A Machine Learning Based Comparative Analysis of Accident Severity Prediction Mechanism in USA

MSc Research Project
Data Analytics

Soham Mohire
Student ID: 19225491

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Soham Mohire
Student ID:	19225491
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	15/12/2022
Project Title:	A Machine Learning Based Comparative Analysis of Accident Severity Prediction Mechanism in USA
Word Count:	6967
Page Count:	25

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:	14th December 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):	

A Machine Learning Based Comparative Analysis of Accident Severity Prediction Mechanism in USA

Soham Mohire
19225491

Abstract

Traffic accidents that occur worldwide are a concerning issue as it results in major deaths and injuries. This burden of casualties tends to be higher in developing countries. Hence, a model to predict the occurrence of accidents is a significant challenge. However, one of the substantial ways to predict the severity of such accidents is through the implementation of machine learning algorithms. Therefore, the primary aim of the proposed thesis is to automate the process of accident detection by analysing the levels of severity and filtering a set of impactful factors that might result into a road accident. For this purpose, a dataset from Kaggle repository is obtained that contains a countrywide car accident data of USA from Dec 2016 to Dec 2021. Theoretical concepts of SMOTE is implemented to balance the dataset and thereby handle data imbalance. Later, the dataset is used to develop a framework based on four machine learning algorithms and one stacking algorithm. Finally, an analysis is done based on factors such as weather conditions and different severity levels that might lead to the occurrence of road accidents. The experimental analysis so conducted in the study indicates that the random forest model has performed better in comparison to all the implemented models by generating an accuracy of 74 percent.

1 Introduction

One of the most undesirable and unforeseen event occurring to road user would be a road accident. In recent times, large number of such accidents has been witnessed in many parts of the USA. Occurrence of fatalities and injuries has created a huge impact on the economy of the country. This has not only led to untimely deaths, but also loss of property damage at social levels. In a survey conducted by WHO in 2017 [1] states that 1.5 million drivers die on a yearly basis due to road accidents and car crashes. In addition to this they stated, that due to negligence in traffic rules, the number of deaths is more likely to increase by 2030. The statistics given by the survey has concluded that 47 road users died on an everyday basis, which led to a 3 percent decrease in the GDP. In another survey report by Michigan Traffic [2] they have registered an estimation of 314,921 death occurrences due to road accidents in 2017. The numerical estimation led to a loss of 230 billion dollars with a massive drop in the economy of the country. Such devastating statistics has eventually become a matter of rising concern, for not only government officials

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] https://www.michigan.gov/msp/divisions/cjic/traffic-crash-reporting-unit

but also research scholars and accident experts belonging to the same field. Extensive research has been conducted to determine all the factors that would highly be responsible for such statistics. Unfortunately, the surveys are conducted through questionnaires and not much implementation of statistical tools has been witnessed. Hence, the primary aim of research scholars is to not only to analyse the factors responsible for such accidents; but also resolve the severity of accidents from a behavioural angle of road users.

Since accidents are highly unpredictable and unforeseen; drawing observations from them is a significant challenge. In addition to the observations being made, drawing conclusions with 100 percent accurate data and results is quite an impossible task. To overcome this limitation, the thesis proposes to detect the severities that might lead to road accidents by using and implementing technical methods of Machine Learning. ML is observed to be as one of the most advanced and technical method to predict the occurrence of such mishaps. Hence, the aim is to build a classification model that can predict road accidents through the experimentation analysis of five classifiers. The classifiers would further undergo a set of processing techniques and make smart decision s by gaining insights from historical data. In addition to the implementation of the selected classifiers, the thesis also presents a conceptual theory of SMOTE; wherein the data would be balanced so that the classification algorithms can perform on them in a significant manner.

1.1 Background

The facilitation of transportation in US is primarily fulfilled through three means: road, air and railways. The preference to cover long distances is gratified through air; whereas road means are majorly used for shorter distances. In such a scenario, comprehending traffic rules is expected to be a significant countermeasure so that traffic collisions could be avoided. Despite the fact of an increasing number of deaths that occur due to road accidents; a rise in injuries, social and economic loss still remains a major concern. This has majorly targeted the GDP of the country. In addition to the economic factors that are laid on the country; accidents and road collisions tend to have a negative impact on the society as a whole. In scenario of external parameters such as service delay of ambulance is also an essential factor to be considered while determining the factors of traffic collision. Hence, many research scholars have been constantly studying on severity reasons and filtering the influential parameters that are responsible for such accidents. For this reason, experts and research scholars have minimalized traffic collisions by initiating different prediction models that could determine collision types and severity levels. The severity levels so created are further taken into consideration as countermeasures of the procedure so that a collision situation could be handled. For instance, if a junction experiences frequent accidents; the respective authorities can install a STOP signal board to prevent the occurrence of collisions.

However with recent advancements in technology, multiple accident detection techniques have been evolved wherein research scholars try to gather accident data with respect to their duration time so that its prediction can be made in an accurate manner. In such a scenario, statistical models and soft based approaches have been widely adopted to analyse the accident time and its occurrence. Apart from the implementation of machine learning and deep learning based models; probabilistic distribution based models such as Structure Equation Model [SEM] (Harb et al.; 2008), Hazard Based Distribution [HBDM] (Pakgohar et al.; 2011) have also been implemented. Implementation of machine learning based models include the execution of algorithms such as SVM, NB, Linear Regression etc. whereas implementation of deep learning models to predict the same includes the execution of algorithms such as Artificial Neural Networks [ANN] (Beshah and Hill; 2010) and Genetic Algorithms (Yin et al.; 2015).

In addition to external parameters as mentioned above, there are several other factors that are yet undetermined which results into road accidents. Such factors are considered to be heterogeneous in nature. Therefore, such scenarios result into acquiring an existing dataset with multiple historical records so that the prediction accuracy of the proposed model could be enhanced. Below mentioned are some of the reasons for road accidents:

- Exceeding the speed limit: a vehicle moving at higher acceleration might have more chances of colliding with another vehicle. In addition to this; over speeding might also result into misjudgement of the road track.

- Drunken driving: this is considered to be as the second most influential reason that results into an occurrence of a road accident.

- Distractions: attending phone calls while driving tend to deviate the attention of drivers from the road and hampers their judgement of driving.

- Unfollowing the traffic rules: drivers majorly tend to skip signals that leads to collision at the intersection.

- Weather conditions: weather conditions in USA such as rain, snow and fog majorly affects the speed of driving on roads.

1.2 Research Problems

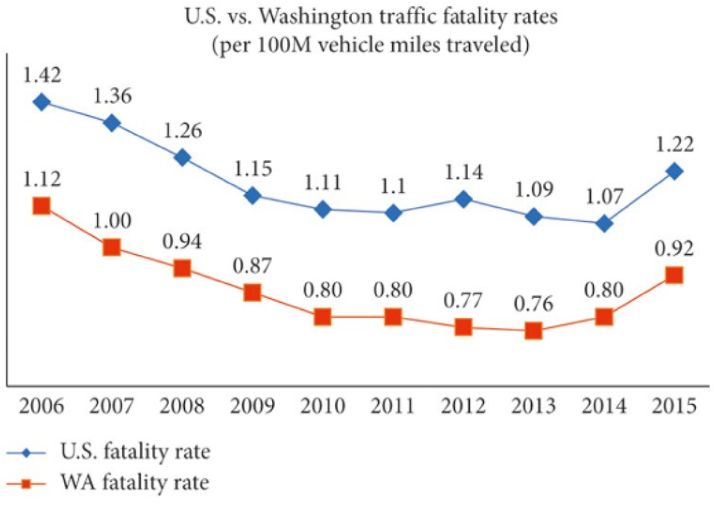


Figure 1: Statistics of road accidents occurring in US and Washingtonn

Taking respective measures to ensure traffic security in the US has been identified as the utmost necessary hotspots around the globe. A statistical analysis of the same is illustrated in figure 1 that depicts the fatality rates occurring in US and Washington. The statistical number has been constantly increasing with 117,053 crashes being identified by Washington alone that includes an approximate value of 500 fatal cases being confirmed [3]. An indication of 8 collisions has been witnessed to occur in every 5 minute that results into a crash every 16 hours [4].

Considering the mentioned statistics above, following are the research problems narrated with respect to the thesis:

- Despite the considerations of several factors such as behavioural, mental and weather conditions; the real and root cause of accidents remains unknown
- Historical datasets with regards to occurrence of road accidents is limited on repositories. With minimal access to dataset and limited number of samples, accurately predicting the reason of road accidents becomes a challenging problem
- Training and testing a smaller dataset might not give higher results of accuracy
- Several factors are responsible for the occurrence of a road accident. In addition to the mentioned reasons; an accident might also occur in scarcity with respect to space and time. This can be considered to be as an unknown factor that might influence the mishap to occur
- Determining real reasons for severity prediction in a country like US is a significant challenge since multiple assumptions and underlying's could be made that might trigger to misled findings
- External parameters such as service delay cannot be monitored
- Most of the existing detection techniques are either based on classification or regression algorithms; that limits down the overall performance of the system model
- Scouting a research work that considers all the parameters for severity prediction is a remarkable problem

1.3 Motivation

The occurrence of road accidents leads to fatal deaths on a regular basis. This results into huge amount of loss for the country. Therefore it is necessary to ensure that certain safety precautions are adopted so that road accidents and collisions can be surpassed. Therefore, this becomes an essential motivational factor for the proposed study; wherein all the available parameters are taken into consideration so that a large portion of financial loss is retrieved. In addition to the influential parameters; proactive approaches and real time operations should also be taken into consideration. Proactive approaches primarily involves the process of resolving road safety issues by analysing the risks associated with it; whereas real time operations include an analysis of scenarios that might result into accidents. Comprehending the motivational factors, below mentioned are the contributions of the thesis:

- Implementation of US dataset to predict the severity of accidents

[3] https://wsdot.wa.gov/MapsData/crash/collisionannual.htm

[4] https://wsdot.wa.gov/MapsData/crash/collisionannual.htm

- Application of SMOTE to balance the dataset so acquired
- Proposal of machine learning algorithms to satisfy the optimal threshold of existing classification algorithms

1.4 Resesrch Objective

The primary aim of the thesis is to develop an automated model that could predict the severity of accidents occurring on the road. For this purpose, the thesis operates on five machine learning algorithms that are deployed on the database acquired from Kaggle repository. The database consists of information with respect to the occurrence of mishaps taking place in the US. Parameters such as weather conditions, monthly and daily analysis of road accidents is taken into consideration. Other amenities such as fog, snow and rain are also taken into consideration. Following are the research objectives of the proposed thesis:

- To develop a model that could predict collisions based on occurrence of similar events in the past
- To carry out statistical analysis and establish a relationship between different factors that influence the occurrence of mishaps
- To analyse the limitations of the existing techniques and implement a model that could overcome them
- To correlate occurring accidents with respect to multiple severity factors
- To manage real time data using traffic API's
- To perform feature engineering techniques so that high levels of accuracy is obtained

1.5 Research Question

There are multiple factors and parameters that are responsible for the occurrence of road accidents. Some involve behavioural factors whereas other includes external parameters such as delay in ambulance services. In such a scenario it is mandatory to list all the possible influential factors that might add more severity to the existing parameters. Hence, this narrates the fundamental research question of the proposed study:

RQ1: What are the existing parameters that are likely to result in a road accident? The availability of accidental datasets is limited on respective repositories. This makes it challenging for research scholars to conduct their experiments using machine learning and deep learning techniques. In addition to a limited dataset; the sample of accident reports so received is imbalanced in nature. A limited and imbalanced dataset has higher chances of not generating optimized levels of accuracy. Hence, this narrates the secondary research question of the proposed study

RQ2: How can the acquired dataset be balanced so that the proposed model can generate better accuracy? Multiple research algorithms have been implemented by research scholars; that includes the implementation of ML, DL, AI and transfer learning. In addition to the conventional algorithms and techniques so used; there are various other approaches that have been highlighted which focuses on accident prediction. However, there are certain drawbacks in the existing techniques. Hence, this narrates the second secondary research question of the proposed study

RQ3: What are the limitations of the existing work and how does the proposed thesis tends to overcome them?

1.6 Organization of Thesis

Chapter 1 includes a summary on accidents caused due to multiple severities. It also includes the problems associated with the research followed by research objectives. Chapter 2 summarizes a thorough literature survey being performed by multiple authors from the same domain. It includes the research work being executed in recent years. Chapter 3 summarizes on the methodologies and algorithms used to implement the thesis. Chapter 4 includes design specifications along with the architectural flow of the proposed work. Chapter 5 includes the workflow of the same along with a brief on implementation details. Chapter 6 mentions the parameters used for evaluating the system model. Chapter 7 includes the results so obtained followed by conclusions and references.

2 Related Work

2.1 Detection of Crash Severities using Statistical and Analytical Modelling

The application of statistical modelling has been widely adopted to predict crash severity as it serves the purpose of being a good indicator and help into easy interpretation of the results. Implementation of regression algorithms have been considered to be as a commonly used technique in statistical modelling. Authors (Vajari et al.; 2020) have conducted their research work in order to determine all the possible factors that might contribute to a road accident. In a similar work proposed by authors (Yuan et al.; 2021) they implemented the techniques of latent class clustering analysis to predict the severity involved in road accidents.

Through the process of the literature survey, it was observed that the evolution of statistical methods was highly implemented in comparison to machine learning methods. However, statistical algorithms performed with less computational complexity; but the methods they used to analyse crash research was commendable. In a research work proposed by author (Mannering and Bhat; 2014) he analysed the highway data that resulted into car crash. For this purpose he used the techniques of statistical modelling that involved implementation of clustering analysis. Since the dataset had information on the reasons so as to why a location was more prone to accidents; the author used this data to analyse reasons on highway crash. The usage of clustering model helped to cluster specific reasons that would majorly result into an accident. In this way, the author analysed the steps and later provided methodological directions so that respective safety countermeasures could be taken and the overall crash severity could be reduced.

In addition to clustering analysis, K-Nearest Neighbor (KNN) is also considered to be as a partially non parametric statistical model that could perform regression tasks. Authors (Cover and Hart; 1967) contributed their work into predicting the real artifacts that could be considered as a major reason for car crash. The respective artifacts were collected using KNN and converted to their equivalent variables. The generated variables later responded only to those observations that were close to the distance between the generated values. Once the values were predicted using KNN an assumption was made and a local structure of crash data was created.

The implementation of statistical modelling however required certain assumptions to be made so that a probabilistic distribution could be established. This distribution further established a relationship between dependant and independent variables. It was simultaneously observed that the implementation of machine learning algorithms boosted along with that of statistical modelling. Since, machine learning did not require an assumption to be made between variables; its implementation did not require a model of underlying mechanism. However, in certain scenarios such as the work proposed by (Cover and Hart; 1967) an overlap of statistical modelling and machine learning took place. This was due to the fact, that both the concepts dealt with analysis being performed on the dataset. However, a major difference that distinguished the two techniques; was their inference with respect to variables. Statistical analysis required a relationship to be established between the variables; whereas this was not the case with machine learning. Machine learning does not demand on assumptions to be made or neither the establishment of underlying relationships.

Figure 2 depicts a pie chart which illustrates the distribution of research works being made in the same domain with the adoption of statistical modelling and machine learning algorithms:

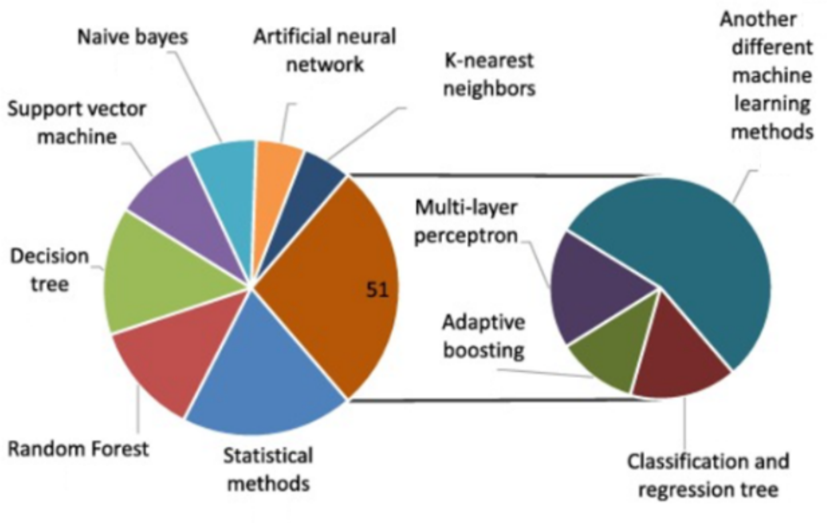


Figure 2: Distribution of Algorithms to detect Accident Severity

2.2 Detection of Crash Severities using Machine Learning

The authors (Wang and Kim; 2019) proposed an accident prediction method that included 201,581 crashes occurring on the roads in Maryland, USA. The prediction method included various factors that contributed to identifying the severities attached to road crashes. The author proposed the generation of two algorithms namely; random forest and decision trees. The prediction accuracy of the model was based on precision and recall factors so obtained. In addition to this, the evaluation parameters also included sensitivity and specificity values. However, through the implementation it was concluded that the model conducted using random forest generated higher levels of accuracy.

In another research work proposed by authors (Labib et al.; 2019) a total of four machine learning based algorithms were used to predict accident severities. The dataset used for model prediction included accidents that occurred in Bangladesh from 2001 to 2015. This dataset included 43,089 road crashes. For the purpose of enhanced prediction, feature selection methods including univariate and bivariate functionalities were taken into consideration. Recursive feature elimination was also used that included the process of essential feature extractions. In addition to the implementation of machine learning algorithms, AdaBoost was also used to increase the overall accuracy levels. However, the boosting mechanism proved to attain an optimized detection of accident severities.

Authors (Jiang et al.; 2019) proposed the implementation of five predictive classifiers that could detect the injuries related to car crashes taking place in Washington, USA. Five classifiers included the implementation of decision trees, logistic regression, SVM, KNN and Naïve Bayes. All the five classifiers were evaluated on the basis of precision and recall factors. The model however, underwent the issues of overfitting. In order to overcome this issue; cross validation technique was used in addition to training and testing of the model. A total of five cross validations were implemented and all the accuracies so obtained were compared for an optimized model. The implementation of SVM resulted into generating higher accuracies with optimized weighted F1 score. The prediction model also stated various reasons that were considered as the primary reason for car crashes.

In another work proposed by authors (Pillajo-Quijia et al.; 2020) they conducted a survey on victims that were run due to drivin g heavy vehicles. Throughout the literature survey conducted in the research, the author tried to highlight that heavier the vehicle; more likely were its chances to collide. For the purpose of implementation; the dataset was acquired from Kaggle repository that included heavy truck crashes that took place in Spain between the years of 2000 and 2008. The severity prediction model included various reasoning's that were given by the author which included the possible reasons of accident severity. However, the implementation was carried out using random forest, decision trees, KNN and SVM. A total of 25 variables were selected as the basis for accident prediction. The final evaluation of the model was done using accuracy and precision factors. It was observed that the SVM model depicted higher superiority in terms of the accuracy so obtained. In addition to this, the author also considered external factors such as not wearing a seatbelt while driving to be as one of the most influential factors that resulted into accidents.

Another machine learning based research work included the work proposed by authors (Umer et al.; 2020) wherein they considered the car crashes that took place in 49 different states on US. The dataset was acquired from Kaggle repository and included car crashes from Feb 2018 to June 2020. A larger dataset resulted in optimizing the process of accident detection since machine learning algorithms work well on huge datasets. A total of six classifiers were used namely; logistic regression, KNN, SVM, decision trees, Naïve Bayes and random forest. The overall performance of the model was evaluated and compared using accuracy and precision factors. In addition to these external parameters such as ambulance services were also taken into consideration while delivering and dealing with road mishaps. However, in comparison to the six classifiers used; SVM generated higher results of precision with an accuracy score of 97 percent.

Figure 3 below gives an overall methodological workflow that was observed in accident prediction using machine learning.

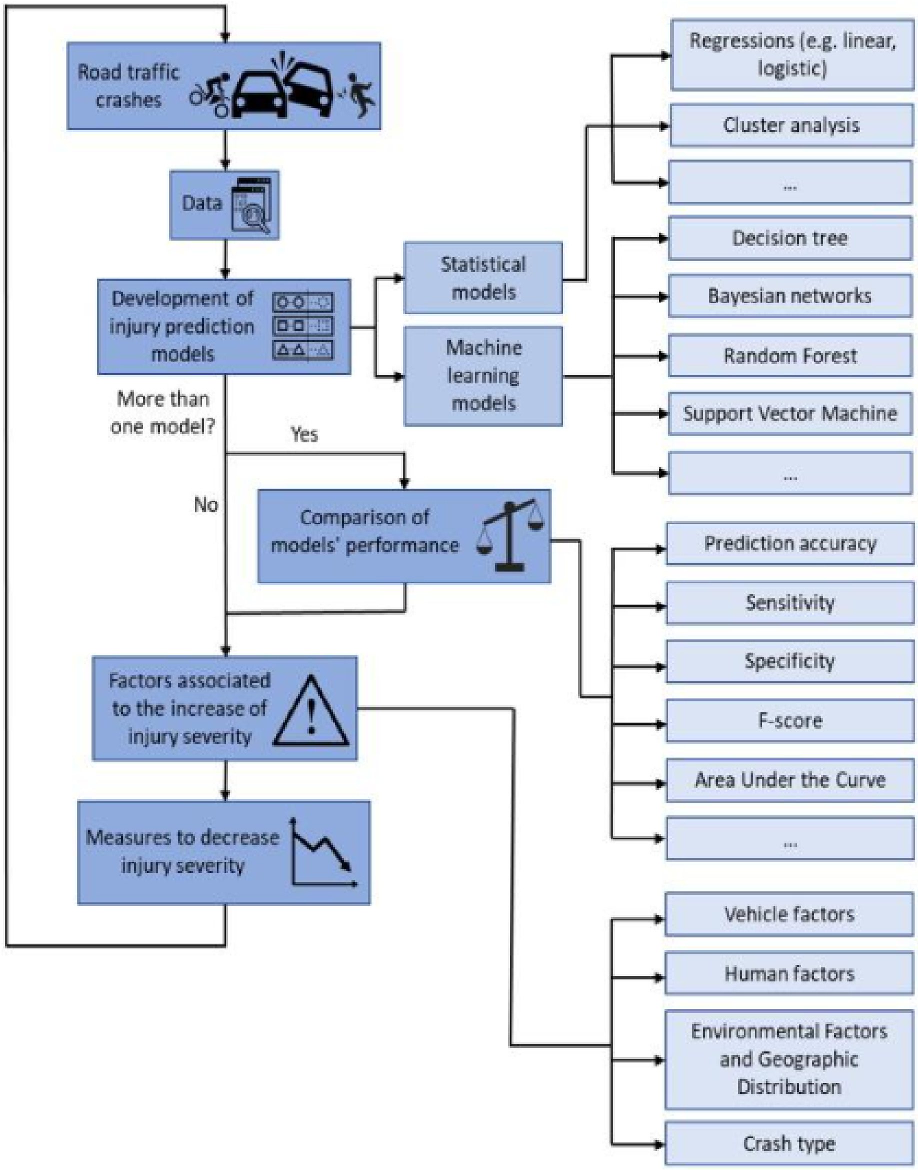


Figure 3: Methodological Workflow Followed by Multiple Research Scholars

2.3 Research Niche

In the study performed by authors (Chen and Chen; 2020) They compared the performance of Logistic Regression, Decision Tree and Random Forest considering factors such as gender, age, vehical type, alcohol consumption, journey type, occupation, weather condition, location, speed limit and obstacles in between to be the most influential that affect the severity of the accidents. They found that the Random Forest was the best-performing out of the three algorithms. Another study by authors (Singh et al.; 2018) conducted research on accident severity in India on the input variables using Random

Forest and Decision Tree ML models here the the parameters considered were crash position, the action of steering, and the occupation of the driver. In this research the problem of data imbalance was also addressed, but the data set used was very small and contained only fewer values. Random Forest was found to be the performing better than decision tree.

In another study by (Jiang et al.; 2020) authors explores the Association Rule Mining (ARM) a method used for data mining and identifying critical factors and mentioned the limitations of this technique. Further, their work extended in developing a framework based on ARM method of data mining and overcoming its drawback. Authors also used Geographical Information System (GIS) to explore relation between affecting key factors and severitity of the injury. Another study by (Liao et al.; 2018) talks about how autonomous vehicles can help to lower the severity of the crash, by making decisions during critical situations, the research explores 3 variations of SVM algorithm to evaluate the speed of the accelerator so that minimum crash injury could be recorded. In another study by (Jeong et al.; 2018) a hybrid method for crash classification was explored to improve the performance gradient boosting and naive classifier was used in this approach. Bagging aggregation was implemented using geometric mean for classification purpose the factors that were highly considered in this research were Disability factors in the driver , Weather conditions such as rainy and cloudy and Light conditions with respect to night time and daylight.

In another study by authors (Mafi et al.; 2018) research was done to determine the level of injury or the severity of injury in people of different age groups and genders, for this study data mining techniques were used, the factors such as the drivers permit information, the reason for the trip and the type of vehicle were considered. Historical data of five years was used to predict crash and shortfalls of incorrect injuries was addressed. Another study by(Zhang et al.; 2018) talked about the criticism faced by the machine learning algorithm and their performance. In this study statistical methodes were compared with four machine learning algorithms. The severity of the crash, and speed limits were considered along with the working and hospital zones near accident prone areas were taken into consideration. The authors aimed to compare the performance of this model on the historical data of the state of Florida. Random forest was found to be the best performing model among all. In research by (Mokhtarimousavi et al.; 2019) authors talked about how important it is to identify the work crash zones, for this purpose they investigated the severity and contributing factors using a parametric approach based on logit modelling framework and no parametric approach based on SVM machine learning model. The found that non parametric approach along with other metaheuristic algorithms give better performance. They successfully identified the crash zones with respect to parametric approaches so that countermeasures could be taken. In another study by (Mokoatle et al.; 2019) discussed accidents occurring at places of interest such as malls, theaters, schools and other points of interest using multivariate logistic regression and XG Boost classification method. This study successfully pinpointed factors such as curves and turning points at this location causing crashes. The accident report of South Africa was analysed and impact injuries were endured upon. Authors (Rezapour et al.; 2020) analyzed the two-wheeler (motorcycle) crashes in this paper. Severity of crash in motorcycle is higher and sometimes fatal when compared to car crash. Author aim to identify the cause of these accidents bu using binary logistic regression as parametric and classification tree as non parametric approach. Some of the influential parameters found were surface condition of roads, speed compliance of driver and alcohol involvement.

2.4 Observational Drawbacks

Throughout the literature survey so conducted, following are the drawbacks witnessed in the existing systems:

- Dataset acquired from the repository contains imbalanced form of data. The NULL values are however removed and cleaned; but the dataset still remains imbalanced and therefore consists of redundant noise. Very less amount of research has been observed to be contributed wherein the issue of imbalanced data could be resolved
- The existing models have witnessed certain constraints and limitations in the form of input so given to the model. The input factors demand pre-requisites; and absence of any requirement would thereby result in misled predictions being generated
- Another limitation in the existing systems includes the number of accident records being maintained on the training dataset. With fewer records, it becomes difficult for an ML to solve dynamic calculations

Therefore, the above drawbacks are taken into consideration and the thesis is thereby proposed wherein the issue of data imbalance is resolved using the conceptual theory of SMOTE. SMOTE is a feature selection technique and the author proposes the implementation of random forest to resolve the same. In addition to data imbalance problem, all the available and respective influential factors are looked upon to deploy a model that could generate accurate predictions. Finally, the challenge of limited dataset can further be resolved using data augmentation techniques. However, the implementation of data augmentation technique is considered to be as the scope of the thesis and is further placed in the section of future works.

3 Methodology

This section of the thesis summarizes the methodologies used to implement the proposed model.

3.1 Logistic Regression

Logistic regression is considered to be a sub category of supervised learning algorithm that comes under the category of classification techniques. It works on the basis of target variables and helps to predict the class values. However, the nature of the target variable is dichotomous that eventually indicates that the final results would have only two possible outcomes. As a result of which, the dependant variable is declared to be binary in nature that would decode the prediction to be as either 1 (YES) or 0 (NO). The calculation of model prediction is done using P(Y=1) as a function of X. In addition to binary logistic regression; when an output variable generates more than two target variables; it is termed as multinomial logistic regression (tutorialspoint.com). However, when there is a possibility of generating more than three classification target variables; it is termed as ordinal logistic regression. It is also worthy to note here that logistic regression is considered to be as the most simplest form of machine learning algorithm that can be applied in cases where a research scholar has to make predictions with an output generation of two variables.

3.2 Decision Tree

Unlike the implementation of logistic regression; decision trees tend to be a sub category of predictive modelling which is applicable in many areas. The working implementation begins with an algorithmic approach that tends to split the initial dataset with respect to different and varying conditions being provided to it. It is one of the most powerful algorithms that come under the conceptual theory of supervised learning. However; its implementation can follow classification as well as regression issues. A decision tree primarily consists of decision nodes that further get split into leaves and branches. The outcome from the branches is regarded to as the prediction generated by a decision tree. For instance; predicting a person is healthy with respect to his age can be done by his personal information being provided in the form of age, habits and exercise. Figure 4 below gives an illustration of the same:

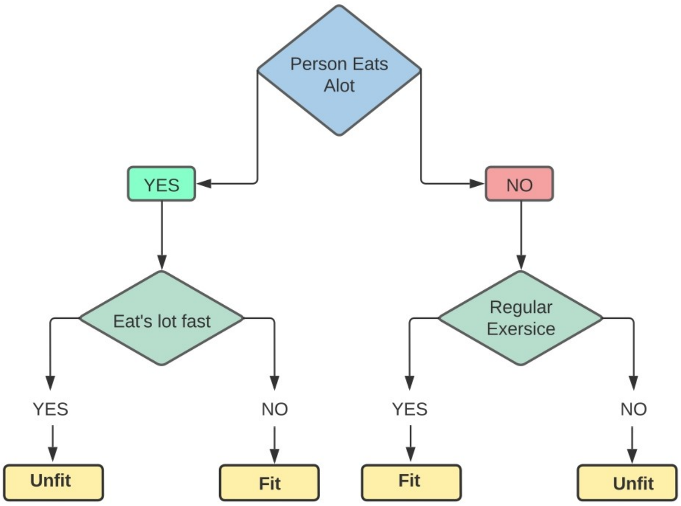


Figure 4: Working of a Decision Tree

3.3 Random Forest

Another supervised learning based classification algorithms is the implementation of random forests. However, its working can be used for predicting classification as well as regression problems. A random forest is considered to be a collection of individual trees that are more robust in nature. The algorithm works on initially responding to one specific decision tree followed by the random forest. The overall implementation of a random forest tends to overcome the major issue of over-fitting that however occurs in a decision tree. However, the predictions in each tree are done through the process of voting. Following are the steps to be taken during the implementation of random forests:

- Select a dataset
- Construct a single decision tree from it
- Execute the process of voting
- Iterate the process of voting until multiple decision trees are formed

- Select the most voted result as the final prediction made by the algorithm

Figure 5 below [5] illustrates the working process of a random forest:

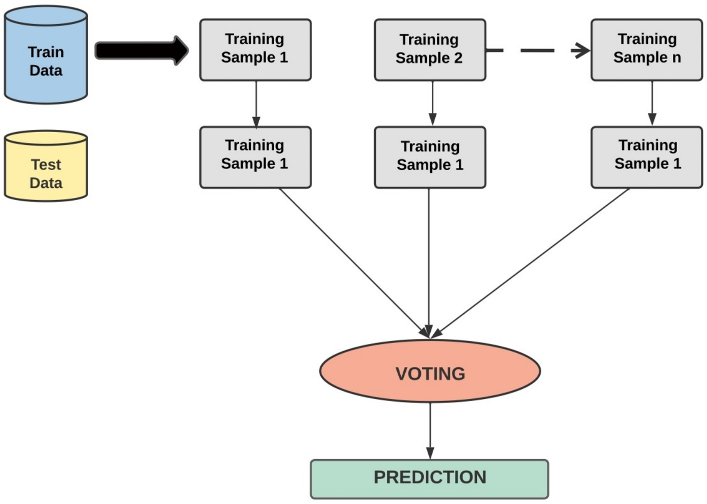


Figure 5: Working of a Random Forest

3.4 Ensemble Learning

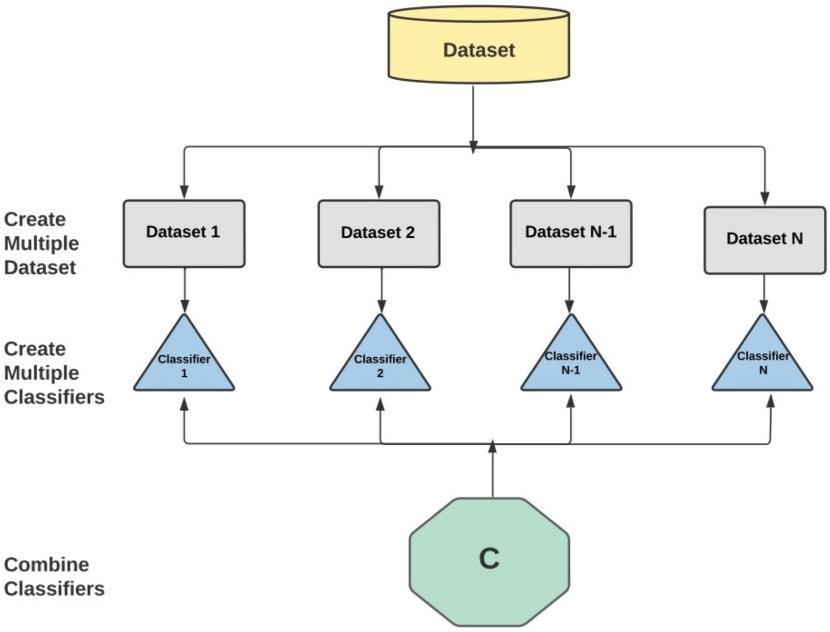


Figure 6: Working of an Ensemble Model

The working procedure of ensemble learning follows the implementation of stacking multiple algorithms together so that the overall accuracy and precision factors of a system

[5] https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_random_forest.htm

model could be retained. A stacking algorithm can henceforth be combined using estimators and Meta estimators. For the purpose of implementation of the proposed thesis, following are the estimators and Meta estimators used:

- Estimators: logistic regression and decision trees
- Meta Estimators: random forest

The working implementation of ensemble learning is depicted in figure 6.

4 Design Specification

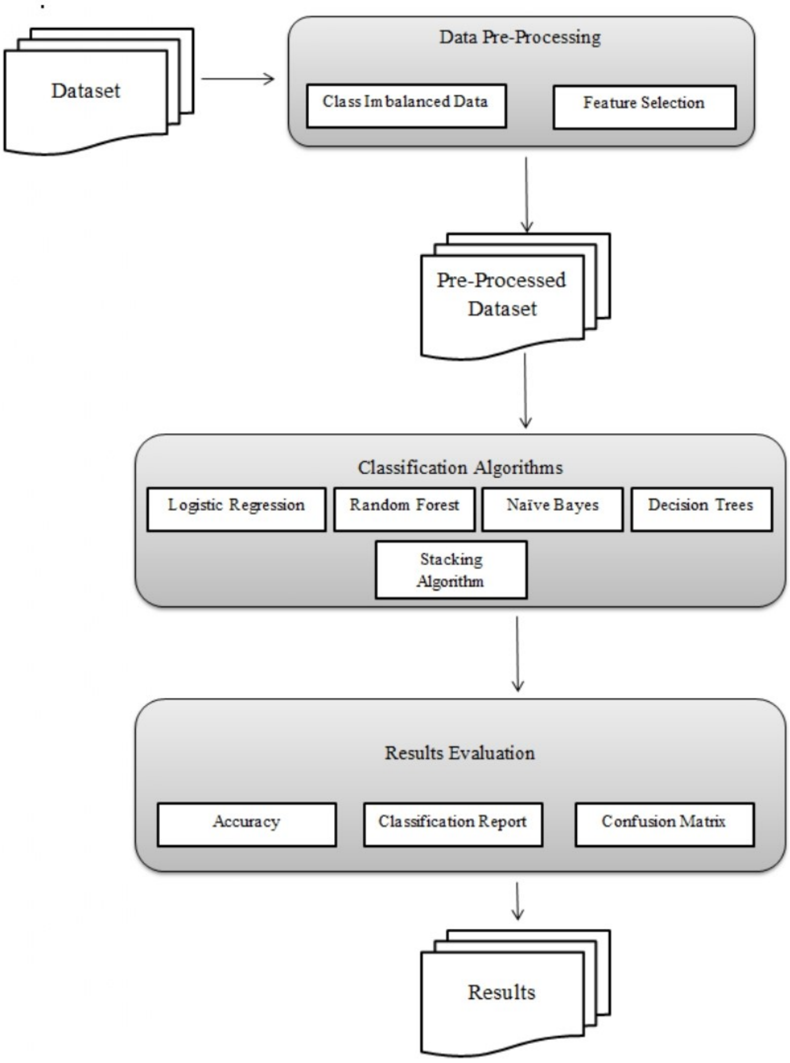


Figure 7: Architecture of the Proposed System

The primary objective of the methodology is to build a system model that could predict the occurrence of accidents that might happen on the road due to certain factors that might contribute to the mishap. For this purpose, the thesis focuses implements the methodologies as explained in the previous section. However, this section of the thesis presents the overall design of the research. The architectural diagram in figure 7 depicts the various phases involved in the system design of the model.

In the primary phase, the data is collected from a respective repository and pre-processed to filter out the most influential factors contributing to severity of accidents. In the next stage, four classification algorithms and one stacking algorithm is applied to build the predictive model. Finally in the evaluation phase; certain metrics are used so as to determine the accuracy and generate an optimized model.

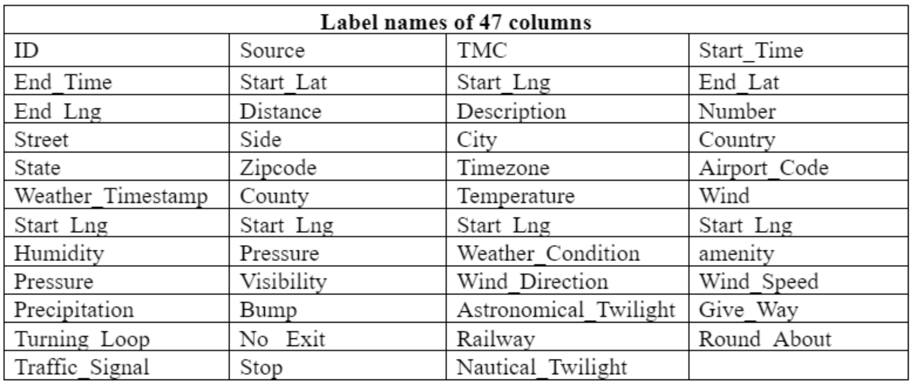
In this manner; the proposed system model is build and developed so that machine learning algorithms could be applied and a knowledge based system could be built so as to fulfil the purpose of the research.

5 Implementation

This section of the thesis highlights on the process of implementation that occurs on each stage.

5.1 Dataset Used

The dataset used in the study is taken from Kaggle [6] repository and contain information on car accidents taking place in 49 states in USA from Feb 2016 to Dec 2021. However, the data in Kaggle repository is collected through multiple traffic API's that tends to capture accident records through network sensors and traffic cameras. An approx. of 4.2 million records of car crashes are mentioned in the dataset. The distribution of the dataset is done on a .CSV file consisting of 47 columns. The table in figure 8 highlights the labels of each column:



Label names of 47 columns			
ID	Source	TMC	Start_Time
End_Time	Start_Lat	Start_Lng	End_Lat
End_Lng	Distance	Description	Number
Street	Side	City	Country
State	Zipcode	Timezone	Airport_Code
Weather_Timestamp	County	Temperature	Wind
Start_Lng	Start_Lng	Start_Lng	Start_Lng
Humidity	Pressure	Weather_Condition	amenity
Pressure	Visibility	Wind_Direction	Wind_Speed
Precipitation	Bump	Astronomical_Twilight	Give_Way
Turning_Loop	No_Exit	Railway	Round_About
Traffic_Signal	Stop	Nautical_Twilight	

Figure 8: Dataset Distribution Over Columns

The number of records against each target class is depicted in figure 8 wherein the value 1 indicates less severity whereas the value 4 indicates maximum severity.

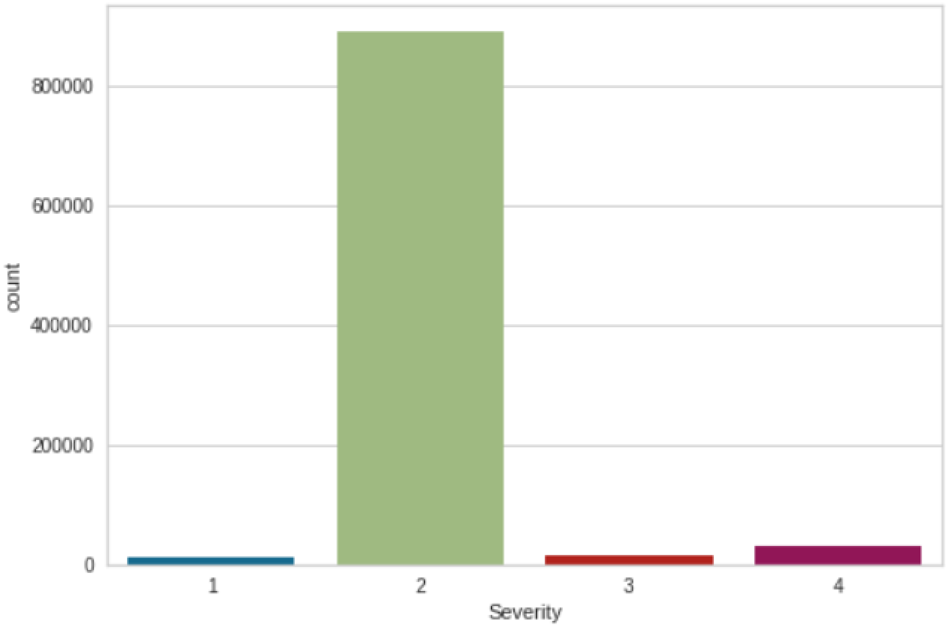


Figure 9: Dataset Distribution with respect to Severity on an Imbalanced Dataset

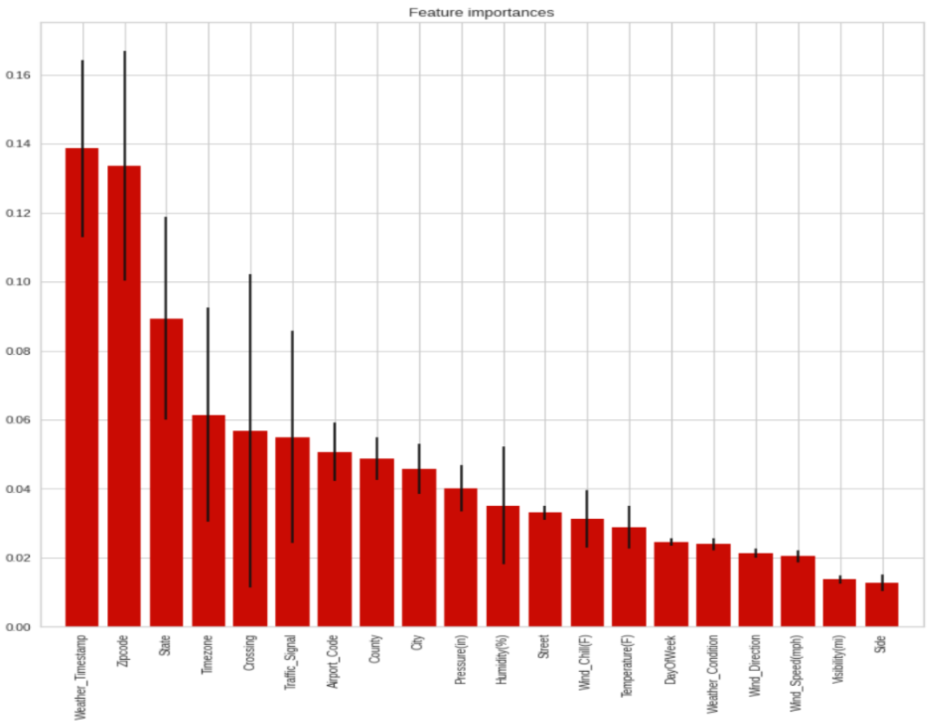


Figure 10: Reduction of 47 columns to 20 columns using feature selection through Random Forest

5.2 Data Pre-Processing

The obtained dataset gathered from the Kaggle repository consists of raw, unstructured and imbalanced data. This data tends to increase the computational complexity of the system and also adds on unnecessary training time to the dataset. Hence, this redundancy in the data must be removed. This process of removing irrelevant data from the dataset is termed as data pre-processing and thereby plays a significant role in enhancing the overall computational process of the model. This stage of data pre-processing majorly involves cleaning the dataset so that the redundancy is removed; normalizing the dataset so that NULL values are dropped; feature selecting the dataset so that only respective columns and their attributes are selected; and balancing the imbalanced dataset. However, for the purpose of implementation of the proposed thesis; data cleaning is performed so that NULL values are dropped and feature selection is performed so that only respective columns of the dataset are used and SMOTE is performed to balance the dataset. For the purpose of feature selection; the thesis implements the concept of random forest; wherein only 20 columns with respect to their attributes are selected and the rest 27 columns with respect to their attributes are discarded. Figure 10 depicts the generated bar graph wherein only 20 columns from 47 columns are selected from the dataset using random forest.

5.3 Data Balance

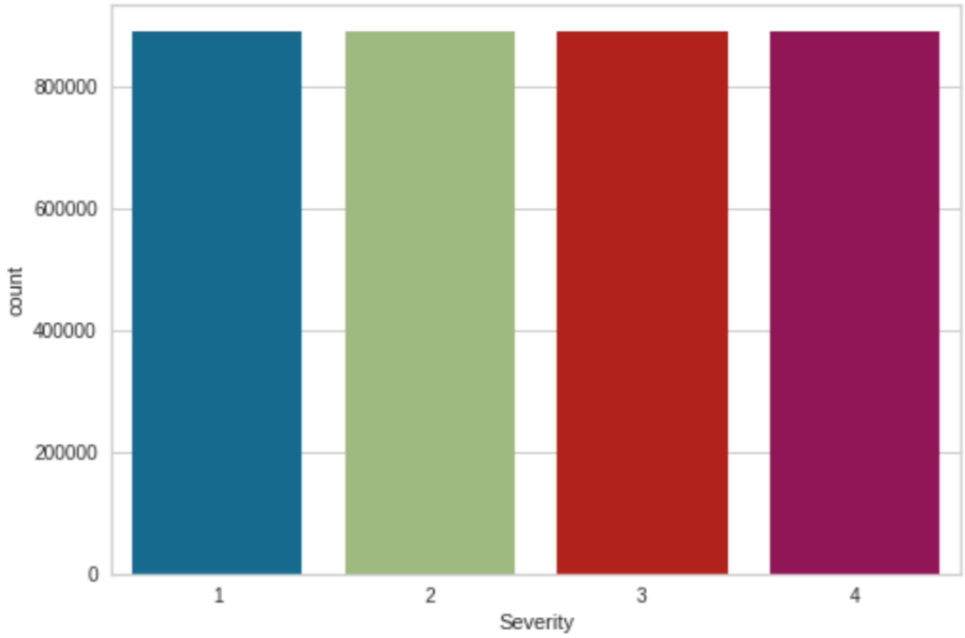


Figure 11: Distribution with respect to severity on a balanced dataset using SMOTE

The data obtained from the repository is imbalanced in nature; and hence needs to be balanced so that respective algorithms can be performed on them. For this purpose, a simple approach of duplicating minority class is performed using SMOTE. SMOTE is

[6] https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

abbreviated to Synthetic Mining Oversampling Technique. The problem of data imbalance is resolved in the dataset by oversampling the minority classes of the dataset; so that the generated samples could perfectly fit into the model and balance the dataset. This duplication technique helps to synthesize the created samples so that they can be matched with the samples that are in minority. Through the implementation of SMOTE technique; the bar graph obtained in figure 9 could further be balanced to be depicted as shown in figure 11.

5.4 Data Visualization

The process of data visualization helps to analyse patterns by gaining historical insights from the dataset. Visualizing this data and illustrating it using bar graphs, pie charts etc. tends to provide more details on every attribute from the columns of the dataset. The bar graphs below in figure 12 depicts the count of accidents with respect to 4 severities as per months and days of weeks.

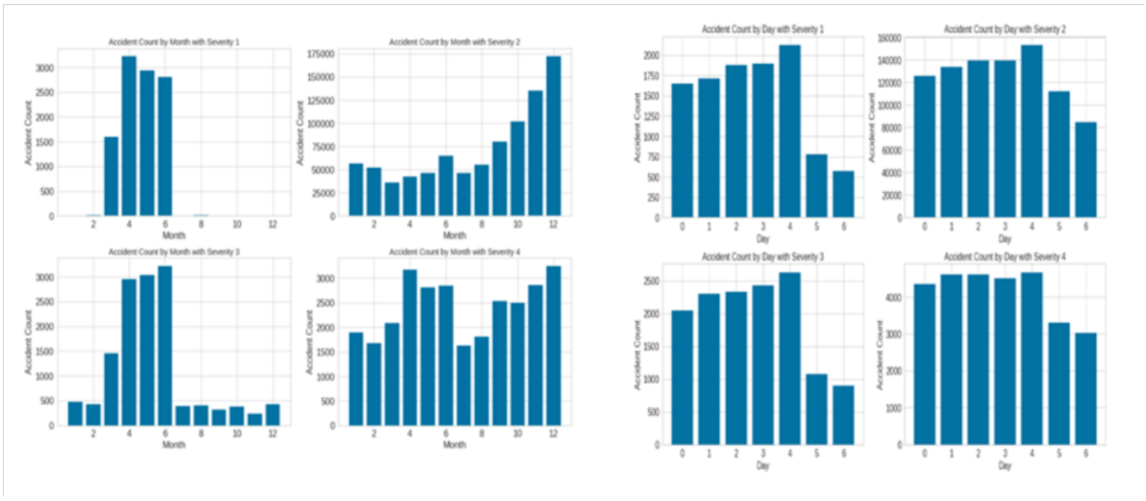


Figure 12: 4 Types of Severities as per Months and as per days of Weeks

The bar graph in figure 13 depicts the count of accidents with respect to their weather conditions.

5.5 Data Train, Test and Split

The effectiveness of the model so created is determined when the dataset undergoes a process of training and testing. For this purpose, a set of algorithms are used and a portion of the dataset is split. This is done so that the developed model can be enhanced so as to reach higher levels of accuracy. For this purpose, the thesis splits the data into 80 percent for training purpose and 20 percent for testing purpose. In addition to this; the model can accomplish more robustness on application of the validation technique. The implementation of the thesis is initially carried out by resampling the imbalanced data using SMOTE so that a small percentage of fatal and serious injuries data are

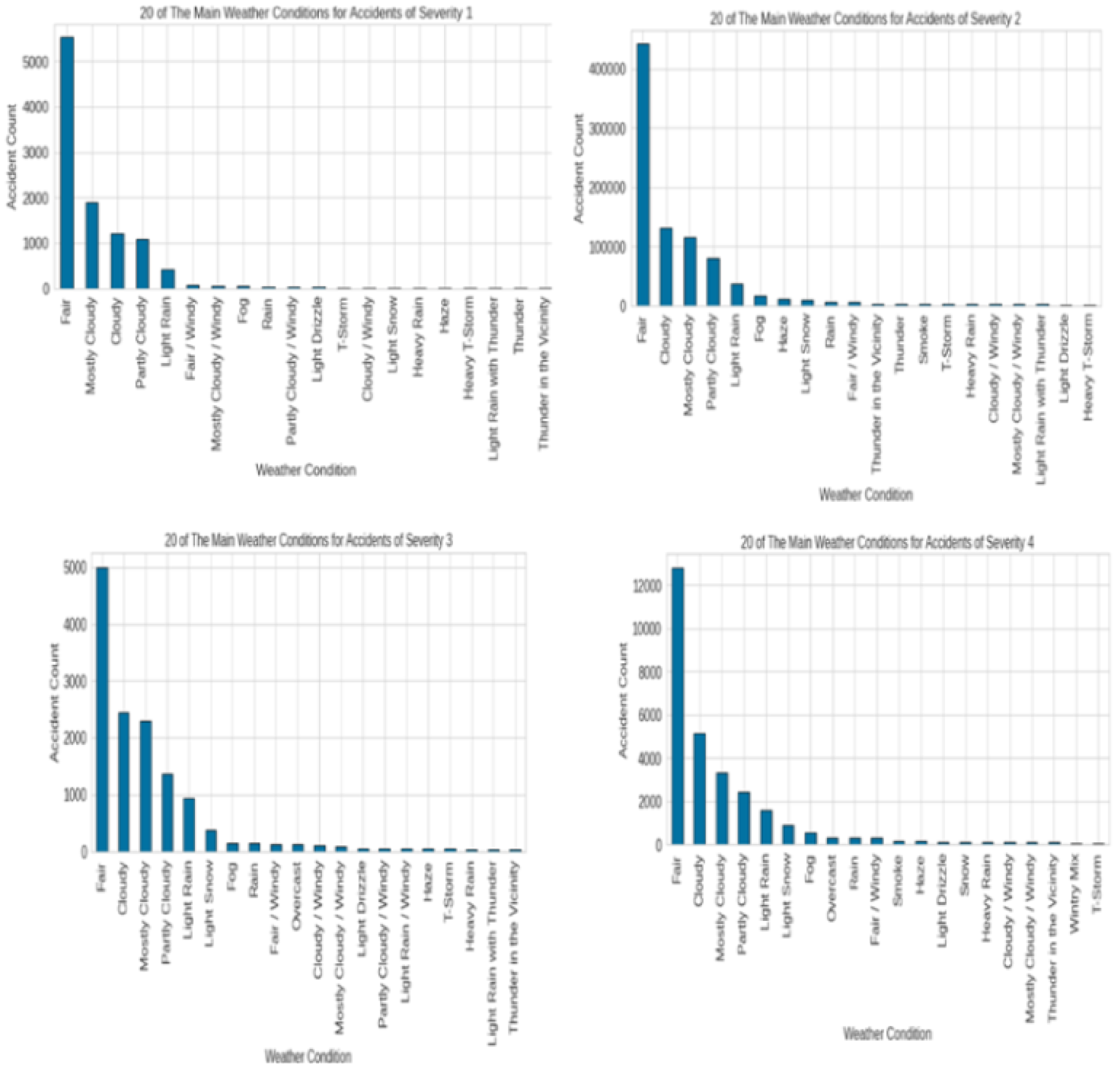


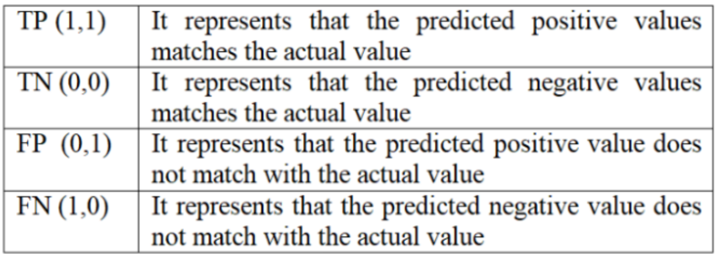
Figure 13: Accident Count with respect to Weather Conditions (cloudy, hazy, windy, rainy, snow, storm and fog.) with 4 Severities

duplicated. After the implementation of SMOTE; the model undergoes training, testing and validating the data for 10 fold-cross validation. The train, test and split of this data are done through algorithms so selected. For this thesis; four algorithms and one stacking algorithm are used to determine the optimization accuracy of the model. Finally, a set of evaluation parameters are executed so as to determine and compare which algorithm performed better in terms of precision.

6 Evaluation Parameters

The creation of the assessment matrix is a requirement that must be completed in order to set the evaluation parameters for improved outcomes. The system's entire performance is represented by the matrix and the evaluation parameters, allowing the efficiency of the system to be calculated and enhanced. However, each system model's parameters must be established in accordance with the system's methodological workflow. The list of evaluation criteria for the proposed thesis' execution is as follows:

- Confusion Matrix: The results collected from the confusion matrix are used to calculate the performance of any system model. It typically takes the form of a tabular representation with precise numbers filled in for both the actual results and the outcomes that were projected. The four fundamental features of a confusion matrix are as follows:



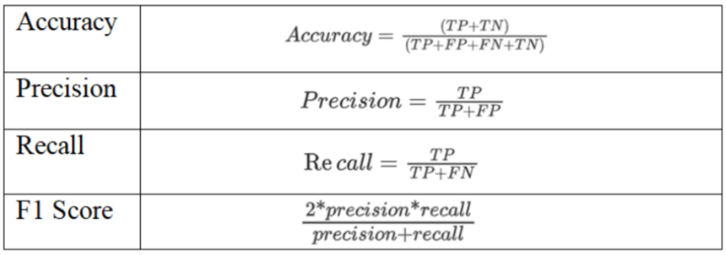
TP (1,1)	It represents that the predicted positive values matches the actual value
TN (0,0)	It represents that the predicted negative values matches the actual value
FP (0,1)	It represents that the predicted positive value does not match with the actual value
FN (1,0)	It represents that the predicted negative value does not match with the actual value

Figure 14: Confusion Matrix Features Table

- Classification Report: To provide information on the values derived from the accuracy, recall, F1Score, and precision factors, a classification report is used. The formulas used to determine each parameter are shown in the table below:

7 Experimental Analysis and Results

The experimental analysis that was done on the system model to get the desired accuracy using the previously mentioned evaluation parameters is included in this portion of the thesis.



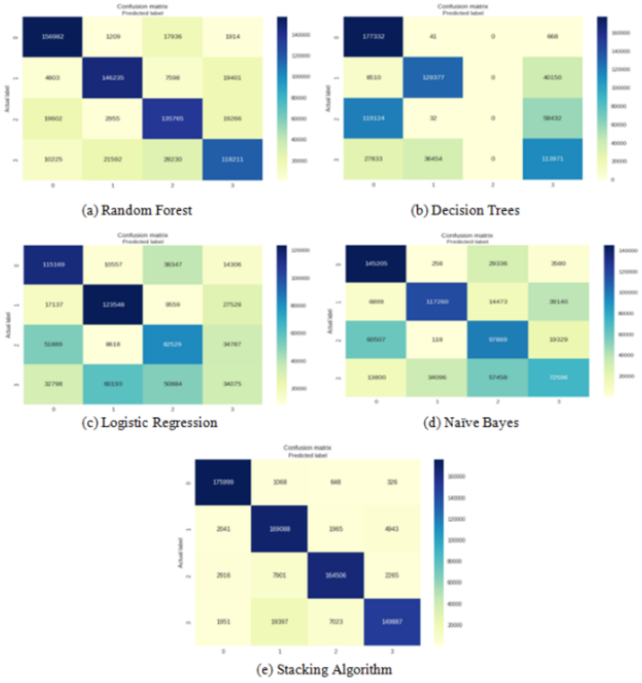
Accuracy	$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$
Precision	$Precision = \frac{TP}{TP+FP}$
Recall	$\text{Re}\,call = \frac{TP}{TP+FN}$
F1 Score	$\frac{2*precision*recall}{precision+recall}$

Figure 15: Accuracy, Precision, Recall and F1 Score

7.1 Algorithms depicting Confusion Matrix

The confusion matrix so generated for the proposed thesis is based on the concepts of multi-class classification technique; wherein only one output is more likely to get generated from a set of a respective class. Hence, with four severities into consideration the confusion matrix shall give real as well as predicted values. However, the outputs so generated shall be positive predicted for only one severity at one time.

The confusion matrix of four algorithms with one stacking algorithm is depicted in figure 16



(a) Random Forest

(b) Decision Trees

(c) Logistic Regression

(d) Naïve Bayes

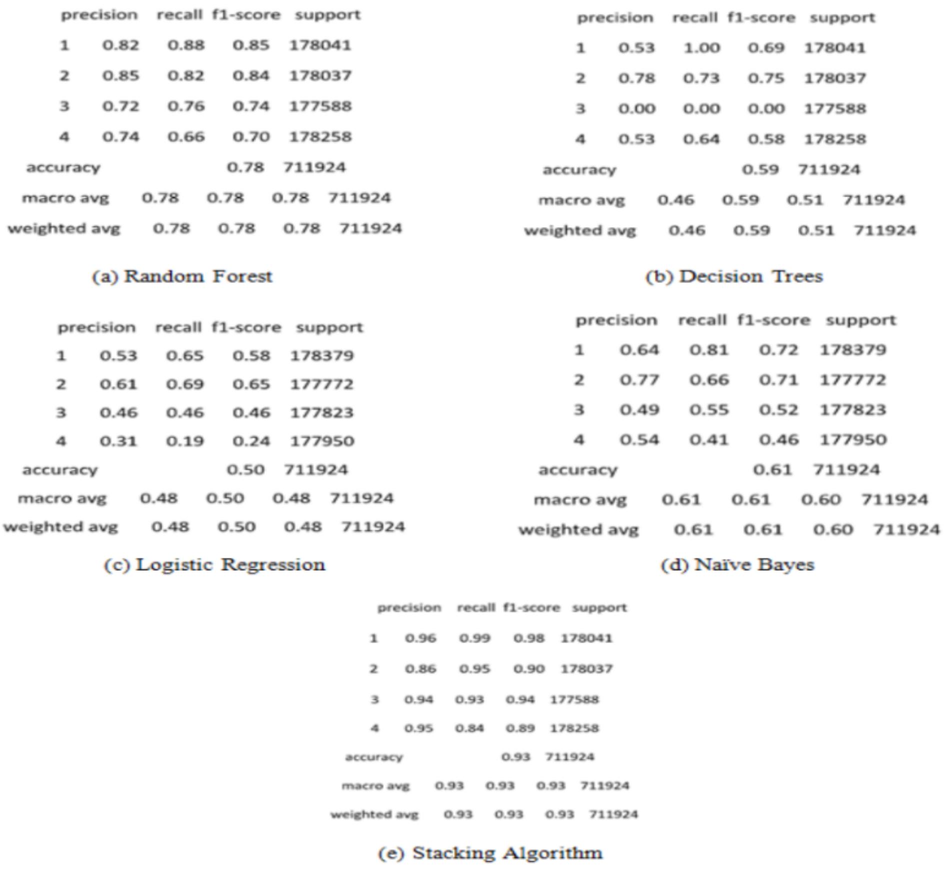
(e) Stacking Algorithm

Figure 16: Confusion Matrix

A total of four severities are taken into consideration for the implementation of the proposed thesis; with severity 1 depicting least severe and severity 4 depicting maximum sever reason for an accident to occur. The indexing of a confusion matrix begins with 0; hence for this purpose, severity 1 shall be indexed as "0" and so on. Considering the confusion matrix of random forest; confusion matrix (a) from figure 16, severity "0" gives a prediction of 156982 cases (blue highlighted) of positive prediction; whereas all the other values in the same column such as 4803, 19602 and 10225 are falsely predicted by random forest for severity “0”. In the same manner, random forest correctly predicts 146235 instances of positive predictions for severity “1”; 135765 for severity “2” and 118211 for positive prediction of severity “3”. Rest all the values so generated by random forest for severities 0, 1, 2 and 3 are wrong predictions made by the algorithm. Hence, the same observations can be made for confusion matrix so generated by decision trees, logistic regression, Naïve Bayes and the stacking algorithm.

7.2 Algorithms depicting Classification Reports

Detailed information on the accuracy, precision, recall, and F1 Factor of the values thus acquired is provided in a classification report. The accuracy values achieved by using the appropriate algorithms are shown in Figure 17



```
        precision   recall  f1-score   support
  1       0.82       0.88     0.85     178041
  2       0.85       0.82     0.84     178037
  3       0.72       0.76     0.74     177588
  4       0.74       0.66     0.70     178258
 accuracy                     0.78     711924
 macro avg      0.78     0.78     0.78     711924
weighted avg     0.78     0.78     0.78     711924
```

(a) Random Forest

```
        precision   recall  f1-score   support
  1       0.53       1.00     0.69     178041
  2       0.78       0.73     0.75     178037
  3       0.00       0.00     0.00     177588
  4       0.53       0.64     0.58     178258
 accuracy                     0.59     711924
 macro avg      0.46     0.59     0.51     711924
weighted avg     0.46     0.59     0.51     711924
```

(b) Decision Trees

```
        precision   recall  f1-score   support
  1       0.53       0.65     0.58     178379
  2       0.61       0.69     0.65     177772
  3       0.46       0.46     0.46     177823
  4       0.31       0.19     0.24     177950
 accuracy                     0.50     711924
 macro avg      0.48     0.50     0.48     711924
weighted avg     0.48     0.50     0.48     711924
```

(c) Logistic Regression

```
        precision   recall  f1-score   support
  1       0.64       0.81     0.72     178379
  2       0.77       0.66     0.71     177772
  3       0.49       0.55     0.52     177823
  4       0.54       0.41     0.46     177950
 accuracy                     0.61     711924
 macro avg      0.61     0.61     0.60     711924
weighted avg     0.61     0.61     0.60     711924
```

(d) Naïve Bayes

```
        precision   recall  f1-score   support
  1       0.96       0.99     0.98     178041
  2       0.86       0.95     0.90     178037
  3       0.94       0.93     0.94     177588
  4       0.95       0.84     0.89     178258
 accuracy                     0.93     711924
 macro avg      0.93     0.93     0.93     711924
weighted avg     0.93     0.93     0.93     711924
```

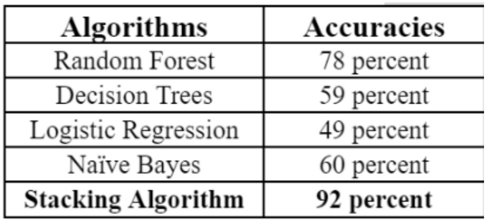
(e) Stacking Algorithm

Figure 17: Classification Report

The precision factor, recall, F1 Score, support and accuracy that is generated for each severity by random forest is depicted in figure 7.2 labelled as classification report. The values generated for severity 1 are 0.82, 0.88, 0.85 and 178041 respectively for precision, recall, F1-score and support. The overall accuracy for all the severities that is generated by an implementation of random forest is observed to be 78 percent. In this way, the precision factor of each severity is mentioned in the classification report. Hence, the same observations can be made for classification reports so generated by decision trees, logistic regression, Naïve Bayes and the stacking algorithm.

7.3 Algorithms depicting Accuracies

The table in the figure 18 depicts the accuracies obtained from respective algorithms.



Algorithms	Accuracies
Random Forest	78 percent
Decision Trees	59 percent
Logistic Regression	49 percent
Naïve Bayes	60 percent
Stacking Algorithm	**92 percent**

Figure 18: Algorithms and their Accuracy's

Throughout the implementation of four algorithms and one stacking algorithm, it can be observed that the stacking algorithm performed better with respect to accuracies and the precision factors so generated.

8 Conclusion and Future Work

The primary aim of the thesis is to build an automated system that could detect the severities of an accident and further predict its occurrence. For this purpose, the thesis presents an implementation of four machine learning algorithms along with one stacking algorithm. The study also includes the conceptual theory of SMOTE to balance the dataset; since the data obtained from repository was imbalanced and redundant in nature. Through the conduction of literature survey it was observed that factors such as alcohol consumption, weather conditions, location of the accident, speed of the vehicle, disability portrayed by the driver etc. had a significant impact for any accident to occur. Hence, all the influential factors were taken into consideration throughout the executional process of the thesis. However, it was observed that random forest generated an accuracy of 78 percent; which was considered to be as the highest among the four machine learning algorithms. In addition to this, the stacking algorithm that combined the concepts of logistic regression and decision trees as estimators and random forest as Meta estimator; generated the highest accuracy of 92 percent. Hence, in comparison to all the five algorithms; the implementation working of the stacking algorithm is thereby chosen to be as the optimized model with highest generating accuracy.

The proposed thesis can however be used to predict the occurrence of road accidents by rampaging through the severities caused. However, a limitation of the thesis is the availability of the dataset with respect to the influential parameters (such as mental condition of the driver, pedestrians on the road etc.) that might further contribute to enhance the overall accuracy of the system. Hence, the future work of the same study would be the implementation of data augmentation technique so that multiple copies of the data could be generated.

References

Beshah, T. and Hill, S. (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in ethiopia, *AAAI Spring Symposium: Artificial Intelligence for Development*.

Chen, M.-M. and Chen, M.-C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest, *Information* **11**(5): 270.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE transactions on information theory* **13**(1): 21–27.

Harb, R., Radwan, E., Pande, A. and Abdel-Aty, M. (2008). Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression, *Journal of Transportation Engineering-asce - J TRANSP ENG-ASCE* **134**.

Jeong, H., Jang, Y., Bowman, P. J. and Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data, *Accident Analysis Prevention* **120**: 250–261.

Jiang, F., Yuen, K. K. R. and Lee, E. W. M. (2020). Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and gis technology, *Journal of Safety Research* **75**: 292–309.

Jiang, L., Xie, Y. and Ren, T. (2019). Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis, *Proceedings of the Transportation Research Board 98th Annual Meeting, Washington, DC, USA*, pp. 13–17.

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)*, IEEE, pp. 1–5.

Liao, Y., Zhang, J., Wang, S., Li, S. and Han, J. (2018). Study on crash injury severity prediction of autonomous vehicles for different emergency decisions based on support vector machine model, *Electronics* **7**(12): 381.

Mafi, S., Abdelrazig, Y. and Doczy, R. (2018). Machine learning methods to analyze injury severity of drivers from different age and gender groups, *Transportation research record* **2672**(38): 171–183.

Mannering, F. L. and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions, *Analytic methods in accident research* **1**: 1–22.

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A. and Hadi, M. (2019). Improved support vector machine models for work zone crash injury severity prediction and analysis, *Transportation research record* **2673**(11): 680–692.

Mokoatle, M., Vukosi Marivate, D. and Michael Esiefarienrhe Bukohwo, P. (2019). Predicting road traffic accident severity using accident report data in south africa, *Proceedings of the 20th annual international conference on digital government research*, pp. 11–17.

Pakgohar, A., Tabrizi, R., Khalili, M. and Esmaeili, A. (2011). The role of human factor in incidence and severity of road crashes based on the cart and lr regression: A data mining approach, *Procedia CS* **3**: 764–769.

Pillajo-Quijia, G., Arenas-Ramírez, B., González-Fernández, C. and Aparicio-Izquierdo, F. (2020). Influential factors on injury severity for drivers of light trucks and vans with machine learning methods, *Sustainability* **12**(4): 1324.

Rezapour, M., Molan, A. M. and Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models, *International journal of transportation science and technology* **9**(2): 89–99.

Singh, G., Sachdeva, S. and Pal, M. (2018). Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of indian highways., *Advances in transportation studies* **45**.

Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N. and Madni, H. A. (2020). Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction, *arXiv preprint arXiv:2010.14921* .

Vajari, M. A., Aghabayk, K., Sadeghian, M. and Shiwakoti, N. (2020). A multinomial logit model of motorcycle crash severity at australian intersections, *Journal of safety research* **73**: 17–24.

Wang, X. and Kim, S. H. (2019). Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models, *Transportation research record* **2673**(9): 640–653.

Yin, C., Xiong, Z., Chen, H., Wang, J., Cooper, D. and David, B. (2015). A literature survey on smart cities, *Science China Information Sciences* **58**(10): 1–18.

Yuan, Y., Yang, M., Guo, Y., Rasouli, S., Gan, Z. and Ren, Y. (2021). Risk factors associated with truck-involved fatal crash severity: Analyzing their impact for different groups of truck drivers, *Journal of safety research* **76**: 154–165.

Zhang, J., Li, Z., Pu, Z. and Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods, *IEEE Access* **6**: 60079–60087.