study on vehicular accidents in Iowa utilized Convolutional Long Short-Term Memory (ConvLSTM) models, combining spatial and temporal features for accident prediction. While providing valuable insights, gaps in feature engineering and model interpretability are identified, along with questions regarding spatial and temporal resolution. Addressing these concerns is essential for advancing our understanding and enhancing safety measures in traffic accident analysis using ConvLSTM models. In a 2006–2013 study of vehicle accidents in Iowa, the researchers used a ConvLSTM setup. The data included RWIS reports, rainfall data, and Iowa Department of Transportation collision data (Yuan, Zhou, & Yang, 2018).

2.1.28 Modern Approaches to Traffic Accident Prediction:

Traffic accidents are a significant global concern, prompting research into modern approaches for prediction to implement effective preventive measures. Gutierrez-Osorio and Pedraza (2020) reviewed recent studies on traffic accident prediction, emphasizing the use of deep learning techniques, such as neural networks, which leverage complex architectures to learn from diverse data sources and improve accuracy. While providing valuable insights, gaps in feature engineering and model interpretability are noted, along with questions regarding temporal considerations and real-time implementation. Addressing these challenges is crucial for advancing our understanding and enhancing road safety through the application of deep learning methods and neural networks in traffic accident prediction. Gutierrez-Osorio and Pedraza reviewed recent studies on traffic accident prediction. The scientists discovered that deep learning techniques and neural networks demonstrated excellent accuracy and precision when integrating a variety of data sources (Gutierrez-Osorio, & Pedraza, 2020).

2.1.29 Data Mining Techniques for Road Accident Analysis:

Analyzing road accidents is essential for improving safety measures and transportation systems. Kumar and Toshniwal (2016) emphasized the significance of data mining techniques, such as clustering, classification, and association rule mining, in accident analysis. While these methods provide valuable insights, gaps in feature selection and segmentation quality are identified. Questions remain regarding temporal aspects and geospatial considerations, which could reveal patterns and accident hotspots over time and space. Addressing these challenges is crucial for enhancing our understanding of road accidents and implementing targeted interventions for prevention. Kumar and Toshniwal state that data mining techniques like clustering algorithms, classification, and association rule mining, along with the identification of accident-prone regions, are highly useful in assessing the various relevant factors of road accidents and in analyzing the various circumstances of accident occurrences (Kumar, & Toshniwal, 2016).

2.1.30 Hotspot Identification Techniques for Road Safety Analysis:

Identifying road sections with high crash risk known as black spots is crucial for effective road safety management. Montella (2010) conducted a comparative analysis of hotspot identification techniques, highlighting the effectiveness of the Empirical Bayes (EB) method. While various methods were evaluated including Crash Frequency, Equivalent Property Damage Only Crash Frequency, and Proportion Method, the EB method consistently outperformed others in reliability and accuracy. However, gaps in feature selection and practical implementation remain, along with questions regarding segment length considerations and integration into safety management processes. Leveraging the reliability of the EB method and addressing these gaps will enhance targeted safety interventions and resource allocation for road safety improvement. Several popular HotSpot IDentification (HSID) techniques have been compared by Montella. Empirical Bayes method (EB) is one of the techniques; it is the most dependable and consistent approach, outperforming other HSID techniques (Montella, 2010)

2.2 Main Takeaways from the Literature Review

The main takeaways and learnings from the literature review that were successfully applied to the research are as follows:

- The literature highlighted how important it is to find relevant attributes. To identify important variables affecting accident severity, such as weather, time of day, and road conditions, feature selection approaches can be used.
- Many machine learning techniques have been investigated such as random forests, SVMs, and neural networks for event identification and prediction by drawing on ideas from the literature. The accuracy of the chosen model was greatly affected by the methods used.
- Feature engineering, proper data cleansing, and dealing with missing values were essential. In order to guarantee the accuracy and integrity of our dataset, the best procedures were committed.
- The research highlighted how crucial it is to select the right evaluation criteria. I evaluated the performance of the model using accuracy scores, and classification reports.
- Accurate forecasting requires an understanding of the particular setting as well as considering local factors.

Chapter 3 Project Description

3.1 Data Sources

The primary data source for this project is the "US Accidents" dataset, which is available on Kaggle. A dataset sample has been used for this project. There is a sampled version of the dataset with 500,000 accidents available for those who want something more manageable and smaller. This sample was extracted from the original dataset to facilitate management and analysis. Traffic events that happened within the continental United States during the past three years are now included in the US-incidents database. Numerous intrinsic and contextual details, including location, weather, time of day, time, natural language description, and places of interest, are included in every accident record with 46 attributes. The dataset is collected from various sources including traffic APIs, traffic cameras, and government agencies. It is making it a comprehensive resource for analyzing road traffic incidents (Moosavi, Samavatian, Parthasarathy, & Ramnath, 2019).

3.2 Machine Learning Techniques

3.2.1 Random Forest

The project utilizes Random Forest machine learning algorithm for predicting the severity of road traffic incidents. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is well-suited for classification tasks and offers several advantages, including robustness to overfitting, handling of large datasets with high dimensionality, and automatic feature selection.

Random Forests belongs to the ensemble learning family. They combine multiple decision trees to create a robust and accurate model. RFs are known for their high accuracy due to the aggregation of multiple decision trees. They reduce overfitting by averaging out individual tree predictions. RFs provide feature importance scores, allowing us to understand which features contribute most significantly to predicting traffic incidents. This insight aids in feature selection and model interpretation. Traffic incidents often exhibit nonlinear relationships. RFs can capture complex interactions between features, making them suitable for this task. RFs handle missing data and outliers well, which is crucial when dealing with real-world traffic datasets.

Random Forest is a method for both regression and classification. Essentially, it is an assemblage of decision tree classifiers. Because random forests correct the tendency of overfitting to their training set, they have an advantage over choice trees in this regard. Following the construction of a decision tree, each node is split based on a randomly chosen feature from the whole feature set. Every single tree is trained using a subset for the training data. Because each tree in a random forest is trained independently of the others, training happens amazingly quickly, even for big data sets with a variety of characteristics and data occurrences. It has been discovered that the Random Forest method offers a decent estimation of the generalization error and is resistant to overfitting (Vanitha & Swedha, 2023).

3.2.2 Decision Tree

A decision tree approach builds a tree-like diagram or model of alternatives and potential outcomes using conditional statements of control to forecast the ultimate decision. Decision trees are easy to interpret. Their hierarchical structure allows us to visualize the decision-making process. Decision trees naturally select relevant features by splitting nodes based on feature importance. Decision trees can model nonlinear relationships between features and outcomes. Decision trees can handle both categorical and numerical features, which is common in traffic incident datasets. Decision trees are computationally efficient and can handle large datasets.

A learned function is used to represent a decision tree, which is a method for addressing discrete-valued target functions. These algorithms have been successfully applied to several tasks and are widely known for their ability to support inductive learning. A route from the root node toward the transaction's outcome or class label is shown once the decision tree is compared to the transaction value. To ascertain if a new transaction is real or fraudulent, this procedure is carried out for each one (Vanitha, 2023).

3.2.3 K-Nearest Neighbor (KNN)

One popular machine learning technique for classification and regression issues is the KNN. The KNN method looks at the k nearest observations in the feature space in order to categorize each observation. The class of the most recent observation is determined by taking the class that includes most of the k-nearest observations. As a result, KNN applies the class of the k closest set of already classified points to an unclassified sample point. When employing KNN, the modeler requested for two hyperparameters: the number of the k nearest neighbors and the distance function to be used to generate a metric for the distance between observations in the feature space (Fiorentini, & Losa, 2020).

KNN is an instance-based learning algorithm. It stores the entire training data set and makes predictions based on the similarity of instances. KNN captures local patterns in the data. For traffic incidents, considering nearby neighborhoods (similar road segments) is essential. KNN adapts well to changes in the data distribution. As traffic conditions vary over time, KNN can adjust accordingly. KNN makes no assumptions about the underlying data distribution, making it versatile for various scenarios. KNN excels in short-term prediction, which aligns with real-time traffic incident forecasting.

KNN is based on the assumption that related things are near together. As a result, it begins by establishing the value of k, which is the number of neighbors followed by the Euclidean distance between them (Ahmed, Hossain, Bhuiyan, & Ray, 2021).

3.2.4 State-of-the-Art Alternatives

Neural Networks (NNs): While NNs can handle complex relationships, they require large amounts of data and extensive tuning. For traffic incidents, where data availability may be limited, simpler models like RFs and Decision Trees perform well.

Support Vector Machines (SVMs): SVMs work well for binary classification but may not be ideal for multiclass traffic incident prediction. They also require careful hyperparameter tuning.

Deep Learning: Deep learning models like CNNs, LSTMs are powerful but often overkill for traffic incidents. They shine in image recognition and sequence data but may not provide significant advantages here. Deep Learning Models are good for complex patterns but require large data and resources. Graph Neural Networks (GNNs) are effective for structured data.

LSTM are excellent for time-series data e.g., traffic incidents over time, and captures temporal dependencies. These are used in sequence-to-sequence prediction. Convolutional Neural Networks (CNNs) are ideal for image-based features e.g., road camera images as these learn hierarchical features. CNN requires large, labeled datasets.

In summary, Random Forests, Decision Trees, and KNN strike a balance between accuracy, interpretability, and adaptability by making them excellent choices for predictive analytics of road traffic incidents. Their ability to handle nonlinear relationships, feature selection, and real-time prediction aligns well with the challenges posed by traffic data.

3.2.4 Performance Evaluation Measures

Accuracy

The assessment of models is a critical step in classification that involved the representation of several parameters. The assessment criteria that were most often utilized in this study are F-score, accuracy, precision, and recall.

Precision

Precision is the measure of a classifier's accuracy. It shows what percentage of all tuples having a positive label are positive.

F1-Score

The F-score is an analysis of statistical measures used for classification that considers the recall and accuracy of the classifier to produce a score that ranges from 0 to 1.

Recall

Recall shows the percentage of true positive tuples that have been correctly categorized and can also be referred to as an indicator of completeness.

3.3 Proposed Road Accident Prediction System

In this project, machine learning methods named Random Forest classifier, Decision Tree and K-Nearest Neighbor (KNN) are trained on the dataset from the US Accidents dataset sample from 2016-2023. The models have been trained to learn patterns and relationships between various features like weather conditions, road characteristics, and accident severity levels. The purpose of these models is to predict which situations are more likely to lead to accidents. After being trained, the model can predict the severity of new incidents based on their feature attributes. Finally, we can analyze which machine learning model performed well. Random forest model has an accuracy of approx. 84.1%, Decision Tree model has an accuracy of approx. 76% and K-Nearest Neighbor (KNN) model has an accuracy of approx. 83%. Hence, random forest model performed well.

Chapter 4 Analysis

4.1 Data Preprocessing and Exploration

4.1.1 Description of The Dataset

Between February 2016 and March 2023, around 2.25 million traffic accidents occurred in the continental United States. Accident records include location, time, description of language, weather, period of day, and points of interest. US accidents can be utilized for real-time accident prediction, hotspot analysis, causality analysis, and analysis of the impact of environmental factors on accident occurrences. The large dataset provides valuable insights for urban planning and transportation infrastructure improvement. The US Accidents dataset can be used with other traffic and meteorological events to identify patterns spanning large-scale geo-spatiotemporal data. Our findings provided insights into propagation and important patterns.

4.1.1.1 Real-time Traffic Data Collection

US accidents dataset had collected real-time traffic data from "MapQuest Traffic" and "Microsoft Bing Map Traffic". These APIs broadcast traffic events captured by various entities including the US and state departments of transportation, law enforcement agencies, traffic cameras, and sensors in road networks. Dataset had collected data every 90 seconds from 6am to 11pm and 150 seconds from 11pm to 6 am. Between February 2016 and March 2023 the gathered traffic accident instances were 2.27 million including 1.7 million from MapQuest and 0.54 million from Bing.

4.1.1.2 Integration of data

To integrate the data, duplicate instances were removed from both sources to create a consistent dataset. Two occurrences had been classified as duplications if the Haversine distance and recorded times were less than a heuristic criterion of 250 meters and 10 minutes, respectively. These conservative parameters had been chosen to minimize the probability of duplication. Using these parameters, it was discovered that around 24,600 duplicated accident entries, or 1% of the total data. After deleting duplicates, the final dataset had 2.25 million accidents.

4.1.1.3 Data Augmentation

Reverse Geo-Coding

Raw traffic accident reports only contain GPS data. The Nominatim program had been used to reverse geocode GPS coordinates into addresses, including street number, name, relative side, city, county, state, nation, and zip code. This approach is equivalent to point-wise map-matching.

Adding Weather data

Weather information provides valuable context for traffic incidents. Weather ground API was used to gather weather data for each accident. Raw weather data was gathered from 1,977 stations at airports around the US. Raw data comprises observation recordings with several variables including temperature, humidity, wind speed, pressure, precipitation in millimeters, and condition. Daily data was obtained from recordings from each weather station and reported any notable changes in the recorded weather parameters.

Supplementing with Points of Interest

Points of interest (POI) are map annotations that indicate facilities, traffic signals, and crossings. These annotations are linked to nodes on a road network. This study focuses on one category of 13 POI kinds linked with a node. Annotations for the United States are sourced from the most current dataset supplied by Open Street Map (OSM) extracted in April 2023. POI annotation for a traffic accident depends on the nearest POI within a threshold distance (τ). several threshold values were evaluated to discover the one that best associates a POI alongside an accident.

Regular expression patterns

27 regular expression patterns were identified for traffic accidents 16 from MapQuest and 11 using Bing data (Moosavi, Samavatian, Parthasarathy, & Ramnath, 2019).

4.1.2 Characteristics of The US Accidents Dataset

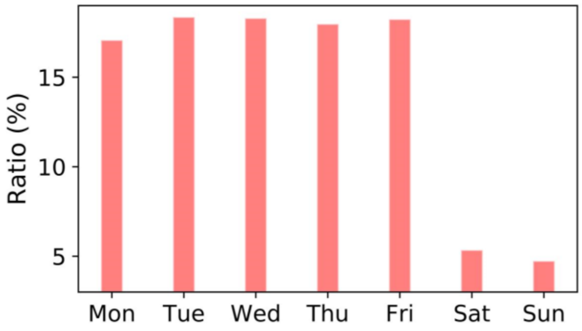


Figure 1. 1 Day of Week Analysis

The graph represents the number of accidents per day of the week according to the US Accidents Dataset.

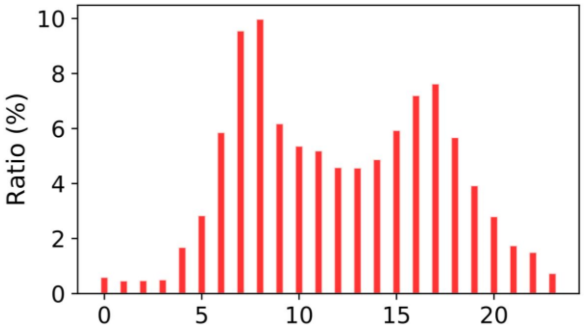


Figure 1. 2 Hour of The Day for Weekdays Analysis

The graph represents the variations in accident severity during different hours of the day according to the US Accidents Dataset.

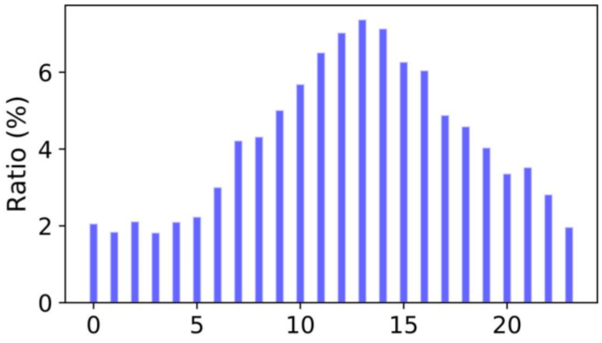


Figure 1. 3 Hour of The Day for Weekends Analysis

The graph represents the variations in accident severity monthly or in seasonal patterns according to the US Accidents Dataset.

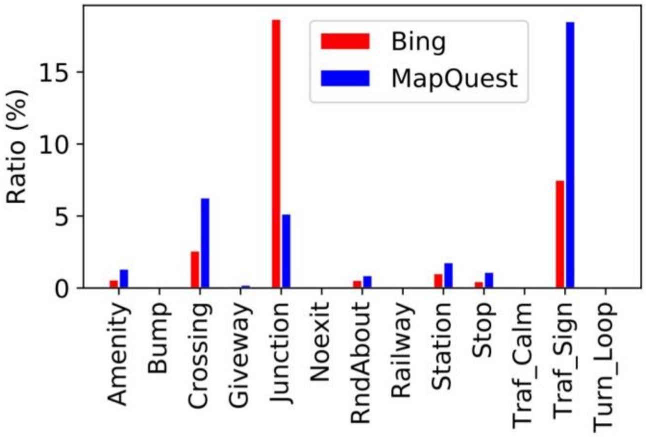


Figure 1. 4 Points of interest Distribution Analysis

This analysis examines how accidents correlate with specific landmarks. Points of interest (POIs) include intersections, schools, hospitals, or other relevant places. The distribution reveal accident hotspots near certain POIs. It reveal certain areas have higher accident rates due to nearby landmarks.

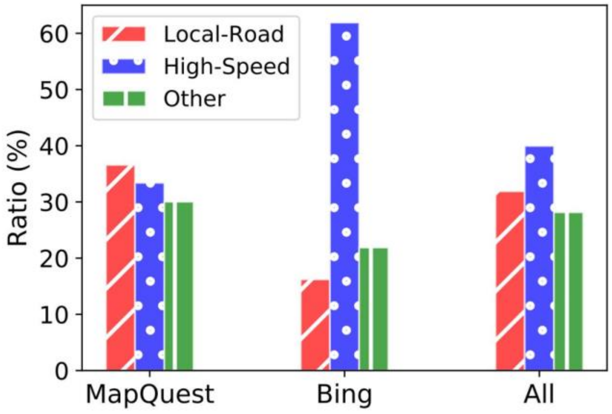


Figure 1. 5 Road Type Distribution Analysis

This graph examines the types of roads where accidents occur. It shows the proportion of accidents on highways, local roads, and all other road types.

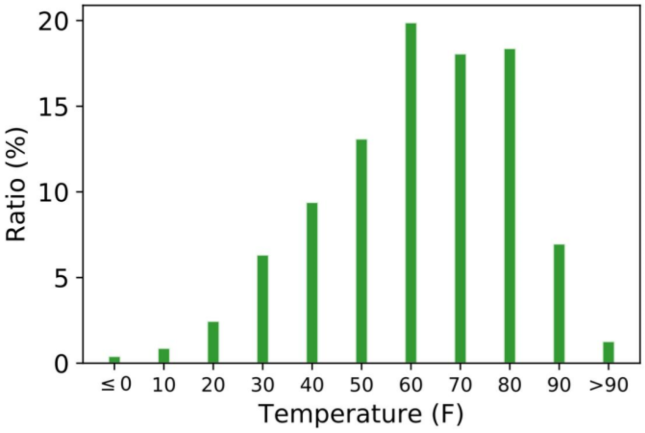


Figure 1. 6 Temperature Distribution Analysis

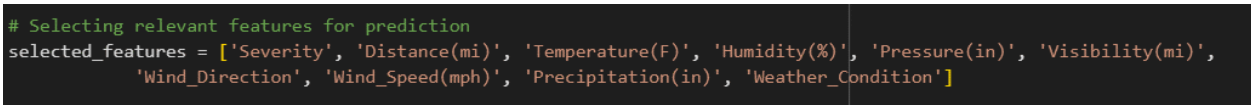
This graph shows the relationship between temperature and accidents. It shows accidents are more common during extreme temperatures (Moosavi, Samavatian, Parthasarathy, & Ramnath, 2019).

4.1.3 Data Collection for Project

The dataset sample was downloaded from the Kaggle website related to road traffic incidents. A dataset sample has been used for this project. There is a sampled version of the dataset with 500,000 accidents available for those who want something more manageable and smaller. This sample was extracted from the original dataset to facilitate management and analysis.

4.1.4 Feature Selection

Relevant Features were identified for development models and selected critical features. The selected features include Severity, Distance(mi), Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), and Weather_Condition.



```
# Selecting relevant features for prediction
selected_features = ['Severity', 'Distance(mi)', 'Temperature(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)',
            'Wind_Direction', 'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition']
```

Figure 1. 2 Selected Features

4.2 Machine Learning Model Development

4.2.1 Random Forest Classifier

Selected Random Forest (RF) due to its robustness and ability to handle complex relationships. The random forest classifier is made up of tree classifiers each of which is constructed utilizing a random vector sampled separately from the vector that was input and each of them casts a single vote for which is the most popular class to categorize the vector of inputs. The random forest classifier employed in this work grows a tree by selecting randomly or combining features at each node. Bagging a method for generating a training data set by freely drawing by replacement N samples, in which N is the total number of samples from the initially selected training set was employed with every feature combination chosen.

Any instances are categorized by selecting which is the most popular voted class across all predictors of trees in the forest. The design of a decision tree necessitated the selection of an attribute selecting measure and pruning process. There are several techniques for selecting characteristics for decision tree induction with the majority assigning a measure of quality directly to the attribute. The Information Gain Ratio criteria and the Gini Index are the two most commonly utilized attribute selection methods in decision tree induction. The Gini Index is used as an attribute choosing measure by the random forest classifier and it quantifies an attribute's impurity in relation to the classes. Each time a tree is developed to its highest level on new training data utilizing a number of features. These full-grown trees have not been trimmed. This is one of the primary benefits of the random forest classifier over previous decision tree approaches presented by Quinlan.

According to the research, the choice of pruning strategies rather than attribute selection measures affects the effectiveness of tree-based classifiers. According to Breiman, as the number of trees rises the generalization error usually converges regardless of pruning the tree, and overfitting is not an issue due to the Strong Law of Large Numbers. To construct a random forest classifier, two user-defined parameters must be specified as the number of characteristics utilized at each node to produce a tree and the total number of trees to be developed. At each node, just certain characteristics are sought for the optimum split. Therefore, the random forest classifier is made up of N trees where N is the total amount of trees to be produced and can be any value specified by the user. To categorize a new data set, each instance is assigned to one of the N trees. For that particular situation, the forest selects the class with the greatest votes out of N (Pal, 2005).

4.2.2 Decision Tree

By building a tree-like graph of alternatives and its potential outcomes, the decision tree approach makes use of conditional control statements to forecast the ultimate decision. The decision tree approach is a popular data mining method for creating systems of classification based on several variables or generating prediction algorithms for a specific variable. This approach divides the population into branch-like segments that form an inverted tree with root, internal, and leaf nodes. The method is non-parametric, meaning it can handle huge, complex datasets without enforcing a sophisticated parametric framework. When the sample size is high enough the research data can be separated into datasets for training and validation. Utilizing the data set for training to create a decision tree model and the validation dataset to determine the ideal tree size for the final model. A basic decision tree model has one binary goal variable Y (0 or 1) and two variables that are continuous x1 and x2 which vary from 0 to 1. The main elements of a decision tree model include nodes and branches, and the most significant processes in model construction are splitting, halting, and pruning (Song, & Ying, 2015).

4.2.3 K-Nearest Neighbor (KNN)

Using the k nearest observations in the feature space as a guide, the KNN algorithm attempts to categorize an observation. The KNN technique in pattern recognition uses training samples to categorize objects based on their similarity in feature space. KNN is a sort of instance-based or lazy learning that approximates functions locally and defers computation until classification.

The KNN is the most basic classifier for data with unknown distribution. This rule preserves the full training set throughout learning and allocates each query to a class based on the majority of the label of its k-nearest neighbors in the training set. When K = 1 the most basic form of KNN is the Nearest Neighbour (NN) rule. This approach requires that each sample be categorized similarly to the surrounding samples. If a sample's classification is unknown, it can be inferred based on the classification given to its nearest neighbors. A training set and an unknown sample can be used to compute distances between them. The training sample closest to the unidentified sample has the shortest distance. Therefore, the unidentified sample may be identified based on the categorization of its nearest neighbor. A KNN classifier's performance is mostly governed by its K and distance metric. The estimate is influenced by the sensitivity of the neighborhood size K. The radius of the local region is defined by the distance of the Kth nearest neighbor to the query, and different K values result in varied conditional class probabilities. Smaller K values result in poor local estimates due to scarce data and confusing points (Imandoust, & Bolandraftar, 2013).

4.3 Validation and Evaluation Procedures

4.3.1 Dataset Partitioning

It is the process of partitioning a dataset into training and testing sets that is required to assess the performance of machine learning models. Splitting the dataset into training 80% and testing 20% sets. Common methods for data splitting include:

i) Train-test Split: Separate the dataset into a training and testing set to evaluate the model's performance. Trained the RF, DT, and KNN models on the training data. Evaluated the RF, DT, and KNN models using the testing set.

ii) Cross-validation: Cross-validation involves splitting the dataset into subsets or folds and iteratively training and evaluating the model to ensure that every data point is used for training and testing.

4.3.2 Evaluation Metrics Used to Assess Model Performance

The measures used to evaluate the performance of machine learning models. It covers measurements like accuracy, precision, recall, and F1-score. Accuracy Score represents the overall correctness of predictions for the RF, DT, and KNN models. The classification Report represents the Precision, recall, and F1-score for each severity level of the RF, DT, and KNN models.

i) Classification Metrics

Classification metrics evaluate the effectiveness of machine learning models on classification tasks. They want to allocate each input data point to a single one of many predetermined categories.

ii) Accuracy

Accuracy is a key parameter for evaluating a classification model's overall performance. It is the ratio of successfully predicted occurrences to all occurrences in the dataset.

iii) Precision and Recall

Precision and recall are important assessment variables in machine learning for determining the trade-off between false positives and false negatives. Precision (P) is the percentage of genuine positive forecasts to all positive predictions. It measures how accurate positive forecasts are. Recall (R) referred to as true positive rate (TPR) is the fraction of genuine positive predictions out of all positive cases. It evaluates the classifier's capability to accurately identify positive cases.

iv) F1-score

The F1-score is a harmonic average of accuracy and recall, creating a metric that integrates both metrics. It is useful when working with unbalanced datasets in which one class is considerably more common than the other (Shah, 2023).

Chapter 5 Results

5.1 Results Interpretation

5.1.1 Accuracy

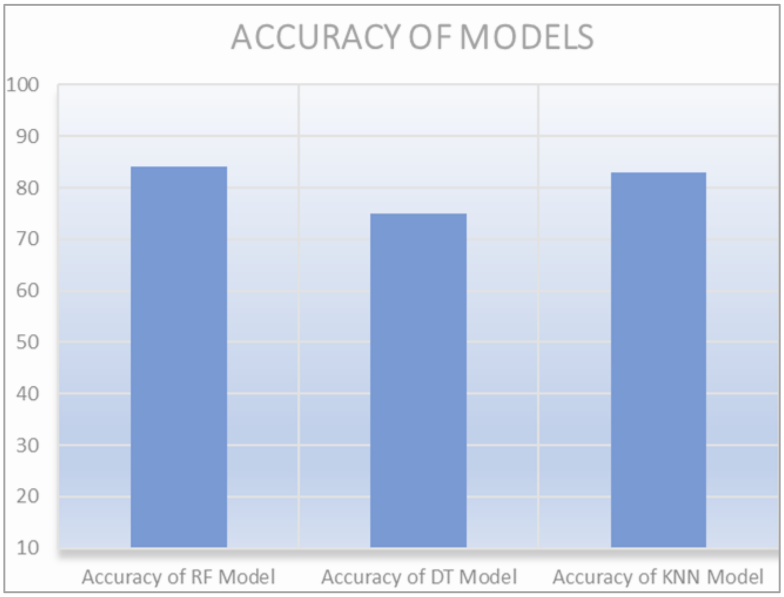


Figure 1.10 accuracy of models

The figure above shows the accuracy of the three models, where the Random Forest model has an accuracy of approximately 84%. The decision Tree model has an accuracy of approx.76%. K-Nearest Neighbor (KNN) model has an accuracy of approx. 83%. Hence, the analysis of accuracy for all models shows the random forest model performed well.

5.1.2 Classification Report

Random Forest model

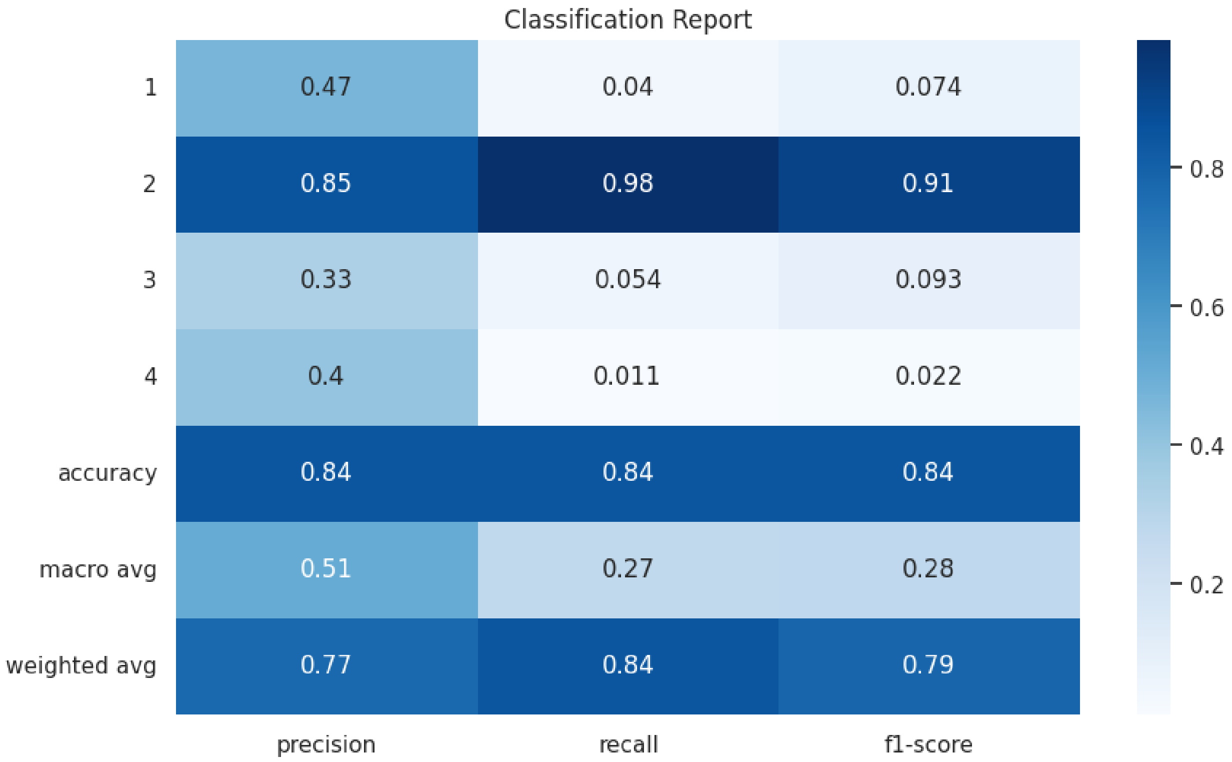
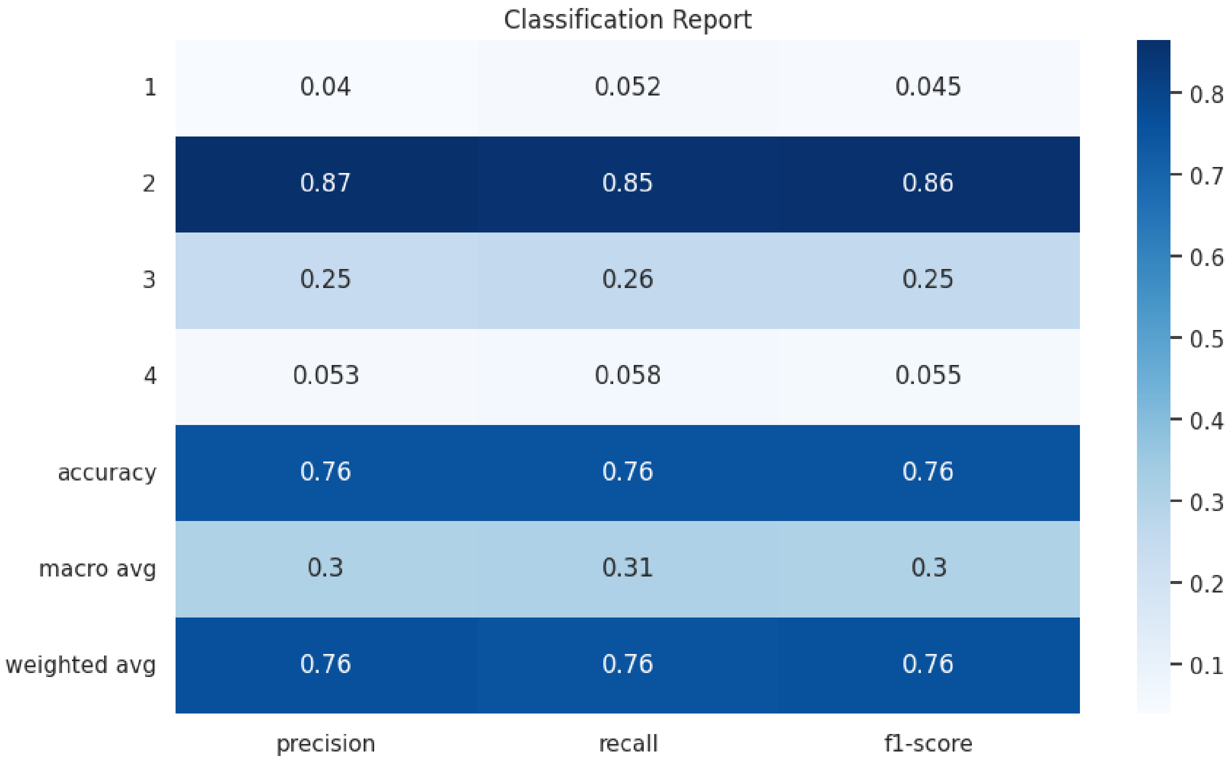


Figure 1. 9 Classification Report of Random Forest model

The Random Forest model results are as follows:

- Precision: Precision for all classes are 47%, 85%, 33%, and 40% respectively.
- Recall: Recall for all classes are 4%, 98%, 5%, and 1% respectively.
- F1-score: F1-score for all classes are 7%, 91%, 9% and 2% respectively.

Decision Tree model



Classification Report

	precision	recall	f1-score
1	0.04	0.052	0.045
2	0.87	0.85	0.86
3	0.25	0.26	0.25
4	0.053	0.058	0.055
accuracy	0.76	0.76	0.76
macro avg	0.3	0.31	0.3
weighted avg	0.76	0.76	0.76

Figure 1. 10 Classification Report of Decision Tree model

Decision Tree model results are as follows:

- Precision: Precision for all classes are 4%, 87%, 25%, and 5% respectively.
- Recall: Recall for all classes are 5%, 85%, 26%, and 6% respectively.
- F1-score: F1-score for all classes are 5%, 86%, 25% and 6% respectively.

K-Nearest Neighbor (KNN) model

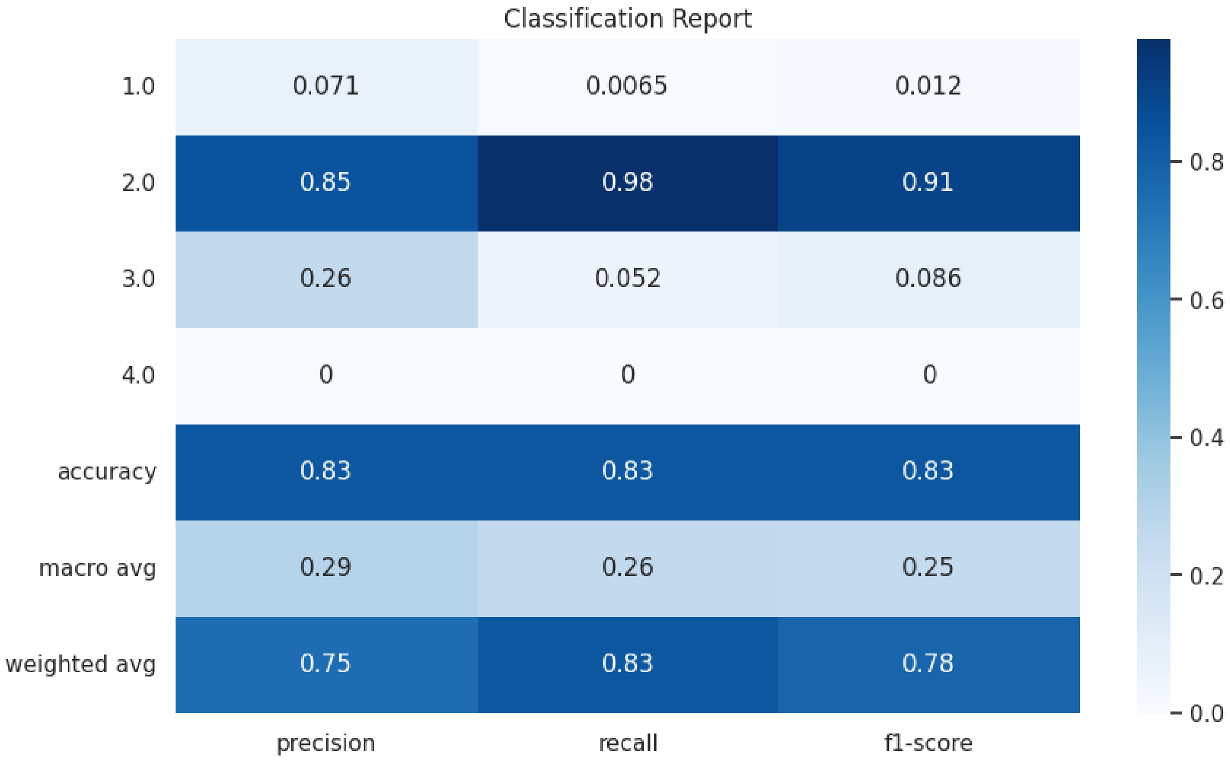


Figure 1. 11 Classification Report of K-Nearest Neighbor (KNN) model

K-Nearest Neighbor (KNN) results are as follows:

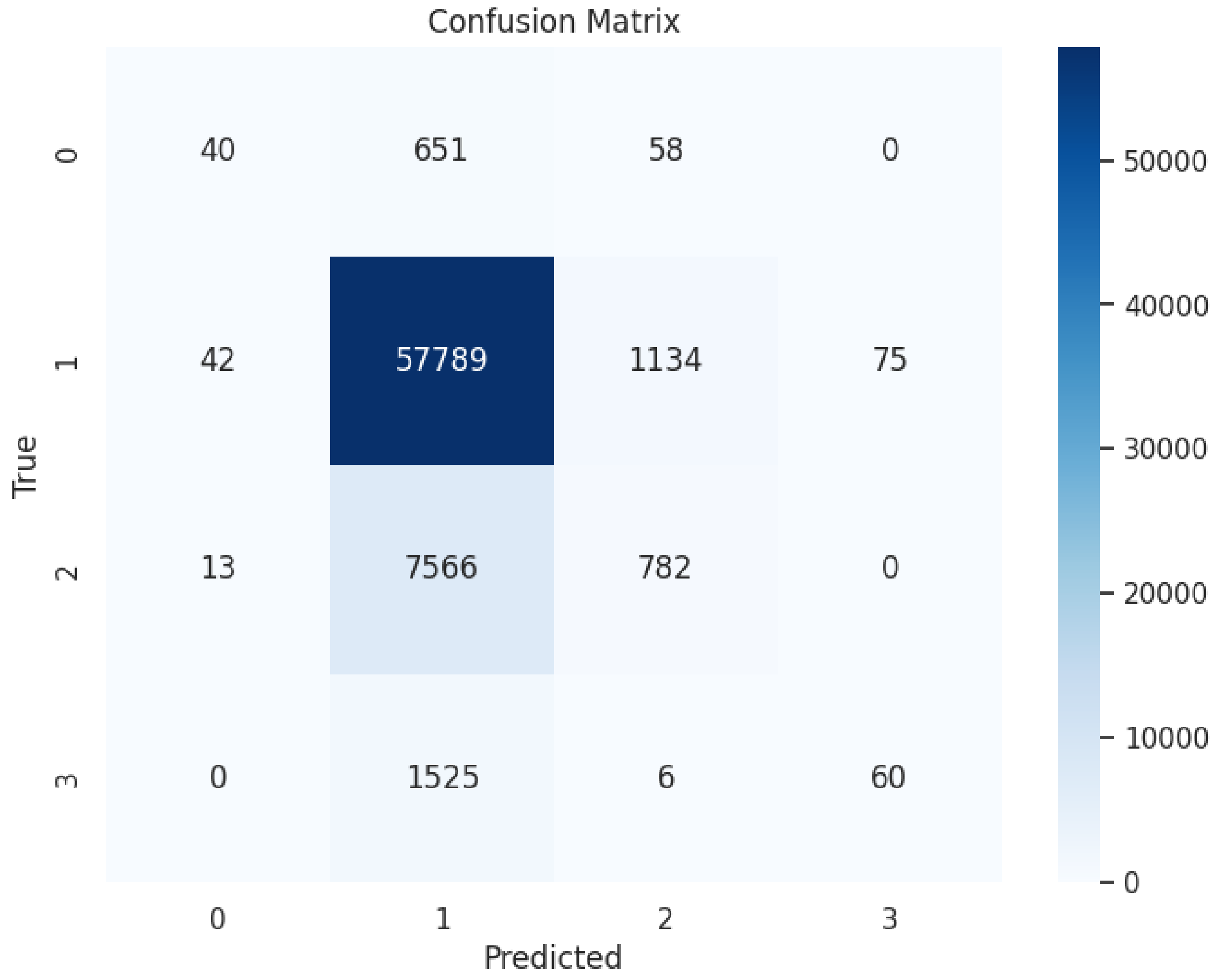
- Precision: Precision for all classes are 7%, 85%, 26%, and 0% respectively.
- Recall: Recall for all classes are 1%, 98%, 5%, and 0% respectively.
- F1-score: F1-score for all classes are 1%, 91%, 9% and 0% respectively.

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	84	0.77	0.84	79
Decision Tree	76	0.76	0.76	76
KNN	83	0.75	0.83	78

In conclusion, Random Forest Classifier model performs admirably across various metrics. Random Forest Classifier emerges as the top performer in terms of accuracy, precision, recall, and F1-score indicating its ability to accurately identify road accidents effectively.

5.1.3 Confusion Matrix

Random Forest model:

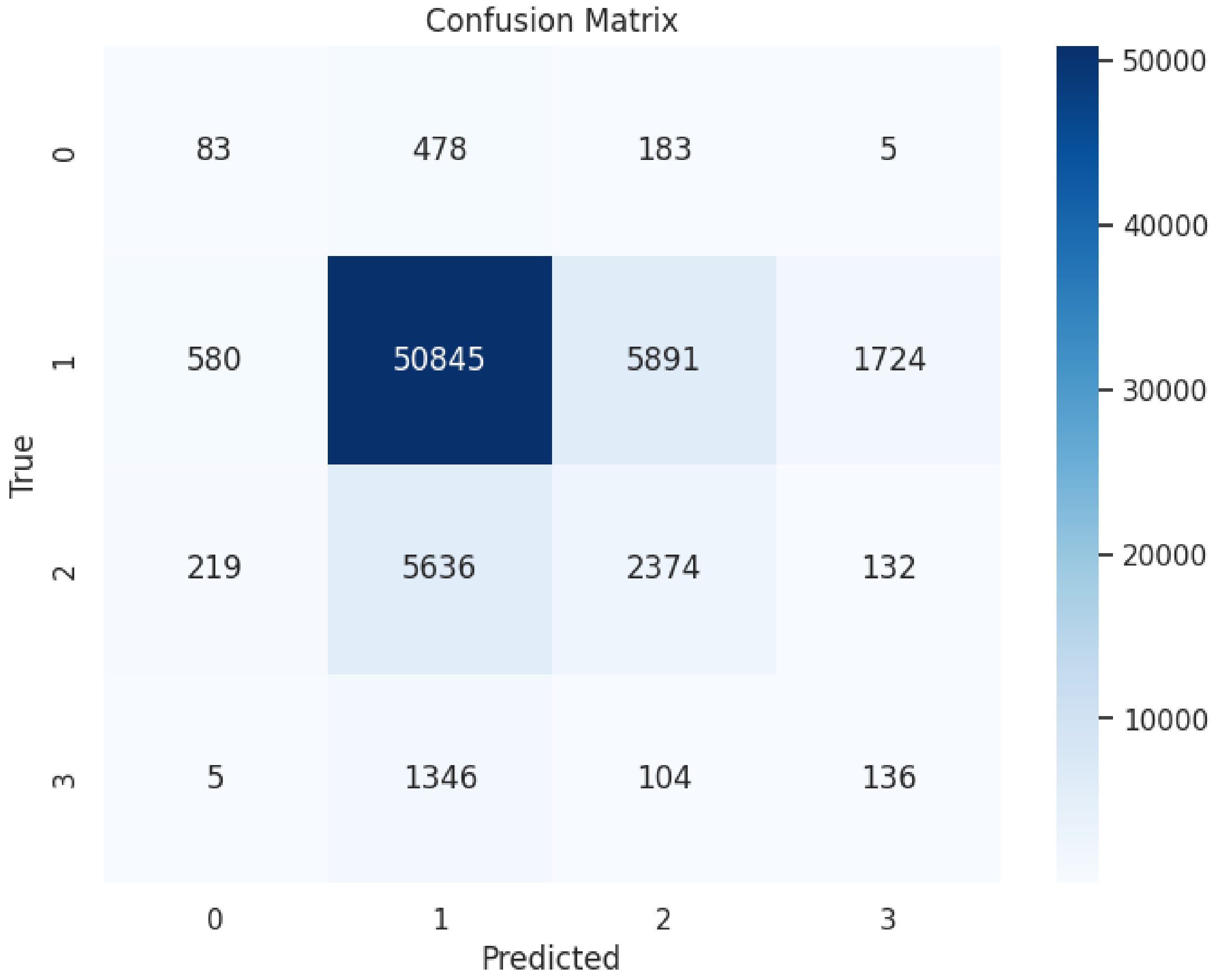


Confusion Matrix

True \ Predicted	0	1	2	3
0	40	651	58	0
1	42	57789	1134	75
2	13	7566	782	0
3	0	1525	6	60

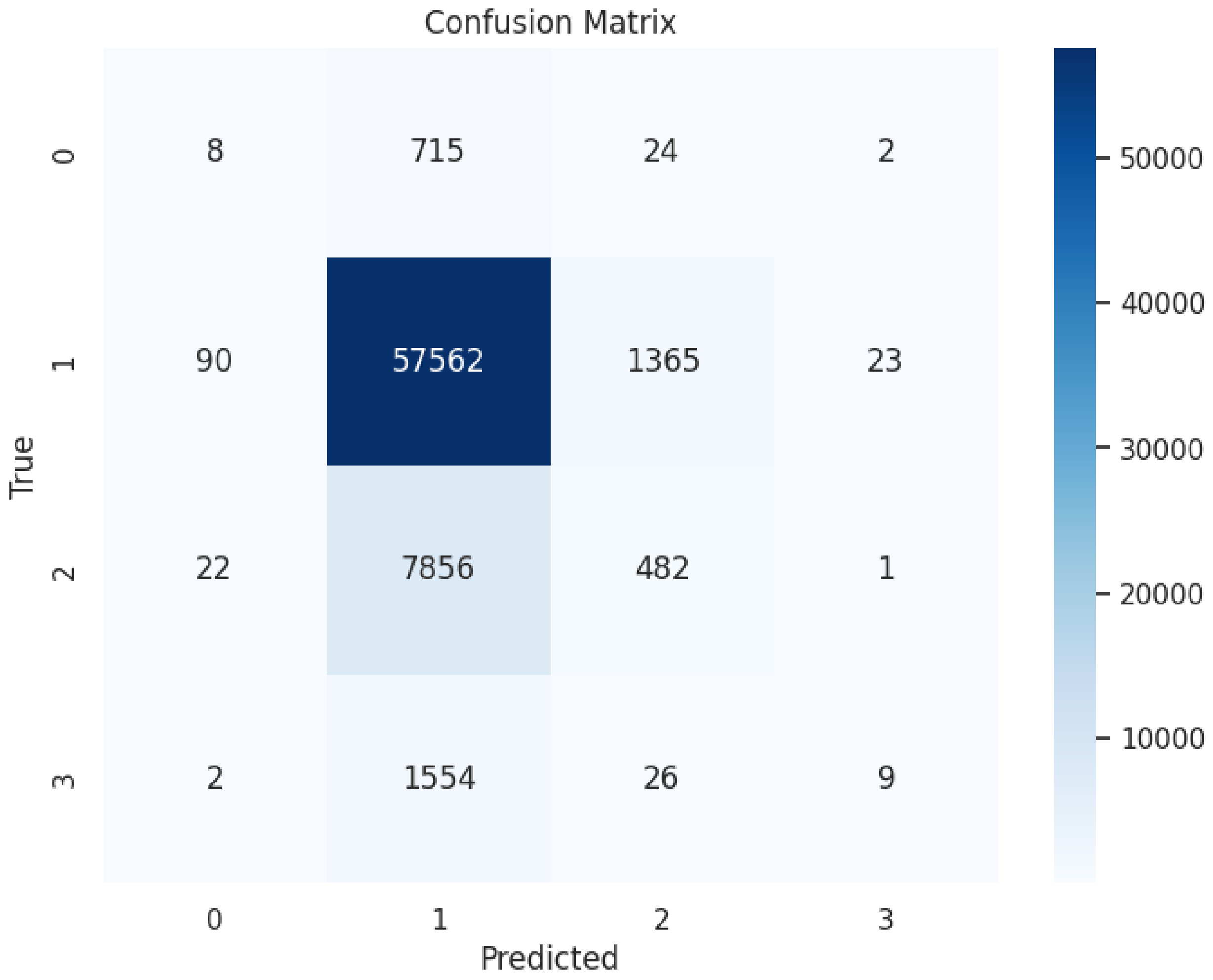
The confusion matrix shows that 57789 predictions are true by the Random Forest model.

Decision Tree:



The confusion matrix shows that 50845 predictions are true by the Decision tree model.

KNN Classifier:

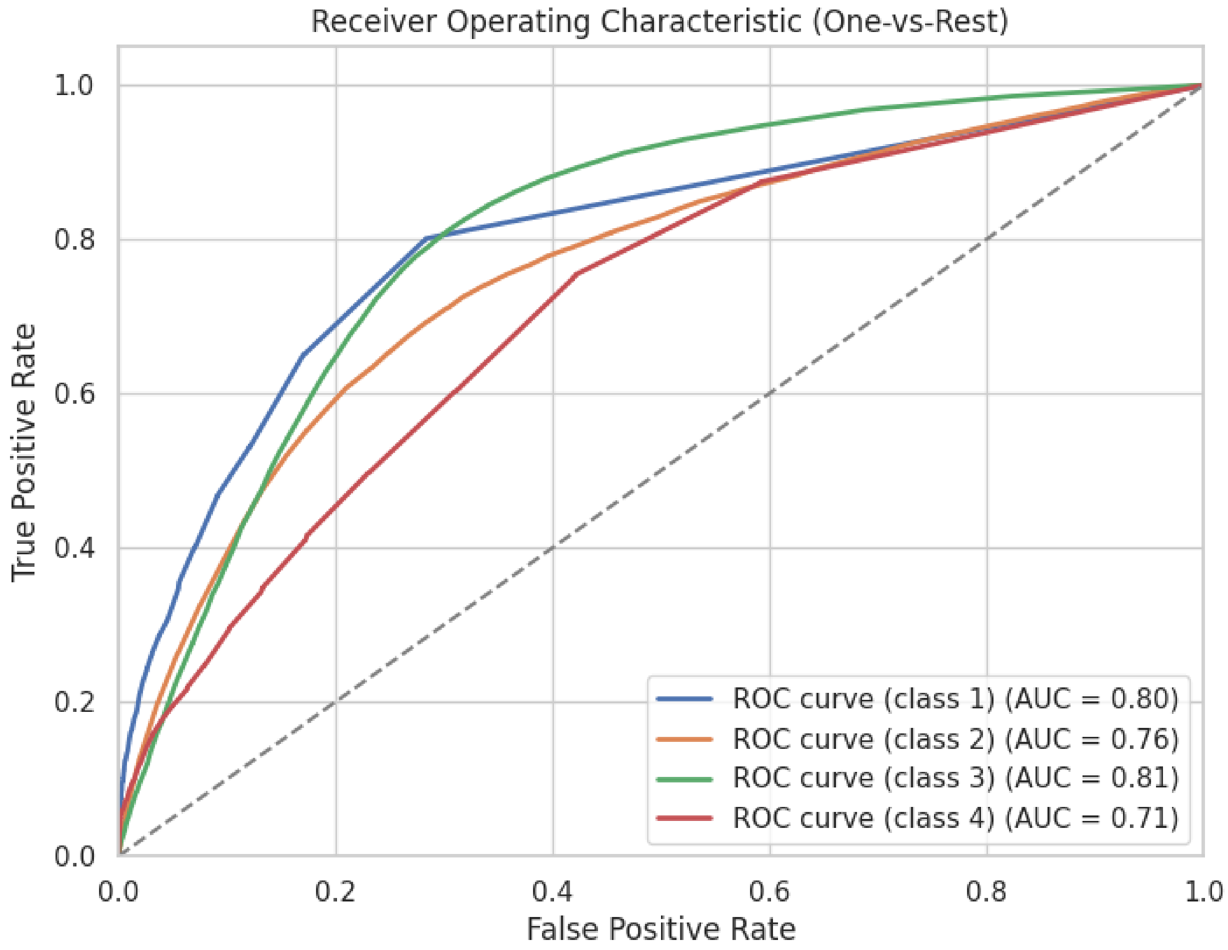


	0	1	2	3
0	8	715	24	2
1	90	57562	1365	23
2	22	7856	482	1
3	2	1554	26	9

The confusion matrix shows that 57562 predictions are true by the KNN model.

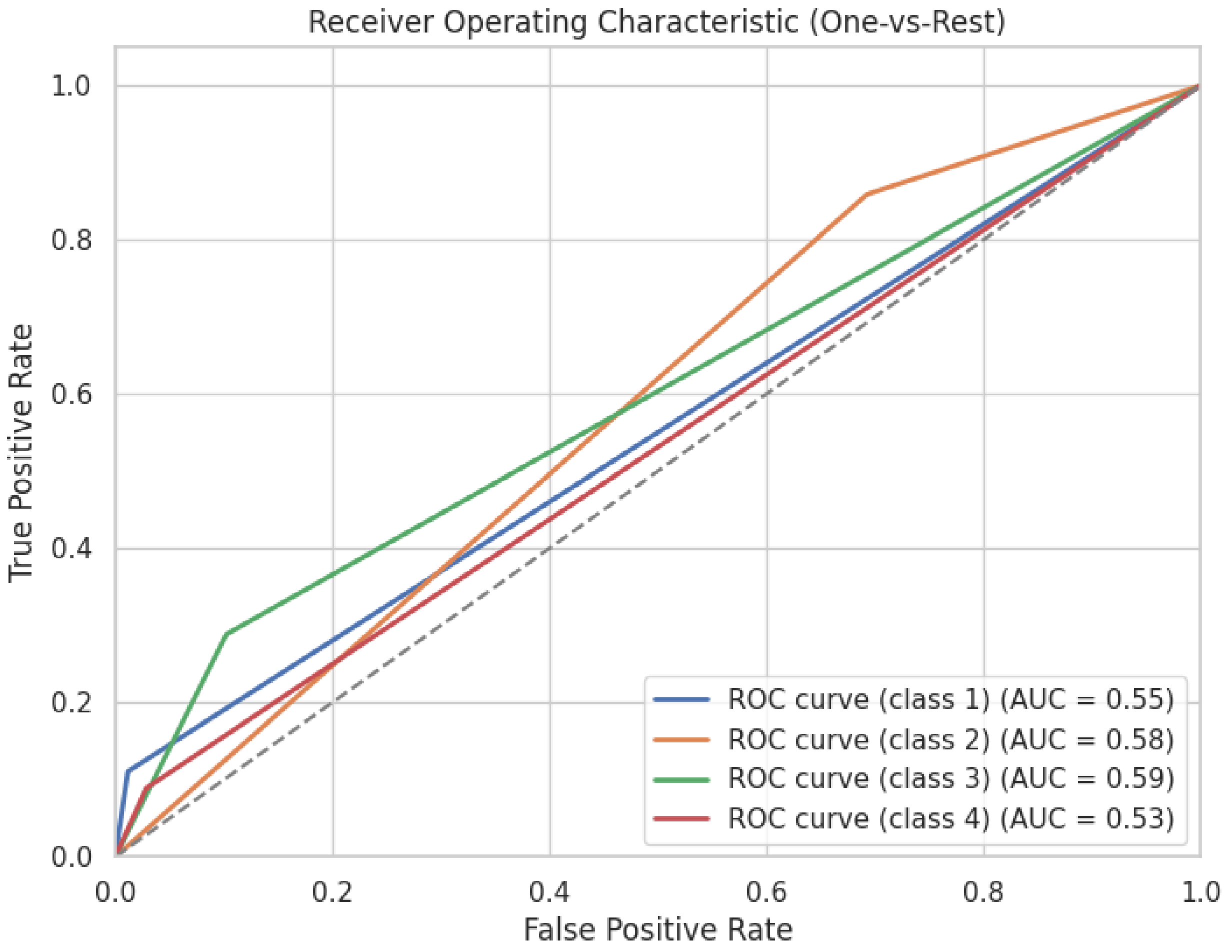
5.1.4 ROC Curve and AUC Area

Random Forest model:



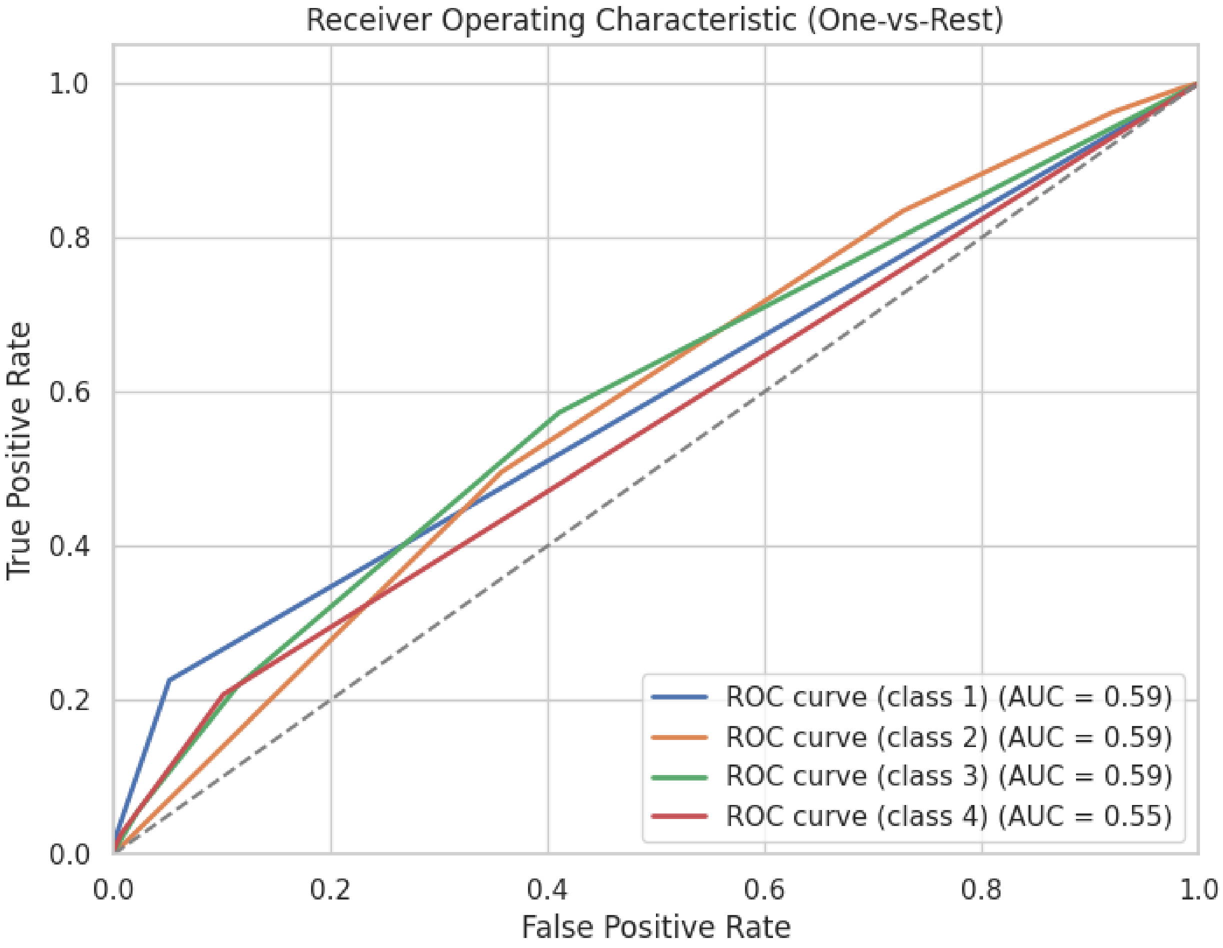
The x-axis shows the False Positive Rate (FPR), which is the proportion of negative instances that were incorrectly classified as positive. The y-axis shows the True Positive Rate (TPR), which is the proportion of positive instances that were correctly identified. The Area Under Curve (AUC) for each class is a numerical summary of the model's performance with a value between 0 and 1. Note that a perfect model would have an AUC of 1. The closer the ROC curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. From the graph it is obvious that Class 3 performed well with AUC = 0.81.

Decision Tree



The x-axis shows the False Positive Rate (FPR), which is the proportion of negative instances that were incorrectly classified as positive. The y-axis shows the True Positive Rate (TPR), which is the proportion of positive instances that were correctly identified. The Area Under Curve (AUC) for each class is a numerical summary of the model's performance with a value between 0 and 1. Note that a perfect model would have an AUC of 1. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. From the graph it is obvious that Class 3 performed well with AUC = 0.59.

KNN Classifier



The x-axis shows the False Positive Rate (FPR), which is the proportion of negative instances that were incorrectly classified as positive. The y-axis shows the True Positive Rate (TPR), which is the proportion of positive instances that were correctly identified. The Area Under Curve (AUC) for each class is a numerical summary of the model's performance with a value between 0 and 1. Note that a perfect model would have an AUC of 1. The closer the ROC curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. From the graph it is obvious that Class 1, 2 AND 3 performed equally with AUC = 0.59.

Hence, according to the ROC curve and AUC values the random forest classifier performed better than decision tree and KNN classifiers.

Chapter 6 Conclusions & Recommendations

6.1 Conclusions

This study predicts the severity of traffic accidents using a machine learning model called KNN, Random Forest, and Decision Tree. The research's experimental results show that Random Forest classification operates better than Decision Tree and KNN. The model performs well overall, with an accuracy of 84%, but there is definitely room for improvement, particularly when it comes to managing minority classes. The Random Forest algorithm, Decision Tree, and KNN-based predictive analytics investigation on the USA accidents dataset produced promising findings for anticipating the severity of road traffic occurrences. The algorithms were able to identify occurrences with a high degree of accuracy using important criteria such as temperature, visibility, wind speed, weather, dawn, and sunset. The classification reports highlighted possible problems and opportunities for development by offering insights into the model's performance across various severity levels. It can be argued that Random Forest is the most effective and efficient model of all because it regularly surpasses all other models in forecasting the severity of accidents.

6.2 Recommendations

It is recommended to examine other attributes or feature combinations that could improve the ability of the models to predict the future. Adding data on past traffic, road infrastructure, or local demographics might yield insightful information. To maximize model performance, experiment with additional machine learning strategies. Improving the model's ability to identify the minority classes would likely improve overall performance.

In order to address class imbalance and enhance model generalization, methods like oversampling, undersampling, or the use of class weights can be used. This is because the severity levels of accidents may not be spread equally throughout the dataset.

To help with road traffic incident management and prevention, when the models have been improved and verified, think about implementing the best one in real-time systems. This can entail creating a stand-alone application for consumers. It will help by integrating the best model into real-time traffic management systems.

6.3 Future Work

In order to take proactive action, more resources with ongoing forecasts and alerts might be supplied to the police for each area at regular intervals in the future. It can be integrated with Google Maps, allowing authorities to follow it in real-time. It is possible to provide a completely functional web application for real-time use by users and transportation authorities or police.

Analysis of the geographical and temporal patterns of traffic events can be used to pinpoint areas and accident peak times. To identify the underlying causes of accidents, this can entail time series modeling and geographic analysis.

To improve the timeliness and precision of accident prediction, investigate the integration of real-time data sources such as traffic cameras, weather stations, and social media feeds. This could make it possible to take preventative action to lessen the effects of accidents.

To obtain a thorough knowledge of road traffic occurrences expand the analysis to include different data modalities such as text data from accident reports, photos from traffic cameras, and sensor data from automobiles. By determining which factors have the most influence on the model's predictions. Interpretability analysis may assist in identifying possible areas for development and contribute to the creation of better interpretable models.

The model can be better able to capture the fundamental trends of all classes, especially the minority ones if the dataset is larger and more diverse. Investigating cutting-edge machine learning methods can improve the model's capacity to identify and make use of intricate patterns in the data.

References

[1] Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. Computers, 10(12), 157.

[2] Vanitha, R. (2023). Prediction of Road Accidents using Machine Learning Algorithms. Middle East Journal of Applied Science & Technology (MEJAST), 6(2), 64-75.

[3] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.

[4] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42).

[5] Li, R., Pereira, F. C., & Ben-Akiva, M. E. (2018). Overview of traffic incident duration analysis and prediction. European transport research review, 10(2), 1-13.

[6] Hireche, S., & Dennai, A. (2020). Machine learning techniques for road traffic automatic incident detection systems: A review. Smart Energy Empowerment in Smart and Resilient Cities: Renewable Energy for Smart and Sustainable Cities, 60-69.

[7] Sahu, S., Maram, B., Gampala, V., & Daniya, T. (2022). Analysis of Road Accidents Prediction and Interpretation Using KNN Classification Model. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 2 (pp. 163-172). Singapore: Springer Nature Singapore.

[8] Ferreira-Vanegas, C. M., Vélez, J. I., & García-Llinás, G. A. (2022). Analytical methods and determinants of frequency and severity of road accidents: a 20-year systematic literature review. Journal of advanced transportation, 2022.

[9] Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences, 2, 1-13.

[10] (Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M. A., Mohabbati-Kalejahi, N., Vinel, & Megahed, 2020). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. Sensors, 20(4), 1107.

[11] Razali, N. A. M., Shamsaimon, N., Ishak, K. K., Ramli, S., Amran, M. F. M., & Sukardi, S. (2021). Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning. Journal of Big Data, 8(1), 1-25.

[12] Sharma, B., Katiyar, V. K., & Kumar, K. (2016). Traffic accident prediction model using support vector machines with Gaussian kernel. In Proceedings of Fifth International Conference on Soft Computing for Problem Solving: SocProS 2015, Volume 2 (pp. 1-10). Springer Singapore.

[13] Wahab, L., & Jiang, H. (2020). Severity prediction of motorcycle crashes with machine learning methods. International journal of crashworthiness, 25(5), 485-492.

[14] Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24, 62-72.

[15] Li, L., Shrestha, S., & Hu, G. (2017, June). Analysis of road traffic fatal accidents using data mining techniques. In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 363-370). IEEE.

[16] AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT) (pp. 272-276). IEEE.

[17] Tiwari, P., Kumar, S., & Kalitin, D. (2017). Road-user specific analysis of traffic accident using data mining techniques. In Computational Intelligence, Communications, and Business Analytics: First International Conference, CICBA 2017, Kolkata, India, March 24–25, 2017, Revised Selected Papers, Part II (pp. 398-410). Springer Singapore.

[18] Regassa, Z. (2009). “Determining the degree of driver’s responsibility for car accident: the case of addis ababa traffic office. Unpublished Master’s Thesis. Addis Ababa University.

[19] Getnet, M. (2009). Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic accidents: the case of Addis Ababa city. Addis Ababa Addis Ababa University.

[20] Hébert, A., Guédon, T., Glatard, T., & Jaumard, B. (2019, December). High-resolution road vehicle collision prediction for the city of montreal. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 1804-1813). IEEE.

[21] Siam, Z. S., Hasan, R. T., Anik, S. S., Dev, A., Alita, S. I., Rahaman, M., & Rahman, R. M. (2020). Study of machine learning techniques on accident data. In Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12 (pp. 25-37). Springer International Publishing.

[22] Theofilatos, A., Graham, D., & Yannis, G. (2012). Factors affecting accident severity inside and outside urban areas in Greece. Traffic injury prevention, 13(5), 458-467.

[23] Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention, 108, 27-36.

[24] Lin, L., Wang, Q., & Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies, 55, 444-459.

[25] Chang, L. Y., & Chen, W. C. (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research, 36(4), 365-375.

[26] Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. Accident Analysis & Prevention, 39(4), 657-670.

[27] 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.

[28] Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. Journal of safety research, 61, 9-21.

[29] Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing machine learning and deep learning methods for real-time crash prediction. Transportation research record, 2673(8), 169-178.

[30] Ren, H., Song, Y., Wang, J., Hu, Y., & Lei, J. (2018, November). A deep learning approach to the citywide traffic accident risk prediction. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 3346-3351). IEEE.

[31] Yuan, Z., Zhou, X., & Yang, T. (2018, July). Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 984-992).

[32] Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. Journal of traffic and transportation engineering (English edition), 7(4), 432-446.

[33] Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24, 62-72.

[34] Montella, A. (2010). A comparative analysis of hotspot identification methods. Accident Analysis & Prevention, 42(2), 571-581.

[35] Ahmed, S., Hossain, M. A., Bhuiyan, M. M. I., & Ray, S. K. (2021, December). A comparative study of machine learning algorithms to predict road accident severity. In 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS) (pp. 390-397). IEEE.

[36] Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), 61.

[37] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

[38] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[39] Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, 3(5), 605-610.

[40] Shah, D. (2023). Top Performance Metrics in Machine Learning: A Comprehensive Guide. https://www.v7labs.com/blog/performance-metrics-in-machine-learning.