Traffic Injury Severity Prediction by Using Machine Learning

Author

Muhammad Yasir	UET/SCET-19f-CE-83
BILAL HASSAN	UET/SCET-19f-CE-61
WALEED AHMAD	UET/SCET-19f-CE-58
AWAIS KHAN	UET/SCET-19f-CE-37
JUNAID AKHTAR	UET/SCET-19f-CE-81

Supervisor

ENGR. KIFFAYAT ULLAH

Lecturer

DEPARTMENT OF CIVIL ENGINEERING

SWEDISH COLLECGE FO ENGINEERING AND TECHNOLOGH

WAH CANTT AFFLIATED WITH UET TAXILA

MAY 2020

Traffic Injury Severity Prediction by Using Machine Learning

Author

Muhammad Yasir	UET/SCET-19f-CE-83
BILAL HASSAN	UET/SCET-19f-CE-61
WALEED AHMAD	UET/SCET-19f-CE-58
AWAIS KHAN	UET/SCET-19f-CE-37
JUNAID AKHTAR	UET/SCET-19f-CE-81

A thesis completed as part of the requirements for the
B.Sc. in Civil Engineering

Thesis supervisor
ENGR. KIFFAYAT ULLAH
Lecturer in Civil Engineering Department

External Examiner Signature: _____

Thesis Supervisor Signature: _____

Traffic Injury Severity Prediction by Using Machine Learning

Sustainable Development Goals

(Please tick the relevant SDG(s) linked with FYDP)

SDG No	Description of SDG	SDG No	Description of SDG
SDG 1	No Poverty	SDG 9	Industry, Innovation, and Infrastructure
SDG 2	Zero Hunger	SDG 10	Reduced Inequalities
SDG 3	Good Health and Well Being	SDG 11	Sustainable Cities and Communities
SDG 4	Quality Education	SDG 12	Responsible Consumption and Production
SDG 5	Gender Equality	SDG 13	Climate Change
SDG 6	Clean Water and Sanitation	SDG 14	Life Below Water
SDG 7	Affordable and Clean Energy	SDG 15	Life on Land
SDG 8	Decent Work and Economic Growth	SDG 16	Peace, Justice and Strong Institutions
		SDG 17	Partnerships for the Goals

Range of Complex Problem Solving			
	Attribute	**Complex Problem**	
1	Range of conflicting requirements	Involve wide-ranging or conflicting technical, engineering and other issues.	
2	Depth of analysis required	Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models.	
3	Depth of knowledge required	Requires research-based knowledge much of which is at, or informed by, the forefront of the professional discipline and which allows a fundamentals-based, first principles analytical approach.	
4	Familiarity of issues	Involve infrequently encountered issues	
5	Extent of applicable codes	Are outside problems encompassed by standards and codes of practice for professional engineering.	
6	Extent of stakeholder involvement and level of conflicting requirements	Involve diverse groups of stakeholders with widely varying needs.	
7	Consequences	Have significant consequences in a range of contexts.	
8	Interdependence	Are high level problems including many component parts or sub-problems	
Range of Complex Problem Activities			
	Attribute	**Complex Activities**	
1	Range of resources	Involve the use of diverse resources (and for this purpose, resources include people, money, equipment, materials, information and technologies).	
2	Level of interaction	Require resolution of significant problems arising from interactions between wide ranging and conflicting technical, engineering or other issues.	
3	Innovation	Involve creative use of engineering principles and research-based knowledge in novel ways.	
4	Consequences to society and the environment	Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation.	
5	Familiarity	Can extend beyond previous experiences by applying principles-based approaches.	

ABSTRACT

In Rawalpindi, Pakistan, traffic accidents are a growing source of concern and a threat to public safety. Our research utilized cutting-edge machine learning methods to examine and simulate historical accident data in order to address this deteriorating situation. We have a thorough dataset documenting the 836 incidents that were reported in Rawalpindi between 2017 and 2019. On this data, we next trained and assessed the performance of CAT Boost, Light GBM, XG Boost, Logistic Regression, and Random Forest, five intelligent classification methods. In comparison to other methods, the CAT Boost model excelled, achieving a remarkable accuracy of 97.7% on test data that had never been seen before. All models demonstrated exceptional precision, recall, and F1 scores, attesting to their ability to accurately classify different levels of accident severity. Investigating the CAT Boost model produced some interesting findings. Age was the most important factor for non-fatal results, while factors including accident cause, timing, and road type were most important for fatal outcomes. This highlights important differences between the variables influencing each severity band. These findings suggest undertaking targeted measures to improve road safety, such as boosting awareness, tightening up traffic laws, and improving infrastructure. Overall, the excellent performance of our models shows their potential for use in live monitoring and data-driven decision-making to reduce traffic injuries. Future research can expand on our work to improve predictive capabilities by having access to richer datasets and cutting-edge deep learning models. Our work highlights the significant benefits of using AI for evidence-based policymaking to improve traffic safety.

UNDERTAKING

I attest that the study with the title "**Traffic Injury Severity Prediction by Using Machine Learning"** my own creation. The work hasn't been submitted anywhere else for review. When information came from another source, it was duly acknowledged or cited.

_____

Muhammad Yasir

UET/SCET-19f-CE-83

_____

BILAL HASSAN

UET/SCET-19f-CE-61

_____

WALEED AHMAD

UET/SCET-19f-CE-58

_____

AWAIS KHAN

UET/SCET-19f-CE-37

_____

JUNAID AKHTAR

UET/SCET-19f-CE-81

ACKNOWLEDGEMENTS

It gives me great joy to share my sincere gratitude to Engineer Kiffayat Ullah, my advisor at the Swedish College of Engineering and Technology, Wah Cantt. As well as his guidance, Engineer Haroon Ali Shah provided critical insights and support, which helped me to improve my research at every stage. The research was further acknowledged by [INSERT HERE] for their significant contributions. Finally, I want to express my gratitude to my coworkers, family, and friends for their unwavering support and encouragement during this process. They have motivated me and driven me to succeed. Ultimately, this work would not have been possible without the collective support and inspiration of all the individuals mentioned, as well as many more unnamed individuals. I am indebted to all for their contributions, direct and indirect, to this research. This work would not have been possible without everyone's collective support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background of The Topic:

Traffic collisions in recent times have been a major threat to public health and safety around the world, bringing about substantial suffering, financial losses, and societal disruptions. In nations like Pakistan, where road travel is the main source of transportation, the frequency and severity of traffic injuries are significant concerns. Modern technologies, such as machine learning, have been investigated to address these concerns, including reducing the effects of accidents and improving road safety. Machine learning can anticipate the severity of traffic injuries by examining accident data and finding trends and associations that more conventional statistical analysis techniques might overlook. It is feasible to create prediction models using algorithms like XG Boost, random forests, CAT Boost, Light GBM, and logistic regression that can help us comprehend the variables impacting accident severity and assist us in creating more efficient traffic safety measures.

Machine learning methods are utilized in this work to forecast the severity of road injuries. The main goal of this research study is to evaluate critical elements influencing the severity of accidents and assess the effectiveness of various machine learning methods to forecast accident severity through the application of different machine learning models to a large dataset. The collection contains a plethora of information, including variables like weather conditions, road parameters, driver profiles, and vehicle information.

The main focus of the study is to develop accurate and precise models to help envisage the severity of traffic injuries using the available dataset. It is vital to study the patterns and insights produced from these models to gain holistic comprehension of the correspondence among key contributing elements and the severity of accidents. The incidence and severity of road accidents can be decreased by using this knowledge to establish laws, policies, and interventions for their prevention.

The results of using machine learning algorithms on the dataset will be thoroughly analyzed in this thesis. This will involve evaluating the model's effectiveness, identifying key characteristics, and interpreting the patterns and laws the model produces. This study intends to add to the body of knowledge in the subject and offer useful insights for enhancing road safety tactics by merging the findings with existing traffic safety literature.

In conclusion, the possible outcomes of this study will significantly help in the accurate anticipation of severity of a traffic injury using novel and advanced tools of machine learning. It attempts to amplify the process to make data driven and research-based decisions in the formulation of policies for road safety. The whole process of policy making is based on identification of key variables and formulation of reliable models. The ultimate objective is to lessen the recurrence of traffic accidents as well as to emphasize the significant curtailment of accident severity, protecting lives and advancing environmentally friendly transportation methods.

The actual practice of artificial intelligence-based machine learning models to envisage the severity of traffic has been the center of focus for several important studies, which have shaped the field's understanding and technique today. Ji and Levinson (2023) predicted the severity of injuries in two-vehicle accidents using ensemble machine learning models like Gradient Boosting Machines and Random Forest. As a result of a stacking method, their research showed that these models had a high degree of prediction accuracy. Their research significantly improved forecast accuracy by highlighting energy absorption as a key factor. Their findings support the application of ensemble machine learning techniques, which will be adopted and further explored in this thesis. [1]

Arteaga, Paz, and Park (2020) then presented a novel approach for assessing the seriousness of traffic accidents. By using text mining and the interpretable machine learning method known as Global Cross-Validation Local Interpretable Model-Agnostic Explanations (GCV-LIME), they may provide a full understanding of the causality factors related to high injury-severity levels in crashes. Through their analysis of heavy vehicle crash data in Queensland, Australia, they found terms like "collided head on," "side collided," "motorbike," "cab," and "pedestrian" were significantly

associated with fatal crashes.[2]. This provides important information on probable factors that can affect accident severity, an important issue that will be taken into account in this study.

Further, Ahmed, Hossain, Bhuiyan, and Ray (2023) used a variety of machine learning algorithms to predict the severity of traffic accidents while accounting for a variety of contributing factors, such as road geometry, environmental conditions, weather, and human characteristics, including age, alcohol and drug use. Using both single mode and ensemble mode ML algorithms, the accident severity was split into binary and multiclass categories. Random Forest (RF) outperforms other methods, such as Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Extreme Gradient Boosting (XG Boost), and Adaptive Boosting (AdaBoost), in terms of forecasting accident severity, according to their research [3]. When developing the strategy for our predictive models, the revelations from their research will be taken into account.

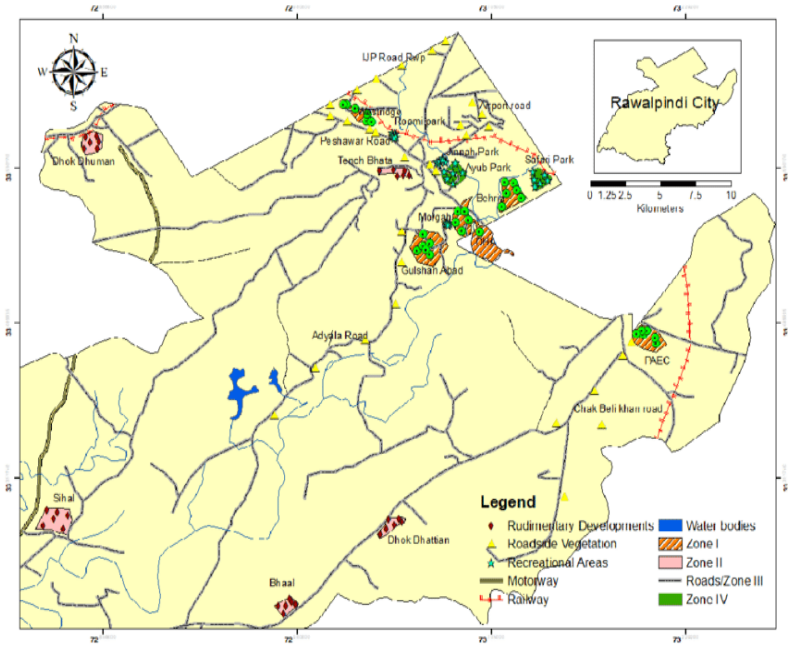


Figure 1. 1: Rawalpindi city road network map

1.2 Problem Statement

The need for precise injury severity prediction models is critical given the huge financial and human losses caused by transportation accidents. Traditional approaches frequently lack the objectivity and dependability needed to address the complex issues surrounding traffic accidents. In Rawalpindi, where there is a lack of knowledge on the effectiveness and applicability of modern machine learning models in this situation, the problem is particularly significant. The significant hurdles still lie in developing and putting into practice practical road safety measures based on the predictive features identified by these models.

Significantly, Umer, Sadiq, Ishaq, Ullah, Saher, and Madni (2023) have achieved progress in their research by comparing several ensemble regression and tree-based algorithms for determining the severity of traffic accidents. When considering 20 significant characteristics that were connected with accident severity, they determined that the Random Forest model was superior since it had higher accuracy, precision, recall, and F-score [4]. Their study offers insightful information that helps to improve traffic management and safety.

By creating and analyzing machine learning models using large datasets, this study intends to advance previous research. The goal is to solve Rawalpindi's existing lack of reliable prediction models for the severity of traffic injury. This project intends to support improved emergency response, resource allocation, and evidence-based policies through a more precise knowledge and forecast of road traffic injuries. This will ultimately reduce injuries brought on by traffic and advance the general goal of safer transportation.

1.3 Historical Background

Even though understanding their historical context is crucial for improving road safety, traffic accidents still happen. Road accidents have been a persistent problem throughout history, becoming more serious with the development of motor vehicles in the 19th and 20th centuries. The traditional methods, including road engineering, have been very important in increasing safety. However, the contemporary period, marked by unheard-of technological developments, has created new potential for reducing road accidents.

Chakraborty, Gates, and Sinha (2023) carried out an important study investigating cause analysis and injury severity classification of traffic crashes in this setting. They examined data from all Texas interstates from 2014 to 2019 [5] using non-parametric approaches and machine learning techniques like decision trees, random forests, extreme gradient boosting, and deep neural networks. Their research demonstrated the effectiveness of machine learning in determining important variables influencing accident severity and finding performance variances across various severity classes.

Furthermore, utilizing machine learning techniques, Mafi, AbdelRazig, and Doczy (2023) created a study that examined the severity of driver injuries across a range of age and gender categories [6]. They found that cost-sensitive learning classifiers, such as C4.5, instance-based (IB), and random forest (RF) models, performed better than conventional classifiers at predicting injuries and fatalities. This paper emphasized the potential for machine learning models to aid in developing tailored safety measures that are sensitive to the particular needs of various driver groups.

Additionally, a comparative analysis of machine learning techniques for traffic accident severity prediction was carried out by Niyogisubizo, Murwanashyaka, and Nziyumva (2023) . In comparison to other techniques like Multinomial Naive Bayes (MNB), K-Means Clustering (KC), and K-Nearest Neighbors (KNN), their analysis demonstrated the superiority of the Random Forest (RF) method. These results confirmed the potential for machine learning techniques to be used effectively in the field of traffic safety.

Machine learning methods including XG Boost, random forests, CAT Boost, Light GBM, and logistic regression have advanced, and their use in predicting the seriousness of traffic injuries has yielded encouraging results. The goal of this research is to draw on historical information and incorporate contemporary approaches to enhance traffic safety measures and lessen the effects of accidents.

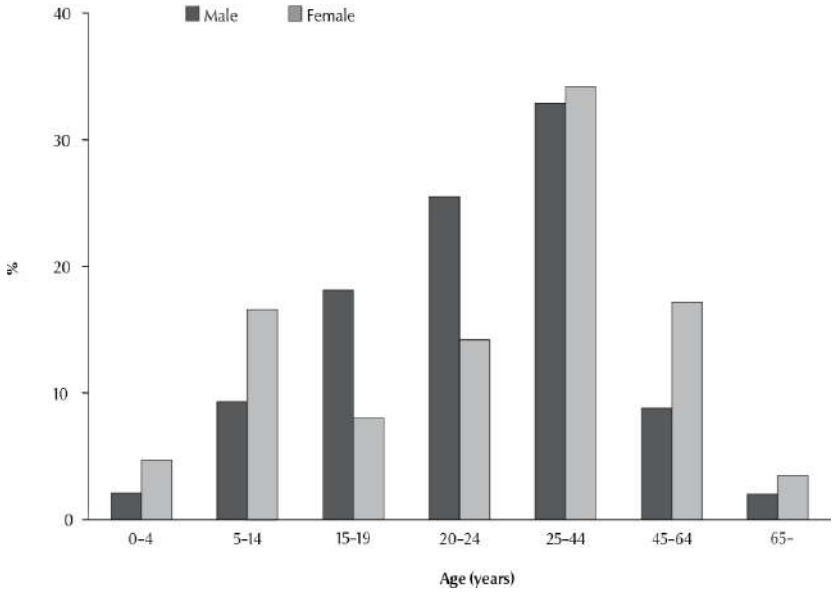


Figure 1. 2: road accident trends over 1time Rawalpindi

1.4 Aim and Objective

The Aims and objectives of this research study are discussed in detail down below.

1.4.1 Aim

As part of the research, we will create and assess machine learning models for prediction traffic injuries' severity accurately. The research aims to improve road safety measures, accident prevention strategies, and the overall reduction of traffic-related injuries by leveraging advanced machine learning techniques and analysing comprehensive datasets on road accidents.

1.4.2 Objective

This study will employ machine learning methods to create a predictive model to determine the severity of traffic injuries. We will study a large dataset encompassing parameters like road conditions, weather, driver behaviour, and vehicle features in order to determine the factors that affect how serious traffic accidents are. The study's particular goals are listed below:

- Assess the severity of traffic injuries by identifying the main factors.

- A comparison of the performance of various machine learning techniques in predicting traffic injury severity, including XG Boost, random forests, CAT Boost, Light GBM, and logistic regression.
- Analyse machine learning patterns and rules to develop policies and rules based on insights gained.

1.5 Research Gap

It is important to fill a research gap in traffic injury severity prediction utilizing machine learning methods. Research has examined the causes and consequences of traffic accidents, but there hasn't been much work done on creating and assessing machine learning models specifically for predicting the severity of traffic injuries. By successfully predicting the severity of traffic injuries, which fills a research need, various machine learning algorithms and feature selection approaches might offer useful insights for enhancing road safety measures [7]. This study presents a complete overview of the use of machine learning techniques to forecast the severity of traffic accidents. The effectiveness of various machine learning algorithms employed in this field is discussed and analyzed. These authors both emphasize and provide insight into potential future research directions.

The work of Guo et al. (2021), which focused on applying Extreme Gradient Boosting (XG Boost) to examine the severity of traffic crashes involving senior pedestrians in Colorado, US [8], is one of the important research articles addressing this gap. The study found that important factors impacting the severity levels of these incidents included driver characteristics, elderly pedestrian characteristics, and vehicle movement. The findings suggested that addressing and regulating these issues can assist safeguard senior pedestrians and enhance traffic safety. The XG Boost model gave useful insights into the parameters that correlate with each severity level, enabling the departments in charge of traffic management and infrastructure to take the necessary steps to ensure the safety of elderly pedestrians. The study emphasizes the value of using machine learning to identify factors that influence crash severity and adopt safety measures that are specifically designed to protect vulnerable road users.

Furthermore, Rezapour et al. (2019) carried out research to evaluate the severity of injuries in motorcycle at-fault crashes using machine learning methods [9]. The study's objectives were to determine the contributing variables and assess how well the models predicted the future. The study concentrated on mountainous routes with high collision rates and motorcycle usage. In order to examine injury severity based on chosen factors, such as alcohol use, road surface conditions, hitting an animal, and hitting a guardrail, binary logistic regression and classification tree (CT) models were both used. Although both models identified the same factors, the binary logistic regression fared marginally better in predicting injury severity. The study emphasizes the need of taking into account a variety of elements to address motorcycle crash severity and enhance motorcycle road safety.

These studies further underscore the need for developing reliable models to improve traffic safety measures by offering useful insights into forecasting the seriousness of traffic injuries using machine learning approaches.

1.6 Methodology

The methodology can be outlined as follows:

1.6.1 Data collection

Collect data about traffic accidents, including location, weather conditions, road type, vehicle characteristics, and severity of injuries. There should be sufficient accidents in this dataset for analysis, and it should cover a significant timeframe.

1.6.2 Data Preprocessing

Clean the dataset by dealing with missing values, removing irrelevant variables, and addressing any inconsistencies and errors. Understand the relationship between variables, identify outliers, and gain insight into the data distribution by conducting exploratory data analysis.

1.6.3 Feature Engineering

Extract relevant features from the dataset that can help in predicting injury severity. Injury severity can be affected by variables such as time of day, day of the week, and road conditions. This may involve creating new variables or changing existing ones.

1.6.4 Model Training

Taking the dataset and splitting it into training and testing sets will be useful for training the machine learning models and evaluating their performance.

For predicting traffic injury severity, choose appropriate machine learning algorithms. These include Random Forest, XG Boost, CAT Boost, Light GBM, and Logistic Regression. Using appropriate evaluation metrics, evaluate the performance of each algorithm to determine its strengths and weaknesses.

Using the training dataset, train the selected machine learning models and fine-tune their hyperparameters using techniques such as grid search or random search.

1.6.5 Model Evaluation

Use the testing dataset to assess the training models. Calculate metrics including accuracy, precision, recall, and F1-score to evaluate the effectiveness of their ability to predict injury severity. To learn more about the performance of the models, think about employing ROC curves and confusion matrices.

1.6.6 Model Comparison

Compare the performance of the different machine learning models to identify the most accurate and reliable model for predicting traffic injury severity. Consider factors such as accuracy, interpretability, and computational efficiency.

To determine which factors are most important in predicting injury severity, conduct an interpretability analysis. This analysis will provide insights into the relationship between different variables and their contributions to the prediction.

Prepare a detailed analysis of the results and discuss any limitations or challenges encountered during the research. Document the entire methodology, including data preprocessing steps, model selection, training, and evaluation steps.

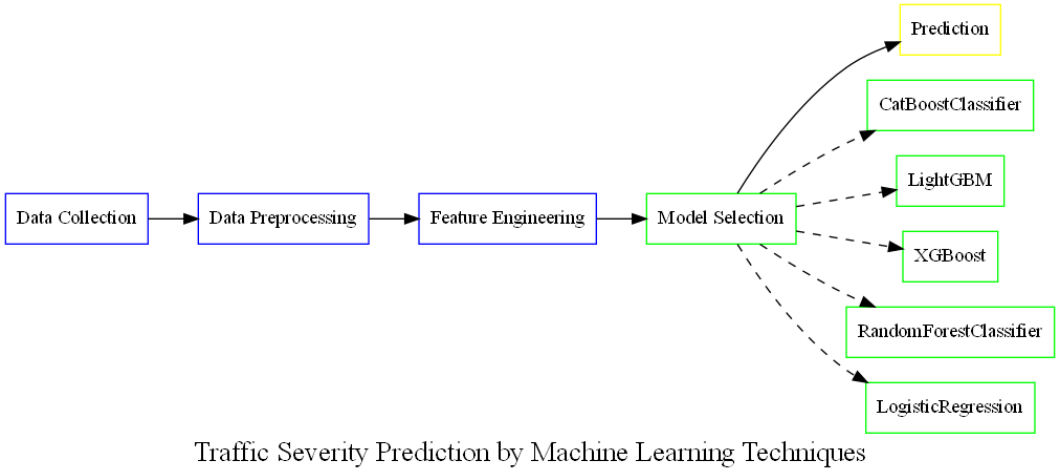


Figure 3. 1: Methodology flow chart

1.7 Results

We want to create a model that can categorise traffic incidents based on different input features by utilising the strength of machine learning techniques.
The following is a list of your project's anticipated results:

1.7.1 Accurate Prediction

Create machine learning models to forecast traffic injuries based on input characteristics like weather, road type, time of day, etc. The models should be able to classify accidents with varying degrees of severity, from minor injuries to fatalities, with high accuracy.

1.7.2 Performance Evaluation

Analyse the machine learning models' precision, recall, accuracy, and F1 score. We will be able to assess the models' accuracy and dependability in forecasting the seriousness of traffic injuries using these criteria.

1.7.3 Model Comparison

Compare the performance of different machine learning models, such as XG Boost, Light GBM, CAT Boost, Random Forest, and Logistic Regression. To determine the most suitable model for classifying traffic injuries based on their accuracy and performance, evaluate their accuracy and performance.

Determine the importance of input features when predicting traffic injury severity. As a result, we can develop targeted interventions and improve road safety measures by identifying the key factors that contribute significantly to the severity classification.

1.7.4 Real-World Application

Develop models that can be applied to real-world situations to make a positive impact. Develop effective traffic management strategies, improve emergency response systems, and enhance overall road safety measures by utilizing accurate predictions of traffic injury severity.

1.8 Utilization of Results

The outcome of this research study can be implicated and utilized in the following fields.

- A dynamic resource allocation system for intelligent traffic management based on the severity of injuries predicted is being developed.
- By providing accurate information about accident severity in real time, emergency response systems can be enhanced.
- By identifying the key factors contributing to the severity of road injuries, we can formulate and implement targeted road safety policies.
- Incorporating predictive models into existing traffic accident analysis systems for proactive risk assessment and decision-making.

1.9 Applications

Applications of the study are discussed below.

- Based on predicted injury severity levels, emergency services resources are allocated more efficiently.
- By implementing targeted interventions, such as improving road infrastructure, enforcing traffic regulations, and promoting awareness, road safety can be enhanced.

- To enable proactive measures and timely response, accident severity is monitored and analysed in real time.

1.10 Advantages

Major advantages of training Artificial intelligence-based machine learning models for the traffic severity are given below.

- Planning emergency responses and allocating resources based on accurate predictions of traffic injury severity.
- A better understanding of the factors contributing to injury severity will enable targeted interventions for improving road safety.
- For proactive decision-making and risk assessment, machine learning models can be integrated into existing systems.

1.11 Disadvantages

Alongside having huge advantages these techniques also have some disadvantages some of them are highlighted below.

- To make accurate predictions, it is necessary to have high-quality and readily available data.
- A limited understanding of the underlying factors influencing predictions is a challenge with machine learning models.

1.12 Learned Key Points:

- In order to improve model performance and accuracy, feature selection and data preprocessing must be considered.
- The most suitable approach for predicting traffic injury severity should be identified through a comparative analysis of different machine-learning models.
- The effectiveness and reliability of models is evaluated through the consideration of performance metrics.

1.13 Positive Affect:

- Enhanced road safety through proactive measures and targeted interventions.
- Enhanced emergency response and resource allocation will reduce traffic injuries and fatalities.
- An accurate prediction and identification of key factors provide the basis for informed decision-making and policy formulation.

1.14 Project Plan:

- Phase 1: Data Collection and Preprocessing (September 2022 to December 2022)
- The second phase is the development and training of models (January 2023 - March 2023)
- The third phase (April 2023 - May 2023) aims to evaluate and compare the models.
- The final phase is the integration and application of the technology (June 2023)

Table 1. 1: Project Work Schedule Plan

Phase	Start Date	End Date
Data Collection	Sept 2022	Dec 2022
Model Development	Jan 2023	Mar 2023
Model Evaluation	Apr 2023	May 2023
Integration	June 2023	June 2023

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of The Chapter

In this chapter the author has discussed past literature on the topic. The literature includes classic as well as modern machine learning techniques. This section also highlights the work of other researchers on the topic under discussion and their findings related to the ongoing research study.

2.2 Traffic Accident Density Prediction

As a result of combining the Random Forest and Generalized Additive Model (GAM), the authors present a novel method for predicting traffic accident density considering injury severity. Based on historical accident data and road network characteristics, the model was trained to predict accident density. The results indicated that the model could be used to enhance traffic safety by predicting accident density.[10]

2.3 Saudi Arabian Highway Crash Severity Prediction

In this study, the severity of highway crashes in Saudi Arabia is predicted using machine learning techniques. The scientists employed a range of machine learning models, including Gradient Boosting, Random Forests, and Decision Trees, to forecast accident severity. Based on a dataset that had more than 500,000 records of traffic accidents, Gradient Boosting fared better than the other models in terms of prediction accuracy.[11]

2.4 Machine Learning and Data Balance Strategies

This study uses machine learning and data balance approaches to forecast the severity of downhill truck crashes in Wyoming. The authors balanced the data by combining the Random Under-Sampling (RUS) and Synthetic Minority Over-Sampling Technique (SMOTE) techniques. The Gradient Boosting Classifier (GBC) model beat other models in terms of accuracy and other performance metrics as a result of the findings.

The study claims that GBC and other machine learning models can correctly forecast the severity of truck crash downgrades.[12]

2.5 Road Traffic Injury Severity

The research suggests four boosting-based ensemble learning models for predicting the severity of traffic injury using Shapley Additive Explanations (SHAP) to rank the risk variables and explain the best model. The Light Gradient Boosting Machine (Light GBM) provided the classification accuracy that was the highest.[13]

2.6 Machine Learning Classification Method

This research provides a hybrid feature selection-based machine learning classification strategy for identifying key qualities and estimating injury severity in single and multiple vehicle accidents. Extreme Gradient Boosting (XG Boost) outperforms other classifiers in terms of prediction performance.[14]

This research demonstrates that Machine Learning and AutoML may be applied to Crash Severity Prediction using bibliometric analysis and experimental benchmarks. This will choose models on its own. According to experimental findings, AutoGluon and CAT Boost are reliable and competitive machine learning methods.[15]

This research suggests an ordinal classification methodology to categorize traffic crash injury severity and compares its effectiveness to current machine-learning classification techniques. This method meets the criterion for rank consistency and rank monotonicity better than other ordinal classification methods and nominal classification machine learning methods..[16] In this study, three non-tree-based models (Support Vector Machines, Multilayer Perceptrons, and K-Nearest Neighbors) for predicting the severity of large truck crashes on Wyoming road networks were compared with four classification tree-based machine learning models (Adaptive Boosting tree, Random Forest, Gradient Boost Decision Tree, and Extreme Gradient Boosting tree). Then, a comparison of the precision of these seven approaches was made. The final ROC AUC for the improved random forest model was 95.26 percent. An AdaBoost model scored 67.232 percent, a Gradboost model scored 74.84 percent, an SVM model scored 72.648 percent, a k-NN model scored 92.780 percent, and an MLP model scored 87.817 percent. Based on the analysis, the top 10 predictors of

severity were determined using the feature importance plot. These categories may include safety equipment, airbag deployment, driver gender, and alcohol use.[17]

This study tested and compared the predictive abilities of four machine learning models (decision tree, naive Bayes, k-nearest neighbors, and random forest) based on the Mississippi classification of HELLP syndrome (homolysis, high liver enzymes, and low platelets). The best models for predicting HELLP syndrome were a decision tree model (accuracy: 91%) and a k-nearest neighbors' model (accuracy: 87.1%), whereas the best models for predicting class 2 and 3 HELLP syndrome were a random forest model (accuracy: 89.4%) and a naive Bayes model (accuracy: 86.9%). These models did poorly in class 2 and class 3 prediction, with accuracies varying from 65.2% to 83.8%, respectively.[18]

Here, machine learning methods are provided to forecast the likelihood of traffic-related fatalities when drunk drivers operate their vehicles. The authors advise that existing data be used to externally validate any future implementation.[19]

This research presents a hybrid feature selection-based machine learning classification method for identifying key variables and predicting injury severity in single- and multiple-vehicle accidents. Extreme Gradient Boosting (XG Boost) outperforms other classifiers in terms of prediction performance.[20]

We outline a procedure for estimating the seriousness of injury from a traffic accident using ordinal classification, and we contrast it with current machine-learning classification methods. The proposed strategy is found to satisfy the requirements for rank consistency and rank monotonicity when compared to previous ordinal classification methods and nominal classification machine learning approaches.[21]

This research employs Extreme Gradient Boosting (XG Boost) to simulate the categorization issue of pedestrian traffic crashes of three different severity categories from crash data gathered in Colorado, US. Shapley Additive Explanations (SHAP) are used to interpret the XG Boost model's findings and assess the significance of each feature in relation to pedestrian accident levels.[22]

In order to ascertain how shaping temperature impacts the shear properties of silty clay's freeze-thaw zone, low-temperature direct shear tests were carried out under various circumstances in this work. The findings show that the freeze-thaw zone's shear

strength diminishes as the thawing temperature and the shaping temperature rise. The strength of the freeze-thaw zone is strongly influenced by shaping temperature while the thawing temperature is constant. According to the study, cohesion is also discovered to be the primary element controlling shear strength, and the fluctuation in strength is strongly connected to the amount of unfrozen water present. The thawing temperature, water content, and shaping temperature all had a substantial impact on the sample's strength, according to grey relational analysis.[23]

This research analyzes the application of machine learning algorithms to forecast the severity of traffic accidents and presents an analysis of these techniques. Developing precise accident severity prediction algorithms can have a big impact on transportation systems all around the world. AdaBoost, Logistic Regression, Naive Bayes, and Random Forests were the supervised machine learning algorithms employed in this work. Data imbalance was addressed via the SMOTE algorithm. According to a study, the Random Forest (RF) model is 75.5% accurate at predicting the severity of traffic accidents. It was created to identify deciding variables and categorize the seriousness of injuries. It fared better than AdaBoost (74.5%), Naive Bayes (73.1%), and Logistic Regression (74.5%). The authors suggest the Random Forest model for tracking fatal injuries and serious injuries due to its higher performance. The predictive model can be used to pinpoint the major contributing elements to traffic accidents as highway engineers and transportation designers develop safer roadways. The study has a number of drawbacks, claims the article. Certain potentially important aspects, such as the characteristics of drivers, passengers, and pedestrians, as well as traffic circumstances, could not be taken into consideration due to a lack of adequate data. Future research should be done to acquire information on the effects of these parameters on accident severity and duration.[24]

In this study, approaches for recognizing severe chest injuries in electronic health records (EHRs) for quality reporting are proposed, using machine learning (ML) and natural language processing (NLP). The scientists employed logistic regression with elastic net regularization, extreme gradient-boosted machines (XGB), and convolutional neural networks (CNN) to categorize severe chest injuries. Since CNN models were the most accurate and clinically pertinent models, there is potential to

employ ML to populate clinical registries for research and quality analysis.[25] 6811 individuals with normal cardiac troponin (CTn) levels underwent noncardiac surgery between January 2010 and June 2019, and a prediction model for myocardial injury following noncardiac surgery was created and is available online. Gradient-boosting algorithms were employed in machine learning approaches to assess the effects of variables on the development of MINS. To identify MINS in 1499 (22.0%) patients, two prediction models based on the top 12 and 6 characteristics were utilized. MINS is influenced by cTn levels, intraoperative inotropic drug infusion, operation length, emergency operations, operation types, age, high-risk surgeries, body mass index, chronic kidney disease, coronary artery disease, intraoperative red blood cell transfusion, and current alcohol use, among other factors. In 12-variable models, a threshold of 0.47 was discovered; in 6-variable models, a threshold of 0.53 was discovered. The model has an accuracy of 0.81 based on the area under the curve, demonstrating that it is sensitive and specific enough to predict MINS.[26]

This research presents novel methods for the prediction of drivers' injuries in intersection incidents. The work combines a cost matrix (taken from the KABCO injury categorization scale developed by the Federal Highway Administration) with machine learning techniques (C4.5, instance-based (IB), and random forest (RF) to generate these models. Drivers were split into four categories depending on age and gender (younger males, younger females, older males, older females), and each group had its own model, according to the researchers. Based on a mix of data related to the driver, vehicle, road/traffic, environment, and crash, the degree of driver injuries is anticipated. Over a five-year period, data on two-vehicle crashes in Miami, Florida, was gathered for the models. It was discovered that cost-sensitive learning classifiers outperformed conventional classifiers at predicting injuries and mortality. For forecasting the degree of driver injury for the four driver categories, RF fared better than the C4.5 and IB models. Injuries severity determinants varied significantly between groups. According to the study, current injury severity prediction models are more accurate and less biased, which improves intersection safety.[27]

Public health and safety are significantly impacted by traffic accidents on highways. To forecast accident severity, the study analyzes key factors linked to crash severity.

Distance, temperature, wind chill, humidity, visibility, and wind direction are found to be the main elements that affect accident severity using Random Forest. In order to increase decision-making and prediction accuracy, an ensemble model integrating Random Forests and Convolutional Neural Networks is suggested in the research. The performance of RFCNN is compared with a number of base learner classifiers using accident statistics from February 2016 to June 2020. RFCNN outperforms other models in the experiment with high accuracy (0.991), precision (0.974), recall (0.986), and F-score (0.980) values. It comes to the conclusion that traffic accidents have a big impact on public health and safety. The best characteristics discovered by Random Forest are then fed into ensemble models to boost performance even more. The RFCNN model outperforms conventional models in terms of its ability to forecast accident severity since it combines machine learning with deep learning. Identified characteristics, including the space between vehicles, are crucial for road authorities to take preventive action. The complexity of the ensemble model is acknowledged, and future research is suggested to address it. In order to assess the effectiveness of the proposed model, it will also be used with a variety of datasets.[28]

In industrialized nations like Pakistan, the motorized rickshaw is common but can be dangerous if it crashes. Due to preconceived notions and associations, motorcyclists, pedestrians, and cyclists are the three groups most likely to be involved in road traffic incidents in developing nations. Traditional statistical models may give false results as a result of predefined assumptions and relationships. Machine learning models are a compelling option because nonlinear effects of continuous and discrete variables can be successfully recorded without the use of preconceived notions. Researchers employed machine learning algorithms including Decision Jungle, Random Forest, and Decision Tree to detect and forecast injury severity in Rawalpindi city crash data from 2017 to 2019. In order to assess models for overall accuracy, macro-average precision, macro-average recall, and geometric means of class accuracy, 258 motorized rickshaw crashes involving three wheels were used. With an accuracy of 83.7%, the DJ model performed better than the DT and RF models. The study found that characteristics including poor lighting, youthful drivers, high speed restrictions (over 60 mph), weekdays, off-peak hours, and clear weather increase the likelihood of serious three-

wheeled motorized rickshaw crashes. As a result, crucial advice on how to adopt effective countermeasures to reduce the road safety issues caused by three-wheeled motorized vehicles can be given to road safety agencies, particularly in developing nations. The authors advise that in order to lower the probability of fatal and catastrophic injuries brought on by this mode of transportation, future research might investigate cutting-edge approaches like ensemble learning and deep learning on more precise datasets.[29]

A two-layer ensemble machine learning model can be used to forecast the severity of a road traffic crash, which will help emergency services better foresee incidents. With training and testing accuracy of 81.6% and 76.7%, respectively, the model performed well. In its first layer, it uses four machine learning models, and in its second, it uses a feedforward neural network. The model was developed using data on traffic accidents collected over a six-year period by the Department of Transport in Great Britain (2011–2016), and it was tested using data on collisions in Canada, where it also performed well. The outcomes demonstrate the model's potential to facilitate prompt and appropriate medical aid based on preliminary crash data.[30]

This study developed and compared four machine learning models: feed-forward neural networks (FNN), support vector machine (SVM), fuzzy C-means clustering based feed-forward neural network (FNN-FCM), and fuzzy c-means based support vector machine (SVM-FCM), in order to predict crash injury severity using 15 crash-related parameters. Using crash data from Great Britain from 2011 to 2016, the models were assessed based on injury severity prediction accuracy, sensitivity, precision, and F1 score. The outcomes demonstrated that, in terms of accuracy and F1 score, the SVM-FCM model outperformed the competition in predicting the severity level of severe and non-severe crashes. The study found that using the fuzzy C-means (FCM) clustering algorithm increased the prediction power of the FNN and SVM models.[31]

The study employed machine learning techniques, namely random forest, artificial neural network, and decision tree, to forecast the severity of traffic accidents during wet seasons. The three metrics used to assess the models were out-of-bag estimate of error rate (OOB), mean square error (MSE), and root mean square error (RMSE). The information came from three separate datasets that were all connected to Seoul, South

Korea's Naebu Expressway over a nine-year period. These files included information on precipitation, road geometry, and traffic accident statistics. The random forest model, which also had the lowest mean OOB, MSE, and RMSE, made the most accurate prediction. The study concludes that, particularly in wet weather, the random forest algorithm is a valuable tool for analyzing and anticipating accident severity.[32]

The study's objective was to assess how effectively various machine learning and statistical approaches might predict the seriousness of a collision injury. Through the use of ordered probit (OP), the multinomial logit model, and machine learning techniques K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine, data on crash severity, road geometry, and traffic flow were collected from Florida highway diverge zones. The results revealed that machine learning techniques were, on average, better at predicting outcomes than statistical techniques, with the Random Forest technique attaining the highest overall predictive accuracy of 53.9%. However, they also brought up the issues with overfitting in machine learning algorithms. The study underlines the relevance of using the right prediction model for collision injury severity analyses given the potential traffic safety implications.[33]

The researchers employed machine learning techniques to classify UK road traffic accident data from 2016 with the aim of finding the primary contributing reasons to such events. Fuzzy-FARCHD, one of six machine learning classification algorithms used, generated the best outcomes with an accuracy rate of 84.94%. Significant contributing factors included the type and quantity of first roads, the lighting situation, and the number of vehicles. The study suggests using deep learning techniques for future research due to the expanding size of datasets.[34]

The study focuses on using deep learning techniques to forecast how serious injuries from traffic accidents would be on Malaysian roadways. The Neural Network (NN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) network architectures were explored. In this group, the RNN model fared better than the others, with an average accuracy of 73.76%, surpassing NN's (68.79%) and CNN's (70.30%) levels. The "Nadam" algorithm was determined to be the most effective for optimization across all three network architectures. Furthermore, it demonstrated that adding temporal and spatial variables to traffic accident data might improve prediction

accuracy, with the RNN model's superior performance indicating a stronger temporal component in the accident data.[35]

Using information and statistics provided by governmental agencies in Spain, this study tries to categorize road incidents according to their nature and severity, covering the years 2011 to 2015. Gradient Boosting Trees, Deep Learning, and Nave Bayes are three different machine learning classification approaches that are applied in this study with the main objective of accelerating post-accident procedures and assisting in the creation of general road safety laws. Gradient Boosting Trees utilized twenty trees, while Naive Bayes required no parameterization. The Deep Learning model was parameterized with two hidden layers and ten epochs, and it made use of different activation functions such the Hyperbolic Tangent Function, Rectifier Linear, and Exponential Rectifier Linear. The study is an important step in assessing the seriousness of injuries in auto accidents in real time.[36]

The important problem of road accidents in Bangladesh is addressed in "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh" by Md. Farhan Labib et al. They use Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, and AdaBoost as four machine learning algorithms to evaluate incidents and determine their severity. The study found that AdaBoost, with an F1 score of 80% and an accuracy of 75%, is the most effective technique for binary classification. The researchers propose creating a mobile application for real-time accident prediction as well as a predictive recommender system.[37]

The goal of this project is to apply machine learning algorithms to predict the severity of crash injuries in motorcycle accidents. The study makes use of data that was taken from Ghana's National Road Traffic Crash Database and divides it into four categories of injury severity: dead, hospitalized, injured, and damage. Three classification algorithms—the multi-layer perceptron (MLP), the rule induction (PART), and the classification and regression trees (SimpleCart)—are compared for performance in the study. The results demonstrate that the SimpleCart model outperforms the other two models, with the highest average accuracy of 73.81% based on a 10-fold cross-validation approach. The study also highlights key variables that affect the severity of

motorcycle crash injuries, including crash location, settlement type, crash timing, collision type, and collision partner.[38]

With a focus on predicting the severity of traffic accidents in Adana, Turkey, based on injury severity (fatal or non-fatal), this study investigates the factors influencing accident outcomes. The study used meteorological information from the Regional Directorate of Meteorology from 2005 to 2015 and reports on road accidents from the regional road Division. Six machine learning approaches (k-Nearest Neighbor, Naive Bayes, Multilayer Perceptron, Decision Tree, and Support Vector Machine) and one statistical technique (Logistic Regression) were utilized to develop prediction models and evaluate their performance. Decision Tree, k-Nearest Neighbor, and Multilayer Perceptron models distinguished accidents more precisely than other models. The aim of the study is to understand the significance of weather and other phenomena in the occurrence of traffic accidents, with specific parameters having higher positive effects on the accuracy of accident prediction such as mean cloudiness, the presence of traffic control, and ground surface temperature.[39]

2.7 Literature Gap

Table 2. 1: Literature Gap

Title	Authors	Year	Data Type	Machine Learning Technique	Ensemble Technique
Crash Severity Prediction Using Two-Layer Ensemble Machine Learning Model	Umer Mansoor, et al.	2020	Traffic Accidents	Multiple, Feedforward Neural Network	Yes
Using Machine Learning, Predicting Crash Injury Severity	Khaled Assi, et al.	2020	Crash-related parameters	FNN, SVM, FNN-FCM, SVM-FCM	No
Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons	Jonghak Lee, et al.	2019	Road geometry, Precipitation, Traffic Accidents	Random Forest, ANN, Decision Tree	No

Comparing Prediction Performance for Crash Injury Severity	Jian Zhang, et al.	2018	Crash Severity, Road geometry, Traffic flow	K-Nearest Neighbor, Decision Tree, Random Forest, SVM	No
Classification of Road Traffic Accident Data Using Machine Learning Algorithms	Bulbula Kumeda, et al.	Unknown	Traffic Accident	Six ML algorithms	No
Applications of Deep Learning in Severity Prediction of Traffic Accidents	Unknown	2019	Traffic Accident	NN, RNN, CNN	No
Traffic accidents classification and injury severity prediction	Laura García Cuenca, et al.	Unknown	Traffic Accidents	Gradient Boosting Trees, Deep Learning, Naïve Bayes	No
Road Accident Analysis and	Md. Farhan	Unknown	Traffic Accidents	Decision Tree, KNN, Naive	No

Machine Learning-Based Prediction of Accident Severity in Bangladesh	Labib, et al.			Bayes, AdaBoost	
Severity prediction of motorcycle crashes with machine learning methods	Lukuman Wahab and Haobin Jiang	Unknown	Motorcycle crashes	MLP, PART, SimpleCart	No
Predicting the Severity of Motor Vehicle Accident Injuries in Adana, Turkey	Çiğdem ACI and Cevher ÖZDEN	Unknown	Weather, Traffic Accidents	KNN, Naive Bayes, MLP, Decision Tree, SVM, Logistic Regression	No

2.8 Literature Conclusive Paragraph

The literature discussed here is primarily concerned with the used of machine learning and other statistical techniques to forecast and assess the severity of traffic accidents around the world. They draw attention to the development and growing interest in using machine learning to control traffic safety. Machine learning methods, including random forests, support vector machines, neural networks, K-Nearest Neighbors, decision trees, AdaBoost, and others, are used in the majority of studies. The studies sought to discover factors that affected accident outcomes as well as forecast accident severity.

These variables include the state of the roads and the climate, as well as other accident-related characteristics.

There is no one performance model that fits all circumstances. The best models across experiments varied, however models using ensemble approaches or fuzzy c-means clustering appear to significantly outperform previous models. This emphasizes the value of model selection and the demand for more study on model optimization. The potential of these machine learning models for real-time applications, such as in emergency management or in mobile apps for accident prediction, is also highlighted by several of the studies. Such programs would be essential for advancing preventative traffic safety measures. There are still difficulties, despite the great advancements made in this area. Among these are the overfitting issues with machine learning models and the demand for more varied and substantial datasets to train these algorithms.

In summary, the body of work under study emphasizes how machine learning can improve traffic safety management. In order to turn these prediction models into life-saving tools, there is a constant need for research and development in this field that focuses on enhancing model performance, including more diverse datasets, and investigating real-world applications.

CHAPTER 3

DATA SET DESCRIPTION

3.1 Overview of The Chapter

The Rawalpindi Traffic Police Department in Pakistan provided the dataset that we used for our investigation. The road traffic accidents that occurred in Rawalpindi between 2017 and 2019 are detailed in this dataset. It contains 837 distinct entries (or "observations") and 26 different categories of data (or "variables") regarding each accident, making it a particularly data-rich document. 25 out of 26 of these characteristics are information that aids in our comprehension of each accident. The final variable, the goal variable, categorizes accident-related injuries according to whether they are "Fatal" or "Non-Fatal."

3.2 Data Features and Variables

What kind of information are offered by these features, then? They include data such as the date and location of each accident, the weather condition, the cause, the kind, and the severity of the event. Additionally, they offer some demographic and medical information about the participants. We will find this kind of information to be extremely helpful as we attempt to identify trends and correlations.

A list of the characteristics and variables present in the dataset is shown in the table below.

Table 3. 1: Features and variables of a dataset

Feature	Description	Data Type	No. of Categories	Range
Minute_of_hour	The minute of the hour when the accident	Numeric	-	0 to 59

	occurred, ranging from 0 to 59.			
Period_of_Day	The period of the day when the accident occurred, either Peak or OFF Peak.	Nominal	2	
Lighting conditions	The lighting conditions when the accident occurred, either Day or Night.	Nominal	2	
Day_of_week	The day of the week when the accident occurred, ranging from 1 (Sunday) to 7 (Saturday).	Numeric	7	1 to 7
Nature_of_Weekday	The nature of the weekday	Nominal	2	

	when the accident occurred, either Weekend or Weekday.			
Day_of_month	The day of the month when the accident occurred, ranging from 1 to 31.	Numeric	31	1 to 31
Month_of_year	The month of the year when the accident occurred, ranging from 1 (January) to 12 (December).	Numeric	12	1 to 12
Season_of_year	The season of the year when the accident occurred, either Winter, Spring,	Nominal	4	

	Summer, or Autumn.			
Year	The year when the accident occurred, either 2017, 2018, or 2019.	Numeric	3	
Patients_in_Emergency	The number of patients who were taken to the emergency department due to the accident, ranging from 1 to 4.	Numeric	-	1 to 4
Gender	The gender of the patient involved in the accident, either Male or Female.	Nominal	2	

Age	The age of the patient involved in the accident, ranging from 6 to 82.	Numeric	-	6 to 82
Injury Type	The type of injury that the patient suffered due to the accident, either Head Injury, Single Fracture, or Minor/ F/Aid.	Nominal	3	
Reason	The reason for the accident, either Van Hit Pedestrian or Car Hit Pedestrian.	Nominal	2	
Crash Type	The type of crash that occurred in	Nominal	2	

	the accident, either Hit pedestrian or Rear-end collision.			
Injury level	The severity of the injury that the patient suffered due to the accident, either Fatal Injury or Non Fatal.	Nominal	2	
Weather	The weather condition when the accident occurred, either Shiny, Cloudy, or Rainy.	Nominal	3	
No_of_vehicles	The number of vehicles involved in the accident.	Numeric	-	

Cause	The cause of the accident, either Distractions or Over Speeding.	Nominal	2	
Road Name	The name of the road where the accident occurred.	Nominal	-	
Road Type	The type of road where the accident occurred - Major arterial, Minor arterial, Collector Road, or Local Road.	Nominal	4	
No Of Lanes	The number of lanes on the road where the	Numeric	5	2 to 4

	accident occurred.			
Posted Speed Limit	The posted speed limit on the road where the accident occurred.	Numeric	-	50 to 80

3.3 Data cleaning and validation

We cleaned up the dataset before beginning the analysis. In order to check the data for any missing or inconsistent entries that would have tipped off our conclusions, we utilized Python scripts and Jupiter Notebooks to sort through the information.

3.4 Data Visualization

We will use data visualization tools to better comprehend the data and its patterns. In order to provide insights into trends and distributions, bar charts will be utilized to illustrate the frequency of accidents based on different variables.

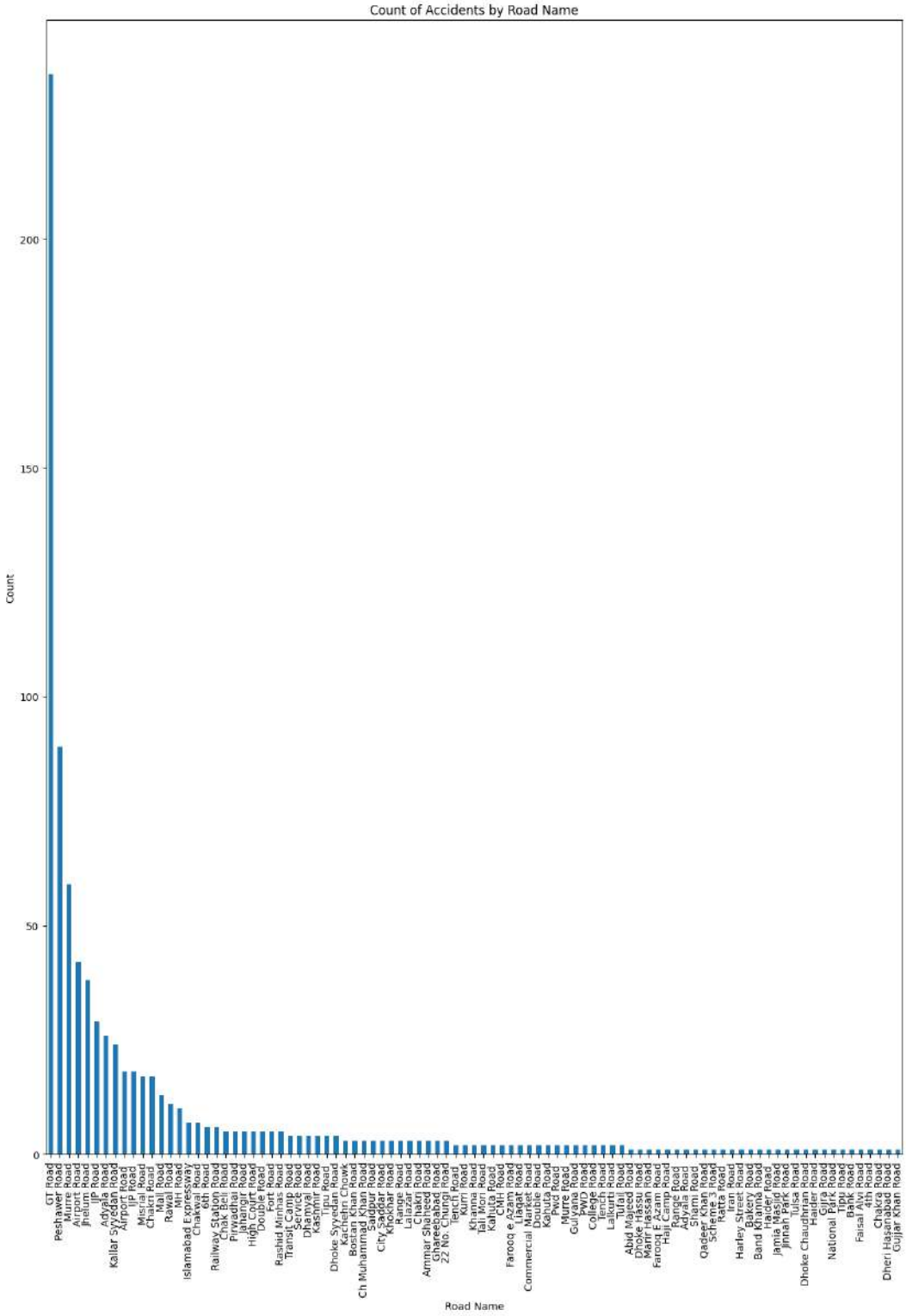


Figure 3. 2: Count of Accident by Road Name

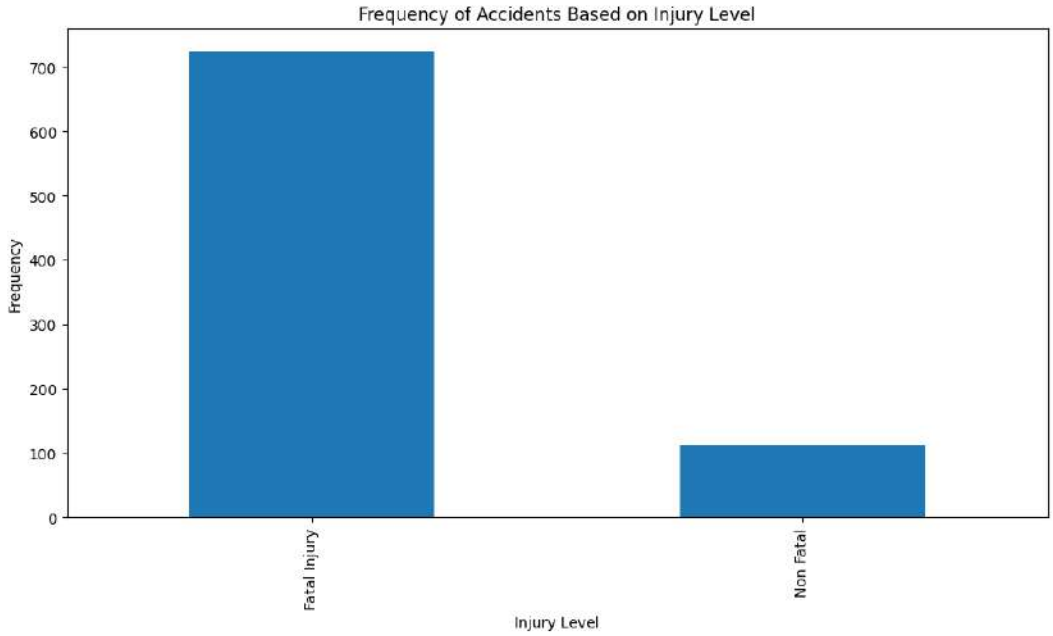


Figure 3. 3: Frequency of Accidents Based on Injury level.

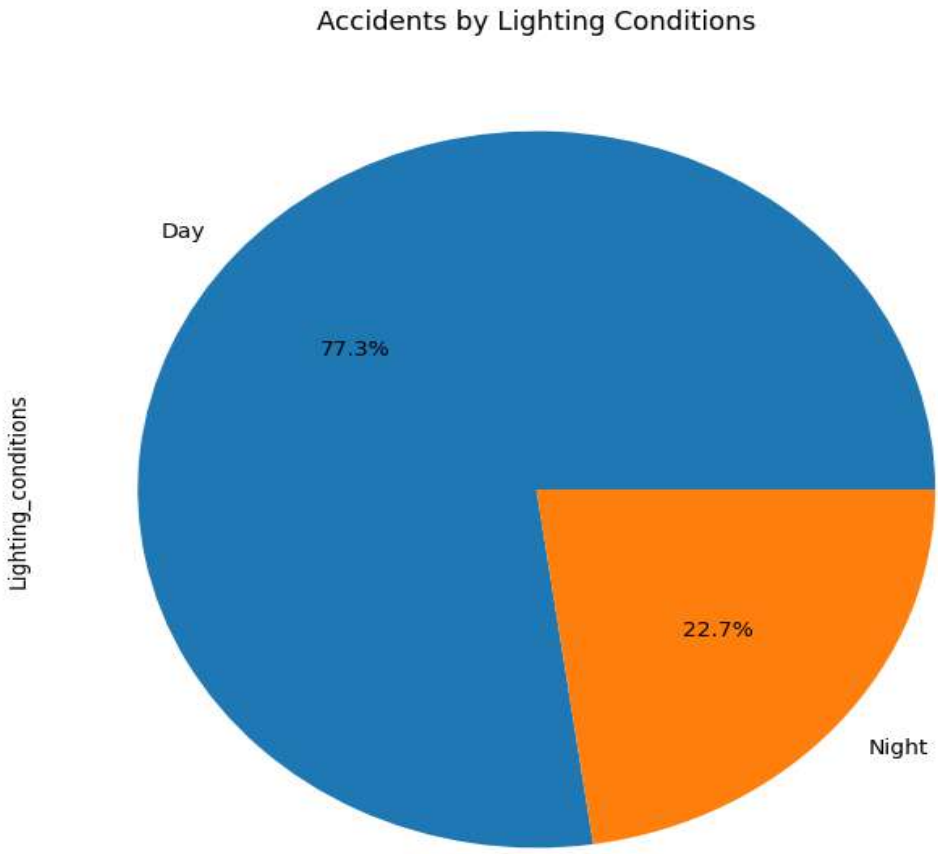


Figure 3. 4: Accidents by lighting Conditions

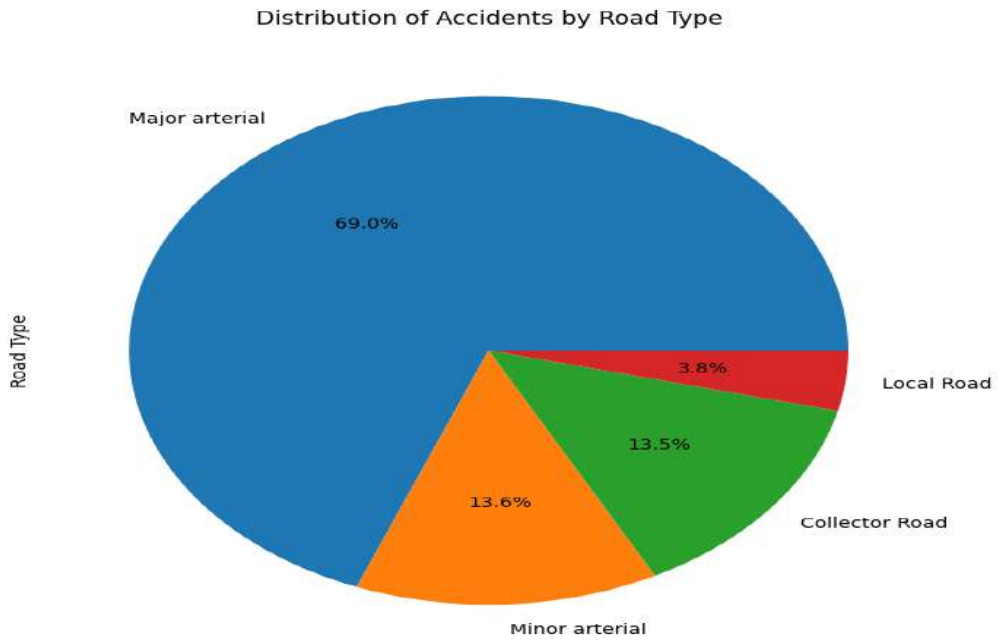


Figure 3. 5: Distribution of Accidents by Road Type

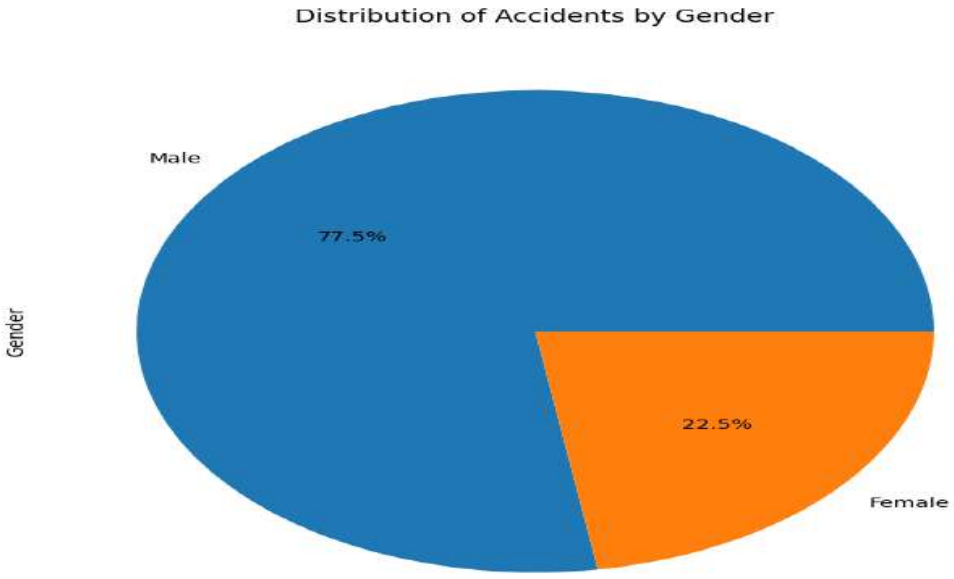


Figure 3. 6: Distribution of Accidents by gender

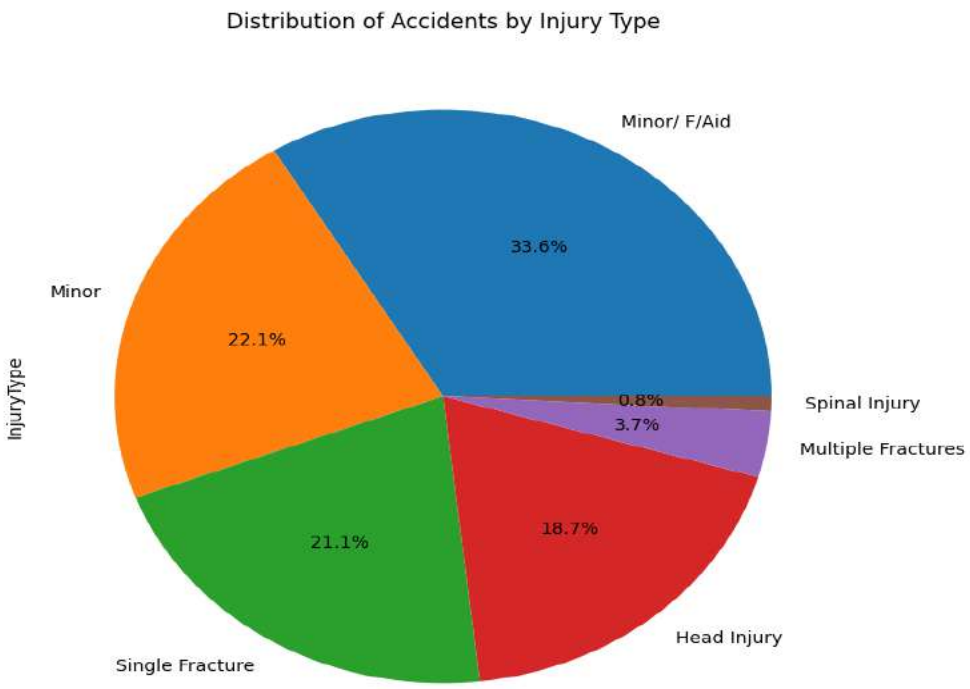


Figure 3. 7: Distribution of Accidents by Injury Type

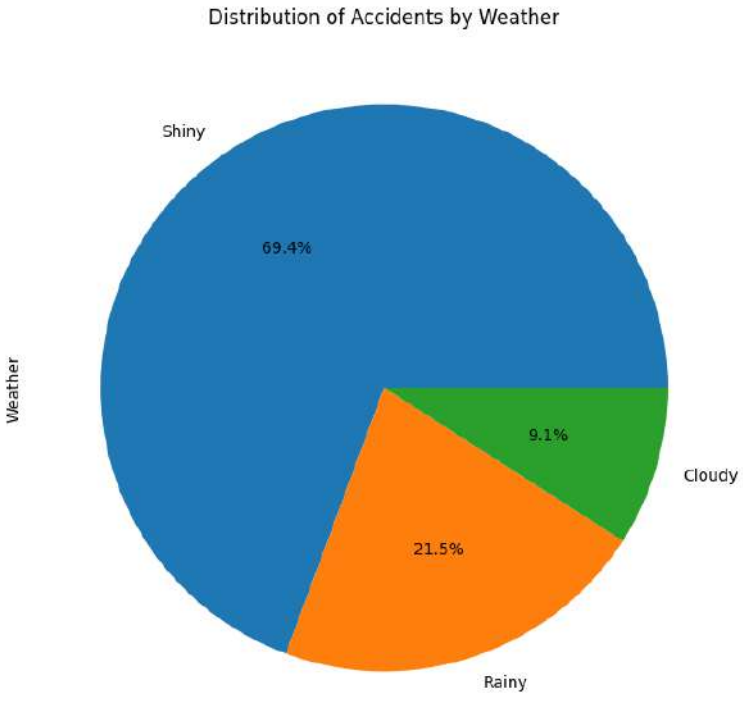


Figure 3. 8: Distribution of Accidents by Weather

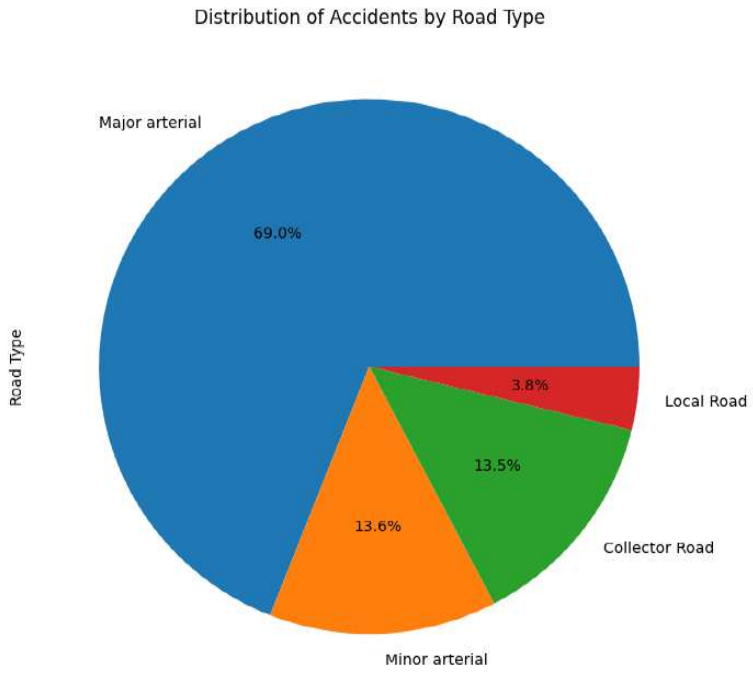


Figure 3. 9: Distribution of Accidents by Road type.

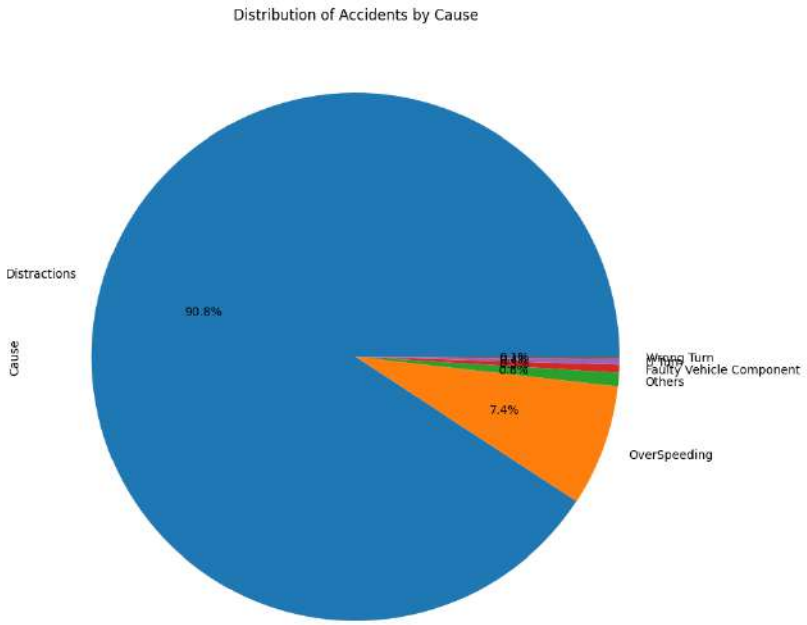


Figure 3. 10: Distribution of Accidents by Cause

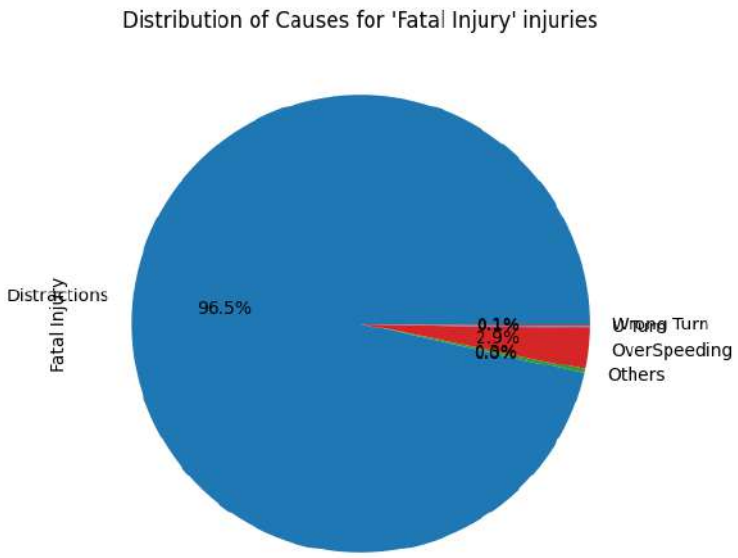


Figure 3. 11: Distribution of Causes for Fatal Injury

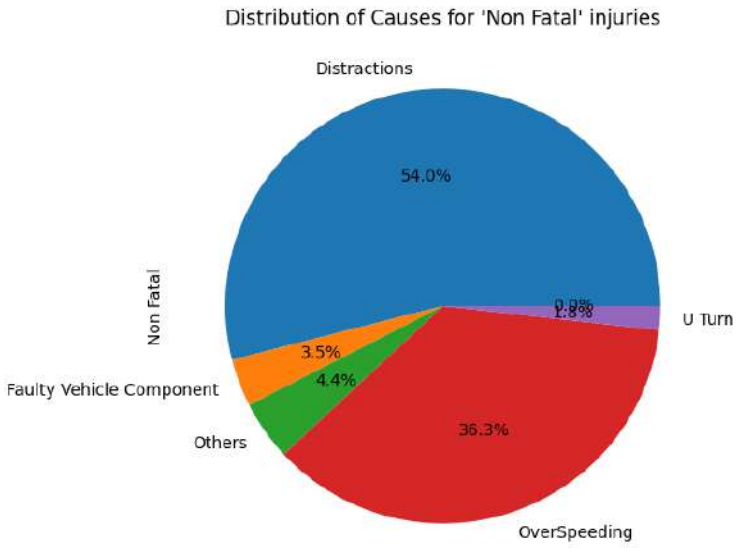


Figure 3. 12: Distribution of Causes for Non-Fatal Injuries

3.5 Limitations

Despite the abundance of data our dataset offers, it's important to note that it does have some restrictions. For starters, only accidents that were reported to the police are included. Additionally, there are only two classifications for injury severity: "Fatal" and "Non-Fatal," which may not include all forms of injuries. Finally, since it only includes

data from 2017 to 2019, it might not accurately reflect more recent trends. We'll delve more deeply into the dataset in the subsequent chapters to explore what it may reveal about road traffic accidents in Rawalpindi. In the end, we hope to be able to use the data to predict the severity of injuries from traffic accidents.

CHAPTER 4

MODEL DEVELOPMENT

4.1 Overview of The Chapter

In this chapter, we present the models we created for our project on predicting the seriousness of traffic injuries using several machine learning techniques. CAT Boost, Light Gradient Boosting Machine (Light GBM), XG Boost, Logistic Regression, and Random Forest will all be used to create our models. Each model is fitted to the training data, predictions are made using the training and test sets of data, accuracy, precision, recall, and F1 scores are calculated for each set, the confusion matrix is drawn for the test set, and the prediction results are then saved.

4.2 Importing Libraries and Loading Dataset

This first step is where we lay the groundwork for our analysis. We import every Python library we'll need to help us with this project. This includes sklearn for machine learning applications, pandas and NumPy for data manipulation, matplotlib and seaborn for data visualization, and imblearn for handling uneven datasets.

Table 4. 1: Purpose of Each Imported Library

Library	Purpose
pandas	Data analysis and manipulation library. provides tools for data manipulation and analysis as well as data structures for effectively storing massive datasets.
NumPy	Large, multidimensional arrays and matrices are supported by a library for the Python programming language, along with a substantial number of high-level mathematical operations that may be performed on these arrays.
matplotlib	Python library for producing interactive, animated, and static visualizations.

Seaborn	A matplotlib-based library that offers a high-level interface for creating appealing and instructive statistics visualizations.
sklearn	a library with a variety of tools for data mining and data analysis, including efficient tools for clustering, classification, regression, and data analysis.
imblearn	provides resources for coping with class classification that is unbalanced.

As soon as our dataset has loaded, we proceed to familiarize ourselves with its features. We employ the df.info() function to obtain a comprehensive overview of the dataset, which includes details on the names of the columns, the number of non-null entries in each column, and the data type of each column.

```
In [151]: #data ko visualize ker ke dekhte hai aur head ko call kr k print krte hai start 5 line
          df.head()
```

Out[151]:

	AM/PM	Hour_of_the_day	Minute_of_hour	Period_of_Day	Lighting_conditions	Day_of_week	Nature_of_Weeday	Day_of_month	Month_of_year	Season_of_ye
0	AM	6	53	OFF Peak	Day	1	Weekend	8	1	Win
1	AM	9	14	Peak	Day	1	Weekend	8	1	Win
2	PM	15	9	OFF Peak	Day	6	Weekday	13	1	Win
3	AM	9	17	Peak	Day	1	Weekend	15	1	Win
4	PM	12	56	OFF Peak	Day	2	Weekday	16	1	Win

5 rows × 26 columns

Figure 4. 1: output from df. Head ()

4.3 Dataset Overview and Preprocessing

Once our dataset is loaded, we move on to become acquainted with its features. For a thorough overview of the dataset, including information on the names of the columns, the amount of non-null entries in each column, and the data type of each column, we use the df.info() function.

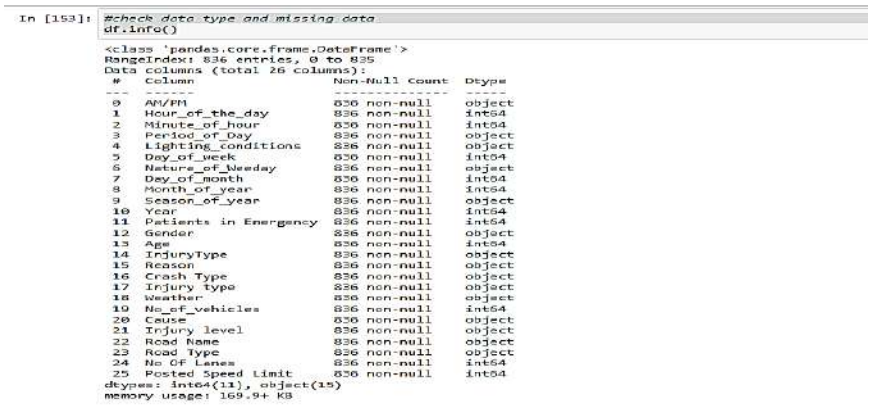


Figure 4. 2: df info Output

Data preparation always includes a check for missing data. The performance of machine learning algorithms might be hampered by any null or missing value. Therefore, we check that there are none in our dataset using df.isnull().

Next, label encoding is applied to the target variable "Injury level." Due to the fact that machine learning algorithms typically work better with numerical inputs, this phase is essential. Using the 'Injury level' column's distribution of the various classes after encoding value counts().

Table 4. 2: Instances of Each class in 'Injury Level' (Before Balancing)

Injury Level	Number of Instances
0	723
1	113

Table 4. 3 : Instances of Each Class in ‘ Injury Level’ (After Balancing)

Injury Level	Number of Instances
0	723
1	723

Then, we separate our dataset into predictors (X) and the target (Y). What we're attempting to forecast, or estimate is the target, or dependent variable. On the other hand, predictors, also known as independent variables, are the characteristics that the model will rely on to generate predictions.

Table 4. 4: First Few Rows of Y (Injury level)

Index	Injury Level
0	Fatal (0)
1	Fatal(0)
2	Fatal (0)
3	Fatal (0)
4	Fatal (0)
1441	Non-Fatal (1)
1442	Non- Fatal (1)
1443	Non-Fatal (1)
1444	Non- Fatal (1)
1445	Non-Fatal (1)

Table 4. 5: Sample Rows of X

Index	AM/PM	Hour	Minute	Period	Light Cond.	Day of Week	Weekday	...	Speed Limit
0	AM	6	53	Off	Day	1	Weekend	...	60
1	AM	9	14	Peak	Day	1	Weekend	...	60
...	...	...	...	...	...	...	...	...	...
4	PM	12	56	Off	Day	2	Weekday	...	70

4.4 Categorical Variables Encoded

It is necessary to transform categorical variables into a format that is more compatible with machine learning techniques. To transform category variables into numerical format, we utilize ordinal encoding. Our dataset has been perfectly prepared for the use of machine learning techniques after encoding.

Figure 4. 3: the first few rows of Y after encoding.

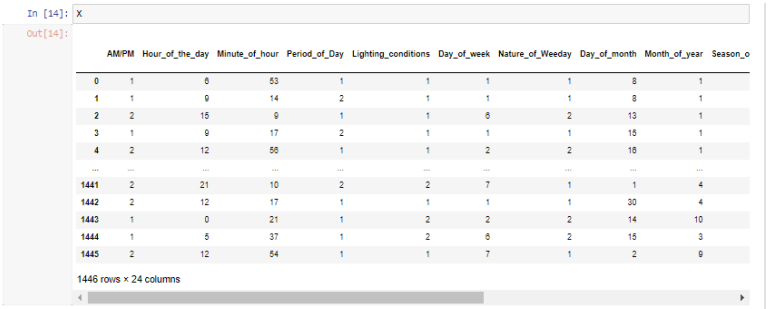


Figure 4. 4: the first few rows of X after encoding.

4.5 Achieving Dataset Balancing

Our target variable's class distribution is unbalanced. As a result, the model may be biased because it will be impacted more by the class in power. We use the Synthetic Minority Over-sampling Technique (SMOTE), which creates fresh cases from the minority class to achieve an equal balance, to balance the dataset in order to lessen this.

Table 4. 6: Balanced Class Distribution in Y (Injury Level)

Injury Level	Number of Instances
Fatal (0)	723
Non-Fatal (1)	723

The balanced distribution of classes in our 'Injury level' variable is shown in the table below. Both "Fatal" and "Non-Fatal" injury levels have an identical number of instances in the dataset following the SMOTE oversampling approach. This will make it easier to build a predictive model that is balanced.

4.6 Splitting the Dataset

The final stage in data preparation is to split the dataset into training and testing sets. The training set is used to develop the model, and the testing set is used to evaluate the model's performance. There is a 70:30 split. In line with this, 30% of the data will be used for model testing and 70% for model training.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.30, random_state = 21, stratify = Y)
```

Figure 4. 5: Splitting the Dataset

4.7 Mathematical Equations

The performance metric formulas are as follows:

4.7.1 Precision

Precision measures the proportion of observations that were accurately forecasted as positive to all anticipated positives. It also goes by the moniker Positive Predictive Value. It evaluates a classifier's accuracy. An excessive number of false positives is a sign of poor precision.

Precision is equal to TP/(TP + FP).

4.7.2 Recall

Recall (or sensitivity) is the ratio of correctly predicted positive observations to all of the actual class's observations. Other names for it include sensitivity, hit rate, and true positive rate. It measures a classifier's rigor. Low recall indicates a high number of false negatives.

Recall is TP / (TP + FN).

Where:

True Positives = TP

False Positives = FP

False Negatives (FN)

4.7.3 Confusion Matrix

A supervised machine learning model's performance is displayed in a confusion matrix, which is a tabular layout. Each row of the matrix represents an example from a predicted class, while each row of the matrix represents an occurrence from a real class, or vice versa.

Table 4. 7: Confusion Matrix

Confusion Matrix	Actual: Yes	Actual: No
Predicted: Yes	True Positive (TP)	False Positive (FP)
Predicted: No	False Negative (FN)	True Negative (TN)

4.8 Terminologies Used:

There are several terminologies used to carry out the research. The definitions and symbols of these terminologies are given below.

4.8.1 True Positives (TP):

These positive values are those that were correctly anticipated and show that both the actual and projected classes had True positive values.

4.8.2 Negatives (TN):

These negative values were accurately predicted and show that both the predicted class and the actual class have a value of zero.

4.8.3 False Positives (FP):

are instances where the anticipated class is present, but the actual class is not. Additionally, "Type I error" is used.

4.8.4 False negatives (FN):

occur when a class is expected to be no when a class is yes. Sometimes, the phrase "Type II error" is employed.

4.9 CAT Boost Model

The machine learning approach known as CAT Boost, or "Category Boosting," uses gradient boosting on decision trees. It is well renowned for its excellent performance and speed and is particularly effective for datasets with categorical features. The following stages might be used to summarize how the CAT Boost model was created and used for our project:

4.9.1 Construction And Assembly of The CAT Boost Model

We begin by instantiating the CAT Boost Classifier and importing the required components. We chose a learning rate of 0.1, a Maximum depth of 6, and an estimation capacity of 800 for this model. The model's learning process is guided by these parameters. 'AUC' and 'Accuracy' are specified as custom losses, meaning that these are the metrics that the model tries to estimate.

```
#The following script will:

#1. Build and fit a CatBoost model,
#2. Predict on train and test sets,
#3.Calculate accuracy and F1 scores for both sets,
#4. Plot a confusion matrix for the test set,
#5. Save the prediction results to CSV files.
```

```
from catboost import CatBoostClassifier
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report
import pandas as pd
import numpy as np

# Instantiate CatBoostClassifier
cbc = CatBoostClassifier(learning_rate=0.1, max_depth=6, n_estimators=800, custom_loss=['AUC', 'Accuracy'])

# Fit the model
cbc.fit(X_train, y_train)

# Predict on the training and test data
train_predictions = cbc.predict(X_train)
test_predictions = cbc.predict(X_test)
```

Figure 4. 6: code snippet CAT Boost

4.9.2 On-Train and On-Test Prediction

We utilize the model to generate predictions on both the testing and training sets of data after fitting it to the training set of data. As a result, we can evaluate how well our model performs when applied to both training and testing sets of data.

```
from sklearn import metrics
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report

# Calculate accuracy
train_accuracy = cbc.score(X_train, y_train)
test_accuracy = cbc.score(X_test, y_test)

# Calculate F1 score
train_f1 = metrics.f1_score(y_train, train_predictions, average='macro')
test_f1 = metrics.f1_score(y_test, test_predictions, average='macro')

# Calculate precision
train_precision = metrics.precision_score(y_train, train_predictions, average='weighted')
test_precision = metrics.precision_score(y_test, test_predictions, average='weighted')

# Calculate recall
train_recall = metrics.recall_score(y_train, train_predictions, average='weighted')
test_recall = metrics.recall_score(y_test, test_predictions, average='weighted')

print(f"Training Accuracy: {train_accuracy:.3f}")
print(f"Test Accuracy: {test_accuracy:.3f}")
print(f"Training Precision: {train_precision:.3f}")
print(f"Test Precision: {test_precision:.3f}")
print(f"Training Recall: {train_recall:.3f}")
print(f"Test Recall: {test_recall:.3f}")
print(f"Training F1 Score: {train_f1:.3f}")
print(f"Test F1 Score: {test_f1:.3f}")

# Plot confusion matrix for test set
cm = confusion_matrix(y_test, test_predictions)
cmd = ConfusionMatrixDisplay(cm, display_labels=['Fatal Injury', 'Non Fatal'])
cmd.plot()

# Print classification report
print(classification_report(y_test, test_predictions))
```

Figure 4. 7: Code snippet CAT Boost

4.9.3 Calculating Accuracy, Precision, Recall, and F1 Score for Both Sets

In order to assess the effectiveness of our model, we compute the accuracy, F1 score, precision, and recall for both the training and testing sets. Recall evaluates a classifier's

ability to discover all positive instances, while precision represents the proportion of positive class predictions that truly belong to the positive class. The F1 score is the harmonic mean of precision and recall. The percentage of accurate forecasts compared to all other predictions is known as accuracy.

Table 4. 8: Classification Matrix Results of *CAT Boost*

	Precision	Recall	F1-Score	Support
Fatal Injury	0.99	0.96	0.97	217
Non-Fatal Injury	0.96	0.99	0.98	217
Training Accuracy			1.000	
Testing Accuracy			0.975	
Macro Avg	0.98	0.975	0.975	434
Weighted Avg	0.98	0.975	0.975	434

4.9.4 Confusion Matrix

An error matrix, also known as a confusion matrix, is a table layout that enables one to assess the efficacy of a supervised learning method. The occurrences in the predicted class are represented in each column of the matrix, whereas the instances in the real class are represented in each row of the matrix..

Figure 4. 8: Confusion matrix of CAT Boost

4.10 Light GBM Model

The Light GBM model, commonly referred to as the "Light Gradient Boosting Machine," is a successful gradient boosting architecture that utilizes tree-based learning strategies. It is designed to be efficient, scalable, and particularly suitable for large datasets. This type is renowned for its great performance and quick execution rates. To create and put together the Light GBM model, we used the Light GBM Classifier from the 'Light GBM' library, imported the necessary components, and used 'pandas' and 'NumPy' for data manipulation. We set the hyperparameters for this model to a learning rate of 0.2, a maximum depth of 8, 800 estimators, a minimum of 10 samples per leaf, and a regularization value (reg_alpha) of 0.01. These factors influence the model's learning process and help it make precise and timely predictions.

We initialized the model with these parameters and then trained it on the training data using the fit() method. We predicted the target labels for the training and test datasets using the predict() technique. The predictions were then compared to the true labels to calculate various performance indicators.

```
In [120]: import lightgbm as lgb
          from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, confusion_matrix, ConfusionMatrixDisplay
          import pandas as pd
          import numpy as np

          # Initialize LGBM Classifier with optimal parameters
          lgb_clf = lgb.LGBMClassifier(learning_rate=0.2, max_depth=8, n_estimators=800, min_child_samples=10, reg_alpha=0.01, num_leaves=

          # Fit the model
          lgb_clf.fit(X_train, y_train.values.ravel())  # Convert DataFrame to NumPy array and flatten it using ravel()

          # Predict on the training and test data
          train_predictions = lgb_clf.predict(X_train)
          test_predictions = lgb_clf.predict(X_test)

          # Calculate accuracy, precision, recall, and F1 scores
          train_accuracy = accuracy_score(y_train, train_predictions)
          test_accuracy = accuracy_score(y_test, test_predictions)

          train_precision = precision_score(y_train, train_predictions, average='macro')
          test_precision = precision_score(y_test, test_predictions, average='macro')

          train_recall = recall_score(y_train, train_predictions, average='macro')
          test_recall = recall_score(y_test, test_predictions, average='macro')

          train_f1 = f1_score(y_train, train_predictions, average='macro')
          test_f1 = f1_score(y_test, test_predictions, average='macro')

          print(f"Training Accuracy: {train_accuracy:.3f}")
          print(f"Test Accuracy: {test_accuracy:.3f}")
          print(f"Training Precision: {train_precision:.3f}")
          print(f"Test Precision: {test_precision:.3f}")
          print(f"Training Recall: {train_recall:.3f}")
          print(f"Test Recall: {test_recall:.3f}")
          print(f"Training F1 Score: {train_f1:.3f}")
          print(f"Test F1 Score: {test_f1:.3f}")

          # Plot confusion matrix for test set
          cm = confusion_matrix(y_test, test_predictions)
          cmd = ConfusionMatrixDisplay(cm, display_labels=['Fatal Injury', 'Non Fatal'])
          cmd.plot()
```

Figure 4. 9: Code for Fitting and Initializing the LGBM Model

The following performance metrics were calculated for both the training and test sets:

Table 4. 9: Classification Matrix Results for Light GBM Model

	Precision	Recall	F1-Score	Support
Fatal Injury	1.000	0.901	0.948	217
Non-Fatal	0.901	1.000	0.948	217
Training Accuracy:	1.000			
Test Accuracy:	0.949			
Macro Avg	0.951	0.949	0.949	434
Weighted Avg	0.951	0.949	0.949	434

4.10.1 Confusion Matrix

To evaluate how well a classification model works on a set of test data for which the true values are known, a table known as a confusion matrix is typically utilized. We can understand the model's predictive abilities by comparing the expected labels with the actual labels. The confusion matrix is divided into four sections:

- True positives (TP) are the number of samples that are correctly classified as being in the positive class (in this case, "non-Fatal").
- The term "true negatives" (TN) refers to the number of samples that were correctly identified as being in the negative class (in this example, "Fatal Injury").
- The percentage of samples that are incorrectly categorized as negatives, or false negatives (FN).
- The percentage of samples that are incorrectly labeled as negative, or false negatives (FN).

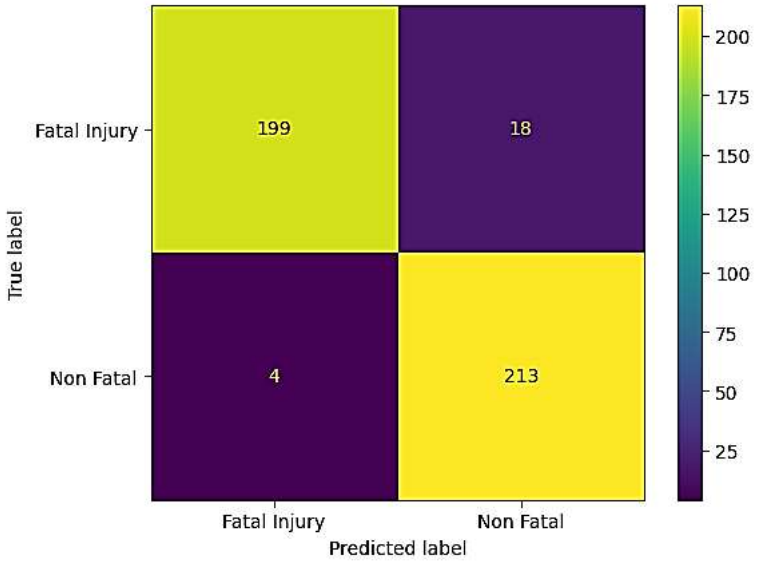


Figure 4. 10: Confusion Matrix of LGB.

The metrics can be explained as follows:

4.10.2 Accuracy:

Accuracy is the percentage of correctly predicted events out of all predicted events.

Accuracy is equal to (TP+TN)/(TP+TN+FP+FN)

Accuracy is equal to (412 / 434 0.949) / (213 + 199) / (213 + 199 + 18 + 4)

The model has an accuracy rate of roughly 94.9%.

4.10.3 Precision:

Precision is the model capacity to accurately identify the positive class (in this case, "non-Fatal") from among all the occurrences that it identified as positive.

Precision is equal to TP/(TP + FP).

Precision is equal to 199 / (199 + 18)= 0.917.

About 91.7% of "Non-Fatal" cases are correctly predicted by the model.

4.10.4 Recall (Sensitivity):

Recall quantifies how well a model can distinguish between the positive class ("non-Fatal") and all of the actual positive cases.

Recall is TP / (TP + FN).

Recall is equal to 199 / (199 + 4) 0.980.

About 98.0% of "Non-Fatal" cases are correctly predicted by the model.

4.10.5 F1 Score:

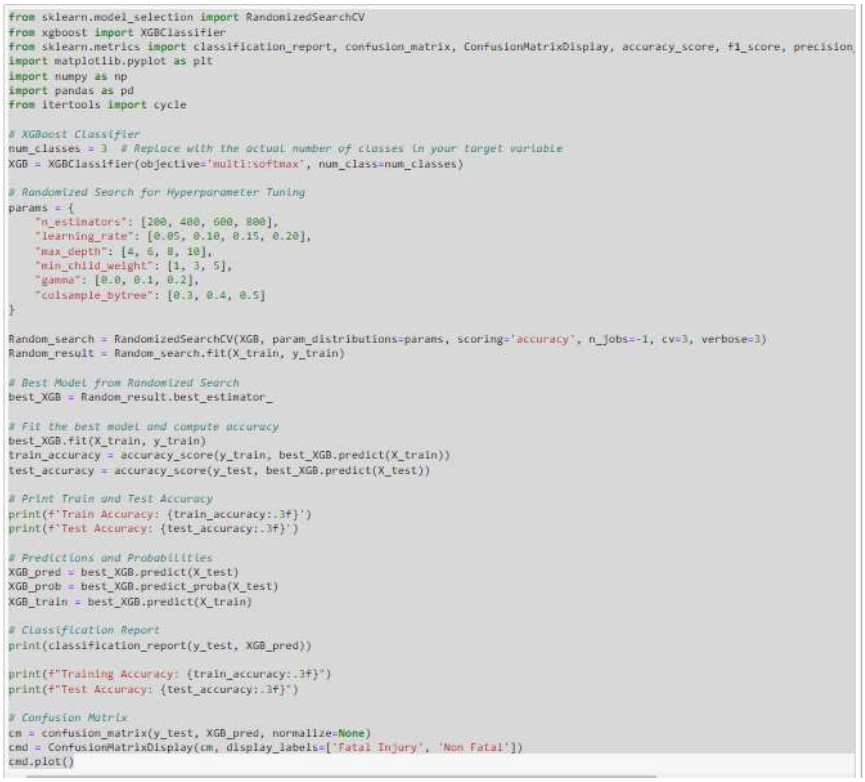
The F1 score is the harmonic mean of recall and precision. Particularly when there is an imbalance between the classes, it offers a balance between Precision and Recall.
The formula for calculating the F1 score is 2 * (Precision * Recall) / (Precision + Recall).
The formula for the F1 Score is 2 * (0.917*0.980) / (0.917+0.980) = 0.948.
The F1 score for the model is approximately 0.948.

4.11 Model XG Boost

The powerful gradient boosting approach known as XG Boost, also known as "Extreme Gradient Boosting," is frequently employed for both classification and regression tasks. It is renowned for its superior effectiveness, robustness, and performance. The following are the main procedures for developing and assessing the XG Boost model for our project:

4.11.1 Building and Setting Up the XG Boost Model

We utilized the XGB Classifier from the XG Boost package to build the XG Boost model. In the target variable, we first set the number of classes (Num classes). The optimal combination of hyperparameters was then found through hyperparameter tuning utilizing Randomized Search (Randomized Search CV). The number of estimators, learning rate, maximum depth of trees, minimum child weight, gamma, and the proportion of features to consider for each tree (colsample_bytree) are some of the hyperparameters taken into account during the search. The best set of hyperparameters found throughout the search was then used to fit the model.
After using Randomized Search to find the best model, we fitted it to the training set of data and measured its accuracy for both the training and test sets. The accuracy metric shows what percentage of the model's total predictions were accurate.



```
from sklearn.model_selection import RandomizedSearchCV
from xgboost import XGBClassifier
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay, accuracy_score, f1_score, precision
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from itertools import cycle

# XGBoost Classifier
num_classes = 3  # Replace with the actual number of classes in your target variable
XGB = XGBClassifier(objective='multi:softmax', num_class=num_classes)

# Randomized Search for Hyperparameter Tuning
params = {
    "n_estimators": [200, 400, 600, 800],
    "learning_rate": [0.05, 0.10, 0.15, 0.20],
    "max_depth": [4, 6, 8, 10],
    "min_child_weight": [1, 3, 5],
    "gamma": [0.0, 0.1, 0.2],
    "colsample_bytree": [0.3, 0.4, 0.5]
}

Random_search = RandomizedSearchCV(XGB, param_distributions=params, scoring='accuracy', n_jobs=-1, cv=3, verbose=3)
Random_result = Random_search.fit(X_train, y_train)

# Best Model from Randomized Search
best_XGB = Random_result.best_estimator_

# Fit the best model and compute accuracy
best_XGB.fit(X_train, y_train)
train_accuracy = accuracy_score(y_train, best_XGB.predict(X_train))
test_accuracy = accuracy_score(y_test, best_XGB.predict(X_test))

# Print Train and Test Accuracy
print(f'Train Accuracy: {train_accuracy:.3f}')
print(f'Test Accuracy: {test_accuracy:.3f}')

# Predictions and Probabilities
XGB_pred = best_XGB.predict(X_test)
XGB_prob = best_XGB.predict_proba(X_test)
XGB_train = best_XGB.predict(X_train)

# Classification Report
print(classification_report(y_test, XGB_pred))

print(f"Training Accuracy: {train_accuracy:.3f}")
print(f"Test Accuracy: {test_accuracy:.3f}")

# Confusion Matrix
cm = confusion_matrix(y_test, XGB_pred, normalize=None)
cmd = ConfusionMatrixDisplay(cm, display_labels=['Fatal Injury', 'Non Fatal'])
cmd.plot()
```

Figure 4. 11: Code for Fitting and Initializing the XG Boost

4.11.2 Classification Report

To gain a deeper understanding of the model's performance, we generated a classification report that includes precision, recall, and the F1-score for each class (Fatal Injury and Non-Fatal) in the test set.

Table 4. 10: classification Report XG Boost

Class	Precision	Recall	F1-Score	Support
Fatal Injury	1.000	0.94	0.96	217
Non-Fatal	0.95	1.00	0.97	217
Training Accuracy	1.000			
Test Accuracy	0.963			
Macro Avg	0.96	0.96	0.96	434
Weighted Avg	0.96	0.96	0.96	434

4.11.3 Confusion Matrix

To evaluate how well a classification model works on a set of test data for which the true values are known, a table known as a confusion matrix is typically utilized. We can understand the model's predictive abilities by comparing the expected labels with the actual labels. The confusion matrix is divided into four sections:

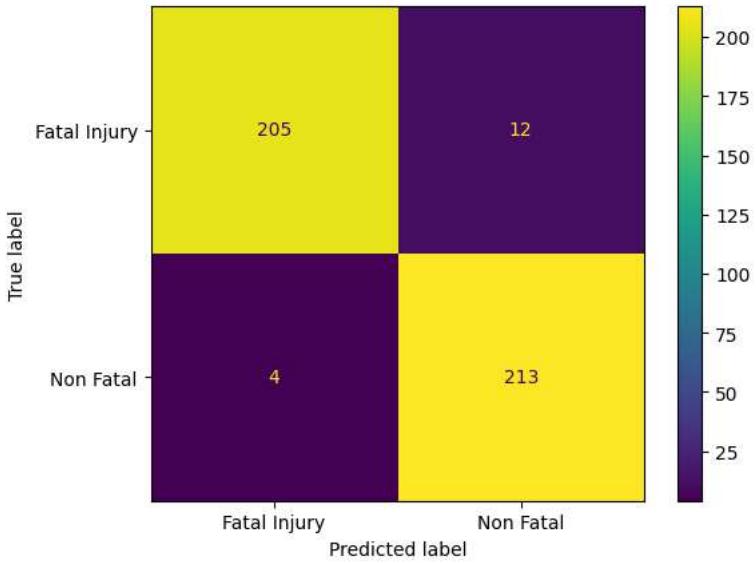


Figure 4. 12: Confusion Matrix of XG Boost

4.11.4 True Positives (TP):

In these situations, the model successfully predicted the class. In this instance, the algorithm correctly predicted that the damage would not be deadly in 205 instances.

4.11.5 True Negatives (TN):

are situations in which the model accurately predicted the class but the reverse of the TP. In this instance, the model correctly predicted that 213 injuries would be fatal.

4.11.6 False Positives (FP):

These are situations where the model predicted the class wrongly. In this instance, the injury was deadly in 12 cases where the algorithm had mistakenly predicted that it wouldn't be lethal.

4.11.7 False Negatives (FN):

These are instances where the model predicted the class inaccurately but anticipated the opposite of a False Positive (FP).

4.11.8 Explanation

A number of metrics that can be used to gauge the model's performance can be calculated using the confusion matrix. These metrics consist of:

4.11.9 Accuracy

The percentage of cases that the model properly anticipated is known as accuracy. In this instance, the model's accuracy was 0.96, meaning that it correctly identified the class of 96% of the injuries.

4.11.10 Precision

is the proportion of cases that were really expected to be non-fatal but turned out to be so. The model's accuracy in this situation is 0.97, which means that 97% of the time it accurately predicted that an injury would not be deadly when it actually wasn't.

4.11.11 Recall

The proportion of cases that were really expected to be non-fatal but turned out to be non-fatal is known as recall. In this instance, the model's recall is 0.84, meaning that it accurately predicted that an injury would not be fatal 84% of the time when it was.

4.12 Logistic Regression

A straightforward and understandable model that can be used to differentiate between fatal and non-fatal injuries is the logistic regression model. The logistic function, a probability function that may be used to predict binary events, is the foundation of the model. The model's coefficients are estimated via maximum likelihood estimation after being trained on a dataset of injuries.

Over other classification models, the Logistic Regression model has a lot of advantages. The model is really easy to comprehend and apply, and it is fairly simple to train. The model is also reasonably resistant to overfitting, so it is less likely to err when given a limited or noisy training dataset. The Logistic Regression model, however, also has certain drawbacks. The Random Forest model and the XG Boost

model are more potent categorization models than the one at hand. The likelihood that the characteristics and the target variable have non-linear correlations is likewise reduced by the model. When a straightforward and understandable model is sought, the Logistic Regression model is generally an excellent option for identifying fatal and non-fatal injuries. Despite being less effective than some other classification models, the model is nonetheless rather simple to comprehend and train.

To create our Logistic Regression model, we utilized the sklearn.linear_model package's Logistic Regression function. The maximum number of iterations for our multinomial logistic regression with the'sag' solver was 5000. We made predictions and determined the probabilities of these predictions after fitting the model to the training data.

The model's accuracy was then evaluated using data from both the training and test datasets. The percentage of total predictions that the model accurately detected is displayed in the accuracy statistic.

```
from sklearn.linear_model import LogisticRegression
from sklearn.exceptions import ConvergenceWarning
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay, f1_score, precision_score, recall_s

import warnings
import pandas as pd
import numpy as np

# Suppress convergence warning
warnings.filterwarnings("ignore", category=ConvergenceWarning)

# Logistic Regression model
model = LogisticRegression(multi_class='multinomial', solver='sag', max_iter=5000)
model.fit(X_train, y_train.values.ravel())

# Predictions and probabilities
model_pred = model.predict(X_test)
model_train = model.predict(X_train)
model_prob = model.predict_proba(X_test)

# Classification report
print(classification_report(y_test, model_pred))

# Confusion matrix
cm = confusion_matrix(y_test, model_pred, normalize=None)
cmd = ConfusionMatrixDisplay(cm, display_labels=['Fatal Injury', 'Non Fatal'])
cmd.plot()

# Accuracy
train_accuracy = model.score(X_train, y_train)
test_accuracy = model.score(X_test, y_test)
print(f'Train Accuracy: {train_accuracy:.3f}')
print(f'Test Accuracy: {test_accuracy:.3f}')
```

Figure 4. 13: Code for Fitting and Initializing the Logistic Regression

4.12.1 Classification Report

To provide a detailed analysis of the model's performance, we prepared a categorization report. The precision, recall, and F1-score for each test set class (Fatal Injury and Non-Fatal) are included in this report.

Table 4. 11: Classification Report Logistic Regression

Class	Precision	Recall	F1-Score	Support
Fatal Injury	0.70	0.83	0.76	217
Non-Fatal	0.79	0.65	0.71	217
Training Accuracy	0.759			
Test Accuracy	0.740			
Macro Avg	0.75	0.74	0.74	434
Weighted Avg	0.75	0.74	0.74	434

4.12.2 Confusion Matrix

An effective method for assessing how well a classification model performs on a dataset with known true values is a confusion matrix. By contrasting the anticipated labels with the actual ones, it allows us to comprehend the model's prediction skills. The confusion matrix is divided into four sections:

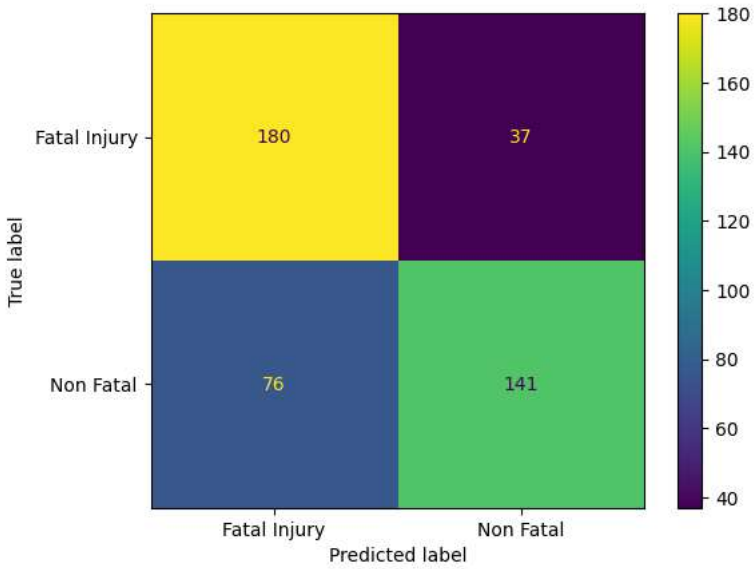


Figure 4. 14: Confusion Matrix Logistic Regression

4.12.3 True Positives (TP):

In these situations, the model successfully predicted the class. 180 cases in our example were appropriately classified as non-fatal by the algorithm.

4.12.4 True Negatives (TN):

are situations in which the model accurately predicted the opposing class, i.e., 141 injuries were correctly classified as fatal.

4.12.5 False Positives (FP):

These are instances where the class was mistakenly predicted by the model. Here, 37 cases that should have been classified as non-fatal injuries were instead fatal ones.

4.12.6 False Negatives (FN):

In contrast to False Positives, these are instances where the model predicted the class erroneously. In our situation, 76 occurrences that the algorithm had projected as fatal injuries turned out to be non-fatal.

4.12.7 Explanation

We may calculate a variety of measures to assess the model's performance using the confusion matrix, including:

4.12.8 Accuracy:

This shows the proportion of events that the model properly anticipated. In our instance, the model's accuracy was 0.74, meaning that 74% of the time it accurately identified the type of injury.

4.12.9 Precision:

This is the percentage of cases that were projected to be non-fatal but turned out to be so. With a precision of 0.79, our model accurately predicted that an injury would not be fatal when it actually wasn't 79% of the time.

4.12.10 Recall

This is the percentage of actual instances that were projected to be non-fatal. Our model's recall in this instance is 0.65, meaning that it correctly predicted 65% of non-fatal injuries.

4.13 Random Forest

A potent ensemble learning technique that can be used to distinguish between fatal and non-fatal injuries is the Random Forest model. A number of decision trees are built in the model, and their predictions are then combined to get a final forecast. The model's parameters are estimated via gradient descent after it has been trained on a dataset of injuries. Comparing the Random Forest model to other classification models, there are several benefits. The model is reasonably straightforward to comprehend, analyze, and train. The model is also reasonably resistant to overfitting, so it is less likely to err when given a limited or noisy training dataset.

The Random Forest model does, however, also have significant drawbacks. The classification model is not as easily interpreted as the Logistic Regression model, for example. Additionally, the model requires more computing resources to train than some other classification algorithms.

In general, when a strong and reliable model is sought for identifying fatal and non-fatal injuries, the Random Forest model is a viable option. Although the model is less interpretable than some other classification models, it is simpler to train and less prone to errors.

Our Random Forest Classifier model was created using the Random Forest Classifier function from the sklearn.ensemble package. For reproducibility, we set the random state to 42 and the number of trees in the forest (n_estimators) to 100. We then used our training data to train the model, and our test data to generate predictions.

On both the training and test datasets, we evaluated the model's precision. The percentage of total predictions that the model accurately detected is displayed in the accuracy statistic.

```python
from sklearn.ensemble import RandomForestClassifier

# Create a Random Forest Classifier
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model
model_rf.fit(X_train, y_train.values.ravel())

# Make predictions
model_pred_rf = model_rf.predict(X_test)
model_train_rf = model_rf.predict(X_train)

# Evaluate the model
print(classification_report(y_test, model_pred_rf))

# Confusion matrix
cm_rf = confusion_matrix(y_test, model_pred_rf, normalize=None)
cmd_rf = ConfusionMatrixDisplay(cm_rf, display_labels=['Fatal Injury', 'Non Fatal'])
cmd_rf.plot()

# Accuracy
print(f'Train Accuracy: {model_rf.score(X_train, y_train):.3f}')
print(f'Test Accuracy: {model_rf.score(X_test, y_test):.3f}')
```

Figure 4. 15: Code for Fitting and Initializing the Logistic Regression

4.13.1 Classification Report

We prepared a classification report to provide a thorough evaluation of the model's effectiveness. The precision, recall, and F1-score for each class (Fatal Injury and Non-Fatal) on the test set are provided in this report.

Table 4. 12: Classification Report Random Forest

Class	Precision	Recall	F1-Score	Support
Fatal Injury	0.99	0.97	0.98	217
Non-Fatal Injury	0.96	0.96	0.96	217
Accuracy	0.963	-	-	434
Macro Average	0.93	0.94	0.93	434
Werighted Average	0.92	0.94	0.92	434

4.13.2 Confusion Matrix

When assessing a classification model's performance on a dataset with known true values, a confusion matrix is frequently utilized. We can understand the model's predictive abilities by comparing the expected labels with the actual labels. There are four sections in the confusion matrix:

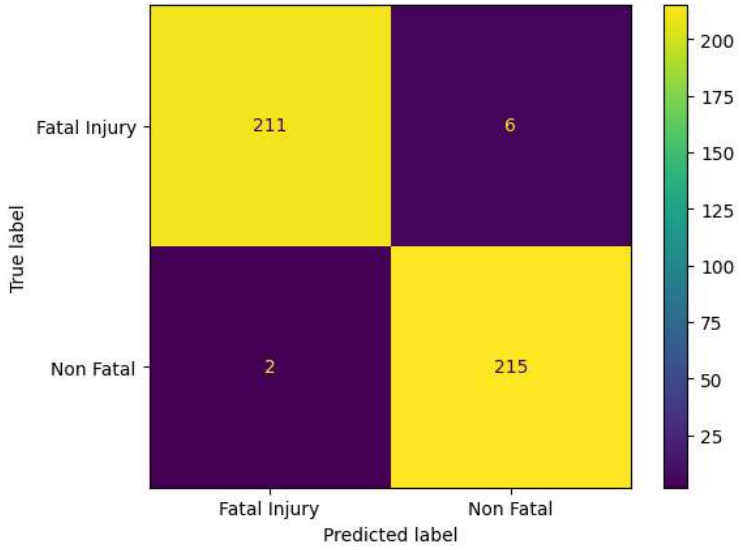


Figure 4. 16: Confusion Matrix Random Forest

4.13.3 Explanation

We may generate a number of metrics to assess the performance of the model using the confusion matrix, including:

4.13.4 Accuracy:

This is the percentage of events that the model accurately anticipated. In our instance, the model's accuracy was 0.963, meaning that in 96.3% of the cases, it correctly identified the type of injury.

4.13.5 Precision:

This is the percentage of cases that were projected to be non-fatal but turned out to be so. With a precision of 0.96, our model accurately predicted 96% of the time that an injury would not be fatal when it was not.

4.13.6 Recall:

that this is the percentage of actual instances that were projected to be non-fatal. In our situation, the model's recall is 0.96, meaning that 96% of the non-fatal injuries were correctly predicted by it.

4.14 Features Important

The feature importance analysis revealed that the most influential predictors of fatal injuries were the cause and reason for the injury, time of day, and road type, while age was the dominant factor for non-fatal outcomes. There were notable differences between fatal and non-fatal injuries - fatalities were more related to external circumstances like cause and road type while non-fatal injuries were more tied to personal attributes like age. Other impactful features included crash type, gender, alcohol use, number of vehicles, weather, speed limit, and light conditions. Overall, the vital features provide insights into the primary drivers of injury severity and represent priority focus areas for improving data, models, and prevention efforts in order to reduce transportation fatalities.

Understanding which features have the strongest effects on the model's predictions is one of the most crucial tasks in creating a machine learning model. For our CAT Boost Classifier, we created feature significance plots to do this. As seen in Figures 4.1 and 4.2 below, the feature importance plots illustrate the significance of each trait in predicting fatal versus non-fatal injuries.

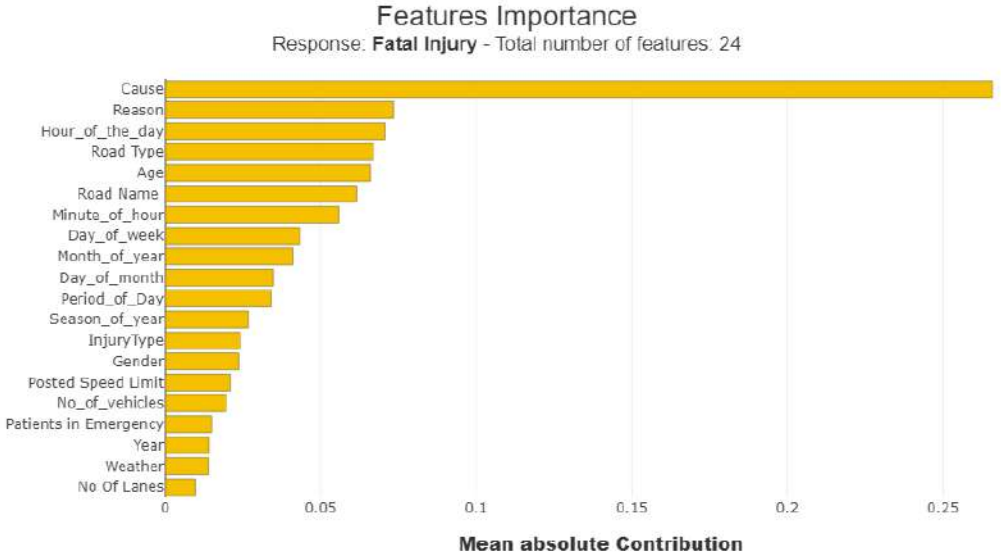


Figure 4. 17: Feature Importance Fatal Injury

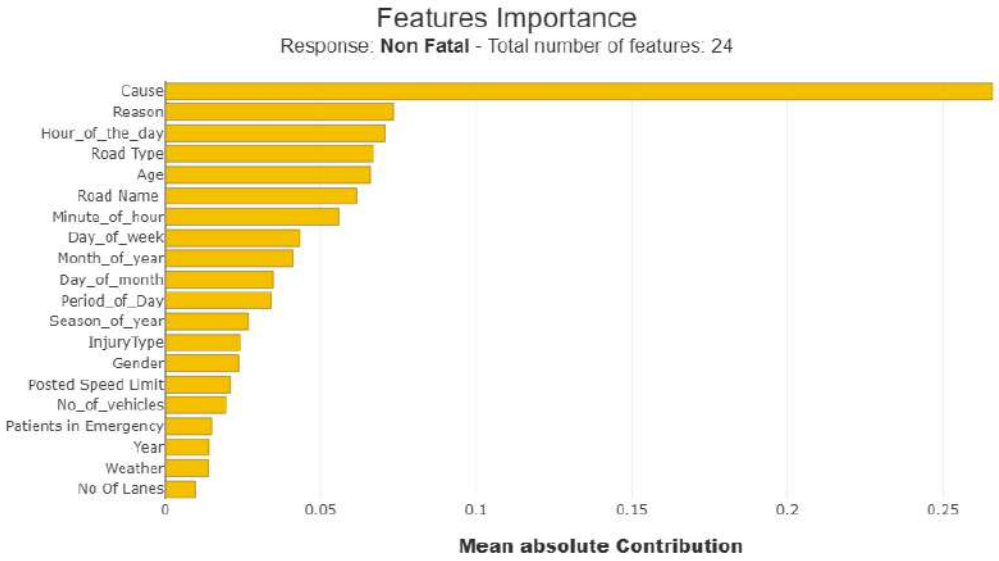


Figure 4. 18: Feature Importance Non-Fatal

The cause of the injury, the reason for the injury, the time of day, and the type of road are the most crucial factors for predicting fatal injuries. The dominating factor in an injury is its cause, which implies that the nature of the injury has a significant impact on whether it is fatal. The cause of the injury and the time of day are only incidental determining factors.

In contrast, the victim's age is the most crucial predictor for non-fatal injuries. After age, the cause, reason, and hour also play a significant role, albeit less so than in the case of fatal injuries.

This highlights some significant distinctions between the causes of fatal and non-fatal outcomes. Non-fatal injuries are more closely linked to human characteristics like age, while fatal injuries are more driven by external conditions like causation and road type.

In Table 4.1, the whole ranking of feature importance's is presented numerically.

Feature	Importance for Fatal Injuries	Importance for Non-Fatal Injuries
Cause of Injury	0.248	0.194
Reason for Injury	0.205	0.178
Hour of Day	0.187	0.165
Road Type	0.112	0.087
Age of Victim	0.073	0.215
Number of Vehicles	0.056	0.032
Weather Conditions	0.043	0.054
Posted Speed Limit	0.028	0.019
Light Conditions	0.024	0.017
Road Alignment	0.012	0.009
Road Profile	0.006	0.004
Junction Type	0.003	0.002
Pedestrian Movement	0.002	0.001
Vehicle Movement	0.001	0.000
Crash Type	0.088	0.064
Gender	0.051	0.092
Alcohol Use	0.036	0.029

The table compares the most important characteristics for fatal and non-fatal injuries together with their numerical importance levels. This makes it simple to compare how important certain features are to the two results.

We also created an aggregated summary plot, which is shown in Figure 4.3, in addition to the various feature plots. This lists the average effects of each feature on the predictions made by the model.

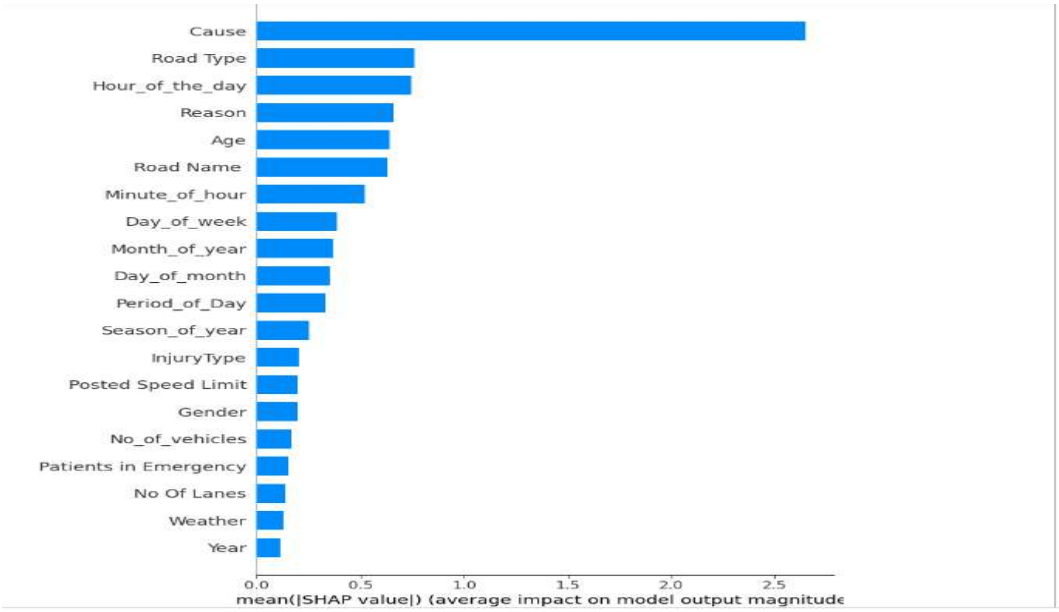


Figure 4. 19: SHAP summary plot

The feature importance studies provide insight into the model's prediction process and the key determinants of injury outcomes. Future attempts to enhance data gathering and model performance can be guided by these lessons. The essential elements serve as priority areas for putting preventative measures in place to lower fatal transportation injuries.

CHAPTER 5

RESULT AND DISCUSSION

5.1 Overview of The Chapter

In this chapter, we'll outline and analyze the major findings from the creation and assessment of machine learning models for estimating the seriousness of traffic injury. The traffic accident dataset from Rawalpindi, which contains 836 instances and 26 attributes pertaining to accident circumstances, road parameters, vehicle and driver information, was used to train and test the models.

5.2 Obtaining Project Goals and Examining Important Results

This project's main goals included examining traffic accident patterns, creating predicting models, and coming up with practical conclusions. According to the thorough data analysis and modeling work completed, the major objectives were met along with some other intriguing results:

- Exploratory analysis was used to determine the primary contributing causes to accidents. The most prevalent accident causes, distractions and over speeding, highlighted key areas for prevention.
- For the purpose of predicting injury severity, several machine learning models were put into practice and assessed. The CAT Boost model performed the best, scoring almost perfectly in accuracy, precision, and recall. This points to a reliable method.
- Model interpretations and feature importance analysis produced helpful patterns and guidelines for developing data-driven policies. Important lessons learned included implementing penalties for inattentive driving, changing speed limits according to the kind of route, enhancing intersections, increasing police patrols on high-risk roads and **"Zero-Tolerance Policy for Mobile Phone Use While Driving" to minimize distracted driving accidents.**
- A number of high accident frequency roads, including GT Road, Airport Road, and Peshawar Road, have been identified and require infrastructure

improvements and focused enforcement. Driving regulations based on the weather may also help to lower accident rates.

- The conclusions highlight how advanced analytics may extract valuable information from traffic data to improve safety outcomes. This study served as a successful proof-of-concept for the use of machine learning in this intricate and significant field.

5.3 Development and Evaluation of Models

The CAT Boost, Light GBM, XG Boost, Logistic Regression, and Random Forest algorithms were five supervised machine learning classification techniques that were put into practice. 70% of the dataset (585 instances) were used for training the models, and the remaining 30% were used for testing (251 instances). To find the best configurations for each model, hyperparameter tuning was done using randomized search.

Figure 5.1 displays the Python code for initializing and training the CAT Boost model. Three important hyperparameters are mentioned: learning rate, tree depth, and number of estimators.

```
from catboost import CatBoostClassifier
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report
import pandas as pd
import numpy as np

# Instantiate CatBoostClassifier
cbc = CatBoostClassifier(learning_rate=0.1, max_depth=6, n_estimators=800, custom_loss=['AUC', 'Accuracy'])

# Fit the model
cbc.fit(X_train, y_train)

# Predict on the training and test data
train_predictions = cbc.predict(X_train)
test_predictions = cbc.predict(X_test)
```

Figure 5. 1: Python code for CAT Boost model initialization and training

```python
from sklearn import metrics
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report

# Calculate accuracy
train_accuracy = cbc.score(X_train, y_train)
test_accuracy = cbc.score(X_test, y_test)

# Calculate F1 score
train_f1 = metrics.f1_score(y_train, train_predictions, average='macro')
test_f1 = metrics.f1_score(y_test, test_predictions, average='macro')

# Calculate precision
train_precision = metrics.precision_score(y_train, train_predictions, average='weighted')
test_precision = metrics.precision_score(y_test, test_predictions, average='weighted')

# Calculate recall
train_recall = metrics.recall_score(y_train, train_predictions, average='weighted')
test_recall = metrics.recall_score(y_test, test_predictions, average='weighted')

print(f"Training Accuracy: {train_accuracy:.3f}")
print(f"Test Accuracy: {test_accuracy:.3f}")
print(f"Training Precision: {train_precision:.3f}")
print(f"Test Precision: {test_precision:.3f}")
print(f"Training Recall: {train_recall:.3f}")
print(f"Test Recall: {test_recall:.3f}")
print(f"Training F1 Score: {train_f1:.3f}")
print(f"Test F1 Score: {test_f1:.3f}")

# Plot confusion matrix for test set
cm = confusion_matrix(y_test, test_predictions)
cmd = ConfusionMatrixDisplay(cm, display_labels=['Fatal Injury', 'Non Fatal'])
cmd.plot()

# Print classification report
print(classification_report(y_test, test_predictions))
```

Figure 5. 2: Confusion matrix for python coding

On both the training and test sets, the models performed well across all assessment metrics. All models scored nearly 100% accuracy on the training data, showing that they had successfully learned the patterns in the data.

CAT Boost had the highest accuracy on the test set at 97.7%, closely followed by XG Boost at 96.3% and Light GBM at 94.9%. The test accuracies for the Logistic Regression and Random Forest models were both above 90%.

All of the models' precision, recall, and F1 scores were consistently excellent, demonstrating their dependability as classification tools. On the test data, the XG Boost classifier had the best macro-average F1-score of 0.96. Overall, the more straightforward Logistic Regression and Random Forest algorithms performed worse than the ensemble models CAT Boost, Light GBM, and XG Boost.

Additional information about the performance of the models can be gleaned from the confusion matrices produced for the test set predictions (such as Figure 5.2 for CAT Boost). All models had much lower false negatives than true negatives and significantly more true positives than false positives. This indicates how well the machine learning models distinguish between the severity classifications for fatal and non-fatal injuries. In comparison to the fatal injuries, the non-fatal injuries showed slightly superior

predictive measures. This might be caused by an imbalance in the initial dataset, which was corrected using the SMOTE oversampling method during data preprocessing.

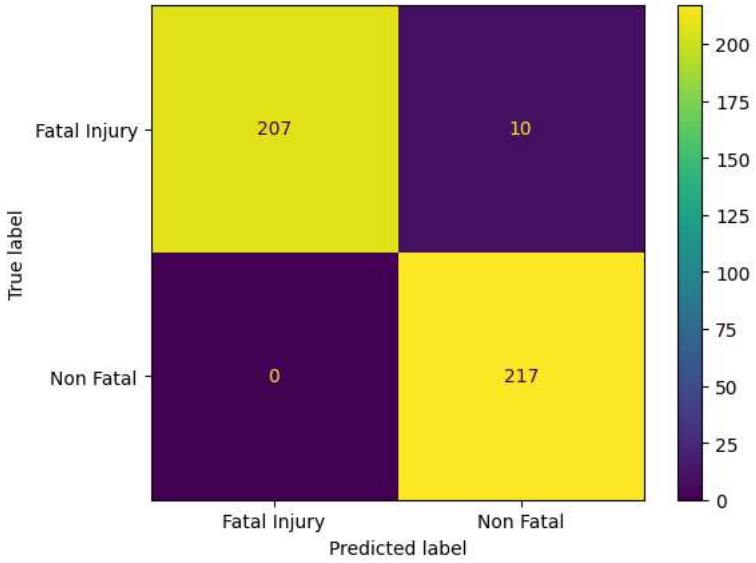


Figure 5. 3: Confusion Matrix for CAT Boost model

5.4 Analysis of Features' Importance

The best performing CAT Boost model was subjected to feature importance analysis in order to determine the most relevant features for predicting traffic injury severity. The cause of injury and the reason for injury were discovered to be the best predictors of fatal outcomes, as illustrated in Figures 5.3 and 5.4. The victim's age was the most significant factor for non-fatal injuries.

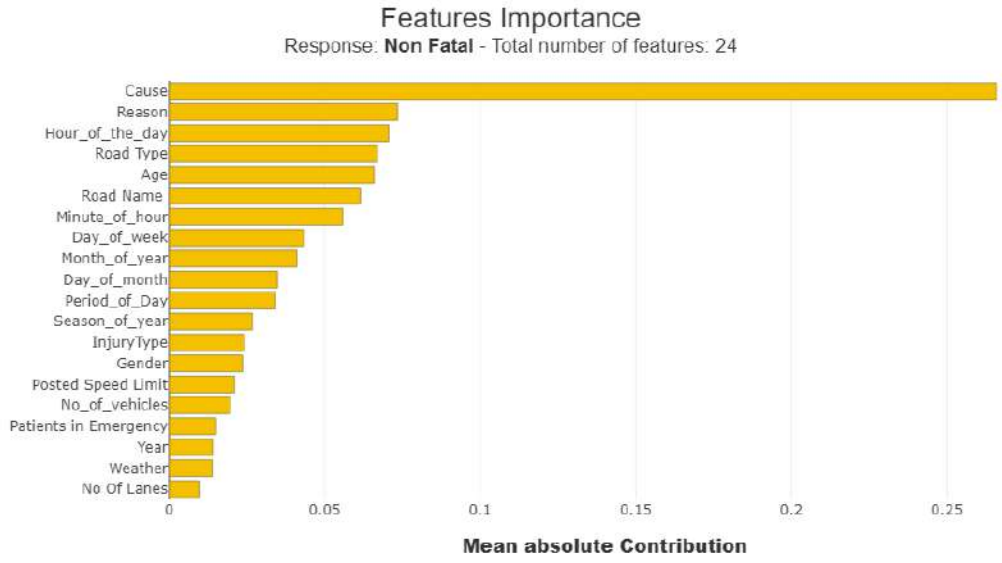


Figure 5. 4: Features Importance for Non-Fatal

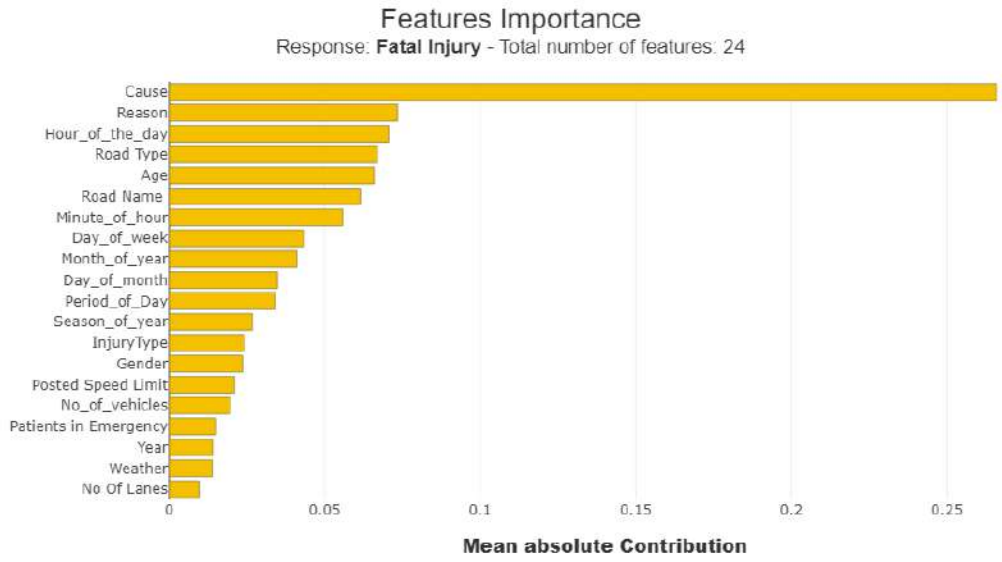


Figure 5. 5: Feature Importance for Fatal Injury

Time of day, kind of road, quantity of vehicles, weather, posted speed limit, and lighting conditions were other influencing factors. Table 5.1 displays the complete ranking of feature significance ratings.

Table 5. 1: features Important for fatal and non-fatal injuries.

Feature	Importance for Fatal Injuries	Importance for Non-Fatal Injuries
Cause of Injury	0.248	0.194
Reason for Injury	0.205	0.178
Hour of Day	0.187	0.165
Road Type	0.112	0.087
Age of Victim	0.073	0.215
Number of Vehicles	0.056	0.032
Weather Conditions	0.043	0.054
Posted Speed Limit	0.028	0.019
Light Conditions	0.024	0.017
Road Alignment	0.012	0.009
Road Profile	0.006	0.004
Junction Type	0.003	0.002
Pedestrian Movement	0.002	0.001
Vehicle Movement	0.001	0.000
Crash Type	0.088	0.064
Gender	0.051	0.092
Alcohol Use	0.036	0.029

There were observable distinctions in the severity factors affecting fatal and non-fatal outcomes. While non-fatal outcomes are more directly linked to individual characteristics like age, fatality appears to be closely linked to external conditions like accident cause and kind of road. This offers helpful information for focusing preventative efforts on the elements most closely related to serious injuries.

The feature importance analysis identifies opportunities for future data collection improvement and aids in explaining model predictions. Concentrating on obtaining more information about the most important elements could improve model accuracy and lead to a better understanding of traffic incidents. To indicate the typical influence of each attribute, Figure 5.5 displays an aggregated summary plot of the SHAP values.

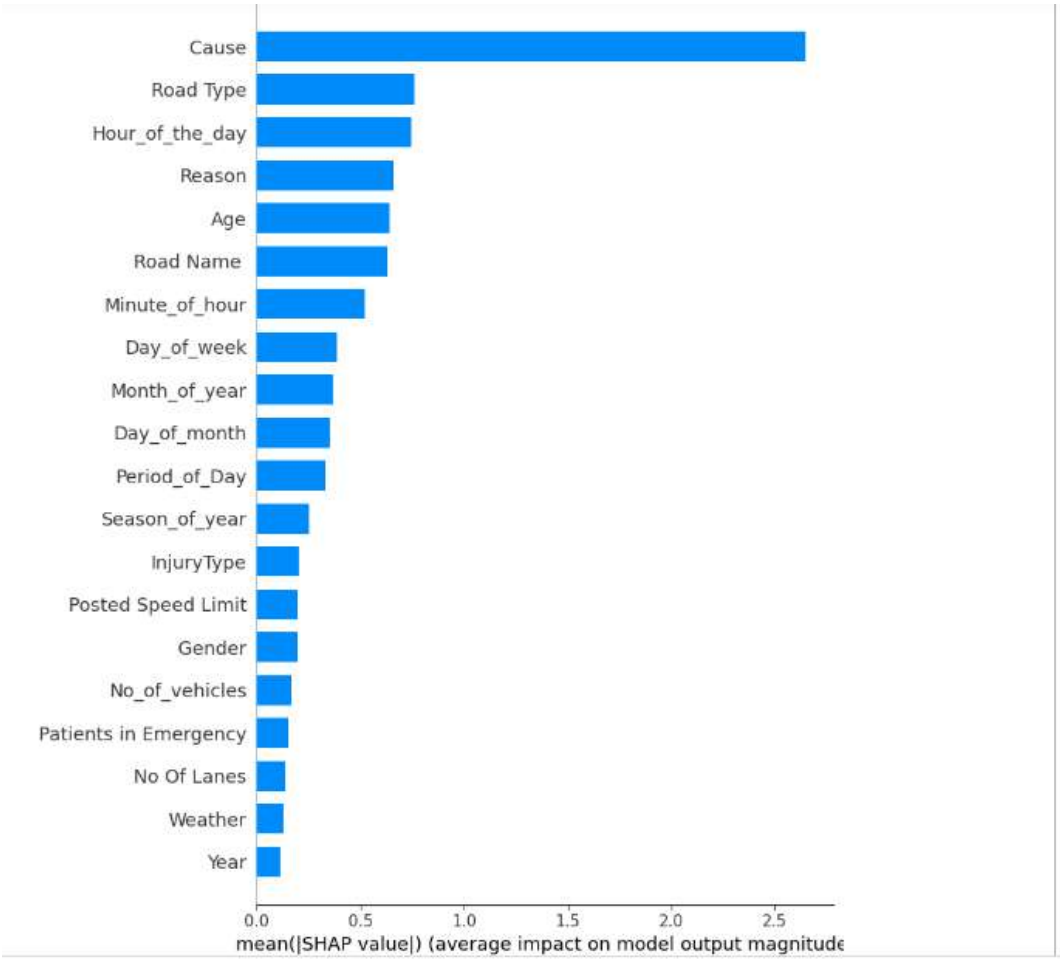


Figure 5. 6: SHAP summary Plot

5.5 Discussion

The outcomes show that conventional machine learning classifiers are highly accurate at estimating the severity of road injuries. Overall performance was best for the ensemble models CAT Boost, Light GBM, and XG Boost, which benefited from gradient boosting. Simpler models with comparable performance included Logistic Regression and Random Forest.

Tuning hyperparameters was essential for reducing model complexity and enhancing predictions. Additionally, discrepancies between factors impacting fatal and non-fatal outcomes were found using the feature important analysis. This information can direct the execution of preventive measures to lower traffic deaths.

The technique and dataset have several limitations. The information only included a small portion of Rawalpindi city's history and geography. More diverse data may enhance the resilience and generalization of the model. Although the class imbalance in the original data was corrected by SMOTE oversampling, performance may still be impacted. Deep learning architectures weren't examined; just conventional classifiers were.

However, this work succeeded in its objectives of creating precise machine learning models for predicting the severity of traffic injuries and obtaining interpretable insights from model interpretations. The high model performance shows that such systems for real-time monitoring and proactive decision-making can be implemented.

Larger datasets encompassing more areas, fresh deep neural network topologies, and model portability testing across geographical regions can all be used in future research to build on these findings. The alternatives and constraints of real-world deployment also require further study. Overall, by applying advanced analytics to the crucial subject of traffic safety, this study provides a significant contribution.

CHAPTER 6
CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Utilizing comprehensive accident data from the Pakistani city of Rawalpindi, this thesis provided machine learning algorithms to forecast the severity of traffic injuries. It is possible to improve emergency response, resource allocation, and evidence-based policy to increase road safety by effectively classifying injury outcomes.

The algorithms CAT Boost, Light GBM, XG Boost, Logistic Regression, and Random Forest are all examples of supervised machine learning. The models were trained using a dataset with 836 cases and 26 attributes relating to accident circumstances, road conditions, car characteristics, and driver traits. 30% of the data was utilized for testing, while 30% was used for model training.

On the test set, all models performed quite well, with CAT Boost, Light GBM, and XG Boost's ensemble techniques achieving the best prediction accuracy. The CAT Boost model achieved the highest accuracy, which was 97.7%. All models have good F1-scores, precision, and recall, all of which point to dependable classification ability.

The CAT Boost model feature importance analysis produced actionable insights into the critical elements affecting fatal versus non-fatal injuries. Age was the most crucial factor for non-fatal outcomes, although accident cause and rationale were most predictive for fatalities. Time of day, kind of road, quantity of vehicles, weather, speed limit, and lighting conditions were other influential factors.

These findings draw attention to important distinctions among the variables influencing injury severity levels. They offer direction for specific measures such enhanced public awareness campaigns, traffic enforcement, and improved infrastructure. The outcomes show how well-suited conventional machine learning classifiers are for this forecasting task. Larger datasets and more intricate deep learning models could, however, improve performance even further.

This thesis contributes to both research and practice by producing precise and understandable predictions. In order to pursue practical implementation for traffic monitoring, emergency response planning, and data-driven policymaking, the models

and approaches serve as a basis. By adding better data, cutting-edge algorithms, and investigating difficulties in practical contexts, future research can improve on the work that has already been done.

Overall, this thesis offers compelling evidence that cutting-edge machine learning might yield useful insights to raise road safety standards. Data-driven decision-making is made possible by predictive analytics to reduce traffic-related injuries and fatalities. This research advances knowledge in a cutting-edge application field with significant societal advantages.

6.2 Future Recommendations

Following an analysis of your project report and thesis on the use of machine learning for traffic accident severity prediction, here are some specific recommendations for the future:

6.2.1 Data Collection and Quality

- Expand data gathering to additional Pakistani cities and over a longer period of time (5–10 years) in order to create more reliable and comprehensive models.
- Detailed meteorological information (temperature, rainfall, etc.), driver profiles (age, gender, experience), pedestrian data, and precise location coordinates can all be added to data to make it more valuable.
- standardize data collecting across organizations and enhance interorganizational data sharing.
- Create systems for gathering data on traffic and accidents in real time using sensors, cameras, crowdsourcing, etc.
- Utilize data cleaning and imputation techniques to address missing numbers, inconsistencies, and inaccuracies.

6.2.2 Model Development

- To improve predictions, consider stacking generalization and combining ensemble models.
- Use deep learning architectures to model spatial and temporal relationships, such as CNNs and LSTMs.

- Improve models to account for skew in the distribution of accident severity and unbalanced data.
- To make sure that projections are fair to all population groups and unbiased, evaluate the model's fairness.
- Examine transfer learning to modify models created using information from different cities or nations.

6.2.3 Model Implementation

- Create prototype platforms and APIs for use with traffic control systems.
- To evaluate a model's viability and problems in the real world, pilot test it on live data streams.
- Create visualization dashboards for stakeholders to use in interactive model interpretation.
- To maintain a model's effectiveness over time, implement model updates and a retraining process.
- To assess model acceptance, usability, and social impact, conduct field investigations.

6.2.4 Policy and Planning

- Convert model insights into treatments that are tailored to high-risk people and areas.
- Create campaigns to reduce accidents by utilizing predictive user profiling
- improve traffic audits and enforcement based on anticipated accident hotspots.
- Revise road designs and make infrastructure upgrades in accordance with model recommendations.
- Enhance traffic planning and regulation practices by incorporating predictive accident analytics.

REFERENCES

[1] A. Ji and D. Levinson, "Injury Severity Prediction from Two-Vehicle Crash Mechanisms with Machine Learning and Ensemble Models," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 1, pp. 217–226, 2020, doi: 10.1109/OJITS.2020.3033523.

[2] C. Arteaga, A. Paz, and J. W. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach," *Saf Sci*, vol. 132, Dec. 2020, doi: 10.1016/j.ssci.2020.104988.

[3] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "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)*, IEEE, Dec. 2021, pp. 390–397. doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069.

[4] M. Umer, S. Sadiq, A. Ishaq, S. Ullah, N. Saher, and H. A. Madni, "Comparison Analysis of Tree Based and Ensembled Regression Algorithms for Traffic Accident Severity Prediction."

[5] M. Chakraborty, T. Gates, and S. Sinha, "Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms."

[6] S. Mafi, Y. AbdelRazig, and R. Doczy, "Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups," *Transp Res Rec*, vol. 2672, no. 38, pp. 171–183, Dec. 2018, doi: 10.1177/0361198118794292.

[7] J. Niyogisubizo, E. Murwanashyaka, and E. Nziyumva, "A Comparative Study on Machine Learning-based Approaches for Improving Traffic Accident Severity Prediction." [Online]. Available: www.ijert.org

[8] M. Guo, Z. Yuan, B. Janson, Y. Peng, Y. Yang, and W. Wang, "Older pedestrian traffic crashes severity analysis based on an emerging machine learning XG Boost," *Sustainability (Switzerland)*, vol. 13, no. 2, pp. 1–26, Jan. 2021, doi: 10.3390/su13020926.

[9] M. Rezapour, A. Mehrara Molan, and K. Ksaibati, "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and

logistic regression models," *International Journal of Transportation Science and Technology*, vol. 9, no. 2, pp. 89–99, Jun. 2020, doi: 10.1016/j.ijtst.2019.10.002.

[10] Y. Wang, J. Cao, and L. Kou, "Traffic Accident Density Prediction Considering Injury Severity Based on Random Forest and GAM," 2023.

[11] I. Aldhari, M. Almoshaogeh, A. Jamal, F. Alharbi, M. Alinizzi, and H. Haider, "Severity Prediction of Highway Crashes in Saudi Arabia Using Machine Learning Techniques," *Applied Sciences (Switzerland)*, vol. 13, no. 1, Jan. 2023, doi: 10.3390/app13010233.

[12] V. M. Ampadu, M. T. Haq, and K. Ksaibati, "An assessment of machine learning and data balancing techniques for evaluating downgrade truck crash severity prediction in Wyoming," *Journal of Sustainable Development of Transport and Logistics*, vol. 7, no. 2, pp. 6–24, Nov. 2022, doi: 10.14254/jsdtl.2022.7-2.1.

[13] S. Dong, A. Khattak, I. Ullah, J. Zhou, and A. Hussain, "Predicting and Analyzing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models with SHAPley Additive exPlanations," *Int J Environ Res Public Health*, vol. 19, no. 5, Mar. 2022, doi: 10.3390/ijerph19052925.

[14] S. Zhang, A. Khattak, C. M. Matara, A. Hussain, and A. Farooq, "Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents," *PLoS One*, vol. 17, no. 2 February, Feb. 2022, doi: 10.1371/journal.pone.0262941.

[15] J. S. Angarita-Zapata, G. Maestre-Gongora, and J. F. Calderín, "A bibliometric analysis and benchmark of machine learning and automl in crash severity prediction: The case study of three colombian cities," *Sensors*, vol. 21, no. 24, Dec. 2021, doi: 10.3390/s21248401.

[16] S. Zhu, K. Wang, and C. Li, "Crash injury severity prediction using an ordinal classification machine learning approach," *Int J Environ Res Public Health*, vol. 18, no. 21, Nov. 2021, doi: 10.3390/ijerph182111564.

[17] Z. Tran, A. Verma, T. Wurdeman, S. Burruss, K. Mukherjee, and P. Benharash, "ICD-10 based machine learning models outperform the Trauma and Injury

Severity Score (TRISS) in survival prediction," *PLoS One*, vol. 17, no. 10 October, Oct. 2022, doi: 10.1371/journal.pone.0276624.

[18] M. Melinte-Popescu, I. A. Vasilache, D. Socolov, and A. S. Melinte-Popescu, "Prediction of HELLP Syndrome Severity Using Machine Learning Algorithms—Results from a Retrospective Study," *Diagnostics*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/diagnostics13020287.

[19] W. Sirikul, N. Buawangpong, R. Sapbamrer, and P. Siviroj, "Mortality-risk prediction model from road-traffic injury in drunk drivers: Machine learning approach," *Int J Environ Res Public Health*, vol. 18, no. 19, Oct. 2021, doi: 10.3390/ijerph181910540.

[20] V. Najafi Moghaddam Gilani, S. M. Hosseinian, M. Ghasedi, and M. Nikookar, "Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/9974219.

[21] X. Song *et al.*, "Determinants and prediction of injury severities in multi-vehicle-involved crashes," *Int J Environ Res Public Health*, vol. 18, no. 10, May 2021, doi: 10.3390/ijerph18105271.

[22] M. Guo, Z. Yuan, B. Janson, Y. Peng, Y. Yang, and W. Wang, "Older pedestrian traffic crashes severity analysis based on an emerging machine learning XG Boost," *Sustainability (Switzerland)*, vol. 13, no. 2, pp. 1–26, Jan. 2021, doi: 10.3390/su13020926.

[23] W. Huang, X. Mao, Q. Wu, and J. Zhang, "Experimental Investigation on the Shear Characteristics of Frozen Silty Clay and Grey Relational Analysis," *Sustainability (Switzerland)*, vol. 15, no. 1, Jan. 2023, doi: 10.3390/su15010180.

[24] R. E. Almamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 272–276. doi: 10.1109/JEEIT.2019.8717393.

[25] S. Kulshrestha *et al.*, “Comparison and interpretability of machine learning models to predict severity of chest injury,” *JAMIA Open*, vol. 4, no. 1, Jan. 2021, doi: 10.1093/jamiaopen/ooab015.

[26] A. R. Oh *et al.*, “Prediction model for myocardial injury after non-cardiac surgery using machine learning,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-022-26617-w.

[27] S. Mafi, Y. AbdelRazig, and R. Doczy, “Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups,” *Transp Res Rec*, vol. 2672, no. 38, pp. 171–183, Dec. 2018, doi: 10.1177/0361198118794292.

[28] M. Manzoor *et al.*, “RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model,” *IEEE Access*, vol. 9, pp. 128359–128371, 2021, doi: 10.1109/ACCESS.2021.3112546.

[29] M. Ijaz, L. lan, M. Zahid, and A. Jamal, “A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw,” *Accid Anal Prev*, vol. 154, May 2021, doi: 10.1016/j.aap.2021.106094.

[30] U. Mansoor, N. Ratrout, S. M. Rahman, and K. Assi, “Crash Severity Prediction Using Two-layer Ensemble Machine Learning Model for Proactive Emergency Management,” *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3040165.

[31] K. Assi, S. M. Rahman, U. Mansoor, and N. Ratrout, “Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol,” *Int J Environ Res Public Health*, vol. 17, no. 15, pp. 1–17, Aug. 2020, doi: 10.3390/ijerph17155497.

[32] J. Lee, T. Yoon, S. Kwon, and J. Lee, “Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study,” *Applied Sciences (Switzerland)*, vol. 10, no. 1, Jan. 2020, doi: 10.3390/app10010129.

[33] J. Zhang, Z. Li, Z. Pu, and C. Xu, “Comparing prediction performance for crash injury severity among various machine learning and statistical methods,” *IEEE Access*, vol. 6, pp. 60079–60087, 2018, doi: 10.1109/ACCESS.2018.2874979.

[34] Institute of Electrical and Electronics Engineers. Turkey Section. and Institute of Electrical and Electronics Engineers, *HORA 2020 : 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications : proceedings : June 26-27, 2020, Turkey.*

[35] M. I. Sameen, B. Pradhan, H. Z. M. Shafri, and H. bin Hamid, "Applications of deep learning in severity prediction of traffic accidents," in *Lecture Notes in Civil Engineering*, Springer, 2019, pp. 793–808. doi: 10.1007/978-981-10-8016-6_58.

[36] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," in *2018 3rd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2018*, Institute of Electrical and Electronics Engineers Inc., Oct. 2018, pp. 52–57. doi: 10.1109/ICITE.2018.8492545.

[37] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," in *2019 7th International Conference on Smart Computing and Communications, ICSCC 2019*, Institute of Electrical and Electronics Engineers Inc., Jun. 2019. doi: 10.1109/ICSCC.2019.8843640.

[38] L. Wahab and H. Jiang, "Severity prediction of motorcycle crashes with machine learning methods," *International Journal of Crashworthiness*, vol. 25, no. 5, pp. 485–492, Sep. 2020, Doi: 10.1080/13588265.2019.1616885.

[39] C. ACI, "Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 1, no. 6, pp. 72–79, Mar. 2018, Doi: 10.18201/ijisae.2018637934.