VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi , Karnataka, INDIA

A Project Report
on

ROAD ACCIDENT ANALYSIS USING KNN MACHINE LEARNING ALGORITHM

Submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering
in
Computer Science and Engineering

Submitted By

AKHILADEVI M	**1GA18CS017**
AMRUTHA M	**1GA18CS023**
CHANDANASHREE K R	**1GA18CS045**
AMRUTHA K	**1GA18CS194**

Under the Guidance of

Dr. ANITHA K
Associate Professor

Department of Computer Science and Engineering
Accredited by NBA(2019-2022)

GLOBAL ACADEMY OF TECHNOLOGY

Rajarajeshwarinagar, Bengaluru - 560 098
2021 – 2022

GLOBAL ACADEMY OF TECHNOLOGY

Department of Computer Science and Engineering

Accredited by NBA(2019-2022)

CERTIFICATE

Certified that the Project Entitled **"Road Accident Analysis Using KNN Machine Learning Algorithm"** carried out by **AKHILADEVI M**, bearing **USN 1GA18CS017, AMRUTHA M**, bearing **USN 1GA18CS023, CHANDANASHREE K R**, bearing **USN 1GA18CS045, AMRUTHA K**, bearing **USN 1GA18CS194,** bonafide students of Global Academy of Technology, is in partial fulfillment for the award of the **BACHELOR OF ENGINEERING** in **Computer Science and Engineering** from Visvesvaraya Technological University, Belagavi during the year 2021-2022. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the report submitted to the department. The Partial Project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Dr. Anitha K Associate Professor Dept. of CSE GAT, Bengaluru.	Dr. Bhagyashri R Hanji Professor & HOD Dept. of CSE GAT, Bengaluru.	Dr. Ranapratap Reddy. N Principal GAT, Bengaluru.

GLOBAL ACADEMY OF TECHNOLOGY

Rajarajeshwarinagar, Bengaluru – 560 098

DECLARATION

We, **AKHILADEVI M**, bearing **USN 1GA18CS017, AMRUTHA M**,bearing **USN 1GA18CS023, CHANDANASHREE K R**, bearing **USN 1GA18CS045, AMRUTHA K** ,bearing **USN 1GA18CS194**, students of Seventh Semester B.E, Department of Computer Science and Engineering, Global Academy of Technology, Rajarajeshwarinagar Bengaluru, declare that the Project Work entitled "**Road Accident Analysis Using KNN Machine Learning Algorithm"** has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree in **Bachelor of Engineering** in **Computer Science and Engineering** from **Visvesvaraya Technological University, Belagavi** during the academic year **2021-2022**.

1.AKHILADEVI M	1GA18CS017
2.AMRUTHA M	1GA18CS023
3.CHANDANASHREE K R	1GA18CS045
4. AMRUTHA K	1GA18CS194

Place: Bengaluru

Date:31-01-2022

ABSTRACT

Today, one of the top priorities for governments is traffic safety. Given the importance of the subject, identifying the causes of road accidents has become the primary goal in reducing the damage caused by traffic accidents. In used machine learning and data mining concepts to identify the various factors that influence road accidents and their severity. The application uses a Machine Learning algorithm (K Nearest Neighbor) to calculate the severity of a possible accident on a scale of 1 to 5 based on various inputs such as weather conditions, road conditions, time of day, and so on (1 being the least and 5 being the most severe). This information can be used to analyse future inputs and improve the system's output accuracy. This model can be improved further to send the accident report to the appropriate authorities, such as hospitals, ambulances, and insurance companies, and can thus be very useful in reducing accident fatality rates in the country.

Methodology: K-nearest neighbors, Logistic Regression

Key Terms: Machine Learning, KNN Algorithm, Severity, Weather Conditions.

ACKNOWLEDGEMENT

The satisfaction and the euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. The constant guidance of these persons and encouragement provide, crowned our efforts with success and glory. Although it is not possible to thank all the members who helped for the completion of the phase - 1 of project individually, we take this opportunity to express our gratitude to one and all.

We are grateful to management and our institute **GLOBAL ACADEMY OF TECHNOLOGY** with its very ideals and inspiration for having provided us with the facilities, which made this, phase - 1 project a success.

We express our sincere gratitude to **Dr. Ranapratap Reddy. N,** Principal, Global Academy of Technology for the support and encouragement.

We wish to place on record, our grateful thanks to **Dr. Bhagyashri R. Hanji** Professor and HOD, Department of CSE , Global Academy of Technology, for the constant encouragement provided to us.

We are indebted with a deep sense of gratitude for the constant inspiration, encouragement, timely guidance and valid suggestion given to us by our guide **Dr. Anitha K, Associate Professor**, Department of CSE, Global Academy of Technology.

We are thankful to all the staff members of the department for providing relevant information and helped in different capacities in carrying out this phase -1 project.

Last, but not least, we owe our debts to our parents, friends and also those who directly or indirectly have helped us to make the phase - 1 project work a success.

AKHILADEVI M	**1GA18CS017**
AMRUTHA M	**1GA18CS023**
CHANDANASHREE K R	**1GA18CS045**
AMRUTHA K	**1GA18CS194**

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

SRS	Software Requirement Specification
DFD	Data Flow Diagram
TCP	Transmission Control Protocol
KNN	K-nearest neighbors
LR	Logistic regression

CHAPTER 1

INTRODUCTION

1.1 DEFINITIONS

1.1.1 MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that improve themselves automatically over time. It is thought to be a subset of artificial intelligence. Machine learning algorithms construct a mathematical model from sample data, referred to 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 range of applications, such as email filtering and computer vision, where developing traditional algorithms to perform the required tasks would be difficult or impossible.

Machine learning is closely related to computational statistics, which focuses on using algorithms to make predictions. Mathematical optimization research contributes 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. Machine learning is also known as predictive analytics when applied to business problems. Machine learning is experience-based learning. For quite some time now, Machine Learning, a prominent topic in the Artificial Intelligence domain, has been in the spotlight.

This field may present an appealing opportunity, and establishing a career in it is not as difficult as it may appear at first glance. It is not a problem if you have no prior experience with math or programming. The most important factor in your success is your own desire to learn all of those things. As an example, consider someone who learns to play chess by watching others play. Computers will also follow the same manner. Programmed through the provision of information on which they are trained, acquiring the ability to identify elements or their characteristics with a high degree of certainty.

In machine learning, tasks are generally classified into broad categories. Fig 1.1 shows the classification of machine learning , these categories are based on how learning is received or how feedback on the learning is given to the system developed. Three of the most widely adopted machine learning methods are:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Unsupervised learning: In this case, your machine is only given a limited set of input data. Following that, it is up to the machine to determine the relationship between the entered data and any other hypothetical data. Unlike supervised learning, where the machine is given some verification data to learn from, independent unsupervised learning assumes that the computer will discover patterns and relationships between different data sets on its own. Clustering and association are two subtypes of unsupervised learning.

Supervised learning: This refers to the computer's ability to recognize elements based on samples provided. Based on this data, the computer studies it and develops the ability to recognize new data. You can, for example, train your computer to filter spam messages based on previously received data.

Reinforcement learning: It is an area of Machine Learning concerned with how intelligence agents ought to take actions in an environment in order to maximize the notation of cumulative reward. Reinforcement learning is one of the three basic machine learning paradigms, alongside supervised and unsupervised learning.

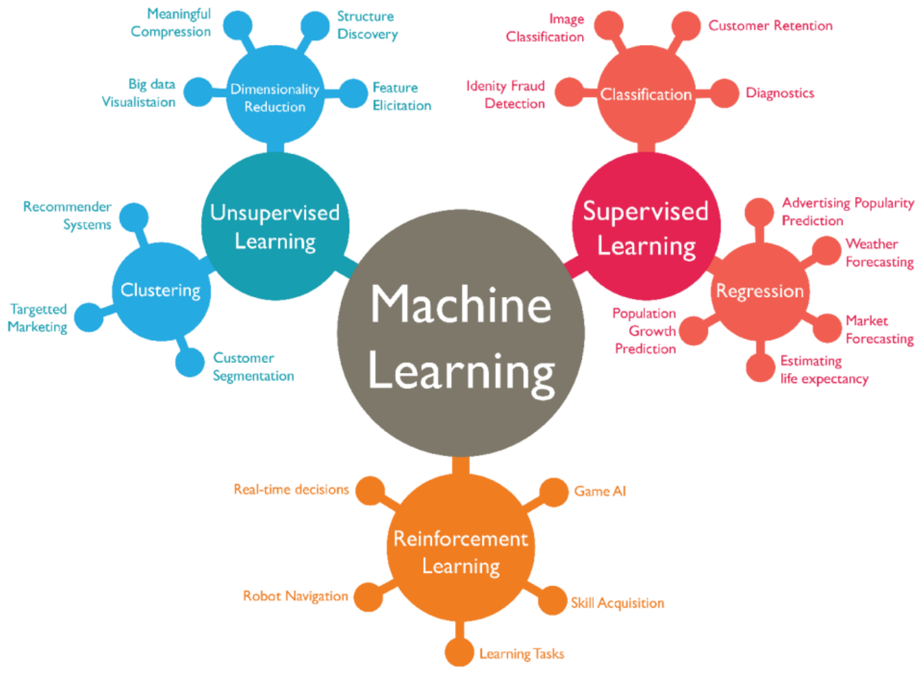

Fig 1.1: Machine Learning and its classification

- **Classification:** When inputs are divided into two or more classes, the learner must create a model that assigns unseen inputs to one or more of these classes (multi-label classification). This is usually dealt with under supervision. Classification is used in spam filtering, where the inputs are email (or other) messages and the classes are "spam" and "not spam."
- **Regression:** A supervised problem in which the outputs are continuous rather than discrete. It consists of mathematical methods that allow data scientists to predict a continuous outcome(y) based on the value of one or more predictor variables(x).
- **Clustering:** The division of a set of inputs into groups. Unlike classification, the groups are not known ahead of time, so this is typically an unsupervised task.

Various algorithms and computation techniques used in supervised machine learning processes include:

- Decision trees
- Logistic Regression
- Random Forest
- k-nearest neighbors
- linear regression

1.1.2 DATA MINING

"Data Mining", that mines the data. It is the process of understanding data through cleaning raw data, finding patterns, creating models, and testing those models. Data mining is also called Knowledge Discovery in Database (KDD). The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

There are different algorithms for different tasks. The function of these algorithms is to fit the model. These algorithms identify the characteristics of data. There are 2 types of models.

- Predictive model
- Descriptive model

Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

Descriptive modeling in unsupervised learning, the data mining algorithms describe some intrinsic property or structure of data and hence are sometimes called descriptive models. On the other hand, supervised learning techniques typically use a model to predict the value or behavior of some quantity and are hence called predictive models.

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 further investigation.
- 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.
- 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 structures to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function that models the data with the leasterror 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.

1.2 PROJECT REPORT OUTLINE

1.2.1 PROBLEM STATEMENT

Road condition monitoring is a challenging worldwide problem in the field of transportation. Poor road surface conditions create a risk of damage to vehicles and increase the chances of accidents. Using machine learning the model aims to make roads safer and accident-free. In this study, datasets containing details about previous accidents in various regions are studied and analysed, and a model that can be used to predict and prevent road accidents is developed. By comparing two scenarios based on out-of-sample forecasts, it is possible to demonstrate how a statistical method based on directed graphs works. The model is used to identify statistically significant factors that can predict the likelihood of crashes and injuries and can be used to perform a risk factor and reduce it. The database used is a public one that many institutes and government websites have access to. The data collected will be analysed, integrated, and grouped together based on different constraints using the best-suited algorithm.

1.2.2 OBJECTIVES

Traffic accidents have a significant impact on society due to the high cost of fatalities and injuries. In recent years, there has been an increase in research interest in determining the factors that significantly affect the severity of the driver's injuries caused by road accidents. Accident analysis is based on accurate and comprehensive accident records. The effective use of accident records is dependent on several factors, including data accuracy, record retention, and data analysis. There are numerous approaches to studying this problem that has been applied to this scenario.

There are several issues with the current practices for preventing accidents in the communities. Many institutes and government websites make the database we'll be using publicly available. Using the best-suited algorithm, the collected data will be analyzed, integrated, and grouped based on various constraints. This estimate will be useful in analyzing and identifying the flaw and the causes of the accidents. It will also be useful as a reference when building roads and bridges to avoid the same problems that were encountered previously. The predictions will be extremely useful in planning the management of such problems.

Models are created for this purpose using accident data records, which can aid in understanding the characteristics of many features such as driver behaviour, roadway conditions, light conditions, weather conditions, and so on. By comparing two scenarios based on out-of-sample forecasts, it is possible to demonstrate how a statistical method based on directed graphs works. The model is used to identify statistically significant factors that can predict the likelihood of crashes and injuries and can be used to perform a risk factor and reduce it.

The primary goal of the road accident prediction system is:

- To design and develop the model which will be cost-effective, simple, and determine the severity of accidents using KNN Algorithm.
- To analyse the previously occurred accidents in the locality which will help us to determine the most accident-prone area.
- To make predictions based on constraints like weather, pollution, road structure, etc.

CHAPTER 2

REVIEW OF LITERATURE

2.1 SYSTEM STUDY

- **Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model (Sep 2021):** Traffic accidents on highways are a leading cause of death despite the development of traffic safety measures. The burden of casualties and damage caused by road accidents is very high for developing countries. Many factors are associated with traffic accidents, some of which are more significant than others in determining the severity of accidents. Data mining techniques can help in predicting influential factors related to crash severity. In this study, significant factors that are strongly correlated with the accident severity on highways are identified by Random Forest. 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 proposed approach is compared with several base learner classifiers. The data used in the analysis include accident records of the USA from February 2016 to June 2020. Obtained results demonstrate that the RFCNN enhanced the decision-making process and outperformed other models with 0.991 accuracies, 0.974 precision, 0.986 recall, and 0.980 F-score using the 20 most significant features in predicting the severity of accidents.
- **Road Accident Analysis using Machine Learning (May 2020):** Abstract: Road accidents are one of the disturbing events that constitute a major loss. In India, it has become a major problem as it is claiming the lives of innocent people. Controlling road accidents has become a crucial task. Despite all that has been done to prevent road accidents, there are always areas that are accident-prone. The main aim of this model is to predict the accident-prone areas considering various factors causing accidents. This model uses data mining technique-apriori and machine learning concept- K-Means to identify the factors causing accidents.

- **A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors (Mar 2020):** Road surface quality is essential for improving the driving experience and reducing traffic accidents. Traditional road condition monitoring systems are limited in their temporal (speed) and spatial (coverage) responses needed for maintaining overall road quality. Several alternative systems have been proposed that utilize sensors mounted on vehicles. In particular, with the ubiquitous use of smartphones for navigation, smartphone-based road condition assessment has emerged as a promising new approach. In this paper, we propose to analyze different multiclass supervised machine learning techniques to effectively classify road surface conditions using accelerometer, gyroscope and GPS data collected from smartphones. Our work focuses on the classification of three main class labels- smooth road, potholes, and deep transverse cracks. We hypothesize that using features from all three axes of the sensors provides more accurate results as compared to using features from only one axis. We also investigate the performance of deep neural networks to classify road conditions with and without explicit manual feature extraction.
- **Road Accident Analysis and Prediction using Machine Learning (Jan 2020):** Engineers and researchers in the automobile industry have tried to design and build safer automobiles, but traffic accidents are unavoidable. Patterns involved in dangerous crashes could be detected by developing a prediction model that automatically classifies the type of injury severity of various traffic accidents. These behavioral and roadway patterns are useful in the development of traffic safety control policy. It is important that measures be based on scientific and objective surveys of the causes of accidents and the severity of injuries. The system presents some models to predict the severity of injury that occurred during traffic accidents using machine-learning approaches. We considered networks trained using learning approaches. Experiment results reveal that among the machine learning paradigms considered various paradigms approaches.
- **Road Accident Prediction using Machine Learning Algorithm (Mar 2019):** The traffic has been transformed into the difficult structure in points of designing and managing by the reason of an increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country

economy and done the studies on the solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need for accession to information from this large calibrated data obtained the cornerstone of data mining. In this study, the assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

- **Detecting Road Surface Condition using Smartphone Sensors and Machine Learning (2019):** Low maintenance of the roads is one of the extensive cause of increasing road accidents and vehicle breakage. Mostly roads contain potholes, wreckage or detritus which is best to be bypassed however if not sometimes lead to severe road causalities. Many car or bus accident happens over a wrecked bridge, as it slips and overturns. To avoid this kind of mishap and improve the safety of the road we have proposed a machine learning based model, MagTrack, to detect these road conditions and inform apriori to the responsible authority for fast reparation as well as other passers-by to drive carefully or take an alternative route. We have collected various road surface conditions such as smooth roads, uneven roads, potholes, speed breakers, and rumble strips data using magnetometer and accelerometer sensor embedded in our Smartphone and analyzed using various classification algorithms like Random Forest (RF), Random Tree (RT) and Support Vector Machine (SVM). The classification has done after performing the feature selection using Greedy Stepwise, Ranker, and best first Algorithms considering minimum, maximum, median and standard deviation as statistical features. 92% of accuracy to detect the road surface condition has been achieved by MagTrack.

2.2 PROPOSED WORK

Models are built from accident data records, which can aid in understanding the characteristics of many features such as roadway conditions such as cracks, potholes, smooth road. Light conditions, weather conditions such as rainy, sunny, fog etc. roadside, and so on. Machine learning will aid in the training of the model. After training, the model is trained using a decision tree to predict the severity.

By comparing two scenarios based on out-of-sample forecasts, it is possible to demonstrate how a statistical method based on directed graphs works. The model is used to identify statistically significant factors that can predict the likelihood of crashes and injuries and can be used to perform a risk factor and reduce it.

The proposed System contains a detailed solution procedure as mentioned below:

- Data Acquisition: Extraction and importing of data.
- Data pre-processing: Cleaning of data and feature extraction/selection.
- Machine Learning Training: Decision Tree, Neural Network and Regression Algorithms.
- Model Evaluation: Testing.
- Output: Prediction of severity.

2.2.1 ADVANTAGES

- The architecture of RFCNN is more efficient, accurate and gives highest performance.
- Prototypes are built which provides the precautions to the people in a particular area.
- Classifiers trained using features from all axis proved to be more accurate when compared to features from only one axis.
- Analyses and simplifies the accidental data and gives a proper and effective visualized report.
- Algorithm is triggered automatically as user inputs the data.
- MagTrack worked well in various distinguishable road condition with an average accuracy of 92%.

2.2.2 DISADVANTAGES

- Implementation of these methodologies increases the Complexity.
- Difficult to provide confidence and support values continuously.
- Sensor technology can be expensive and also SVM consumes huge amount of time.
- Concentrates only on Traffic conditions
- The application completely depends on the user inputs
- Smartphone app for data Visualization is not implemented.

2.3 SCOPE OF THE PROJECT

The primary goal of this project is to assess the utility of a J48 decision tree classifier, random forest (RF), and instance-based learning with parameter k (IBk) model for predicting and categorizing motorcycle crash severity. The models' performance was evaluated and compared to that of a multinomial logic model (MNLM). This study also identified and investigated factors that may be related to the severity of injuries in motorcycle accidents. One of the most important tasks in traffic safety is identifying factors that have a significant impact on crash severity. Based on this, the policy can be developed to reduce the number of fatalities and injuries caused by crashes. The main features include:

- A demonstration of how to open datasets can be combined to obtain meaningfulfeatures for road accident prediction.
- A high spatial and temporal resolution road accident prediction model for theisland of Montreal,
- A comparison of three algorithms dealing with data imbalance in the context ofroad accident prediction.

This research contributes to vehicle safety in three ways. First, it fills a gap in the lack of application of machine learning in motorcycle crash severity analysis. It is a novel study because an extensive review of the existing literature revealed that this is the first time the J48. Decision Tree Classifier, RF, and IBk models have been used to predict the severity of a motorcycle crash. Second, investigating contributing factors associated with motorcycle crash severity in Ghana is under-researched; thus, this study contributes to the motorcycle safety literature by filling this gap.

CHAPTER 3

SYSTEM REQUIREMENT SPECIFICATION

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 the accident happened, at what time a 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 in 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 tocalculate accident severity
- 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 NON-FUNCTIONAL REQUIREMENTS

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. This should be contrasted with functional requirements that define specific behavior or functions. The plan for implementing functional requirements are detailed in the system design. The plan for implementing non- functional requirements is detailed in the system architecture. Other terms for non-functional requirements are "constraints", "quality attributes", "quality goals", "quality of service requirements" and "non-behavioral requirements". Non-functional requirements persists

- Scalability
- Maintainability
- Accessibility
- Portability

Non-Functional Requirements include:

- **Product requirement**
  This model will work at any time, according to given data set it will predict why road accident happened in particular area, city etc.
- **Organization requirement**
  This system can be used by the government employ to know the why accidenthappened and can take precaution steps while constructing the roads.
- **External requirement**
  To access the system employees have register their information will be secured. Only the employ of the government can be used.

3.3 HARDWARE REQUIREMENTS

• Processor	: Any Processor above 500 MHz
• RAM	: 512Mb
• Hard Disk	: 10 GB
• Input device	: Standard Keyboard and Mouse
• Output device	: VGA and High-Resolution Monitor

3.4 SOFTWARE REQUIREMENTS

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, large scale 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
- Spyder
- Google Colab

- **JUPYTER**

Project Jupyter is a nonprofit organization created to "develop open-source software, open standards, and services for interactive computing across dozens of programminglanguages". 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 thediscovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and Jupyter Lab, 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 Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text and its logo is as shown in Fig 3.1. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

Fig 3.1: Jupyter

- **SPYDER**

Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language and its logo is as shown in Fig 3.2. 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 ArchLinux, 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.

Fig 3.2: Spyder

- **COLAB**

Google Colaboratory (also known as Colab) is a free Jupyter Notebook hosted environment from Google and its logo is as shown in Fig 3.3. It allows one to run a notebook wholly in the cloud and store the code and data in Google Drive. Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs. Colab notebooks are stored in Google Drive, or can be loaded from GitHub. Colab notebooks can be shared just as you would with Google Docs or Sheets. Simply click the Share button at the top right of any Colab notebook, or follow these Google Drive file sharing instructions. Colab notebooks are stored in Google Drive, or can be loaded from GitHub. Colab notebooks can be shared just as you would with Google Docs or Sheets. Simply click the Share button at the top right of any Colab notebook, or follow these Google Drive file sharing instructions

Fig 3.3: Colab

CHAPTER 4

SYSTEM DESIGN

4.1 DESIGN OVERVIEW

The Proposed solution perform the following sequence of tasks:

- The data must be collected from the various sources and users.
- Pre-processing of data must be done to retrieve the data in the expected format.
- The data must then be checked if all the required data is available.
- All unwanted data must be eliminated to reduced processing time.
- Models must be trained from collected datasets using various algorithms.
- The models are then tested with data not used during training.
- The best algorithm with the highest accuracy must be chosen as the modelthat will be used for classification.
- Once the model is trained and tested it can be used to classify the datagiven by users from social media websites.
- The model will then process the given data and provide only the messagesthatit deems to be fraudulent/criminal.
- This data will then be passed to an algorithm that will iterate through the data and provide the sources of all the messages while eliminating the forwarded messages.
- The algorithm with the fastest response will be chosen as the algorithmthat will be used to retrieve the source of the messages.
- The algorithm will be chosen based on its complexity.
- Since most data received will be huge, an algorithm with the fastestresponse is the most preferable.

4.2 SYSTEM ARCHITECTURE

System Design presents a simple view of the overall working process as shown in Fig 4.1. Decision Tree, KNN, and Logistic regression are the machine learning techniques used for road accident analysis. Data Gathering a large number of traffic accident records with complete information are required to train using the proposed approaches for accurate prediction of accident severity. Pre-processing of data all of the accident records in this dataset were written in formal language. Based on the feature, we properly organize the entire dataset. Feature Selection In order to obtain a more accurate prediction, feature selection is an important factor to consider.

The proposed approaches' methodology. The Decision Tree It is a graphical representation of all possible solutions to a problem/decision given certain conditions. KNN is a feature-similarity-based classification algorithm. It analyses the data, calculates the distance and similarities between them, and groups them into clusters based on K values. Distance can be calculated in a variety of ways; for this study, we will use the Euclidean distance measurement. The process of modelling the probability of a discrete outcome given an input variable is known as logistic regression. The most widely used logistic regression models

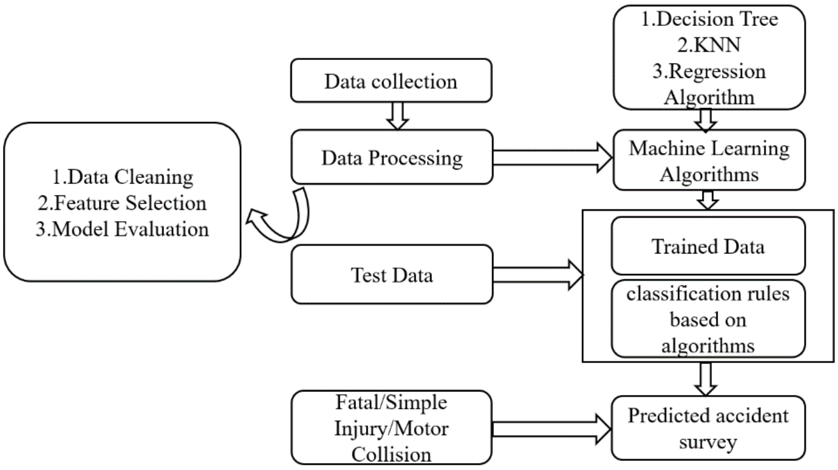

Fig 4.1: System design of Road Accident Analysis

4.3 DATA FLOW DIAGRAM

A data-flow diagram (DFD) is a method of representing the data flow of a process or system (usually an information system). The DFD also includes information about each entity and the process's outputs and inputs. A data-flow diagram lacks control flow, decision rules, and loops. A flowchart can represent specific operations based on data. Data-flow diagrams can be displayed using a variety of notations. A process must contain at least one of the endpoints (source and/or destination) for each data flow. Another data-flow diagram can be used to refine a process's representation by subdividing it into sub-processes.

4.3.1 DATA FLOW DIAGRAM-LEVEL 0

It is also known as a context diagram. It's designed to be an abstraction view as shown in Fig 4.2, which is the system that show as single process with its relationship to external entities. It represents the entire system as a single bubble with input and output data indicated by incoming/outgoing arrows.

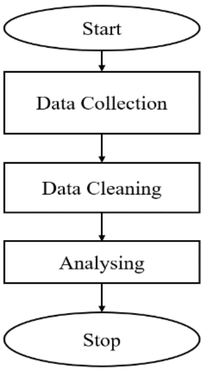

Fig 4.2: Level 0 Data Flow Diagram of Road Accident Analysis

4.3.2 DATA FLOW DIAGRAM-LEVEL 1

In level-1 DFD as shown in Fig 4.3, the context diagram is decomposed into multiple bubbles/processes. In this level, we highlight the main functions of the system and breakdown the high-level process of level-0 DFD into sub processes.

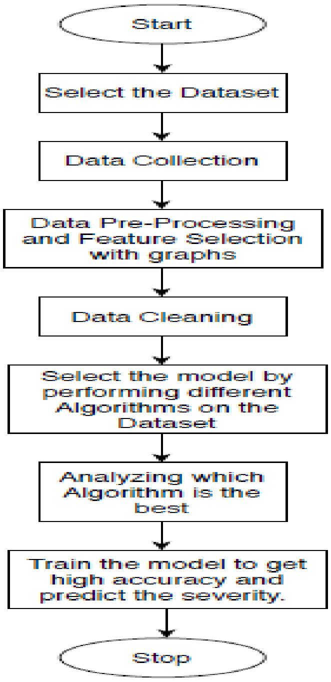
Start
Select the Dataset
Data Collection
Data Pre-Processing and Feature Selection with graphs
Data Cleaning
Select the model by performing different Algorithms on the Dataset
Analyzing which Algorithm is the best
Train the model to get high accuracy and predict the severity.
Stop

Fig 4.3: Level 1 Data Flow Diagram of Road Accident Analysis

4.4 USE CASE DIAGRAM

A UML use case diagram is the primary form of system/software requirements for an undeveloped software program as shown in Fig 4.4. The use case specifies the expected behaviour (what) rather than the exact method of achieving it (how). Once defined, use cases can have both textual and visual representations (i.e, use case diagram). A key concept of use case modelling is that it allows us to design a system from the perspective of the end-user. By specifying all externally visible system behavior, it is an effective technique for communicating system behavior in user terms.

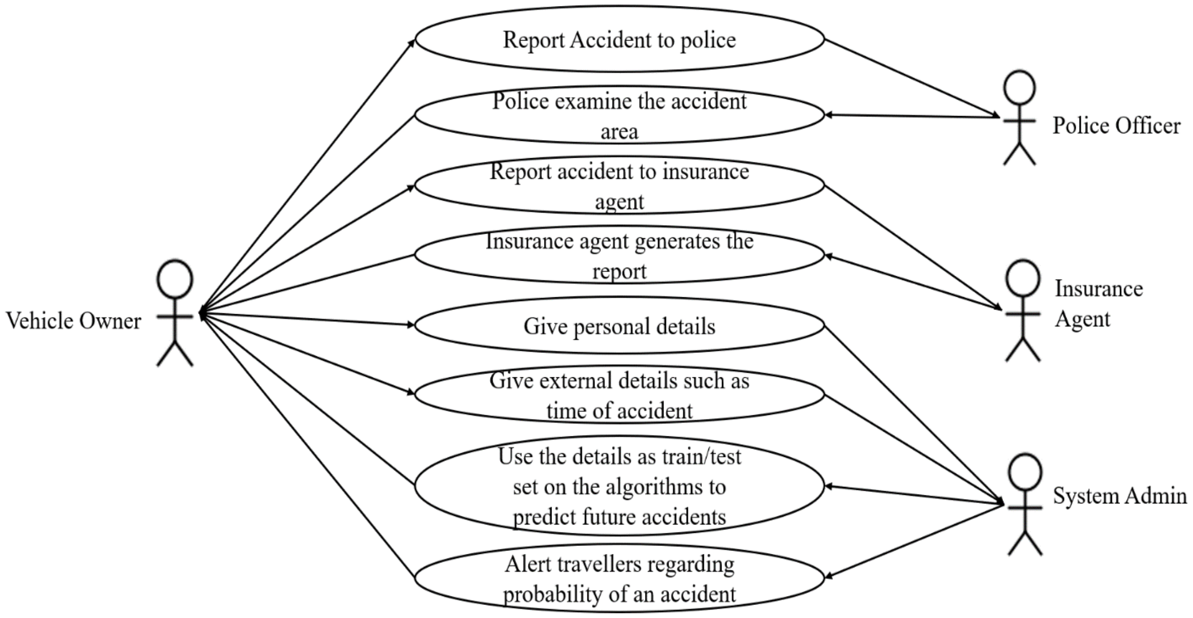
Report Accident to police
Police examine the accident area
Report accident to insurance agent
Insurance agent generates the report
Give personal details
Give external details such as time of accident
Use the details as train/test set on the algorithms to predict future accidents
Alert travellers regarding probability of an accident
Vehicle Owner
Police Officer
Insurance Agent
System Admin

Fig 4.4: Use case diagram of Road Accident Analysis

4.5 CLASS DIAGRAM

A static diagram is a class diagram. It represents an application's static view. A class diagram is used not only for visualising, describing, and documenting various aspects of a system, but also for building executable code for a software application. A class diagram describes a class's attributes and operations, as well as the constraints imposed on the system. Because they are the only UML diagrams that can be mapped directly to object-oriented languages, class diagrams are widely used in the modelling of object-oriented systems. A class diagram is a visual representation of a collection of classes, interfaces, associations, collaborations, and constraints. It is also referred to as a structural diagram.

Purpose of Class Diagrams the purpose of class diagram is to model the static view of an application. As shown in Fig 4.5, Class diagrams are the only diagrams which can be directly mapped with object-oriented languages and thus widely used at the time of construction. UML diagrams like activity diagram, sequence diagram can only give the sequence flow of the application, however class diagram is a bit different. It is the most popular UML diagram in the coder community.

The purpose of the class diagram can be summarized as −

- Analysis and design of the static view of an application.
- Describe responsibilities of a system.
- Base for component and deployment diagrams.
- Forward and reverse engineering.

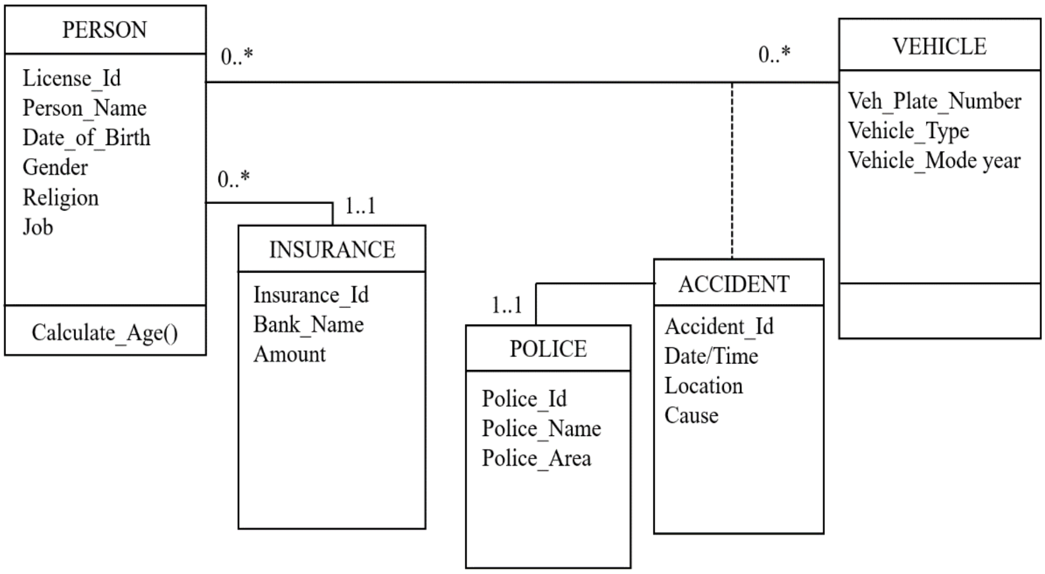

Fig 4.5: Class Diagram of Road Accident Analysis

4.6 SEQUENCE DIAGRAM

A sequence diagram simply depicts the interaction of objects in sequential order, i.e. the order in which these interactions occur as shwon in Fig 4.6. Sequence diagrams show how and in what order the objects in the system work. Businesspeople and software developers frequently use these diagrams to document and understand requirements for new and existing systems. They document the interaction of objects in the context of a collaborative effort. Sequence diagrams are time-focused, and they visually represent the order of the interaction by using the vertical axis of the diagram to represent time, what messages are sent, and when. Sequence diagrams depict the interaction that occurs in a collaboration that either realizes a use case or an operation (instance diagrams or generic diagrams) as well as high-level interactions between the system's user and the system, the system and other systems, or subsystems (sometimes known as system sequence diagrams).

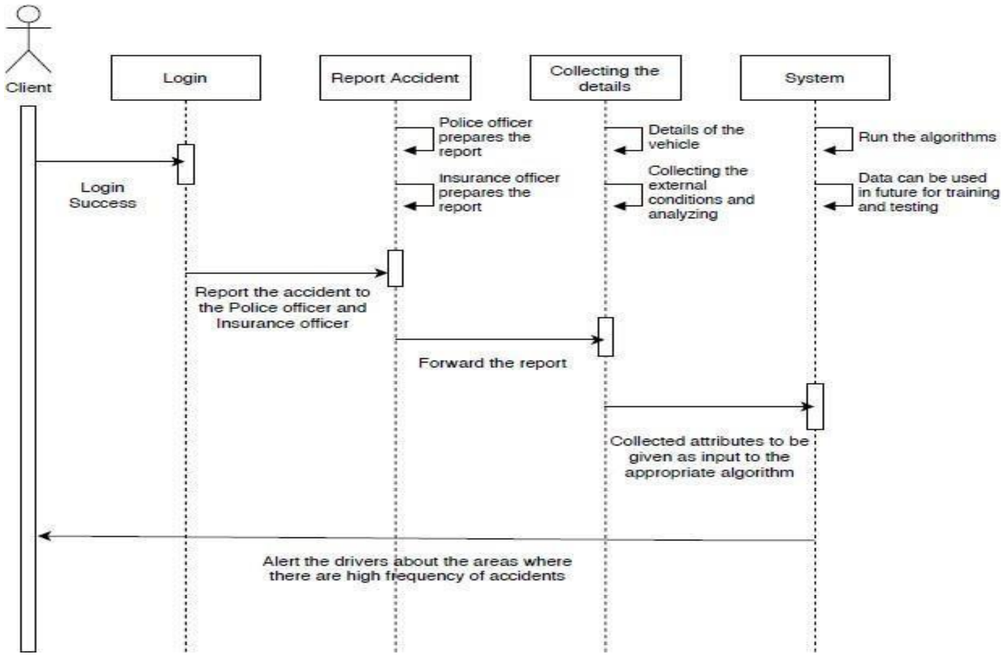
Client
Login
Report Accident
Collecting the details
System
Police officer prepares the report
Insurance officer prepares the report
Details of the vehicle
Collecting the external conditions and analyzing
Run the algorithms
Data can be used in future for training and testing
Login Success
Report the accident to the Police officer and Insurance officer
Forward the report
Collected attributes to be given as input to the appropriate algorithm
Alert the drivers about the areas where there are high frequency of accidents

Fig 4.6: Sequence Diagram of Road Accident Analysis

4.7 ACTIVITY DIAGRAM

Activity diagrams are used to depict the flow of control in a system and to refer to the steps involved in carrying out a use case. Activity diagrams are used to represent sequential and concurrent activities. It essentially uses an activity diagram to visually depict workflows. As shown in Fig 4.7, an activity diagram focuses on the state of flow and the order in which it occurs. Using an activity diagram describes or depicts what causes a specific event. UML primarily models three types of diagrams: structure diagrams, interaction diagrams, and behavior diagrams. An activity diagram is a behavioral diagram, which depicts a system's behavior. An activity diagram depicts the control flow from a starting point to a finishing point, highlighting the various decision paths that exist while the activity is being performed. Using an activity diagram, we can depict both sequential and concurrent processing of activities. They are commonly used in business and process modelling to depict the dynamic aspects of a system.

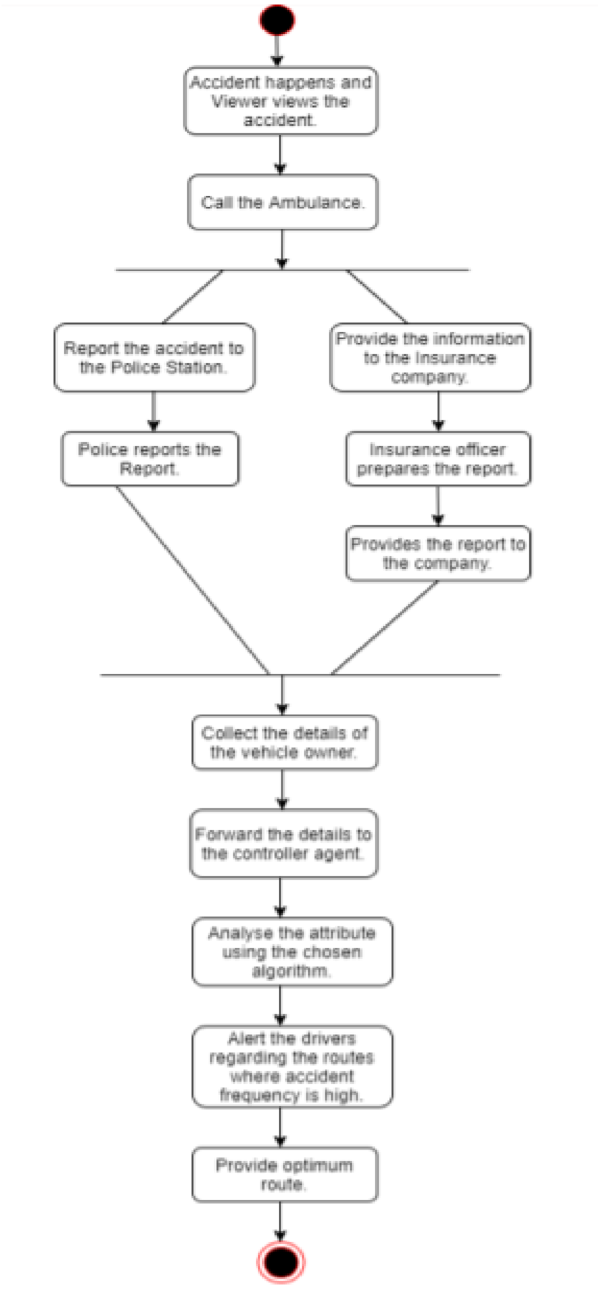
Accident happens and Viewer views the accident.
Call the Ambulance.
Report the accident to the Police Station.
Provide the information to the Insurance company.
Police reports the Report.
Insurance officer prepares the report.
Provides the report to the company.
Collect the details of the vehicle owner.
Forward the details to the controller agent.
Analyse the attribute using the chosen algorithm.
Alert the drivers regarding the routes where accident frequency is high.
Provide optimum route.

Fig 4.7: Activity Diagram of Road Accident Analysis

4.8 MODULE SPLIT-UPS

4.8.1 DATA ACQUISITION

Data acquisition is the process of collecting data, including what data is acquired, how, and why. Data management begins with data acquisition: from the moment that the University is in possession of data, it has a responsibility for managing it appropriately, including complying with laws and regulations that may apply to that data. The same is true for units and individuals; everyone has a responsibility to appropriately manage the data entrusted to their care.

Guidelines for data acquisition:

- Collect only necessary data.
- Be aware of restrictions on data or its collection.
  Legal, regulatory, privacy, or other restrictions can apply to data collection. This may include:
    - providing notice of data collection
    - obtaining consent to data collection
    - collecting only certain data
    - Obtaining a contract or agreement prior to data collection.
- Consider the source of the data.
- Manage data creation.

4.8.2 DATA PREPROCESSING

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Need of Data pre-processing

- For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, the Random Forest algorithm does not support null values, therefore to execute a random forest algorithm null values have to be managed from the original raw data set.
- Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and the best out of them is chosen.

4.8.3 MACHINE LEARNING MODELS

4.8.3.1 KNN

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.

Working of KNN Algorithm

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of the following steps –

- For implementing any algorithm, we need a dataset. So during the first step of KNN, we must load the training as well as test data.
- Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.
- For each point in the test data do the following –
    - Calculate the distance between test data and each row of training data with the help of any of the methods namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
    - Now, based on the distance value, sort them in ascending order.
    - Next, it will choose the top K rows from the sorted array.
    - Now, it will assign a class to the test point based on a most frequent class of these rows.
- End

4.8.3.2 LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Types of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

4.8.4 MODEL EVALUATION

To evaluate the performance of the classifiers various performance evaluation metrics are used for machine learning models. For each of the classifiers, we consider relevant and important parameters which best enable us to derive a conclusion on its performance. A confusion matrix is a specific tabular representation of the performance of a supervised machine learning algorithm. Each column represents the number of instances of the predicted class while each row represents the number of instances of an actual class. Most classification metrics are derived from the confusion matrix based on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). A classifier's accuracy, precision, and recall are described.

There are four different outcomes that can occur when your model performs classification predictions:

True positives occur when your system predicts that an observation belongs to a class and it actually does belong to that class.

True negatives occur when your system predicts that observation does not belong to a class and it does not belong to that class.

False positives occur when you predict an observation belongs to a class when in reality it does not. Also known as a type 2 error.

False negatives occur when you predict an observation does not belong to a class when in fact it does. Also known as a type 1 error.

Following Evaluation Metrics:

Precision: This refers to the proportion (total number) of all observations that have been predicted to belong to the positive class and are actually positive. The formula for Precision Evaluation Metric is as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall: This is the proportion of observation predicted to belong to the positive class, which truly belongs to the positive class. It indirectly tells us the model’s ability to randomly identify an observation that belongs to the positive class. The formula for Recall Evaluation Metric is as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1 Score: This is an averaging Evaluation Metric that is used to generate a ratio. The F1 Score is also known as the Harmonic Mean of the precision and recall Evaluation Metrics. This Evaluation Metric is a measure of overall correctness that our model has achieved in a positive prediction environment i.e., of all observations that our model has labeled as positive, how many of these observations are actually positive. The formula for F1 Score Evaluation Metric is as follows:

$$\text{F1 Score} = \frac{2\text{*Precision * Recall}}{\text{Precision + Recall}}$$

CHAPTER 5

CONCLUSION

Accidents on the road are caused by a variety of factors. After reviewing all of the research papers, it is possible to conclude that factors such as vehicle types, driver age, vehicle age, weather condition, road structure, and so on have a significant impact on road accident cases. As a result, we developed an application that provides accurate predictions of road accidents based on the aforementioned factors. After reviewing all of the research papers, it is possible to conclude that factors such as vehicle types, driver age, vehicle condition, and road structure have a significant impact on road accident cases. As a result, we created an application that predicts road accidents using machine learning.

Proposed solutions for road damage recognition achieved an 87% accuracy, the expected output aims to achieve accuracy of within 85-90%. An efficient solution will be developed for Road Accident Analysis.

BIBLIOGRAPHY

[1]. Mubariz Manzoor, Muhammad Umer, Saima Sadiq Saleem Ullah, Hamza Ahmad Madni, abid Ishaq, and Carmen bisogni.”Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model.” IEEE Access 10.1109/ACCESS.2021.3112546

[2]. N. Sridevi, M.V. Keerthana, Monisha V. Pal, T.R. Nikshitha, P. Jyothi. ”Road Accident Analysis Using Machine Learning.”International Journal of Research in Engineering, Science and ManagementVolume-3, Issue-5, May-2020 www.ijresm.com|ISSN(Online):2581-5792

[3]. Akanksh Basavaraju, Jing Du, Fujie Zhou, and JimJi. ”A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors.”IEEE SENSORS JOURNAL, VOL.20, NO.5, MARCH1, 2020.”

[4]. Sahil Dabhade, Sai Mahale, Avinash Chitalkar, Pushkar Gawhad, Vicky Pagare.”Road Accident Analysis and Prediction using Machine Learning. ”International Journal for Research in Applied Science & Engineering Technology (IJRASET)ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177 Volume 8 Issue I, Jan2020-Available at www.ijraset.com

[5]. Vipul Rana, Hemant Joshi, Deepak Parmar, Pradnya Jadhav, Monika Kanojiya. ”Road Accident Prediction using Machine Learning Algorithm.” International Research Journal of Engineering and Technology (IRJET) Volume: 06Issue:03|Mar2019 www.irjet.net .

[6]. Meenu Rani Dey, Utkalika Satapathy, Pranali Bhanse, Bhabendu Kr. Mohanta, Debasish Jena. ”Detecting Road Surface Condition using Smartphone Sensors and Machine Learning. ”Information Security Lab IIIT Bhubaneswar Odisha, India a117004@iiit-bh.ac.in 978-1-7281-1895-6/19/$31.00 c2019 IEEE.

[7]. Rewa Desale, Khushboo chudhari, Harshada Pawar, Arati Patil, Bhushan Nandwalkar **“**A Lifesaver: Healthcare System for Road Accident and Emergency Patient”. IOSR Journal of Computer Engineering (IORS-JCE) Volume 23, Issue 3, Ser. 1 (May-June 2021) www. Iosrjournals.org.

[8]. Jayesh Patil; Mandar Prabhu; Dhaval Walavalkar; Vivian Brian Lobo "Road Accident Analysis using Machine Learning". IEEE Pune Section International Conference (PuneCon) 10.1109/PuneCon50868.2020.9362403, 01 March 2021 .

[9]. J. Lepine, V. Rouillard, and M. Sek, "On the use of machine learning to detect shocks in road vehicle vibration signals," Packaging Technol. Sci., vol. 30, no. 8, pp. 387–398, Aug. 2017.

[10]. K. Chen, M. Lu, X. Fan, M. Wei, and J. Wu, "Road condition monitoring using on-board three-axis accelerometer and GPS sensor," in Proc. 6th Int. ICST Conf. Commun. Netw. China (CHINACOM), Aug. 2011, pp. 1032–1037.

[11]. F. Benedetto and A. A. Benedetto Tedeschi, "GPR image and signal processing for pavement and road monitoring on Android smartphones and tablets," in Proc. EGU Gen. Assem. Conf. Abstr., May 2014.

[12]. J. Masino, J. Thumm, M. Frey, and F. Gauterin, "Learning from the crowd: Road infrastructure monitoring system," J. Traffic Transp. Eng., vol. 4, no. 5, pp. 451–463, Oct. 2017.