

ANALYSIS OF MACHINE LEARNING ALGORITHM WITH ROAD ACCIDENTS DATA SETS

Submitted in partial fulfillment of the requirements for the award of

Bachelor of Engineering degree

In

Computer Science and Engineering

By

Madireddy Divya (38110286)
ABBURI SUSHMA (Reg No. 38110003)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI - 600 119

APRIL - 2020



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai –
600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Madireddy Divya (38110286) ABBURI SUSHMA (Reg No. 38110003)**

, who carried out the project entitled **"ANALYSIS OF MACHINE LEARNING ALGORITHM WITH ROAD ACCIDENTS DATA SETS"** under my supervision from **NOVEMBER 2019** to **APRIL 2020**.

Internal Guide

Dr. S. VIGNESHWARI M.E., Ph.D

Head of the Department

Dr.L.Lakshmanan.,M.E.,Ph.D.,

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

We Madireddy Divya (38110286) ABBURI SUSHMA (Reg No. 38110003)

hereby declare that the Project Report entitled **“ANALYSIS OF MACHINE LEARNING ALGORITHM WITH ROAD ACCIDENTS DATA SETS”** done by me under the guidance of Dr. S. VIGNESHWARI M.E., Ph.D is submitted in partial fulfillment of the requirements for the award of **Bachelor of Engineering degree in Computer Science and Engineering.**

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean**, School of Computing, also **Dr. S. VIGNESHWARI M.E., Ph.D., and Dr. L. LAKSHMANAN M.E., Ph.D., Heads** of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S. Vigneshwari M.E, Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Currently, road transport infrastructure failing to cope up with the exponential increase in vehicular population and to computing the fastest driving routes and accidents in the presence of varying traffic conditions is an essential problem in modern navigation systems. To prevent this problem is to investigate the transport department dataset with ensemble learning method for finding the best road selection without accident forecasting by prediction results of best accuracy calculation by comparing supervised machine learning algorithms. In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Additionally, to compare and discuss the performance of various machine learning algorithm measurements from the given transport department dataset with evaluation of GUI based road accident prediction by given attributes.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	V
	LIST OF FIGURES	IX
	LIST OF ABBREVIATIONS	X
1	INTRODUCTION	1
	1.1 EXISTING SYSTEM	3
	1.2 PROPOSED SYSTEM	4
	1.2.1 Ensemble learning	4
	1.2.2 Max Voting	5
	1.2.3 Averaging	5
	1.2.4 Weighted Average	5
	1.2.5 Voting based Ensemble learning	6
	1.3 AIM	6
	1.4 SCOPE	7
	1.5 OBJECTIVE	7
2	LITERATURE SURVEY	8
	2.1 LITERATURE SURVEY	8
3	METHODOLOGY	12
	3.1 METHODOLOGIES	12
	3.1.1 Sequential Ensemble learning (Boosting)	12
	3.1.2 Parallel Ensemble Learning (Bagging)	12
	3.1.3 Stacking & Blending	13
	3.2 FEASIBILITY STUDY	14
	3.2.1 Data Wrangling	14
	3.2.2 Data collection	14
	3.2.3 Preprocessing	14
	3.3 CONSTRUCTION OF A PREDICTIVE MODEL	14
	3.3.1 Dataflow diagram for machine learning	15

3.3.2 Work flow diagram	16
3.3.3 UML Diagram	16
3.3.3.1 Use Case Diagram	16
3.3.3.2 Activity Diagram	17
3.3.4 Sequence Diagram	18
3.4 PROJECT REQUIREMENTS	18
3.5 ENVIRONMENTAL REQUIREMENTS	19
3.5.1 Software Description	20
3.5.2 Anaconda Navigator	20
3.5.3 Conda	21
3.5.4 The Jupyter Notebook	21
3.5.5 Notebook document	21
3.5.6 Jupyter Notebook App	22
3.5.7 Kernel	22
3.5.8 Notebook Dashboard	22
3.6 MODULE-01	24
3.6.1 Data validation process	24
3.6.2 Data Validation	24
3.6.3 Data Pre-processing	25
3.7 MODULE - 02	26
3.7.1 Exploration data analysis of visualization	26
3.7.2 Training the Dataset	28
3.7.3 Testing the Dataset	28
3.8 MODULE - 03	29
3.8.1 Logistic Regression	29
3.8.2 Decision Tree	30
3.9 MODULE -04	31
3.9.1 Support Vector Machines (SVM)	31
3.9.2 Random Forest	32
3.10 MODULE -05	33
3.10.1 K-Nearest Neighbor (KNN)	33

	3.10.2 Naive Bayes algorithm	34
	3.11 MODULE – 06	35
	3.12 MODULE - 07	36
	3.12.1 Accuracy calculation	36
	3.12.2 Comparing Algorithm with prediction result	37
4	RESULTS AND DISCUSSION	40
	4.1 RESULTS	40
	4.2 DISCUSSION	40
5	CONCLUSION AND FUTURE WORK	41
	5.1 CONCLUSION	41
	5.2 FUTURE WORK	41
	REFERENCES	42
	APPENDIX	43
	A.SAMPLE CODE	43
	B.SCREEN SHOTS	48
	C.PUBLICATION WITH PLAGIARISM REPORT	50

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO.
1.1	Process of Machine learning	3
1.2	Ensemble structure	5
1.3	Ensemble model	6
3.1	Process of dataflow diagram	15
3.2	Data flow diagram for Machine learning model	15
3.3	Workflow Diagram	16
3.4	Use Case Diagram	17
3.5	Activity Diagram	17
3.6	Sequence Diagram	18
3.7	Pre-processing Data	25
3.8	Analysis of data	26
3.9	Result of Logistic Regression	29
3.10	Result of Decision Tree Classifier	30
3.11	Result of Support Vector Machine	32
3.12	Result of Random Forest Algorithm	33
3.13	Result of K-Nearest Neighbor	34
3.14	Result of Naive Bayes	35
3.15	Confusion matrix	37

LIST OF ABBREVIATION

ABBREVIATION	EXPANSION
FN	False Negatives
FP	False Positives
GUI	Graphical User Interfaces
NBA	Naive Bayes Algorithm
PEL	Parallel Ensemble Learning
SEL	Sequential Ensemble learning
SVM	Support Vector Machines
TN	True Negatives
TP	True Positives

CHAPTER 1

INTRODUCTION

The amount of accident data stored by traffic management departments has been growing at an ever-increasing rate in the last few years because of numerous road traffic accidents. Government entities and some private sectors were busy collecting accident data at daily bases. Data from accidents is often among the most valuable assets for local authorities and central governments, as it helps in budgeting and implementation of policies. It can also help policymakers to make appropriate decisions pertaining to infrastructure development, planning and social grants allocation. But as the amount of this data is growing, there is high demand and a need of finding methods, technique and tools to analyze such large volumes of data and find a solution to the cause of accidents. Road traffic accidents are a major worldwide threat that continues to cause casualties, injuries and fatalities on roadways on a daily basis, resulting in huge losses both at the economic and social levels. This global problem needs more attention to reduce the severity and the frequency of accident occurrence. The past data about previous accidents represents a formidable opportunity for researchers to identify the most influential factors in such accidents, which in turn play a key role in finding appropriate solutions to mitigate this problem in the future. In recent years, several data mining techniques have been effectively used to extract useful knowledge from large data sets containing information about traffic accidents. Road and traffic accidents are one of the major causes of fatality and incapacitation across the world. Road contingency can be considered as an event in which a conveyance collides with other conveyance, person or other objects. A road contingency not only provides property damage but it may lead to partial or full incapacitation and sometimes can be fatal for human being. Incrementing number of road accidents is not a good sign for the conveyance safety. The only solution requires the analysis of traffic contingency data to identify different causes of road accidents and taking preventive measures. Machine Learning is one of the applications of Artificial Intelligence technique where the machine learns the data implicitly rather than explicit programming. Nowadays, machine learning plays a crucial role in our day to day life. It is used almost in every field like transport, medical, banking etc. in which transport and medical field has more

importance than others as they are related to lives. In the field of transportation, machine learning can be applied in many sectors like traffic flow prediction, accident prediction, tourist place suggestion etc. Machine learning is used not only for automation but also for safety.

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance. Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class or it may be multi-class too. Some examples of classification problems are speech recognition, handwriting recognition, bio metric identification, document classification etc.

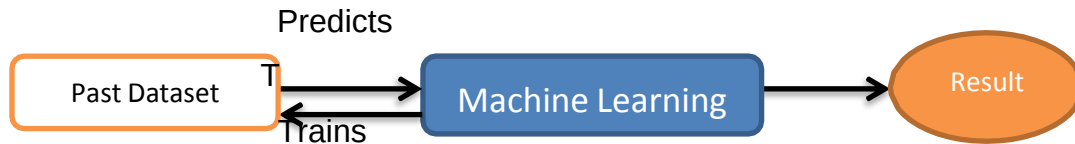


Fig 1.1: Process of Machine learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where we have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is $y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

Existing system:

The traffic signals optimized in a receding horizon framework can quickly adapt to the variations in traffic flows or change in some vehicle behavior. To minimize the total crossing time of all the vehicles, usually the shortest possible duration of signals are generated. Such short signals almost eliminate the stop-delay of the vehicles at any penetration level when the intersection is under saturated. As the portion of the video increases, the traffic density, average speed, stop-delay are improved significantly. The scheme is found to be environment friendly as it significantly reduces the fuel consumption and emissions around the intersection. The videos use distributed

controllers without using the detailed information of its preceding traffic accidents in deciding their driving action.

The traffic signals are optimized in a receding horizon control framework that aims at minimizing the total crossing time of all vehicles, considering their dynamical states. The control scheme ensures comfortable crossing of manually driven vehicles by retaining the basic features of the traditional signal management systems. The optimal signal changing times are broadcasted one cycle ahead, which enables the automated vehicles to tune their speed in order to cross the intersection with minimum stop-delay. More specifically, the framework optimizes the green time of each signal without considering the existing cycle-split concept explicitly. It is observed that the optimization process usually results in the shortest possible green period of each signal that can be realized without reducing the capacity of the intersection at any traffic volumes. Consequently, the resulting short signal cycle which is adaptive to the traffic around the intersection improves the average speed and reduces both the traffic density and the number of idling vehicles. As a consequence.

Drawbacks:

- It can't get videos or any information from transport department legally to predict traffic accidents for user interface through dataset.
- It defined only traffic control scenario information.
- It's not achieves more accurate prediction result of parameters.

PROPOSED SYSTEM

Ensemble learning:

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model and it is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model. In the world of Statistics and Machine Learning, Ensemble learning techniques attempt to make the performance of the predictive models better by improving their accuracy. Ensemble Learning is a process using which multiple machine learning models (such as classifiers) are strategically constructed to solve a particular problem.

Max Voting:

The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which we get from the majority of the models are used as the final prediction.

Averaging:

Similar to the max voting technique, multiple predictions are made for each data point in averaging. In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

Weighted Average:

This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction.

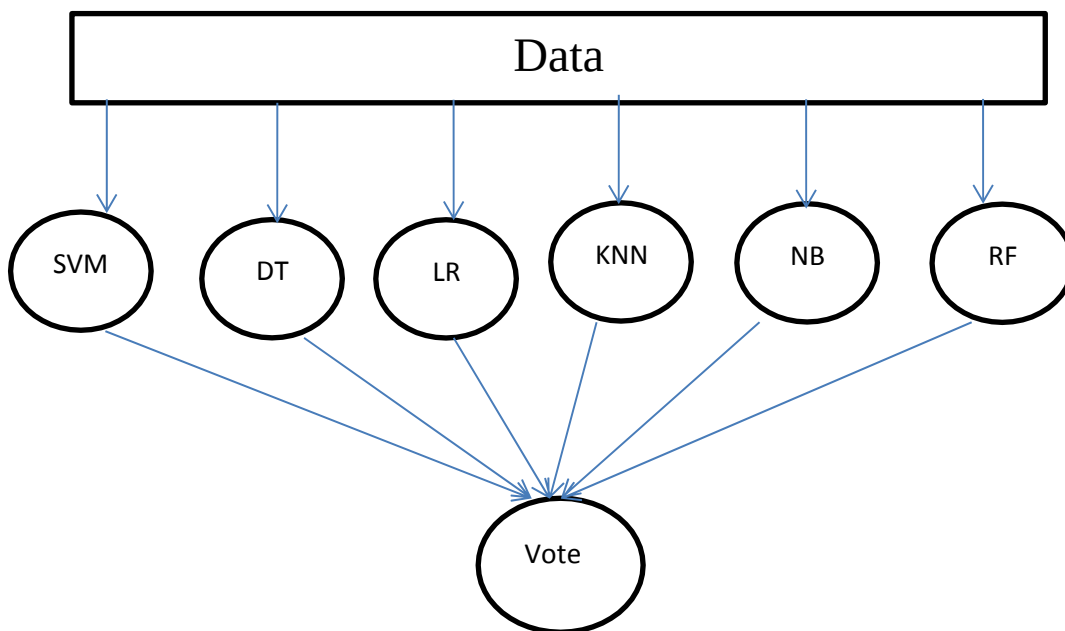


Fig 1.2: Ensemble structure

Important points: (while using ensemble model prediction)

- Better accuracy (low error)
- Higher consistency (Avoids over fitting)
- Reduces bias and variance errors

- Single model over fits
- Results worth the extra training
- It can be used for classification as well as regression

Voting based Ensemble learning:

Voting is one of the most straightforward Ensemble learning techniques in which predictions from multiple models are combined. The method starts with creating two or more separate models with the same dataset. Then a Voting based Ensemble model can be used to wrap the previous models and aggregate the predictions of those models. After the Voting based Ensemble model is constructed, it can be used to make a prediction on new data. The predictions made by the sub-models can be assigned weights. Stacked aggregation is a technique which can be used to learn how to weigh these predictions in the best possible way.

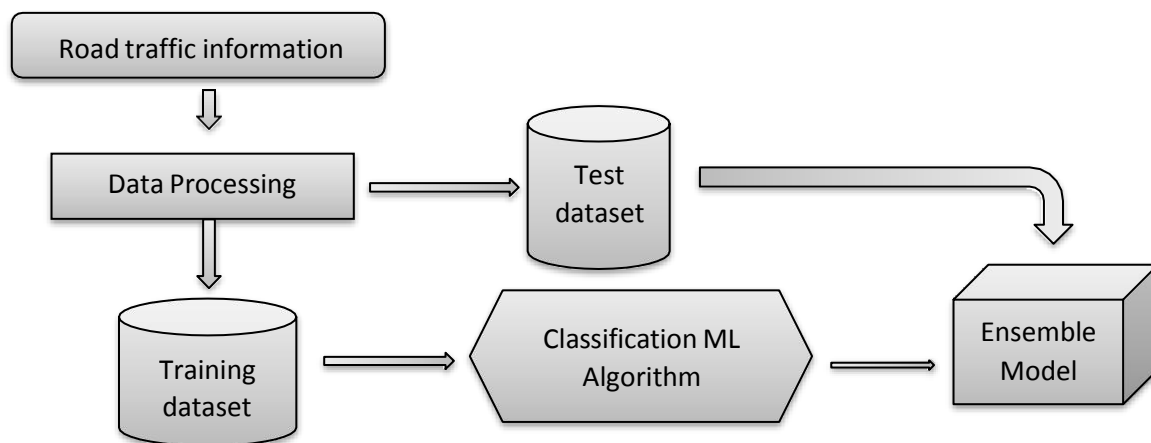


Fig 1.3: Ensemble model

AIM:

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 kaggle 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. The classification and regression model of road accident analysis can be done by a single classifier. Either KNN or SVM is alone to classify and predict the target variable. But we are using both of them together in our approach to make it as ensemble method. Classification problems using ensemble learning model can give better accuracy than others.

SCOPE:

The scope of this project is to investigate a dataset of transport department records using ensemble learning technique and to identifying best road selection by accident reports.

OBJECTIVES:

The goal is to develop a machine learning model for real-time accident forecasting by predicting results in the form of best accuracy by comparing supervised algorithm with mean value of voting classifier results.

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 for them.
- To make predictions based on constraints like weather climate, road structure, etc.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE SURVEY:

A Data Mining Approach for Analyzing Road Traffic Accidents (2019):

- It implemented a data mining framework to identify, analyse and determine attributes contributing to road accidents. The main aim of this research project is to implement a data mining framework for analysing the relationship between accident attributes and make recommendations for preventing the high occurrence of these accidents.
- It evaluated with road accidents data from Khomas region, Namibia. The results demonstrate that the use of such an analytical tool can help in creating a knowledge base. The results find out that male drivers have massively contributed to the higher risk of accidents, especially, at intersections and during daylight.
- It also observed that young drivers are often involved in road traffic accidents happening in clear areas. Proportionally, old aged drivers are most likely to be involved in fatal accidents than in non-fatal accidents. Even though, descriptive data mining methods have clearly been able to uncover reasonable information from Khomas Region accident data sets. However, the results remain at a very general level and do not yet provide a detailed knowledge for traffic accident experts. A detailed data about accidents must be available for the findings to provide accurate and original facts about the accident occurred in any region. Data mining produces very understandable and useful results, but it can only be more effective if all needed details of data are available.
- The data mining approach is more valuable when more attributes and the needed information about the accident are available. Information such as, seasonal speed limits, no passing zones, uphill/downhill degrees, curve radius, graveling, salting, vehicle defects, protective devices, status/type of driving license, the number of years with license, apparent suicide cases, and sleepiness were missing in the data.

- It is important that accident data should contain all needed data or information in order to perfect the modeling of the framework. However, roads must have all needed traffic signs and regulation to prevent road traffic accidents, but they may not be effective enough to reduce accident if knowledge about the cause of the accident is not yet discovered. Therefore, the use of data mining is very important to discover this knowledge.

Forecasting of Road Accident in Kerala: A Case Study (2018)

- Traffic accidents are the foremost reason for death and injuries around the world, fatalities are still on the ascent in many creating nations including India. Data Analysis of road accidents makes a strong impact for taking preventive measure to overcome the mishap. It addressed the prediction problem of road accidents using time series analysis across all districts of Kerala. Time series analysis is useful in discovering the trends in road accidents which enables the prediction of future patterns. In the present MS, it used the time series road accidents data in Kerala, India for the period January 1999 – December 2016 to understand the patterns in the data and to develop appropriate model to predict about future patterns which may enable authorities to take preventive steps.
- It subsisted the data till 2013 December as training data for the model selection and rest of the data is used for model valuation. Two models discussed here are the “Holt-Winters (HW) exponential smoothing” and “Seasonal ARIMA (SARIMA)”. Both the models will provide the forecast values within the confidence interval of the test data. It analyzed the road accident data for total number of accidents in Kerala, number of death cases due to motor accidents and the number of injuries.
- All these subsets of data shows non stationary behavior which was confirmed by ADF test. Then this subset of data became stationary after differencing once. It employed Box- Jenkins and Holt Winter exponential smoothing on the data it got after differencing once. Even though, descriptive data mining methods have clearly been able to uncover reasonable information from Khoums Region accident data sets. However, the results remain at a very general level and do not yet provide a detailed knowledge for traffic accident experts.

Predicting the Number of People for Road Traffic Accident on Highways by Hour of Day (2019)

- The process method type (i) is appropriate to predict number of deaths and injuries female and type (ii) is appropriate to predict number of deaths and injuries male. The predict model and method types can be used to estimate the number of male and female in future which are used as the data to generate the efficient traffic incident management system during any period and to support the developing policy for road accidents reduction in the area of Thailand.
- The volume of vehicles has become crucial factor for the number of road accidents then; this paper is predicting the number of people for traffic accident by hour of day regarding deaths and injuries on highways in Thailand. Time series predicting methods: Multiplicative Decomposition, Multiplicative Hotlines, and Seasonal Autoregressive Moving Integrated Moving Average (SARIMA) model are applied to predict the number of people deaths and injuries. The Predicting error is measured by using the mean absolute percentage error (MAPE).

Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity (2018)

- According to World Health Organization report the number of deaths by road traffic accident is more than 1.25 million people and every year with non-fatal accidents affecting more than 20-50 million people. Several factors are contributed on the occurrence of road traffic accident. In this study, data mining classification techniques applied to establish models (classifiers) to identify accident factors and to predict traffic accident severity using previously recorded traffic data.
- Using WEKA (Waikato Environment for Knowledge Analysis) data mining decision tree (J48, ID3 and CART) and Naïve Bayes classifiers are built to model the severity of injury. The classification performance of all these algorithms is compared based on their results. The experimental result shows that the accuracy of J48 classifier is higher than others. Four data mining classification techniques (J48, ID3, CART and Naïve Bayes) are employed to find out determinant road accident factors.

- The accuracy of J48 classifier is higher than other classifiers, but naïve bayes classification performance is better than the others even if its accuracy is less than J48 and CART classifiers as it is shown in AUC and ROC results. The result of the analysis showed that speed limit, weather condition, number of lane, lighting condition, and accident time are most important and most influential road accident factors. In the other hand, gender, age, accident area, and vehicle type are factors that have less influence on road accident severity.
- Most fatal accident severities are occurred in the raining weather conditions that drive at the mid night and also serious accidents are occurred in the foggy weather condition in one lane roads.

A Survey on Analyses of Factors Related to Road Accidents Using Data Mining Techniques (2018)

- These Analyses made with respect to factors influencing road accidents. Road accidents cause major damages to humans. It can cause lifetime injuries. There is a spike in the number of accidents over the recent years. So it is an important concern for departments that deal with road safety and for the citizens. Road traffic data must be critically analyzed to determine the factors that are closely related to road accidents. Factors like collision type, road condition, light effect, weather and drunk driver must be considered. Places near residential areas, zebra crossing, school regions are the major zone of accidents. This paper makes comparative study on the analyses done so far in this regard. It proposes an effective means for analyzing the road accident dataset.
- Existing work mainly concentrates on only analyzing and finding correlation between the parameters influencing road accidents using the dataset for existing roads. It does not consider the newly planned roads and roads under construction. This paper aims at finding accident prone locations for roads that are planned for future and roads that are under construction.

CHAPTER 3

METHODOLOGY

METHODOLOGIES:

Road and traffic accidents are unsure and undeterminable incidents and their valuation calls for the expertise of the factors affecting them. Road and traffic accidents are defined by a set of variables which can be usually of discrete nature. The essential difficulty in the analysis of coincidence records is its heterogeneous nature.

Classification is machine learning technique that can be used as an initial task to obtain various goals and the classification categorize the accident data into different categories. Some of the commonly used Ensemble techniques are discussed below.

Sequential Ensemble learning (Boosting): (Example: Adaboost, Stochastic Gradient Boosting)

Boosting is a machine learning ensemble meta-algorithm for principally reducing bias, and furthermore variance in supervised learning, and a group of machine learning algorithms that convert weak learner to strong ones. Boosting is a meta-algorithm which can be viewed as a model averaging method. It is the most widely used ensemble method and one of the most powerful learning ideas. This method was originally designed for classification but it can also be profitably extended to regression. The original boosting algorithm combined three weak learners to generate a strong learner and sequential ensemble methods where the base learners are generated sequentially.

Parallel Ensemble Learning (Bagging): (Example: Random Forest, Bagged Decision Trees, Extra Trees)

Bagging is a machine learning ensemble meta-algorithm intended to improve the strength and accuracy of machine learning algorithms used in classification and regression purpose. It additionally diminishes fluctuation of data (variance) and help to from over-fitting. Bagging or Bootstrap Aggregation is a powerful, effective and simple ensemble method. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement and it can be used with any type of model for classification or regression. Bagging is only effective when using unstable non-linear models and parallel ensemble methods where the base learners are generated in parallel

Stacking & Blending: (Example: Voting Classifier)

- Stacking is concerned with combining multiple classifiers generated by using different learning algorithms on a single dataset which consists of pairs of feature vectors and their classifications. This technique consists of basically two phases, in the first phase, a set of base-level classifiers is generated and in the second phase, a meta-level classifier is learned which combines the outputs of the base-level classifiers.
- Stacking is a way of combining multiple models that introduces the concept of a meta learner. It is less widely used than bagging and boosting. Unlike bagging and boosting, stacking may be (and normally is) used to combine models of different types. The procedure is as follows:
 - ✓ Split the training set into two disjoint sets.
 - ✓ Train several base learners on the first part.
 - ✓ Test the base learners on the second part.
 - ✓ Using the predictions from the inputs, and the correct responses as the outputs, train a higher level learner.

Preparing the Dataset (Example given dataset- by your convenient)

The dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new data details filled at the time of application form acts as a test data set.

Table shows details of the datasets:

Table 2.1: Variables in the Dataset

Variable	Description
accident_occurs_Date	Date of accident occurs
Light_Cond	Road light condition
weathconds	Place of weather conditions
mod	Vehicle driver or pedestrian
cctv_footage	CCTV Footage videos is on or off condition
Vehicle_type	Type of vehicle
Route1,2,3	Multi traffic routes
locations	Place of accident

FEASIBILITY STUDY

Data Wrangling:

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions. The given attributes can be used to rebuild the whole process of the accident in a relatively full and objective manner. It provides more than sufficient information and references for road traffic accident analyses. The whole dataset is divided into two parts training and testing dataset. Training data set is used to train the model and to formulate the rules which are used to make decisions. The test dataset is used to validate the rules obtained from trained classifier using new data.

Data collection:

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms, K-Nearest Neighbor and Support vector classifier are applied on the Training set and based on the test result accuracy, Test set prediction is done. Data Collection Road traffic accident is under persuading of many factors, which make it a complicate and as far as information is concerned, there are different databases of traffic accident in different countries.

Preprocessing:

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done.

CONSTRUCTION OF A PREDICTIVE MODEL:

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to preprocess then, what kind of algorithm with model.

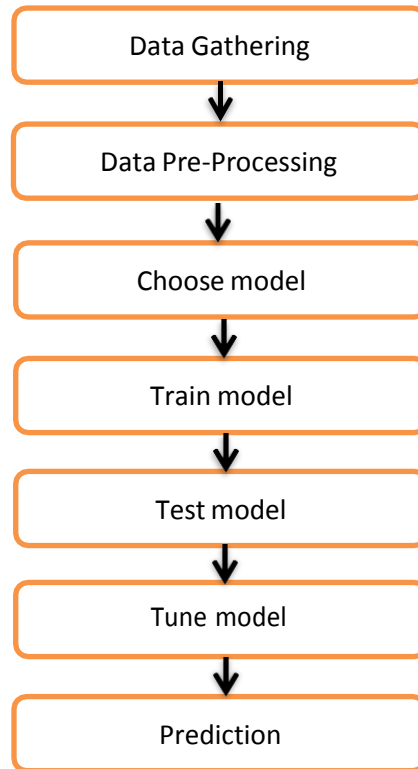


Fig 3.1: Process of dataflow diagram

Dataflow diagram for machine learning:

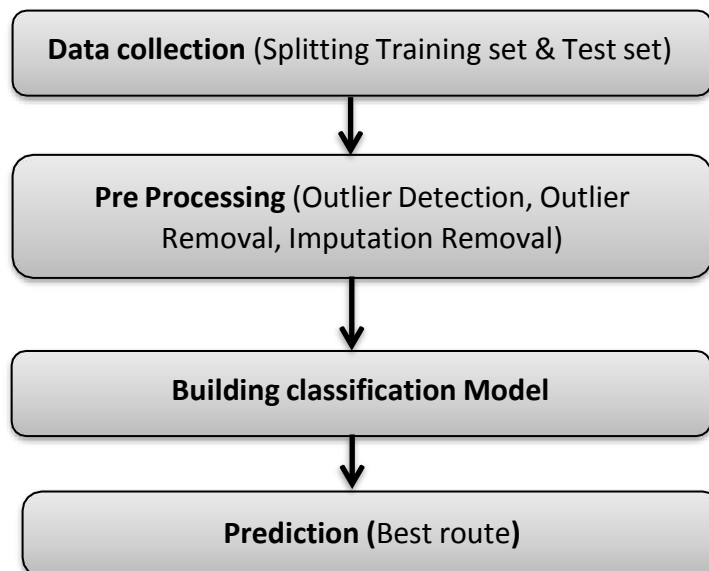


Fig 3.2: data flow diagram for Machine learning model

Work flow diagram:

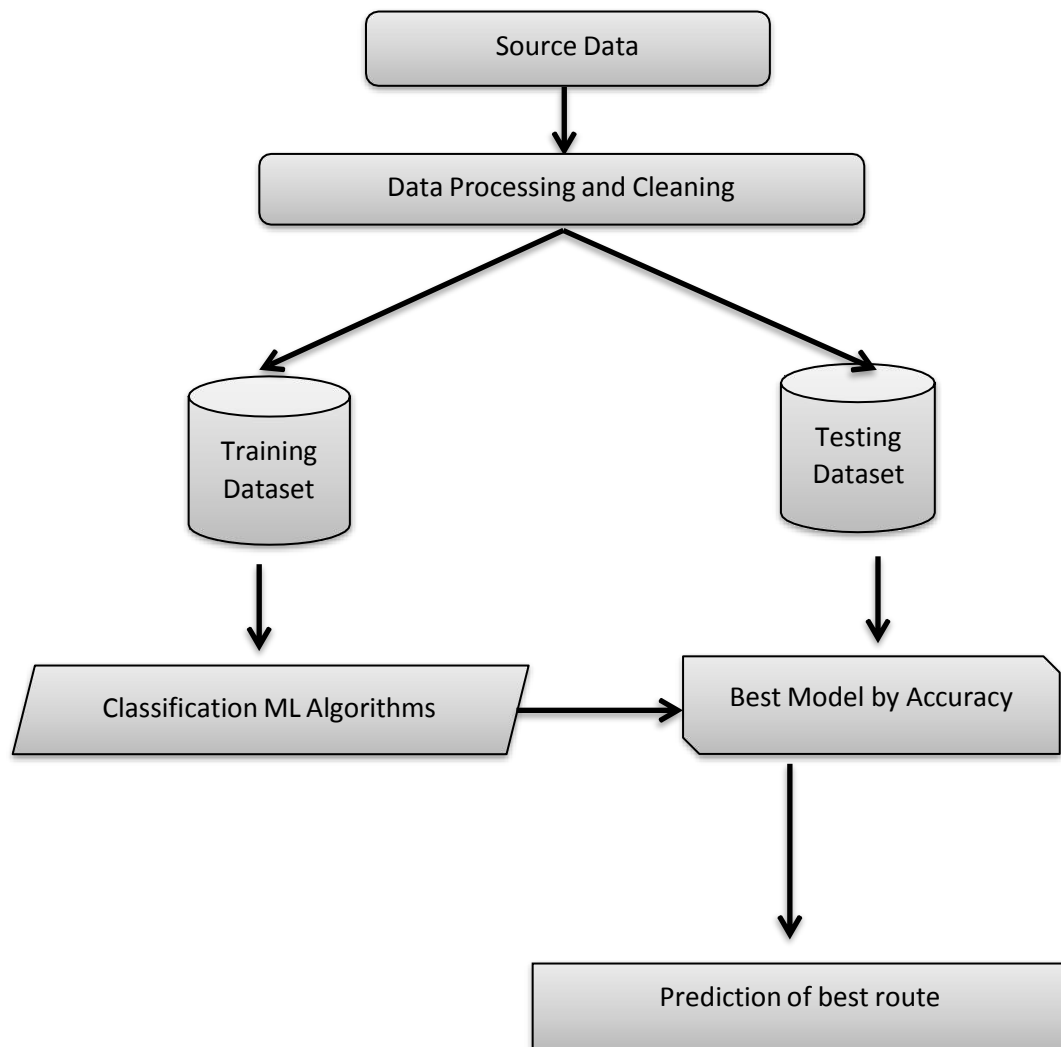


Fig 3.3: Workflow Diagram

UML Diagram:

Use Case Diagram:

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner.

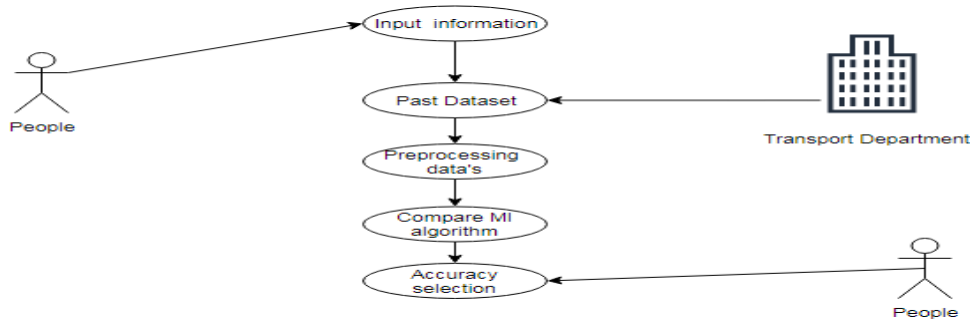


Fig 3.4: Use Case Diagram

Activity Diagram:

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single

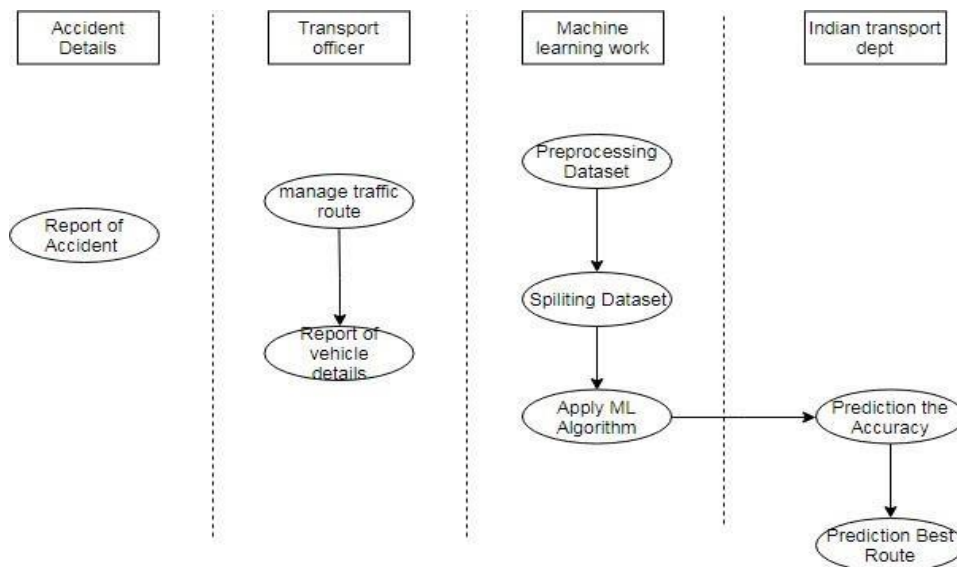


Fig 3.5: Activity Diagram

Sequence Diagram:

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system. Other dynamic modeling techniques include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

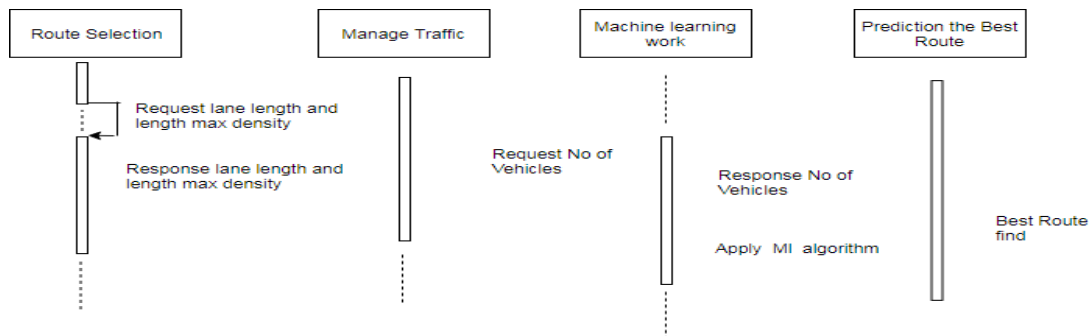


Fig 3.6: Sequence Diagram

PROJECT REQUIREMENTS:

General:

Requirements are the basic constraints that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements
2. Non-Functional requirements
3. Environment requirements
 - A. Hardware requirements
 - B. software requirements

Functional requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and sea born.

Non-Functional Requirements:

Process of functional steps:

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

ENVIRONMENTAL REQUIREMENTS:

1. Software Requirements:

Operating System	: Windows
Tool	: Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor	: Pentium IV/III
Hard disk	: minimum 80 GB
RAM	: minimum 2 GB

Frontend: GUI using python (Tkinter)

Backend: Data analysis using python (Pandas, Numpy etc.)

Library Packages:

- Pandas
- Numpy
- Sk-learn

- Matplotlib
- Seaborn
- Tkinter

Software Description:

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system “Conda”. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the Condo package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the conda install command or using the pip install command that is installed with Anaconda. Pip packages provide many of the features of conda packages and in most cases they can work together. Custom packages can be made using the conda build command, and can be shared with others by uploading them to Anaconda Cloud, PyPI or other repositories. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, you can create new environments that include any version of Python packaged with conda.

Anaconda Navigator:

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda 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:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

Conda:

Conda is an open source, cross-platform, language-agnostic package manager and environment management system that installs, runs and updates packages and their dependencies. It was created for Python programs, but it can package and distribute software for any language (e.g., R), including multi-languages. The Conda package and environment manager is included in all versions of Anaconda, Miniconda, and Anaconda Repository.

The Jupyter Notebook:

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. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Notebook document:

Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc...). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

Jupyter Notebook App:

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet. In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing opening notebook documents or shutting down their kernels.

Kernel:

A notebook kernel is a “computational engine” that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels). When you open a Notebook document, the associated kernel is automatically launched. When the notebook is executed (either cell-by-cell or with menu Cell -> Run All), the kernel performs the computation and produces the results. Depending on the type of computations, the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel is shut-down.

Notebook Dashboard:

The Notebook Dashboard is the component which is shown first when you launch Jupyter Notebook App. The Notebook Dashboard is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown). The Notebook Dashboard has other features similar to a file manager, namely navigating folders and renaming/deleting files.

Working Process:

- Download and install anaconda and get the most useful package for machine learning in Python.
- Load a dataset and understand its structure using statistical summaries and data visualization. Machine learning models, pick the best and build confidence that the accuracy is reliable.

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird's eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely, from loading data, summarizing data, evaluating algorithms and making some predictions.

Here is an overview of what we are going to cover:

1. Installing the Python anaconda platform.
2. Loading the dataset.
3. Summarizing the dataset.
4. Visualizing the dataset.
5. Evaluating some algorithms.
6. Making some predictions.

Frontend: GUI using python (Tkinter)

Backend: Data analysis using python (Pandas, Numpy etc.)

Modules:

- Data validation and pre-processing technique (Module-01)
- Exploration data analysis of visualization and training a model by given attributes (Module-02)
- Performance measurements of logistic regression and decision tree algorithms (Module-03)

- Performance measurements of Support vector classifier and Random forest (Module-04)
- Performance measurements of KNN and Naive Bayes (Module-05)
- Calculating the votes by ensemble learning method (Module-06)
- GUI based prediction of road accidents (Module-07)

MODULE-01:

Data validation process:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset)

Data Validation:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values,

duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Data Pre-processing:

	Reference_id	Grid Ref_Easting	Grid Ref_Northing	Number_of_Vehicles	accident_occurs_Date	Timing	1st_Road_Class	Light_Cond	cctv_footag	weathconds	...	mod	seviars
0	1112091	429175	431904	2	1/1/2013	6:40 PM	Unclassified	Darkness: street lights present and lit	nil	Raining without high winds	...	Driver	Slight
1	1180869	430429	431025	3	1/2/2013	2:30 PM	Motorway	Daylight: street lights present	on	Fine without high winds	...	Driver	Slight
2	1180869	430429	431025	3	1/3/2013	2:30 PM	Motorway	Daylight: street lights present	on	Fine without high winds	...	Driver	Slight
3	11A0238	424660	427582	3	1/4/2013	8:17 AM	Motorway	Daylight: street lights present	on	Fine without high winds	...	Driver	Slight
4	11A0238	424660	427582	3	1/5/2013	8:17 AM	Motorway	Daylight: street lights present	on	Fine without high winds	...	Driver	Slight

Fig 3.7: Pre-processing Data

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing 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. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in

such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

MODULE - 02:

Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

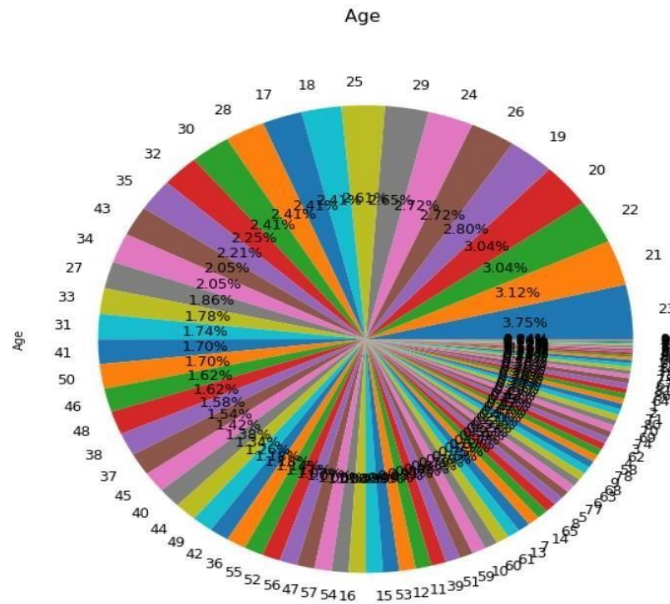


Fig 3.8: Analysis of data

Data Visualization After the classification and regression process the predicted results are visualized in graphical or tabular format for better understanding of the

users. This process is called as Data Visualization. We can also get the summary of the results in numerical format. Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.
- How to summarize the relationship between variables with scatter plots.

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results. Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security. It couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Advantages of train/test split:

1. This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
2. Simpler to examine the detailed results of the testing process.

Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More “efficient” use of data as every observation is used for both training and testing.

Training the Dataset:

- The first line imports iris data set which is already predefined in sklearn module and raw data set is basically a table which contains information about various varieties.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- To encapsulate load_data method in data_dataset variable. Further divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.
- This method divides dataset into training and test data randomly in ratio of 67:33 / 70:30. Then we encapsulate any algorithm.
- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

Testing the Dataset:

- Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output.
- So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

MODULE - 03:

Logistic Regression:

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts $P(Y=1)$ as a function of X .

```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

     0       0.99         1.00         1.00         431
     1       1.00         0.98         0.99         207
     2       1.00         0.99         1.00         122

 accuracy          0.99         760
 macro avg          1.00         0.99         0.99         760
 weighted avg       0.99         0.99         0.99         760

Accuracy result of Logistic Regression is: 99.3421052631579

Confusion Matrix result of Logistic Regression is:
[[431  0  0]
 [ 4 203  0]
 [ 0  1 121]]

Sensitivity : 1.0
Specificity : 0.9806763285024155
```

Fig 3.9: Result of Logistic Regression

Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.

- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

Decision Tree:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

```

Classification report of Decision Tree Classifier Results:

              precision    recall  f1-score   support

     0               1.00      1.00      1.00       431
     1               1.00      1.00      1.00       207
     2               1.00      1.00      1.00       122

 accuracy               1.00
 macro avg              1.00
weighted avg              1.00

Accuracy result of Decision Tree Classifier is 100.0

Confusion Matrix result of Decision Tree Classifier is:
[[431  0  0]
 [ 0 207  0]
 [ 0  0 122]]

Sensitivity :  1.0
Specificity :  1.0

```

Fig 3.10: Result of Decision Tree Classifier

Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an

associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

MODULE -04:

Support Vector Machines (SVM):

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

The following are the basic steps involved in performing the Support Vector Machines:

- How to learn an SVM model from training data. How to best prepare your data for the SVM algorithm. Where you might look to get more information on SVM.
- How to disentangle the many names used to refer to support vector machines. The representation used by SVM when the model is actually stored on disk.

How a learned SVM model representation can be used to make predictions for new data.

```
Classification report of Support Vector Machines Results:

              precision    recall  f1-score   support

     0           0.57         1.00         0.72         431
     1           0.00         0.00         0.00         207
     2           0.00         0.00         0.00         122

 accuracy          0.19
 macro avg          0.19         0.33         0.24         760
weighted avg          0.32         0.57         0.41         760

Accuracy result of Support Vector Machines is: 56.71052631578948

Confusion Matrix result of Support Vector Machines is:
[[431  0  0]
 [207  0  0]
 [122  0  0]]

Sensitivity : 1.0
Specificity : 0.0
```

Fig 3.11: Result of Support Vector Machine

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset. Build a decision tree based on these N records.

- Choose the number of trees you want in your algorithm and repeat steps 1 and 2. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).
- The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

```

Classification report of Random Forest Results:

              precision    recall  f1-score   support

     0               1.00      1.00      1.00        431
     1               1.00      1.00      1.00        207
     2               1.00      0.99      1.00        122

 accuracy               1.00
 macro avg              1.00      1.00      1.00
weighted avg              1.00      1.00      1.00

Accuracy result of Random Forest is: 99.86842105263159

Confusion Matrix result of Random Forest is:
[[431  0  0]
 [ 0 207  0]
 [ 0  1 121]]

Sensitivity : 1.0
Specificity : 1.0

```

Fig 3.12: Result of Random Forest Algorithm

MODULE -05:

K-Nearest Neighbor (KNN):

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors.

Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors. The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it

takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the “k” is the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances. “Closeness” is determined using a proximity measurement (Euclidean) across all features.

```

Classification report of K-Nearest Neighbor Results:
              precision    recall  f1-score   support

     0       0.57         0.72         0.64         431
     1       0.28         0.24         0.26         207
     2       0.38         0.11         0.18         122

 accuracy          0.41
 macro avg          0.36
 weighted avg       0.46

Accuracy result of K-Nearest Neighbor is: 49.21052631578947

Confusion Matrix result of K-Nearest Neighbor is:
[[310 103  18]
 [152  50   5]
 [ 82  26  14]]

Sensitivity :  0.7506053268765133
Specificity :  0.24752475247524752

```

Fig 3.13: Result of K-Nearest Neighbor

Naive Bayes algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method. The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability. Naive Bayes is a

statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

```

Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

     0               1.00      1.00      1.00        431
     1               1.00      1.00      1.00        207
     2               1.00      1.00      1.00        122

 accuracy               1.00              760
 macro avg              1.00              760
weighted avg              1.00              760

Accuracy result of Naive Bayes is: 100.0

Confusion Matrix result of Naive Bayes is:
[[431  0  0]
 [ 0 207  0]
 [ 0  0 122]]

Sensitivity :  1.0
Specificity :  1.0

```

Fig 3.14: Result of Naive Bayes

MODULE - 06:

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants. There will be two kinds of pages like registration user purpose and login entry purpose of users.

MODULE - 07:

Accuracy calculation:

False Positives (FP):

A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN):

A person who default predicted as payer. When actual class is yes but predicted class is no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP):

A person who will not pay predicted as defaulter. These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. If actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN):

A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

It achieved precision, recall, true positive rate (TPR), and false positive rate (FPR) for each classification techniques as it is shown in the above tables and also achieved different interesting confusion matrix for each classification techniques and we can see the classification performance of each classifiers by the help of confusion matrix. We use a confusion matrix to compute the accuracy rate of each severity class. For each class, it demonstrates how instances from that class receive the various

classifications. Here in the next table we have shown instances that are correctly classified and incorrectly classified in accordance with overall accuracy of each classification techniques. All classifiers perform similarly well with respect to the number of correctly classified instances.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

Fig 3.15: Confusion matrix

Comparing Algorithm with prediction result:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization

methods to show the average accuracy, variance and other properties of the distribution of model accuracies. In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness. In the example below 5 different algorithms are compared:

- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Decision tree
- Support Vector Machines
- Naive Bayes

Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output. So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

Advantages:

- These reports are to the investigation of applicability of AI application for traffic accidental forecasting in operational conditions.
- It highlights some observations on future research issues, challenges, and needs.
- It also used to prevent accidents by creating awareness to the civilians.
- Road accident prediction models are used for a variety of purposes: most frequently to estimate the expected accident frequency for various roadway entities (highways, intersections, interstates etc.)

Applications of Ensemble Methods:

- Ensemble methods can be used as overall diagnostic procedures for a more conventional model building. The larger the difference in fit quality between one of the stronger ensemble methods and a conventional statistical model, the more information that the conventional model is probably missing.
- Ensemble methods can be used to evaluate the relationships between explanatory variables and the response in conventional statistical models. Predictors or basis functions overlooked in a conventional model may surface with an ensemble approach.
- With the help of the ensemble method, the selection process could be better captured and the probability of membership in each treatment group estimated with less bias.
- One could use ensemble methods to implement the covariance adjustments inherent in multiple regression and related procedures. One would “residualized” the response and the predictors of interest with ensemble methods.

CHAPTER 4

RESULTS AND DISCUSSION

RESULTS:

The accuracy of each of the models can be calculated individually by using the values of TP, TN, FP and FN from the confusion matrix, which is obtained separately for both the algorithms. The accuracy of Naïve Bayes classifier is found to be more than that of Decision Tree algorithm, and hence, Naïve Bayes algorithm can act as an optimal algorithm for this accident criticality prediction system which serves the purpose of predicting the criticality of an accident with the given attributes. It proves to be optimal when lesser attributes are taken into consideration.

DISCUSSION:

From the above, we can find that the algorithms have the best predict result from the accuracy aspect, the Bagging decision tree has the best predict result from the values, the random forest has the best predict result from the integrated effect. The experiment also reflects the characteristics of the algorithms. Therefore, we can use the Logistic Regression as the model, with the least error value when we the accuracy is more important; we can use the bagging decision algorithm when the appropriate degree is more important; we can use the random forests as the mode when the integrated effect is more important. The appropriate model should be determined by the different usage. Of course, this paper only uses three methods; we don't rule out the existence of other better models. For the current application model, Logistic Regression is the most suitable for traffic accident data prediction.

CHAPTER 5

CONCLUSION AND FUTURE WORK

CONCLUSION

A more flexible traffic signal has been updated for a mixed type of traffic. Mixed traffic consists of both manual and automated vehicles. For AVs, an efficient speed changing procedure is advanced responsively. The proposed system maintains the present traffic to be smooth and content with crossing point of genuinely decided vehicles. A number of parameters are considered in the system. These parameters are the main source on determining whether that traffic is an accident prone region or non-accident prone region. The user interface has been well developed where the user has the facility to choose the algorithm he wants to work with. However, the system is capable of working on all algorithms, a few being K-nearest, Random forest, Logistic Regression, and chooses the one that produces the highest accuracy. This facilitates the user to receive a highly reliable response. Thus the system turns out to be more trustworthy. Based on the result the user can avoid going to the accident prone region. This system works well for even normal people in their day to day routine. It can be developed to a smaller scale for the user's easy accessibility.

FUTUTRE WORK

To optimizes the work to implement in Artificial Intelligence environment. Transport Department needs to automate the recognizing the best course by not disaster from capability process (consistent) in perspective on the record detail. To automate this strategy by show the desire achieves web application or work territory application.

REFERENCES

- [1] K. G. Boldness and S. M. Parapar, "Postponement and fuel utilization at traffic signals," *Inst. Traffic Eng.*, vol. 45, no. 11, pp. 23–27, 1975.
- [2] T.- Q. Tang, Z.- Y. Yi, and Q.- F. Lin, "Impacts of sign light on the fuel utilization and emanations under vehicle following model," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 200–205, Mar. 2017.
- [3] N. H. Gartner, "Advancement and testing of an interest responsive system for traffic signal control" in *Proc. Amer. Control Conf.*, Jun. 1982, pp. 578–583.
- [4] N. H. Gartner, F. J. Pooran, and C. M. Andrews, "Usage of the OPAC versatile control methodology in a rush hour gridlock signal system," in *Proc. Intell. Transp. Syst.*, Aug. 2001, pp. 195–200.
- [5] A. G. Sims and K. W. Dobinson, "The Sydney facilitated versatile traffic (SCAT) framework reasoning and advantages," *IEEE Trans. Veh. Technol.*, vol. VT-29, no. 2, pp. 130–137, May 1980.
- [6] D. I. Robertson and R. D. Bretherton, "Upgrading systems of traffic flags progressively the SCOOT strategy," *IEEE Trans. Veh. Technol.*, vol. 40, no. 1, pp. 11–15, Feb. 1991.
- [7] S. Sen and K. L. Head, "Controlled streamlining of stages at a crossing point," *Transp. Sci.*, vol. 31, no. 1, pp. 5–17, 1997.
- [8] B. Asadi and A. Vahidi, "Prescient journey control: Utilizing up and coming traffic signal data for improving mileage and diminishing excursion time," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 3, pp. 707–714, May 2011.
- [9] M. A. S. Kamal, S. Taguchi, and T. Yoshimura, "Convergence vehicle helpful eco-driving with regards to somewhat associated vehicle condition," in *Proc. Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 1261–1266.
- [10] S. I. Guler, M. Menendez, and L. Meier, "Utilizing associated vehicle innovation to improve the proficiency of crossing points," *Transp. Res. C, Emerg. Technol.*, vol. 46, pp. 121–131, Sep. 2014.
- [11] J. Rios-Torres and A. A. Malikopoulos, "A study on the coordination of associated and mechanized vehicles at crossing points and converging at thruway entrance ramps," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1066–1077, May 2017.

APPENDICES

A. SAMPLE CODE:

```
"from tkinter import *\n",\n"import numpy as np\n",\n"import pandas as pd\n",\n"df = pd.read_csv(\"df.csv\")\n",\n"del df[\"Reference_id\"]\n",\n"del df[\"Grid Ref_ Easting\"]\n",\n"del df[\"Grid Ref_ Northing\"]\n",\n"del df[\"accident_occurs_Date\"]\n",\n"del df[\"Timing\"]\n",\n"del df[\"ratings\"]\n",\n"del df[\"1st_Road_Class\"]\n",\n"del df[\"mod\"]\n",\n"del df[\"Age\"]\n",\n"del df[\"seviors\"]\n",\n"del df[\"classes\"]\n",\n"data": {\n",\n"  \"text/html\": [\"<div>\n",\n"    \"<style scoped>\n",\n"      .dataframe tbody tr th:only-of-type {\n",\n"        vertical-align: middle;\n",\n"      }\n",\n"    \"\n",\n"    .dataframe tbody tr th {\n",\n"      vertical-align: top;\n",\n"    }\n",\n"    \"\n",\n"    .dataframe thead th {\n",\n"      text-align: right;\n",\n"    }\n",\n"    \"\n",\n"  ]\n",\n"}
```

```

"/style>\n",
"<table border=\"1\" class=\"dataframe\">\n",
"  <thead>\n",
"    <tr style=\"text-align: right;\">\n",
"      <th></th>\n",
"      <th>Number_ of_ Vehicles</th>\n",
"      <th>Light_ Cond</th>\n",
"      <th>cctv_ footag</th>\n",
"      <th>weathconds</th>\n",
"      <th>gener</th>\n",
"      <th>Vehicle_ type</th>\n",
"      <th>Route1</th>\n",
"      <th>Route2</th>\n",
"      <th>Route3</th>\n",
"      <th>locations</th>\n",
"    </tr>\n",
"  </thead>\n",
"  <tbody>\n",
"    <tr>\n",
"      <td>0</td>\n",
"      <td>2</td>\n",
"      <td>Darkness: street lights present and lit</td>\n",
"      <td>nil</td>\n",
"      <td>Raining without high winds</td>\n",
"      <td>Male</td>\n",
"      <td>Car</td>\n",
"      <td>accident occurs</td>\n",
"      <td>normal</td>\n",
"      <td>traffic</td>\n",
"      <td>FlowerBazaar</td>\n",
"    </tr>\n",

```

```

" <tr>\n",
"   <td>1</td>\n",
"   <td>3</td>\n",
"   <td>Daylight: street lights present</td>\n",
"   <td>on</td>\n",
"   <td>Fine without high winds</td>\n",
"   <td>Male</td>\n",
"   <td>Car</td>\n",
"   <td>checkpoint</td>\n",
"   <td>accident occurs</td>\n",
"   <td>normal</td>\n",
"   <td>HighCourt</td>\n",
" </tr>\n",
" <tr>\n",
"   <td>2</td>\n",
"   <td>3</td>\n",
"   <td>Daylight: street lights present</td>\n",
"   <td>on</td>\n",
"   <td>Fine without high winds</td>\n",
"   <td>Female</td>\n",
"   <td>Car</td>\n",
"   <td>normal</td>\n",
"   <td>barricade</td>\n",
"   <td>accident occurs</td>\n",
"   <td>Harbour</td>\n",
" </tr>\n",
" <tr>\n",
"   <td>3</td>\n",
"   <td>3</td>\n",
"   <td>Daylight: street lights present</td>\n",
"   <td>on</td>\n",

```

```

"    <td>Fine without high winds</td>\n",
"    <td>Male</td>\n",
"    <td>Car</td>\n",
"    <td>accident occurs</td>\n",
"    <td>normal</td>\n",
"    <td>railway crossing</td>\n",
"    <td>PortMarine</td>\n",
"  </tr>\n",
" <tr>\n",
"   <td>4</td>\n",
"   <td>3</td>\n",
"   <td>Daylight: street lights present</td>\n",
"   <td>on</td>\n",
"   <td>Fine without high winds</td>\n",
"   <td>Male</td>\n",
"   <td>Car</td>\n",
"   <td>No Entry</td>\n",
"   <td>accident occurs</td>\n",
"   <td>normal</td>\n",
"   <td>Washermenpet</td>\n",
" </tr>\n",
" </tbody>\n",
"</table>\n",
"</div>"]],
"text/plain": [
"  Number_of_Vehicles          Light_Cond cctv_footag \\n",
"0          2  Darkness: street lights present and lit      nil  \n",
"1          3    Daylight: street lights present      on  \n",
"2          3    Daylight: street lights present      on  \n",
"3          3    Daylight: street lights present      on  \n",
"4          3    Daylight: street lights present      on  \n",

```

```

"\n",
"          weathconds  gener Vehicle_type      Route1 \\n",
"0 Raining without high winds  Male      Car accident occurs  \n",
"1  Fine without high winds  Male      Car      checkpoint  \n",
"2  Fine without high winds  Female      Car      normal  \n",
"3  Fine without high winds  Male      Car accident occurs  \n",
"4  Fine without high winds  Male      Car      No Entry \n",
"\n",
"      Route2      Route3  locations \n",
"0      normal      traffic FlowerBazaar \n",
"1 accident occursnormal      HighCourt  \n",
"2  barricade  accident occurs  Harbour \n",
"3      normal railway crossing  PortMarine \n",
"4 accident occursnormal Washermenpet " ]
},
"Source": [
"df.head ()" ],
{
"cell_type": "code",
"execution_count": 5,
"metadata": {},
"outputs": [{
"data": {
"text/plain": [
"Index(['Number_ of_Vehicles', 'Light_Cond', 'cctv_footag', 'weathconds',\n",
"      'gener', 'Vehicle_type', 'Route1', 'Route2', 'Route3', 'locations'],\n",
"      dtype='object')"]
}
}],
"source": [
"df.columns"
] },

```

B. SCREEN SHOTS:

Road Accident Prediction

Road Accident Prediction using Machine Learning

(Demo Transport Department Dataset)

Gender Details	None	Type of Vehicles	None
Light Conditions	None	No. Vehicles	None
CCTV Footage Availability	None	Enter Location	None
Weather Conditions	None		
DecisionTree		DecisionTree Algorithm	
RandomForest		Randomforest Algorithm	
LogisticRegression		LogisticRegression Algorithm	
KNeighborsClassif		KNeighborsClassifier Algorithm	
SVC		SVC Algorithm	
NB		NB Algorithm	

Road Accident Prediction

Road Accident Prediction using Machine Learning

(Demo Transport Department Dataset)

Gender Details	None	Type of Vehicles	None
Light Conditions	None	No. Vehicles	None
CCTV Footage Availability	None	Enter Location	None
Weather Conditions	None		
Ensemble Method		Voting Classifier	

Road Accident Prediction

Road Accident Prediction using Machine Learning

(Demo Transport Department Dataset)

Gender Details: Male

Type of Vehicles: Car

Light Conditions: Daylight_street_lights_present

No. Vehicles: 3

CCTV Footage Availability: CCTV_ON

Enter Location: Adyar

Weather Conditions: WC_Fine_with_high_winds

DecisionTree	No_Accidents	DecisionTree Algorithm
RandomForest	No_Accidents	Randomforest Algorithm
LogisticRegression	No_Accidents	LogisticRegression Algorithm
KNeighborsClassifier	No_Accidents	KNeighborsClassifier Algorithm
SVC	No_Accidents	SVC Algorithm
NB	Traffic	NB Algorithm

Road Accident Prediction

Road Accident Prediction using Machine Learning

(Demo Transport Department Dataset)

Gender Details: Male

Type of Vehicles: Car

Light Conditions: Daylight_street_lights_present

No. Vehicles: 3

CCTV Footage Availability: CCTV_ON

Enter Location: Adyar

Weather Conditions: WC_Fine_without_high_winds

Ensemble Method	No_Accidents	Voting Classifier
-----------------	--------------	-------------------

C.PUBLICATION WITH PLAGIARISM REPORT:

www.ijemr.net	ISSN (Online): 2250-0758 ISSN (Print): 2394-6962
	Publication Impact Factor (PIF) for 2019: 5.230
VANDANA PUBLICATIONS	
International Journal of Engineering and Management Research	
(ISO 9001:2008 Organization)	
Ref No: IJEMR/V-10/I-2/03/2020	(Section – A)
Date: 09-04-2020	
Certificate of Publication	
This is to certify that Research Paper title “Analysis of Machine Learning Algorithm with Road Accidents Data Sets”, by “P Sumanth” has been published with the “International Journal of Engineering and Management Research”, Volume-10, Issue-2 of April 2020.	
 (Editor-in-Chief) International Journal of Engineering and Management Research (IJEMR)	
Head Office: UG-4, Avadh Tower, Kaysons Lane, Hazratganj, Lucknow (U.P)-226001, (INDIA) Phone: 0522-4108552, +91-9696045327, +91-6393024152 +91-9839748474 E-Mail: contact@ijemr.net, editor@ijemr.net, ijemrcontact@gmail.com Website: www.ijemr.net www.vandanapublications.com	



Publication Impact Factor (PIF) for 2019: 5.230

VANDANA PUBLICATIONS

International Journal of Engineering and Management Research

(ISO 9001:2008 Organization)

Ref No: IJEMR/V-10/I-2/03/2020

(Section – A)

Date: 09-04-2020

Certificate of Publication

This is to certify that Research Paper title “**Analysis of Machine Learning Algorithm with Road Accidents Data Sets**”, by “**P Sai Anudeep**” has been published with the “International Journal of Engineering and Management Research”, Volume-10, Issue-2 of April 2020.

(Editor-in-Chief)
International Journal of Engineering and
Management Research (IJEMR)



Head Office: UG-4, Avadh Tower, Kaysons Lane, Hazratganj, Lucknow (U.P)-226001, (INDIA)

Phone: 0522-4108552, +91-9696045327, +91-6393024152 +91-9839748474

E-Mail: contact@ijemr.net, editor@ijemr.net, ijemrcontact@gmail.com

Website: www.ijemr.net | www.vandanapublications.com

Submission date: 12-Mar-2020 12:13PM (UTC-0500)

Submission ID: 1274420692

File name:

OF_MACHINE_LEARNING_ALGORITHM_WITH_ROAD_ACCIDENTS_DATA_SETS.docx
(238.16K)

Word count: 3506

Character count: 18965

ORIGINALITY REPORT

8%

SIMILARITY INDEX

5%

INTERNET SOURCES

7%

PUBLICATIONS

4%

STUDENT PAPERS

ANALYSIS OF MACHINE LEARNING ALGORITHM WITH ROAD ACCIDENTS DATA SETS

P SUMANTH^[1], P SAI ANUDEEP^[2], S DIVYA^[3]

[1][2] UG Student, Dept. of Computer Science, Sathyabama University, Chennai, India.

[3] Assistant professor, Dept. of Computer Science, Sathyabama University, Chennai, India.

ABSTRACT: Beginning at now, street transport framework neglect to alter up to the exponential expansion in vehicular masses and to ascertaining the quickest driving courses and catastrophes inside observing differing traffic conditions is a critical issue right presently structures. To upset this issue is to explore the vehicle division dataset with bundle learning technique for finding the best street choice without calamity gauging by want aftereffects of best accuracy count by looking at oversaw AI figuring. In bits of information and AI, bundle strategies utilize diverse learning calculations to give indications of progress prudent execution. The assessment of dataset by facilitated AI technique (SMLT) to get two or three data takes after, factor perceiving proof, univariate evaluation, bivariate and multi-variate appraisal, missing worth medications and separate the information support, information cleaning/organizing and information perception will be done with everything taken into account given dataset. In addition, to look at and talk about the presentation of different AI figuring estimations from the given vehicle division dataset with assessment of GUI based street fiasco want by given attributes.

KEYWORDS: Dataset, Ensemble method, GUI results.

INTRODUCTION

The proportion of setback data set aside by traffic the administrators workplaces has been creating at an ever-growing rate over the latest couple of years because of different road vehicle crashes. Government substances and some private parts were found social affair setback data at step by step bases. Data from accidents is every now and again among the most significant assets for neighborhood authorities and central governments, as it helps in arranging and utilization of approaches. It can in like manner help policymakers with settling on fitting decisions relating to foundation progress, arranging and social distinctions movement. All things considered, as the extent of this information is making, there is demand and a need of discovering strategies, procedure and contraptions to assessments such colossal volumes of information and discover an answer to the reason behind debacles. Street vehicle crashes are a vital as a rule danger that keeps causing challenges, wounds and fatalities on roadways reliably, accomplishing tremendous episodes both at the financial and social levels. This general issue needs more idea as for reduce the truth and the rehash of mishap event. The past information about past mishaps speaks to an impressive open door for analysts to distinguish the most persuasive factors in such mishaps, which thusly assume a key job in finding suitable answers for alleviate this issue later on. Lately, a few information mining systems have

been successfully used to separate valuable information from enormous informational indexes containing data about car crashes. Street and car crashes are one of the significant reasons for casualty and weakening over the world. Street possibility can be considered as an event in which a conveyance collides with other conveyance, person or other objects. A road risk not only does road hazard cause property damage, it can also lead to partial or complete disability, and can sometimes be fatal to humans. Addition number of street mishaps is definitely not a decent sign for the transport security. The main arrangement requires the investigation of traffic possibility information to recognize various reasons for street mishaps and taking preventive measures. AI is one of the utilizations of Artificial Intelligence procedure where the machine learns the information certainly as opposed to express programming. These days, AI assumes an essential job in our everyday life. It is utilized nearly in each field like vehicle, clinical, banking and so forth in which transport and clinical field has more significance than others as they are identified with lives. In the field of transportation, AI can be applied in numerous segments like traffic stream forecast, mishap expectation, and traveler place proposal and so on. AI is utilized for computerization as well as for wellbeing.

II. RELATED WORK

Many different types of systems have come up earlier to predict traffic on lanes, control accidents on busy roads and to detect the rate of traffic too. All of them work in independent forms. They follow their own individual algorithms and methodologies to solve their complexities. One of them is to detect the traffic on the road by taking input from the cc cameras available. This system uses machine learning algorithms. They use the detecting mechanism that has gained high scope lately. By using almost similar techniques, a detection of the number of people on roads is carried out. Rather than focusing on the vehicles, this methodology focuses on the people rate to determine the accident prone regions. Through machine learning prediction algorithm, huge amount of data about the regions, places, kind of accidents, number of deaths, time of accident etc. are given to the system and trained. Thus, this system attains the capability to determine the places that are more prone to accidents. This is one of the traditional ways that has been worked upon.

III. LITERATURE SURVEY

TITLE: A Data Mining Approach for Analyzing Road Traffic Accidents

AUTHOR: Zhaoqiang Chen ; Qun Chen ; Zhanhuai Li

YEAR: 2019

DESCRIPTION:

It completed a data mining structure to recognize, analyze and choose credits adding to road accidents. The guideline purpose of this investigation adventure is to realize a data burrowing structure for dismembering the association between accident characteristics and make proposition for thwarting the high occasion of these setbacks. It evaluated with road incidents data from Khomas area, Namibia. The results show that the use of such a coherent gadget can help in making a data base. The results find that male drivers have enormously added to the higher risk of setbacks, especially, at unions and during light.

TITLE: Forecasting of Road Accident in Kerala: A Case Study

AUTHOR: Ihab F. Ilyas ; Xu Chu

YEAR: 2018

DESCRIPTION:

Vehicle crashes are the central explanation behind death and wounds far and wide, fatalities are still moving in many making countries including India. Information Analysis of street fiascos has a solid effect for taking preventive measure to beat the trouble. It kept an eye out for the guess issue of street difficulties utilizing time strategy evaluation over all zones of Kerala. Time course of action appraisal is valuable in finding the models in street mishaps which connects with the figure of future models. In the present MS, it utilized the time plan street mishaps information in Kerala, India for the period January 1999 – December 2016 to comprehend the models in the information and to make fitting model to foresee about future models which may empower specialists to make preventive advances.

TITLE: Predicting the Number of People for Road Traffic Accident on Highways by Hour of Day

AUTHOR: Xu Chu ; Ihab F. Ilyas ; Paolo Papotti

YEAR: 2019

DESCRIPTION:

Predicting the accidents through the number of vehicles on roads in an already researched approach. However, in this paper, the authors have determined the newer method. They predict the number of people on the roads and how they influence the rate of accidents. The focus has been shifted from vehicles to people. Another important parameter, the time of the day is also considered important. The integrated moving model, multiplicative decomposition model, and holtwines are used to detect the people passing. The screw up rate in the end helps perform the final prediction.

TITLE: Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity

AUTHOR: Joeri Rammelaere ; Floris Geerts ; Bart Goethals

YEAR: 2018

DESCRIPTION:

As demonstrated by World Health Organization report the amount of passing by road fender bender is more than 1.25 million people and reliably with non-deadly setbacks impacting more than 20-50 million people. A couple of segments are contributed on the occasion of road car crash. At this moment, mining

portrayal strategies applied to develop models (classifiers) to recognize accident factors and to anticipate auto collision earnestness using as of late recorded traffic data. The accuracy of J48 classifier is higher than various classifiers, anyway simple Bayes gathering execution is better than anything the others whether or not its precision is under J48 and CART classifiers as it is showed up in AUC and ROC results. The outcome of the examination showed that speed limit, atmosphere condition, number of way, lighting condition, and accident time are commonly huge and most remarkable road disaster factors. In the other hand, sex, age, incident zone, and vehicle type are factors that have less impact on road disaster reality.

TITLE: A Survey on Analyses of Factors Related to Road Accidents Using Data Mining Techniques

2

AUTHOR: Philip Bohannon ; Wenfei Fan ; Floris Geerts ; Xibei Jia ; Anastasios Kementsietsidis

YEAR: 2018

DESCRIPTION:

These Analyses made concerning factors influencing street mishaps. Street catastrophes cause vital harms to people. It can cause lifetime wounds. There is a spike in the measure of misfortunes over the advancing years. So it is a basic worry for working environments that administer street security and for the occupants. Street traffic information must be on an essential level inquired about to pick the fragments that are enduringly identified with street episodes. Fragments like impact type, street condition, light impact, air and alcoholic driver must be considered. Places close to neighborhoods, zebra crossing, and school locale are the basic zone of debacles. This paper makes near appraisal on the assessments done in that limit with regards to the present minute. It proposes sensible methods for analyzing the street debacle dataset. Existing work basically fixates on essentially isolating and discovering association between the parameters influencing street mishaps utilizing the dataset for existing roads. It doesn't consider the starting late coordinated streets and paths a work in progress. This paper targets finding gawky districts for streets that are made game arrangements for future and paths that are being taken a shot at.

IV. EXISTING SYSTEM

There are so many existing systems to control traffic and prevent accidents using previous data sets. Government still follows basic and old traffic control system. There are chances of happening accidents in such cases. To prevent these types of incidents we are going to propose our system. The existing systems have the timer enabled traffic control system, which may lead to accidents in a certain types of conditions like bad weather conditions that is the case that it cannot be controlled manually. It cannot alert the accident zones based on weather condition, type of vehicle. There are some systems that alert the accident zones manually by passing the sign boards through that route but this is an old approach although it can alert the persons who are travelling in that particular route. Automatic systems are not much available in all the places for the prediction of accident zones.

V. PROPOSED SYSTEM

Social occasion learning improves AI results by consolidating two or three models. This methodology permits the creation of better smart execution stood apart from a solitary model and it is the specialty of joining distinctive strategy of understudies (specific models) together to ad lib on the security and insightful power of the model. In the area of Statistics and Machine Learning, Ensemble learning methods endeavor to make the presentation of the farsighted models better by improving their exactness. Outfit learning is a system utilizing which different AI models, are intentionally attempted to manage a specific issue.

VI. MODULE DESCRIPTION

1. DATA VALIDATION AND PRE-PROCESSING TECHNIQUE (MODULE-01)
2. EXPLORATION DATA ANALYSIS OF VISUALIZATION AND TRAINING A MODEL BY GIVEN ATTRIBUTES (MODULE-02)
3. PERFORMANCE MEASUREMENTS OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS (MODULE-03)

4. PERFORMANCE MEASUREMENTS OF SUPPORT VECTOR CLASSIFIER AND RANDOM FOREST (MODULE-04)
5. PERFORMANCE MEASUREMENTS OF KNN AND NAIVE BAYES (MODULE-05)
6. CALCULATING THE VOTES BY ENSEMBLE LEARNING METHOD (MODULE-06)

DATA VALIDATION AND PRE-PROCESSING TECHNIQUE (MODULE-01)

Machine learning validation technology is applied to get the Machine Learning (ML) model error rate close to the real data set error rate. Data validation is not mostly required for all kinds of data. However, a system that needs to work on real time scenarios, need data validation. Real time scenarios consists of huge amount of data. This type of large volume data holding multiple data types definitely needs to be validated. After data validation, we require to gather information about our data. The process of data collection, and its analysis gives better insights of what exactly the data is. It basically standardizes the data. Finds missing data, handles outliers, classifies and creates labels. It further extracts only the data that is required for our process. The dataset is well balanced and this increases the accuracy rate of our research. Therefore, the entire process of data validation and pre-processing gives us an idea of which algorithm the system must work with.

EXPLORATION DATA ANALYSIS OF VISUALIZATION AND TRAINING A MODEL BY GIVEN ATTRIBUTES (MODULE-02)

Information perception is a significant aptitude in applied measurements and AI. Measurements do in reality center around quantitative depictions and estimations of information. Information representation gives a significant suite of apparatuses for increasing a subjective comprehension. This can be useful when investigating and finding a workable pace dataset and can help with recognizing plans, degenerate data, inconsistencies, and generously more. With a little space data, data portrayals can be used to discuss and show key associations in plots and charts that are more intuitive and accomplices than extents of alliance or centrality. Data portrayal and exploratory data assessment are whole fields themselves and it will recommend an increasingly significant bounce into some the books referenced at the end.

PERFORMANCE MEASUREMENTS OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS (MODULE-03)

It is a statistical method for analyzing a data set in which there are one or more distinct outcome-specific variables. The result is calculated using a dichotomous equation (in which only two possible results exist). The relapse is calculated. Its main target is to make the best model that fits to depict the link between result variable and the illustrative. Calculated relapse is a Machine Learning grouping calculation that is utilized to anticipate the likelihood of an all-outward variable. In different to the calculated relapse, the strategic relapse, the parallel variable is assigned with the value 1 or 0 accordingly. It is one of the most impressive and famous calculation. Choice tree calculation falls under the class of directed learning calculations. It works for both consistent just as clear cut yield factors.

Classification report of Logistic Regression Results:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	431
1	1.00	0.98	0.99	207
2	1.00	0.99	1.00	122
accuracy			0.99	760
macro avg	1.00	0.99	0.99	760
weighted avg	0.99	0.99	0.99	760

Accuracy result of Logistic Regression is: 99.3421052631579

Confusion Matrix result of Logistic Regression is:

```
[[431  0  0]
 [ 4 203  0]
 [ 0  1 121]]
```

Sensitivity : 1.0

Specificity : 0.9806763285024155

Figure 1: Result of Logistic Regression

Classification report of Decision Tree Classifier Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	431
1	1.00	1.00	1.00	207
2	1.00	1.00	1.00	122
accuracy			1.00	760
macro avg	1.00	1.00	1.00	760
weighted avg	1.00	1.00	1.00	760

Accuracy result of Decision Tree Classifier is 100.0

Confusion Matrix result of Decision Tree Classifier is:

```
[[431  0  0]
 [ 0 207  0]
 [ 0  0 122]]
```

Sensitivity : 1.0

Specificity : 1.0

Figure 2: Result of Decision Tree Classifier

PERFORMANCE MEASUREMENTS OF SUPPORT VECTOR CLASSIFIER AND RANDOM FOREST (MODULE-04)

A classifier that sorts the informational index by setting an ideal hyper plane between information. I picked this classifier as it is extraordinarily flexible in the quantity of various kernelling capacities that can be applied and this model can yield a high consistency rate. Bolster Vector Machines are maybe one of the most mainstream and discussed AI calculations. They were incredibly main stream around the time they were created during the 1990s and keep on being the go-to strategy for a high-performing calculation with small tuning.

Classification report of Support Vector Machines Results:

	precision	recall	f1-score	support
0	0.57	1.00	0.72	431
1	0.00	0.00	0.00	207
2	0.00	0.00	0.00	122
accuracy			0.57	760
macro avg	0.19	0.33	0.24	760
weighted avg	0.32	0.57	0.41	760

Accuracy result of Support Vector Machines is: 56.71052631578948

Confusion Matrix result of Support Vector Machines is:

```
[[431  0  0]
 [207  0  0]
 [122  0  0]]
```

Sensitivity : 1.0

Specificity : 0.0

Figure 3: Result of Support Vector Machine

Classification report of Random Forest Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	431
1	1.00	1.00	1.00	207
2	1.00	0.99	1.00	122
accuracy			1.00	760
macro avg	1.00	1.00	1.00	760
weighted avg	1.00	1.00	1.00	760

Accuracy result of Random Forest is: 99.86842105263159

Confusion Matrix result of Random Forest is:

```
[[431  0  0]
 [ 0 207  0]
 [ 0  1 121]]
```

Sensitivity : 1.0

Specificity : 1.0

Figure 4: Result of Random Forest Algorithm

PERFORMANCE MEASUREMENTS OF KNN AND NAIVE BAYES (MODULE-05)

K-Nearest Neighbor is a managed AI calculation which stores all cases compare to preparing information focuses in n-dimensional space. At the point when obscure discrete information is gotten, it breaks down the nearest k number of occasions spared (closest neighbors) and returns the most widely recognized class as the expectation and for genuine esteemed information it restores the mean of k closest neighbors. Out yonder weighted closest neighbor calculation, it loads the commitment of every one of the k neighbors as indicated by their division using the going with question giving progressively essential burden to the closest neighbors. The Naive Bayes count is an instinctual method that uses the probabilities of each acknowledge having a spot for each class to make a figure. It is the overseen learning approach you would consider in case you expected to exhibit a judicious showing issue probabilistically. The probability of a class regard given an estimation of a quality is known as the prohibitive probability. By expanding the prohibitive probabilities together for every quality for a given class regard, we have a probability of a data event having a spot with that class. To make a figure we can register probabilities of the case having a spot with each class and select the class a motivating force with the most raised probability.

Classification report of K-Nearest Neighbor Results:

	precision	recall	f1-score	support
0	0.57	0.72	0.64	431
1	0.28	0.24	0.26	207
2	0.38	0.11	0.18	122
accuracy			0.49	760
macro avg	0.41	0.36	0.36	760
weighted avg	0.46	0.49	0.46	760

Accuracy result of K-Nearest Neighbor is: 49.21052631578947

Confusion Matrix result of K-Nearest Neighbor is:

```
[[310 103 18]
 [152 50 5]
 [ 82 26 14]]
```

Sensitivity : 0.7506053268765133

Specificity : 0.24752475247524752

Figure 5: Result of K-Nearest Neighbor

```

Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

     0         1.00        1.00        1.00         431
     1         1.00        1.00        1.00         207
     2         1.00        1.00        1.00         122

 accuracy          1.00
 macro avg          1.00
 weighted avg       1.00

Accuracy result of Naive Bayes is: 100.0

Confusion Matrix result of Naive Bayes is:
[[431  0  0]
 [  0 207  0]
 [  0  0 122]]

Sensitivity : 1.0
Specificity : 1.0

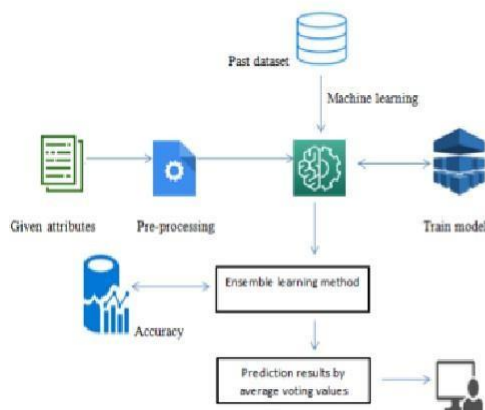
```

Figure 6: Result of Naive Bayes

CALCULATING THE VOTES BY ENSEMBLE LEARNING METHOD (MODULE-06)

Tkinter is a python library for creating GUI (Graphical User Interfaces). We utilize the tkinter library for making a use of UI (User Interface), to make windows and all other graphical UI and Tkinter will accompany Python as a standard bundle, it tends to be utilized for security motivation behind every client or bookkeepers. There will be two sorts of pages like enlistment client reason and login passage motivation behind clients.

VII. SYSTEM ARCHITECTURE



The proposed system initially gathers all the required parameters. All the gathered data is further pre-

processed. Data cleaning, transformation and integration is done one after the other. The pre-processed data is analyzed and then visualized accordingly. This data is used to train the model and then come up with a prediction algorithm. The performances of all these algorithms are calculated. For accuracy, the ensemble learning method is applied.

VIII. FUTURE ENHANCEMENT

Transport Department needs to automate the recognizing the best course by not disaster from capability process (consistent) in perspective on the record detail. To automate this strategy by show the desire achieves web application or work territory application.

IX. CONCLUSION

A more flexible traffic signal has been updated for a mixed type of traffic. Mixed traffic consists of both manual and automated vehicles. For AVs, an efficient speed changing procedure is advanced responsively. The proposed system maintains the present traffic to be smooth and content with crossing point of genuinely decided vehicles. A number of parameters are considered in the system. These parameters are the main source on determining whether that traffic is an accident prone region or non-accident prone region. The user interface has been well developed where the user has the facility to choose the algorithm he wants to work with. However, the system is capable of working on all algorithms, a few being K-nearest, Random forest, Logistic Regression, and chooses the one that produces the highest accuracy. This facilitates the user to receive a highly reliable response. Thus the system turns out to be more trustworthy. Based on the result the user can avoid going to the accident prone region. This system works well for even normal people in their day to day routine. It can be developed to a smaller scale for the user's easy accessibility.

X. REFERENCES

- [1] K. G. Boldness and S. M. Parapar, "Postponement and fuel utilization at traffic signals," *Inst. Traffic Eng.*, vol. 45, no. 11, pp. 23–27, 1975.
- [2] T.- Q. Tang, Z.- Y. Yi, and Q.- F. Lin, "Impacts of sign light on the fuel utilization and emanations under vehicle following model," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 200–205, Mar. 2017.

- [3] N. H. Gartner, "Advancement and testing of an interest responsive system for traffic signal control" in Proc. Amer. Control Conf., Jun. 1982, pp. 578–583.
- [4] N. H. Gartner, F. J. Pooran, and C. M. Andrews, "Usage of the OPAC versatile control methodology in a rush hour gridlock signal system," in Proc. Intell. Transp. Syst., Aug. 2001, pp. 195–200.
- [5] A. G. Sims and K. W. Dobinson, "The Sydney facilitated versatile traffic (SCAT) framework reasoning and advantages," IEEE Trans. Veh. Technol., vol. VT-29, no. 2, pp. 130–137, May 1980.
- [6] D. I. Robertson and R. D. Bretherton, "Upgrading systems of traffic flags progressively the SCOOT strategy," IEEE Trans. Veh. Technol., vol. 40, no. 1, pp. 11–15, Feb. 1991.
- [7] S. Sen and K. L. Head, "Controlled streamlining of stages at a crossing point," Transp. Sci., vol. 31, no. 1, pp. 5–17, 1997.
- [8] B. Asadi and A. Vahidi, "Prescient journey control: Utilizing up and coming traffic signal data for improving mileage and diminishing excursion time," IEEE Trans. Control Syst. Technol., vol. 19, no. 3, pp. 707–714, May 2011.
- [9] M. A. S. Kamal, S. Taguchi, and T. Yoshimura, "Convergence vehicle helpful eco-driving with regards to somewhat associated vehicle condition," in Proc. Int. Conf. Intell. Transp. Syst. (ITSC), Sep. 2015, pp. 1261–1266.
- [10] S. I. Guler, M. Menendez, and L. Meier, "Utilizing associated vehicle innovation to improve the proficiency of crossing points," Transp. Res. C, Emerg. Technol., vol. 46, pp. 121–131, Sep. 2014.
- [11] J. Rios-Torres and A. A. Malikopoulos, "A study on the coordination of associated and mechanized vehicles at crossing points and converging at thruway entrance ramps," IEEE Trans. Intell. Transp. Syst., vol. 18, no. 5, pp. 1066–1077, May 2017.