ROAD ACCIDENTS ANALYSIS USING MACHINE LEARNING

A SEMINAR REPORT

Submitted by

V.VISWA SUBRAHMANYAM (RA1911026020076)
K.NAVEEN CHAITANYA (RA191102602090)
M.NAVEEN (RA1911026020103)

Under the guidance of

Ms Deva Hema M.E

(Assistant Professor, Department of Computer Science and Engineering)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

ABSTRACT

Today, traffic safety is one of the main priorities of governments. Considering the importance of topic, identifying the factors of road accidents has become the main aim to reduce the damage caused by traffic accidents. In this paper, we have applied the concepts of machine learning and data mining to identify the various factors that affect road accidents and its severity. The application takes various inputs such as weather conditions, road conditions, time of day etc. and uses machine learning algorithm (Decision tree algorithm) to calculate the severity of a possible accident on a scale of 1 to 4 (1 being the least and 4 being the most severe). This data can be used for analysis of future inputs and improves the accuracy of the system output. This model can further be improved to send the report of the accident to the concerned authorities, such as hospitals, ambulance and insurance agencies and can therefore prove to be very helpful in reducing accident fatality rates in the country

LIST OF CONTENT

3.4 SCOPE OF PROJECT

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 DOMAIN INTRODUCTION

1.1.1 MACHINE LEARNING

Machine learning (**ML**) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine Learning**,** a prominent topic in Artificial Intelligence domain, has been in the spotlight for quite some time now. This area may offer an attractive opportunity, and starting a career in it is not as difficult as it may seem at first glance. Even if you have zero experience in math or programming**,** it is not a problem. The most important element of your success is purely your own interest and motivation to learn all those things.

Machine learning is learning based on experience. As an example, it is like a person who learns to play chess through observation as others play. In this way, computers can be programmed through the provision of information which they are trained, acquiring the ability to identify elements or their characteristics with high probability.

Road Accident Analysis using Machine Learning

First of all, you need to know that there are various stages of machine learning**:**

- data collection
- data sorting
- data analysis
- algorithm development
- checking algorithm generated
- the use of an algorithm to further conclusions

To look for patterns, various algorithms are used, which are divided into two groups**:**

- Unsupervised learning
- Supervised learning

Unsupervised learning: In this, your machine receives only a set of input data. Thereafter, the machine is up to determine the relationship between the entered data and any other hypothetical data. Unlike supervised learning, where the machine is provided with some verification data for learning, independent Unsupervised learning implies that the computer itself will find patterns and relationships between different data sets. Unsupervised learning can be further divided into clustering and association.

Supervised learning: This implies the computer ability to recognize elements based on the provided samples. The computer studies it and develops the ability to recognize new data based on this data. For example, you can train your computer to filter spam messages based on previously received information.

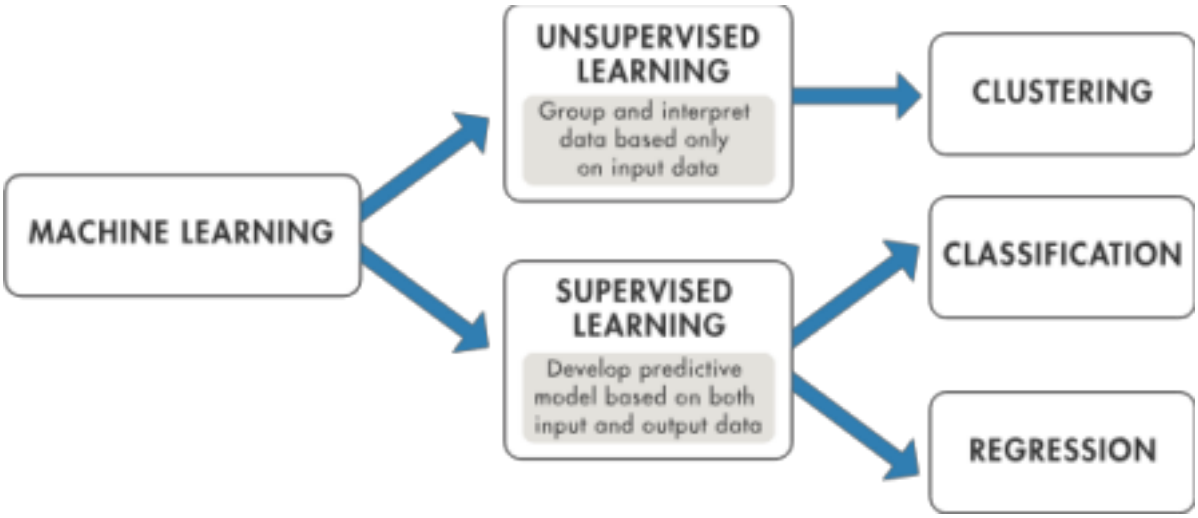

Figure 1.1.1 Machine Learning and its classification

1. **Classification:** When inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.
2. **Regression:** Which is also a supervised problem, A case when the outputs are continuous ratherthan discrete.
3. **Clustering:** When a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

Some Supervised learning algorithms include:

- Decision trees
- Logistic Regression
- Random Forest
- k-nearest neighbors
- linearregression

1.1.2 DATA MINING

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Models are created using accident data records which can help to understand the characteristics of many features like driver's behavior, roadway conditions, light condition, weather conditions and so on. These models can study and used to predict accidents based on weather, time of the day, region etc.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictiveanalytics.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of features, or inputs, in a set of data.

Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget, and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labelling. Reinforcementlearning algorithms are given feedback in the form

of positive or negative reinforcement in a dynamic environment, and are used in autonomous vehicles or in learning to play a game against a human opponent. Other specialized algorithms in machine learning include topic modelling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems. Meta learning algorithms learn their own inductive bias based on previous experience. In developmental robotics, robot learning algorithms generate their own sequences of learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies, and

imitation.

Data mining is a branch which involves looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology. The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc. Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

Many forms of data mining are predictive. For example, a model might predict income based on education and other demographic factors. Predictions have an associated probability (How likely is this prediction to be true?). Prediction probabilities are also known as confidence (How confident can I be of this prediction).

Some forms of predictive data mining generate rules, which are conditions that imply a given outcome. For example, a rule might specify that a person who has a bachelor's degree and lives in a certain neighborhood is likely to have an income greater than the regional average. Rules have an associated support (What percentage of the population satisfies the rule?).

Figure 1.1.2 Data Mining

There is a great deal of overlap between data mining and statistics. In fact, most of the techniques used in data mining can be placed in a statistical framework

However, data mining techniques are not the same as traditional statistical techniques.

Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover,statistical methodstypically donotscale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.

Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.

Data mining involves six common classes of tasks: -

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require furtherinvestigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attemptsto find a function which modelsthe data with the least error that is, for estimating the relationships among data ordatasets.

1.2 PROBLEM STATEMENT

This model aims to make roads more secure and accident-free using machine learning. Here, datasets containing details about previous accidents in various regions is studied and analyzed and a model is developed which can be used to predict and prevent road accidents. It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. the model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it. The database used is a public one, available to many institutes and government websites. The data collected will be analyzed, integrated and grouped together based on different constraints using the bestsuited algorithm.

There are several problems with current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analyzed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be helpful to analyses and identify the flaw and the reasons of the accidents. It will also be helpful while making roads and bridges as a reference to avoid the same problems faced before. The predictions made will be very much useful to plan the management of such problems.

1.3 OBJECTIVES

There is a huge impact on the society due to traffic accidents where there is a great cost of fatalities and injuries. In recent years, there is an increase in the researches attention to determine the significantly affect the severity of the driver's injuries which is caused due to the road accidents. Accurate and comprehensive accident records are the basis of accident analysis. the effective use of accident records depends on some factors, like the accuracy of the data, record retention, and data analysis. There are many approaches applied to this scenario to study this problem.

There are several problems with the current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analyzed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be helpful to analyze and identify the flaw and the reasons of the accidents. It will also be helpful while making roads and bridges as a reference to avoid the same problems faced before. The predictions made will be very much useful to plan the management of such problems.

Design and control of traffic by advanced systems come in view as the important need. Assumption on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous.

For this, models are created using accident data records which can help to understand the characteristics of many features like driver's behavior, roadway conditions, light condition, weather conditions and so on. It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. the model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to performa riskfactor and reduce it.

The main objective of the road accident prediction system:

- Analyze the previously occurred accidents in the locality which will help us to determine the most accident-prone area and help us to set up the immediate required help forthem.

- To make predictions based on constraints like weather, pollution, road structure, etc.

1.4 SCOPE OF THE PROJECT

The features of this project include:

- A demonstration of how open datasets can be combined to obtain meaningful features for road accident prediction.
- A high spatial andtemporalresolution roadaccidentpredictionmodelforthe island of Montreal,
- A comparison ofthree algorithmsdealingwith data imbalance in the context of road accident prediction.
- An implementation of Balanced Random Forest in Apache Spark for efficient distributed training.

The primary objective of this work is to evaluate the application of a J48 decision tree classifier, random forest (RF), and instance-based learning with parameter k (IBk) model for predicting and classifying motorcycle crash severity. The performances of the models were assessed and compared to that of a multinomial logit model (MNLM). This study also identified and examined factors that are potentially significant to injury severity in motorcycle crashes. Identifying factors that significantly affect crash severity is one of the most critical tasks in traffic safety. Based on this, policy can be formulated to mitigate the number of fatalities and injuries resulting from crashes. The contributions of this research to motorcycle safety are threefold: firstly, to fill in the gap in the lack of application of machine learning in motorcycle crash severity analysis. It is an innovative study because the extensive review of existing literature revealed that this is the first time the J48.

Decision Tree Classifier, RF and IBk models are employed to predict motorcycle crash severity. Secondly, investigating contributing factors associated with motorcycle crash severity in Ghana is under-researched; this study, therefore, contributes to the literature on motorcycle safety by fill in this gap.

CHAPTER 2

Literature survey

s.no	title	Author name Jounal name year	objective	Name of the techniques used and dataset	Limitations
1	Using Machine Learning to Predict Car Accident Risk	Daniel Wilson, medium.com, 2018 Link: https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d5	Making use of machine learning to save lives and prevent accident	Dataset: Static features Weather Time features Speed limit Road curvature Average traffic volumes etc. Techniques used: ArcGIS platform (location-focused platform-as-a-service - (PaaS)	As this theory is published on considering Static features, real time traffic information in which we can plan better routes is not recorded
2	Analyzing road accident data using machine learning paradigms	1) Priyanka A. Nandurge, 2) Nagaraj V. Dharawadkar IEEE, 2017	To determine the main factors associated with road traffic accidents and to prevent them accordingly.	Dataset: Time Number injured Light condition Weather condition Type of area Road type, etc. Techniques used: K-means clustering and association	It requires to specify the number of clusters (k) in advance and it can not handle noisy data and outliers.

3	Development of traffic accident prediction models by traffic and road characteristics in urban areas.	Dahee Hong, citeseerx.ist.psu.edu, 2012link:https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.563.2040&rep=rep1&type=pdf	To develop accident prediction model that reduces the risk of accidents by physical characteristics of road types through a survey of characteristics of roads and accidents in urban areas.	Dataset:Road typeNumber of accidents/kmNumber of IntersectionsNumber of Connection roads Traffic VolumeNumber of injured/kmNumber of death/kmTechniques used:Calculation mechanisms using the above datasets.	This survey does not use any modern use cases like geo location, weather, speed of vehicle etc.,. that are main reasons for the cause of accidents.

s.no	title	Author name Jounal name year	objective	Name of the techniques used and dataset
4	Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning	Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine. 7th International Conference on Smart Computing & Communications (ICSCC),2019	This research paper has been done to analyze traffic accidents more deeply to determine the intensity of accidents by using machine learning	DATA SET 1.Accident Severity 2.Weight 3.Vehicle type 4.Road Geometry 5.Divider 6.Location type Techniques used Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes and AdaBoost
5	A Framework for Analysis of Road Accidents using machine learning	Shristi Sonal Saumya Suman. Proceedings International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)2018.	The road accident data analysis use data mining and machine learning techniques, focusing on identifying factors that affect the severity of an accident.	Dataset 1.No. of victims 2.Date of accident 3.Time of accident 4.Road type 5.Weather 6.condition 7.Lighting on road 8.Accident severity 9.Age group Techniques used a powerful N-dimensional array object,R programming language

title	Author name Jounal name year	objective	Name of the techniques used and dataset	Limitation
Machine Learning Approach to Prediction of Passenger Injuries on Real Road Situation	Yongbeom Lee, Eungi Cho, Mingyu Park. National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning)2018	In the paper, we propose a model for prediction of the severity of traffic accident injuries applicable to the automatic notification system of traffic accident	Dataset: 1.The details of a crash 2.The components of a harmful event 3.vehicle information 4.General occupant information 5.Occupant injury information Techniques usedNASS-CDS, machine learning, decision tree, prediction of injury grad	It requires to specify the number of clusters in advance andIt can not handle noisy data and outliers

s.no	title	Author name Jounal name year	objective	Name of the techniques used and dataset	Limitation
7	Prediction of Road Accident and Severity of Applying Machine Learning Techniques	IEEE 8th R10Humanitarian TechnologyConference (R10-HTC)	Prior work on this issue mainly focused on either the possibility of an accident or the degree of severity. Some works showed low accuracy due to deficiency in record which is a major problem	DATA SET 1.Accident Severity 2.Weight 3.Vehicle type 4.Road Geometry 5.Divider 6.Location type Techniques used Decision Tree, K-Nearest Neighbors (KNN), Naïve	its hard to find the proper areas and correct point in the areas due to certain road conditions and lack of information of small locations and narrow road
8	Road Accident Analysis and Hotspot Prediction using Clustering	Jayesh Patil Vaibhav Patil Dhaval Walavalkar Vivian Brian Lobo 2021 6th International Conference on Communication and Electronics Systems (ICCES) 2009	This tech era,everything isalmostbecomingpossible. Machinelearning is used to analyze variousalgorithmsthroughexpert analysis	DATA SET 1.Accident-prone area 2. Condition of vehicle 3.Traffic conditions Techniques used k-means clustering algorithm; Machine learning; potholes; prediction model; unsupervised learning	As this theory is published on considering Static features, real time traffic information in which we can plan better routes is not recorded

LITERATURE SURVEY :

The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additionalresources.

- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the mostrelevant

  characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal. Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly.

  machine learning with AI and cognitive technologies can make it even more

effective in processing large volumes of information. Data mining is a branch which involves looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology. The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc. Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc. Many forms of data mining are predictive. For example, a model might predict income based on education and other demographic factors. Predictions have an associated probability (How likely is this prediction to be true?). Prediction probabilities are also known as confidence (How confident can I be of this prediction?). Some forms of predictive data mining generate rules, which are conditions that imply a given outcome. For example, a rule might specify that a person who has a bachelor's degree and lives in a certain neighborhood is likely to have an income greater than the regional average. Rules have an associated support (What percentage of the population satisfies the rule?). There is a great deal of overlap between data mining and statistics. In fact, most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques.

Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover,statistical methodstypically do notscale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.

Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.

Data mining involves six common classes of tasks: -

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require furtherinvestigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attemptsto find a function which modelsthe data with the least error that is, for estimating the relationships among data or datasets. • Summarization – providing a more compact representation of the data set, including visualization and report generation.

2.2 EXISTING SYSTEM

1. Sachin Kumar: used data mining techniques to identify the locations where high frequency accidents are occurred and then analyses them to identify the factors that have an effect on road accidents at that locations. The first task is to divide the accident location into k groups using the k-means clustering algorithm based on road accident frequency counts. Then, association rule mining algorithm applied in order to find out the relationship between distinct attributes which are in accident data set and according to that know the characteristics of location

2. S.Shanthi: proposed data mining classification technology based on gender classification, in which RndTree and C4.S use AdaBoost Meta classifier to provide high-precision results. From the Critical Analysis Reporting Environment (CARE) system provided by the Fatal Analysis Reporting System (FARS) used by the training data set.

3. Tessa K. Anderson: proposed a method of identifying high-density accident hotspots, which creates a clustering technique that determines that stochastic indices are more likely to exist in some clusters, and can therefore be compared in time and space. The kernel density estimation tool enables the visualization and manipulation of density-based events as a whole, which in turn is used to create the basic spatial unit of the hotspot clusteringmethod.

   The severity of damage occurring during a traffic accident is replicated using the performance of various machine learning paradigms, such as neural networks trained using hybrid learning methods, support vector machines, decision trees, and concurrent mixed models involving decision trees and neural networks. The experimental results show that the hybrid decision tree neural network method is better than the single method in machine learning paradigms.

4. Lukuman Wahab, Haobin Jiang:

   • Traffic safety is a global issue that is progressing at an alarming rate. It severely affects developing, as well as developed countries.

   • The current studies on road traffic crashes have focused primarily on the analysis of fatal crashes, and those involving pedestrians.

- However, where attempts have been made to study motorcycle crashes specifically the focus has typically been on helmet usage and commercial motorcycle operations without consideration of the factors that influence crash severity.
- In this research machine learning based algorithms is proposed to predict motorcycle crash severity.
- Traffic safety is a global issue that is progressing at an alarming rate. It severely affects developing, as well as developed countries.
- The current studies on road traffic crashes have focused primarily on the analysis of fatal crashes, and those involving pedestrians.
- However, where attempts have been made to study motorcycle crashes specifically the focus has typically been on helmet usage and commercial motorcycle operations without consideration of the factors that influence crash severity.
- In this research machine learning based algorithms is proposed to predict motorcycle crash severity.

5. IJIRCCE-Samila Kumar, M Prashanthi:

- In this paper, data mining technique is used.
- Data mining allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationshipsidentified.
- Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.
- After selecting dataset, pre-processing is be done which consists of attribute selection, data cleaning and data transformation.
- Data gathered from different sources was consolidated, mapped and scrutinized. Some of the data that is not pertinent to the data mining exercise was ignored.
- For dataset association rule is applied which defines correlation among a set of items, the relationship is defined in terms of the frequency of co occurrence or appearance of the items together in each transaction

- Here random forest method is used to select the attributes that are mostly affecting the road accidents, apriority algorithm is applied on selected attributes to find the most important rules by discarding the irrelevant rules.

6. Vipul Rana, Hemant Joshi, Deepak Parmar:

- One of the most complicated and difficult daily needs is overland transportation. In India, more than 150,000 people are killed each year in traffic accidents.

- Data Mining is a field of study within machine learning and focuses on exploratory data analysis through unsupervised learning. It is a branch which involves looking for valid and potentially useful patterns in huge datasets. So, it is the process of analyzing data from different perspectives and summarizing it into useful information.
- Machine learning is the process of training the machine to perform certain task without the explicit knowledge of programming. So, given a dataset we need to train the model by performing certain analysis using the algorithm.
- Data setselection.
- Data set cleaning and Data Transformation.
- Data Processing and Algorithm.
- Output and user side experience.

2.3 PROPOSED SYSTEM

Models are created using accident data records which can help to understand the characteristics of many features like latitude, longitude, roadway conditions, light condition, weather conditions, road side and so on. Machine learning will help to train the model. Model is trained by using decision tree after training helps to predict the severity.

It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. The model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it.

Here the road accident study is done by analyzing some data by giving some queries which is relevant to the study. The queries like what is the most dangerous time to drive,

2.4 ADVANTAGES AND DISADVANTAGES

ADVANTAGE:

- Help to know the severity according to the giving condition
- User friendly GUI
- Results can be used to avoid accident in a particular area
- Helps how roads should be constructed to avoid accident
- Model is trained well so it gives accurate severity.

DISADVANTAGE:

- If accident details are not recorded then it difficult to provide which area has chance of more accident
- Algorithm takes numerical values.
- Few rows where deleted due to null values which can be replaced by mean or mode value.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 FUNCTIONAL REQUIREMENTS

In software engineering, a functional requirement defines a function of a software system or its component. A function is described as a set of inputs, the behavior, and outputs. Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish. Behavioral requirements describing all the cases where the system uses the functional requirements are captured in use cases.

In this project, we take data set from the government website. This data will help to predict why accident happened, at what time maximum number of accidents will happened etc. The model which we will develop will help to predict the reasons why accident is happening. This will help which construction of road. According to result new rule can me introduced for the safety of human being.

Here, the system has to perform the following tasks:

- Take user id and password and allow users to login
- Logged in users can input the necessary conditions in the appropriate fields to calculate accidentseverity
- On clicking on “Submit” button, the values are passed onto to the python code and severity is predicted using the appropriate algorithm and displayed.

3.2 HARDWARE REQUIREMENTS

- Processor : Any Processor above 500 MHz
- RAM : 512Mb
- Hard Disk : 10 GB
- Input device :Standard Keyboard and Mouse

3.3 SOFTWARE REQUIREMENTS

3.3.1 SOFTWARE

Anaconda:

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing like data science, machine learning applications, largescale data processing, predictive analytics, etc. that aims to simplify package management and deployment.

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage anaconda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install

them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator:

- Jupyter Notebook
- QtConsole
- Spyder
- PyQt

For this project two application will be using are Jupyter and

Spyder: **1. Jupyter**

Project Jupyter is a nonprofit organization created to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". Spun-off from IPython in 2014 by Fernando Pérez,Project

Jupyter supports execution environments in several dozen languages. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab, the next-generation version of Jupyter Notebook.

Jupyter Lab is a web-based interactive development environment for Jupyter notebooks, code, and data. Jupyter Lab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. Jupyter Lab is extensible and modular: write plugins that add new components and integrate with existing ones.

The JupyterNotebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

Figure 3.3.1.1 Jupyter

2. 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. It is released under the MIT license. Spyder is extensible with first- and

third-party plugins,[6] includes support for interactive tools for data inspection and embeds Python-specific code quality assurance and introspection instruments, such as Pyflakes, Pylint and Rope. It is available cross-platform through Anaconda, on Windows, on macOS through MacPorts, and on major Linux distributions such as Arch Linux, Debian, Fedora, Gentoo Linux, openSUSE and Ubuntu. Spyder uses Qt for its GUI, and is designed to use either of the PyQt or PySide Python bindings. QtPy, a thin abstraction layer developed by the Spyder project and later adopted by multiple other packages, provides the flexibility to use either backend.

Features include:

- An editor with syntax highlighting, introspection, code completion. •

  Support for multiple IPython consoles.

- The ability to explore and edit variables from a GUI.

- A Help pane able to retrieve and render rich text documentation on functions, classes and methods automatically or on-demand

- A debugger linked to IPdb, for step-by-step execution

- Static code analysis, powered by Pylint

- A run-time Profiler, to benchmark code

- Project support, allowing work on multiple development effortssimultaneously.

- A built-in file explorer, for interacting with the filesystem and managing

projects.

- A "Find in Files" feature, allowing full regular expression search over a specified scope.

- An online help browser, allowing users to search and view Python and package documentation inside the IDE
- A history log, recording every user command entered in each console.
- An internal console, allowing for introspection and control over Spyder's own operation.

Figure 3.4.1.2 Spyder

3. PyQt

PyQt is a Python binding of the cross-platform GUItoolkit Qt, implemented as a Python plug-in. PyQt is free software developed by the British firm Riverbank Computing. It is available under similar terms to Qt versions older than 4.5; this means a variety of licenses including GNU General Public License (GPL) and commercial license, but not the GNU Lesser General Public License (LGPL).

PyQt supports Microsoft Windows as well as various flavours of UNIX, including Linux and MacOS (or Darwin).

PyQt implements around 440 classes and over 6,000 functions and methods including:

• A substantial set of GUI widgets

• Classes for

accessing SQL databases (ODBC, MySQL, PostgreSQL, Oracle,

SQLite) • QScintilla, Scintilla-based rich text editor widget

• Data aware widgets that are automatically populated from a database •

An XML parser

• SVG support

To automatically generate these bindings, Phil Thompson developed the tool SIP, which is also used in other projects.

In August 2009, Nokia, the then owners of the Qt toolkit, released PySide, providing similar functionality, but under the LGPL,[8] after failing to reach an agreement with Riverbank Computing to change its licensing terms to include LGPL as an alternative license.

4. Qt Designer

Qt Designer is the Qt tool for designing and building graphical user interfaces (GUIs) with Qt Widgets. You can compose and customize your windows or dialogs in a what- you-see is-what-you-get (WYSIWYG) manner, and test them using different styles and resolutions. Widgets and forms created with Qt Designer integrate seamlessly with programmed code, using Qt's signals and slots mechanism, so that you can easily assign behavior to graphical elements. All properties set in Qt Designer can be changed dynamically within the code. Furthermore, features like widget promotion and custom plugins allow you to use your own components with Qt Designer.

Note: You have the option of using Qt Quick for user interface design rather than widgets. It is a much easier way to write many kinds of applications. It enables a completely customizable appearance, touch-reactive elements, and smooth animated transitions, backed up by the power of OpenGL graphics acceleration.

If you are new to Qt Designer, you can take a look at the Getting To Know Qt Designer document. For a quick tutorial on how to use Qt Designer, refer to A Quick Start to Qt Designer.

Qt Creator is a cross-platform C++, JavaScript and QML integrated development environment which simplifies GUI application development. It is part of the SDK for the Qt GUI application development framework and uses the Qt API, which encapsulates

host OS GUI function calls. It includes a visual debugger and an integrated WYSIWYG GUI layout and forms designer. The editor has features such as syntax highlighting and autocompletion. Qt Creator uses the C++ compilerfrom the GNU Compiler Collection on Linux and FreeBSD. On Windows it can use MinGW or MSVC with the default install and can also use Microsoft Console Debugger when compiled from source code. Clang is also supported

CHAPTER 4

4. SYSTEMARCHITECTURE

4.1 Accident Data Set

Description of the Dataset:

This study used data from the National Automotive Sampling System (NASS) General Estimates System (GES). The GES datasets are intended to be a nationally representative probability samples from the annual estimated 6.4 million accident reports in the United States. The initial dataset for the study contained traffic accident records from 1995 to 2000, a total number of 417,670 cases. According to the variable definitions for the GES dataset, this dataset has drivers' records only and does not include passengers' information. The total set includes labels of year, month, region, primary sampling unit, the number describing the police jurisdiction, case number, person number, vehicle number, vehicle make and model; inputs of drivers' age, gender, alcohol usage, restraint system, eject, vehicle body type, vehicle age, vehicle role, initial point of impact, manner of collision, rollover, roadway surface condition, light condition, travel speed, speed limit and the output injury severity. The injury severity has five classes: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. In the original dataset, 70.18% of the cases have output of no injury, 16.07% of the cases have output of possible injury, 9.48% of the cases have output of no incapacitating injury, 4.02% of the cases have output of incapacitating injury, and 0.25% of the cases have fatal injury. Our task was to develop machine learning based intelligent models that could accurately classify the severity of injuries (5 categories). This can in turn lead to greater understanding of the relationship between the factors of driver, vehicle, roadway, and environment and driver injury severity. Accurate results of such data analysis could provide crucial information for the road accident prevention policy. The records in the dataset are input/output pairs with each record have an associated output. The output variable, the injury severity, is categorical and (as described above) has five classes. A supervised learning algorithm will try to map an input vector to the desired output class.

4.2 Data Preparation

When the input and output variables are considered there are no conflicts between the attributes since each variable represents its own characteristics. Variables are already categorized and represented by numbers. The manner in which the collision occurred has 7 categories: no collision, rear-end, head-on, rear-to-rear, angle, sideswipe same direction, and sideswipe opposite direction. For these 7 categories the distribution of the fatal injury is as follows: 0.56% for non-collision, 0.08% for rear-end collision, 1.54% for head-on collision, 0.00% for rear-to-rear collision, 0.20% for angle collision, 0.08% for sideswipe same direction collision, 0.49% for sideswipe opposite direction collision. Since head-on collision has the highest percent of fatal injury; therefore, the dataset was narrowed down to head-on collision only. Head-on collision has a total of 10,386 records, where 160 records show the result as a fatal injury; all of these 160 records have the initial point of impact categorized asfront.

4.3 UML DIAGRAMS

A UML diagram is a diagram based on the UML (Unified Modelling Language) with the purpose of visually representing a system along with its main actors, roles, actions, artefacts or classes, in order to better understand, alter, maintain, or document information about the system. The different types of UML diagrams used to design the system are given below:

4.3.1 Use Case Diagram

A UML use case diagram is the primary form of system/software requirements for a new software program underdeveloped. Use cases specify the expected behavior (what), and not the exact method of making it happen (how). Use cases once specified can be denoted both textual and visual representation (i.e. use case diagram). A key concept of use case modelling is that it helps us design a system from the end user's perspective. It is an effective technique for communicating system behavior in the user's terms by specifying all externally visible system behavior.

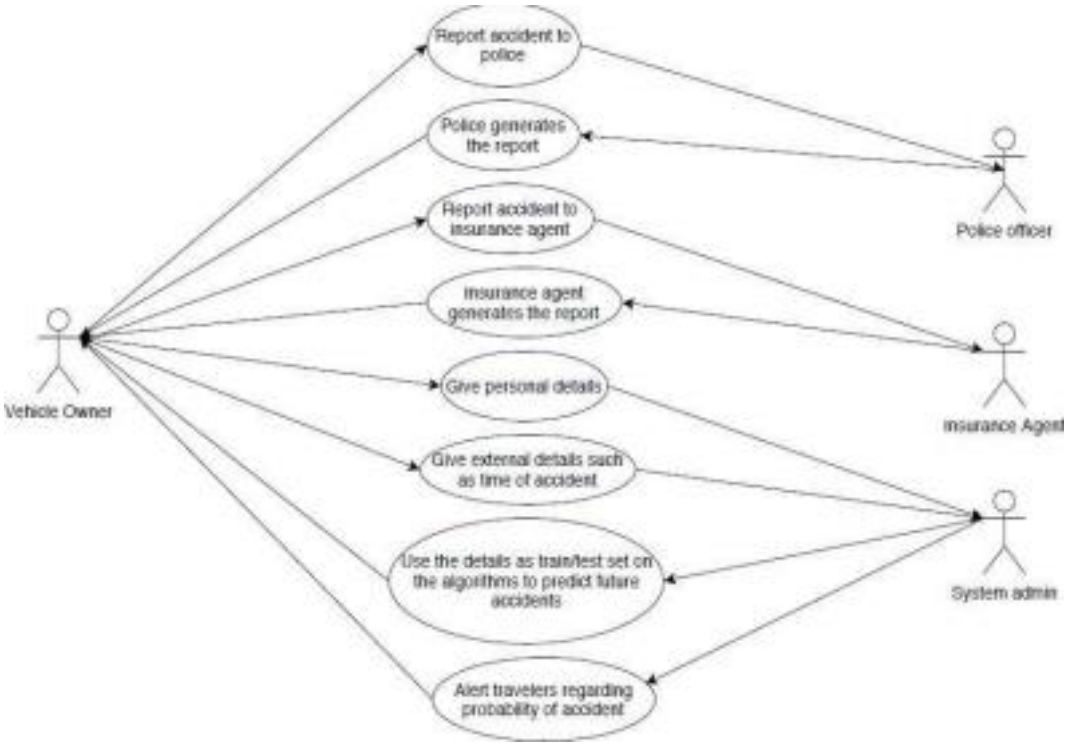

Figure 4.3.1.1 Use Case Diagram

4.3.2 Activity Diagram

Activity diagrams are used to illustrate the flow of control in a system and refer to the steps involved in the execution of a use case. Sequential and concurrent activities are modelled using activity diagrams. It basically depicts workflows visually using an activity diagram. An activity diagram focuses on condition of flow and the sequence in which it happens. It describes or depicts what causes a particular event using an activity diagram. UML models basically three types of diagrams, namely, structure diagrams, interaction diagrams, and behavior diagrams. An activity diagram is a behavioral diagram i.e. it depicts the behavior of a system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed. We can depict both sequential processing and concurrent processing of activities using an activity diagram. They are used in business and process modelling where their primary use is to depict the dynamic aspects of a system.

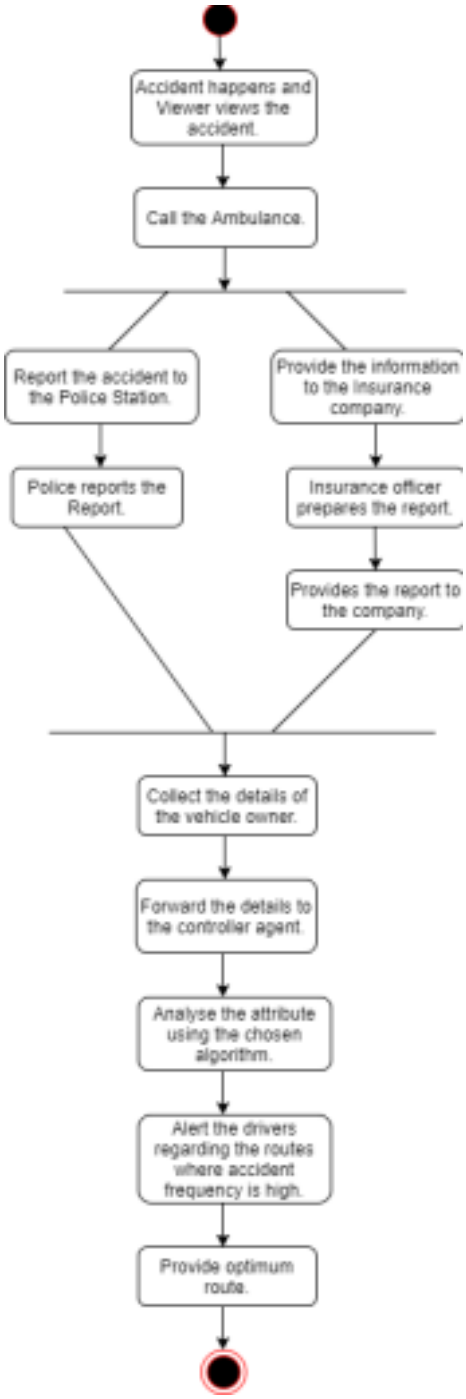

Figure 4.3.2.1 Activity Diagram

4.3.3 Data Flow Diagram

A data-flow diagram (DFD) is a way of representing a flow of a data of a process or a system (usually an information system). The DFD also provides information about the

outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and no loops. Specific operations based on the data can be represented by a flowchart. There are several notations for displaying data flow diagrams. For each data flow, at least one of the endpoints (source and / or destination) must exist in a process. The refined representation of a process can be done in another data-flow diagram, which subdivides this process into sub-processes. The data flow diagram is part of the structured-analysis modeling tools.

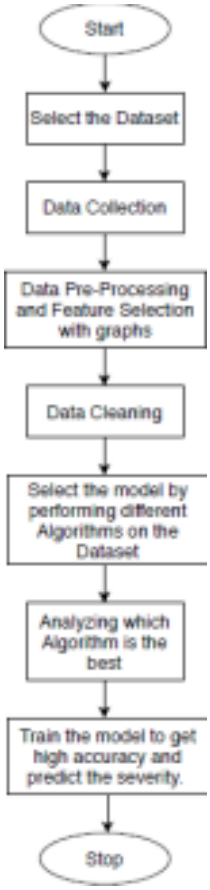

Figure 4.3.3.1 Data Flow Diagram

4.3.4 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to

document and understand requirements for new and existing systems. They capture the interaction between objects in the context of a collaboration. Sequence diagrams are time focus and they show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when. Sequence diagrams captures the interaction that takes place in a collaboration that either realizes a use case or an operation (instance diagrams or generic diagrams) and high-level interactions between user of the system and the system, between the system and other systems, or between subsystems (sometimes known as system sequence diagrams).

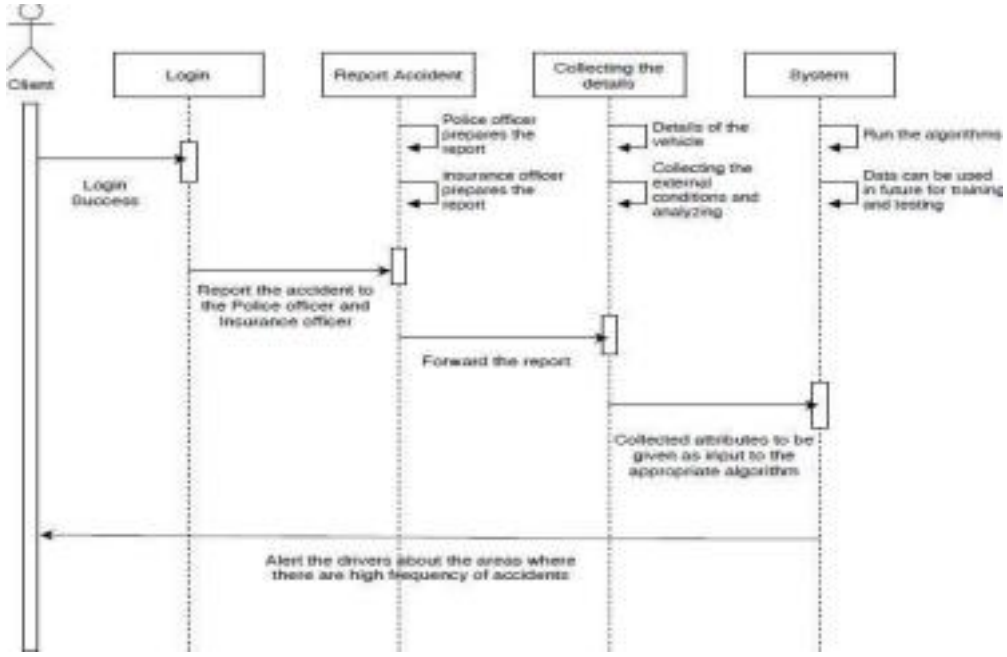

Figure 4.3.4.1 Sequence Diagram

CHAPTER 5

IMPLEMENTATION

5.1 ALGORITHMS USED

To predict the severity of an accident under the given conditions, we decided compared the accuracy of 5 algorithms and choose the one that gave the highest score. The algorithms chosen were, Decision Tree, KNN, Random Forest, Linear Regression and Logistic Regression.

5.1.1 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problemstoo.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a classlabel.

Decision Tree Algorithm Pseudocode:

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets**.** Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

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 record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

We continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modelled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.

Decision Tree Algorithm Advantages and Disadvantages:

Advantages:

1. Decision Trees are easy to explain. It results in a set ofrules.
2. It follows the same approach as humans generally follow while making decisions.
3. Interpretation of a complex Decision Tree model can be simplified by its visualizations. Even a naive person can understand logic.
4. The Number of hyper-parameters to be tuned is almost null.

Disadvantages:

1. There is a high probability of overfitting in Decision Tree.
2. Generally, it gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
3. Information gain in a decision tree with categorical variables gives a biased response for attributes with greater no. of categories.
4. Calculations can become complex when there are many classlabels.

5.1.2 KNN (K Nearest Neighbors)

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

- **KNN Algorithm Pseudocode:**

Step 1 − For implementing any algorithm, we need dataset. So, during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following –

- **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

- **3.2** − Now, based on the distance value, sort them in ascending order. •

3.3 − Next, it will choose the top K rows from the sorted array.

- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4 – End

KNN Algorithm Advantages and Disadvantages:

Advantages:

- It is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well asregression.
- It has relatively high accuracy but there are much better supervised learning models than KNN.

Disadvantages:

- It is computationally a bit expensive algorithm because it stores all the training data.
- High memory storage required as compared to other supervised learning algorithms.
- Prediction is slow in case of big N.
- It is very sensitive to the scale of data as well as irrelevantfeatures.

5.1.3 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Random Forest Algorithm Pseudocode:

- **Step 1** − First, start with the selection of random samples from a givendataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample.

Then it will get the prediction result from every decision tree.

- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result.

Random Forest Algorithm Advantages and Disadvantages:

Advantages:

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance then single decision tree.
- Random forests are very flexible and possess very high accuracy.
- Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.
- Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

Disadvantages:

- Complexity is the main disadvantage of Random forest algorithms.
- Construction of Random forests are much harder and time-consuming than decision trees.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.
- The prediction process using random forests is very time-consuming in comparison with other algorithms.

5.3 EXPECTED OUTCOME

In the road accident prediction project use the dataset is in terms of values and some data is plain English word so, the numerical values data is easily predicted and also the calculation are easily done but, the normal word are display as it is or the non predicted data are drop in the table. So, this dataset are many columns and rows and all numbers of null values will be fulfilling in forward fill method and also use the classification algorithm entire dataset. In that classification algorithm we will use Logistic Regression Algorithm The logistic algorithm will make the prediction in terms of percentage, to find accuracy level in percentage and Error percentages. This Algorithm is only for the yes and no type of result or successful and unsuccessful. The equation for combinations of all 15 input variables. The classification algorithm of the entire dataset. In the Road Accident prediction final result is to find the percentage of accident in particular area. Having lower number of features helps the algorithm to converge faster and increases accuracy. In the Road Accident prediction final result is to find the percentage of accident in particular area. Then we apply logistic regression on these features and obtain the least error

CHAPTER 6

CONCLUSION

Road Accidents are caused by various factors. By going through all the research papers, it can be concluded that Road Accident cases are hugely affected by the factors such as types of vehicles, age of the driver, age of the vehicle, weather condition, road structure and so on. Thus, we have built an application which gives efficient prediction of road accidents based on the above-mentioned factors.

By going through all the research paper, it can be concluded that road accident cases are hugely affected by the factors such as types of vehicles, age of the driver, vehicle condition and road structure. Thus, we have built an application which gives efficient prediction of road accidents based on machine learning.

CHAPTER 7

REFERENCES

[1] Peden, M. (2004) "World report on road traffic injury prevention". Geneva: World Health Organization.

[2] M. Chang, L. Y., & Chen, W. C. (2005). "Data mining of tree- based models to analyse freeway accident frequency". Journal of Safety Research, 36(4), 365-375.

[3] Tesema, T. B., Abraham, A.,& Grosan, C. (2005). "Rulemining and classification of road traffic accidents using adaptive regression trees". International Journal of Simulation, 6(10), 80-94.

[4] Maze, T. H., Agarwai, M., & Burchett, G. (2006). "Whether weather matters to traffic demand, traffic safety, and traffic operations and flow"