# Navigation System To Avoid Accident Prone Areas Using Machine Learning Techniques

MSc Research Project
Data Analytics

Saikrishnan Murali
Student ID: 20217200

School of Computing
National College of Ireland

Supervisor: Prashanth Nayak

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Saikrishnan Murali |
| **Student ID:** | 20217200 |
| **Programme:** | Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prashanth Nayak |
| **Submission Due Date:** | 15/08/2022 |
| **Project Title:** | Navigation System To Avoid Accident Prone Areas Using Machine Learning Techniques |
| **Word Count:** | 6923 |
| **Page Count:** | 21 |

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:** | |
| **Date:** | 18th September 2022 |

## 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): | |

# Navigation System To Avoid Accident Prone Areas Using Machine Learning Techniques

Saikrishnan Murali
20217200

**Abstract**

One of the most common ways in which individuals are affected is by road traffic accidents, which are more likely to occur in areas with excessive traffic or poor roadway management. The best way to avoid accidents is to avoid these regions and travel along a safer route. The aim of this research is to develop a navigation system using machine learning that avoids accident-prone areas. The research data comprises of daily accident reports with location details acquired from the City of Chicago website, which is an open repository. By using classification models such as Logistic Regression, Decision Trees, and Random Forests, the severity of the accident can be determined. The optimal model determined is Random Forest which has a accuracy, precision, and recall, scores of 94%, 94% and 91% respectively. Once the severity is determined, the parameters that have the greatest effect on severity will be identified. The risk score for each location will be determined on the basis of those parameters, using multiple linear regression. Using the KNN clustering technique, distinct areas or clusters will be identified on the basis of the co-ordinates and the risk score of the location. The machine learning model's outputs will be fed into the open route service (ORS), and depending on the mode of transportation, the route map, instructions, and time required for each stage of the journey will be provided. The model developed in this research project can be implemented in real time to ensure people travel safely and minimize traffic accidents.

**Keywords**: Accidents, Logistic Regression, Decision Tree Classifier ,Random Forest Classifier, Severity, Multiple Linear Regression, Risk Score, K-Means Clustering, Accident Prone Area, Navigation System.

## 1 Introduction

In recent years, industrialization and growth have contributed significantly to the production of vehicles at a decent price, resulting in an increase in road users, which increased the chances of them being careless, leading to a rise in traffic fatalities and injuries.It is estimated that in the year 2016, around 1.35 million people have died in traffic accidents. After considering the role that traffic accidents play in the global population, it becomes clear that they are among the most common causes of death. Accidents are sudden and unexpected occurrences; one way to assess their severity is by measuring their effects as injuries, losses, or property damage. By combining various machine learning approaches, it is possible to assess both the severity of an accident and to create a model that can calculate a risk score for a location based on the severity.

Several studies such as the work of Zou et al. (2021), take a selective approach to accident data with an emphasis on severity of fatal accidents, however, this study shall examine all accidents that occur on a day-to-day basis. Additionally, little research has been conducted on the relevance of utilizing the risk score to identify locations prone to accidental events rather than only focusing on the frequency of road traffic accidents and the factors leading to them such as the work of Klinjun et al. (2021), which uses the Haddon matrix[1] to identify the risk factors. Unlike previous research, this study suggests using supervised machine learning techniques to estimate the severity of an event and the risk score of a location, as well as using unsupervised machine learning to identify clusters or areas based on the proximity of locations.

The objective of this study is to provide a solution to the research question ***Can a risk score based on machine learning be used to provide safer travel routes to an user?*** The following sets of particular study objectives were developed to answer the research topic.

- ***To design:***
  - A machine learning model to identify the severity using classification techniques such as Logistic , Decision Tree and Random Forest.
  - Using Multiple linear regression to create a risk score based on the factors that determine the severity of the accident.
  - Use K-Means clustering algorithm to classify the locations involved in the data into a specific group based on the computing the distances between them.
  - Build a navigation system extracting road attributes from map data and integrating them with accident prone zone information, which were identified in the study using the risk score.
- To demonstrate the implementation of the above design, three datasets which describe each traffic crash in the year 2021 is used,which is available on the open Chicago repository.
- The classification models will be evaluated using the classification report which shows the precision, recall and f1-score and regression models will be evaluated based on R Square/Adjusted R Square ,Mean Square Error(MSE)/ Root Mean Square Error(RMSE) and Mean Absolute Error(MAE)

The majority of the study focuses on applying machine learning technology to detect accident prone locations, but for the navigation system, the website openrouteservice [2] or ORS is used, which provides route service by using the Open Standards and Open Geodata.For the purposes of this research, ORS's free API provides global spatial services by utilizing OpenStreetMap's geographic data. The study will allow users to take a risk-free route based on data compiled from accidents that occur every day.

[1] The Haddon Matrix is the commonly used in the injury prevention sector for treatment and preventive strategy development

[2] https://openrouteservice.org/

The remainder of the report is divided into sections that provide a framework for the project. The Related works section gives us an understanding of the research objective by providing a comprehensive review of the published literature that uses or is comparable to the background of the study. The Methodology and Design Specification describes the overall techniques used in the research and helps us in understanding the approach and the overall architecture of the model.The model Implementation helps in implementing the model that will provide the solution to the research problem and the discussion based on its results is explained in is made upon the result is explained in and Evaluation.The final observation and future works the research will be covered in the Conclusion .

# 2 Related Work

## 2.1 Introduction

The literature review section includes previous research publications on accident prediction and analysis, utilizing various machine learning techniques. This section is divided into subdivisions that are linked to the methodologies and procedures applied in the study.

## 2.2 Analysis of the Factors in Traffic Accidents

This section focuses on understanding the potential factors or parameters available in the research data, such as weather and temperature on the day of the accident, which are necessary. Since the current research also focuses on identifying the different ways in which columns and their corresponding values can be reduced or cleaned, this section focuses on the analytical methods or processes by which they are detected. The research works done by Pervez et al. (2021), focuses on Identifying Factors Contributing to accidents that are taken place due to motorcycles and the columns details consists of information which are related to demographic features of the victims, the vehicles involved, the cause and type of accident, the time and location of the crash. The initial method followed is to perform descriptive analysis to determine the values count percentage of each columns for determining the missing values and in this research there are none. In the research work performed by Zhu (2021), the data included in the study is connected to a vehicle-bicycle collision. Since the study will be conducted using the gradient boosting approach, the data should not contain unbalanced classes, hence the data re-sampling process is used. As a result, the original data is re-sampled using the Synthetic Minority Over-sampling Technique (SMOTE). This is done to deal with the unbalanced class data, but the problem of using this approach is that when the data is balanced, it may discard useful insights and over-sample the data, which may be required for developing rule-based classifiers such as Random Forests. Therefore this technique is not used in the current research where classifiers are used. In the researches done by Stigson et al. (2021) and Yan, Chen, Wang, Zhang and Zhao (2021), the data was classified according to various factors which were related to the gender, age, accident type and injury type. In the former research technique the chi-squared test was performed to determine the proportional difference between the categories and the latter used Fish-bone diagram of the causal factors contributing to the accident. In order to understand the reported injuries in the data, the columns were reviewed based by both the research, on the nature of accident and the analysis revealed with the various types of injuries. Therefore this type of analysing the columns and the values are used in the current research to remove redundant and useless columns. One of the main issue is the method of handling missing data which is the main focus in the research paper Elhassan et al. (2022). The method of Inductive Learning Algorithm is used by framing rules or condition based on which the data can be converted into two possible classification values. Therefore when a missing value is present instead of removing it is replaced with in the existing values which are present in the similar combination. These data processing procedures are utilized in accordance with the current research's application to generate data needed for implementing the machine learning techniques.

## 2.3 Prediction of Accidents using Supervised Machine Learning

The machine learning techniques used in the research involve both classification and regression techniques. Therefore the inclusion of research paper involving all of them together and separately is necessary. The research paper Castillo-Botón et al. (2022), involves the fog events prediction where the machine learning technique which can be implemented as both regressor and classifier are used such as Random forest is used as regression or decision trees. The methodology used in the research has divided the problems involved into regression and different classification tasks. prediction which involves in determining the level of visibility are assigned to regression and classification problem involves the use of identifying the hyper-parameters which is done using grid search. Research using classifiers and regression techniques will be discussed in the following subsection

### 2.3.1 Prediction using Classification Techniques

The research paper Yan, He, Zhang, Liu, Qiao and Zhang (2021), involves the crash risk analysis, where the machine learning approach is centered on the usage of tree-based and non-tree-based models, and therefore a total of 10 classification models are used to predict the crash severity. The tree based models are used for implementing the classification trees to identify severity of the accident and emphasizes the usage of larger data for improving the threshold at node level. The Other non tree models involve the use of directions and distance by converting the user data to a linear and vector subspace.When the training accuracy of tree-based models was compared to that of non-tree-based models, the average training accuracy of tree-based models was 99.37% and latter was slightly less with 99.11% Another research paper Tamakloe et al. (2021), focused to determine the factors impacting buses, which are utilized as public transportation in the majority of countries. Factors such as the road quality and the lighting situation on the day of the accident are taken into account and they are further classified as good and bad. To investigate the influence of different factors on severity outcomes, the research focuses on heterogeneity and distinct models such as the conventional ordered Logit model and the random parameters ordered Logit model.

### 2.3.2 Prediction using Regression Techniques

The use of regression is explored in the study Li et al. (2021), to determine Traffic flow where the emphasis is on recreating data from estimated travelling time and traffic flows using machine learning algorithm Gaussian Process Regressor.By recording the travel time with respect to the travel flow on several days using the google maps. The process works on using the regressor to estimate traffic flows from travel duration and therefore

Construct and train multi-models to reduce estimated traffic flow error. The model's output is then compared to that measured by real-world sensors in order to assess the performance the GPR trained models. Similar to the research study on classification models to predict severity, the research performed in Hong et al. (2021), is to analyse it severity of a traffic accident and the factors involved in it. The study's aim is to apply multiple linear regression and a multi-collinearity test to determine the relationship between traffic accident severity and physical and environmental accident parameters. The regression technology utilized in this study is R-studio programming language that gives the option of performing step-wise variable selection to pick the influencing variable and performing multiple linear regression with the dependent variables.

## 2.4 Unsupervised Cluster Analysis of Spatial Data

The unsupervised models are used to identify a pattern or a design based on the analysis of the data unlike the earlier models where a target variable is predicted. Researchers Islam et al. (2021) and Dadwal et al. (2021), conducted a study on clustering algorithms to analyze road traffic crashes, with the former utilizing conventional techniques such as K-mean,Mini Batch K-means,DBSCAN,OPTICS and the latter applying an Adaptive Clustering analysis. Despite the fact that both were used to analyze road accidents, the approach is different. The adaptive clustering methodology predicts traffic accidents using numerous temporal and geographical characteristics, whereas the latter use data based on real-time comparative performance analysis with the focusing parameter being distance. The results in the conventional research technique is done based on the time taking for simulation and internal cluster validation. The structural comparison shows that the DBSCAN and OPTICS are better than the K-means but the simulation duration is quite long, and it grows as more points are provided. The adaptive clustering analysis is tested by comparing it to the standard model such as DBSCAN and exhibits a 4% improvement in performance. Another study that performed by Abdullah et al. (2022), is K-Means clustering to determine the areas that were affected by Covid-19 in the country of Indonesia. The object of the study was to identify the key features of each area that might be utilized to forecast the future. The information utilized are based on the area's infected count and are divided into three categories: confirmed, recovered, and death. Using the R-studio software the analysis was done with distances measurement done using Euclidean distances. The number of clusters generated using elbow and other standard approaches was three, and grouping active cases in an area is significant for making inferences about the disease effect.

## 2.5 Navigation simulation using Machine Learning and OpenRouteService

The machine learning technique that produced output in various data forms were explained before, however in this section, the models are utilized to provide navigation simulation by merging with OpenRouteService (ORS), an OpenStreetMap (OSM) platform. The two use cases to be addressed are a navigation simulation model and one model to estimate Average Speed in Travel. A simulation of pre-hospital emergency medical service is described in the study done by Olave-Rojas and Nickel (2021), with various scenarios and comparisons to present an accurate representation of the reality. Providing a connection between call-takers, ambulance crews, and emergency doctors for life-threatening emergencies is the main focus of this research. The other simulation approach done by Keller et al. (2020), is focused on estimating Average Speed in Rural Road Networks. To generate new input features, the ORS Directions data is utilized as a reference, combined with supervised and unsupervised ML models and dimensionality reduction. The procedure consists of dimensionality reducing the ORS input to supply it as input to machine learning models that produce a better result based on the R-square score, such as SVM, Linear Regression, the Ridge Model, and the SOM. The model with a greater R-square value estimates average speed values more accurately.

## 2.6 Conclusion

An analysis of recent publications revealed that there is no clear approach to offering a navigation system based on machine learning which can assist in avoiding accident-prone regions, which were calculated using a risk score based on the severity of incidents in a particular location. Certain data cleaning and machine learning approaches utilized in this research are inspired by concepts in the literature section described above, however the overall models generated in this research are novel and achieve the predicted outcomes with high accuracy.

# 3 Methodology

One of the most essential components of research is methodology, and in this section, the step-by-step process of research is presented using the KDD methodology (Figure 1)

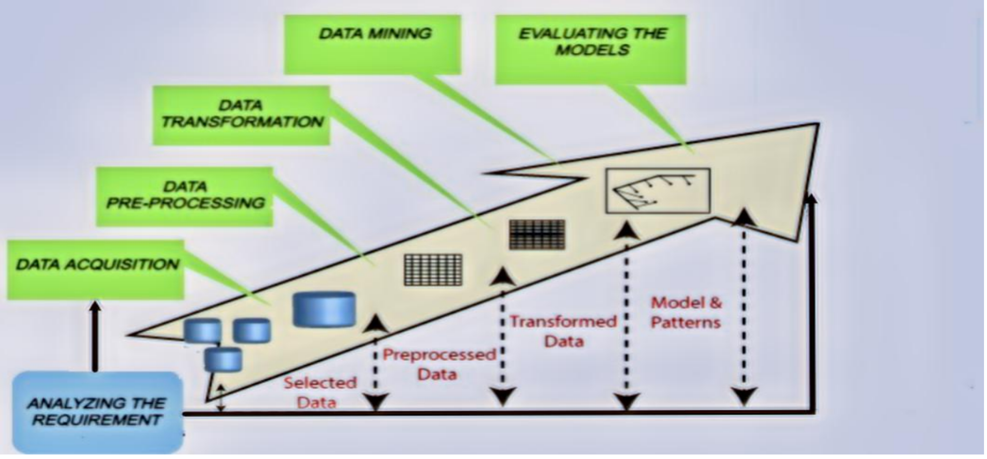


Figure 1: KDD Methodology
This is an edited version of the image[3]

## 3.1 Analysing the Requirement

The idea behind the research is to create a navigation system on the risk score of a location based on the severity of the accident. There are several criteria that may be used to evaluate the severity of an accident, but the research requires data from accidents that occur on a daily basis. Therefore it is necessary to identify a repository of daily

[3] https://static.javatpoint.com/tutorial/data-mining/images/kdd-process-in-data-mining.png

accidents in a particular city along with information about its geographical location, weather conditions, and a number of other factors that determine the severity of an accident, such as the value of property damage and the number of injuries or deaths involved. The severity is predicted using these details and then compared to the actual data to measure the model's efficiency. Following this, the risk score is computed based on identifying the key factors for estimating the severity of the accident. Once this is done the final information are imported into a medium which will helps in providing the navigation for the user.

## 3.2 Data Acquisition

Based on the research objective of identifying the severity of an accident on a daily basis, the best approach will be to examine accident data from a particular Country or city. Therefore, the accidents that have taken place in the city of Chicago is explored using the open Chicago repository[4].The repository contains data on Chicago's neighborhoods, which are cataloged into several categories so that users can easily explore them.Various datasets are accessible based on the category names, but only three are relevant to the research and are linked to each other. These three datasets contain accident data from 2013 to 2022, and because the need is just for 2021, only that year is filtered. The common details are linked to the IDs used to record the accident and the date when it occurred. The dataset ”Traffic Crashes - Crashes” provides details such as the location , weather and other pertinent information, which gives us an overall understanding of the accidents that takes place daily. It also provides us with an overall view of the damages that have happened in the accident, such as financial loss and the total amount of individuals affected in each of the accidents. The ”Traffic Crashes - Vehicles” and ”Traffic Crashes - People” datasets provide specific information about the vehicles and people involved in the accidents, respectively. Once all three datasets have been acquired, they are combined and subjected to data pre-processing.

## 3.3 Data Pre-Processing

The final data that is obtained after the three datasets are merged using their common columns. The primary step is to perform the data cleaning and missing values are to be identified. The remaining columns that are vital and provide information about the location and severity of the accidents are identified in stages. Following this, the redundant and non-information columns which were detected are eliminated. The missing values in vital category features are handled by replacing them with 0's and with values that do not significantly affect the data. The rows with missing values are not eliminated as it represents the number of accidents on a daily basis.

## 3.4 Data Transformation

The process of data transformation is very important in this research as there are several column names and values that are to be modified. The details of these activities as explained below.

- Columns with values stored in a different datatype are prioritized. The variables with a definite number of values, such as weather and lighting condition columns,

[4] https://data.cityofchicago.org/

are transformed to polychotomous variables and values which can be identified with 0's and 1's are transformed into binomial.

- The Bucketing or binning are done to the variables which are to be changed from a numeric series into fixed categorical ranges. There are also values which are closely related and hence they are merged as a whole.

- Feature engineering is performed by adding an additional column which improves the value of the information and enhances the performance of machine learning models.

- The polychotomous variables are converted to its corresponding binary values by changing each category to a feature in the data. For example, the weather_condition column which has a value 'clear' gets converted to weather_condition_clear and its corresponding values become 1's if the feature with clear is present else it becomes 0.

### 3.5 Data Mining

The data obtained after transformation process is split into training and testing data in the ratio of 3:1 respectively. The separated data is fed into data mining algorithms including classification, regression, and clustering. In the study, these three techniques are interdependent, i.e., the regression is conducted based on the classification output, and the clustering is performed based on the regression outcome. The Classification involves the use of models like Logistic Regression, Decision Tree Classifier and Random Forest. These models are used to determine the severity of the accident.The Multiple Linear Regression model is used to determine the risk scores for each location based on the severity of the accidents and K-Means clustering is used to establish boundaries using the locations co-ordinates .

### 3.6 Evaluating the models

Since three types of machine learning algorithms are utilized, each model is evaluated in a variety of ways. The research papers Muhammad et al. (2021) and Sammouda and El-Zaart (2021), works on supervised and unsupervised machine learning models, and helps in understanding the necessary techniques to evaluate the models used in the study. The classification models are evaluated by metrics such as Accuracy, Recall, Precision, ROC etc. These metrics can be obtained by using methods such as Confusion matrix,ROC-AUC curve. The Multiple Linear Regression model are evaluated by using metrics such as R Square/Adjusted R Square , (MSE)/Root Mean Square Error(RMSE) and finally Mean Absolute Error(MAE) . The elbow method is used to evaluate the K-Means clustering model.This is used to determine the optimal number of clusters required for model structure.

## 4 Design Specification

An ideal design specification will be used to describe the operating principle of the research in both technical and logical terms. This section will explain the brief descriptions provided in the methodology.

## 4.1 Overview of the Design

The overall design as show in Figure 2 explains that three datasets are merged to form the dataframe of size 238068 rows and 144 columns. The three datasets have common columns such as crash_record_id and crash_date which is unique and does not have any missing data. The detailed explanation of how the final data is obtained , will be done in section 5.2.Once the merged data is obtained, a combination of machine learning algorithms must be applied, with the results used to provide navigational guidance. The overall design components include the use of classification, regression, and clustering to develop a machine learning model that will be used to identify risk-free navigation routes utilizing Open Route Service (ORS), a location-based service generated from Open Street Map.

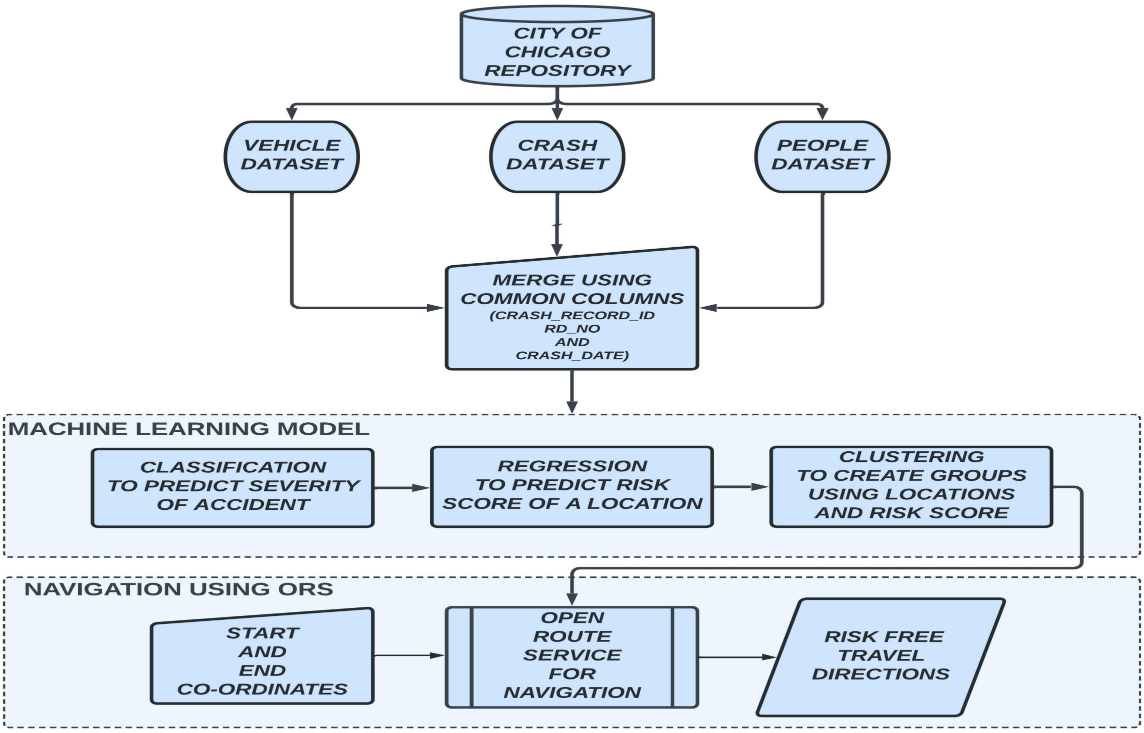

CITY OF CHICAGO REPOSITORY
VEHICLE DATASET
CRASH DATASET
PEOPLE DATASET
MERGE USING COMMON COLUMNS (CRASH_RECORD_ID RD_NO AND CRASH_DATE)
MACHINE LEARNING MODEL
CLASSIFICATION TO PREDICT SEVERITY OF ACCIDENT
REGRESSION TO PREDICT RISK SCORE OF A LOCATION
CLUSTERING TO CREATE GROUPS USING LOCATIONS AND RISK SCORE
NAVIGATION USING ORS
START AND END CO-ORDINATES
OPEN ROUTE SERVICE FOR NAVIGATION
RISK FREE TRAVEL DIRECTIONS

Figure 2: Overall Design of the Project

## 4.2 Design of Machine learning model

The three machine learning models used involve the both supervised and unsupervised technique. The explanation for the purpose of these techniques in this research study is given below.

- The supervised technique consists of both classification and regression model. Classification models such as logistic,random forest and decision tree classifier are used to identify if an accident is severe or not. The binary output explains the severity of daily accidents, with 0 indicating non-severe and 1 indicating severe. This model's important characteristics are used as independent variables in the regression model.
- The multiple linear regression gives us a risk score to identify, if a location is accident prone or not. The risk score are split into three categories which indicate the threat

level as high, medium and high. The details explanation can be found in the implementation section.

- The Unsupervised K-Means Clustering model is used to determine clusters or the various groups using the locations of the final data.The locations are divided into three groups based on analysis of elbow method. The clusters are then combined with the three risk score categories, resulting in nine groups, three of which have a high risk score and should be avoided.

### 4.3 Risk free navigation using ORS

The open Route Service, or ORS, provides travel directions based on a given location. By using its API service, the research will develop a navigation service where in users will able to travel when the start and end locations are provided. The direction along with the time taken and the distance to be covered can also be obtained using the ORS's API services. Once all the coordinates of three high-risk score areas are provided, the ORS will be able to provide an alternate route which is risk free. By using the folium library function, the navigation can be viewed in maps. The detailed explanation of this implementation will be given in the section 5.4.

## 5 Implementation

### 5.1 Exploratory Data Analysis

Since the data is based on one year, it's important to understand when most accidents occur. There are three time-related columns that offer information on the month, day, and hour of the event and hence the analyzing is performed using those. As per Figure 3, the majority accidents occur at 4 p.m. on Sundays and in the month of March. The research focuses on the factors that influence accidents, feature analysis is not necessary.

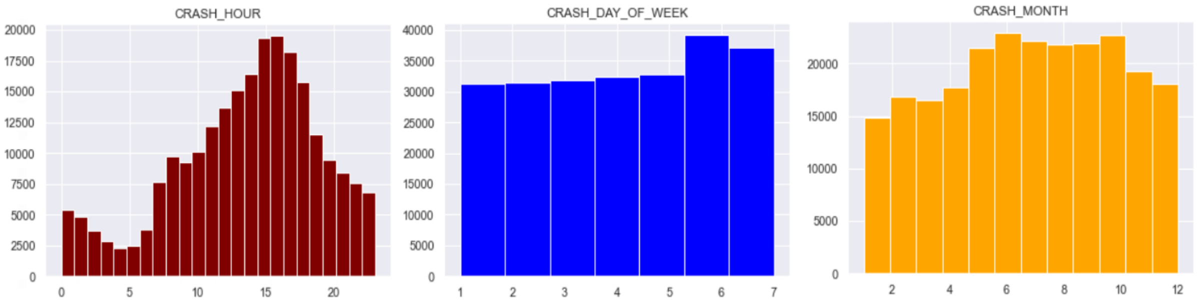


Figure 3: Analysis on Hour, Day and Month of the Accident

### 5.2 Preparation of the Final Data

The data to be utilized for the research study has numerous columns, thus the most difficult task is to reduce the number of columns and the values of certain columns.The functions that will performed to obtain the final data, are listed below.

### 5.2.1 Reducing Column Size

- The primary step is to identify columns that have more than 70% of missing values are removed. Therefore from 144 columns the data reduced to 74.

- The next step is to check columns that are interdependent. Columns such as AREA_01_I' and 'AREA_12_I' can be removed as their AREA related columns were removed.

- The next stage is to identify categorical columns and, if not, significant column values that can be changed to categorical columns.As a result, columns with values that are not missing but do not have significant information must be identified. The easier way that can be done is by checking if columns have values such as 'unknown' and 'NONE' on a higher percentage, they do not contribute to any details and hence can be removed. Few examples are Columns such as 'EJECTION' , 'BAC_RESULT' which have values similar to 'unknown'.

- Next step is to identify columns that have redundant information i.e columns which have similar information. Columns such as 'BEAT_OF_OCCURRENCE' denote the area of police patrols and since the location co-ordinates is present, it can be removed.

- Once the irrelevant columns are removed, the next step is to reduce the values and also make sure there are no more null values.

### 5.2.2 Feature Engineering

Feature engineering is done to remove columns as well as add column to improve the model performance.

- **Feature Engineering To Remove Columns:** The columns are further reduced by performing feature engineering using existing column values. There are columns such as 'STREET_NO', STREET_DIRECTION and 'STREET_NAME' which can be used to create an address of the location. Once it is done these three columns can be removed.

- **Feature Engineering To Add Columns:** Because the data is for a year, the seasons column may be created using the month value. Using CRASH_MONTH, the four seasons 'Winter,' 'Spring,' 'Summer,' and 'Fall' are generated based on their occurrence.

### 5.2.3 Reducing Columns Values by grouping

According to earlier research paper Zou et al. (2020), models becomes more effective when the column values are grouped. Therefore there are several columns which have values that can be grouped such as 'POSTED_SPEED_LIMIT' where the speed limit values are grouped into '0-15', '16-25', '26-40', '41+'. There are a total of 10 columns which are similar to this and hence all of them are changed. Once this is done, using pandas get_dummies, the values are converted to its corresponding binary values as explained in the transformation section in methodology. Once this is done the data is ready for data mining process

## 5.3 Implementation of machine learning models

The final data after performing data manipulation is given as an input to machine learning models. As per Figure 4 the machine learning models consists of three models such as classification, regression and clustering. The explanation for each of the machine learning models will be explained in the below section.

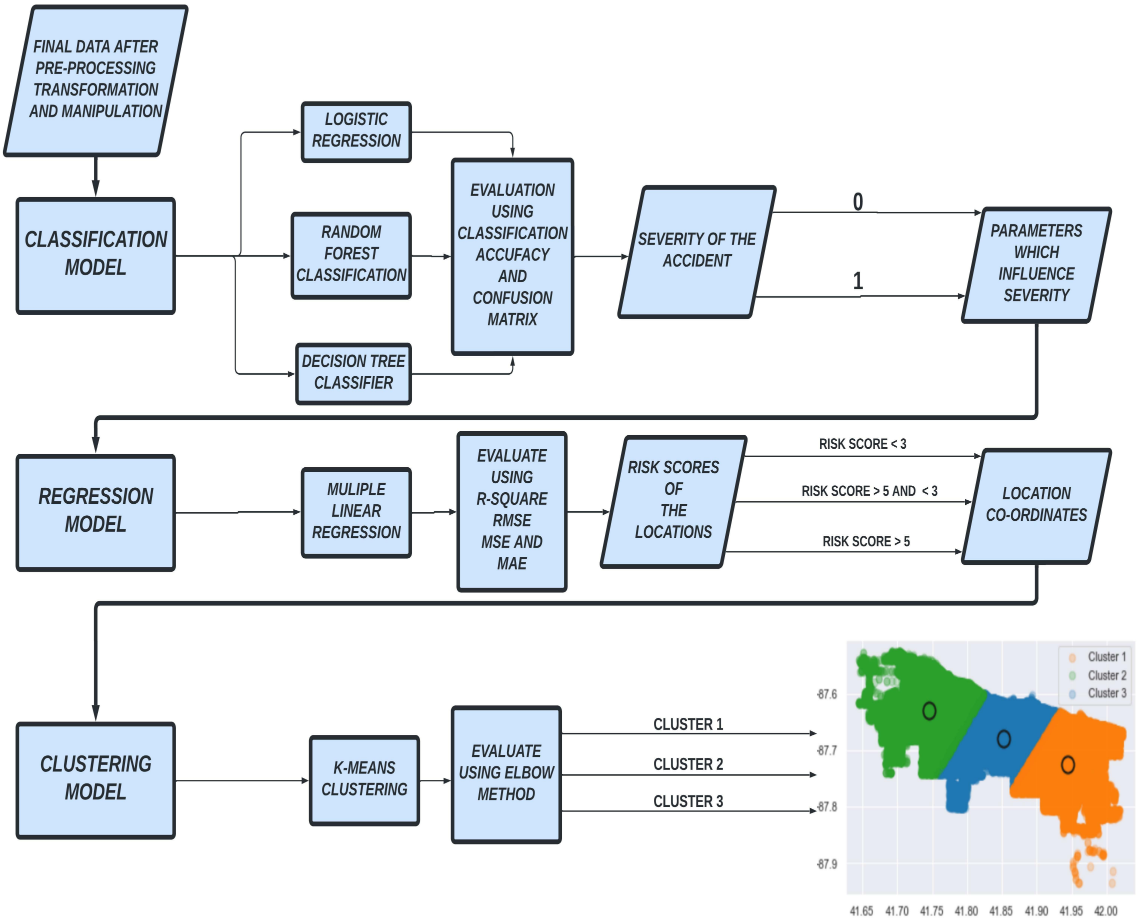


Figure 4: Machine Learning Model

### 5.3.1 Accident severity prediction using Classification models

The data is split into test and train data randomly in the ratio of 3:1 and given to the classification model. By utilizing the variables available in the final data it provides an output that categorizes the accidents into two categories: severe or not. To identify the model which can perform the best, the research will be using logistic, random forest and decision tree. The model is trained using the larger data (Training data), then using the testing data it is predicted and evaluated via the confusion matrix and evaluation metric which will be explained in detail in the evaluation section.

The Random Forest has the best accuracy followed by Decision tree, one of these models will be chosen to determine the features that influence the severity of the accident. The best model features are determined by utilizing the 'GridSearchCV' function, which is often used to identify the optimal hyperparameters of a model. Because the GridSearch in Random Forest takes longer to run than the Decision Tree, the latter will be applied. The Gridsearch decision tree model is implemented to discover the top parameters that produce the best results. This approach is then used to discover the features that will be used as criteria to determine the risk score. The Figure 5 shows the top features which influence the severity of an accident.

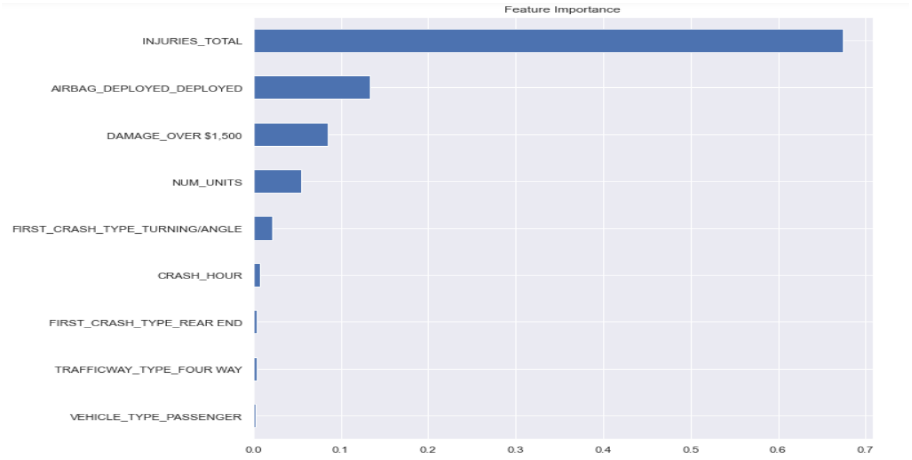


Figure 5: Top Parameters affecting the Severity

#### 5.3.2 Risk score prediction using Regression model

The linear regression model is used to determine a location's risk score using three criteria that represent the accident's mortality count, injury count, and financial impact. The analysis to show the level of influence of these 3 parameters have on the severity and it is explained in detail in the evaluation section 6.4 .The risk score ranges from 0 to 10, with individual parameters weighted at 2, 1.5, and 1.5 respectively. Data are divided into testing and training groups in a 3:1 ratio, and since no target variable is present for prediction, the equation of multiple linear regression is used. Since accidents can occur more than once in a location, the final risk score is calculated by averaging the scores of all accidents in that location. Once this is completed, the total number of distinct locations is 72681, each with its own risk score.

#### 5.3.3 Area Segregation using Clustering models

The previous two models were used to define a category and numerical variable, respectively, but the purpose of this model is to create a meaningful structure or clusters utilizing the 72681 distinct locations. The K-means clustering is used because it implements the Euclidean distance to locate the ”k nearest points” of a given sample point. In this case the algorithm works with the accident location's latitude and longitude coordinates, thus there should be no missing data in these columns. The number of clusters must be determined before proceeding with the k-means analysis. The elbow method is applied to provide a optimal number of clusters and the detailed explanation of how it works will be explained in the evaluation section 6.3. The model produced three clusters based on the elbow method, and the data was divided based on the distance feature, with each cluster having a center.

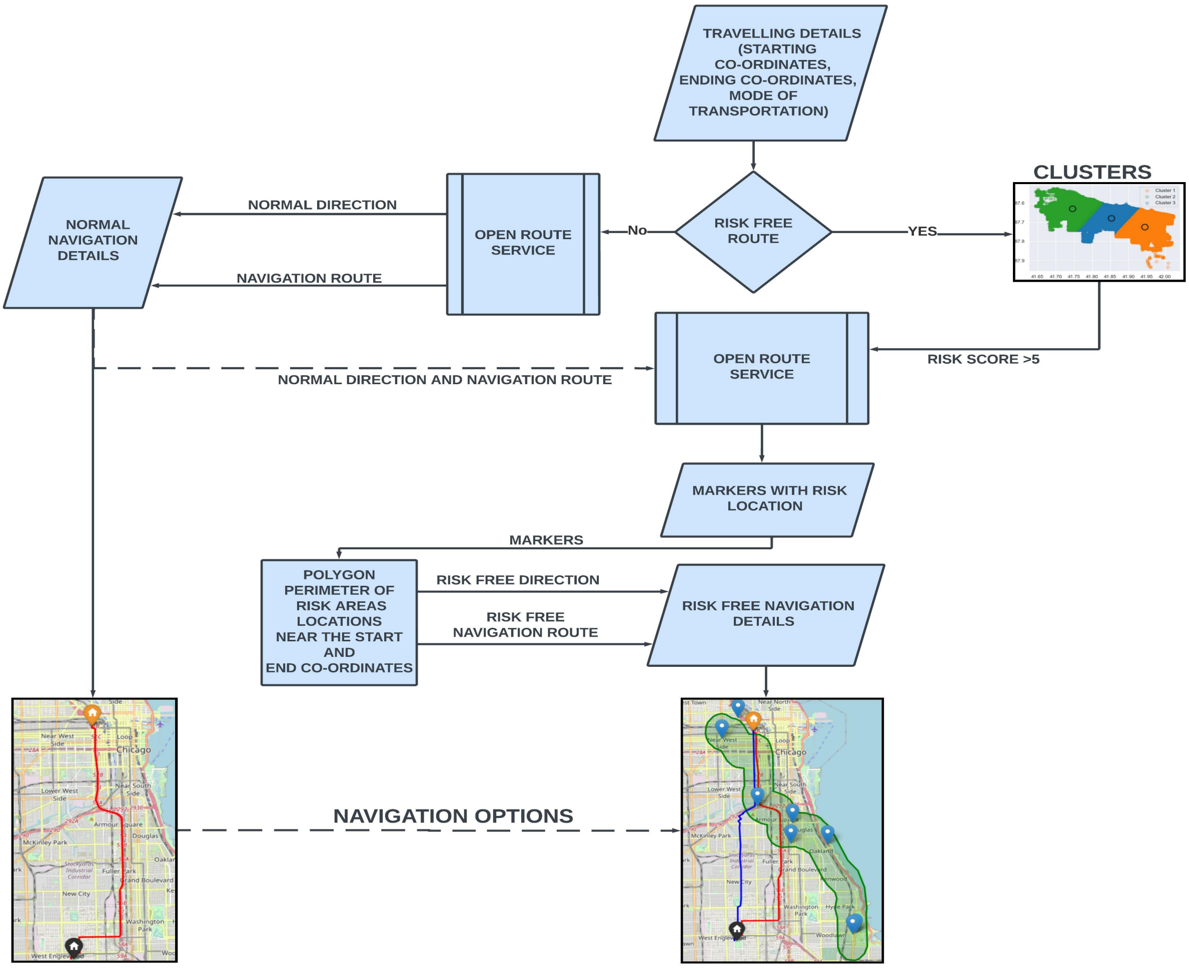

TRAVELLING DETAILS (STARTING CO-ORDINATES, ENDING CO-ORDINATES, MODE OF TRANSPORTATION)
RISK FREE ROUTE
YES
No
CLUSTERS
OPEN ROUTE SERVICE
NORMAL DIRECTION
NAVIGATION ROUTE
NORMAL NAVIGATION DETAILS
RISK SCORE >5
OPEN ROUTE SERVICE
NORMAL DIRECTION AND NAVIGATION ROUTE
MARKERS WITH RISK LOCATION
MARKERS
POLYGON PERIMETER OF RISK AREAS LOCATIONS NEAR THE START AND END CO-ORDINATES
RISK FREE DIRECTION
RISK FREE NAVIGATION ROUTE
RISK FREE NAVIGATION DETAILS
NAVIGATION OPTIONS

Figure 6: Navigation Flow

## 5.4 Navigation model using Open Route Service

The three clusters formed have their locations categorized based on local proximity, and each individual has their own risk score. As a result, they are divided into three categories such as risky (risk score greater than 5), cautious (risk score greater than 3 but less than 5), and safe (risk score less than 3) areas. The navigation works by utilizing the Openrouteservice's Location Based Service, which uses open street map data to give us with both the direction and the navigation route.By using the folium library function, the navigation can be viewed in maps. In order to access these feature in ORS API, an token key is created. As per Figure 6 the details required for standard navigation are the mode of travel, start and end location co-ordinates, however for a route that must be devoid of accident prone sites, the following parameters are required.

- The primary details necessary for risk-free navigation are the co-ordinates of accident-prone areas, which can be collected from the output of the final clustering model. Since there are 3 clusters, the risk locations of clusters that are near the start and the end co-ordinates are selected.

- Once the risk locations are collected, ORS accepts in co-ordinates in reverse(longitude, latitude) form, to generate a map consisting of markers to denote only risk location.

- Using the risk co-ordinates around the navigation zone, a perimeter surrounding the navigation may be created using the Buffer builder feature. It creates a polygon around the navigation path at a user-specified distance.

- The final navigation map includes an adjusted path that will use the risk locations near the normal navigation route and detour to an altered risk-free navigation route. The direction, distance, and time taken (depending on the mode of transportation) for both normal and risk-free travel are acquired using the ORS API's direction function.

- The Figure 7 shows the Normal vs Risk Free Route where the red color denotes the normal route and blue denotes risk free route. The green polygon denote the area or perimeter of risk zones, which are denoted by blue markers.

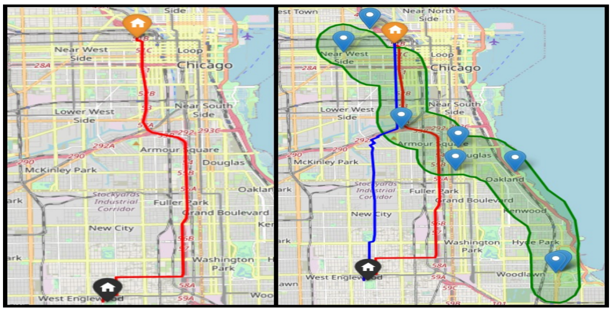


Figure 7: Normal Route vs Risk Free Route

# 6 Evaluation

The following experiments were performed on each machine learning models and their results were analyzed based as described in section 3.6

## 6.1 Evaluation of Classification Models

The classification model which is used to predict the severity consist of a data which is 236258 rows × 378 columns. The data is divided into training and testing sets randomly in the ratio of 3:1, with the training set used for trained and the testing set used for prediction and evaluation. Since there are three classification models, the model with the best accuracy will be utilized as the final model. The way in which these three models are evaluated are using the Confusion Matrix, Classification metrics and ROC-AUC curve which are explained below.

### 6.1.1 Evaluation using Confusion Matrix

A confusion matrix is a summary of a classification algorithm's prediction outputs. The number of correct and wrong predictions is summarized using count of the values and divided by class. The Evaluation metrics helps us to understand the overall precision,recall,f1-score and also gives us the accuracy of the three models. classifier's precision explains how well the model avoids labeling instances as positive when they really aren't, and its recall is how well it finds all positive instances. F1 scores are computed using a weighted harmonic mean of precision and recall, with 1.0 being the finest and 0.0 being the worst. The Support is the number of instances the value is specified in the data. The Figure 8 is of the confusion matrix which illustrates the split where the total number of 0s (38236) and 1s (20829) are accurately predicted and the metrics of it is shown in Table 1. The total number of Non-Severe values predicted by logistic regression, decision tree, and random forest is 37383, 36015, and 37697, respectively, whereas the total number of 1s predicted by them is 14537,18781, and 17530.

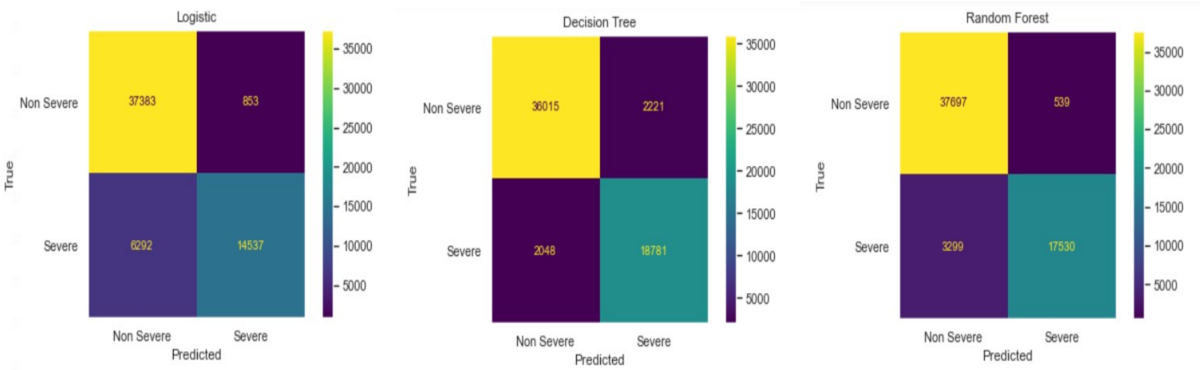
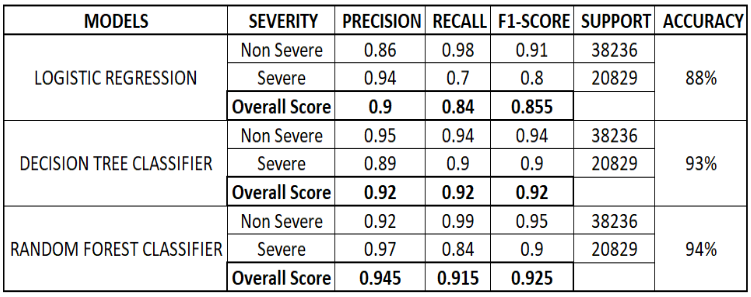


Figure 8: Confusion Matrix of all the Classification Models



| MODELS | SEVERITY | PRECISION | RECALL | F1-SCORE | SUPPORT | ACCURACY |
|---|---|---|---|---|---|---|
| LOGISTIC REGRESSION | Non Severe | 0.86 | 0.98 | 0.91 | 38236 | 88% |
| | Severe | 0.94 | 0.7 | 0.8 | 20829 | |
| | **Overall Score** | **0.9** | **0.84** | **0.855** | | |
| DECISION TREE CLASSIFIER | Non Severe | 0.95 | 0.94 | 0.94 | 38236 | 93% |
| | Severe | 0.89 | 0.9 | 0.9 | 20829 | |
| | **Overall Score** | **0.92** | **0.92** | **0.92** | | |
| RANDOM FOREST CLASSIFIER | Non Severe | 0.92 | 0.99 | 0.95 | 38236 | 94% |
| | Severe | 0.97 | 0.84 | 0.9 | 20829 | |
| | **Overall Score** | **0.945** | **0.915** | **0.925** | | |

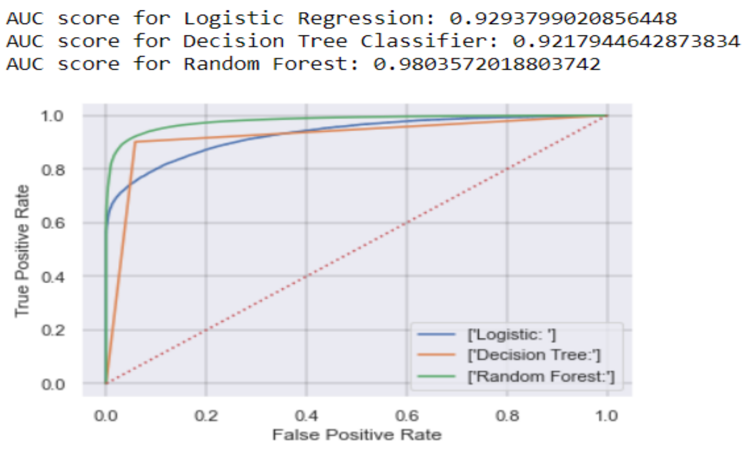
Table 1: Evaluation Metrics of all the Classification Models

### 6.1.2 Evaluation using ROC and AUC score

The ROC curve is a probability curve that describes performance by merging confusion matrices at all threshold levels. The AUC measures the degree of separability and how well the model can differentiate between classes. The higher the AUC, the better the model is at predicting 0 classes as non-severe and 1 classes as severe. Figure 9 shows that Random Forest has the highest AUC which is 98% followed by Logistic and Decision Tree with 92.9% and 92.1% respectively.

## 6.2 Evaluation of Regression Model

The regression model is used to predict the risk score, which ranges from 0 to 10, and the Scatter plots of Actual versus Predicted values in Figure 10 illustrate it graphically. If the model is efficient and has a high R Square value, all of the points should be around the red regression line, which shows the regression link between the dependent and independent variables.The evaluation metrics to be used for regression models are R Square, Adjusted R Square, MSE, RMSE, MAE.



```
AUC score for Logistic Regression: 0.9293799020856448
AUC score for Decision Tree Classifier: 0.9217944642873834
AUC score for Random Forest: 0.9803572018803742
```

Figure 9: ROC AUC of all the Classification Models

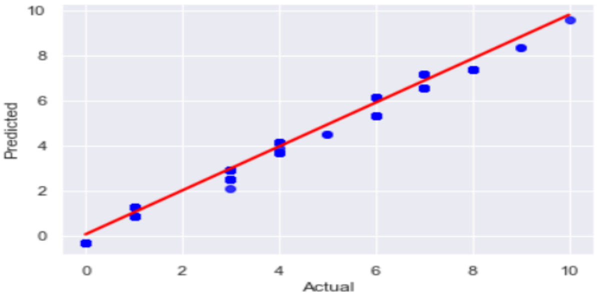


Figure 10: Actual vs Predicted Graph

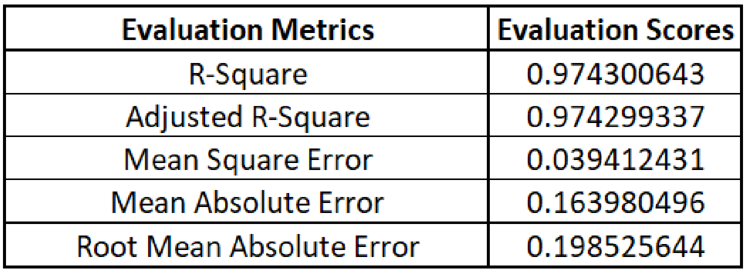
The final model is implemented in the testing data and as predicted by the graph, With R-squared: 0.974300643 and Adjusted R-squared: 0.974299337, the model can be regarded to be effectively fitted. The mean absolute error and mean squared error measure the average of the absolute and square difference between the actual and projected values, respectively, and both metrics evaluate the residuals, where the former measures the average and the latter the variance.A regression model with a lower MAE, MSE, and RMSE value is more accurate and hence The model is regarded accurate with values such as 0.039412431, 0.163980496, and 0.198525644.

## 6.3 Evaluation of Clustering Model

The K-Means clustering model has the objective of partitioning 72681 locations into k clusters and for identifying the cluster count, the elbow method is used. It involves calculating the sum of squared distances between each location and the centroid and determining how many clusters to apply by using the curve's elbow. Figure 11, shows that the cluster has reduced slowly and we can see the elbow point at k=3.

## 6.4 Discussion

As part of the overall study, three classification models, one multiple linear regression model, and one clustering model were developed. The overall metrics score of classification and regression models is shown in Table 1 and Table 2 respectively. Each categorization evaluation metric is divided into two as they are calculated and determined



| Evaluation Metrics | Evaluation Scores |
|---|---|
| R-Square | 0.974300643 |
| Adjusted R-Square | 0.974299337 |
| Mean Square Error | 0.039412431 |
| Mean Absolute Error | 0.163980496 |
| Root Mean Absolute Error | 0.198525644 |

Table 2: Evaluation Metrics of Multiple Linear Regression Models

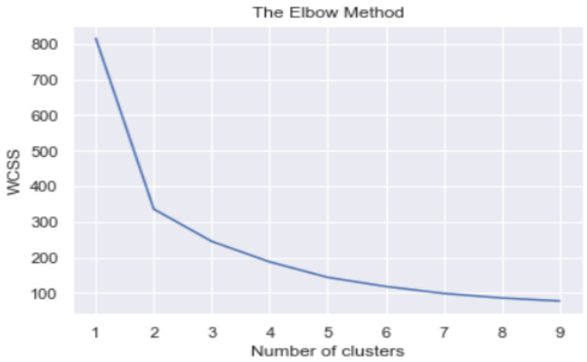


Figure 11: Elbow Method for K-Means Clustering

separately. This is due to the fact that in some circumstances, a model is necessary to determine just severe or non-severe accidents. The Random Forest classifier has the highest scores in practically all measures, with a decrease in the recall score determining severe accidents. The data in the classification model used are based on the training and testing data which were generated randomly and hence they are imbalanced. As the data includes accidents occurring every day and most of them are not severe, the data was not balanced using techniques such as SMOTE (synthetic minority oversampling technique).This is because by balancing the data it could be falsified and may result in discarding the insights needed to develop rule-based classifiers, such as Random Forests.

The figure 12 shows all the factors which are used as independent variables in the calculation of risk score using multiple linear regression. The risk score of many areas that are risky are less compared to that of safe is because the fatal injury has occurred in a very less count (545) which is less compared to the size of the data. The figure 12(a) shows that the majority of non severe accidents occur when there are no injuries. The figure 12(c) shows that majority of the accidents irrespective of the severity have the cost of damage more than 1,500.

The Figure 13 shows the clustering outcome of all the 3 risk areas along with the overall accident prone area. On comparison with others, risk area in figure 15 does not complete the whole pattern since there are less risk areas compared to safe and cautious. Another distinction is that there are four clusters, as opposed to three in the others. The regression and clustering models are used to identify the risky areas and locations which are provided to the Open Route Service (ORS).

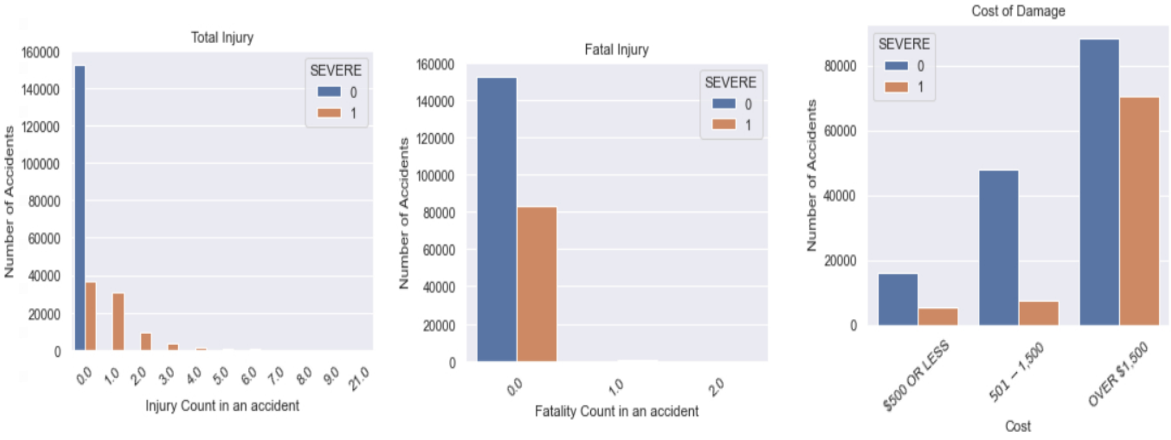


Figure 12: (a) Severity vs Total Injury (Top Left), (b) Severity vs Fatal Injury (Top Right) and Severity vs Damage Cost (Bottom)

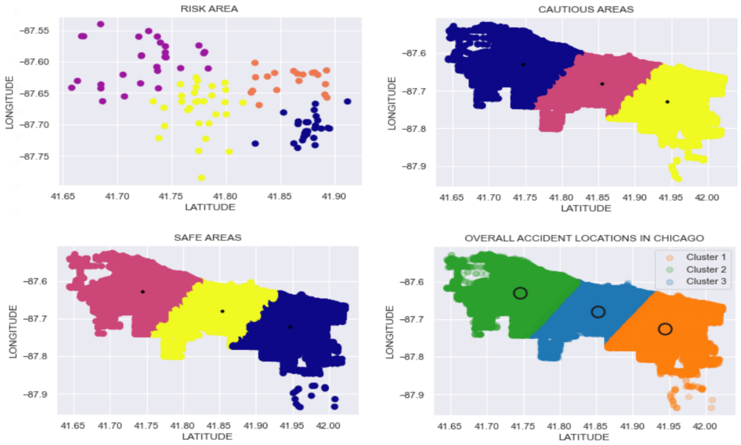


Figure 13: (a) Risky Locations (Top Left), (b) Cautious Location (Top Right), (c) Safe Area (Bottom Left) and (D) Overall Accident Area

# 7 Conclusion and Future Work

The research objective is to provide a navigation which is free of accident prone zone. In order to achieve this objective the research was separated in to 4 parts where three consists of machine learning models and one uses openrouteservice website which provides route service when the locations are provided. The first model used has three classification models to determine the severity and it was successfully implemented with Random forest model being the highest performing model with an accuracy of 94%. Using the multiple regression model, the risk scores were predicted.The final location was then given as the input for clustering model and hence the 3 clusters were obtained. The cluster and the risk scores were gives a final 9 clusters, 3 for each of the risk level. Since the research focused only on avoiding the risky areas, only the risky locations were provided to the Open Route Service, and ideally the required output consisting of a normal and an alternate route avoiding the risky areas was obtained

The domain offers plenty of room for future research. Although, the proposed model produced very good predictions, the data used is only for a year 2021, despite the data being available for 5 years. With the usage computational resources, this could be implemented. The data coverage is constrained inside the city of Chicago, therefore by getting accident data from different places, the model may be enhanced and the scale for navigation can be expanded. In the paper Tang et al. (2019), Random Forest and other models were not used, as it takes longer time to run and therefore gridsearchcv with the decision tree model was used. Similar to Islam et al. (2021), further enhancement can be done by using other clustering models such as DBSCAN and OPTICS, and comparison can be done based on internal cluster validation metric and execution time. By including the cautious locations along with risky one, the scale of the navigation model can also be improved.

# Acknowledment

I would like to take this opportunity to express my sincere gratitude to everyone who has helped me throughout my Postgraduate career, especially my supervisor Prashanth Nayank. His assistance and advice during the research period were extremely valuable in carrying out this project.

# References

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R. and Hidayat, R. (2022). The application of k-means clustering for province clustering in indonesia of the risk of the covid-19 pandemic based on covid-19 data, *Quality & Quantity* **56**(3): 1283–1291.

Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P., Deo, R. and Salcedo-Sanz, S. (2022). Machine learning regression and classification methods for fog events prediction, *Atmospheric Research* **272**: 106157.

Dadwal, R., Funke, T. and Demidova, E. (2021). An adaptive clustering approach for accident prediction, *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, pp. 1405–1411.

Elhassan, A., Abu-Soud, S. M., Alghanim, F. and Salameh, W. (2022). Ila4: Overcoming missing values in machine learning datasets–an inductive learning approach, *Journal of King Saud University-Computer and Information Sciences* **34**(7): 4284–4295.

Hong, A., Noh, D. and Choi, J.-H. (2021). An analysis of elderly drivers' traffic accidents influential factors using multiple linear regression, *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, pp. 1722–1724.

Islam, M. R., Jenny, I. J., Nayon, M., Islam, M. R., Amiruzzaman, M. and Abdullah-Al-Wadud, M. (2021). Clustering algorithms to analyze the road traffic crashes, *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, IEEE, pp. 1–6.

Keller, S., Gabriel, R. and Guth, J. (2020). Machine learning framework for the estimation of average speed in rural road networks with openstreetmap data, *ISPRS International Journal of Geo-Information* **9**(11): 638.

Klinjun, N., Kelly, M., Praditsathaporn, C. and Petsirasan, R. (2021). Identification of factors affecting road traffic injuries incidence and severity in southern thailand based on accident investigation reports, *Sustainability* **13**(22): 12467.

Li, J., Boonaert, J., Doniec, A. and Lozenguez, G. (2021). Multi-models machine learning methods for traffic flow estimation from floating car data, *Transportation Research Part C: Emerging Technologies* **132**: 103389.

Muhammad, L., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C. and Mohammed, I. A. (2021). Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset, *SN computer science* **2**(1): 1–13.

Olave-Rojas, D. and Nickel, S. (2021). Modeling a pre-hospital emergency medical service using hybrid simulation and a machine learning approach, *Simulation Modelling Practice and Theory* **109**: 102302.

Pervez, A., Lee, J. and Huang, H. (2021). Identifying factors contributing to the motorcycle crash severity in pakistan, *Journal of advanced transportation* **2021**.

Sammouda, R. and El-Zaart, A. (2021). An optimized approach for prostate image segmentation using k-means clustering algorithm with elbow method, *Computational Intelligence and Neuroscience* **2021**.

Stigson, H., Malakuti, I. and Klingegård, M. (2021). Electric scooters accidents: Analyses of two swedish accident data sets, *Accident Analysis & Prevention* **163**: 106466.

Tamakloe, R., Lim, S., Sam, E. F., Park, S. H. and Park, D. (2021). Investigating factors affecting bus/minibus accident severity in a developing country for different subgroup datasets characterised by time, pavement, and light conditions, *Accident Analysis & Prevention* **159**: 106268.

Tang, J., Liang, J., Han, C., Li, Z. and Huang, H. (2019). Crash injury severity analysis using a two-layer stacking framework, *Accident Analysis & Prevention* **122**: 226–238.

Yan, M., Chen, W., Wang, J., Zhang, M. and Zhao, L. (2021). Characteristics and causes of particularly major road traffic accidents involving commercial vehicles in china, *International journal of environmental research and public health* **18**(8): 3878.

Yan, X., He, J., Zhang, C., Liu, Z., Qiao, B. and Zhang, H. (2021). Single-vehicle crash severity outcome prediction and determinant extraction using tree-based and other non-parametric models, *Accident Analysis & Prevention* **153**: 106034.

Zhu, S. (2021). Analysis of the severity of vehicle-bicycle crashes with data mining techniques, *Journal of safety research* **76**: 218–227.

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

Zou, Y., Zhu, T., Xie, Y., Li, L. and Chen, Y. (2020). Examining the impact of adverse weather on travel time reliability of urban corridors in shanghai, *Journal of Advanced Transportation* **2020**.