CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Road Traffic Accident (RTA) is an unexpected event that unintentionally occurs on the road which involves vehicle and/or other road users that causes casualty or loss of property. Over 90% of the world's fatalities on roads occur in low and middle income countries which account for only 48% of world's registered vehicles. The financial loss, which is about US$518 billion, is more than the development assistance allocated for these countries. While developed rich nations have stable or declining road traffic death rates through coordinated correcting efforts from various sectors, developing countries are still losing 13% of their gross national product (GNP) due to the endemic of traffic casualties. World Health Organization (WHO) fears, unless immediate action is taken, road crashes will rise to the fifth leading cause of death by 2030, resulting in an estimated 2.4 million fatalities per year.

Accidental severity analysis involves three factors that are number of injuries, number of casualties, and destruction of property. Authors take severity level independently and consider four options like light injury, severe injury, fatal, and property damage. Road traffic accidents are a major cause of injuries, deaths,permanent disabilities and property loss. It not only affects the economy it also affects the healthcare system because it puts a burden on the hospitals. Statistics shown by the ministry of public security of china from the years 2009 and2011, traffic accidents caused an average of 65123 people to lose their life and 255540 got injuries annually. Identifi-cation of primary factors affecting road accident severity is required to minimize the level of accidental severity. Acci-dental Severity does not happen by chance; there are patterns that can be predicted and prevented. Accidental events can be analyzed and avoided. Being one of the major issues of accident management, accident severity prediction plays an important role to the rescuers in evaluating the level of severity in traffic accidents, their potential impact, and in implementing efficient accident management procedures.Accidental severity analysis involves three factors that are number of injuries,

number of casualties, and destruction of property. Authors take severity level independently and consider four options like light injury, severe injury, fatal, and property damage. The accidental severity level is definedas injury, possible injury, and property damage .

The fast-growing population, increasing motorization, and rapid urbanization have made road users more vulnerable to accidents which are on the rise globally and are a significant challenge to deal with. Road traffic accidents impose a substantial human, financial, and economic burden on society. According to the report published by the World Health Organization (WHO), road crashes result in an average annual human death of approximately 1.3 million and 20 to 50 million serious and minor injuries. Countries and international organizations have designed technologies, systems, and policies to prevent accidents. Whilst several measures have been adapted to make the roads safe for users, there is no such real-time warning system available to guide the user about the probability of an accident. Similar to multiple other countries in the world, New Zealand also has road signs, such as high crash areas, slippery when wet, speed limits, etc., which warns users regarding road safety. Apart from the road signs, prediction of accident rates on road links is an important input to road safety improvement. The prediction model explores the relationships between crash severity injury categories and contributing factors such as driver behavior, vehicle characteristics, roadway geometry, road-environment conditions and cause of accidents, which could then be addressed by transportation policies. Vehicular crash data can be used to model both the frequency of crash occurrence and the degree of crash severity. Crash frequency refers to the prediction of the number of crashes that may occur on a specific road segment or intersection over a certain time frame.

The accidental severity level is defined as injury, possible injury, and property damage. In the last two decades, accidental severity is one of the popular research areas. Researchers were applying different statistical approaches for road accident classification. These techniques help in analyzing the cause of road accidents. To uncover the underlying relationship between a road accident and the contributing factors, ML based models have been studied in recent days. The primary focus of this study is to analyze a set of widely used

ML models in terms of their prediction accuracy with the variation of contributing factors. Therefore, this research aims to study different ML algorithms and compare the performance of these algorithms which can be considered to predict road accidents and its severity.

1.2 ENVIRONMENT FACTORS CAUSING ACCIDENT

Environmental factors causing accidents is a term used to capture all the causes of a workplace incident that can be directly attributed to the working environment. These include features of the natural environment, aspects of workplace design, as well as the machinery or equipment used on the jobsite. Some of the environmental factors that are contributing to the accident are listed below.

i. Road Surface

Road surface or pavement distress is a condition of road damage such as cracking, patching and potholes, surface deformation, and surface deflect. This project discusses the effect of pavement distress on the risk of accidents based on collective damage conditions of pavement surfaces. Road surface condition is a matter of concern especially for traffic safety. Poor road surface conditions (large potholes and deep cracks exist, discomfort at slow speeds) can lead to high rates of accidental fatality, especially in high speed roads in single vehicle and multi vehicle accident types.

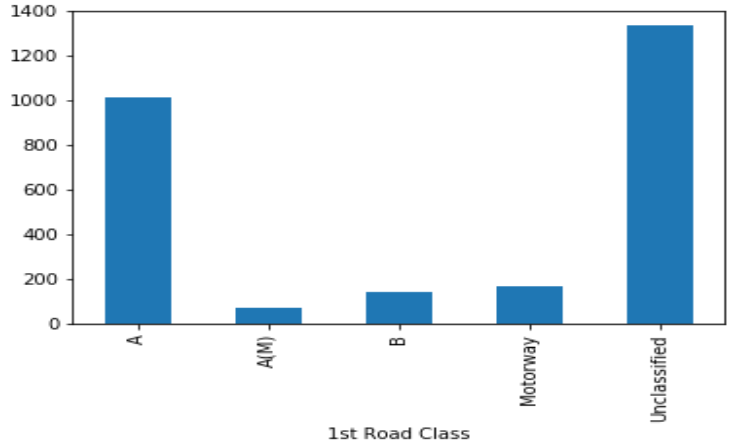

Fig 1.1 Number of accident in different road type

The figure 1.1 mimics the number of accidents that happen on different types of road surfaces in Europe. The A, A(M), B and the motorway are the roads that are maintained by the highway authorities. The unclassified are the soil roads in the rural and metro areas without the maintenance, which kinds of roads led to the most of the accidents.

ii. Lighting Condition

Inadequate lighting can lead to serious injuries resulting from slips and falls, collisions or other similar incidents which can break bones, cause serious ligament damage, concuss victims or even threaten their lives. In addition to headaches, your workers may also experience neck, back and shoulder pain if they have to hunch over and strain to see items as a result of incorrect lighting design.

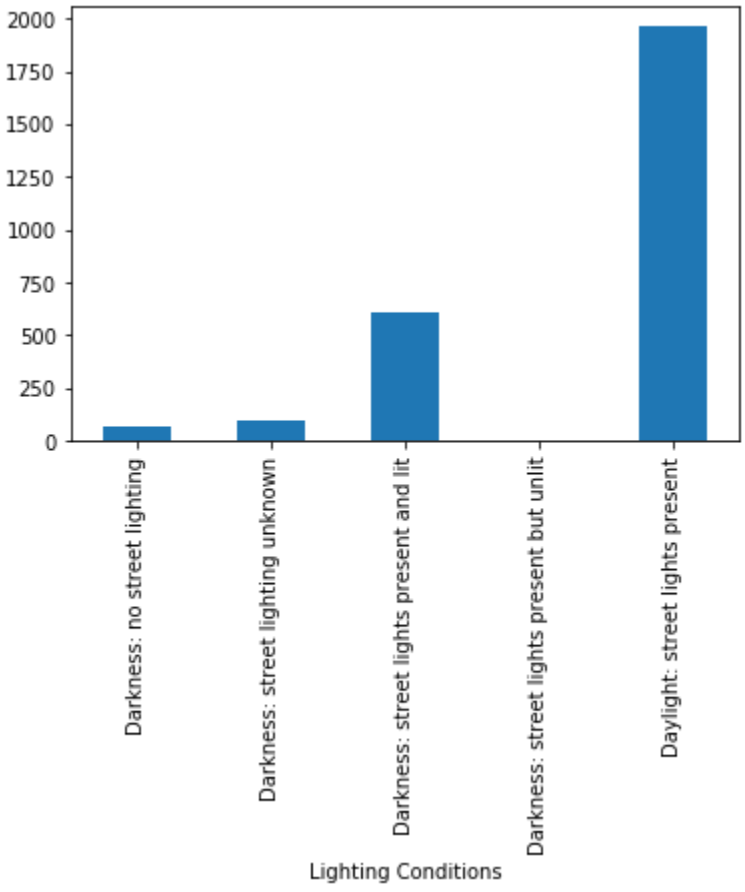

Fig 1.2 Number of accident on different lighting conditions

The figure 1.2 most of the accidents happen in the daylight but when street lights are present it is in the middle day-night evening times .

iii. Weather Condition

Weather acts through visibility impairments, precipitation, high winds, and temperature extremes to affect driver capabilities, vehicle performance (i.e., traction, stability and maneuverability), pavement friction, roadway infrastructure, crash risk, traffic flow, and agency productivity.

Weather conditions can play a part in many accidents, even those concerning other drivers who may have caused your accident and resultant injuries. Sometimes, it's not until you're involved in an accident that you realize just how dangerous it can be to drive in adverse weather conditions.

Black Ice : Black ice is caused by rain or snow melting and freezing on a road or footpath. It can also form due to a road not being designed or maintained well enough to drain water effectively. It's hard to see and even harder to drive on. As a result, it plays a leading role in many weather-related accidents.

Rain : Rain can affect the ability to drive safely in many ways. It might impact driver visibility, causing you to be caught up in an accident that was out of your control. It can also cause roads to be slippery and even flood when road drainage systems are poorly designed or maintained. If you must travel in the rain, pay attention to the weather forecast of the areas you're traveling through and drive to the conditions.

Wind : Driving in high winds can be challenging for many people. The windier it is, the more hazards there may be on the road, and the harder your vehicle may be to control. High-sided vehicles like trucks and vans may be more at risk than others, alongside motorcycles and inexperienced operators. Wind can also cause debris to fly onto roads, directly into the path of fast-moving traffic. This can also lead to expensive and painful accidents.

Snow : While many people have mastered driving in the snow in the area they live in, it doesn't mean accidents don't happen. Snow can act unpredictably while also hiding layers of treacherous ice that can lead to spin outs and collisions. City officials may arrange for roads to be regularly maintained for safety, but it's generally advised for people to avoid driving in such weather conditions if they don't need to.

Sun : Many people use sun blindness as an excuse for causing or being involved in a car accident. Sun blindness involves poor visibility due to the sun being directly in your eyes while driving.

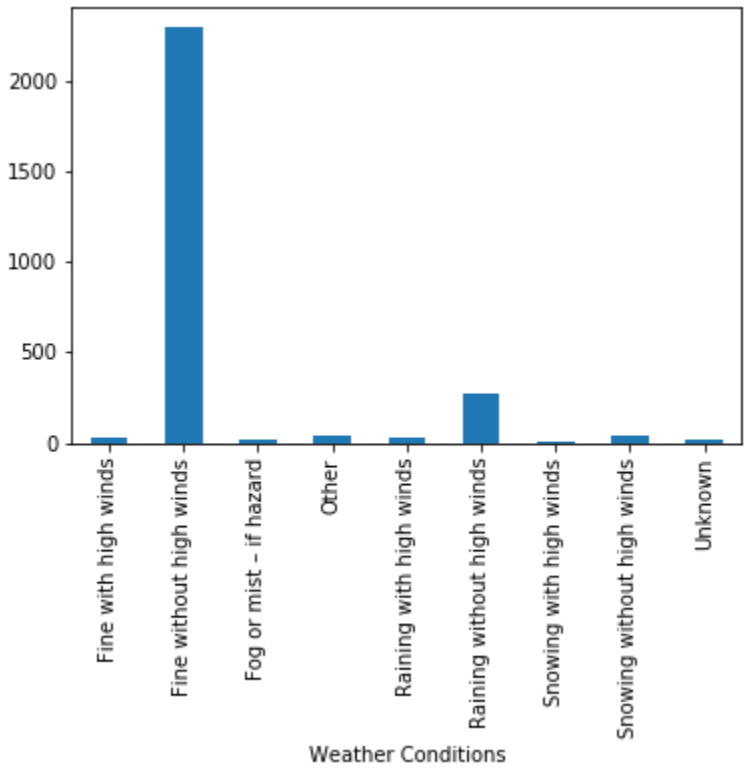

Fig 1.3 Number of accident on different weather conditions

The weather can wreak havoc on the roads, and accidents happen. These weather events above may be some of many contributing factors to an accident you or a loved one were in, and you may wish to discuss them with your chosen personal injury attorney. Weather conditions can play a part in many accidents, even those concerning other drivers who may

have caused your accident and resultant injuries. Sometimes, it's not until you're involved in an accident that you realize just how dangerous it can be to drive in adverse weather conditions.

The general objective was to disaggregate the road accident problem using mass accident data to find groups of road users, vehicles and road segments which would be suitable targets for countermeasures. Successful development of a countermeasure requires a clear understanding of where it can potentially break the chain of events leading to traumatic injury on the road. A countermeasure is a measure which attempts to break the road trauma chain before one of the undesirable steps can occur (eg. accident involvement, injury or death). A target group for a countermeasure is a group of entities (humans, vehicles or roads) for which the chain can be broken effectively and, desirably, cost-effectively.

Mass accident data needs to be analyzed to find target groups for countermeasures in a way which maximizes the chances that the countermeasure will be cost-effective. The project has developed general principles for analysis which meet this aim. These have led to four specific methods of mass data analysis, depending on the nature of the road trauma problem being addressed in the search for countermeasure target groups, namely:

- High Risk Groups (groups with high rates of accident involvement per opportunity to be involved)
- High Severity Groups (groups with high rates of severe injury per accident involvement)
- Accident Involvement Clusters (groups involved in accidents who are homogeneous on a number of factors relevant to countermeasure design and as large as possible)
- Severe Injury Clusters (groups associated with severe injury who are homogeneous on a number of factors relevant to injury countermeasure design and as large as possible).

1.3 IMPACTS OF ROAD ACCIDENT

Road traffic accidents (RTA) are a global problem resulting in deaths, physical injuries, psychological problems and financial losses. These financial damages have immediate consequences and long term consequences on the victims and their families. Road traffic injuries are the leading cause of death globally and the principal cause of death in the age group of 15 to 49 years. Every year the lives of approximately 1.3 million people are cut short globally as a result of a road traffic crash.

i. Social Impacts

The social consequences of road traffic accidents include loss of productivity of the victims, the cost of the legal system, the cost of pain and suffering and loss of quality of life of the victim and their family. The loss of productivity represents a significant proportion of the total social costs.

ii. Economical Impacts

In economic terms, the cost of road crash injuries is estimated at roughly 1% of the gross national product in low‑income countries, 1.5% in middle‑income countries and 2% in high‑income countries.

1.4 OBJECTIVE

There are many factors that contribute to road accidents, especially those related to the environment, vehicles and the travelers. Despite many precautionary measures to reduce road accidents, it remains one of the uncontrollable. Many factors are associated with traffic accidents, some of which are more significant than others in determining the severity of accidents. Identifying the accident severity and its factors can be given a solution for reducing the risk of future road accidents. Successful development of a countermeasure requires a clear understanding of where it can potentially break the chain of events leading to

traumatic injury on the road. The use of big traffic data and artificial intelligence may help develop a promising solution to predict or reduce the risk of road accidents. Machine learning techniques are playing a vital role in predicting the future events based on the historical data. Developing the machine learning classifier for the road accident severity prediction can help to identify the accident severity and developing countermeasures and alerting systems for the future preventions.

To uncover the underlying relationship between a road accident and the contributing factors, ML based models have been studied in recent days. The primary focus of this project is to analyze a set of widely used ML models in terms of their prediction accuracy with the variation of contributing factors. Therefore, this research aims to study different ML algorithms and compare the performance of these algorithms which can be considered to predict road accidents and its severity.

The project's major purpose is to:

- Identify the group of factors lead to the road accident
- Develop a system to identify the road accident severity
- Deploying a model with higher accuracy and lower error rate
- Developing countermeasures for accident injuries
- Developing E-mail based road accident tracing system
- Developing an intelligent AI based application which can notify the user about their zonal accident severity based on their live location environmental factor

CHAPTER 2

LITERATURE SURVEY

1. TRAFFIC ACCIDENT'S SEVERITY PREDICTION: A DEEP-LEARNING APPROACH-BASED CNN NETWORK

In a traffic accident, an accurate and timely severity prediction method is necessary for the successful deployment of an intelligent transportation system to provide corresponding levels of medical aid and transportation in a timely manner. Existing traffic accident's severity prediction methods mainly use shallow severity prediction models and statistical models. To promote the prediction accuracy, a novel Traffic accident's severity Prediction-Convolutional Neural Network (TASP-CNN) model for traffic accident's severity prediction is proposed that considers combination relationships among traffic accident's features. Based on the weights of traffic accident's features, the Feature Matrix to Gray Image (FM2GI) algorithm is proposed to convert a single feature relationship of traffic accident's data into gray images containing combination relationships in parallel as the input variables for the model. Moreover, experiments demonstrated that the proposed model for traffic accident's severity prediction had better performance. Top features affecting accidental severity include distance, temperature, wind_Chill, humidity, visibility, and wind direction. This study presents an ensemble of machine learning and deep learning models by combining Random Forest and Convolutional Neural Network called RFCNN for the prediction of road accident severity. The performance of the model is compared with several base learner classifiers.

2. A DATA MINING FRAMEWORK TO ANALYZE ROAD ACCIDENT DATA

One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, the heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data. In this paper, we proposed a framework that used K-modes

clustering technique as a preliminary task for segmentation of 11,574 road accidents on the road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining is used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generating association rules. Further a trend analysis has also been performed for each cluster and EDS accidents which finds different trends in different clusters whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.. Association rule mining is used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generating association rules.

3. A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS TO PREDICT ROAD ACCIDENT SEVERITY

In this work the author developed a machine learning model for identifying the severity of the accident by considering the human factors such as alcohol, drug, age, and gender are often ignored when determining accident severity. In this work single and ensemble mode machine learning (ML) methods compared their performance in terms of prediction accuracy, precision, recall, F1 score, area under the receiver operator characteristic (AUROC). Road accidents are a global issue that cause deaths and injuries besides other direct and indirect losses. Countries and international organizations have designed technologies, systems, and policies to prevent accidents. The use of big traffic data and artificial intelligence may help

develop a promising solution to predict or reduce the risk of road accidents. Most existing studies examine the impact of road geometry, environment, and weather parameters on road accidents. However, human factors such as alcohol, drug, age, and gender are often ignored when determining accident severity. In this work, we considered various contributing factors and their impact on the prediction of the severity of accidents. For this, we studied a set of single and ensemble mode machine learning (ML) methods and compared their performance in terms of prediction accuracy, precision, recall, F1 score, area under the receiver operator characteristic (AUROC).

4. ACCIDENT DATA ANALYSIS TO DEVELOP TARGET GROUPS FOR COUNTERMEASURES

This work analyzes the accident data to develop target groups for developing the accident countermeasures. The High risk group, high accident severity groups and accident involved clusters are identified and used to develop the accident countermeasures. The author identified each category such as vehicle, driver, passenger, environmental factors and the leading accident factors among them. All the high risk factors are clustered to create a countermeasure for each category.

5. ROAD TRAFFIC ACCIDENTS INJURY DATA ANALYTICS

Road safety researchers working on road accident data have witnessed success in road traffic accidents analysis through the application of data analytic techniques, though little progress was made into the prediction of road injury. This paper applies advanced data analytics methods to predict injury severity levels and evaluates their performance. The study uses predictive modeling techniques to identify risk and key factors that contribute to accident severity. The study uses publicly available data from the UK department of transport that covers the period from 2005 to 2019. The paper presents an approach which is general enough so that it can be applied to different data sets from other countries. The results identified that tree based techniques such as XGBoost outperform regression based ones, such

as ANN. In addition to the paper, it identifies interesting relationships and acknowledges issues related to quality of data.

6. A REVIEW OF DATA ANALYTIC APPLICATIONS IN ROAD TRAFFIC SAFETY. PART 1: DESCRIPTIVE AND PREDICTIVE MODELING

This work presents a comprehensive review on data analytic methods in road safety. Analytics models can be grouped into two categories: predictive or explanatory models that attempt to understand and quantify crash risk and (b) optimization techniques that focus on minimizing crash risk through route/path selection and rest-break scheduling. Their work presented publicly available data sources and descriptive analytic techniques (data summarization, visualization, and dimension reduction) that can be used to achieve safer-routing and provide code to facilitate data collection/exploration by practitioners/researchers. reduce the start-up burden of data collection and descriptive analytics for statistical modeling and route optimization of risk associated with motor vehicles. From a data-driven bibliometric analysis, we show that the literature is divided into two disparate research streams: (a) predictive or explanatory models that attempt to understand and quantify crash risk based on different driving conditions, and (b) optimization techniques that focus on minimizing crash risk through route/path-selection and rest-break scheduling.

Translation of research outcomes between these two streams is limited. To overcome this issue, we present publicly available high-quality data sources (different study designs, outcome variables, and predictor variables) and descriptive analytic techniques (data summarization, visualization, and dimension reduction) that can be used to achieve safer-routing and provide code to facilitate data collection/exploration by practitioners/researchers. Then, the author done the statistical and machine learning models used for crash risk modeling.

7. TRAFFIC ACCIDENT PREDICTION MODEL USING SUPPORT VECTOR MACHINES WITH GAUSSIAN KERNEL

Road traffic accident prediction models play a critical role in the improvement of traffic safety planning. The focus of this study is to extract key factors from the collected data sets which are responsible for the majority of accidents. In this paper urban traffic accident analysis has been done using support vector machines (SVM) with Gaussian kernel. Multilayer perceptron (MLP) and SVM models were trained, tested, and compared using collected data. The results of the study reveal that the proposed model has significantly higher prediction accuracy as compared with traditional MLP approach. There is a good relationship between the simulated and the experimental values. Simulations were carried out using LIBSVM (library for support vector machines) integrated with octave. In this work, applied support vector machines with different Gaussian kernel functions for crash to extract important features related to accident occurrence. The paper compared neural networks with support vector machines. The paper reported that SVMs are superior in accuracy. However, the SVMs method has the same disadvantages of ANN in traffic accident severity prediction.

8. PREDICTING FREEWAY TRAFFIC CRASH SEVERITY USING XGBOOST-BAYESIAN NETWORK MODEL WITH CONSIDERATION OF FEATURES INTERACTION

This project proposes the XGBoost based framework which analyzes the relationship between collision, time and environmental and spatial factors and fatality rate. Results show that the proposed method has the best modeling performance compared with other machine learning algorithms. The paper identified eight factors that have an impact on traffic fatality. In the field of freeway traffic safety research, there is an increasing focus in studies on how to reduce the frequency and severity of traffic crashes. Although many studies divide factors into “human-vehicle-road-environment” and other dimensions to construct models which

show the characteristic patterns of each factor's influence on crash severity, there is still a lack of research on the interaction effect of road and environment characteristics on the severity of a freeway traffic crash. This research aims to explore the influence of road and environmental factors on the severity of a freeway traffic crash and establish a prediction model towards freeway traffic crash severity.

Firstly, the obtained historical traffic crash data variables were screened, and 11 influencing factors were summarized from the perspective of road and environment, and the related variables were discretized. Furthermore, the XGBoost (eXtreme Gradient Boosting) model was established, and the SHAP (SHapley Additive exPlanation) value was introduced to interpret the XGBoost model; the importance ranking of the influence degree of each feature towards the target variables and the visualization of the global influence of each feature towards the target variables were both obtained.

Then, the Bayesian network-based freeway traffic crash severity prediction model was constructed via the selected variables and their values, and the learning and prediction accuracy of the model were verified. Finally, based on the data of the case study, the prediction model was applied to predict the crash severity considering the interaction effect of various factors in road and environment dimensions.

The results show that the characteristic variables of road side protection facility type (RSP), road section type (LAN), central isolation facility (CIF), lighting condition (LIG), and crash occurrence time (TIM) have significant effects on the traffic crash prediction model; the prediction performance of the model considering the interaction of road and environment is better than that of the model considering the influence of single condition; the prediction accuracy of XGBoost-Bayesian Network Model proposed in this research can reach 89.05%. The identification and prediction of traffic crash risk is a prerequisite for safety improvement, and the model proposed and results obtained in this research can provide a theoretical basis for related departments in freeway safety management.

9. PREDICTING TRAFFIC ACCIDENT SEVERITY USING MACHINE LEARNING TECHNIQUES

Road accidents, harming countries' economies, national assets as well as people's lives, are one of the major problems for countries. Thus, investigating contributing factors to the accidents and developing an accurate accident severity prediction model is critical. Using the traffic accident data collected in Austin, Dallas, and San Antonio city of Texas between 2011 and 2021, the primary contributing factors in crashes are probed and the performance of a deep learning model and five different machine learning techniques, such as Logistic Regression, XGBoost, Random Forest, KNN, and SVM, are investigated. The finding shows that the Logistic Regression algorithm shows the best performance among the others with an accuracy of 88% in classifying accident severity.

10. A STUDY ON ROAD ACCIDENT PREDICTION AND CONTRIBUTING FACTORS USING EXPLAINABLE MACHINE LEARNING MODELS: ANALYSIS AND PERFORMANCE

Road accidents are increasing worldwide and are causing millions of deaths each year. They impose significant financial and economic expenses on society. Existing research has mostly studied road accident prediction as a classification problem, which aims to predict whether a traffic accident may happen in the future or not without exploring the underlying relationships between the complicated factors contributing to road accidents.

A number of researches have been done to date to explore the importance of road accident contributing factors in relation to road accidents and their severity, however, only a few of those research have explored a subset of ensemble ML models and the New Zealand (NZ) road accident dataset. Therefore, in this paper, they have evaluated a set of machine learning (ML) models to predict road accident severity based on the most recent NZ road accident dataset. We have also analyzed the predicted results and applied an explainable ML (XML) technique to evaluate the importance of road accident contributing factors.

To predict road accidents with different injury severity, this work has considered different ensembles of ML models, like Random Forest (RF), Decision Jungle (DJ), Adaptive

Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (L-GBM), and Categorical Boosting (CatBoost). New Zealand road accident data from 2016 through 2020 obtained from the New Zealand Ministry of Transport is used to perform this study. The comparison results show that RF is the best classifier with 81.45% accuracy, 81.68% precision, 81.42% recall, and 81.04% of F1-Score.

Next, they have employed the Shapley value analysis as an XML technique to interpret the RF model performance at global and local levels. While the global level explanation provides the rank of the features' contribution to severity classification, the local one is for exploring the use of features in the model. Furthermore, the Shapley Additive exPlanation (SHAP) dependence plot is used to investigate the relationship and interaction of the features towards the target variable prediction.

Based on the findings, it can be said that the road category and number of vehicles involved in an accident significantly impact injury severity. The identified high-ranked features through SHAP analysis are used to retrain the ML models and measure their performance. The result shows 6%, 5%, and 8%, increase, respectively, in the performances of DJ, AdaBoost, and CatBoost models.

CHAPTER 3

ROAD ACCIDENT SEVERITY ANALYSIS

3.1 METHODOLOGY

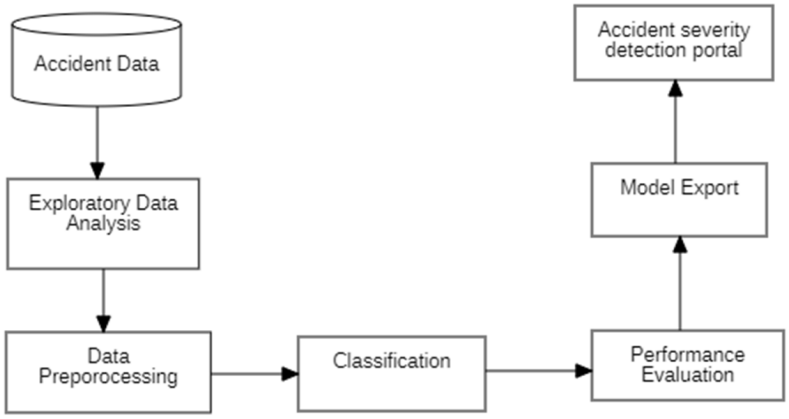

Fig 3.1 : Block diagram of the proposed system

This project classifies the road accident as slight or serious. The flow of the project is shown in figure 3.1. This project consists of five main phases.

1. Conducting descriptive study on the accident data

2. Pre-processing the data using grouping and label encoding

3. Building the machine learning classification model

4. Performance evaluation

5. Developing intelligent web portal for accident severity classification 6

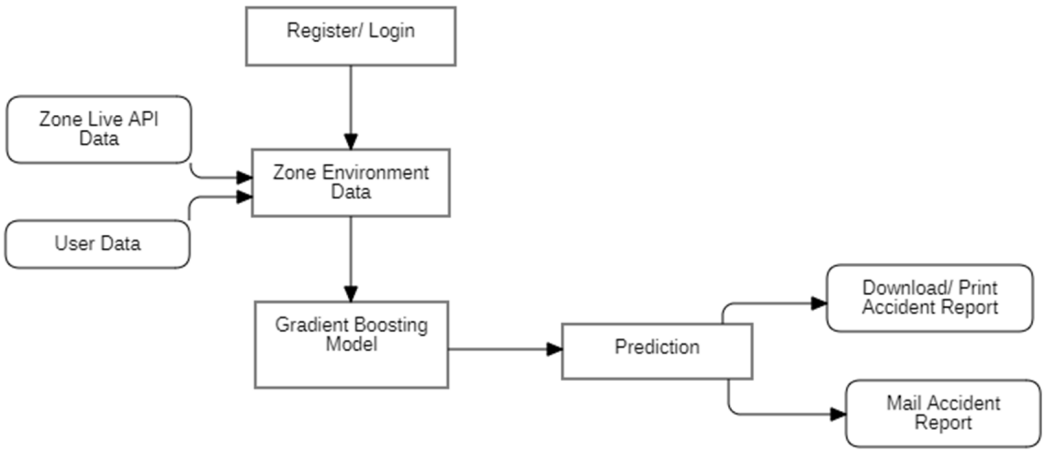

Fig 3.2 : Block diagram of web portal

The steps involved in the intelligence web portal are described in the figure 3.2.

3.1.1 Dataset

This project was implemented using the Accident Severity dataset from the kaggle. The dataset contains 3057 accident samples. Each sample contains 14 predictive attributes and 1 target attribute. The dataset contains the accident data from 01-01-2009 to 31- 12-2009. There are no null values present in the dataset. The target attribute contains two values: slight accident and serious accident. The dataset contains 321 serious data and 2736 slight accident data.

The figure 3.3 describes the predictive attributes that are available in the accident data and its description.

Variables	Description	Data Type	Scale	Null Value
Reference No	Accident identity number	Integer	Serial number	No
Easting	Easting point	Integer	Map point	No
Northing	Northing point	Integer	Map point	No
Number of Vehicles	Number of vehicle in the spot	Integer	Vehicle count	No
Accident Date	Date of accident happened	Date	Date	No
Time (24hr)	Time of accident happened	Time	Time	No
1st Road Class	Road type	Varchar	Category	No
Road Surface	Surface type of the road	Varchar	Category	No
Lighting Conditions	Lighting condition in the spot	Varchar	Category	No
Weather Conditions	Weather condition in the spot	Varchar	Category	No
Casualty Class	Casuality type (Driver..ect)	Varchar	Category	No
Sex of Casualty	Sex of the casuality	Varchar	Category	No
Age of Casualty	Age of the casuality	Integer	Age in years	No
Type of Vehicle	Type of the vehicle	Varchar	Category	No
Severity	Accident severity	Varchar	Category	No

Table 3.1 Accident Dataset Attributes

3.1.2 Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

Three data preprocessing tasks were conducted in the accident data

1. Data reduction

2. Encoding the categorical data

3. Dropping the unwanted columns

Some attribute values can be combined into a single variable. By grouping those values into the categories will help to reduce the number of attributes. The same thing is applicable to

reduce the number of categories in an attribute. Consider the ‘Type of vehicle’ column in the accident data [M/cycle 50cc and under', 'Motorcycle over 125cc and up to 500cc', 'Motorcycle over 50cc and up to 125cc', 'Motorcycle over 500cc] these are all the different kinds of two wheeler motorcycles. Hence, all the data can be grouped into a single category as ‘Motorcycle’.

ii. Data Encoding

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models. In the field of data science, before going for modeling, data preparation is a mandatory task. Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. The figure 3.3 describe how the label encoding performed on the accident data.

Road Surace	Weather Conditions	Casualty Class
0-Dry 1-Frost / Ice- 2-Flood 3-Snow 4-Wet / Damp	0-Fine without high winds 1-Fog or mist – if hazard 2-Fine with high winds 3-Other 4-Raining without high winds 5-Raining with high winds 6-Snowing without high winds 7-Snowing with high winds 8-Unknown	0-Driver 1-Passenger 2-Pedestrian

Table 3.2 Label Encoding

iii. Drop Unwanted Columns

Some of the attributes are usually stored for reference purposes. Those kinds of attributes are not used for the prediction. That data might reduce the accuracy of the classifier. Hence, those columns are removed from the attribute list.

3.1.3 Build machine learning model

Four machine learning classifiers such as Decision tree learning, K-Nearest neighbor, Random forest classifier and the Gradient boosting classifiers are used to develop machine learning models. These models are imported using the Sci-kit learn machine learning library.

3.1.4 Training and Testing

The pre-processed data splitted into two sets. 70% of the data considered as the training data and the rest 30% used to test the model.

3.1.5 Result Analysis

The machine learning models are evaluated using the standard metrics such as accuracy, precision, recall and ROC curve. Based on the metrics of each classifier, the comparative study was conducted over the models. The result shows that the ensembler can give the best accuracy among the traditional single machine learning classifiers.

3.2 GRAPHICAL USER INTERFACE

To deploy the model in the web browser the user interface was designed using HTML, CSS and Javascript. Flask framework used to provide the interface between the Mysql database and the web browser.

3.2.1 Save the machine learning model

To save the model, we simply need to pass the model object to Pickle's dump() function. This will serialize the object and convert it into a "byte stream" that can be stored in the model.pkl file. From the result analysis the Gradient Boosting Classifier provides better accuracy.

3. 3 TOOLS USED

The tools used for the project:

- ·Python
- Jupyter Notebook
- Flask
- HTML
- CSS
- JavaScript
- MySql
- Spyder

3. 3. 1 Python

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard.

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0.[35] Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle-detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions.

i. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

ii. Scikit Learn

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

iii. Collections

Collections in Python are containers that are used to store collections of data, for example, list, dict, set, tuple etc. These are built-in collections. The collections module provides alternatives to built-in container data types such as list, tuple and dict. Several modules have been developed that provide additional data structures to store collections of data

iv. MatplotLib

It is an amazing visualization library in Python for 2D plots of arrays. It is a multiplatform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

v. NumPy

NumPy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

3.3.2 Jupyter Notebook

The Jupyter Notebook is an open-source web application for creating and sharing documents with live code, equations, visualizations, and text. Jupyter Notebook is administered by the Project Jupyter team. The Jupyter Notebooks project is a spin-off of the IPython project, which previously had its own IPython Notebook project. Jupyter derives its name from the three primary programming languages it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which enables you to write your programmes in Python, but you can also use more than 100 other kernels. Jupyter notebooks are intended to provide a more user-friendly interface for code used in digitally-supported research or education.

Jupyter, an open-source environment compatible with a variety of programming languages, has gained traction in a variety of fields.It is useful for documenting code, teaching programming languages, and providing students with a space to easily experiment with provided examples. Jupyter Notebooks can be executed in two major environments: Jupyter Notebook and the more recent JupyterLab. Jupyter Notebook is widely used and well-documented; it provides a simple file browser and an environment for creating, editing, and executing notebooks. Jupyter Lab is more complex and resembles an Integrated Development Environment in its user interface.

The Jupyter Notebook is not only useful for teaching and learning programming languages like Python, but also for sharing data. Google and Microsoft each offer their own

version of the Notebook, which can be used to create and share documents at Google Colaboratory and Microsoft Azure Notebooks, respectively. JupyterLab incorporates the Jupyter Notebook into a browser-based Integrated Development type Editor. JupyterLab can be viewed as an advanced version of Jupyter Notebook. In addition to Note, JupyterLab allows you to run terminals, text editors, and code consoles in your browser.

3.3.3 Flask

Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on the WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine.

i. HTTP Methods

GET - This is used to send the data in and without encryption of the form to the server.

POST - Sends the form data to the server. Data received by POST method is not cached by the server.

ii. Routing

Nowadays, the web frameworks provide routing techniques so that users can remember the URLs. It is useful to access the web page directly without navigating from the Home page. It is done through the following *route()* decorator, to bind the URL to a function.

iii. Handling Static Files

A web application often requires a static file such as javascript or a CSS file to render the display of the web page in the browser. Usually, the web server is configured to set them, but during development, these files are served as static folders in your package or next to the module.

3.3.4 HTML

HTML is the standard markup language for documents intended for display in a web browser. It can be aided by Cascading Style Sheets (CSS) and scripting languages such as JavaScript. Web browsers receive HTML files from a web server or local storage and convert them into multimedia web pages. HTML describes the semantic structure of a web page and originally included hints for the document's appearance.

HTML elements are the fundamental constituents of HTML pages. Images and other objects, such as interactive forms, can be embedded in the rendered page using HTML constructs. HTML enables the creation of structured documents by assigning structural semantics to text elements such as headings, paragraphs, lists, links, and other elements.

3.3.5 CSS

CSS is a style sheet language that is used to describe the presentation of a document written in a markup language such as HTML. In addition to HTML and JavaScript, CSS is a fundamental technology for the World Wide Web. CSS is designed to separate presentation from content, including layout, colors, and fonts. This separation can improve content accessibility; provide greater flexibility and control in the specification of presentation

characteristics; enable multiple web pages to share formatting by specifying the relevant CSS in a separate.css file, which reduces complexity and repetition in the structural content; and enable the.css file to be cached to improve page load speed for pages that share the file and its formatting.

3.3.6 JavaScript

JavaScript, also known as JS, is a programming language that, along with HTML and CSS, is one of the core technologies of the World Wide Web. Over 97 percent of websites use JavaScript for client-side web page behavior, with third-party libraries frequently incorporated. All major web browsers include a dedicated JavaScript engine for executing code on user devices. JavaScript is a high-level, often just-in-time, ECMAScript-compliant, compiled programming language. It features dynamic typing, prototype-based object orientation, and functions with first-class status. Event-driven, functional, and imperative programming styles are supported. APIs for working with text, dates, regular expressions, standard data structures, and the Document Object Model are available (DOM). Initially, JavaScript engines were only used in web browsers, but they are now integral components of many servers and applications.

3.3.7 MySql

MySQL is an open-source relational database management system (RDBMS). Its name is a combination of "My", the name of co-founder Michael Widenius's daughter and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data may be related to each other; these relations help structure the data.

SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a

relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

3.3.8 Spyder

Spyder is an open-source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates with a number of prominent packages in the scientific Python stack, including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, as well as other open-source software.

3.3.9 Weather Stack API

API covers global weather data across the board — from a multi-year history all the way to live information and accurate weather forecasts. The first step to using the API is to authenticate with your weatherstack account's unique API access key, which can be found in your account dashboard after registration. To authenticate with the API, simply use the base URL below and pass your API access key to the API's access_key parameter.

3.3.10 Flask-Session

Flask-Session is an extension for Flask that supports Server-side Session to the application. The Session is the time between the client logs in to the server and logs out of the server. The data that is required to be saved in the Session is stored in a temporary directory on the server. The data in the Session is stored on the top of cookies and signed by the server cryptographically. Each client will have their own session where their own data will be stored in their session.

3. 4 MACHINE LEARNING ALGORITHM

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine

learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc. Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning

I. Supervised Machine Learning

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labeled" dataset, and based on the training, the machine predicts the output. Here, the labeled data specifies that some of the inputs are already mapped to the output. More precisely, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the shape & size of the tail of cat and dog, Shape of eyes, color, height (dogs are taller, cats are smaller), etc. After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, color, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how

the machine identifies the objects in Supervised Learning. The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

II. Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labeled, and the model acts on that data without any supervision. The main aim of the unsupervised learning algorithm is to group or categorize the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

Let's take an example to understand it more precisely; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects. So, now the machine will discover its patterns and differences, such as color difference, shape difference, and predict the output when it is tested with the test dataset.

III. Semi Supervised Machine Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labeled training data) and Unsupervised learning (with no labeled training data) algorithms and uses the combination of labeled and unlabeled data sets during the training period.

Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced. The main aim of semi-supervised learning is to effectively use all the available data, rather than only labeled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labeled data. It is because labeled data is a comparatively more expensive acquisition than unlabeled data.

IV. Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explores its surroundings by hitting & trail, taking action, learning from experiences, and improving its performance. Agents get rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agents is to maximize the rewards.

3.4.1 Decision Tree Learning

The Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Terminologies Of The Decision Trees

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

Leaf / Terminal Node: Nodes that do not split are called Leaf or Terminal nodes.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

The figure 3.5 describes the general structure of the decision tree.

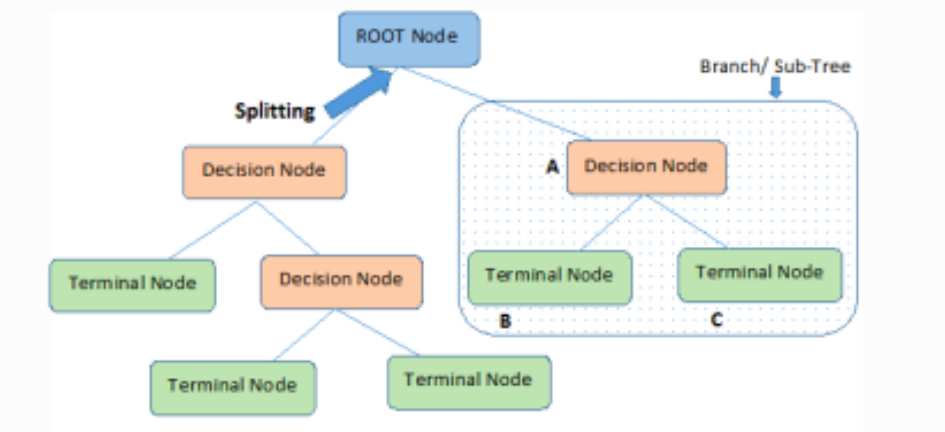

Fig 3.5 : Decision Tree

Steps

- It begins with the original set S as the root node.
- On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute. · It then selects the attribute which has the smallest Entropy or Largest Information gain.
- The set S is then split by the selected attribute to produce a subset of the data. · The algorithm continues to recur on each subset, considering only attributes never selected before.

Entropy

Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. Flipping a coin is an example of an action that provides information that is random.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad \ldots(1)$$

Where, P_i = Probability of randomly selecting an example in class I

Information Gain

Information gain or IG is a statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy.

Information Gain = 1 – Entropy …(2)

3.4.2 K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms that utilizes the Supervised Learning technique.It assumes the similarity between the new case/data and existing cases and places the new case in the category that is the most similar to the existing categories.The K-NN algorithm stores all available data and classifies a new data point on the basis of similarity. This implies that when new data becomes available, the K- NN algorithm can easily classify it into a suitable category.It can be used for both Regression and Classification, although Classification is its primary application.

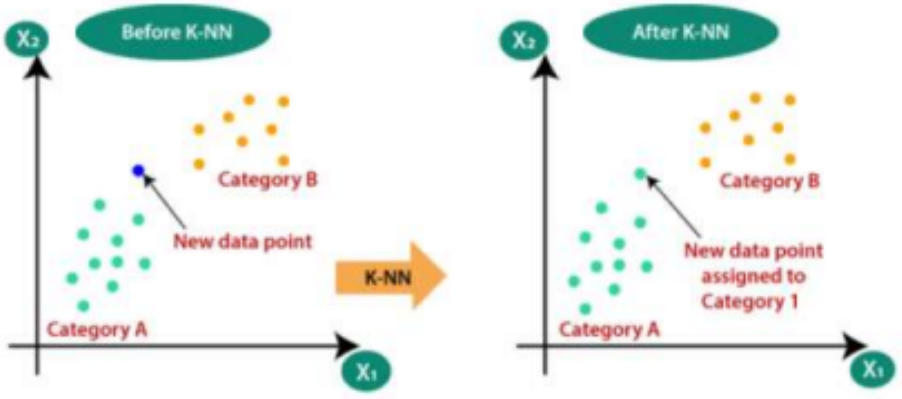

Fig 3.6 : K-nearest Neighbor

K-NN is a non-parametric algorithm, meaning it makes no assumptions about the data it is analysing.It is also referred to as a lazy learner algorithm because it does not immediately learn from the training set. Rather, it stores the dataset and, at the time of classification, performs an action on the dataset. During the training phase, the system simply stores the dataset, and when it receives new data, it classifies it into a category that is highly similar to the original category.

3.4.3 Random Forest Classifier

Popular machine learning algorithm that belongs to the supervised learning technique is Random Forest. It is applicable to both Classification and Regression problems in Machine Learning. It is based on ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the model's performance.

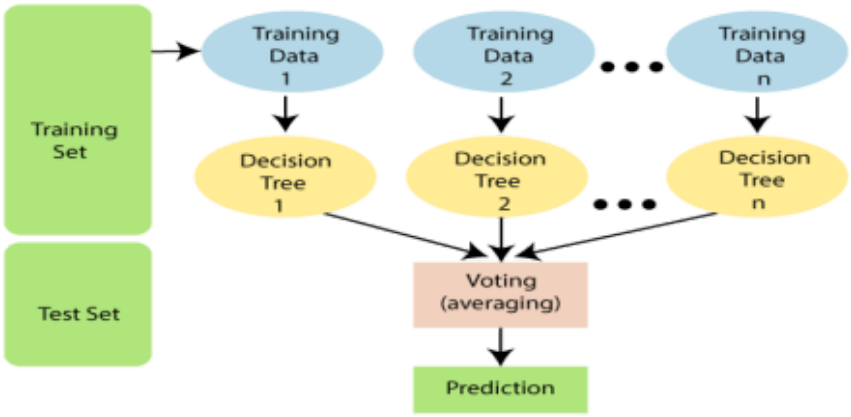

Fig 3.7 : Random Forest Classifier

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier: There should be some actual values in the feature variable of

the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

Steps :

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

Step 1: Select random K data points from the training set.

Step 2: Build the decision trees associated with the selected data points (Subsets). **Step 3**: Choose the number N for decision trees that you want to build.

Step 4: Repeat Step 1 & 2.

Step 5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

3.4.4 Gradient Boosting Classifier

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

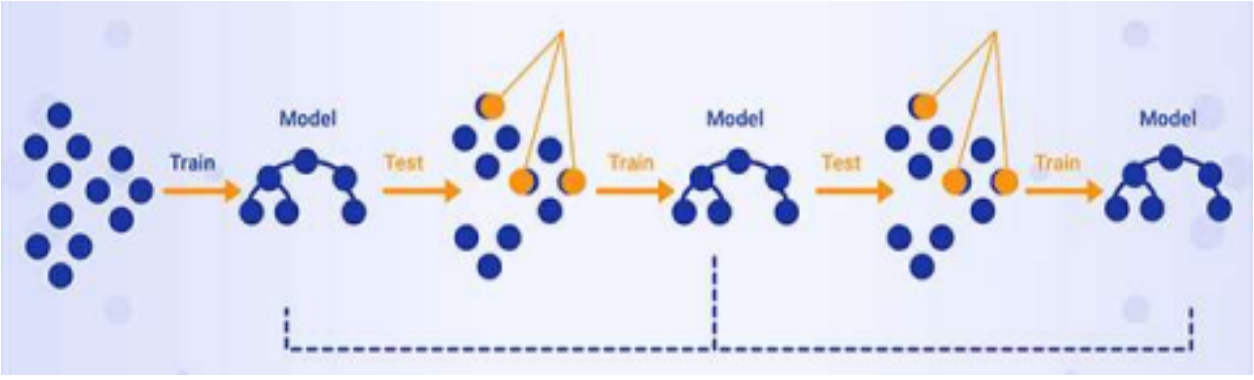

Fig 3.8 : Gradient Boosting Classifier

Gradient boosting algorithms can be used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

Steps:

Step 1: Calculate the average of the target label

Step 2: Calculate the residuals : *residual = actual value – predicted value*

Step 3: Construct a decision tree

Step 4: Predict the target label using all of the trees within the ensemble **Step 5**: Compute the new residuals

Step 6: Repeat steps 3 to 5 until the number of iterations matches the number specified by the hyperparameter (i.e. number of estimators)

Step 7: Once trained, use all of the trees in the ensemble to make a final prediction as to the value of the target variable

CHAPTER 4

RESULTS ANALYSIS

4.1 DATA PREPROCESSING RESULT

Figure 4.1 describes the dataset structure before preprocessing. The dataset downloaded from the kaggle and data preprocessing conducted inorder to transform into the machine readable form.

```
accidentData.head(5)
```

	Reference Number	Easting	Northing	Number of Vehicles	Accident Date	Time (24hr)	1st Road Class	Road Surface	Lighting Conditions	Weather Conditions	Casualty Class	Casualty Severity	Sex of Casualty	Age of Casualty	Type of Vehicle
0	3309	429093	436258	1	01-Jan-09	55	Unclassified	Dry	Darkness: street lights present and lit	Fine without high winds	Pedestrian	Slight	Male	44	Car
1	2609	434723	435534	1	02-Jan-09	2335	Unclassified	Dry	Darkness: street lights present and lit	Fine without high winds	Driver	Serious	Female	23	Car
2	2809	441173	433047	1	02-Jan-09	1645	Unclassified	Dry	Darkness: street lights present and lit	Fine without high winds	Pedestrian	Slight	Female	12	Car
3	3809	428487	431384	1	02-Jan-09	1723	A	Dry	Darkness: street lights present and lit	Fine without high winds	Pedestrian	Slight	Male	15	Car
4	3909	425928	435480	2	02-Jan-09	1350	Unclassified	Dry	Daylight: street lights present	Fine without high winds	Driver	Slight	Female	34	Car

Fig 4.1 : Dataset before preprocessing

Figure 4.2 describes the dataset after performing the pre-processing such as data reduction, label encoding and dropping unwanted attributes. The dataset will be transformed as shown in figure. The data is stored as a csv file and used in the training and testing phases. 30% data used for the testing purpose and the 70% of the data utilized in the training purpose. Each model has been trained and tested using the same data.

The label encoding outperforms while converting the text data into the numerical form. Since some attributes contain multiple values under the same group of attributes and those are grouped based on its similarity.

```
accidentData.head(5)
```

	Number of Vehicles	1st Road Class	Road Surface	Lighting Conditions	Weather Conditions	Casualty Class	Casualty Severity	Sex of Casualty	Age of Casualty	Type of Vehicle
0	1	4	0	2	1	2	0	1	44	6
1	1	4	0	2	1	0	1	0	23	6
2	1	4	0	2	1	2	0	0	12	6
3	1	0	0	2	1	2	0	1	15	6
4	2	4	0	4	1	0	0	0	34	6

Fig 4.2 : Dataset after preprocessing

4.2 PERFORMANCE METRICS

The following metrics are used to evaluate the performance of the classification.

4.2.1 Confusion Matrix

A confusion matrix, also known as an error matrix, is a summarized table used to assess the performance of a classification model. The number of correct and incorrect predictions are summarized with count values and broken down by each class. The confusion matrix is as shown in figure 4.3.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 4.1 Confusion matrix

Positive (P): Observation is positive

Negative (N): Observation is not positive.

True Positive (TP): Outcome where the model correctly predicts the positive class. True Negative (TN): Outcome where the model correctly predicts the negative class.

False Positive (FP): Also called a type 1 error, an outcome where the model incorrectly predicts the positive class when it is actually negative.

False Negative (FN): Also called a type 2 error, an outcome where the model incorrectly predicts the negative class when it is actually positive.

4.2.2 Accuracy

Accuracy is defined as the ratio of correctly predicted examples by the total examples. Accuracy of the classification is calculated using Eq. (3)

$$Accuracy = TP + TN / (TP + TN + FP + FN) \quad ...(3)$$

4.2.3 Error Rate

Error rate is calculated by eliminating the accuracy from the total accuracy

$$\text{Error rate} = 1 - Accuracy \quad ...(4)$$

4.2.4 Precision

Precision is also called Positive predictive value. Precision is also known as positive predictive value and is the proportion of relevant instances among the retrieved instances.The ratio of correct positive predictions to the total predicted positives. Precision of the classification is calculated using Eq. (5)

$$Precision = TP / (TP + FP) \quad ...(5)$$

4.2.5 Recall

Recall also called Sensitivity, Probability of Detection, True Positive Rate. The ratio of correct positive predictions to the total positive examples. Recall of the classification is calculated using Eq. (6)

$$Recall = TP / (TP + FN) \ldots(6)$$

4.2.6 Specificity

Specificity is defined as the proportion of actual negatives, which got predicted as the negative. Specificity of the classification is calculated using Eq. (7)

$$Specificity = TN / (TN + FP) \ldots(7)$$

4.2.7 ROC Curve

A ROC curve (receiver operating characteristic curve) graph shows the performance of a classification model at all classification thresholds. It plots two Parameters namely True Positive (TPR) rate and False positive rate (FPR).

TPR(Eq. 8) and FPR(Eq. 9) of the classification is calculated as follows

$$True\ Positive\ Rate = TP/ (FP + TN) \ldots (8)$$

$$False\ positive\ Rate = FP / (FP + TN) \ldots(9)$$

4.2.8 AUC

AUC stands for Area under the ROC Curve. A perfect classifier would have an AUC of 1. Usually, if your model behaves well, you obtain a good classifier by selecting the value of the threshold that gives TPR close to 1 while keeping FPR near 0.

4.2.9 Performance comparison over the classifiers

The figure 4.4 describes the accuracies that are gained by the considered classifiers. Gaining high accuracy is the important thing to be noticed. The Decision tree learning has got 89% of accuracy. The KNN got 89.4 % of accuracy and worked better than the KNN. Nearly the Random forest algorithm got 90% of accuracy. The best accuracy gained by the Gradient boosting model. Since it is an ensemble classifier, it employs many weak learners in order to learn the severity classification. Hence, its performance is better than the classifications that are considered.

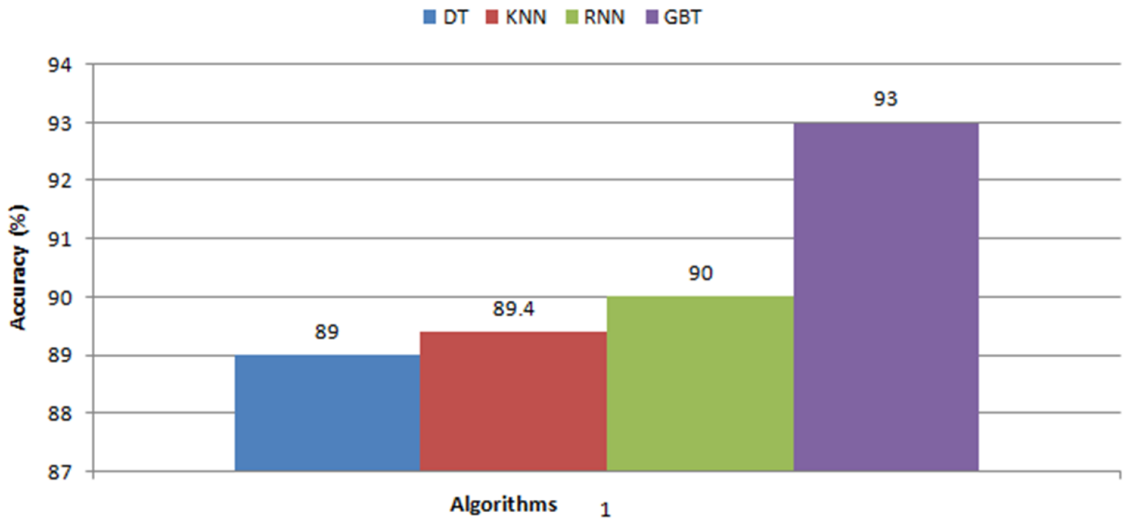

Fig 4.3 Accuracy score

The figure 4.5 describes the recall that is achieved by the considered classification algorithms. Recall score is used to measure the model performance in terms of measuring the count of true positives in a correct manner out of all the actual positive values. Precision-Recall score is a useful measure of success of prediction when the classes are very imbalanced.

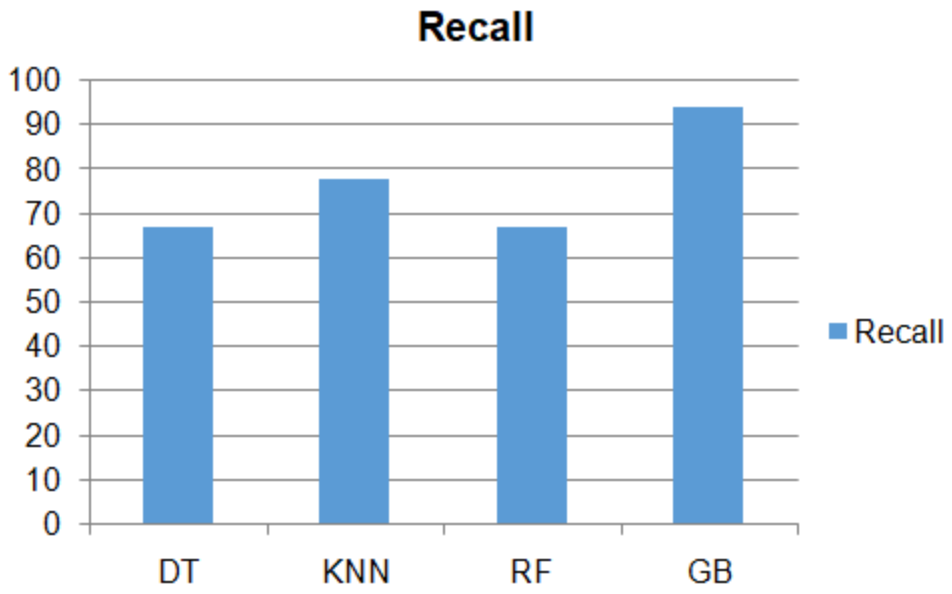

Fig 4.4 Recall score

From the figure 4.4 the Gradient boosting classifier gives the best recall results than other considered classifiers. This mimics the gradient boosting classifier and can give the best recall predictions while considering the road accident data. This shows that the gradient boosting classifiers can work well even if the data is imbalanced. Since it is the binary classification the gradient boosting classifier outperforms while considering the accuracy and recall.

The ROC curve is another performance metric used to evaluate the machine learning model. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

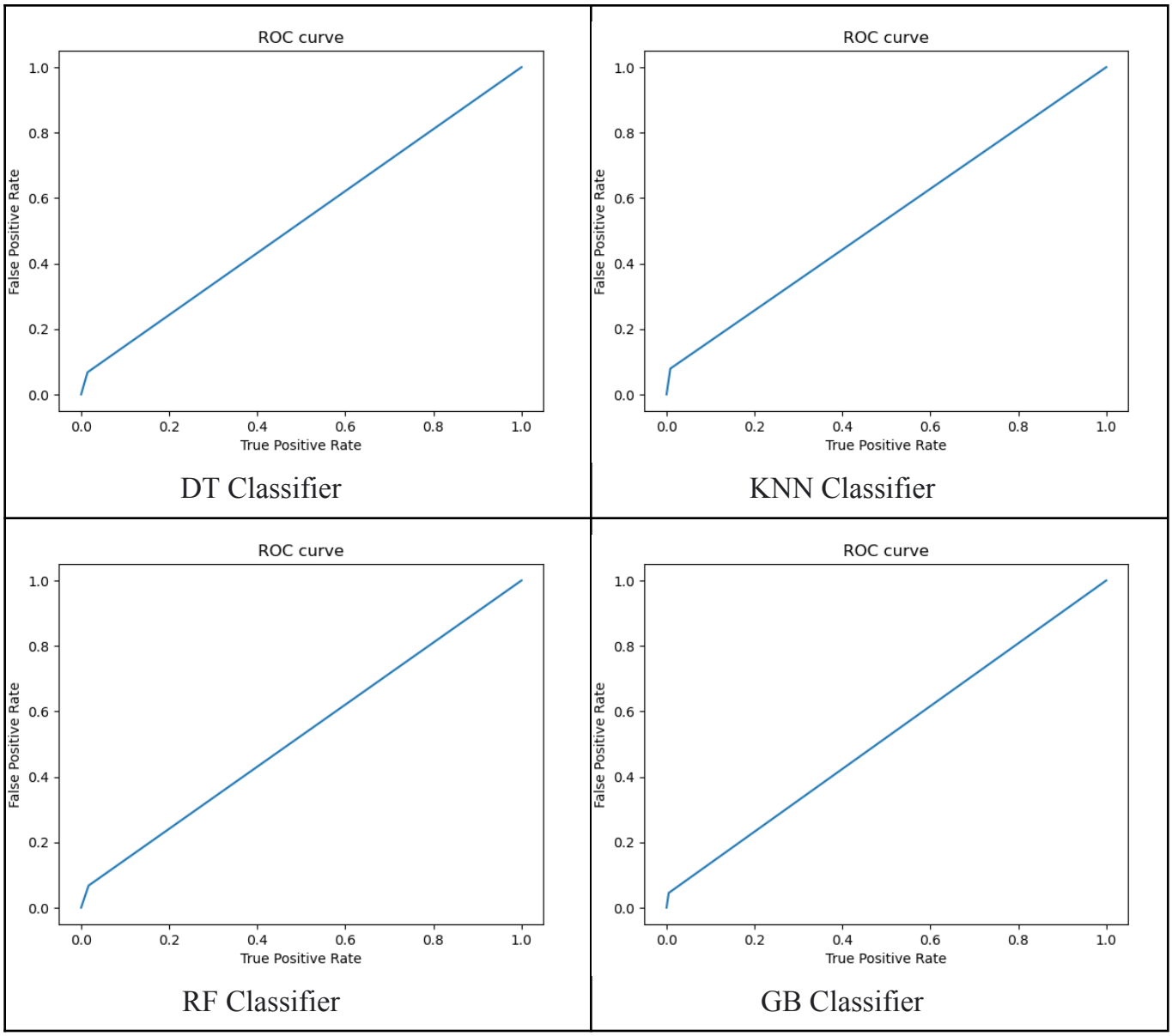

Table 4.2 ROC Curve

The above table 4.2 describes the ROC Curves that are achieved by the considered classifier. The gradient boosting classifier has its curve from less false positive rate.

4.3 GRAPHICAL USER INTERFACE

The main objective of developing a user interface is to deploy the extracted machine learning model in a website. By this implementation any end users without any technical knowledge about machine learning also can use the model to predict the accident severity.

The application consists of three phases.

- Front End
- Programming Interface
- Database Server

i. Front End

The front end of this project is developed using HTML, CSS and JavaScript. The friend end designed with the below functionalities

1. User Registration
2. User Login
3. Accident Severity Detection in the User Live Location
4. Accident Severity Detection in the User Preferred Location
5. Accident Severity Detection between two Locations
6. Downloading the Prediction as Report
7. Send the Report as Mail

ii. Programming Interface

Flask is used as a backend . It is used to provide the interface between the web browser and the MySql database. The user details, accident details and location details are saved and retrieved from the database

iii. Database Server

Xampp server is used to provide a server based mechanism for the MySql Database. The below are some of the images from the intelligent accident severity detection portal.

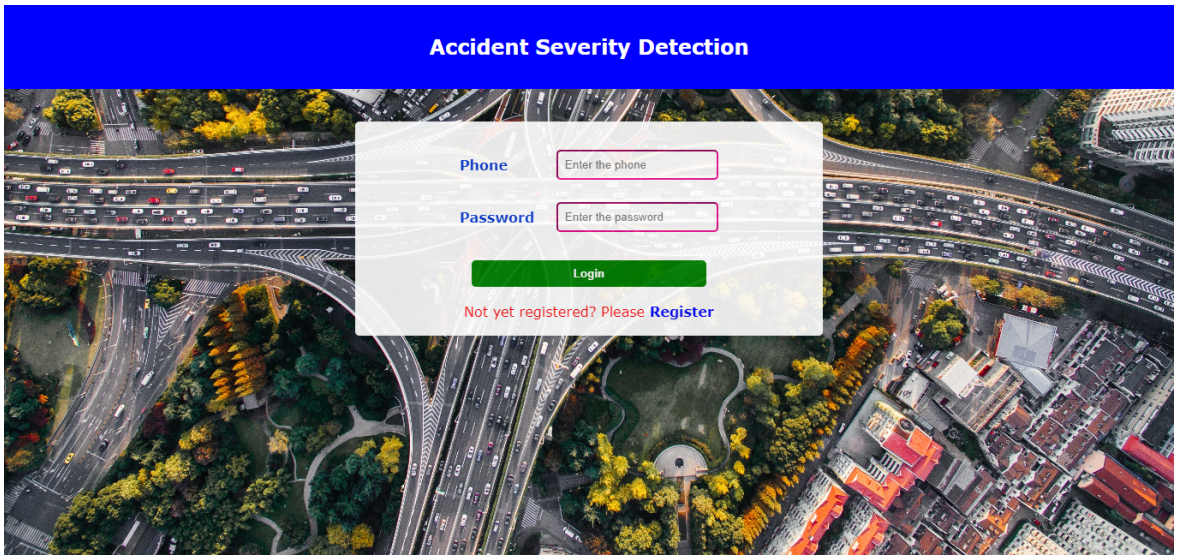

Fig 4.5 User Authentication

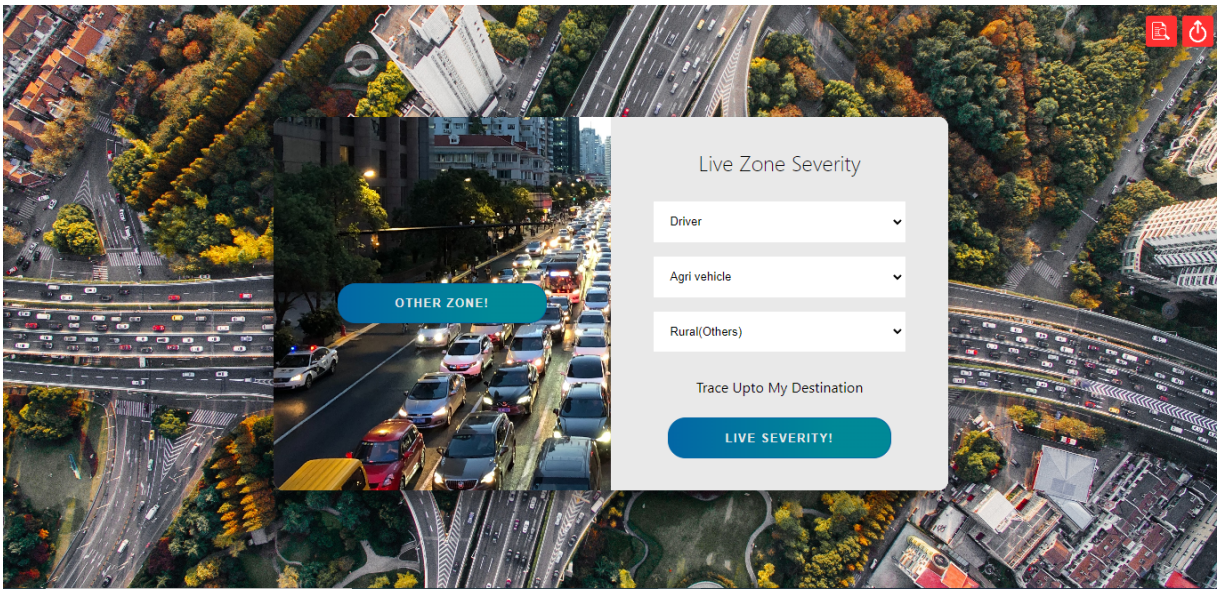

Fig 4.6 Live Accident severity Detection

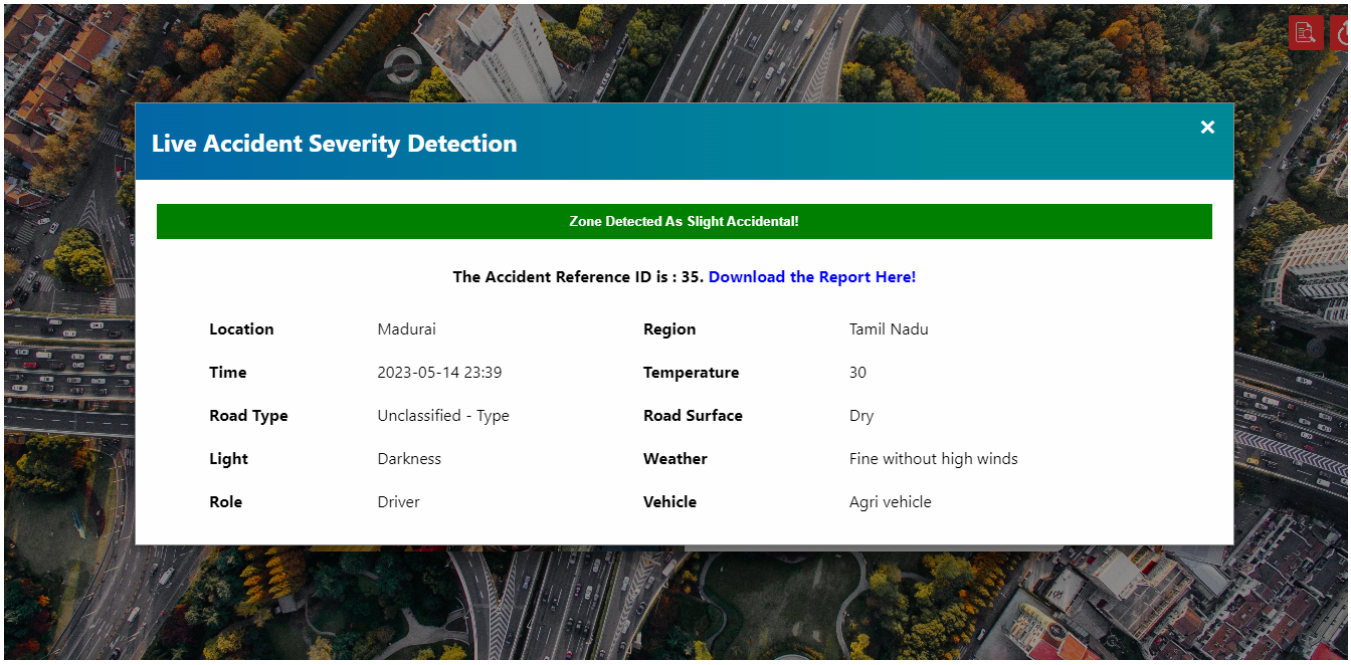

Fig 4.7 Live Accident severity Detection Result

The web application has been tested with various test data that are available and unavailable in the training and test dataset. The system predicts accurately whatever the data provided. Since it is a live data based prediction, it works well in various geographical locations also.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Road accidents are one of the most regrettable hazards in this hectic world. Road accidents lead to numerous casualties, injuries, and fatalities each year, as well as significant economic losses. Predicting the accident severity is one of the major tasks. The Gradient boosting model outperforms while considering the accident data and it achieved 93% of accuracy. Number of Vehicles, Road Class, Road Surface, Lighting Conditions, Spot Weather Conditions, Casualty Class, Sex of Casualty, Age of Casualty, Type of Vehicle are used to predict the accident severity. This is extraordinarily beneficial for the highway authorities, police departments and for journalists. The key applications of this project are Early accident severity prediction, No expert knowledge required, Can access the model anytime and anywhere, Can access the previous predictions, Can send mail immediately to the respective authority.

5.2 FUTURE SCOPE

The methodology can be used for pre prediction also. Whenever the driver, traveler or passenger starts a journey in a particular area they can predict the accidents happening in that area and the severity of the accident. This can be further used to identify the risk factors, countermeasures. The previous predicted data also can be used to predict the future accident severities and to improve the efficiency of the model.

REFERENCES

- Mubariz mansoor, Muhammad umar , Saima sadiq , Abid isaq, Saleem ullah, Hamza, and Carmen, "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model", Digital Object Identifier 10.1109/ACCESS.2021.3112546.

- Sachin Kumar and Durga Toshniwal, "A data mining framework to analyze road accident data", DOI 10.1186/s40537-015-0035-y

- Shakil Ahmed, Md Akbar Hossain, Md Mafijul Islam Bhuiyan, Sayan Kumar Ray, "A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity", 978-1-6654-6667-7/21/$31.00 ©2021 IEEE DOI 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

- Accident Data Analysis to Develop Target Groups For Countermeasures, Max Cameron

- Mohamed K Nour, Atif Naseer, Basem Alkazemi, Muhammad, "Road Traffic Accidents Injury Data Analytics", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020

- A. Mehdizadeh, M. Cai, Q. Hu, M. A. A. Yazdi, N. Mohabbati-Kalejahi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, “A review of data analytic based applications in road traffic safety. Part 1: Descriptive and predictive modeling,” Sensors (Switzerland), vol. 20, no. 4, pp. 1–24, 2020.

- Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, M. A. A. Yazdi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, “A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling,” Sensors (Switzerland), vol. 20, no. 4, pp. 1–19, 2020.

- J. Ma, Y. Ding, J. C. Cheng, Y. Tan, V. J. Gan, and J. Zhang, “Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective,” IEEE Access, vol. 7, pp. 148 059–148 072, 2019.

- N. Zagorodnikh, A. Novikov, and A. Yastrebkov, “Algorithm and software for identifying accident-prone road sections,” Transp. Res. Procedia, vol. 36, pp. 817– 825, 2018. [Online]. Available: https://doi.org/10.1016/j.trpro.2018.12.074.

- L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, “Traffic Accidents Classification and Injury Severity Prediction,” in 2018 3rd IEEE Int. Conf. Intell. Transp. Eng. ICITE 2018, 2018, pp. 52–57