National College of Ireland

Prediction of Accident Severity Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Dhruv Vimal Shah
Student ID: X21121087

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Dhruv Vimal Shah
Student ID:	X21121087
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	15/12/2022
Project Title:	Prediction of Accident Severity Using Machine Learning Algorithms
Word Count:	7106
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Dhruv Vimal Shah
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction Of Accident Severity Using Machine Learning Algorithms

Dhruv Vimal Shah
X21121087

Abstract

Vehicle accidents are among the most terrifying experiences a person can have, and at times leaves a lifelong mark on the victim. Accidents and collisions are a regular phenomenon, that keeps happening frequently as a direct result of the recklessness of drivers, road conditions, and other environmental conditions. Using crash severity prediction models, various government agencies can get insights into the variables that influence the incidence, allowing them to forecast the severity of an accident. With the aid of accident data, machine learning algorithms can help to find patterns that might help predict the severity of an accident, like fatalities, serious injuries, or just minor injuries. The fundamental purpose of the research was to provide a way to use machine learning algorithms to predict the level of damage caused by an accident. In this study, we made a prediction framework and used three different machine learning algorithms—random forest, logistic regression, and decision tree—to figure out how bad the accident's impact could be. This study carried out three experiments using a publicly available dataset collected from the Kaggle repository, originally released by the UK Department of Transport. Each of the algorithms was fine-tuned with hyperparameters to boost the classification prediction to gain the best possible results for the study. The random forest model was 86.23% accurate, the logistics regression model was 85.60% accurate, and the decision tree model was 86.23% accurate. The results demonstrate that Random Forest and Decision Tree were the best algorithms in terms of accurately predicting all three accident severity classes, as opposed to Logistic Regression, which only predicted the third class. Following the construction of the models, the results of the experiments were analysed with performance metrics such as accuracy, precision, recall, f1 score, and confusion matrix. With the use of random forests and decision tree algorithms, the proposed solution will help improve road safety and help the authorities in charge of road maintenance come up with plans for reducing accidents.

1 Introduction

Regardless of how often they occur, road accidents are among the most terrible things that can happen to a car driver. Most people who use roads are at least somewhat knowledgeable of the general safety measures and laws that must be followed when doing so. Nonetheless, accidents and crashes are still caused solely by the carelessness of road users. Accidents can be caused by a number of things, such as poor visibility, dangerous road conditions, or other drivers who aren't paying attention. Accidents can have a wide range of causes, but the injuries they inflict, the damage they do to property, and the

wreckage they leave behind in vehicles almost always share similar characteristics. Every year, there are thousands of road accidents that lead to loss of life, physical injuries, disabilities, and/or mental trauma, among other disastrous consequences. Visual impairments, precipitation, high winds, and temperature extremes are some ways that weather can impact driving skills, the performance of vehicles, surface friction on roadways, the infrastructure of roadways, the risk of collisions, and the flow of traffic. Due to the generally unpredictable nature of the weather, even the most cautious and experienced drivers run the risk of being involved in a collision. A driver's ability to see the road ahead may be hindered by several factors in addition to fog, such as other potential hazards. For example, when there is a heavy downpour, it might be difficult for vehicles to see the road well. Accidents can be brought about by unfavourable weather conditions such as sleet, snow, rain, fog, gusts of wind, or slippery pavement. After a rainstorm, when the roads are wet and slippery, there are more accidents than after any other severe storm. Extreme weather conditions can induce fluctuations in speed, which can both increase the chance of a traffic collision and make it more difficult for drivers to notice other people who are using the road. Because motor vehicle collisions are one of the top causes of death across the country, so improving road safety is of the utmost importance. Due to the growing number of cars on our roads, everyone must always be aware of their surroundings, know and follow all safety rules, and treat others with respect.

1.1 Research Question

"How accurately can machine learning predict the severity of an accident?"

1.2 Research Objective

In this line of research, we will use machine learning techniques like Random Forest, Logistics Regression and Decision trees. We need a solution that is proven and tested, can be implemented, and can be customized to solve the research question that was provided above. Moreover, the emergency services and insurance firms that deal with motor vehicle collisions would benefit directly from this technology. The police, fire services, and hospitals will receive fewer calls for emergencies if there are fewer accidents, which will also mean fewer claims for insurance companies to pay out.

1.3 Report Structure

The structure of the paper is broken down into the following sections: In Section 2 of this study, we present a summary of the relevant prior research. More specifically, our data mining approach is discussed in Section 3 of this paper. The research project's implementation is discussed in Section 4. The evaluation of the different approaches was presented in Section 5, where the best ones were ranked according to their accuracy, F1 score, precision, and recall. Finally, in Section 6, we present the conclusion and suggestions for further research.

2 Related Work

In this section, we will have a look at several studies that help in the process of acquiring domain knowledge by grasping and mastering a variety of research approaches that are

provided by researchers. This section is further broken down into the following subsections: 2.1 The forecast of accidents using machine learning; 2.2 The forecast of accidents using deep learning.

2.1 Prediction of accidents using machine learning

This research (Augustine and Shukla; 2022) recommends setting up an accident prediction system that can do things like aid in the analysis of potential safety concerns and tell you whether an accident will happen. This research relies heavily on the official accident records of a district in India for the years 2018–2020. Researchers compared many machine learning algorithms to find out which model is best at predicting accidents. This research applied several machine learning models to forecast future accidents. These models included logistic regression, random forest, decision tree, K-nearest neighbour, XGBoost, and support vector machines. The best performance came from the Random Forest algorithm, which attained an accuracy of 80.78%. However, KNN's accuracy of 65.17% was the lowest of all the algorithms tested. The study only made use of a relatively small portion of the available data, which covered the years 2018–2020. A large data set is required. The research paper presented by (Gowda; 2020) made use of predictions that were obtained via classification models such as ensemble logistic regression, random forest, ensemble (logistic, decision trees, SVM), XGBoost, and AdaBoost classifiers. These models were used to make the predictions. When correctly recognising accidents based on location and time, the algorithms Random Forest and XGBoost had the greatest performance, each with a 78% success rate. However, while it is an impressive method, this research has not yet reached its full potential in terms of being able to predict bad things that happen because of bad weather. (Najafi Moghaddam Gilani et al.; 2021) conducted a study on the severity of accidents by making use of machine learning algorithms such as multiple logistic regression and pattern recognition in artificial neural networks. These methods determined the most important factors in determining the frequency of accidents. The dataset includes reported incidents that occurred in Rasht's metropolitan areas during 2019 and 2020. The severity of an accident was worsened by factors like inadequate lighting, poor weather, and risky, low-quality vehicles. The accuracy of the model that uses machine learning is significantly higher (98.1%) than that of the logistic technique. The logistic regression model demonstrates that darkness in the sky and head-on collisions increase the likelihood of an accident due to the driver's reduced ability to see well and maintain cognitive attention. The study has some shortcomings, such as the fact that the data will only be collected for one calendar year, beginning in March 2019 and ending in March 2020, and that the researchers will only take into account predetermined weather conditions such as clear, cloudy, and rainy weather rather than real-time weather. The statistics on China's yearly traffic accidents are analysed in this study by (Zhang et al.; 2020) It does this with a method called the Long Short-Term Gradient Boosted Regression Tree for Memory Networks (LSTM-GBRT), which gives a model for predicting accidents. When compared to other models like the standard regression model, the normal back propagation neural network model, the LSTM neural network model, and the GBRT model, the LSTM-GBRT model fits the data the best. In addition to that, this research is limited by a few other factors. Environmental influences are not included in this analysis. Accidents involving vehicles on the road are notoriously difficult to anticipate, as their frequency is determined by a diverse set of contributing factors. Since it is hard to gather and evaluate information about the weather, it is not included

in this model either. This is because of the difficulty associated with doing so. (Wu and Wang; 2020) In the field of traffic studies, one of the most important things to study is how to predict tourist road accidents, and researchers have chosen China's excellent reputation for traffic safety as their case study. Based on the nonlinear nature of road traffic injuries in China, two ways to compare neural networks were made: one for support vector regression neural networks and the other for back propagation neural networks. Also, the size of the data from the matching records of road site visitors in China was cut down by figuring out what was most important. The authors of this research paper (Labib et al.; 2019) use machine learning to figure out how many traffic accidents there are in Bangladesh. In Bangladesh, there were 43,000 auto accidents between 2001 and 2015. Decision trees, KNN, Naive Bayes, and adaptive boosting (AdaBoost) were used to assess traffic accidents in Bangladesh. This study divided accidents into four categories: death, severe injury, minor injury, and vehicle collision. Eleven elements impacting most accidents in Bangladesh were selected as criteria to categorise the seriousness of every traffic accident into these four classifications. AdaBoost had the highest accuracy, at 80%. In the research they did in 2019, (AlMamlook et al.; 2019). looked at four machine-learning techniques for building precise classifiers. Logistic Regression (LR), AdaBoost, Naive Bayes (NB), and Random Forest (RF). The confusion matrix F1-score values demonstrate that the Random Forest outperformed the other models. This study found that random forest algorithms were 76% accurate in predicting accidents. This study lacks consideration for variables including the environment, traffic, pedestrians, and passengers. Each has an impact on the severity and frequency of accidents, but none have been put into practice due to a lack of data. Machine learning is used in the study by (Alagarsamy et al.; 2021). to alert travellers to dangerous regions. China's Taihuyuan served as the source of the dataset. "Black pots" are accident hotspots that have been found. Users who are travelling see alert information as a black area on a map. Users are alerted about dangerous or accident-prone areas. In high-risk scenarios, the combination of a random forest with a Gaussian distribution may save several lives. In hazardous areas, BF-GD issues warnings to travellers. The advised procedures provide alerts and save lives. The recommended strategy outperforms current ones in terms of accuracy, at 93.4%. There are gaps in the research, including how the weather was not considered. (Reddy et al.; 2022) This study investigates the causes of traffic accidents in the United States and makes recommendations for how to cut down on their incidence and severity. This study makes predictions about the severity of accidents that occur in Virginia by utilising logistic regression, K-nearest neighbours, and random forests. The accuracy provided by Random Forest was 82%. The data will be examined from several different perspectives. They need an analysis of the accident geography in the United States. Evaluate the frequency and severity of recent incidents involving concentrations. The weather component of the evaluation considers both intensity and climate. Analyze the factors that may be manipulated to either raise or decrease the number of accidents. In this study, the researchers used Python's Pandas module to determine whether the number of traffic accidents in the United States is on the rise or falling, as well as the times of day when they occur most frequently. Most accidents take place regardless of whether it is raining or not. The addition of additional algorithms and conditions observed in real-time proves the accuracy of future forecasts.

2.2 Prediction of accidents using deep learning

Deep learning-based innovative road traffic accident prediction using "convolutional neural networks" is proposed in this study by (Thaduri et al.; 2021). By considering the light, the weather, and the flow of traffic, it creates a traffic state matrix and a CNN model. In comparison to traditional machine learning, CNN's prediction method has both a lower rate of loss and a higher level of accuracy. The study solely looked at rain and snow; there was no consideration given to sunshine, darkness, wind, humidity, or any other factor. Research is required to make accurate predictions regarding accidents. This research Chandar et al. (2020) introduces a novel method that identifies geographical and environmental elements that influence road safety and predicts accident proneness based on these qualities. Time, weather, and location all have a role. They made their predictions about a road's safety index by using a graph neural network (GNN). On untrained samples, graph neural networks achieved an accuracy of 65%. The lack of success in achieving even more accuracy may be due to the unpredictability that is associated with human error in many occurrences. Because the study only achieved a moderate level of accuracy in the end, they need to continue working on the issue of accident prediction and try to enhance the accuracy rate. The research seems to have a lot of potentials, but it only has a moderate level of accuracy. (Mizan et al.; 2020) are developing an intelligent vehicle management system with the goal of reducing the number of automobile collisions. This prototype of an intelligent vehicle management system consists of three elements. To begin, a drowsiness detector will determine whether the driver is fatigued while they are behind the wheel. Any traces of alcohol that may be present in the body of the driver can be detected by an alcohol sensor. The overload detector of the vehicle will finally indicate whether it is overloaded. The study had certain shortcomings since it focused on factors such as drinking and driving, excessive tiredness, and overloaded cars rather than other factors that might contribute to accidents. This highlights the need for more studies to be conducted so that accidents caused by other situations can be predicted. (Viswanath et al.; 2021) conducted a study in which they investigated the relationships between traffic accidents, the quality of the roads, and the influence that environmental elements have on the probability of an accident occurring. The Apriori algorithm and Support Vector Machines were used in the construction of their accident prediction models, which were constructed via data mining strategies. This analysis made use of the datasets for the road accidents that occurred in Bangalore between 2014 and 2017 that were available online to the public. To have an accurate accident prediction system, the model must have many additional components added to it. The most important of these new factors should be the weather factor, as it is extremely important. This article by (Malik et al.; 2021) anticipated that the United Kingdom would sustain collision damage with ML. SMOTE was used to correct the class imbalance that existed in the dataset. The levels of accuracy for the Decision Tree, the Random Forest Bagging, and the Logistic Regression models are respectively 87.88%, 98.80%, and 84.12%. The ability of the framework to forecast crashes was increased by employing SMOTE and standardising the data. High winds and low visibility are the most common causes of accidents. Based on this data, 18.5% of the events took place during clear skies, whereas 5.7% took place during storms. Most accidents are caused by dry roadways. Statistics on traffic accidents and trends in their frequency are used to guide the development of new traffic legislation. Accuracy can be improved with techniques such as deep learning, representation learning, and logic and reasoning algorithms. This research does not

concentrate on improving model parameters or analysing specific weather conditions like cloudy, sunny, or rainy days. The report written by (Ma et al.; 2022) evaluates 2005–2015 UK automotive accident data to assess accident intensity. It shows that the most serious events occurred during the week, on one-way streets, roads with a speed limit of 30 kilometres per hour, and in broad daylight. Bayesian classifiers need accurate, exhaustive, and precise descriptions of the factors involved in a traffic accident. In the actual world, the likelihood of getting into a traffic accident can be increased by several factors, including the kind of road, the speed restrictions, and the weather. There are a few factors that affect the occurrences. This method of predicting real-world events considers a wide range of scene circumstances. The Bayesian approach may be used to model automobile collisions. Depending on the elements, complexity can be skilfully managed to varying degrees. The currently available studies do not cover all aspects of the traffic problem. The research by (Theofilatos et al.; 2016) uses a Logistic Regression Model to predict Traffic Accidents on the Road. It treats accidents as uncommon occurrences and applies corrective procedures to the resultant probabilities to account for the low frequency of positive samples (accidents). The model is restricted to using traffic data as its predictors (highways in the city of Athens). In developing their models, they used a variety of techniques, including logistic regression and the treatment of RTAs as an uncommon occurrence.

3 Methodology

The study is broken down into a series of phases that are structured according to the KDD process methodology (Knowledge Discovery in Databases). Because KDD has more of an emphasis on implementation than systems engineering, it is ideally suited for problems involving detection and prediction. The approach consists of multiple different processes, all of which are broken down into six distinct categories. Data selection, data pre-processing, exploratory data analysis, transformation, modeling, and evaluation. The figure below shows the flow of the KDD methodology.

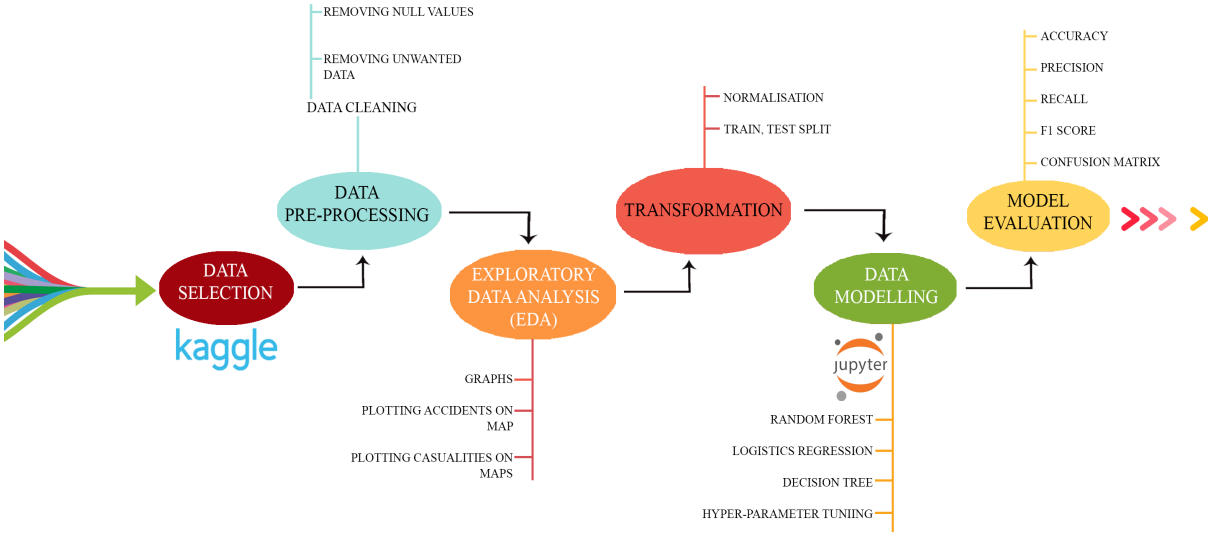

Figure 1: Flow of KDD Methodology

3.1 Data Selection

KDD begins with the identification of suitable data, which is subsequently subjected to data mining techniques. It is crucial at this stage to comprehend and develop the necessary project-plan knowledge. This will assist in the correct selection of data, as a core understanding of the domain will make it simpler to comprehend the columns of the dataset and eventually finalise the appropriate dataset. The data selected was the "UK Car Accidents 2005-2015" dataset from Kaggle[1]. The data was initially extracted from the United Kingdom's Department of Transport. Hence there was not any breach of any copyright issues. The data contains 3 files (Accident0515, Casualties0515, and Vehicles0515) in (.csv) format. Accident0515 is the main file, and through the Accident_Index column, it has links to Casualties0515 and Vehicles0515. The Accident0515 file comprises 1780653 rows and 31 columns. The Casualties0515 file contains 2216720 rows and 14 columns, and the Vehicles0515 file has 3004425 rows and 21 columns.

3.2 Data Pre-processing

The next step in the KDD process is preprocessing the data. This is done to get the best results from machine learning models. Before putting the models to use, it is important to clean and prepare the data so that the machines can understand it better. Data preprocessing involves figuring out if the data is categorical or numerical if there are any missing values, and if there are any outliers. After determining these variables, the data will be cleaned. This dataset has two types of missing values: "-1" and "Nan". After looking at each column, there are a total of 138 entries with missing values in the columns Location_Easting_OSGR, Location_Northing_OSGR, Longitude, and Latitude. There are 151 blanks in the column Time and 129471 blanks in the column LSOA_of_Accident_Location. Then, I ran an outer join query on the vehicles dataset so that all data frames could be joined. The data will then be cleaned up and set up for future work. This includes getting rid of the parameters Location_Easting_OSGR, Location_Northing_OSGR, LSOA_of_Accident_Location, Junction_Control, and 2nd_Road_Class.

3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis, more often known as EDA, is used in this section of the study to examine the data in detail. This analysis involves expressing the data in a diagram to make it easier to comprehend. EDA is utilized to investigate the data and compile a summary of the most significant discoveries. Finding trends or patterns may be accomplished with its assistance through the use of statistical insights and visualizations. To plot the visualizations, you'll need the packages seaborn and pyplot, which are both available in the matplotlib library. In this study, we did several EDAs on the data to find out how many accidents happened each day of the week. Even down to the hour of the day or night, we were able to establish the total number of accidents. The age group of the drivers involved in the accidents was also determined. The day of the week when the accident took place is depicted in the histogram in Figure 2. It indicates that Thursday was the day with the highest number of accidents. The following histogram in figure 3 presents data on the time of day and night when accidents took place. It appears that most incidents took place in the afternoon, specifically between 15 and 16

[1]Dataset Link: `https://www.kaggle.com/datasets/silicon99/dft-accident-data`

hours. It is safe to suppose that this hour of the day experiences the highest volume of moving traffic, such as people leaving their workplaces. The following part of the histogram in Figure 4 presents the number of persons in each age group that were part of the accidents. The histogram is broken down into 11 distinct age categories, represented by digits. According to the histogram, the majority of those who have been injured in accidents are between the ages of 26 and 35. In the course of my research, we have even made use of open street maps (OSM), which rely on the Folium package for their mapping capabilities. Folium is a robust Python module that assists in the development of a variety of different Leaflet maps. By default, Folium builds a map in a different HTML file. Because Folium's results can be changed, this library is a great place to find information for making dashboards. The function "folium.Popup" has been used to make a text output when clicking on an item on the map. The number of accidents in the UK may be seen in Figure 5. Based on the information provided by longitude and latitude, we can figure out which area has the most accidents. However, it is dependent on how much traffic there is in that particular region. The number of people killed or injured in each accident is displayed in the following Figure 6. It assigns distinct colour codes to the various types of casualties. The colour blue indicates that there was one person who was affected by that particular hotspot. The presence of the colour orange indicates that there were two people affected at the hotspot, while the inclusion of the colour red indicates that the location is extremely prone to accidents and that there were more than two people hurt there.

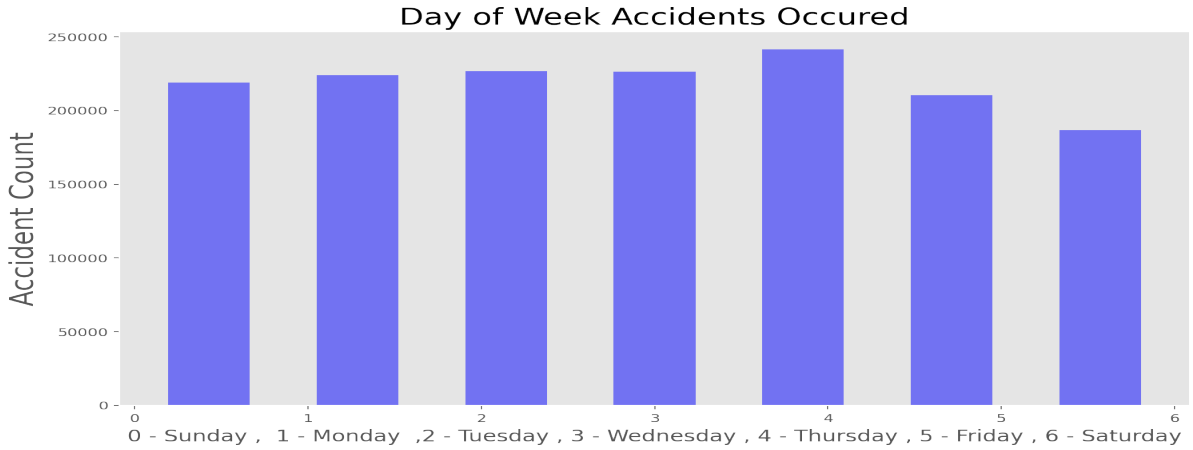

Figure 2: Day of Week Accident Occurred

3.4 Transformation

The process of KDD then advances to the next phase, which is data transformation. Data transformation is an important step in the process of gaining knowledge because it gets your data ready for modeling. The normalisation process is carried out in this research. There are just 2 columns that need to be normalised to ensure that our machine-learning algorithms are not adversely affected. In the dataset, the age of the drivers ranges from 18 to 88 years old, and we can standardise the data. In addition, the age of the car is likewise on a scale from 0 to 100, and it has the potential to skew the results of the machine learning model. However, we will normalise this predictor as well. Figure 7

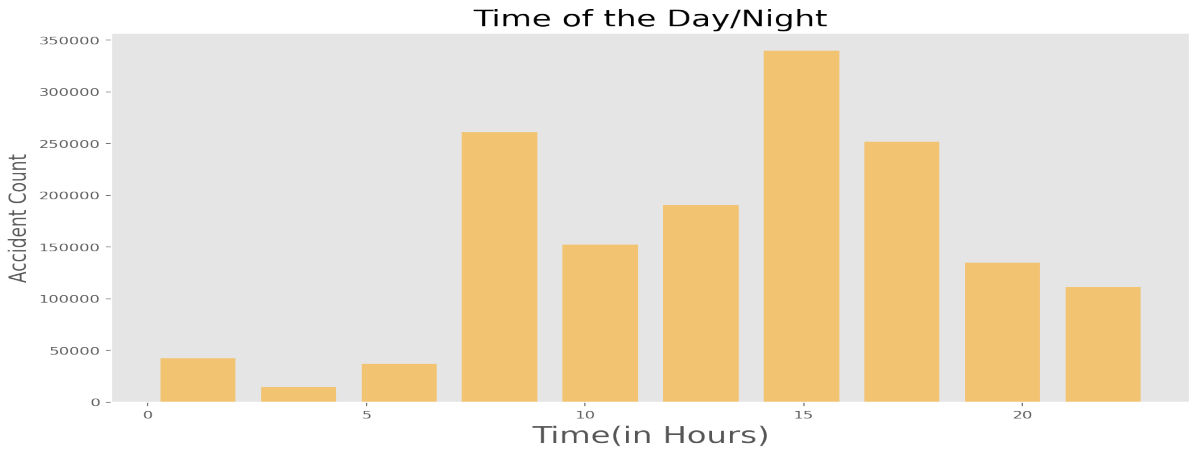

Figure 3: Time of Day/Night Accident Occurred

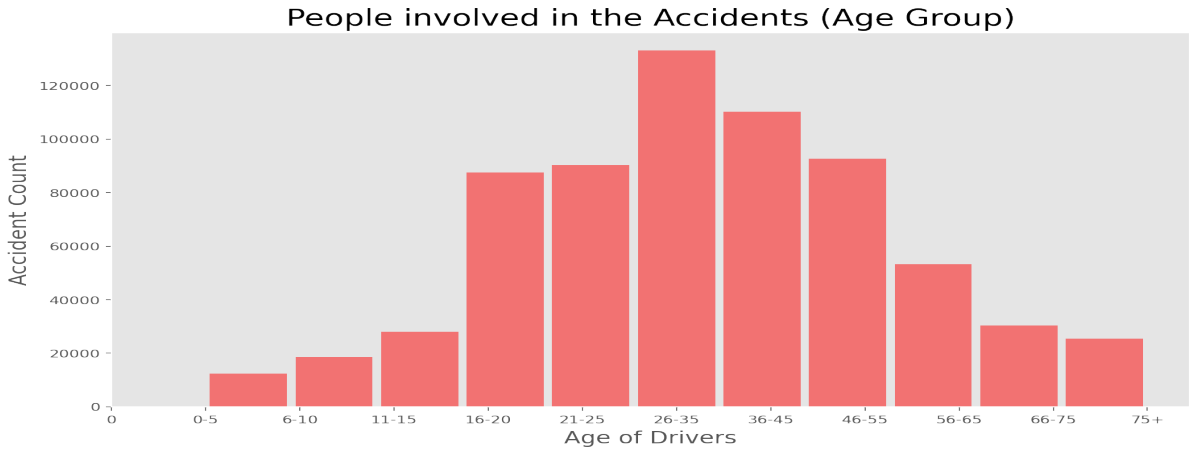

Figure 4: Age Group of people involved in Accidents

shows the before normalization of the factors (Age_of_Driver) & (Age_of_Vehicle). And Figure 8 shows after normalization of the factors (Age_of_Driver) & (Age_of_Vehicle). Also, the data is split into 80% training data and 20% test data with a random state of 99.

3.5 Data Modelling

Following the completion of the preprocessing of the dataset and the transformation of the data, the modeling phase will start. This part of the project is very important because it requires understanding a lot of the project's parts, being familiar with the literature reviews done by other researchers, and figuring out the best way to do things to get a good result. The following algorithms are going to be utilized to make an accurate prediction of the severity of an accident.

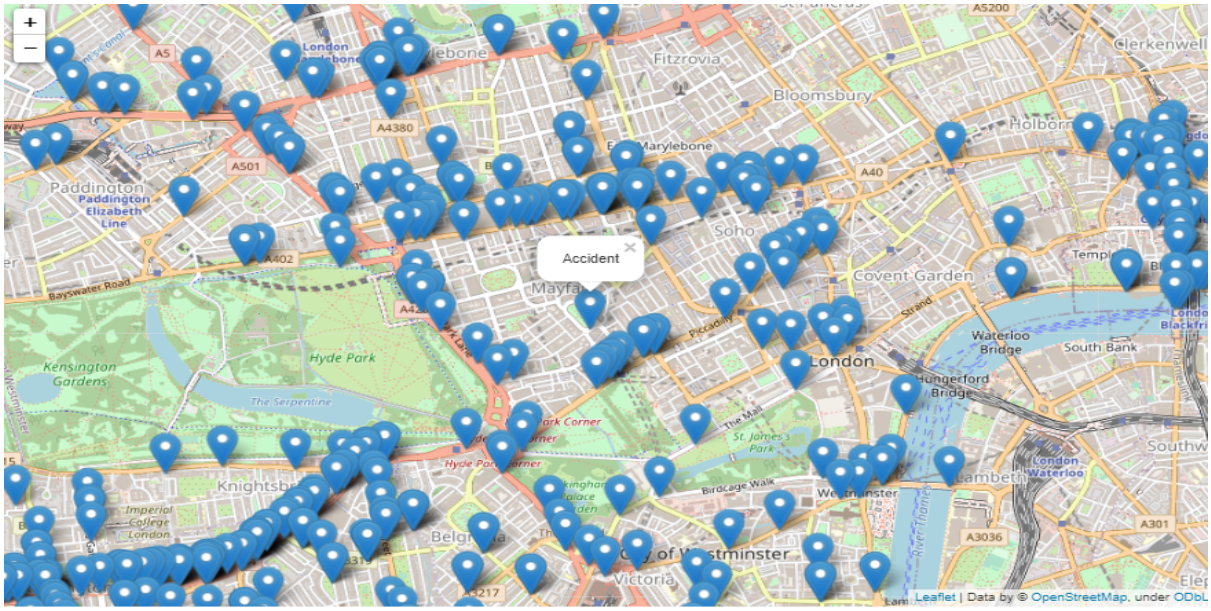

Figure 5: Accidents Plotted on Open Street Maps

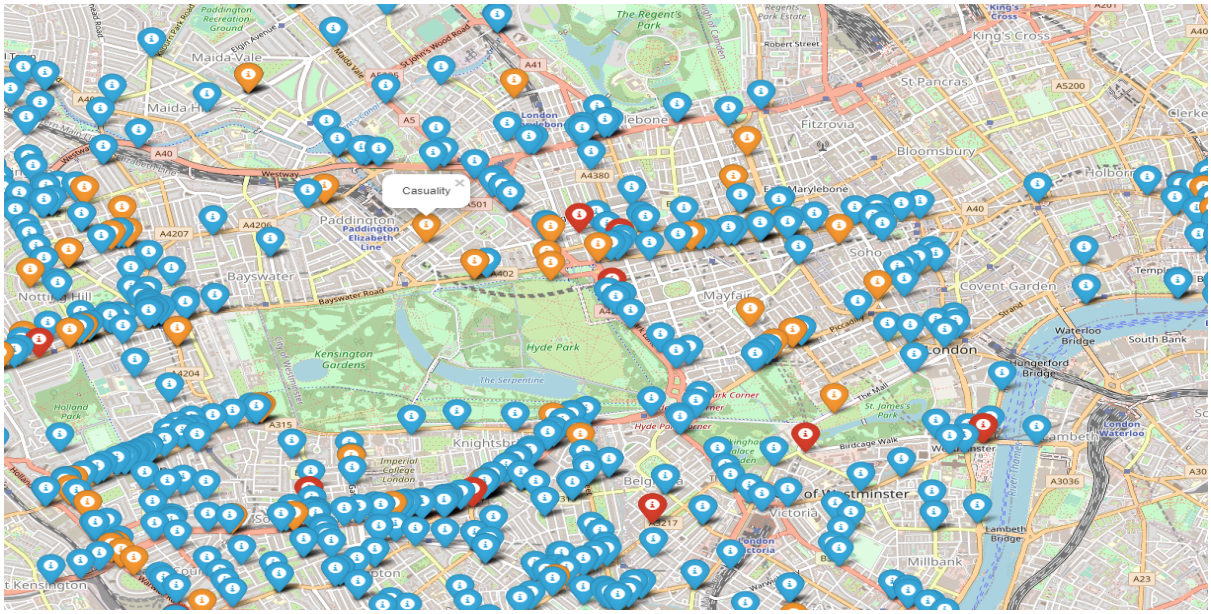

Figure 6: Causalities Plotted on Open Street Maps

3.5.1 Random Forest with & without hyperparamter tuning

Several independent decision trees combine to form a random forest, each of which provides an estimate of the likelihood of the dependent variable. After then, the ultimate result is determined by taking the average of all of the probabilities. Using this method, the first samples of a dataset are created by selecting data points using replacement. After that, decision trees are made by using only a few of the available input variables instead of all of them. In this particular model, the prediction was initially performed on the initial train-test sets, which were denoted by the notations "X train and Y train". To

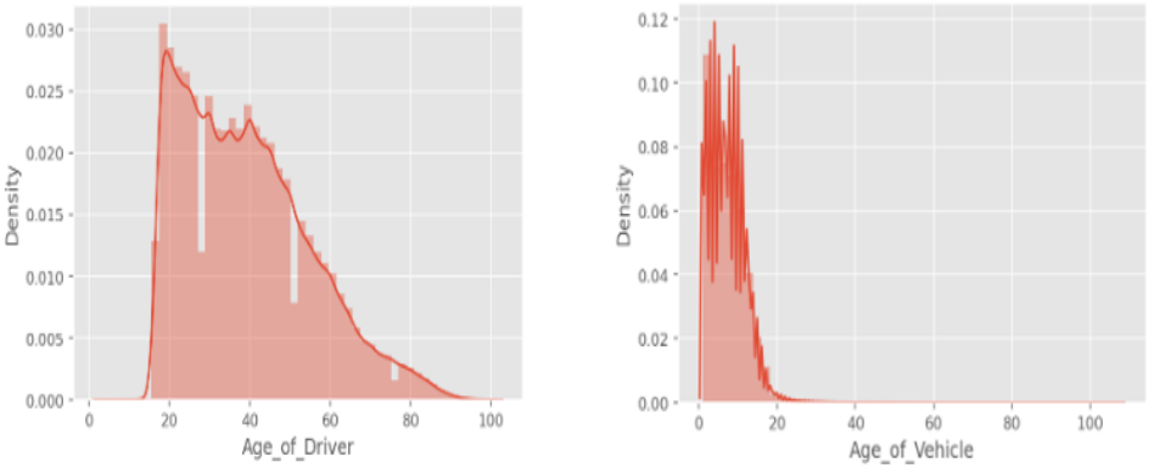

Figure 7: Before Normalization (Age_of_Driver) & (Age_of_Vehicle)

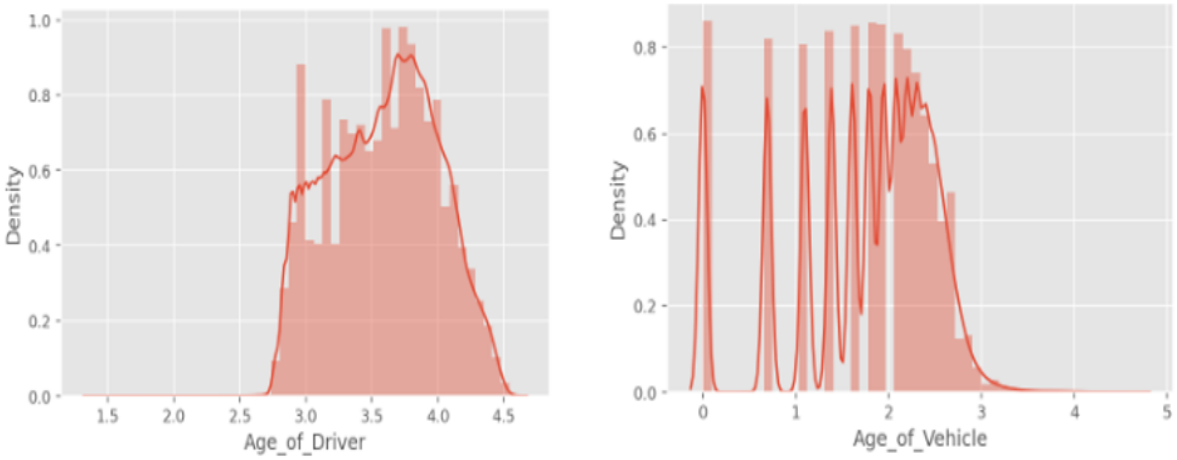

Figure 8: After Normalization (Age_of_Driver) & (Age_of_Vehicle)

accomplish this, the model was analysed after the 'RandomForestClassifier' classifier was brought in from the 'sklearn.ensemble' package. The hyperparameter tuning process was performed to improve the model's performance. The number of decision trees that make up a random forest as well as the number of characteristics that are taken into account by each tree when deciding how to divide a node are examples of hyperparameters in a random forest. Before beginning the training process for the model, certain hyperparameters should be modified so that the accuracy of the model may be improved. In this particular instance, the "RandomizedSearchCV" function from 'ScikitLearn' library was loaded, and a matrix of the hyperparameter values were specified.

3.5.2 Logistics Regression with & without hyperparamter tuning

Logistic regression is a type of supervised machine learning that uses past data to predict the expected outcome of tasks that involve binary classification. As the dependent variable of interest in the dataset is binary, logistic regression may be applied. The ability of logistic regression to manage a large number of features while still being effective in the prediction of an outcome with two possible outcomes is one of the method's primary benefits. After separating the training and test data, the logistic regression model was applied to the dataset, and the model was tested using 'LogisticRegression' from the sklearn library. The logistic regression was imported with the help of the function "sklearn.linear". Hyperparameter optimization was also done on the logistic regression model to make it more accurate. In this particular scenario, the "LogisticsRegressionCV" function that can be found in the ScikitLearn package was loaded, and a grid containing the values for the hyperparameters was provided.

3.5.3 Decision Tree with & without hyperparamter tuning

It's a tree-shaped classifier where each node represents a feature of a dataset and each branch indicates a set of decision rules that led to that feature being classified. A decision tree's "root" node acts as the starting point for making predictions about the dataset's category. To determine whether or not to proceed to the next node in the tree, the algorithm compares the elements of the root property with those of the record (the real dataset). From the sklearn library, the package known as "DecisionTreeClassifier" was brought in and used for this model. The model was applied to and evaluated on the original training and test sets. Even this model was tuned with hyperparameter adjustments.

3.6 Evaluation

A number of evaluation metrics, such as accuracy, precision, recall, the F1 score, and a plot of the confusion matrix, are used to compare the classification models that were built. This helps figure out how likely it is that the predicted classes will match the real classes.

3.6.1 Accuracy

Accuracy is the number of accurately predicted samples. It illustrates how well the prediction was accomplished regarding the values that were collected. The percentage

that was reached shows how well the model worked, which can be used to judge how well the classification works.

$$Accuracy = \frac{True\ Positive\ Classes + True\ Negative\ Classes}{Total\ Prediction\ Classes}$$

3.6.2 Precision

The term "precision" refers to the accuracy that is determined by considering only the classes that have a positive value. It provides information on the percentage of times a positive evaluation of a class is justified.

$$Precision = \frac{True\ Positive\ Classes}{True\ Positive\ Classes + False\ Positive\ Classes}$$

3.6.3 Recall

The ratio of accurately predicted positive classes to all observations in the actual class is used to calculate recall. The macro-averaged recall is taken into consideration when calculating recall for specific classes and averaging the results.

$$Recall = \frac{True\ Positive\ Classes}{True\ Positive\ Classes + False\ Negative\ Classes}$$

3.6.4 F1 Score

The F1 Score is the most significant measure to use in the evaluation. It is a type of measurement that combines precision and recall into one metric for evaluation. To account for this, both false positives and false negatives are taken into account when making this score. Despite this, F1 is almost always better than accuracy, especially when there is an uneven distribution of classes.

$$F1Score = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

3.6.5 Confusion Matrix

The confusion matrix is a table that summarises the number of correct and incorrect predictions produced by a classifier. It is used to determine how well a categorization model performs. You may evaluate the efficiency of a classification model by computing its accuracy, precision, recall, and F1 score. When both the observed and predicted values are positive, we have a true positive (TP). When both the actual value and the forecast come out to be negative, we get what statisticians call a "true negative." A false positive occurs when there is a discrepancy between the actual state of affairs and the predicted one (FP). The Type 1 mistake is another name for this. It's an example of a false negative, which occurs when the facts support a good outcome yet a negative prediction is made (FN). In other words, this is the same thing as Type 2 error.

4 Implementation

The dataset that will be used for this research comes from the online repository known as Kaggle. Three files in (.csv) format are available: Accidents0515, Casualties0515, and Vehicles0515. The primary file is called Accident0515, and it has connections to the Casualties0515 and Vehicles0515 files via the "Accident_Index" column. An 80:20 ratio is used to divide the data of the train to test. In this research project, 3 experiments are performed for deciding the appropriate model such as Experiment 1(Random Forest with without hyperparameter tuning), Experiment 2(Logistics Regression with without hyperparameter tuning), and Experiment 3(Decision Tree with without hyperparameter tuning). These three experiments correspond to three distinct models, each of which has a unique set of parameters. The output of the model may be modified to some extent by modifying various aspects of the model's parameters.

4.1 Environment Setup

This section contains a list of all of the tools and software that were used to complete the project successfully. This investigation was carried out using a machine having a 64-bit processor, the Windows operating system, and 8 GB of RAM. Python is the programming language that was utilised for the development of the models since it is capable of scripting and executing machine learning models within a web browser. Jupyter Notebook, version 6.4.5, which is supported by Anaconda, was used to carry out the code's execution. Python's most recent release (version 3.8.8) was utilised during the coding process.

4.2 Experiment 1 - Random Forest with & without hyperparameter tuning

Initially, a model was constructed by employing a random forest algorithm with default parameters and a n_estimator value of 200. This resulted in the generation of a model. After the model has been trained using the model's default parameters, it is trained once again with the train split to identify the parameters that produce the best results. After finding the optimal parameters, the model was hyperparameter tuned with the following parameters: 'bootstrap' = True, 'max_depth': [80, 90, 100, 110], 'max_features': [4, 5], 'min_samples_leaf': [5, 10, 15],'min_samples_split': [8, 10, 12], 'n_estimators': [100, 200, 300]. Following the completion of the training, the test set is then forecasted and evaluated using several classification evaluation metrics, including accuracy, precision, recall, the f1 score, and crosstabs.

4.3 Experiment 2 - Logistics Regression with & without hyperparameter tuning

The second model to be produced and put into use was a logistic regression, which was constructed with the default configuration. After the model has been trained using the model's default parameters, it is trained once again with the train split to identify the parameters that produce the best results. The model was hyperparameter tuned with the following parameters (cross-validation)'cv=3', random_state=0,multi_class='multinomial'. After the training is done, the test set is predicted and evaluated using a number of

classification assessment metrics, such as accuracy, precision, recall, the F1 score, and crosstabs.

4.4 Experiment 3 - Decision Tree with & without hyperparameter tuning

A decision tree was used as the model that was put into use in this. In the first step of the process, it was constructed using merely the default parameters. After the model has been trained using the model's default values, it is trained once again with the train split to determine the parameters that produce the best results. To obtain the best possible score, all that remains for us to do is find the optimal values for the least number of sample leaves and the maximum number of attributes. So the model was tuned with the 'min_samples_leaf=12', 'max_features=4'. Following the completion of the training, the test set is then predicted and evaluated using several classification assessment metrics, including accuracy, precision, recall, the f1 score, and crosstabs.

5 Evaluation

This section presents the metrics that were used for the evaluation. The most important part of the research is the evaluation measures because they show how well the model works. The various metrics of evaluation that were applied to this research and the findings of those measures are detailed below. The three models are assessed based on a number of different assessment measures, including accuracy, precision, recall, the f1 score, and the confusion matrix. Even the model that was ultimately used to accurately determine the severity of an accident is detailed in this section.

5.1 Experiment 1 - Random Forest with & without hyperparameter tuning

In the random forest model with default parameters, it is observed that it had an accuracy of 84.59%. Class 1 had a precision score of 5%, recall of 0 and f1 score of 1%. Class 2 had a precision score of 23.17%, recall of 5% and an f1 score of 9%. The last class of severity had a precision score of 86%, recall of 97% and an f1 score of 91%. Figure 9 shows the classification report of the following. After finding the optimal parameters, the model was hyperparameter tuned. Random forest took lots of time to tune the hyperparameter. The result of it is observed that it has an accuracy of 86.23%. Class 1 had a precision score of 0%, recall of 0 and f1 score of 0%. Class 2 had a precision score of 44.3%, recall of 2% and an f1 score of 3%. The last class of accident severity had a precision score of 86.4%, recall of 99.7% and an f1 score of 92.6%. Figure 10 below shows the classification report of the following. The accuracy of the random forest was increased after tuning it with a hyperparameter.

5.2 Experiment 2 - Logistics Regression with & without hyperparameter tuning

In the logistics regression model with default parameters, it is observed that it had an accuracy of 86.23%. Class 1 and 2 had a precision score of 0%, recall of 0% and f1 score

Accuracy 84.59

	precision	recall	f1-score	support
1	0.053114	0.007054	0.012454	4111
2	0.231720	0.056486	0.090831	38151
3	0.866542	0.972663	0.916541	264697
accuracy			0.845862	306959
macro avg	0.383792	0.345401	0.339942	306959
weighted avg	0.776748	0.845862	0.801808	306959

Predicted	1	2	3	All
Actual				
1	29	313	3769	4111
2	113	2155	35883	38151
3	404	6832	257461	264697
All	546	9300	297113	306959

Figure 9: Random Forest Classification Report & Crosstab Matrix

Accuracy 86.23

	precision	recall	f1-score	support
1	0.000000	0.000000	0.000000	4111
2	0.444380	0.020104	0.038468	38151
3	0.864674	0.997091	0.926173	264697
accuracy			0.862311	306959
macro avg	0.436351	0.339065	0.321547	306959
weighted avg	0.800857	0.862311	0.803439	306959

Predicted	2	3	All
Actual			
1	189	3922	4111
2	767	37384	38151
3	770	263927	264697
All	1726	305233	306959

Figure 10: Hyperparameter tuned Random Forest Classification Report & Crosstab Matrix

of 0%. The last class (3) of severity had a precision score of 86.23%, recall of 99.99% and an f1 score of 92.6%. Figure 11 shows the classification report of the following. After finding the optimal parameters, the model was hyperparameter tuned. Even after tuning it the accuracy, recall, precision and f1 score of all the classes remained the same. As we can see Logistic regression still didn't predict two classes of accident severity out of 3. Even though it is showing 86.2% accuracy. Figure 12 shows the classification report of the following.

Accuracy 86.23

	precision	recall	f1-score	support
1	0.000000	0.000000	0.000000	4111
2	0.000000	0.000000	0.000000	38151
3	0.862323	0.999928	0.926042	264697
accuracy			0.862258	306959
macro avg	0.287441	0.333309	0.308681	306959
weighted avg	0.743599	0.862258	0.798545	306959

Predicted	1	3	All
Actual			
1	0	4111	4111
2	4	38147	38151
3	19	264678	264697
All	23	306936	306959

Figure 11: Logistics Regression Classification Report & Crosstab Matrix

Accuracy 86.23

	precision	recall	f1-score	support
1	0.000000	0.000000	0.000000	4111
2	0.000000	0.000000	0.000000	38151
3	0.862319	0.999989	0.926065	264697
accuracy			0.862311	306959
macro avg	0.287440	0.333330	0.308688	306959
weighted avg	0.743595	0.862311	0.798565	306959

Predicted	1	3	All
Actual			
1	0	4111	4111
2	0	38151	38151
3	3	264694	264697
All	3	306956	306959

Figure 12: Hyperparameter tuned Logistics Regression Classification Report & Crosstab Matrix

5.3 Experiment 3 - Decision Tree with & without hyperparameter tuning

In the decision tree model with default parameters, it is observed that it had an accuracy of 75.36%. Class 1 had a precision score of 3%, recall of 4% and f1 score of 3%. Class 2 had a precision score of 16%, recall of 18.8% and an f1 score of 17.3%. The last class of severity had a precision score of 87.1%, recall of 84.6% and an f1 score of 85.8%. Figure 13 shows the classification report of the following. After finding the optimal parameters, the model was hyperparameter tuned. The result of it is observed that it has an accuracy of 85.69%. Class 1 had a precision score of 15%, recall of 0 and f1 score of 0%. Class 2 had a precision score of 31.6%, recall of 4% and an f1 score of 7.7%. The last class (3) of accident severity had a precision score of 86.6%, recall of 98.7% and an f1 score of 92.3%. Figure 14 shows the classification report of the following. The accuracy of the random forest was increased after tuning it with a hyperparameter.We didn't find much of a distinction between Accident Severity classes 1 and 2. Nevertheless, we were able to increase the accuracy of the severity of the class 3 accident severity. The accuracy rate increased from 75.1% to 85.69% as a result.

Accuracy 75.36

	precision	recall	f1-score	support
1	0.034483	0.042569	0.038101	4111
2	0.160482	0.188750	0.173472	38151
3	0.871364	0.846069	0.858531	264697
accuracy			0.753612	306959
macro avg	0.355443	0.359129	0.356701	306959
weighted avg	0.771803	0.753612	0.762399	306959

Predicted	1	2	3	All
Actual				
1	175	907	3029	4111
2	918	7201	30032	38151
3	3982	36763	223952	264697
All	5075	44871	257013	306959

Figure 13: Decision Tree Classification Report & Crosstab Matrix

5.4 Discussion

The fundamental objective of this research is to design an algorithm for machine learning that is capable of forecasting the extent of damage that will be caused in an automobile accident in the future, with the expectation of achieving a safer driving environment.

Accuracy 85.69

	precision	recall	f1-score	support
1	0.153846	0.000973	0.001934	4111
2	0.316212	0.044376	0.077830	38151
3	0.866592	0.987340	0.923034	264697
accuracy			0.856932	306959
macro avg	0.445550	0.344230	0.334266	306959
weighted avg	0.788642	0.856932	0.805650	306959

Predicted	1	2	3	All
Actual				
1	4	329	3778	4111
2	3	1693	36455	38151
3	19	3332	261346	264697
All	26	5354	301579	306959

Figure 14: Hyperparameter tuned Decision Tree Classification Report & Crosstab Matrix

This has the potential to give instant benefits to emergency services as well as to insurance companies that are dealing with such circumstances. If there are fewer accidents, then fewer people will have a need to contact emergency services like the police, the fire department, and hospitals. This will also result in fewer claims being filed against insurance companies, which will reduce the amount of money those businesses have to pay out. This goal was successfully accomplished with the help of the research study that was conducted. On the other hand, this research initially presented a number of challenges to overcome. The investigation utilised several different types of models, including random forests, logistic regression, and decision trees. The models were divided into two groups for the train test with a ratio of 80:20. In addition to that, hyperparameter tuning was performed on each of the models. As can be seen, the logistic regression method performed rather well in terms of accuracy. When we examine the confusion matrix in further detail, it is very clear that the Decision Tree algorithm performed far better, with an accuracy of 85.69%. The Random Forest model failed to forecast the first class, but it predicted the second and third classes (the majority class) quite well, indicating that the model outperformed the others. The logistic regression model was unable to predict the first and second-class outcomes, but it did predict the third class and had an overall accuracy rate of 86.23%. This indicates that it only predicted the major class (3rd class) and that the majority class, which is dominating, mislead the accuracy. Because the minority class is important in logistics regression, the class imbalance caused a model to favour strong recall, resulting in a lower F1 score. In the decision tree algorithm, the model accurately predicted each of the three classes and didn't affect the F1 score. The accuracy, precision, recall, and f1 score are some of the assessment metrics that are utilised in order to assess the quality of the output produced by these algorithms. Therefore, random forests and decision trees proved to be the most effective prediction algorithms for the severity of an accident.

6 Conclusion and Future Work

The research's main goal is to develop a method for employing machine learning algorithms to forecast the extent of damage caused by an accident. The three methods of machine learning that were used were random forest, logistical regression, and decision trees. Even the hyperparameters have been fine-tuned across the board for every model. The data used for this research was fetched from Kaggle. The dataset that was used is rather extensive. The results of the study would cut down on the number of accidents

that happen and give emergency services and insurance companies that deal with these kinds of situations an immediate advantage. If there are fewer casualties, then fewer people will have a need to contact emergency services like the police, the fire department, and hospitals. This will also result in fewer claims being filed against insurance companies, which will reduce the amount of money those businesses have to pay out. The results of the study showed that the Random Forest model and the Decision Tree model were much better at predicting the different classes of accident severity than any of the other models. These models have received a really high accuracy such as a random forest with an accuracy of 86.23% and a decision tree with an accuracy of 85.60%. These are the best algorithms that can be used to correctly predict the severity of an accident. The report was divided into six different sections. The first section provided an introduction to the research topic, while the second section discussed relevant work that had been done in the specific field. In Section 3, we covered the KDD approach that was applied during the course of the research. The following section, number 4, details the actions that were taken to implement the model. The following component, number 5, details the evaluation metrics that were used to evaluate the model. Finally, the last section provided a conclusion and discussed future research that needs to be done in the domain. The future scope will have a more narrow emphasis and may be augmented with other predictors like the density of population, volume of traffic, number of stores, number of tourist attractions, and so on. Even the addition of real-time weather data could be quite beneficial in the future, as it would allow drivers to get accurate information about their route that also took into account the impact that the weather would have on their travel.

References

Alagarsamy, S., Malathi, M., Manonmani, M., Sanathani, T. and Kumar, A. S. (2021). Prediction of road accidents using machine learning technique, *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1695–1701.

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

Augustine, T. and Shukla, S. (2022). Road accident prediction using machine learning approaches, *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 808–811.

Chandar, S., Reddy, A., Mansoor, M. and Jamadagni, S. (2020). Road accident proneness indicator based on time, weather and location specificity using graph neural networks, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1527–1533.

Gowda, M. L. (2020). *Traffic accidents prediction using ensemble machine learning approach*, Master's thesis, Dublin, National College of Ireland.

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K. and Nawrine, F. (2019). Road accident analysis and prediction of accident severity by using machine learning in

bangladesh., *2019 7th International Conference on Smart Computing Communications (ICSCC), Smart Computing Communications (ICSCC), 2019 7th International Conference on* pp. 1 – 5.

Ma, M., Xie, S. and Jia, P. (2022). A predictive analysis model research based on the bayesian classifier: Taking united kingdom vehicle accidents as an instance, *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, pp. 372–379.

Malik, S., El Sayed, H., Khan, M. A. and Khan, M. J. (2021). Road accident severity prediction — a comparative analysis of machine learning algorithms, *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, pp. 69–74.

Mizan, A. M., Tahmidul Kabir, A. Z. M., Zinnurayen, N., Abrar, T., Ta-sin, A. J. and Mahfuzar (2020). The smart vehicle management system for accident prevention by using drowsiness, alcohol, and overload detection, *2020 10th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, pp. 173–177.

Najafi Moghaddam Gilani, V., Hosseinian, S. M., Ghasedi, M. and Nikookar, M. (2021). Data-driven urban traffic accident analysis and prediction using logit and machine learning-based pattern recognition models., *Mathematical Problems in Engineering* pp. 1 – 11.

Reddy, S. S., Chao, Y. L., Kotikalapudi, L. P. and Ceesay, E. (2022). Accident analysis and severity prediction of road accidents in united states using machine learning algorithms, *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–7.

Thaduri, A., Polepally, V. and Vodithala, S. (2021). Traffic accident prediction based on cnn model, *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1590–1594.

Theofilatos, A., Yannis, G., Kopelias, P. and Papadimitriou, F. (2016). Predicting road accidents: a rare-events modeling approach, *Transportation research procedia* **14**: 3399–3405.

Viswanath, D., K, P., R, N. and R, B. (2021). A road accident prediction model using data mining techniques, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1618–1623.

Wu, D. and Wang, S. (2020). Comparison of road traffic accident prediction effects based on svr and bp neural network, *2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA)*, Vol. 1, pp. 1150–1154.

Zhang, Z., Yang, W. and Wushour, S. (2020). Traffic accident prediction based on lstm-gbrt model, *Journal of Control Science and Engineering* **2020**.