Accident Prediction Using Machine Learning: Analyzing Weather Conditions, and Model Performance

University Of Oulu
Faculty of Information-Technology and Electrical Engineering
Master's Thesis
Muhamad Shahroz Abbas
30th May 2023

Acknowledgement

It is with immense gratitude that I acknowledge the support and help of those whose efforts have made it possible for me to complete this thesis. The journey, though challenging, has been rewarding and enlightening, and I owe a deep debt of gratitude to everyone who has played a part in this process. First and foremost, I wish to express my sincere thanks to the University of Oulu, for providing a conducive environment that fostered learning, growth, and the pursuit of knowledge. The enriching experiences I've had here in the Information Processing Science and Software Engineering program have been invaluable and will guide me throughout my future career.

My thesis on "Predicting Accidents" stands as a testament to the mentorship and guidance of my supervisors, Dr. Ella Peltonen and Tero Päivärinta. Their profound expertise in Artificial Intelligence and Machine Learning has been crucial in navigating the complexities of this research. Their patient guidance, constructive criticisms, and unwavering faith in my potential have continually driven me towards excellence. The intellectual challenges they posed along this journey have fostered my critical thinking abilities and spurred intellectual growth. They were always there to assist and support me. I kept advancing and getting better because to her support. They not only assisted me with the research for this thesis, but also taught me how to do research and approach problems in a scientific manner. For this, I am deeply grateful.

I am also thankful for the support of my fellow students and friends, especially Numan Akbar who have been a constant source of encouragement, companionship, and inspiration. Their supportive words and actions, especially during challenging periods, have been instrumental in keeping me focused and motivated. A heartfelt thanks goes to my family, whose unwavering support and belief in me have been the bedrock upon which I stand. Their sacrifices, patience, and love have been my strength throughout this journey.

Finally, I am deeply thankful to all the researchers and authors whose works have been referred to in this thesis. Their significant contributions to the field of artificial intelligence have not only formed the foundation of this work but also spurred my interest and learning in the subject. To all of you, I offer my sincere thanks and appreciation. Your collective effort has made this journey not only possible but also a truly enriching experience.

Abstract

The transportation industry has undergone a technological revolution. The necessity of traveling safely, staying informed, and being updated has multiplied. People may rely on modern vehicles with the newest technology for reliability and safety. However, a tense scenario is made worse by late rescue efforts at accident place, which also significantly raises the accident mortality rate in remote places. This thesis investigates the use of machine learning approaches to forecast the likelihood of accidents, with a particular emphasis on understanding the impact that weather and road conditions have in determining the severity of accidents. Specifically, this thesis investigates how weather and road conditions affect the likelihood of accidents occurring. The purpose of the study is to contribute to the development of accident prevention techniques that are more effective and data-driven by determining the most influential elements that lead to accidents and evaluating the effectiveness of various machine learning models in properly predicting accident likelihood. This will be accomplished by identifying the factors that are most likely to lead to accidents. An exhaustive examination of the relevant literature was carried to determine the current state of the art in accident prediction and to identify the primary factors related with the weather and the road conditions that have an effect on the severity of accidents.

In order to determine whether or not machine learning is useful for predicting the chance of an accident, many models were trained and evaluated on a sizable dataset consisting of actual traffic collisions. Several different metrics, including accuracy, precision, recall, and F1-score, were used to evaluate the performance of the model. A comprehensive analysis of the models was carried out in order to highlight the benefits and drawbacks of each methodological approach. According to the findings of this research, machine learning models are able to accurately forecast the likelihood of accidents with a high level of accuracy when they are presented with an adequate amount of data and features that have been carefully selected. Additionally, the results offer valuable insights into the relationships between weather and road conditions and accident severity, which can inform future transportation planning, infrastructure development, and safety measures.

Keywords: Machine learning, predictions, statistical analysis.

Content

Acknowledgement 2
Abstract 3
1 Introduction 6
2 Literature review 8
2.1 Comparative Analysis of ML Models For Accident Predictions 8
2.2 Weather and its effects on road accidents 16
2.3 Statistical techniques 20
2.3.1 Logistic regression 20
2.3.2 Correlation 21
2.3.3 Tobit regression 22
2.3.4 Chi-square 22
2.4 Handling of Imbalanced Data 23
2.5 Detecting the accidents 24
2.6 Open Data 26
2.6.1 Open Data in Traffic Management Systems 27
2.6.2 Other Datasets and Motivation for Chosen Dataset 28
3 Research methodology 29
3.1 Working of machine learning model 29
3.2 Data selection and description 31
3.3 Categorized the weather variable 36
3.4 One-hot encoding 37
3.5 Data balancing technique 38
3.6 Machine learning (ML) models 39
3.6.1 Naive Bayes 39
3.6.2 Random Forest 40
3.6.3 Extreme Gradient Boost (XGBoost) 41
3.6.4 K-Neighbors Classifier 41
3.6.5 Decision Tree 42
3.7 performance matrix 43
3.7.1 Confusion Matrix 43
3.7.2 Precision 44
3.7.3 Recall 45
3.7.4 F1-score 45
3.8 Statistical techniques used to answer the first question 46
3.8.1 Chi-square test 46
3.8.2 Cramer's V 46
5 Results 48
5.1 Statistical analysis 48
5.1.1 Conclusion 50
5.2 Comparing the Performance of Machine Learning Models in Predicting

Accident Likelihood 53
5.2.1 Random forest 53
5.2.2 Decision Tree 54
5.2.3 K-Neighbours Classifier 54
5.2.4 Extreme Gradient Boost (XGBoost) 55
5.2.5 Naive Bayes 55
5.2.6 Conclusion 56
6 Discussion 57
6.1 Limitation and area of improvement 58
6.2 Potential application areas and examples 58
7 Summary 59
8 References 60

1 Introduction

The transportation system is an important part of our lives in today's fast-paced world. Where it is serving the mankind, it also has some disadvantages. The number of accidents is rising as there are more cars on the road. According to the World Health Organization (WHO), traffic accidents are the ninth most common cause of death worldwide and the number one killer of people between the ages of 15 and 29. Road accidents often result in injuries, property damage, and financial losses in addition to fatalities (World Health Organization 2020).

The first important field of research that can aim in lowering the number of accidents on the road is accident prediction using artificial intelligence (AI). Statistical techniques, which had a limited degree of accuracy and efficiency, were used in the conventional methods of accident prediction. Yet, new opportunities for accident prediction using machine learning methods have emerged as a result of recent developments in artificial intelligence (AI). The use of AI in accident prediction involves the development of predictive models that can analyze huge amounts of datasets to find out patterns and predict potential accidents. These models can take into account a range of factors such as weather conditions, location coordinates, accident severity, temperature, wind chill, humidity, air pressure, wind speed, precipitation, and visibility. By analyzing these factors, the models can provide insights into the likelihood of accidents occurring and suggest strategies to mitigate the risk. The goal of the thesis is to present a thorough analysis of AI accident prediction machine learning models and advanced statistical analysis techniques. The thesis will examine the literature that has already been written on the subject, study the various machine learning algorithms utilized for accident prediction, and assess each technique's efficacy. The thesis will also examine the difficulties and restrictions associated with utilizing AI to forecast accidents and make recommendations for solutions.

The primary objective of this thesis is to predict accidents and answer the following research question:

- *RQ1: How do weather and road conditions affect severity of accident and what are the most important factors that contribute to the likelihood of accident?*
- *RQ2: Can machine learning models be used to accurately predict the likelihood of accidents, and how do different models compare in terms of accuracy and reliability?*

To achieve these objectives, an open-source accident dataset will be used and pre-processed using various techniques such as variable selection, missing data elimination, and data balancing through oversampling using the Synthetic Minority Over-sampling Technique (SMOTE). The pre-processed dataset will be used to train and test different machine learning models, including decision trees, random forests, naive bayes, extreme gradient boost, and neural networks. While a considerable amount of research has been conducted on accident prediction using machine learning, most of the literature has only considered the overall accuracy of these models to predict the accident. Likewise, in the statistical analysis, the existing body of work has primarily centered on different statistical analysis such as chi-square, correlation, or logistic regression. Therefore, this thesis will strive to fill these gaps by not only examining the overall accuracy of various machine learning models but also considering other performance metrics like precision, recall, and F1 score. Additionally, this thesis will also augment chi-square analysis with other statistical measures, such as p-value, degree of freedom, and Cramer's V to find out the

elements that contribute to the likelihood of accidents, and to determine how weather conditions affect the severity of accidents.

The results of these analyses will provide insights into the factors that contribute to accidents, how different machine learning models perform in predicting accidents, and how weather and road conditions affect the severity of accidents. The significance of this research lies in its potential to contribute to the development of more accurate and efficient accident prediction systems that can enhance road safety and minimize the number of fatalities and injuries. The thesis will be valuable to researchers, decision-makers, and stakeholders in transportation who are making efforts to increase traffic safety and minimize the number of accidents on the road.

2 Literature review

Due to its ability to both avoid and lessen the effects of accidents, accident prediction and detection is a crucial undertaking in the field of safety engineering. It has become more popular in recent years to use artificial intelligence (AI) methods to anticipate accidents and increase traffic safety. AI-based accident prediction system uses machine learning algorithms to evaluate past data and find patterns that might forecast the likelihood of upcoming accidents. AI accident prediction has been the subject of numerous investigations, with encouraging outcomes. In this thesis, the literature review will explore the different AI techniques used for accident prediction, detection, and their applications. Machine learning is one of the methods of AI that is most frequently used to anticipate accidents. Predictions can be made utilizing patterns and relationships that machine learning algorithms can identify in data. Various machine learning algorithms such as decision trees, neural networks, support vector machines, and random forests can be used for accident prediction.

2.1 Comparative Analysis of ML Models For Accident Predictions

In a study the researchers discussed the idea of ensemble learning, which entails creating numerous classifiers and combining their output for improved performance. Boosting and tagging are two common approaches to group learning. Boosting involves adding more weight to points that earlier predictors misclassified, whereas bagging builds each tree separately using a bootstrap sample of the data. After that, the article presents the concept of random forests, which Breiman introduced in 2001. When building classification or regression trees, random forests, an extension of bagging, introduce an additional level of randomness. In standard trees, the best split across all variables is used to divide each node, whereas in random forests, the best split among a subset of predictors that was arbitrarily selected at that node is used to divide each node. It has been demonstrated that this method outperforms other classifications, such as discriminant analysis, support vector machines, and neural networks, and that it is resistant to over fitting (Liaw et al., 2002).

In the next article authors proposed a deep-learning method using a traffic accident's severity prediction-convolutional neural network (*TASP-CNN*) model for traffic accident severity prediction. The suggested method effectively funds the latent traffic accident's severity feature representation, such as the feature combination and deeper feature correlations from traffic accident's data, in contrast to the prior methods that only take into account the shallow structure of traffic accidents. Using data from traffic accidents that occurred over an eight-year span (2009–2016) provided by Leeds City Council, the performance of the proposed TASP–CNN model was assessed. With deep learning techniques, the authors hope to overcome the drawbacks of conventional approaches by autonomously learning high-level features from unprocessed input data. The dataset used for the researcher's experiments is described by the authors and contains details about the weather, the type of road, and the number of vehicles that were engaged in the collision. Next, they describe the structure of their suggested CNN model, which is made up of a number of convolutional and pooling layers followed by fully linked layers. Using a variety of measures, including accuracy and F1 score, the model was assessed after being trained on the dataset. The experiment's findings demonstrated that the proposed CNN model worked better than conventional approaches for estimating the severity of traffic accidents. Its performance was compared to that of the NBC, KNN, LR, DT, GB, SVC,

Conv1D, NN, and LSTM–RNN models. The writers contend that their method can increase traffic safety by accurately predicting accident severity and enabling emergency services to react more quickly and effectively (M. Zheng et al. 2019).

Zong et al. (2013) compares the performance of Bayesian network models and regression models in predicting the severity of traffic accidents. The authors note that accurate prediction of accident severity is critical for emergency services to respond effectively and reduce the number of fatalities and injuries. The dataset used in the author's experiments, which contains variables like the type of road, the type of weather, and the speed of the vehicle, is described by the authors. After that, they go over how to build and train Bayesian networks and regression models, as well as how to assess their success using a variety of metrics like accuracy and error rate. The experiment's findings indicate that both the Bayesian network and regression models can be useful for foretelling the severity of traffic accidents, but that the Bayesian network model performs better in terms of precision than the regression model. According to the authors, the Bayesian network model can be a helpful tool for determining accident severity and enhancing traffic safety.

Recurrent neural network is another machine learning model used for prediction purpose. That's why Sameen and Pradhan (2017) developed a machine-learning model using a recurrent neural network (RNN) that can predict the severity of traffic accident based on 1,130 accident records that have occurred on the North-South Expressway (NSE), Malaysia from 2009 to 2015. The multilayer perceptron (MLP) and Bayesian logistic regression (BLR) models were contrasted with the suggested RNN model to better understand its benefits and drawbacks. The research used a grid search to find the ideal network architecture, which consisted of a Long-Short Term Memory (LSTM) layer, two fully-connected layers, and a Softmax layer. The RNN model was chosen because of its efficiency in processing sequential data. Using a stochastic gradient descent algorithm and dropout approach, the model was trained. The RNN model was compared to Bayesian Logistic Regression and Multilayer Perceptron models, and a sensitivity analysis was also carried out. In comparison to the MLP and BLR models, the RNN model outperformed them, with a validation accuracy of 71.77%, according to the research. The findings of the comparison studies demonstrated that the RNN model performed better than the MLP and BLR models. The RNN model had a validation accuracy of 71.77%, while the MLP and BLR models only managed 65.48% and 58.30%, respectively. The RNN model can be a useful tool for estimating the severity of traffic accidents, according to the study's findings, when deep learning frameworks are used.

F. N. Ogwueleka et al. (2014) discussed the topic of road traffic accidents (RTAs) and their effects on human life, with a focus on emerging nations where the rate of RTAs is rising as a result of things like inefficient drivers and bad road conditions. The authors use Nigeria as an example to demonstrate how an Artificial Neural Network (ANN) model can be used to evaluate and forecast accident rates in developing nations. In addition to highlighting ANN's capacity to model complex, nonlinear relationships without making any assumptions beforehand, the research outlines the benefits of using ANN over conventional programming for solving complex and non-algorithmic problems. The paper is divided into sections that each address a case study, related research, and data analysis. The authors conclude that the ANN model they developed performs better than other statistical methods for predicting accident rates.

Similarly, Chen et al. (2020) proposed machine learning techniques to forecast the risk of traffic accidents. According to the authors, Data collection and selection, preprocessing, and the application of mining algorithms are the three stages in the procedure. The study makes use of information from the Portuguese National Guard database as well as other

openly accessible databases to find trends related to the frequency of accidents. According to the study's findings, there are the most accidents happened between the hours of 17:00 and 20:00, and they found that rain is the meteorological element that has the biggest impact on the likelihood of accidents. Additionally, it demonstrates that compared to other days of the week, Friday is the day with the highest number of incidents. The results of the research can be used to help those in charge of making decisions about how to allocate resources for traffic surveillance in the most efficient way. By combining information from the past with predictions for the future, the research creates a novel method for estimating the number of accidents that will probably happen in the future. This method helps to identify locations where there will be a more risk of traffic accidents happening.

Since different variables can have varying degrees of impact on traffic accidents, the severity of accidents can be a key indicator of the damage they cause. Yang, j et al. (2023) uses data from vehicle traffic accidents from the Chinese National Automobile Accident In-Depth Investigation System to address the prediction of traffic accident severity. The author's main goals were to develop new methods for predicting the seriousness of traffic incidents and to identify the key variables that have a significant impact on their seriousness. Using the random forest algorithm, the authors of the paper ranked the significance of 12 accident features, including engine size, engine age, vehicle age, month of the year, day of the week, age range of drivers, vehicle maneuver, speed limit, accident location, accident form, road information, and collision speed. Accident location, accident form, road information, and collision speed were also added as additional accident characteristics that were not included in the significance ranking. The goal of the article was to develop a prediction model of traffic accident severity with better accuracy by comparing various algorithms and optimizing the outcomes. The writers draw the conclusion that as society has developed quickly recently and as there have been more cars on the road, there have also been more traffic accidents, causing significant economic and human losses. As a result, traffic science and intelligent vehicle study are currently focused on preventing traffic accidents and determining their severity. Authors shows that random forest algorithm is best to predict the severity of traffic accidents based on its high performance in data classification compared with back propagation (BP) neural network, Support vector machines (SVM) and Radial Basis Function (RBF) Neural Network.

Similarly, Al-Mistarehi et al. (2022) uses ML techniques to analyze the factors that significantly affect each level of crash severity, distribute hot spots, identify the causes and conditions of crashes, forecast the risk factors that affect these levels, and determine how these factors affect pedestrian safety. In light of the factors such as highway, vehicle, and environment, the findings demonstrated that the random forest model was the most appropriate technique to predict minor, medium, and severe injuries. There were significant factors that contributed to various injuries and fatalities, such as the type of crash (collision), the road's characteristics (flat straight), its type (flexible pavement), its surface (dry), its lane configuration (two ways with median), the weather (clear), the vehicle category (small passenger car), the driver error (failing to take the necessary safety precautions while driving), the time of day (Thursday), and the range of driver age (18–36 years).

Yan and Shen (2022) suggests a hybrid model, known as BO-RF, that combines Bayesian optimization (BO) with random forest (RF) to forecast the severity of traffic accidents on urban roadways. The fundamental predictive model is RF, and the parameters of RF are adjusted using BO to enhance the model's performance. The suggested model offers more accurate results than traditional algorithms and outcomes that may be understood through

relative importance and a partial dependence plot. The partial dependence plot aids in examining how various influences on traffic accident severity affect each other, while the relative importance enables for the identification of significant influencing elements for the severity of traffic accidents. The study emphasizes the significance of traffic accident severity prediction for managing and controlling traffic safety. Traffic accidents are a serious menace that have caused severe human suffering and significant financial losses. Emergency responders can estimate probable effects and quickly put accident management plans into place with the help of the forecast of traffic accident severity.

Statistical models and artificial intelligence (AI) models both are discussed in this article which are the two main categories of models for determining the severity of traffic accidents. Statistical models strictly assume the explanatory and response variables, but AI models are devoid of assumptions and are capable of managing complicated nonlinear relationships. However, the bulk of AI models are opaque, and the results they produce are confusing. RF, on the other hand, is an ensemble model based on decision trees that provides relative importance and partial dependence plots, making the results easy to comprehend. The performance of RF is, however, significantly influenced by hyperparameter choices. In order to enhance RF performance, the study uses BO to determine the best parameter values. BO is a useful technique for selecting high-quality parameters for problems involving machine learning and is suitable for the optimization of objective functions that are characterized by either the absence of analytical expressions or the high cost of evaluation.

Dias et al. (2023) demonstrates a method for estimating the likelihood of traffic accidents. The program employs data mining techniques and algorithms to extract knowledge from accident-related data housed in the Portuguese National Guard database as well as from other databases that are accessible to the general public, such as meteorological data sources and the annual calendar. Three components make up the system: pre-processing, the usage of mining techniques, and the selection and gathering of data. According to the research, accidents are most common between the hours of 17:00 and 20:00, and rain is the weather condition that has the biggest impact on the likelihood that an accident will occur. They also came to the conclusion that accidents happen more frequently on Fridays than any other day of the week. The study is unique since it attempts to forecast the amount of incidents that will probably happen in the future. A neural network was used to get the best outcome, with several models being created for each collection. The findings have consequences for those making decisions about how best to allocate resources for traffic surveillance.

Going further, Ali et al. (2017) used data for single-vehicle accidents in Chicago from 2004 to 2012 to apply a random parameters logit model (with heterogeneity in means) and investigate the effect(s) of passengers on driver-injury severity levels. Three separate sub-populations were taken into account in the analysis, including vehicles with one occupant (the driver), vehicles with two occupants (the driver and a passenger), and vehicles with three occupants, with potential driver-injury severity outcomes of no injury, minor injury, and serious injury (driver and two passengers). Additionally, the analysis took into account a broad range of additional potential variables that might influence the severity of driver-injury cases, including driver's characteristics, weather, vehicle, and roadway characteristics. Statistically significant differences between subpopulations in the presence or lack of passenger data are revealed by the model estimation results. Result shows that, random parameter and significant heterogeneity in means for two-occupant vehicles , having a younger passenger ride with a younger driver (both less than 25 years old) has highly heterogeneous effects (with higher likelihoods of both no injuries and severe injuries), particularly if the driver is male. Authors also discovered that in the two-

occupant vehicle case, having a younger driver (less than 25 years old) and a peer passenger (15–25 years old) together increased the likelihood of severe injury, highlighting a higher baseline risk with a younger driver and a younger passenger. The findings of our three-occupant model, however, indicate that the likelihood of a serious driver injury is significantly reduced when two of the passengers are younger than 15 years old. They discover incredibly complicated relationships when they take the impact of passenger/driver gender interactions into account. However, a few things stick out very clearly. A female motorist traveling with two passengers and having only female passengers generally results in more serious driver injuries. Comparatively to having one male and one female passenger, having all male passengers increased the risk of both no injury and severe injury for male drivers with numerous passengers, while having all female passengers decreased the risk of both.

In further studies on predicting injury severity level, a number of prediction models, including NN, SVM, and Decision Tree, have been widely used (DT). Iranitalab and Khattak (2017), describes a study that used statistical and machine learning techniques to predict crash severity. The goal of the study was to compare the performance of four prediction techniques, Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM), and Random Forests (RF)—as well as look into how K-means Clustering (KC) and Latent Class Clustering (LCC) for data clustering affected the performance of crash severity prediction models. The research extracted two-vehicle crashes as the analysis data from reported incident data from Nebraska, United States, from 2012 to 2015. Training/estimation (2012–2014) and validation (2015) sections of the dataset were created. The training/estimation dataset was used to train/estimate the four prediction methods, and the validation dataset was used to determine the correct prediction rates for each crash severity level, the overall correct prediction rate, and a suggested crash costs-based accuracy measure. According to the research, MNL was the least effective method, and NNC had the best prediction performance in both general and more serious crashes, followed by RF and SVM. MNL, NNC, and RF all performed better due to KC, and MNL and RF also did better due to LCC, but NNC perform worse because of LCC.

The research also created a method for comparing crash severity prediction techniques based on crash costs. According to factors like crash severity, the method takes into account the costs of crashes imposed on a community, the severity of injuries sustained by crash victims, and the costs of potential crashes in which insurance companies may be involved. The application of the prediction in reality determines how the final comparison results should be interpreted. While a hospital or emergency department would prefer a model with the lowest SPE value, safety managers who need to forecast yearly crash costs would prefer the combination of NNC and KC as the method with the lowest value.

Aldhari et al. (2022) talked on Saudi Arabia's issue of road safety, which has been cited as a key area of emphasis for the nation's Vision 2030 objectives. In order to build machine learning-based models to forecast the severity of accidents, the study analyzed data from the Qassim Province, which has one of the highest rates of traffic accidents in the nation. A resampling strategy was utilized to solve the problem of data imbalance, and three classifiers were deployed, including two ensemble machine learning techniques. To rank the elements influencing the severity of accident injuries, the SHapley Additive exPlanations (SHAP) analysis was performed. The findings demonstrated that the XGBoost classifier beat the other classifiers, with accuracy, precision, recall, F1-scores, and an area curve for multi-category classifications of 71%, 70%, and 0.87, respectively. With an accuracy of 94% and a binary classification area curve of 0.98, the same classifier fared better than the competition. The results of the study showed that the kind of road

and lighting conditions were two of the most important factors influencing injury severity outcomes, while the temporal parameters, such as the month and day of the week, as well as the road type, were associated with severe injuries. The study's conclusions should help policymakers in the Qassim Region and other Saudi Arabian regions establish safety mitigation initiatives.

The effects of sample size on the multinomial logit, ordered probit, and mixed logit models of crash severity were discussed by Ye and Lord (2014). The connection between the severity of an accident and its contributing factors, such as the characteristics of the driver and the vehicle, the state of the road, and other elements of the road environment, is investigated using crash severity models. The research employs a Monte-Carlo analysis based on both simulated and observed data and examines samples with a range of 100–10,000 observations. According to the research, regardless of the approach taken, using small sample sizes has a significant impact on the creation of crash severity models. In addition, the sample size requirements for the ordered probit model and the multinomial logit model are intermediate, with the mixed logit model requiring the highest sample size and the ordered probit model requiring the lowest sample size. According to the study's findings, given the data's size and characteristics, the information could aid transit safety analysts in selecting the right model.

Three different machine learning models (logistic regression, decision tree, and random forest) are evaluated in terms of how well they perform at predicting the results of specific tasks. The study discovered that, in terms of accuracy and sensitivity, the random forest model outperformed the other two models, Chen M and Chen C (2020). As opposed to sensitivity, which assesses a model's capacity to correctly detect instances of success, accuracy refers to the model's capacity to predict an outcome of a task. (i.e., instances where the task outcome is positive). The accuracy rates for the decision tree and logistic regression models were 72.8% and 71.1%, respectively, while the random forest model had a success rate of 77.6%. In comparison to decision tree and logistic regression models, the paper contends that the random forest model may be a better option for tasks requiring precision and sensitivity. It's crucial to keep in mind that the effectiveness of these models might vary depending on the precise task and dataset employed, so it's always a good idea to try a few different models and evaluate their performance before deciding on one.

Iranitalab and Khattak (2017) compared the effectiveness of four methods—MNL, Nearest Neighbor Classification (NNC), SVM, and RF—in predicting the severity levels of accidents in a dataset that contains 68,448 two-vehicle crashes from 2012 to 2015 in Nebraska, the United States. The original dataset's response variable, or the severity of the crashes, had five categories, with fewer observations in the categories for catastrophic injury and death collisions. The authors aggregated the observations in these two categories and used four categories as the four classes of the dependent variable in order to handle the imbalanced data. Each machine learning model also included the implementation of two clustering techniques, K-mean clustering (KC) and Latent class clustering (LCC), to address the occurrence of 19 unobserved heterogeneity in the dataset. MNL outperformed the other three models with an overall accuracy of 64.17%, SVM with a score of 61.52%, RF with a score of 59.43%, and NNC with a score of 54.74%. The approaches used for clustering did not increase accuracy in general. The authors also suggested a strategy based on crash costs to look into the models' overall prediction cost error (OPE). They discovered that, despite the fact that clustering had no effect on the machine learning models' accuracy, KC and LCC enhanced the OPE outcomes for MNL, NNC, and RF. The best OPE, 26.05%, was produced using NNC with KC clustering.

In 3,185 rollover crashes that occurred in New Mexico between 2010 and 2011, Chen et al. (2016) looked into trends of driver injury severity. Machine learning classification and regression trees (CART) was used to determine the important contributing. To assess the effectiveness in predicting the severity, factors and SVM were utilized. They divided the original five distinct severity levels into three 20-category subsets to address the impact of unbalanced data. The results of CART showed that driver seatbelt use was the most important factor influencing the outcome of injury severity in rollover collisions out of a total of 22 predictor variables in the dataset. The quality of the lighting and the slope of the road were judged to be unimportant. 18 relevant variables were later used as inputs for an SVM learning method. With an accuracy of 58.77% for the non-injury category and 50.46% for the non-incapacitating injury category, the SVM model fared best in these categories. With an accuracy of 22.67%, the model did the worst for incapacitating injury and mortality categories.

In a study, Shi, X. et al. (2019) provides a structure for analyzing the behavior of drivers when operating motor vehicles and forecasting the degrees of risk. Feature extraction and selection, unsupervised risk rating, and imbalanced data resampling are all incorporated into the system. More than 1,300 driving behavior features are retrieved from the trajectory data for each vehicle. These driving behavior features provide in-depth and multi-view measurements on each behavior. Estimating the possible dangers posed by automobiles on the road is done through the use of unsupervised data labeling. In order to mitigate the disparity between the risky and safe classes, vehicles are categorized into a number of groups, each of which is given a risk rating, and under-sampling of the data pertaining to the safe group is carried out. Key characteristics are chosen based on feature importance ranking and recursive removal, and XGBoost is utilized to construct links between behavior features and the relevant risk levels. The levels of danger posed by automobiles when driving can be estimated based on a selection of their essential attributes. An overall accuracy of 89% is reached for behavior-based risk prediction by using NGSIM trajectory data as a case study. Fuzzy C-means is used to cluster four risk categories, 64 essential behavior traits are selected, and NGSIM trajectory data are used as the basis for the case study. This method delivers an accurate forecast of danger levels and is helpful in identifying relevant elements for driving assessments.

In order to deal with the unbalanced data, two sampling techniques can be used, the synthetic minority over-sampling technique (SMOTE) and randomize class balancing (RCB). Zhang, J. et al. (2018) compare the accuracy of several statistical and machine learning models in predicting collision injury severity. The models were created utilizing crash data gathered at motorway diverging locations to forecast the severity of each crash's associated injuries. According to the study, machine learning models were generally more accurate at predicting collision injury severity than statistical models. Particularly, it was discovered that the RF and KNN models, with total forecasting accuracy of 53.9% and 52.9%, respectively, were the most accurate. These results were in line with earlier studies. The linear structure of utility functions and the distribution assumptions of error terms, which may not always hold for accident severity data, were blamed for the statistical models' poorer performance. On the other hand, no presumptions regarding the characteristics of the data's distribution or the relationship between the dependent and independent variables were necessary for machine learning models. As a result, they were able to learn functional forms from the training data and make data-driven predictions. Sensitivity analysis was also used in the study to determine the significance of explanatory factors on crash severity. The four machine learning models demonstrated that they have taken the sequence of crash seriousness into account. In the dataset, the link between injury severities and explanatory variables was better captured by the MNL model than by the OP model. The study evaluated the variable importance

of statistical and machine learning models and discovered that, occasionally, they calculated the variable importance on explanatory factors substantially differently. This was likely caused by the diverse procedures used by various machine learning techniques to explore the built-in characteristics of the data, which produced varying estimates of parameter relevance. The study came to the conclusion that when using machine learning techniques for the inference of variable importance on accident injury severity, one should exercise caution. The evaluation of a variable's relevance may not always be more accurate if prediction accuracy is higher. In order to better comprehend the variations between machine learning models, the study intends to carry out a thorough evaluation of sensitivity analysis of variable relevance.

In a different paper, writers discuss a study that used artificial intelligence to forecast the severity of traffic accidents. Four different models were created by the authors using four different types of input: feed-forward neural networks (FNN), support vector machines (SVM), fuzzy C-means clustering-based feed-forward neural network (FNN-FCM), and fuzzy c-means-based SVM. (SVM-FCM). The study was based on the Great Britain accident database from 2011 to 2016. In terms of accuracy and F1 score, the study found that the SVM-FCM model outperformed the other models in predicting the severity level of severe and non-severe collisions. The ability of the FNN and SVM models to forecast was reportedly enhanced by the FCM clustering method (Assi, K. et al. 2020).

Mansoor, U. et al. (2020), Create machine learning models that accurately predict the severity of traffic collisions using information that is easily accessible and quick to gather from the scene of the collision, such as the local speed limit, the type of intersection control, the weather, and the types of vehicles involved. In the study, various machine learning models' abilities to forecast the severity of traffic crashes were compared. KNN, DT, AdaBoost, SVM, FNN, and a two-layer ensemble model were among the models used. In order to develop and evaluate these models, the researcher’s analyzed data on road traffic collisions collected over a six-year period, from 2011 to 2016, from the British Department of Transport. The dataset was divided into training and testing groups at random using a 7:3 ratio. Accuracy, precision, recall, and F1 score for each model were assessed in order to better understand how well it worked. In terms of accuracy and F1 score, AdaBoost surpassed all other individual models in the study's results while not overfitting. The least trustworthy model was KNN, which had the lowest F1 ratings for both severity levels. The ability to forecast accident severity levels, however, was greatly enhanced by the proposed two-layer ensemble model. With testing accuracy of 76.7% and F1 scores of 0.75 and 0.77 for severe and non-severe crashes, respectively, the two-layer ensemble model outperformed all basic models. Accuracy was greatly increased for both training and testing. The researchers then assessed the transferability of these models using a crash dataset they downloaded from the online National Collision Database (NCDB) in Canada. The models performed similarly to how they had done with Great Britain's dataset, with an accuracy of 79.3% and F1 scores of 0.78 for fatal collisions and 0.80 for non-fatal crashes, respectively. All other models were inferior to the two-layer ensemble model in performance. The results of the study imply that if similar models are generally extended to other accident datasets, high accuracy of crash severity prediction can be anticipated.

2.2 Weather and its effects on road accidents

Between 2005 and 2007, researchers conducted the National Motor Vehicle Crash Causation Survey (NMVCCS), which sought to gather data directly from the scene of light vehicle crashes. The study covered 5,470 collisions over a 2.5-year span, which is typical of 2,189,000 collisions across the country. 3,945,000 drivers, 4,031,000 vehicles, and 1,982,000 passengers are thought to have been involved in these collisions.

In 94% of collisions, the driver was blamed for the crash, 2% of crashes involved a vehicle component failure or degradation, and 2% of crashes involved the environment (slippery roads, bad weather, etc.). The "critical reason" for the crash, or the final event in the causal chain, was ascribed to the driver in all three cases. Recognition mistakes (41% of crashes), judgment errors (33% of crashes), and performance errors (11% of crashes) were the "critical reason" categories that occurred most frequently when drivers were involved.

The most prevalent "critical reason" for crashes involving automobiles was tire issues (35% of crashes), followed by brake issues (22% of crashes). Slick roads were the "critical reason" for the majority of environment-related crashes (50%) and were followed by glare (17% of environment-related crashes). The NMVCCS data has limitations, which should be noted. For example, it only includes collisions that take place between 6 a.m. Furthermore, the identification of a "critical reason" does not place the responsibility for the accident on the shoulders of the driver, the vehicle, or the surroundings (Singh, S. 2015).

Data gathered in US department of transportation under road weather management program from 2007 to 2016 reveals that each year, roughly 5,891,000 vehicle crashes occur. Of these, about 21% (or around 1,235,000) are caused by adverse weather conditions or hazardous road surfaces. These harsh conditions result in the loss of approximately 5,000 lives and injuries to over 418,000 individuals every year. A deeper dive into the data indicates that most weather-related accidents (70%) happen on wet roads and nearly half (46%) take place during rainfall. Conversely, winter conditions cause fewer accidents: snow or sleet (18%), icy pavement (13%), and snowy or slushy pavement (16%). Fog is the least contributing factor, causing only 3% of such incidents. Furthermore, adverse weather or slick roads are a factor in 15% of deadly crashes, 19% of crashes involving injury, and 22% of crashes that only cause property damage. In numbers, these conditions lead to almost 4,900 fatal crashes, over 301,100 crashes causing injury, and nearly 919,700 crashes with property damage alone every year.

Using a negative binomial model and a log-change model, Zou et al. (2021) investigates how the influence of a variety of factors contributes to fatal car accidents in the states of California and Arizona. Indicators of social development and climate as well as the frequency of fatal traffic accidents are included in these categories. Both models accurately fit the data, with climate variables (average temperature, precipitation) and non-climate variables (beer consumption, rural vehicle miles travelled ratio, and vehicle performance) strongly increasing the incidence of fatal traffic accidents. The authors investigate that, the number of automobile collisions will rise by 4.0% in the state of California and by 3.6% in the state of Arizona if the annual mean temperature rises by 1 degree Fahrenheit. The number of people who lose their lives in car accidents is expected to rise by 4.8% in California and 4.6% in Arizona if there is a 0.5 standard deviation increase in the amount of precipitation that falls over 24 months. Hail, wind, and other forms of inclement weather have less of an impact on the number of accidents that occur

on the roads.

Moreover, a rise in the number of people killed in car accidents has been linked to a number of factors in the state of California, including a rise in the consumption of beer, a fall in the ratio of miles traveled by rural vehicles to total miles traveled, increases in temperature and precipitation, and hailstorms. There was a correlation between increased vehicle performance and a decrease in the frequency of these accidents. In the state of Arizona, a greater GDP, median income, gasoline prices, highway capital spending, temperature, precipitation, and wind conditions led to an increase in the number of fatal accidents. On the other hand, increased vehicle performance was associated to a drop in the number of these incidents.

Bergel-Hayat et al. (2013) focused on examining the connection between different types of weather and the likelihood of accident. The research looked at data from France, Netherlands, and Athens, applying time series analytic methods to the information. The results showed that there were strong relationships between meteorological variables and the number of road injury accidents.

According to the findings of the researchers, rainfall not only had a direct impact on the accident rates on highways, but also had an indirect impact on the accident rates on main roads due to changes in exposure, as was seen in France. In all of the cases that were analyzed, a positive correlation was found between temperature and the number of accidents; however, the importance of this correlation varied according to the time of year and the location. On the other hand, in Athens, an inverse association was found between the amount of rainfall and the number of accidents. This suggests that rainfall contributed to a reduction in the number of traffic accidents, most likely as a result of a decreased volume of traffic. It was discovered that extreme weather conditions, in particular extremely low temperatures and heavy amounts of precipitation, had a major impact on accident rates, notably in the city of Athens.

The findings of the study indicate that additional research is required to fully comprehend the interplay between climate and vehicular traffic as it relates to the incidence and severity of accidents. In addition to this, it highlighted the potential advantages of combining average and extreme weather data into accident data models for the purpose of gaining a more in-depth comprehension of the ways in which the weather influences accident risk. The data can assist influence initiatives for prevention such as awareness campaigns, infrastructure improvements, and local warning systems. Additionally, they can help facilitate the examination of the impact that weather conditions have on road safety on a national level.

The purpose of this study conducted by Finnish Meteorological Institute (2012) is to investigate the relationship between inclement weather, specifically snowfall and cold temperatures, and the incidence of car accidents in the county of Kymenlaakso, which is located in the southern part of Finland. It analyzes data collected throughout the winters beginning in 2002/2003 and continuing through 2007/2008 in an effort to determine whether or not there is a connection between particular weather patterns and the number of automobile accidents. An examination of the data gathered on the weather revealed that the incidence of accidents tended to rise if the temperature dropped to 0 degrees Celsius or below and whenever snowfall was observed. To be more specific, a snowfall of more than 10 centimeters virtually doubled the daily average number of accidents, which is a strong sign of the dangers posed by winter circumstances. Furthermore, an assessment of days with heavy snowfall, defined as roughly 5 centimeters or more, often

found increased accident rates. This was the case despite significant day-to-day variability in the accident rates.

The researchers also brought up the difficulty of precisely predicting accident risks due to the many other factors that are at play in addition to the meteorological conditions. These include things like the road conditions, the behavior of drivers, the state of vehicles, and the local traffic regulations, all of which can have a substantial impact on the chance of accidents. Therefore, despite the fact that meteorological conditions can be a significant cause to traffic accidents, they are merely one aspect of the greater collection of elements that affect road safety. According to the findings of the study, traffic safety management systems that are not only adaptable and flexible but also take into account information about the current weather as well as any other relevant elements are extremely important. These findings could help improve traffic safety tactics, particularly in locations with climatic circumstances that are comparable to those in Kymenlaakso County. As a result, there is a possibility that the number of road traffic accidents that occur when weather conditions are unfavorable could be reduced.

The research conducted by Islam et al. (2022) aimed to understand the effect of changing weather on road accidents leading to death in Saudi Arabia, a region experiencing a high frequency of climatic events and road accident-related fatalities. The study made use of annual data from 13 regions within the country, spanning from 2003 to 2013. The investigation discovered that factors impacting accidents during this time included temperature, rainfall, sandstorms, and the quantity of vehicles. According to the study, traffic accidents occur four times more frequently in urban areas than in rural areas, and the average number of injuries (2832) is four times higher than the average number of fatalities (696). Drivers and passengers of motor vehicles were most frequently killed, then pedestrians, motorcyclists, and cyclists. Climate factors like temperature, precipitation, and sandstorms have been found to be dangerous in urban environments.

The findings of the regression analysis showed that the overall number of accidents was significantly positively impacted by average temperature, rainfall, sandstorms, and the number of cars. Only inside-city accidents considerably caused deaths, although both inside- and outside-city accidents significantly caused injuries. It's interesting to note that only motor vehicle accidents were shown to have a statistically significant accident death rate. According to the report, Saudi Arabia's roads are at risk from climate change because it would raise temperatures, increase the frequency of sandstorms, and disrupt rainfall patterns. Possible adaptation strategies, such as warning signs, road improvements, safety campaigns, and raised public knowledge, were suggested to lessen the negative effects of climate change on road safety. The study also stressed the need for enhanced cycling safety measures, better road infrastructure, and efficient traffic regulations in relation to climate-related extreme weather events.

Zeng et al. (2020) uses a Bayesian spatial generalized ordered logit model to examine the effect of current weather conditions on the severity of motorway crashes. The model takes into account elements that have been observed, including wind speed, air temperature, precipitation, visibility, humidity, and other variables. The Kaiyang Freeway in China's 2014 and 2015 crash data are examined. The suggested model takes into account the geographical correlation between nearby crashes and the ordered pattern of crash severity levels. In comparison to a generalized ordered logit model, it exhibits a strong geographical correlation, a better model fit, and more accurate estimation results.

The findings show that more precipitation decreases the likelihood of light and severe crashes while increasing the likelihood of medium crashes. The severity of crashes is also greatly influenced by a number of other variables, including roadway parameters (horizontal curvature and vertical grade), driver characteristics, vehicle attributes, car registered province, collision time, crash type, and emergency medical service response time. In addition to the strategies already in place, engineering countermeasures are recommended to lessen the severity of crashes on wet days. According to the results of estimate and evaluation, the spatial generalized ordered logit model performs better overall than the conventional model in terms of goodness-of-fit. The threshold between medium and severe crashes and the latent severity propensity are both significantly influenced favorably by precipitation. Reduced sliding resistance and more mental effort required of drivers can be linked to a lower chance of light collisions. The reduced likelihood of serious crashes in heavy rain is explained by the risk compensation theory (Zeng et al. 2020).

Khodadadi-Hassankiadeh et al. (2020) conducted a study from 2014 to 2018 to ascertain accident trends in foggy conditions and the connection between the driver, the road, and accident severity in Guilan, Iran. With the use of STATA software and time-series estimators for multivariate regression analysis, it applied a retrospective descriptive-analytical methodology. The findings revealed that compared to other meteorological conditions, foggy situations had a significantly higher death rate from traffic accidents. According to the study, the likelihood of a fatal accident in fog increases with female drivers but decreases with age.

Moreover, the months of December, February, and November saw the greatest number of accidents brought on by fog. Interestingly, the rate of fatalities was much lower at specific times, including 2 AM, 9 AM, 11 AM, 13 PM, and 19 PM, and the number of connected injuries dramatically decreased at 2 AM, 3 AM, and 2 PM. Injury rates were much higher in the cities of Rasht and Anzali. The study indicated that accidents in foggy conditions caused significant damage and injuries in most cities, even if the injury rate was noticeably greater in Rasht and Anzali. It's possible that the installation of fog lights and other road amenities contributed to the association between the rate of fatal accidents and specific distances on particular highways. The study also showed a substantial age-related decline in the death rate from road traffic accidents (RTAs), suggesting an age-specific pattern in accident rates.

Sangkharat, K. et al. (2021) investigate the effect of rainfall on road accidents in Thailand from 2012 to 2018 using emergency data from the National Institute for Emergency Medicine (NIEM. The data were analyzed using a generalized linear model (GLM) and a time-series approach. The results were reported using relative risk (RR) at 95% confidence intervals compared with dry days, with the study controlling for long-term trends, seasonality, days of the week, public holidays, and other meteorological characteristics.

Their findings suggest that high rainfall levels were found to significantly increase the number of traffic accidents in both Thailand's northern and southern regions, with the southern provinces having a larger projected risk than the northern provinces. Surprisingly, however, really heavy rain (more than 20 mm/day) showed a decrease in risk. With an RR of 1.052 for the Northern provinces and 1.062 for the Southern provinces, rainfall amounts between 10 and 20 mm per day demonstrated the highest predicted risk, which was a significant discovery.

The study also looked at other climatic factors, such as temperature and relative humidity, and found that the Southern provinces had greater average temperatures, precipitation, and relative humidity than the Northern provinces. When examining the frequency of traffic accidents, it was found that weekdays experienced more accidents than weekends, with the highest incidences occurring from October to December. The study suggests that in order to solve this problem and enhance the effectiveness of service, ambulance forecast models and warning systems should incorporate rainfall.

Basagaña et al. (2015) conducted a research is to investigate the effect that high ambient temperatures have on automobile collisions, with a particular emphasis on collisions that involve elements linked to driver performance, such as distractions, driver mistake, exhaustion, or sleepiness. Research conducted in Catalonia (Spain) over the warm period from 2000 to 2011 indicated that there was a considerable rise in the crash risk during heatwave days and with each 1°C increase in maximum temperature. This was discovered through the use of a time-series analysis for motor vehicle accidents. The likelihood of collisions increased by 2.9% during heat waves and by 1.1% for every 1°C increase in temperature. The link was greater (7.7%) for collisions that had driver performance difficulties. The findings shed light on how important it is for safety measures involving roads to take into account the prevailing weather, particularly in light of the effects of climate change.

2.3 Statistical techniques

To examine the factors that influence injury severity, regression models are frequently utilized. For example, Khattak et al. (2002) employed the ordered probit modeling technique to examine possible characteristics that affect the severity of injuries suffered by elderly drivers (65 years of age and older) engaged in traffic crashes that occurred in Iowa, United States, between 1990 and 1999.

2.3.1 Logistic regression

Al-Ghamdi, A.S. (2002) discussed the variables that affect how serious car accidents are in Riyadh. In order to assemble accident-related data from traffic police records, logistic regression analysis was used to determine the most significant factors associated with accident severity. The purpose of this study was to look into the elements that affect how serious car accidents are in Riyadh. The study employed logistic regression analysis to identify the most important factors connected to accident severity using accident-related data gathered from traffic police records. According to the study, the two criteria that had the greatest impact on accident severity were the event's location and its underlying cause. The model demonstrated that the probability of a fatal accident in a non-intersection accident was increased by stratifying location-related data into two classes. According to the findings, organizations should concentrate their efforts on traffic accident sites other than intersections in order to make safety improvements more cost-effectively. The research also implies that measures to decrease serious accidents can be prioritized using the chances described in the publication. The likelihood of being involved in a fatal accident at a site other than an intersection as a result of a wrong-way violation is relatively larger than that of any other violation, hence drivers should be cautioned about the potential lethality of wrong-way violations in a particular awareness campaign, for example. The study's conclusions indicate that logistic regression holds promise for producing informative interpretations that may be used to inform upcoming safety upgrades in Riyadh. The study does note that when calculating the possibilities described in the report, no consideration was made for traffic exposure or statistics that were

unavailable in Riyadh or difficult to collect. When such data is made available, the conclusions may serve as a guide for further investigation.

2.3.2 Correlation

In recent years, most of the researcher used statically analyzed techniques to anticipate or prediction. To keep this in mind Zhang, Z. et al. (2015) highlights the difficulties of evaluating multivariate data that has a large number of variables and the significance of correlation analysis in determining links between these variables. The authors present a method for arranging variables into a 2D layout that encodes their pairwise correlations and may be applied for interactive axes sorting in parallel coordinate displays. They then refine this method even further into a correlation map that employs geographical closeness to transmit correlations, making it simpler to understand and manage interactions between variables. The authors also discuss the need for efficient visual interfaces that enable analysts to quickly gain an overview of the overall correlation relationships in the data and easily manipulate the data to reveal hidden relationships via various modes of interactions, including filtering, selection, bracketing, and clustering. They contrast correlation analysis to regression analysis and point out that neither can prove cause-and-effect connections between the variables. The authors also offer novel mechanisms that handle categorical and numerical variables in a unified framework, scalability for large numbers of variables via a multi-scale semantic zooming approach, and visualization of data relations within the sub-spaces spanned by correlated variables by projecting the data into a corresponding tessellation of the map.

There are a lot of factors that can cause the accident. Rodionova et al. (2022) Discusses a study carried out in Saint Petersburg, Russia, that sought to determine the variables influencing the severity of auto accidents that happened there between 2015 and 2021. The study looked at a number of variables, including as lighting and weather conditions, road infrastructure, human factors, accident kinds, and vehicle attributes like category and color, that could have an impact on crash severity. The objective was to comprehend the key causes of traffic accidents and to provide knowledge that might be applied to the creation of efficient road safety regulations. The essay emphasizes that because it directly affects both population expansion and economic growth, road safety is a critical component of sustainable development. In the past, a lot of research has been done with the goal of lowering fatal accidents and catastrophic injuries. However, because the distribution of severity levels vary, with slight severity being the most frequent and fatality being the least, it might be challenging to anticipate fatal results. The Russian government has started the "Safe Quality Roads" nationwide initiative for 2019–2030 with the goal of reducing the number of traffic fatalities per 100,000 people from 13 in 2017 to 8.4 by 2024 and 4 by 2030. However, attaining this goal necessitates a thorough comprehension of the most hazardous drawbacks of the current road system. The study investigated 37,585 observations of accidents that occurred in Saint Petersburg between 2015 and 2021. The factors that affect crash severity were found by the researchers using ordered probit regressions to evaluate the data. The study discovered that a variety of elements, including illumination, weather, road infrastructure, human factors, vehicle type, and car color, have an impact on crash severity. However, run-off-road incidents were discovered to be the most significant accident category, contributing to an 11.2% rise in accidents that result in fatalities. It was discovered that deficiencies in road infrastructure, such as inadequate lighting and road barriers, significantly increased fatal outcomes by 12.6% and 2.8%, respectively.

2.3.3 Tobit regression

Anastasopoulos et al. (2012), highlights the use of tobit regression as a statistical method to examine variables that affect the frequency of automobile accidents on particular stretches of road. Since the accident rates on these segments are continuous data with a zero censor, not all road segments may have seen accidents during the data collection period. The underreporting of less significant collisions, which would lessen the possibility that they would appear in crash databases, is one of several probable causes of this censorship. Regardless of the severity of the injury, traditional tobit-regression investigations have focused on the overall accident rate. The issue of censoring depending on the severity of crashes is not addressed by this strategy, either. The paper recommends a tobit-regression technique to account for accident rates by injury severity level, such as the rate of no-injury, probable injury, and injury accidents per distance traveled. The study uses five years' worth of data from Washington State roadways to estimate a multivariate tobit model of accident-injury-severity rates, which it then uses to analyze the likelihood of differential censoring across injury-severity levels. It also considers the possibility of contemporaneous error correlation caused by unobserved characteristics that are shared by many route segments. The empirical results show that the multivariate tobit model outperforms the univariate tobit model, is nearly identical to the multivariate negative binomial model, and has the potential to provide a more thorough understanding of the factors determining accident-injury-severity rates on specific roadway segments. The article also analyzes the drawbacks of conventional accident-frequency methodologies and emphasizes the possibility of tobit regression as a substitute strategy for examining accident-causing elements. The paper suggests that tobit regression be used in additional areas of transportation safety research in future studies.

2.3.4 Chi-square

The Chi-square test was employed in the study in Denizli, Turkey, to examine 1338 traffic incidents. It sought to explore the important elements influencing these mishaps and discovered that both the traits of the individuals and some environmental factors had a big impact. According to the research, people between the ages of 20 and 29 and 30-39 had the highest probability of being in a car accident. It's interesting to note that between the ages of 40 and 69, accidents happen less frequently. According to reports, men are more likely than women to be involved in accidents. The study discovered that accidents tended to rise toward the weekend, peaking on Saturdays when days and hours were taken into account. The period between 16:00 to 19:59, which corresponds to the end of the workday, was the most accident-prone. The accident rates for various car kinds also varied. Motorcycles and bicycles were the two vehicles most frequently engaged in collisions. Due to their widespread use in public transit, buses and minibuses played a crucial role in the event. Contrary to popular opinion, cars were associated with the fewest accidents. The level of schooling was also found to have a substantial impact on accident rates. People who had only completed their primary education were more likely to be in accidents. The frequency of accidents reduced as education levels rose, suggesting that education may play a preventive role in traffic accidents. According to the study's findings, knowing these elements could aid in the creation of strategies for predicting and even preventing accidents as well as serving as a roadmap for initiatives to improve traffic safety (Sari and Zeytinoğlu 2009).

2.4 Handling of Imbalanced Data

Many of the machine learning models do not works with imbalance data. That is why, Jeong, H. et al. (2018) balanced the data and conduct research that focus on accurately categorize the seriousness of injuries suffered in motor vehicle collisions. The Michigan Traffic Crash Facts (MTCF) dataset from 2016–2017 contained statistics on 297,113 automobile accidents. However, there was an imbalance in the distribution of the various accident severity classes in the MTCF dataset, as is typical with many collision datasets. In order to account for this, the researchers balanced the classes using a variety of methods, including under- and over-sampling. The researchers used five various classification learning models to categorize the levels of injury severity, including logistic regression, decision trees, neural networks, gradient boosting models, and Naive Bayes classifiers. Then, the researchers used Bootstrap aggregation (or bagging) and majority voting, two distinct training-testing techniques, to try and enhance the classification performance of these models. The researchers used the geometric mean (G-mean), a statistical measure that considers both the sensitivity and specificity of the model's forecasts, to assess the performance of these models. When bagging was combined with decision trees and the over-sampling method for imbalanced data, they discovered that the classification performance was at its best. Additionally, by combining under-sampling and bagging, the impact of remedies for the unbalanced data was enhanced. The authors also took into account two additional classification problems, one with two classes and the other with three classes, in addition to the MTCF dataset's initial five injury severity classes. This gave them the opportunity to research the effects of the number of groups on the effectiveness of classification models and to compare their findings to previous research.

A greater comprehension of the connection between crash risks-factors and the seriousness of injuries, according to the researchers, can improve driving safety, lower the number of fatal collisions, and lessen the economic toll of collisions. They also talked about the restrictions of using classification success rate as a performance indicator for models. They observed that datasets on accident severity are frequently unbalanced, with the non-fatal class usually containing disproportionately more data points than the fatal class. As a result, models with high accuracy rates may misclassify groups with greater severity. This was addressed by the researchers by using additional statistical measures to produce more insightful measurements, such as true positive, true negative, false positive, and false negative. To compare the overall effectiveness of various models, they finally used the geometric mean of sensitivity and specificity as a compact evaluation measure.

Data imbalance is a critical issue nowadays. To deal with the data imbalance issue Jeong, H. et al. (2018) uses under- and over-sampling methods to account for unbalanced classes. The aim of this study is to accurately and sensitively categorize the degree of injury in motor vehicle incidents. The Michigan Traffic Crash Facts (MTCF) dataset from 2016–2017 provided the 297,113 vehicle crashes used in the study.. To categorize the degrees of damage severity, five classification learning models—Naive Bayes classifier, Gradient Boosting Model, Decision Tree, Neural Network, and Logistic Regression—were employed. Bootstrap aggregation (also known as bagging) and majority voting are two training-testing techniques used in the study to try and enhance classification performance. When decision trees are used in conjunction with bagging and over-sampling is applied to imbalanced data, the classification performance is at its best. Under-sampling in conjunction with bagging increases the impact of treatments for unbalanced data. The study takes into account two additional classification issues, one

with two classes and the other with three classes, to examine the effect of the number of classes on the effectiveness of classification models. The article also analyzes the limitations of classification accuracy rate in assessing model performance and recommends integrating additional statistical measures, such as true positive, true negative, false positive, and false negative, to produce more insightful metrics. At the end of the day, it is possible to assess the overall performance of several models by using the geometric mean (or G-mean) of sensitivity and specificity. The G-mean has high values when both sensitivity and specificity are high and the gap between the two measures is minimal. It is calculated as the square root of the product of sensitivity and specificity. The study's advantages include recommending extra statistical measures to provide metrics that are more informative and increasing the categorization performance of injury severity in motor vehicle incidents. The study's limitations in terms of how well models function when measured by classification accuracy rate and the requirement to balance the dataset prior to training are its drawbacks.

2.5 Detecting the accidents

On the other hand, the use of sensors technology in accident detection systems can become a popular approach to reduce the number of accidents and minimize their impact. This literature review also focuses on the use of sensors in accident detection systems and presents an overview of the current state-of-the-art.

The accelerometer is one of the sensors that accident detection systems employ the most frequently. It can track variations in acceleration and can spot unexpected variations that might be signs of an accident. They are frequently employed in smartphones to detect screen rotation and have been proven to be successful in identifying car accidents. A system that employs a smartphone accelerometer to identify car accidents was suggested by the authors of a study by Dong et al. (2016) with a 92% accuracy rate and an 8% false positive rate, the system proved effective in detecting accidents.

Another type of sensor commonly used in accident detection systems is the GPS receiver. GPS receivers can detect changes in vehicle speed and location and can be used to determine if a vehicle has been involved in an accident. In a study by Kumar et al. the authors proposed a system that uses GPS and accelerometer sensors to detect vehicle accidents. The system was able to detect accidents with an accuracy of 95.2% and a false positive rate of 4.8% (Wang et al. 2017).

Several types of sensors can be utilized in accident detection systems in addition to accelerometers and GPS receivers. LiDAR sensors, for instance, are able to identify things in the surroundings and can be employed to ascertain whether a vehicle has collided with an object. The authors of a study (Shi et al., 2019) presented a system that uses LiDAR sensors to identify car accidents. With a 97% accuracy rate and a 3% false positive rate, the system was able to identify accidents.

An Automated Accident Detection System: A Hybrid Solution was presented by M. S. Abbas. The proposed technology can swiftly alert the emergency services or a worried family member with the precise location of an accident by using the Short Messaging Service (SMS). Unfortunately, it only has a few limited capabilities, such as the inability to detect fire or any criminal action involving the car (M. S. Abbas et al. 2019).

Using GPS and GSM, Sane et al. reported Real-Time Vehicle Accident Detection and Tracking system. This system is operated manually by pushing a button, and the contact

number to whom an alarm message must be delivered is hardcoded and will not be altered. The system will ask the motorist whether they want to send SMS or not after an accident. When the driver clicks the button, the microcontroller recognizes that the collision was not serious and decides not to send an SMS. Other times, if a collision has been detected and the button hasn't been depressed within the allotted time, the microcontroller will obtain the coordinates of the current location and will send an SMS alert to the driver's family using the GPS and GSM modules that have been installed (N. H. Sane et al. 2016).

Balfaqih, M. et al. (2021) describe an Internet of Things (IoT)-based system for accident detection and classification that can identify and categorize vehicle accidents according to their level of severity and give emergency response providers with critical accident-related information. In order to find the best accurate classifier for the system, many machine learning classifiers were tested. The system uses a microcontroller, GPS, and a number of sensors to identify various physical factors linked to vehicle movements. The system's implementation revealed that the models with the highest precision and recall were the Gaussian Mixture Model (GMM) and Classification and Regression Trees (CART). It was discovered that the g-force value and fire occurrence had a significant impact on how serious an accident was. The paper's main contribution is an efficient system for detecting and categorizing accidents that employs a powerful IoT platform to automatically record events and offer crucial details. The most precise model for classifying accident severity levels was determined through a comparison study of different machine learning classifiers.

Another article analyzes the requirement for an autonomous accident detection system for powered two-wheelers (PTW), such as motorcycles, scooters, and mopeds, in light of the rise in collisions and fatalities involving PTW users. Using factors specific to the vehicle and rider physiological data, the suggested system can identify accidents in real time. There are three steps in it: a system for detecting crucial events, one for detecting accidents, and one for determining the severity of the occurrence. An improved decision tree technique and an adaptive sequence window approach are suggested in the article to validate the existence of accidents based on the sequence of states found. Within five minutes, the system recognizes the fall of the car and the rider, avoiding false positives. Using a combination of three parameters, the Decision Support System (DSS) operating on the On-Board Diagnostic (OBD) unit mounted on the PTW determines the accident's severity after it has been identified. The author draws the conclusion that the faster response time provided by an autonomous accident detection system for PTW can contribute to a decrease in the death rate in PTW accidents (Jackulin Mahariba, A. et al 2022).

Fernandes, B. et al. (2016) proposed intelligent transportation systems (ITS) to lessen traffic accidents, which are a major issue for public injury prevention. The paper introduces HDy Copilot, a program for automatically detecting accidents that is integrated with alert distribution by eCall and IEEE 802.11p. (ITS-G5). The application takes data from the smartphone's accelerometer, magnetometer, and gyroscope in addition to the ODB-II system in the car. The driver can customize the application, get alerts, and remove erroneous accident detections using an Android smartphone as a human-machine interface. Successful collision, rollover, eCall, and Decentralized Environmental Notification Message detection and transmission are all capabilities of the application. (DENM). In order to focus research efforts within the vehicular communications scientific community, the essay underlines the need for standardization. The article focuses on the increase in safety however there are three categories of ITS benefits: improved transportation efficiency and environmental protection. By utilizing smartphones, ODB-II data, and vehicular communications, the HDy Copilot program

offers a cheap and portable alternative to built-in systems. The article stresses the significance of effectively detecting and diagnosing crashes involving motor vehicles because failure to do so could result in the loss of emergency time and resources. This issue is addressed by the program by limiting false positives through a countdown sequence that is started by the accident detection algorithm. The report does not, however, discuss the shortcomings of the HDy Copilot software or the possible risks of relying on smartphones to identify crashes involving motor vehicles. The ODB-II system, which may not be available in all vehicles, and the sensors on the smartphone may have an impact on the application's accuracy. With the gathering and distribution of data via vehicle communications, there might also be privacy issues.

Accidents in various industries can lead to injuries, fatalities, property damage, and economic losses, making it essential to implement measures to prevent them. While there have been significant improvements in safety measures, traditional reactive approaches to safety management are not entirely effective in reducing the frequency and severity of accidents. Hence, a proactive and predictive approach to safety management is necessary, which highlights the importance of developing accurate and reliable accident prediction and detection models. These models can identify potential safety hazards and prevent accidents before they occur. However, existing methods for accident prediction and detection face challenges, such as data quality and model accuracy. Therefore, there is a need to investigate the current state-of-the-art in accident prediction and detection, evaluate the effectiveness of existing methods, and propose novel techniques to improve safety management and prevent accidents (N. H. Sane et al., 2016). The overall objective of the study is to investigate the degree of influence of different variables that contribute to crash by identify the best model for accident prediction and accident detection.

2.6 Open Data

According to the “open knowledge foundation”, open data refers to any content, information, or data that anyone can freely use, re-use, and redistribute without any kind of restriction imposed by the law, technology, or society such as copyrights, patents, or other mechanisms of control. The key features of open data are:

- **Accessibility and availability**: The data should be easy to find, download, and use, preferably at no cost. The format of the data should be easy to work with and adaptable to different purposes. Essentially, anyone should be able to access it without jumping through hoops or paying a fee, and it should be in a format that makes it easy to manipulate and analyze.

- **Reuse and redistribution**: The data should come with permissions that allow people to use it again for different purposes and to share it with others. This includes combining it with other datasets. It should also be machine-readable, which means that computer programs can easily process it. The terms of use shouldn't restrict these activities.

- **Universal participation**: The data should be free to be used, reused, and shared by anyone. There shouldn't be restrictions on its use based on who is using it or for what purpose. This means that whether you're a business, a student, or just an interested individual, you should be allowed to use the data. There shouldn't be limitations based on the nature of usage, such as only for educational purposes or barring its use for commercial activities. Essentially, no one should be excluded

or favored in the use of the data.

Moreover, the movement towards open data is a response to the increasing importance of data in our society, with open data advocates arguing that limitations on access to data hinder innovation and public accountability. Open data practices often stem from public, academic, or non-profit entities, but can also be seen in private sectors with a growing emphasis on data sharing and collaboration. Data can be gathered through various means, such as surveys, experiments, administrative records, sensors, and more. Similarly, there are numerous benefits to open data practices. Not only does it promote transparency and collaboration, but it also has the potential to drive economic growth and innovation. It allows researchers, policymakers, businesses, and citizens to make informed decisions, solve problems, and develop new services and applications.

2.6.1 Open Data in Traffic Management Systems

Open data sets have become increasingly valuable in the field of traffic management systems. They provide useful, actionable information that can be leveraged to optimize traffic flow, enhance public transportation services, and improve the overall safety and efficiency of urban mobility.

- **Understanding Traffic Patterns**: Open data sets can provide detailed insights into traffic patterns. For instance, data on vehicle counts, types, speeds, and directional flows can help traffic managers understand and predict typical traffic behavior, identify hotspots of congestion, and optimize traffic light sequences. Additionally, historical data can help predict future patterns and inform infrastructure planning (Lv et al. 2014).

- **Enhancing Public Transportation**: Open data sets also play a vital role in enhancing public transportation systems. Real-time data about bus and train locations and delays can improve passenger information systems and contribute to more efficient route planning and scheduling (Monzon et al. 2012). Moreover, data on passenger numbers can help transit authorities to distribute resources more effectively and plan for future capacity needs.

- **Improving Road Safety**: Open data can also contribute to road safety. Analysis of accident data can help identify dangerous intersections and road segments and guide interventions aimed at reducing accidents (Anderson, 2009). Similarly, data on driving behavior can inform the development of strategies for promoting safer driving habits.

- **Facilitating Research and Innovation:** Finally, open data sets are a valuable resource for researchers and innovators. The availability of open traffic data can stimulate research into new traffic management approaches and technologies, contributing to the development of smart cities (Lv et al. 2014). It can also support the development of traffic-related applications and services by private sector companies, driving innovation and economic growth. However, it's important to consider the privacy implications of open data in traffic management. While open data can be anonymized and aggregated to protect individual privacy, there's still potential for misuse if not properly managed.

2.6.2 Other Datasets and Motivation for Chosen Dataset

Several open datasets are available for traffic management. One example is the UCI Machine Learning Repository, which has several datasets related to traffic management. Similarly, the EU Open Data Portal, UK's Department for Transport, and various city or state-level open data portals offer relevant data.

The choice of dataset often depends on the research question or application at hand. In comparing each source has its unique strengths and limitations in regards to their accident data offerings. However, for the purpose of comprehensive analysis and answering research questions efficiently and accurately, the US data on Kaggle is the preferred data source. Here are the reasons why:

- **Data Completeness**: The US data on Kaggle presents a more comprehensive data set in terms of the breadth and depth of variables provided. As compared to the UCI repository's 'UrbanGB' dataset which only has location coordinates of accidents, and the EU Open Data Portal's dataset which provides only a limited set of variables and requires tedious integration. This process can be time-consuming and increases the likelihood of data inconsistency and integrity issues. On the other hand, the US data on Kaggle offers a vast array of variables including those relevant to weather conditions, location, and time, as well as specific variables relating to traffic and road characteristics.

- **Data Accessibility**: Unlike UK's Department for Transport, which only provides reports rather than the raw data. The major downside here is that the data has already been processed and presented in a specific way, limiting flexibility and personalization in the analysis. This could be a hindrance for those who wish to conduct novel, independent research or apply different data analysis methodologies. In contrary to that, the US data on Kaggle allows direct access to the data for personalized and detailed analysis. This increases the flexibility and possibilities for data exploration.

- **Data Continuity and Volume**: The US data on Kaggle, being collected continuously since February 2016 and covering 49 states of the United States, boasts around 2.8 million accident records. This gives a rich, vast, and diverse data volume to work with. In contrast, the EU Open Data Portal only provides accident data for ten years from 2009-2019, and each year's data needs to be downloaded and merged separately. Also, the number of accident cases in each dataset is relatively small, approximately 2500, which limits the statistical power and generalizability of the findings.

- **Data Variety and Detail:** The US data on Kaggle provides a wealth of variables ranging from geographical, meteorological, and infrastructural to temporal factors. This allows for a more detailed and complex analysis of the accident data, including the interplay of multiple variables and the investigation of nuanced research questions. Therefore, taking into account all these factors, the US data on Kaggle is the most suitable choice for our analysis. These comprehensive attributes enhance the reliability and validity of the data and facilitate a thorough and precise examination of the research questions.

3 Research methodology

The purpose of this thesis is to predict the road accidents using machine learning and also aims to answer to the following research questions:

RQ1: How do weather and road conditions affect severity of accident and what are the most important factors that contribute to the likelihood of accident?
RQ2: Can machine learning models be used to accurately predict the likelihood of accidents, and how do different models compare in terms of accuracy and reliability?

3.1 Working of machine learning model

Below is a detailed explanation of the working flow of a machine learning model for accident prediction:

- Data selection and pre-processing:

The first stage is to choose appropriate data for accident prediction, which may include elements like the type of route, the amount of traffic, and the time of day. After it has been gathered, the data needs to be pre-processed to make sure it is clean and in a format that is appropriate for machine learning. This could entail eliminating missing values, scalability of numerical features, encoding of categorical variables, and partitioning the data into training and testing sets.

- Feature Selection:

The next stage is to choose the features that will be utilized to train the machine learning model that are the most appropriate. This is significant because duplicated or unnecessary features may have a negative effect on the model's performance.

- Model Selection:

There are many machine learning algorithms that can be used for accident prediction, such as decision trees, random forests, SVMs, and neural networks. The type of problem and the qualities of the data will determine which algorithm is used.

- Training the Model:

After deciding on an algorithm, training data are used to build the model. In order to obtain the greatest performance, this involves feeding the algorithm with the data, modifying the model's parameters, and fine-tuning the model.

- Model Evaluation:

After the model has been trained, it must be tested to see how well it works with untested data. The testing data set is often utilized for this, and metrics like accuracy, precision, recall, and F1 score can be used to evaluate the model's performance.

- Deployment:

Finally, the trained model can be used in a real-time environment to predict accidents in. This could involves integrating the model into a web application, mobile application, or other software system. Jupyter notebook is being used for accident prediction in this thesis. There are various benefits of utilizing Python and Jupyter Notebook for accident prediction such as:

Jupyter Notebook provides an interactive environment that allows us to run code, display data, and quickly evaluate the results. It is ideal for exploratory data analysis due to the ease with which variables can be changed, graphs can be created, and different predictive models can be tested. While conducting the analysis, developing and documenting the code in Jupyter Notebook is simple. Code can offer comprehensive reasons, illustrations, and comments to make it simpler to understand and maintain. Similarly, utilizing well-known frameworks for data visualization, such as Matplotlib and Seaborn. Jupyter Notebook seamlessly integrates. These packages allow us to create interesting visualizations that analyze connections, patterns, and trends in your accident data. For understanding the data and effectively communicating findings, visualizations are crucial. Jupyter It can also integrates Scikit-learn and TensorFlow, two well-known machine learning packages. These libraries include a wide range of pre-implemented tools and techniques for building predictive models. These libraries may be instantly loaded into Jupyter Notebook so we can test out different algorithms to predict accidents based on your data (Real Python, 2023).

Following are some libraries used with ML models:

- Pandas.
- Numpy.
- Sklearn.
- sklearn.tree.
- sklearn.neighbors.
- sklearn.naive_bayes.
- sklearn.ensemble.
- sklearn.model_selection.
- sklearn.metrics.
- xgboost.
- Imblearn.
- Matplotlib.

Following are some API's used with ML models:

- from sklearn.tree import DecisionTreeClassifier.
- from sklearn.neighbors import KNeighborsClassifier.
- from sklearn.naive_bayes.
- import GaussianNB .
- import xgboost as xgb.
- from sklearn.ensemble import RandomForestClassifier.
- from sklearn.model_selection import train_test_split.
- From sklearn.metrics import classification_report, confusion_matrix, accuracy_score.
- from imblearn.over_sampling import SMOTE.

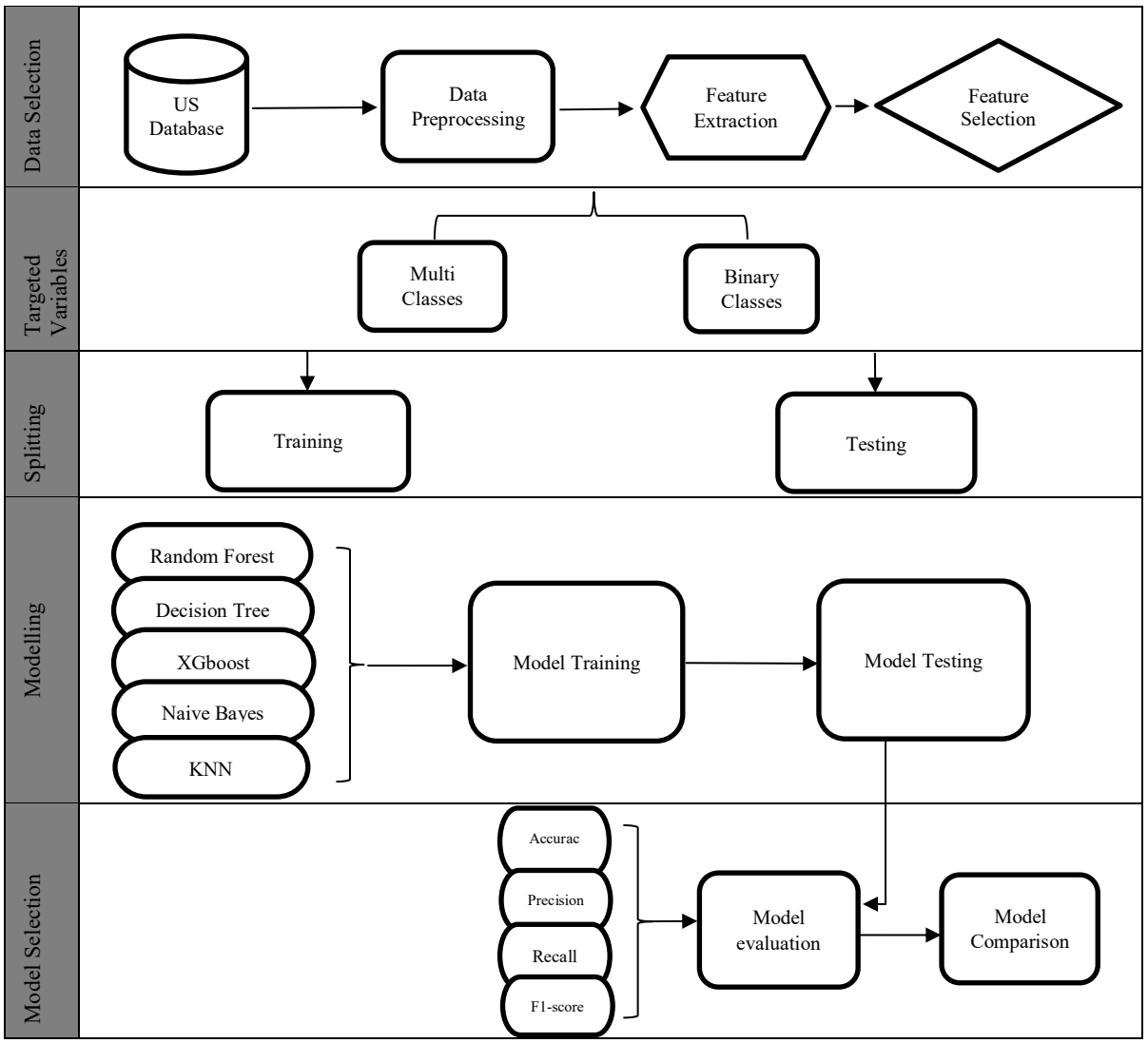

Figure 1: Working flow of ML models

3.2 Data selection and description

To predict the accidents and find out the accuracy of different machine learning models, this research used the freely available data on kaggle website (Moosavi, S. 2022). The dataset includes details on almost 2.8 million car crashes that happened in the US between February 2016 and December 2020. Among 2.8 million accident cases only 1048575 are available in the form of CSV. Police records, traffic cameras, and other public data sources were some of the sources from which the data was gathered. The data includes various features such as the location and time of the accident, weather and road conditions, severity of the accident, and details about the vehicles involved (Moosavi et al., 2019).

In an article, writers emphasize the significance of minimizing traffic accidents as well as the requirement for sizable datasets to forecast and evaluate them. In the past, most research have concentrated on small datasets with restricted coverage, whereas huge datasets have either been private or have excluded crucial contextual information. The authors developed US Accidents, a publicly accessible dataset of accident data. The authors also offer a fresh approach for compiling sizable databases of traffic accidents, and they report a variety of findings using the US Accidents dataset. The scientific

community can use this dataset, according to the authors, to enhance infrastructure for transportation and public transit, as well as to make roadways safer (Moosavi et al., 2019. Latitude, Longitude, Temperature, Wind chill, Humidity, Air pressure, Visibility, Wind speed, and Precipitation are some of the key variables that is used to predict the likelihood of accident based on severity of accidents in this thesis. Initially, data was imbalanced and values of each variables are given below:

Table 1: Demonstrating the values of each variables in number

Variables	**Values**
Latitude	1048575
Longitude	1048575
Temperature (F)	1026829
Wind Chill	834453
Humidity (%)	1025847
Air Pressure	1029990
Visibility (mi)	1026041
Wind Speed (mph)	990888
Precipitation (in)	816892
Severity	1048575

After removing null, unknown values, and cleaning the dataset it reduced to 785863 in total. According to (Moosavi, S. 2022), below is the explanation of each variable in the US-Accidents dataset:

- Latitude: The latitude coordinate of the location where the accident occurred.
- Longitude: The longitude coordinate of the location where the accident occurred.
- Temperature (F): The temperature at the time of the accident, measured in Fahrenheit.

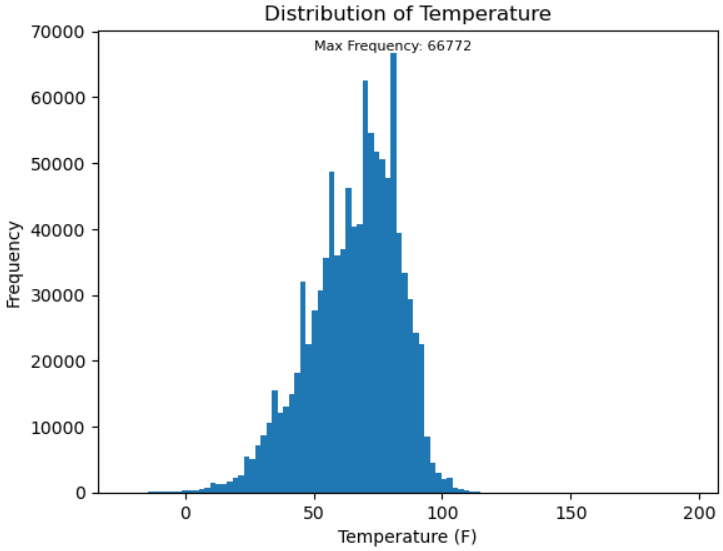

Figure 2: Image shows the number of accidents happened on a particular temperature

The temperature variable ranges from -89.0°F to 196.0°F. The distribution of temperature in the dataset is roughly normal, with a mean temperature of 61.6°F and a median temperature of 64°F. This suggests that most accidents occur under moderate temperature conditions with highest number of accidents occurred when the temperature was 70.60-76.30(F).

- Wind Chill: Wind chill is a weather variable that measures the perceived temperature on exposed skin due to the combined effect of wind and temperature.

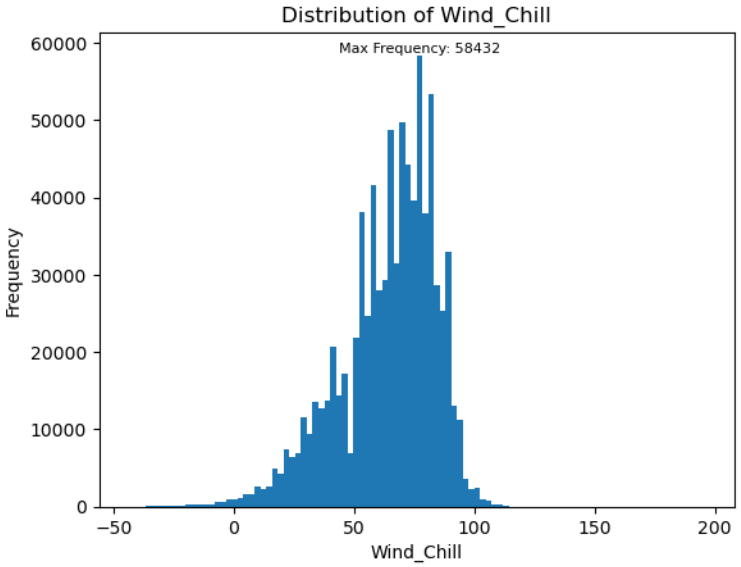

Figure 3: Image shows the number of accidents happened on a particular wind-chill

Wind chill is usually expressed as a temperature equivalent, and it is an important variable to consider when analysing traffic accidents, as it can affect driving conditions and human comfort. It ranges between the -50 and 200 (F). Highest number of accidents occurred when wind chill was 70.60 – 76.30 (F).

- Humidity (%): The relative humidity at the time of the accident, measured as a percentage.

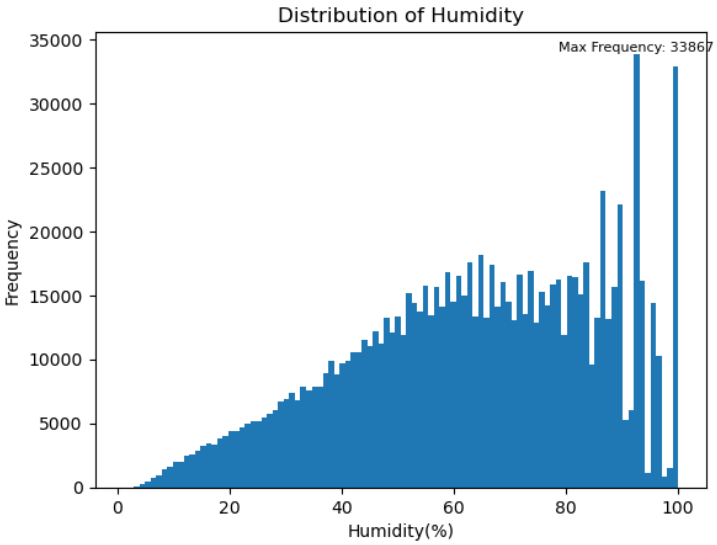

Figure 4: Image shows the number of accidents happened on a particular humidity level

It variable ranges from 0 to 100. The distribution of humidity in the dataset is roughly normal, with a mean humidity of 61.5% and a median humidity of 64%. This suggests that highest number of accidents occur when humidity was between 92.08-94.06.

- Air Pressure: The air pressure at the time of the accident, measured in inches of mercury (inHg).

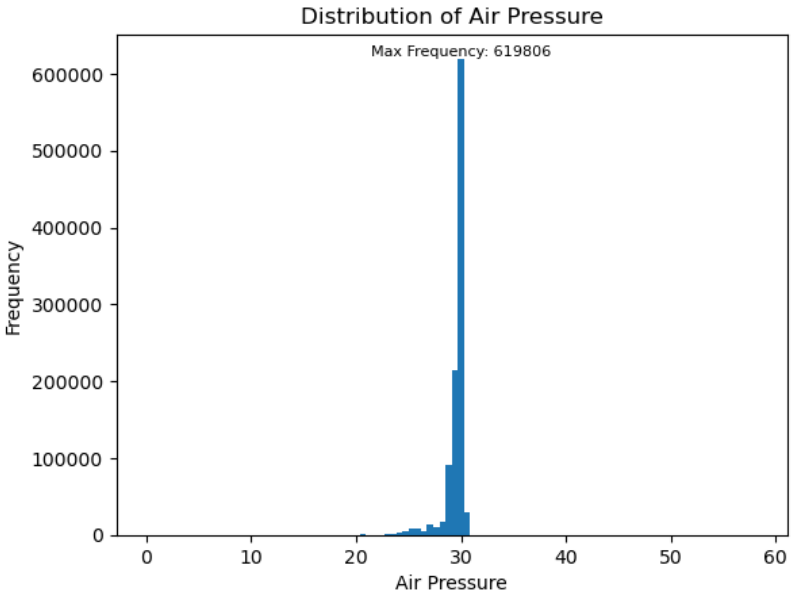

Figure 5: Image shows the number of accidents happened on a particular air pressure

The air pressure variable is continuous and ranges from 0.0 - 60.0 in Hg. The distribution of air pressure in the dataset is normally distributed, with the majority of accidents occurring under moderate air pressure conditions between 29.45 and 30.63 Hg.

- Visibility (mi): The visibility at the time of the accident, measured in miles.

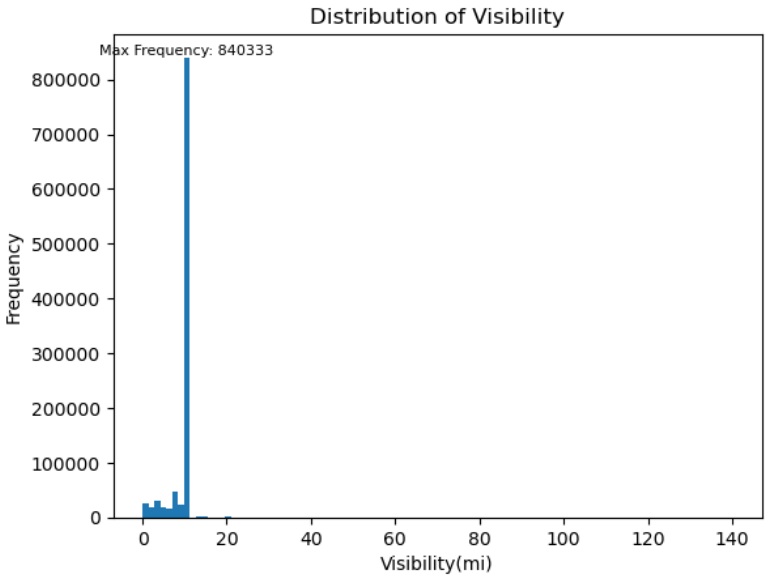

Figure 6: Image shows the number of accidents happened on a particular visibility level

It is continuous and ranges from 0.0 to 140.0 miles. The distribution of visibility in the dataset is skewed to the right, with a mean visibility of 9.19 miles and a

median visibility of 10.00 miles. This shows that most accidents occur in visibility ranges from 8.40 – 11.20 mi.

- Wind Speed (mph): The wind speed at the time of the accident, measured in miles per hour (mph).

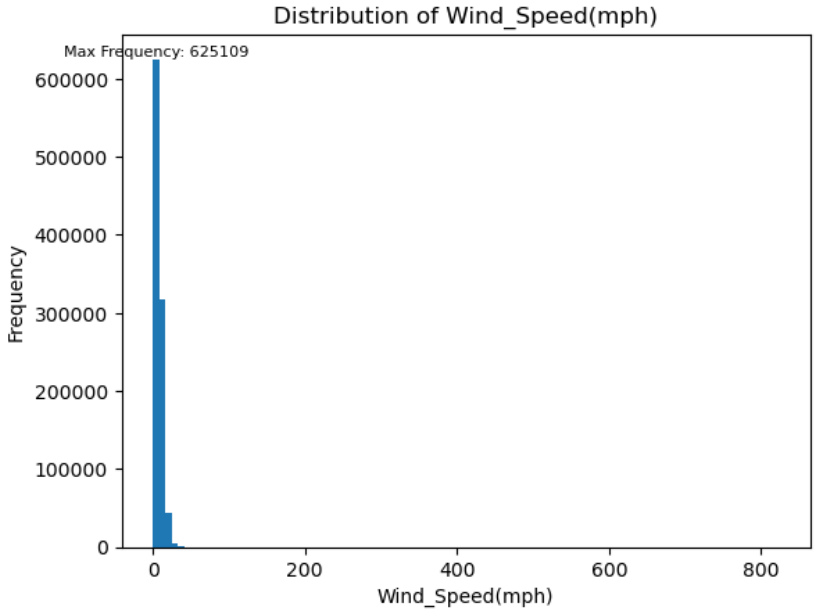

Figure 7: Image shows the number of accidents happened on a particular wind speed

It is also continuous and ranges from 0.0 mph to 1.09k mph. The distribution of wind speed in the dataset is highly skewed, with a majority of accidents occurring under calm wind conditions (i.e., wind speed between 0.0 – 21.76 mah).

- Precipitation (in): Precipitation is a weather metric that gauges how much liquid or solid water falls from the sky and makes it to the earth. Rain, snow, ice, hail, and drizzle are just a few of the different types of precipitation.

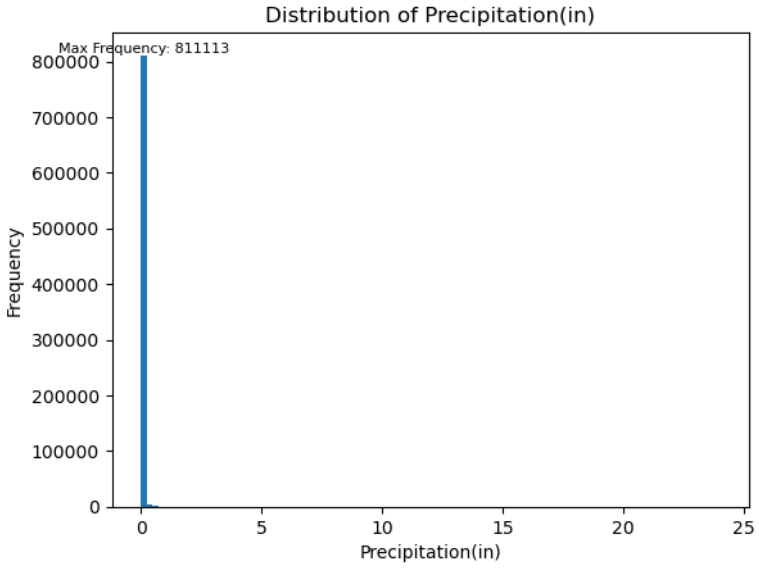

Figure 8: Image shows the number of accidents happened on a particular precipitation level

It is a crucial weather factor to take into account when examining traffic incidents because it can have an impact on traffic flow and driver behavior. The

precipitation variable is continuous and ranges from 0.00 inches to 25.25 inches. The distribution of precipitation in the dataset is highly skewed, with a majority of accidents occurring under dry conditions (i.e., no precipitation). Specifically, about accidents in the dataset occurred when there was no precipitation.

- Severity: A measure of the severity of the accident, ranging from 1 to 4, where 1 indicates a minor accident with short delay and 4 indicates a severe accident with long delay. Below is a brief description of what each severity level means:

    - Severity level 1: This indicates an accident with the least impact or damage. Typically, accidents in this category involve only property damage, such as a collision with a stationary object.
    - Severity level 2: This indicates an accident with a minor impact. Accidents in this category may result in some injuries, but they are not usually life-threatening.
    - Severity level 3: This indicates an accident with a significant impact. Accidents in this category may result in serious injuries and usually life-threatening.
    - Severity level 4: This indicates an accident with the most severe impact or damage. Accidents in this category usually involve multiple vehicles and fatalities.

Below is the table that shows the accidents according to each severity levels:

Table 2: Accidents according to the each severity levels.

Severity levels	Accidents
1	0
2	966042
3	39989
4	42544

3.3 Categorized the weather variable

Categorization is needed when working with features that represent different categories or groups. Categorization is performed on specific weather-related features such as temperature, wind chill, humidity, air pressure, visibility, wind speed, and precipitation. The goal of categorization is to turn numerical or continuous data into discrete groupings. This is done to make the data representation more straightforward and to identify any potential patterns or connections that may exist within each category. In dealing with non-linear relationships, categorization is helpful. It can also make it simpler for machine learning algorithms to find patterns and make predictions (Brownlee, 2020). Below are some of the categories of weather variables found at (National Geographic Society, 2020).

Temperature (F):

- Hot: temperatures above 90°F.
- Warm: temperatures between 70°F - 90°F.
- Cool: temperatures between 50°F - 69°F.
- Cold: temperatures below 50°F.

Wind Chill (F):

- Low: wind chill below 32°F.
- Moderate: wind chill between 32°F - 0°F.
- High: wind chill below 0°F.

Humidity (%):

- Dry: humidity below 30%.
- Comfortable: humidity between 30% and 60%.
- Humid: humidity above 60%.

Pressure (in):

- High pressure: air pressure above 30.00 inHg.
- Normal pressure: air pressure between 29.92 inHg and 30.00 inHg.
- Low pressure: air pressure below 29.92 inHg.

Visibility (mi):

- Good: visibility above 5 mi.
- Moderate: visibility between 2.5 mi and 5 mi.
- Poor: visibility below 2.5 mi.

Wind Speed (mph):

- Light: wind speed below 10 mph.
- Moderate: wind speed between 10 mph and 25 mph.
- High: wind speed above 25 mph.

Precipitation (in):

- Light: precipitation below 0.1 in.
- Moderate: precipitation between 0.1 in and 0.5 in.
- Heavy: precipitation above 0.5 in.

3.4 One-hot encoding

One-hot encoding is the conversion of categorical data to numeric form so that machine learning algorithms can use it to make better predictions. One-hot assigns a binary value of 1 or 0 to each category value after converting it into a new categorical column. A binary vector is used to represent each integer value. Consider a dataset having a "temperature" characteristic that can be one of the following: hot, cool, or cold. We would add the three additional features “temperature _hot," “temperature _cool," and " temperature _cold" to one-hot encoding. If the temperature is hot, we would use the vector [1, 0, 0] to represent it. The values for cool and cold would be [0, 1, 0] and [0, 0, 1], respectively.

One-hot encoding is employed in machine learning since deep learning neural networks and ML algorithms both require numerical input and output variables. We need a technique to convert categorical values from non-numeric formats, which are frequently used for categorical data, into numerical values without adding any extra or inaccurate information to the data. The algorithm might read this as an ordinal relationship (i.e., cold is "greater" than cool, and cool is "greater" than hot), which is not what we want. If we simply mapped the categories to a single numerical characteristic (hot to 1, cool to 2, cold to 3), this could happen. This issue is resolved by one-hot encoding (Brownlee, 2020).

3.5 Data balancing technique

Data imbalance is a problem that can affect the performance of ML models when the number of elements in one class is significantly lower than the number of elements in another class. The problem of data imbalance has been addressed using a variety of strategies, including under sampling the majority class or oversampling (SMOTE) the minority class. These methods seek to optimize the machine learning models' functionality while balancing the distribution of classes in the data set. When tested the performance of decision tree model with under sampling, the model gave the worst performance as shown in figure 9 below:

```
Overall accuracy = 22.22222222222222 %
Classification report:
              precision    recall  f1-score   support

           2       0.33      1.00      0.50         2
           3       0.00      0.00      0.00         6
           4       0.00      0.00      0.00         1

    accuracy                           0.22         9
   macro avg       0.11      0.33      0.17         9
weighted avg       0.07      0.22      0.11         9

Confusion matrix:
[[2 0 0]
 [4 0 2]
 [0 1 0]]
```

Figure 9: Shows the classification report of decision tree model with under sampling

In contrary to that, most of the ML models performed well with SMOTE (Synthetic Minority Over-sampling Technique) as shown in figure 10. That is why SMOTE techniques is being used in this thesis. SMOTE addresses data imbalanced problem by generating synthetic minority class instances. In order to operate, it chooses one element from the minority class and locates its k nearest neighbors in the feature space. Next, it interpolates between the chosen instance and its k neighbors to produce synthetic instances. This results in a larger minority class that is balanced with the majority class. SMOTE is typically used when dealing with classification problems where there is a class imbalance. It can be used with any machine learning algorithm, but it is particularly useful for algorithms that are sensitive to class imbalance, such as decision trees and random forest model (Chawla et al., 2002).

```
Overall accuracy = 97.76339691189827 %
Classification report:
              precision    recall  f1-score   support

           2       0.97      0.97      0.97      2927
           3       0.99      1.00      1.00      2963
           4       0.97      0.97      0.97      2918

    accuracy                           0.98      8808
   macro avg       0.98      0.98      0.98      8808
weighted avg       0.98      0.98      0.98      8808

Confusion matrix:
[[2825   11   91]
 [   3 2959    1]
 [  86    5 2827]]
```

Figure 10: Shows the classification report of decision tree model with SMOTE

3.6 Machine learning (ML) models

Five different machine learning (ML) models (Naive Bayes model, Random Forest, Extreme Gradient Boost, K-Neighbors Classifier, and Decision Tree) were used in our research to compare the performance of the five models based on Confusion Matrix, Precision, recall, and f1-score. Below is the detail of each model:

3.6.1 Naive Bayes

Naive Bayes Classifiers are one type of probabilistic classifier that uses the Bayes theorem to classify data in accordance with a specific set of observed evidence. The classification process is referred to as "naive" if the classifier makes the assumption that the presence or absence of one feature in a class is unrelated to the presence or absence of any other feature in that class. Several real-world applications of the naive Bayes classifier, such as text categorization, prediction, spam filtering, and sentiment analysis, have shown it to be effective despite this oversimplifying assumption (Berrar, D. 2019).

The naive Bayes classifier is well renowned for its effectiveness in lowering misclassification error rates, but it makes the assumption that characteristics are independent, which isn't always true in classification situations that occur in the real world. Nevertheless, the naive Bayes classifier has proven to function superbly even with dependent features. The excellent parallelizability of the conditional probability calculation in the naive Bayes classifier makes it a powerful tool for big data analytics. The article claims that by eliminating strongly correlated features, the classifier's performance can be improved.

Naive Bayes is like a detective. For example, it uses clues or features to make an educated guess about which category, or class, something belongs to. Let's imagine that we are engaged in accident prediction. We have information on things like the time of day, weather, and type of route. We wish to foretell whether or not an accident will result from these circumstances.

Here's how Naive Bayes can help:

- **Weather Conditions:** If dataset indicate that 70% of accidents occur in rainy conditions. This becomes our prior knowledge. Naive Bayes uses this to predict the likelihood of an accident when it's raining.

- **Time of Day**: If dataset indicate that 80% of accidents occur at night time. This is another clue for our "detective".

A Naive Bayes classifier, by virtue of its "naive" nature, now presumes that these features (weather condition and day of time) are not correlated. This may not always be the case in real life. For instance, it's possible that roadways are more likely to be slippery when it rains or that there are more vehicles on the road at that time of day, both of which could have an impact on accident rates. Given these two features, if we want to predict the chance of an accident on a rainy night on a highway, Naive Bayes will simply multiply the probabilities together. This might give us a high likelihood of an accident.

3.6.2 Random Forest

Random Forest is among the most well-known and popular algorithms used by data scientists. "Random forest," a well-liked supervised machine learning technique for categorizing data and forecasting, is used frequently. Decision trees are built using different samples, and the decision trees' majority vote is used to determine the average classification and regression conclusion. Its better performance helps with regression and classification issues. (R, S.E. 2023). Both regression and classification issues are successfully handled by it. The number of trees to be grown and the quantity of randomly sampled candidate variables for each split are the tuning parameters for a random forest. Only a random subset of the predictors is taken into account by the algorithm while creating a random forest, resulting in a high degree of variety that prevents over-fitting (Liaw et al., 2002).

Consider a teacher who must respond to a question that is incredibly challenging. What does the person do? He/she might solicit the opinions of their coworkers before deciding based on their comments. The Random Forest algorithm basically operates in this manner. A decision is made by a collection of decision trees working together. Comparable to asking a group of students in the class a question, each of whom brings a unique set of experiences and viewpoints to the discussion, he/she can frequently obtain a more thorough and accurate response from the group as a whole than from a single student. Let's bring this back to accident prediction. Suppose we have a dataset with various features like weather conditions, time of day, road type, and vehicle type. Each of these can help us predict whether an accident might happen. If we only utilized one decision tree, it might place too much emphasis on one element, such as the weather, while neglecting the others. This could result in overfitting, where the model performs admirably on the training set of data but falls short on the test set. We can prevent this using Random Forest. We guarantee that all features are taken into account by building a large number of decision trees (the "forest"), with each one looking at a random selection of features. The forest's trees each vote on whether an accident will occur, and the decision is made based on the results of the majority.

3.6.3 Extreme Gradient Boost (XGBoost)

Extreme Gradient Boosting (XGBoost) is one of the most widely used machine learning algorithms for classification, predictions, ranking, and regression applications. It is an ensemble strategy that combines a number of ineffective prediction models, such decision trees, to build an effective prediction model (Chen & Guestrin, 2016). A few applications that have employed XGBoost include recommendation systems, computer vision, and natural language processing.

The XGBoost method works by adjusting the weights of each observation in the training data and gradually adding additional decision trees to the ensemble. The overall prediction accuracy rises as a result. In each try, the method calculates the negative gradient of the loss function with respect to the ensemble predictions and then applies the results to build a new decision tree. The weights of the observations are then changed in light of the new hypotheses, and the ensemble is subsequently expanded to accommodate the new tree. The final prediction of XGBoost is the weighted average of all the ensemble forecasts of the decision tree. Each tree's weight is determined by how well it performs on a validation set, which prevents overfitting. (Brownlee, 2020).

For example consider XGBoost as a bit like the captain of a ship, steering it through a storm. The ship is our prediction model, and the storm represents the complexity of our data. The captain doesn't make one big steering turn and then hope for the best. Instead, they make a series of smaller adjustments, continuously correcting the course based on the ship's current position and the state of the storm. This is the essence of gradient boosting, the technique at the core of XGBoost. Now, let's translate this to our accident prediction scenario. Suppose we have data on various factors like weather conditions, time of day, and road conditions, which we want to use to predict the likelihood of an accident.

We start with a very simple model, like a decision tree. It's not perfect, but it gives us a starting point. Then we calculate how far off our predictions are from the real outcomes, which in machine learning lingo is referred to as the gradient of the loss function. This is where the 'boosting' part comes in. We train a new decision tree to predict not the actual outcome, but the error of our previous model. This new tree is like a second mate giving advice to our captain on how to correct the course. We keep adding new mates (or trees) to our ensemble, each one focusing on correcting the mistakes of the combined crew so far. The insight of each mate is weighted based on their accuracy, ensuring that the more accurate mates have more influence on the final decision.

Finally, our prediction is the sum of the insights from the captain and all the mates, weighted by their accuracy. This incremental approach is what makes XGBoost powerful, allowing it to gradually improve its predictions and tackle complex, real-world problems like accident prediction.

3.6.4 K-Neighbors Classifier

Alpaydin (2010) discuss that KNN is a non-parametric technique that classifies incoming data points based on how similar their features are to the features of the closest K training samples. Choosing a value for K is a hyperparameter that can be changed to enhance performance. When K is reduced, the decision boundary becomes more flexible and responsive to local data variations, while overfitting is also a possibility. Data points along the border between classes may, however, end up being misclassified as a result. The

decision border is rounded off and strengthened against noisy data by a higher value of K.

Going further, KNN is a straightforward, non-parametric algorithm for classification and predictions, according to Hastie, Tibshirani, and Friedman (2009). It locates the K nearest neighbors, finds the distance between the new data point and the training data point, then assigns the new data point's class label to the K nearest neighbors who have the highest proportion of that class label. We can understand the KNN with example given bellow.

Imagine Mr 'X' moving to a new city and you're looking for a place to live. He know that he prefer neighborhoods that are quiet, safe, and have a park nearby. One way to find a suitable neighborhood could be to talk to locals and ask them about their neighborhoods. But he doesn't just ask one person, instead ask several people from different neighborhoods. Then, he choose the neighborhood that most of the people you asked recommend. This is essentially how K-Nearest Neighbors (KNN) works. In the context of accident prediction, suppose we're trying to predict the likelihood of an accident at a certain intersection based on features like traffic volume, visibility, and weather conditions. Each intersection in our data is like a person we're asking, and the "recommendation" they give is their accident rate. KNN starts by looking at the 'K' most similar intersections to the one we're interested in, based on their features. The 'K' is something we can adjust. If we set K=1, we're only asking the most similar intersection. If we set K=10, we're asking the 10 most similar intersections.

There's a trade-off here. If K is too small, we're putting a lot of trust in a few intersections, which might mislead us if they're not representative or if the data contains some noise. This is like asking only one local about where to live and trusting their opinion completely, even though they might have peculiar tastes or had a bad day. On the other hand, if K is too large, we might dilute the information from the most similar intersections with less relevant ones. This would be like asking the whole city about where to live, even those living in areas completely different from what you're looking for. Once KNN has identified the 'K' nearest neighbors, it predicts the accident rate at the intersection of interest as the average accident rate of these neighbors. In other words, it "votes" the most common outcome from the K nearest neighbors. So, KNN is a simple but powerful method that can be effective for accident prediction. The key is to find the right balance for 'K', considering the specifics of your data.

3.6.5 Decision Tree

According to (Trevor Hastie et al., 2001) a common approach in deep learning for solving classification and regression problem is the decision tree. It is a hierarchical model that presents decisions and their potential outcomes in the form of tree, with each node or leaf signifying a decision and every branch mark a possible outcome of that decision. At each decision node/leaf, the algorithm selects the best feature to divide the data into groups based on some criterion, such as information gain or Gini index. Once a stopping condition has been met, such as reaching a maximum depth or having a minimal number of samples in a leaf node, the process is then restarted recursively for each subset. After that, the data is split into two or more subsets. The generated tree can be used to make predictions for additional data points by following the path from the root node to a leaf node that matches to the anticipated class or value.

Decision trees handle both category and numerical data, and they are resistant to outliers and missing values. They are also simple to comprehend and visualize. In addition, they

could undergo overfitting, which causes low generalization performance on new data since the tree is very complex and captures noise from the training set. Random forests, gradient boosting trees, and adaptive boosting trees are a few decision tree variants that aim to improve performance and reduce overfitting. Let’s understand this by following example.

Imagine someone planning a road trip. He got a list of decisions to make, like which route to take, when to leave, what to pack, etc. He might even draw a flowchart to help himself. That's basically what a Decision Tree is, but it's used for predicting stuff like accidents. Let's say we're trying to predict whether an accident is likely at a particular intersection. The Decision Tree might start by asking, "Is this a busy intersection?" If the answer is yes, it could then ask, "Are there traffic lights?" Depending on the answers, it will keep asking questions until it gets a clear picture of the situation, and then make a prediction - accident likely, or not. One of the great things about Decision Trees is that they're not picky about data. They can handle different types of data (like categories or numbers), and they're pretty robust when it comes to outliers or missing data points. Plus, they're pretty easy to understand because you can literally see the 'tree' of decisions that the model is making.

But, they're not perfect. Sometimes a Decision Tree can get too focused on the details and lose sight of the bigger picture. It's like planning for every single rest stop on your road trip, but forgetting to check the weather forecast. This is called overfitting, and it's when the model performs well on the data it was trained on, but not so great on new data. Luckily, there are ways to deal with this. Methods like Random Forests, Gradient Boosting Trees, and Adaptive Boosting Trees are like having a team of decision-makers. Instead of relying on one flowchart (or tree), these methods use lots of them and then combine their predictions. This can lead to better, more reliable predictions. So, in a nutshell, Decision Trees are a handy tool for making predictions, as long as you're aware of their limitations and how to handle them. They're like your trusty road map - not always 100% accurate, but definitely useful for navigating the data landscape.

3.7 performance matrix

To compare the performance of the each model the detail of performance matrix such as Confusion Matrix, Precision, recall, and f1-score is given below:

3.7.1 Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model by comparing its predictions to the true labels of the data. It contains four terms: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). TP and TN indicate correct predictions, while FP and FN indicate incorrect predictions (Powers, D. M. 2011).

Here is an example of a confusion matrix for a binary classification problem:

Table 3: Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The number of instances that are both genuinely positive and projected to be positive is known as True Positive (TP). The number of occasions where a result is projected to be negative but is really positive is known as a false positive (FP). The number of situations where a predicted negative outcome actually occurs is known as false negatives (FN). The number of instances that are both genuinely negative and projected to be negative is known as True Negative (TN).

Imagine a model that's guessing whether an accident will happen at an intersection. There are four possibilities:

- The model predicts an accident, and it's right. This is called a True Positive (TP).
- The model predicts no accident, and it's right again. This is a True Negative (TN).
- The model predicts an accident, but it's wrong. That's a False Positive (FP).
- Lastly, the model predicts no accident, but it's wrong - an accident happens. This is a False Negative (FN).

Precision, recall, and F1-score are just a few of the performance indicators for the classification model that can be computed using the confusion matrix. These metrics can be used to compare various models or algorithms and assess the model's overall performance.

3.7.2 Precision

Precision is a statistic used to assess a classification model's performance, primarily in binary classification issues. It is described as the percentage of accurate positive predictions among all the positive forecasts the model made. It is a measure we use to determine just how good our accident prediction model is at its job. Imagine it like a detective trying to solve a case. It's not enough for the detective to gather a massive amount of evidence. They also need to ensure that the evidence they've collected is relevant and points them in the right direction. If they're dealing with a lot of false leads, it's going to be a lot harder to solve the case. It can be calculated as:

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \quad (1)$$

False Positives (FP) are instances that are genuinely negative but were predicted to be positive, whereas True Positives (TP) are cases that are actually positive and predicted to be positive.

In other words, precision assesses the model's ability to distinguish true positives from all other occurrences that it correctly classifies as positive. A low false positive rate means that the model predicts few positive outcomes incorrectly, which is indicated by a high precision. A low precision, on the other hand, means that the model consistently predicts good outcomes. In fields like fraud detection or medical diagnosis, where a false positive can result in pointless treatments or investigations, precision is a valuable indicator when the cost of false positives is large. (Powers, D. M. 2011).

3.7.3 Recall

Recall is a statistic used to assess a classification model's performance, notably in binary classification issues. It is described as the percentage of accurate positive predictions among all instances of real positive data, and can be calculated as:

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (1)$$

Recall quantifies how well the model separates out the real positives from all other positive occurrences in the data, i.e., the examples that truly belong to the positive class. When a model has a high recall, it means that it properly recognizes the majority of positive cases and has a low incidence of false negatives. A low recall, on the other hand, suggests that the model is missing a lot of successful cases. When the cost of false negatives is large, like in medical diagnosis or fraud detection, when a false negative can result in missed diagnoses or fraudulent activity going unnoticed, recall is a useful indicator (Powers, D. M. 2011).

For example, recall in the context of predicting accidents, is like a safety net for our model. It tells us how well the model catches all the actual accident cases from the dataset. To put it simply, it's all about the question: Out of all the actual accidents that occurred, how many did our model successfully predict?

Consider the formula below:

$$\text{Recall} = \text{Correct Accident Predictions} / (\text{Correct Accident Predictions} + \text{Missed Accident Predictions}) \quad (2)$$

Here, "Correct Accident Predictions" are the instances where our model predicted an accident and, unfortunately, an accident did happen. These are our True Positives (TP) means the model got it right. On the other hand, "Missed Accident Predictions" are the instances where our model didn't predict an accident, but an accident actually happened. These are our False Negatives (FN) means the model missed these.

A high recall means that our model is good at catching accidents - it misses very few actual accident cases. This is crucial in scenarios like accident prediction where missing an actual accident can have serious consequences. However, a low recall means our model is frequently missing actual accident cases. This could lead to a lack of preventive measures when they're actually needed, which can be hazardous.

Overall, recall helps us measure the completeness of our accident prediction model. A good model should have high recall, making sure it flags most, if not all, of the accidents that are about to happen.

3.7.4 F1-score

F1-score is a metric used to evaluate the performance of a classification model. It is the harmonic mean of precision and recall, and is calculated as:

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (1)$$

The previously mentioned metrics, recall and precision, assess how successfully the model distinguishes genuine positives from false positives among all positive predictions and true positive cases, respectively. The balance between precision and recall is represented by the F1-score, which combines these two metrics into a single value. It has a value between 0 and 1, with a higher number denoting better performance. An F1-score of 0 implies that the model's predictions are wholly erroneous, whereas an F1-score of 1 suggests flawless precision and recall.

When the distribution of the classes is unbalanced. That is when one class has much more instances than the other. The F1-score is especially helpful. In these situations, accuracy might be a deceptive indicator because a model that merely predicts the majority class will be accurate to a high degree. The F1-score provides a more illuminating evaluation of the model's performance because it accounts for both recall and precision (Powers, D. M. 2011).

3.8 Statistical techniques used to answer the first question

3.8.1 Chi-square test

According to (Gibbons & Chakraborti, 2021), the chi-square test is used to determine the independence of two categorical variables by calculating the chi-square statistic, which follows a chi-square distribution with degrees of freedom determined by the size of the contingency table. Chi-square value measures the statistical significance of the association between two categorical variables. The higher the chi-square value, the more significant the association. In order to perform the test, one must compute the chi-square statistic, which is the sum of the squared differences between the observed and predicted frequencies divided by the expected frequencies. This statistic has a chi-square distribution, which is dependent on the size of the contingency table (i.e., the number of rows and columns) in terms of the number of degrees of freedom.

The chi-square test includes determining a test statistic, represented by the X^2, that quantifies the discrepancy between the observed and expected frequencies. The chi-square distribution that the test statistic follows has a shape that is dependent on the test's degrees of freedom. In the contingency table, the degrees of freedom are determined as the sum of the products of the number of rows minus one and the number of columns minus one.

The alternative hypothesis in chi-square is that there is a substantial difference between the observed and anticipated frequencies, while there isn't one, which is the null hypothesis of the chi-square test. The null hypothesis is rejected if the test statistic is greater than the critical value, which is determined using a critical value from the chi-square distribution and the relevant degrees of freedom.

3.8.2 Cramer's V

Cramer's V is a measure of association between two nominal categorical variables. It is a statistic that ranges from 0 to 1, with higher values indicating a stronger association between the two variables. Cramer's V is based on the chi-squared test statistic and takes into account the number of categories in each variable (Gibbons & Chakraborti, 2021).

The formula for Cramer's V is:

$$V = \sqrt{(\chi 2 \,/\, n * (min(r, c) - 1))} \quad (1)$$

Where $\chi 2$ is the chi-squared test statistic, n is the total number of observations, r is the number of rows, and c is the number of columns in the contingency table of the two categorical variables. Cramer's V can be interpreted as follows:

- 0 indicates no association between the two variables.
- 0.1 to 0.3 indicates a weak association.
- 0.3 to 0.5 indicates a moderate association.
- 0.5 to 1 indicates a strong association.

Cramer's V is commonly used in fields such as psychology, sociology, and market research to analyse the relationship between two categorical variables.

5 Results

5.1 Statistical analysis

To answer the third research question "*How do weather and road conditions affect severity of accident and what are the most important factors that contribute to the likelihood of accident?"*, different statistical techniques such as chi-square analysis with p-value, degree of freedom and cramer's v is used.

Table 4: Chi-square analysis with 10000 samples using natural categories

Weather variables	**Chi-square values**	**P-values**	**Degree of freedom**	**Cramer's V**
Temperature	8973.703	0.000	6	0.452
Visibility	11025.152	0.000	4	0.501
Air Pressure	1772.067	0.000	4	0.201
Wind Speed	4479.024	0.000	4	0.319
Humidity	3306.284	0.000	4	0.274
Precipitation	992.258	0.000	4	0.150
Wind Chill	6732.137	0.000	4	0.391

Above data provides the results of chi-square tests and Cramer's V for the association between each weather variable and the severity of accidents. Below are some conditions (mentioned in the previous section), should be considered while analyzing the results (Gibbons & Chakraborti, 2021).

- Chi-square value measures the statistical significance of the association between two categorical variables. The higher the chi-square value, the more significant the association.
- P-value measures the probability of observing a chi-square statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true. A p-value of less than 0.05 is generally considered statistically significant.
- Degrees of freedom represent the number of categories that can vary in the contingency table without changing the chi-square value.
- Cramer's V is a measure of association between two categorical variables, which takes into account both the chi-square value and the sample size. The range of Cramer's V is between 0 and 1, with values closer to 1 indicating a stronger association.

Analyzing the given data, it is clear that all weather variables are associated with the severity of accidents, with p-values less than 0.05.

Below is the detail of each variable and how it is associated with the severity of accidents based on the given Chi-square statistics.

Temperature: The Chi-square value for temperature is 8973.703, and p-value < 0.05, indicating a strong association between temperature and the severity of accidents. The Cramer's V value of 0.452 also indicates a moderate association. There is moderate association between temperature and severity of accidents which suggests that high or very low temperatures, can impact road conditions and driver behavior, leading to more severe accidents. This conclusion is in line with the outcomes of the research carried out by Basagaña et al. (2015) and Zou et al. (2021), confirming that temperature influences the occurrence of accidents.

Visibility: The Chi-square value for visibility is the highest among all the weather-related variables, at 11025.152. The p-value is also indicating a strong association between visibility and accident severity. The high Cramer's V value of 0.501 indicates a strong association. Therefore, it is clear that poor visibility due to fog, rain, or other weather conditions can make it difficult for drivers to see, leading to more severe accidents. This finding aligns with the results from studies conducted by Khodadadi-Hassankiadeh et al. (2020) and Sangkharat et al. (2021), all of which found a highly increase in traffic accidents during foggy and rainy weather.

Air Pressure: The Chi-square value for air pressure is low as compare to temperature and visibility and p-value is indicating a moderate association between air pressure and the severity of accidents. The Cramer's V value of 0.201 suggests a weak association means road accidents are barely connected with air pressure.

Wind Speed: The Chi-square value for wind speed and the p-value, indicating a moderate association between wind speed and accident severity. The Cramer's V value of 0.319 suggests a moderate association. High wind speeds can impact vehicle stability and control, leading to accidents.

Humidity: The Chi-square and p-value for humidity also indicating a moderate association between humidity and accident severity. The Cramer's V value of 0.274 suggests a week association. High humidity can lead to fog and reduced visibility, while low humidity can make the road surface dry and slippery, both of which can lead to more severe accidents. This conclusion coincides with the findings from research carried out by Singh, S. (2015), Zou et al. (2021), Khodadadi-Hassankiadeh et al. (2020), and Sangkharat et al. (2021). All these studies explored a substantial rise in traffic incidents under conditions of poor visibility and wet or slippery roads.

Precipitation: The low Chi-square and p-value for precipitation, indicating a weak association between precipitation and the severity of accidents. Similarly, the Cramer's V value of 0.150 also suggests a weak association. In results rain, snow, and other forms of precipitation may make the road surface slippery, may become cause of accidents. These results are somehow aligned with the findings of Zeng et al. (2020).

Wind Chill: The Chi-square and p-value for wind chill indicating a moderate association between wind chill and accident severity. The Cramer's V value of 0.391 suggests a moderate association. These findings are matches with the research carried out by Zou et al. (2021).

5.1.1 Conclusion

Chi-square statistics suggest that all the weather-related variables are more or less associated with the severity of accidents, with visibility and temperature emerge as the most critical weather-related factors affecting the severity of road accidents. Therefore, efforts to mitigate the impact of these factors, such as implementing effective fog dispersal systems, heatwave alerts, or improved road maintenance during extreme temperatures, may help reduce the severity of accidents. The other factors, while less significant, should not be overlooked as they can still contribute to the likelihood and severity of accidents. Drivers should be educated about the potential risks associated with these weather conditions, and appropriate safety measures should be put in place.

Going further, in this study, although the primary analytic method employed was the Chi-square test, alternative statistical approaches were also explored, namely correlation and logistic regression. These methods were thoroughly investigated and analyzed, but they were not ultimately utilized in this thesis. Following is the result of correlation analysis that could be helpful for the future researchers:

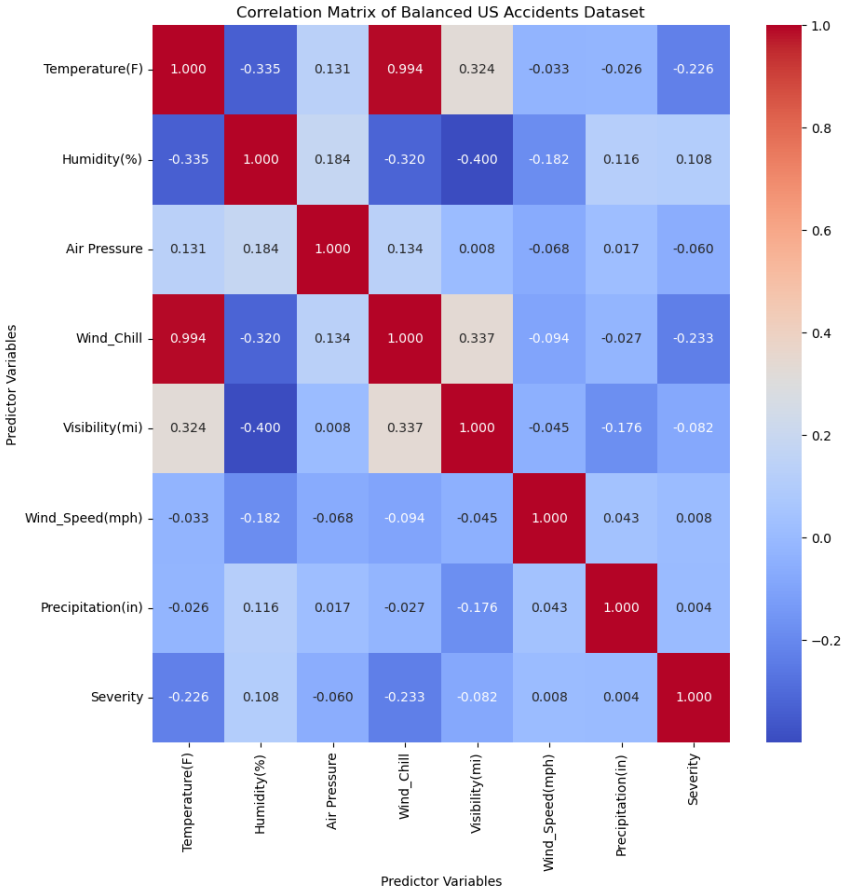

Figure 11: Shows the correlation coefficient of each variable

The correlation matrix show information about the relationship between variables. The values range from -1 to 1, where -1 means a perfect negative correlation, 0 means no correlation, and 1 means a perfect positive correlation. The diagonal values are always 1 because all variables are perfectly correlated with them self.

Table 5: P-values of each variable against severity of accident.

Weather variables	Severity
Temperature	0.000000e+00
Wind-chill	0.000000e+00
Humidity	0.0
Air Pressure	0.000000e+00
Visibility	0.000000e+00
Wind Speed	1.500377e-22
Precipitation	7.324876e-06

Temperature (F): The correlation coefficient between temperature and severity is negative, indicating that as the temperature decreases, the severity of accidents increases. This could be due to several factors, such as icy roads, reduced visibility due to fog, and increased risk of hypothermia or frostbite, which can lead to more severe accidents. The p-value is very low, which means that this correlation is statistically high and prove the indirect link between temperature and severity of accident.

Wind-chill: The correlation coefficient between wind chill and severity is also negative, indicating that as the wind chill decreases, the severity of accidents increases. This makes sense because wind chill is the perceived temperature that takes into account the effect of wind on the body. Again, the p-value is very low between wind chill, indicating a statistically significant correlation.

Humidity (%): The correlation coefficient between humidity and severity is positive, indicating that as humidity increases, the severity of accidents also increases. This could be because high humidity can reduce visibility due to fog, and also cause slippery roads due to moisture. The p-value is very low, indicating a statistically correlation between them.

Air Pressure: The correlation coefficient between air pressure and severity is negative, indicating that as air pressure increases, the severity of accidents decreases. This is because high pressure generally leads to clear skies, which means better visibility and safer driving conditions. Similarly, the p-value is quite low, indicating the high correlation.

Visibility (mi): The correlation coefficient between visibility and severity is negative, indicating that as visibility decreases, the severity of accidents increases. This is because low visibility can make it difficult for drivers to see and avoid obstacles or other vehicles. The p-value is very low, indicating a statistically high correlation.

Wind Speed (mph): The correlation coefficient between wind speed and severity is very low, and the p-value is very low. This means that there is little to no correlation between wind speed and accident severity.

Precipitation (in): The correlation coefficient between precipitation and severity is very low, and the p-value is very low. This means that there is little to no correlation between precipitation and accident severity.

Finally, based on the correlation coefficients and p-values provided, it seems that temperature, wind chill, humidity, and visibility are the most important weather variables that are correlated with accident severity and affect severity of accident.

Coming to the logistic regression, of the output shows the classification report, which evaluates the performance of the logistic regression model on the test data. The report shows the precision, recall, and f1-score for each class (0 and 1) as well as the overall accuracy which is 92%.

Table 6: Performance of logistic regression

Classes	Precision	Recall	F1-score
0	0.92	1.00	0.96
1	0.93	0.92	0.92

The second part of the output shows the importance of each predictor variable in the logistic regression model. The coefficients of the logistic regression model represent the importance of each predictor variable in predicting the outcome variable (in this case, the severity of the accident). The larger the absolute value of the coefficient, the more important the corresponding predictor variable is in predicting the outcome.

For example, the largest coefficient is associated with the "Precipitation (in)" variable, which suggests that this variable has the most influence on the severity of accidents. However, it's important to note that the coefficient of "Precipitation (in)" is positive, which means that as the amount of precipitation increases, the severity of the accidents also tends to increase. On the other hand, the coefficients for "Wind Chill" and "Humidity (%)" are negative, which means that as these variables increase, the severity of the accidents tends to decrease. Latitude and longitude are geographic coordinates that can help identify the location of the accident, which can be useful for emergency services and traffic management. However, their coefficients in the logistic regression model are relatively small, indicating that their influence on the severity of accidents is not as significant as other predictor variables.

Table 7: Importance of each predictor variable

Predictor variables	Output
Latitude	0.055312198479903756
Longitude	0.017367972779213432
Temperature(F)	0.04938336722084942
Wind Chill	-0.054267023704441815
Humidity (%)	-0.009621787826803252
Air Pressure	0.03889089710162015
Visibility(mi)	0.05394165881901363
Wind Speed(mph)	0.026440797382297262
Precipitation(in)	2.073401902914903

Similarly, Temperature (F) and Air Pressure are environmental conditions that can also affect the severity of accidents. A higher temperature may lead to more driver fatigue or vehicle malfunctions, while lower air pressure may affect the vehicle's performance. The coefficients for these variables are positive, indicating that as these variables increase, the severity of accidents tends to increase. Visibility (mi) and Wind Speed (mph) are also important predictor variables that can affect the severity of accidents. Poor visibility due to fog, rain, or snow can increase the likelihood of accidents. Similarly, high wind speeds can cause vehicles to lose control and lead to accidents. The coefficients for these variables are also positive, indicating that as these variables increase, the severity of accidents tends to increase.

In conclusion, all the predictor variables listed in the table are important in predicting the severity of accidents to some degree but precipitation is the most important factor in predicting the severity of accidents. However, their influence and importance may vary depending on the specific conditions and circumstances of each accident.

5.2 Comparing the Performance of Machine Learning Models in Predicting Accident Likelihood

To answer the second research question "*Can machine learning models be used to accurately predict the likelihood of accidents, and how do different models compare in terms of accuracy and reliability?*" this section used the five machine learning (ML) models (Naive Bayes model, Random Forest, Extreme Gradient Boost, K-Neighbours Classifier, and Decision Tree) in our research to compare the performance of the five models based on Precision, recall, and f1-score. Before working with any ML model, over-sampling technique is used to deal with the data imbalance issue. Below is the result of each model.

5.2.1 Random forest

Based on the calculations of random forest model the overall accuracy is 98.53%, which means that the models were able to correctly classify 98.53% of the accidents in the test set. The classification report provides precision, recall, and F1-score for each severity level. For severity level 2, the precision is 0.98, recall is 0.97, and F1-score is 0.98. For severity level 3, the precision, recall, and F1-score are all 1.00, indicating perfect classification. For severity level 4, the precision is 0.98, recall is 0.98, and F1-score is 0.98. All three severity levels (2, 3, and 4) have high precision, recall, and f1-score values, indicating good model performance. In this case, the precision for all three classes is above 0.98, indicating that the model is correctly predicting a high proportion of positive instances. The recall for all three classes is also above 0.97, indicating that the model is correctly identifying a high proportion of actual positive instances. Additionally, the f1-score for all three classes is above 0.98, indicating a good balance between precision and recall.

Overall, the machine learning model performed very well in predicting the severity of accidents, achieving high accuracy and F1-scores with balanced dataset. Similarly, based on the performance of this model and the high overall accuracy and precision/recall scores for all classes, it is reasonable to say that random forest model can be used to accurately predict the likelihood of accidents, at least in this particular dataset and with the chosen features. In this research findings, the random forest model demonstrates a high degree of accuracy at 98.5%. This compares favorably to the accuracy of 77.6% reported in the research conducted by Chen M and Chen C (2020), suggesting that this model provides more precise predictions with dataset used in this research.

Table 8: Classification report of random forest model with accuracy

Severity Level	**Precision**	**Recall**	**F1-score**
2	0.98	0.97	0.98
3	1.00	1.00	1.00
4	0.98	0.98	0.98
Overall accuracy= 98.52717570403603 %			

5.2.2 Decision Tree

Based on the performance metrics of decision tree the overall accuracy of the model is 95.325%, which means that it is able to correctly predict the severity of accidents 95.325% of the time. The classification report provides more detailed performance metrics for each severity class. In this case, the precision, recall, and f1-score for severity level 2 are 0.93, 0.93, and 0.93 respectively. For severity level 3, the precision, recall, and f1-score are 0.99, 1.00, and 0.99 respectively. Finally, for severity level 4, the precision, recall, and f1-score are 0.93, 0.93, and 0.93 respectively.

Therefore, the high precision, recall, and f1-score for all classes indicate that the model has good performance and can be used to accurately predict the likelihood of accidents across all severity classes. The Random Forest model outperforms the Decision Tree model in terms of accuracy. Specifically, the Random Forest model boasts an accuracy rate of 98.5%, which is higher than the 95.3% accuracy rate exhibited by the Decision Tree model which is aligned with results of research conducted by Chen M and Chen C (2020).

Table 9: Classification report of decision tree model with accuracy

Severity Level	Precision	Recall	F1-score
2	0.93	0.93	0.93
3	0.99	1.00	0.99
4	0.93	0.93	0.93
Overall accuracy= 95.32526435678483%			

5.2.3 K-Neighbours Classifier

The K-Neighbor classifier model achieved an overall accuracy of 96.012% on the test data which also support the findings of Iranitalab and Khattak (2017). Looking at the classification report, it is clear that the model performed well on all three classes. For the severity level 2, the precision is 0.99, recall is 0.89, and f1-score is 0.94. This means that out of all the predicted accidents, 99% of them were actually from severity level 2, and out of all the actual severity level 2 accidents, 89% of them were correctly identified by the model. The f1-score is a harmonic mean of precision and recall, which gives an overall measure of the model's accuracy on this severity level. The f1-score of 0.94 indicates that the model performed well on severity level 2.

For class 3, the precision is 0.99, recall is 1.00, and f1-score is 0.99. This means that out of all the predicted accidents, 99% of them were actually from severity level 3, and out of all the actual accidents, 100% of them were correctly identified by the model. The high f1-score of 0.99 indicates that the model performed very well on severity level 3. Going further, for severity level 4, the precision is 0.91, recall is 0.99, and f1-score is 0.95. This means that out of all the predicted accidents, 91% of them were actually from severity level 4, and out of all the actual accidents, 99% of them were correctly identified by the model. The f1-score of 0.95 indicates that the model performed well on this class, but not as well as on classes 2 and 3.

Overall, the KNN classifier model performed well on all three classes, with high precision, recall, and f1-score. This suggests that it is a good candidate for predicting the likelihood of accidents.

Table 10: Classification report of K-Neighbours Classifier with accuracy

Severity Level	Precision	Recall	F1-score
2	0.99	0.89	0.94
3	0.99	1.00	0.99
4	0.91	0.99	0.95
Overall accuracy= 96.01210633804953 %			

5.2.4 Extreme Gradient Boost (XGBoost)

In this case, the overall accuracy of the XGBoost is 92.645%. Looking at the classification report, the precision, recall, and f1-score are reported for three different severity levels: level 2, level 3, and level 4. For severity level 2, the precision is 0.92, which means that out of all the instances that the model predicted as severity level 2, 92% were actually severity level 2. The recall is 0.94, which means that out of all the instances that were actually severity level 2, the model correctly identified 94% of them. The f1-score is 0.93, which is a balanced measure of precision and recall.

Similarly, for severity level 3, the precision is 0.95, which means that out of all the instances that the model predicted as severity level 3, 95% were actually severity level 3. The recall is 0.97, which means that out of all the instances that were actually severity level 3, the model correctly identified 97% of them. The f1-score is 0.96, which is high and indicates a well-balanced performance.

Finally, for severity level 4, the precision is 0.92, which means that out of all the instances that the model predicted as severity level 4, 92% were actually severity level 4. The recall is 0.87, which means that out of all the instances that were actually severity level 4, the model correctly identified 87% of them. The f1-score is 0.89, which is lower than the other severity levels, indicating that the model has some difficulty in accurately identifying severity level 4 incidents. Overall, the performance of the model seems to be good, with high accuracy and good precision and recall scores for most severity levels and based on the given data and performance metrics, it seems that the XGBoost is able to accurately predict the likelihood of accidents with good precision, recall, and f1-score scores.

Table 11: Classification report of Extreme Gradient Boost with accuracy

Severity Level	Precision	Recall	F1-score
2	0.92	0.94	0.93
3	0.95	0.97	0.96
4	0.92	0.87	0.89
Overall accuracy= 92.64512883960523 %			

5.2.5 Naive Bayes

The overall accuracy of the Naive Bayes model is 63.31%, which is relatively lower compared to the other models I discussed earlier. This means that the model correctly predicted all the severity levels for about only 63% of the cases.

Looking at the classification report, it is clear that the precision, recall, and f1-score for

severity level 2 are same 0.59, 0.59, and 0.59 respectively. For severity level 3, the precision, recall, and f1-score are 0.68, 0.78, and 0.73 respectively. Finally, for severity level 4, the precision, recall, and f1-score are 0.62, 0.53, and 0.57 respectively. Therefore, Naive Bayes model performs better for severity level 3 as compared to the other levels. This is because the precision, recall, and f1-score are higher for severity level 3 as compared to the other levels. The model has a high recall value for severity level 3, which means that it correctly identifies most of the actual severity level 3 cases.
However, the model's performance is relatively poor for severity level 4, as the recall value is relatively low. This means that the model is not able to identify all the actual severity level 4 cases. Overall, this machine learning algorithm appears to be inaccurate in predicting accidents based on its performance measures.

Table 12: Classification report of Naïve Bayes with accuracy

Severity Level	**Precision**	**Recall**	**F1-score**
2	0.59	0.59	0.59
3	0.68	0.78	0.73
4	0.62	0.53	0.57
Overall accuracy= 63.308174391646965%			

5.2.6 Conclusion

Based on the classification reports and overall accuracy values for each model, it is clear that machine learning models can be used to predict the likelihood of accidents with varying degrees of accuracy and reliability. The Random Forest model has the highest overall accuracy of 98.53% and perfect precision, recall and F1-score for Severity Level 3, indicating that it is a highly reliable and accurate model for predicting accidents. The Decision Tree model also has a high overall accuracy of 95.33% and good precision, recall and F1-score values for all severity levels, indicating that it is also a reliable and accurate model.

On the other hand, the K-Neighbours Classifier has an overall accuracy of 96.01%, which is also quite high. However, its precision, recall and F1-score values are not as strong as the Random Forest or Decision Tree models for Severity Level 2 and Severity Level 4, which may indicate some limitations in its ability to accurately predict accidents for those severity levels. The Extreme Gradient Boost model has a lower overall accuracy of 92.65%, which indicates that it is less reliable and accurate compared to the other models. However, it still has good precision, recall and F1-score values for Severity Level 2 and Severity Level 3, indicating that it may still be a useful model for predicting accidents within those severity levels. Finally, the Naive Bayes model has the lowest overall accuracy of 63.31% and its precision, recall and F1-score values are significantly lower than the other models across all severity levels. This suggests that it may not be a reliable or accurate model for predicting accidents in this context.

Overall, the findings suggest that machine learning models can be used to accurately predict the likelihood of accidents, with Random Forest and Decision Tree models being the most reliable and accurate models in this context. However, it is important to note that the effectiveness of each model may vary depending on the specific dataset and context in which it is used, and further testing and validation may be required before these models can be implemented in real-world.

6 Discussion

This research set out to investigate the impact of weather and road conditions on accident severity and the potential of machine learning models in predicting accident likelihood. Despite the significant research done on predicting accidents through machine learning, most existing studies have primarily only focused on the model's overall accuracy for predicting accidents. Furthermore, most of the statistical analyses conducted so far have mainly relied on techniques like chi-square, correlation, or logistic regression. This thesis, however, aims to bridge these gaps by not just scrutinizing the overall accuracy of multiple machine learning models, but also taking into account other crucial performance indicators like precision, recall, and the F1 score. In addition to that, this thesis will enrich the chi-square analysis with supplemental statistical metrics, including p-value, degree of freedom, and Cramer's V, to explore the factors influencing accident likelihood, and to understand how weather conditions affect accident severity. The research questions also have been addressed comprehensively, leading to several important findings.

Regarding the first research question, the chi-square analysis with other statistical measures, such as p-value, degree of freedom, and Cramer's V confirmed that weather-related variables influence the severity of road accidents. This echoes previous findings that weather conditions have a profound impact on accident occurrence and severity. Specifically, visibility and temperature were identified as the most crucial factors. Our findings indicate that strategies such as implementing fog dispersal systems, heatwave alerts, or improved road maintenance during extreme temperatures may significantly reduce the severity of accidents. However, it's also important to emphasize that other weather conditions, while less associated, can still contribute to accident likelihood and severity. This suggests that a comprehensive approach, which includes driver education about various weather risks and appropriate safety measures, would be beneficial.

Turning to the second research question, our findings support the use of machine learning models to predict accident likelihood. This aligns with an increasing body of literature advocating the use of advanced computational methods in traffic safety analysis. The Random Forest model demonstrated the highest accuracy in our study, with an impressive overall accuracy of 98.53%, and excellent precision, recall, and F1-score for Severity Level 3. The Decision Tree model also performed well, confirming that these models are robust tools for predicting accidents.

Conversely, the K-Neighbours Classifier, while having commendable overall accuracy, showed some limitations for certain severity levels. The Extreme Gradient Boost model showed some reliability but had a lower overall accuracy, indicating it may not be the best choice in this context. Finally, the Naive Bayes model exhibited the lowest overall accuracy and less satisfactory performance matrix values, indicating its limitations in this context.

These findings underscore the potential of machine learning in traffic safety, but they also highlight the need for careful selection and validation of the model best suited for the specific context. The performance of machine learning models can vary based on the data and context, hence it would be beneficial to perform further testing before implementing these models in real-world scenarios.

In conclusion, this study demonstrates the impact of weather and road conditions on accident severity and the potential of machine learning models in predicting accident likelihood. These findings can contribute to enhancing road safety measures and

developing more accurate predictive models, ultimately contributing to the reduction of road accidents. Future research should consider other potential contributing factors, and continue to evaluate and refine machine learning models to ensure their accuracy and reliability in diverse contexts.

6.1 Limitation and area of improvement

It is essential to recognize the limits of our study, despite the fact that it does yield helpful insights. In the first place, the weather was the primary focus of our investigation, to the exclusion of any other possible elements that could have contributed to the severity of the accident. In order to gain a more complete knowledge of the occurrence, it is recommended that future studies take into account additional aspects such as the road infrastructure, the driver's behaviour, and the features of vehicles. The second limitation of our investigation is that it was based on a particular dataset, which restricts the applicability of our results. The relevance of our findings could be improved by conducting additional research using a variety of datasets originating from a variety of geographic regions. In conclusion, despite the fact that we investigated a number of machine learning models, it is possible that there are alternative models or approaches that could produce superior results when it comes to predicting the chance of an accident. Evaluating and contrasting the various alternative models would be an important focus for research in the future.

6.2 Potential application areas and examples

The insights that we gained from our research have practical applications that can be used to improve road safety measures and construct prediction models with a higher level of accuracy. The adoption of specialized driver education programs that place an emphasis on the dangers associated with driving in various weather conditions is one potential area for application. We can equip drivers to adjust their conduct appropriately in unfavorable weather conditions if we raise awareness of the issue and promote suitable safety measures. In addition, our findings imply that there is a requirement for the incorporation of meteorological information into intelligent transportation systems. Existing applications such as real-time navigation systems e.g. Google map, for instance, could be improved by adding meteorological conditions in order to deliver more precise and individualized routes while also taking into account the possibility of dangers and hazards.

7 Summary

The primary focus of this study was to investigate the impact of weather and road conditions on the severity of accidents and to determine the feasibility of machine learning models in accurately predicting the likelihood of such incidents. The research was centered on two key research questions.

Firstly, the study examined the influence of weather and road conditions on accident severity and identified the most related factors contributing to accidents. We utilized an open-source accident dataset, which was preprocessed using techniques like variable selection, missing data elimination, and data balancing through the Synthetic Minority Over-sampling Technique (SMOTE). Chi-square statistical analysis was performed, suggesting that all weather-related variables are more or less associated with the severity of accidents. Visibility and temperature were found to be the most critical factors affecting the severity of road accidents. Hence, appropriate measures such as implementing effective fog dispersal systems, heatwave alerts, or improved road maintenance during extreme temperatures could help reduce accident severity.

Secondly, the research evaluated the ability of machine learning models including decision trees, random forests, naive bayes, extreme gradient boost, and neural networks to predict accident likelihood. The models' performance was gauged using metrics like accuracy, precision, recall, and F1 score. The Random Forest model emerged as the most reliable and accurate model for predicting accidents, with an overall accuracy of 98.53%. The Decision Tree model also showed high overall accuracy (95.33%), indicating its reliability. However, the Naive Bayes model showed the lowest accuracy (63.31%) and was deemed less reliable in this context.

It is concluded that machine learning models can be effectively used to predict the likelihood of accidents, with models like Random Forest and Decision Tree proving the most effective. However, the effectiveness of each model may vary depending on the dataset and context, necessitating further testing and validation for real-world implementation.

These findings not only provide insight into the factors affecting accident severity but also open a promising avenue in employing machine learning techniques for proactive accident prediction and mitigation. Future studies can aim to refine the models further and potentially integrate them into traffic management systems to enhance road safety.

8 References

1. Han, S., Lee, W., Kim, M., & Lee, S. (2018). Accident detection and prevention using smartphone sensor data and machine learning. Sensors, 18(5), 1633.

2. Meegahapola, L., & Ochieng, W. Y. (2019). Real-time accident detection and management system using GPS and accelerometer sensors. IET Intelligent Transport Systems, 13(5), 794-802.

3. Li, W., Li, L., Li, Q., Li, H., & Li, S. (2020). An intelligent accident detection system based on GPS and gyroscope. Journal of Intelligent & Fuzzy Systems, 38(5), 6225-6237.

4. Zhu, Y., & Xiong, L. (2021). Vehicle rollover detection based on fusion of MEMS sensors and machine learning. Sensors, 21(3), 833.

5. Feng, C., Xu, H., Xu, H., & Wang, J. (2019). A novel accident detection algorithm based on an improved deep learning approach. IEEE Access, 7, 90780-90789.World Health Organization. (2020). Road traffic injuries. Retrieved from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

6. Khattak, A. J., Pawlovich, M. D., Souleyrette, R. R., & Hallmark, S. L. (2002). Factors related to more severe older driver traffic crash injuries. Journal of Transportation Engineering, 128(3), 243-249.
7. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18- 22.

8. M. Zheng, T. Li, R. Zhu, J. Chen, Z. Ma, M. Tang, Z. Cui and Z. Wang, "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network," in IEEE Access, vol. 7, pp. 62201-62210, 2019, doi: 10.1109/ACCESS.2019.2903319.

9. F. Zong, H. Xu, and H. Zhang, ''Prediction for traffic accident severity: Comparing the Bayesian network and regression models,'' Math. Problems Eng., vol. 2013, nos. 2–3, 2013, Art. no. 475194.

10. M. I. Sameen and B. Pradhan, ''Severity prediction of traffic accidents with recurrent neural networks,'' Appl. Sci., vol. 7, no. 6, p. 476, 2017.

11. F. N. Ogwueleka et al., "An artificial neural network model for road accident prediction: A case study of a developing country" (2014) *Acta Polytechnica Hungarica*, 11(5). Available at: https://doi.org/10.12700/aph.11.05.2014.05.11.

12. H., Chen, H., Zhao, Y., & Ma, X. (2020, November 16). *Critical Factors Analysis of Severe Traffic Accidents Based on Bayesian Network in China*. Critical Factors Analysis of Severe Traffic Accidents Based on Bayesian Network in China https://doi.org/10.1155/2020/8878265.

13. Yang, J., Han, S. and Chen, Y. (2023) "Prediction of traffic accident severity based on random forest," *Journal of Advanced Transportation*, 2023, pp. 1–8. Available at: https://doi.org/10.1155/2023/7641472.

14. Al-Mistarehi, B. W., Alomari, A. H., Imam, R., & Mashaqba, M. (2022, March 14). *Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS*. Frontiers. https://doi.org/10.3389/fbuil.2022.860805.

15. Dias, D., Silva, J.S. and Bernardino, A. (2023) "The prediction of road-accident risk through data mining: A case study from Setubal, Portugal," *Informatics*, 10(1), p. 17. Available at: https://doi.org/10.3390/informatics10010017.

16. Ali et al, "The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach". (2017, May 16. Volume 14, Pages 41-53) https://doi.org/10.1016/j.amar.2017.04.001.

17. Iranitalab, A. and Khattak, A. (2017) "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, 108, pp. 27–36. Available at: https://doi.org/10.1016/j.aap.2017.08.008.

18. Ye, F., & Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Analytic methods in accident research, 1, 72-85.

19. Jeong, H. *et al.* (2018) "Classification of Motor Vehicle Crash Injury severity: A hybrid approach for imbalanced data," *Accident Analysis & Prevention*, 120, pp. 250–261. Available at: https://doi.org/10.1016/j.aap.2018.08.025.

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

21. Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accident Analysis & Prevention, 90, 128-139.

22. Zhang, Z. *et al.* (2015) "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Transactions on Visualization and Computer Graphics*, 21(2), pp. 289–303. Available at: https://doi.org/10.1109/tvcg.2014.2350494.

23. Rodionova, M., Skhvediani, A. and Kudryavtseva, T. (2022) "Prediction of crash severity as a way of road safety improvement: The case of Saint Petersburg, Russia," *Sustainability*, 14(16), p. 9840. Available at: https://doi.org/10.3390/su14169840.

24. Anastasopoulos, P.C. *et al.* (2012) "A multivariate Tobit analysis of highway accident-injury-severity rates," *Accident Analysis & Prevention*, 45, pp. 110–119. Available at: https://doi.org/10.1016/j.aap.2011.11.006.

25. Zhang, J. *et al.* (2018) “Comparing prediction performance for Crash Injury severity among various machine learning and statistical methods,” *IEEE Access*, 6, pp. 60079–60087. Available at: https://doi.org/10.1109/access.2018.2874979.

26. Shi, X. *et al.* (2019) “A feature learning approach based on XGBoost for driving assessment and risk prediction,” *Accident Analysis & Prevention*, 129, pp. 170–179. Available at: https://doi.org/10.1016/j.aap.2019.05.005.
27. Assi, K. *et al.* (2020) “Predicting crash injury severity with machine learning algorithm synergized with Clustering Technique: A Promising Protocol,” *International Journal of Environmental Research and Public Health*, 17(15), p. 5497. Available at: https://doi.org/10.3390/ijerph17155497.

28. Mansoor, U. *et al.* (2020) “Crash severity prediction using two-layer ensemble machine learning model for proactive emergency management,” *IEEE Access*, 8, pp. 210750–210762. Available at: https://doi.org/10.1109/access.2020.3040165.

29. Dong, B., Yu, Z., Zhu, W., Xiong, Y., & Xiong, H. (2016). A smartphone-based system for car accident detection using accelerometer and gyroscope sensors. Sensors, 16(10), 1629.

30. Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. Computer networks, 52(12), 2292-2330.

31. Wang, L., Tan, Y., & Han, Z. (2017). Review of vehicle collision detection and automatic alarm technology. IOP Conference Series: Earth and Environmental Science, 85(4), 042035.

32. Kumar, P., Kumar, S., & Kumar, A. (2017). Design and development of vehicle accident detection and reporting system using GPS and accelerometer sensors. Journal of Engineering and Applied Sciences, 12(11), 2783-2787.

33. Luber, S., & Vasudevan, R. (2016). Collision detection using LIDAR and vision sensors. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 905-910). IEEE.

34. Shi, H., Li, S., & Zhang, Y. (2019). Vehicle collision detection system based on LiDAR sensors. Journal of Physics: Conference Series, 1335(5), 052022.

35. Hassan, M. S. Abbas, M. Asif, M. B. Ahmad and M. Z. Tariq, "An Automatic Accident Detection System: A Hybrid Solution," 2019 4th International Conference on Information Systems Engineering (ICISE), Shanghai, China, 2019, pp. 53-57.

36. N. H. Sane, D. S. Patil, S. D. Thakare, and A. V. Rokade, “Real Time Vehicle Accident Detection and Tracking Using GPS and GSM”, International Journal on Recent and Innovation Trends in Computing and Communication, 4(4), pp. 479-482, 2016.

37. Balfaqih, M. *et al.* (2021) “An accident detection and classification system using internet of things and machine learning towards Smart City,” *Sustainability*, 14(1), p. 210. Available at: https://doi.org/10.3390/su14010210.

38. Jackulin Mahariba, A., R., A.U. and Rajan, G.B. (2022) "An efficient automatic accident detection system using inertial measurement through machine learning techniques for powered two Wheelers," *Expert Systems with Applications*, 192, p. 116389. Available at: https://doi.org/10.1016/j.eswa.2021.116389.

39. Jeong, H. *et al.* (2018) "Classification of Motor Vehicle Crash Injury severity: A hybrid approach for imbalanced data," *Accident Analysis & Prevention*, 120, pp. 250–261. Available at: https://doi.org/10.1016/j.aap.2018.08.025.

40. Fernandes, B. *et al.* (2016) "Automatic accident detection with multi-modal alert system implementation for its," *Vehicular Communications*, 3, pp. 1–11. Available at: https://doi.org/10.1016/j.vehcom.2015.11.001.

41. Al-Ghamdi, A.S. (2002) "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, 34(6), pp. 729–741. Available at: https://doi.org/10.1016/s0001-4575(01)00073-2.

42. Chen, M.-M. and Chen, M.-C. (2020) "Modeling road accident severity with comparisons of logistic regression, decision tree and Random Forest," *Information*, 11(5), p. 270. Available at: https://doi.org/10.3390/info11050270.

43. Yan, M. and Shen, Y. (2022) "Traffic accident severity prediction based on random forest," *Sustainability*, 14(3), p. 1729. Available at: https://doi.org/10.3390/su14031729.

44. Aldhari, I. *et al.* (2022) "Severity prediction of highway crashes in Saudi Arabia using machine learning techniques," *Applied Sciences*, 13(1), p. 233. Available at: https://doi.org/10.3390/app13010233.

45. Berrar, D. (2019) "Bayes' theorem and naive bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology*, pp. 403–412. Available at: https://doi.org/10.1016/b978-0-12-809633-8.20473-1.

46. R, S.E. (2023) Understand random forest algorithms with examples (updated 2023), Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (Accessed: April 17, 2023).

47. Chen, T. and Guestrin, C. (2016) "XGBoost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Preprint]. Available at: https://doi.org/10.1145/2939672.2939785.

48. Brownlee, J. (2020). XGBoost With Python Mini-Course. Retrieved from https://machinelearningmastery.com/xgboost-with-python-mini-course/

49. Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA: MIT Press.

50. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.

51. Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer, New York, 2001 "The elements of statistical learning, data mining, Inference, and prediction.. no. of pages: XVI+533. ISBN 0-387-95284-5," *Statistics in Medicine*, 23(3), pp. 528–529. Available at: https://doi.org/10.1002/sim.1616.

52. Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

53. Moore, D.S. (2021) *Introduction to the practice of Statistics*. Freeman & Company, W. H.

54. Alzen, J.L., Langdon, L.S. and Otero, V.K. (2018) "A logistic regression investigation of the relationship between the learning assistant model and failure rates in introductory stem courses," *International Journal of STEM Education*, 5(1). Available at: https://doi.org/10.1186/s40594-018-0152-1.

55. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*. Hoboken, NJ: Wiley.

56. Chawla, N.V. *et al.* (2002) "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, pp. 321–357. Available at: https://doi.org/10.1613/jair.953.

57. Gibbons, J.D. and Chakraborti, S. (2021) *Nonparametric statistical inference*. Boca Raton, FL: CRC Press.

58. Brownlee, J. (2020). Why One-Hot Encode Data in Machine Learning? Machine Learning Mastery.

59. Real Python (2023) *Jupyter Notebook: An introduction*, *Real Python*. Available at: https://realpython.com/jupyter-notebook-introduction/ (Accessed: 17 May 2023).

60. Brownlee, J. (2020) *Why one-hot encode data in machine learning?*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ (Accessed: 17 May 2023).

61. National Geographic Society (2020) Meteorological models, National Geographic Society. Available at: https://www.nationalgeographic.org/activity/meteorological-models/ (Accessed: 17 May 2023).

62. Bishop, C.M. (2006) Pattern recognition and machine learning by Christopher M. Bishop. New York: Springer Science+Business Media, LLC.

63. Sari, M., Mutlu, Ö., & Zeytinoğlu, A. (2009). Effects of Human and External Factors on Traffic Accidents. Pamukkale University, Faculty of Science and Art, Department of Mathematics.

64. Singh, S. (2015). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.

65. *How Do Weather Events Impact Roads? - FHWA Road Weather Management*. Available at: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

66. Zou, Y., Zhang, Y. and Cheng, K. (2021) 'Exploring the impact of climate and extreme weather on fatal traffic accidents', *Sustainability*, 13(1), p. 390. doi:10.3390/su13010390.

67. Bergel-Hayat, R. *et al.* (2013) 'Explaining the road accident risk: Weather effects', *Accident Analysis & Prevention*, 60, pp. 456–465. doi:10.1016/j.aap.2013.03.006.

68. Finnish Meteorological Institute (2012) 'The effect of snowfall and low temperature on road traffic accident rates in Southern Finland'.

69. Islam, M., Alharthi, M. and Alam, Md.M. (2022) *The impacts of climate change on road traffic accidents in Saudi Arabia*. doi:10.31219/osf.io/2p5aj.

70. Zeng, Q. *et al.* (2020) 'Investigating the impacts of real-time weather conditions on freeway crash severity: A bayesian spatial analysis', *International Journal of Environmental Research and Public Health*, 17(8), p. 2768.

71. Khodadadi-Hassankiadeh, N. et al. (2020) The pattern of road accidents in fog and the related factors in north of Iran in 2014-2018 [Preprint]. doi:10.21203/rs.3.rs-73501/v1.

72. Sangkharat, K. *et al.* (2021) 'Determination of the impact of rainfall on road accidents in Thailand', *Heliyon*, 7(2). doi:10.1016/j.heliyon.2021.e06061.

73. Basagaña, X. *et al.* (2015) 'High ambient temperatures and risk of motor vehicle crashes in Catalonia, Spain (2000–2011): A Time-series analysis', *Environmental Health Perspectives*, 123(12), pp. 1309–1316. doi:10.1289/ehp.1409223.

74. *What is open?* Available at: https://okfn.org/opendata/ (Accessed: 30 May 2023).

75. Anderson, T.K. (2009) 'Kernel density estimation and K-means clustering to profile road accident hotspots', *Accident Analysis & Prevention*, 41(3), pp. 359–364. doi:10.1016/j.aap.2008.12.014.

76. Monzon, Andrés, Sara Hernandez, and Rocio Cascajo. (2012) "Real Time Passenger Information systems and quality of bus services." Proceedings of the 12th International Conference Reliability and Statistics in Transportation and Communication (RelStat'12), Riga Latvia, 1–10.

77. Lv, Y. *et al.* (2014) 'Traffic flow prediction with Big Data: A deep learning approach', *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–9. doi:10.1109/tits.2014.2345663.