A

Major Project

On

Road Accident Prediction using Machine Learning Algorithm

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

A.Prem Kumar	(177R1A0501)
G.Tharun Kumar	(177R1A0519)
S.Pranaya	(177R1A0550)

Under the Guidance of

B.P.Deepak Kumar

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CMR TECHNICAL CAMPUS
UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved byAICTE, New Delhi)
Recognized Under Section 2(f) & 12(B) of the UGC Act.1956,
Kandlakoya (V), Medchal Road, Hyderabad-501401.

2017-21

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "Road Accident Prediction using Machine Learning Algorithm" being submitted by A Prem kumar (177R1A0501) G Tharun Kumar (177R1A0519) & S Pranaya (177R1A0550) in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2020-21.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

B P Deepak Kumar
Assistant Professor
INTERNAL GUIDE

Dr.A.Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HoD

EXTERNAL EXAMINER

Submitted for viva voce Examination held on

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

B P Deepak Kumar, Assistant Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Mr. J. Narasimha Rao, Mr.B.P.Deepak Kumar, Mr.K. Murali, Dr. Suwarna Gothane** and **Mr. B. Ramji** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to Dr. K. Srujan Raju, Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to Dr. A. Raji Reddy, Director for being cooperative throughout the course of this project. We would like to express our sincere gratitude to Sri. Ch. Gopal Reddy, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of CMR Technical Campus who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

A.Prem Kumar (177R1A0501)

G.Tharun Kumar (177R1A0519)

S. Pranaya (177R1A0550)

ABSTRACT

The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.There are several problems with current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analysed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be helpful to analyse 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. Injuries due to road accidents are one of the most prevalent causes of death apart from health related issues. The World Health Organization states that road traffic injuries caused an estimated 1.35 million deathsworldwide in the year 2016. That is, a person is killed every 25 seconds. This calls for the need to analyze road accidents and the factors affecting them and come up with a method to reduce the probability of their occurence. The analysis of road accident severity was done by running an accident dataset through several machine learning classification algorithms to see which model performed the best in classifying the accidents into severity classes such as slight, severe and fatal.. It was observed that logistic regression to perform multilabel classification gave the highest accuracy score. It was also observed that factors such as number of vehicles, lighting conditions and road features played a role in determining the severity of the accident.The goal of this project is the investigate what causes not serious and serious accidents in hopes of preventing and decreasing the number of them. The dataset consists of accident records from the UK over the course of 15+ years. I hope to show the causes of these accidents through visualizations and create an algorithm that can predict the severity of accidents.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1. INTRODUCTION

1.1 PROJECT SCOPE

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

1.2 PROJECT PURPOSE

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, pollution, road structure, etc.

1.3 PROJECT FEATURES

After the prediction system is ready to use. The Website is developed for the user. The user just has to fill a form which consists of different options they need to select. They are like the type of climate, the type of vehicle and so on. Once the user submits the form the algorithm is triggered and the input given by the user is passed to the prediction system. The user is given how accident prone the road can be in percentage.

Road accidents have proved to be one of the leading causes of severe injury and has been on the increase over the years. With almost double the number of vehicles on the road compared to a few years ago, road accidents have been at an all time high; thus taking a huge toll on health, finance and property. Although various laws and safety measures have come into effect, there is always a probability of an accident occurring due to a variety of reasons.

Driver neglect, driver recklessness, road conditions, weather conditions, driving skill and a number of other factors influence the safety of both the vehicle and the surroundings.Road accident reports in the UK suggest that driver error has been the leading cause of vehicle collision, with the driver failing to look at his surroundings properly. Driver misjudging distance and speed of both same side and oncoming traffic has found to be a close second cause of accidents with about 80% of these collisions occurring on the same side of the road. Driving with poor maneuvering skills, low visibility, loss of control and driving on slippery surfaces also majorly contributed to the occurence of these accidents. With close to about 50000 cases having been reported in the year 2018, a vast majority of these accidents could have been avoided if the driver took the required precautions while on the road.

One of the most complicated and difficult daily needs is overland transportation. In India, more than 150,000 people are killed each year in traffic accidents. That's about 400 fatalities a day and far higher than developed auto markets like the US, which in 2016 logged about 40,000. Every year over 1 million vehicles are added to traffic averagely. 1.2 million People have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death- laceration ratio will increase.

Design and control of traffic by advanced systems come in view as the important need. Assumption on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous. Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data. Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using information from data and use statistical method.

The costs of fatalities and injuries due to traffic accidents have a great impact on

the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents [29][30].

There are several approaches that researchers have employed to study this problem. These include neural network, nesting logic formulation, log-linear model, fuzzy ART maps and so on. Applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behaviour, roadway condition and weather condition that were causally connected with different injury severity. This can help decision makers to formulate better traffic safety control policies. Roh et al. [22] illustrated how statistical methods based on directed graphs, constructed over data for the recent period, may be useful in modelling traffic fatalities by comparing models specified using directed graphs to a model, based on out-of-sample forecasts, originally developed by Peltzman [23]. The directed graphs model outperformed Peltzman's model in root mean squared forecast error. Ossenbruggen et al. [24] used a logistic regression model to identify statistically significant factors that predict the probabilities of crashes and injury crashes aiming at using these models to perform a risk assessment of a given region.

These models were functions of factors that describe a site by its land use activity, roadside design, use of traffic control devices and traffic exposure. Their study illustrated that village sites are less hazardous than residential and shopping sites. Abdalla et al. [25] studied the relationship between casualty frequencies and the distance of the accidents from the zones of residence. As might have been anticipated, the casualty frequencies were higher nearer to the zones of residence, possibly due to higher exposure. The study revealed that the casualty rates amongst residents from areas classified as relatively deprived were significantly higher than those from relatively affluent areas. Miaou et al. [26] studied the statistical properties of four regression models: two conventional linear regression models and two Poisson regression models in terms of their ability to model vehicle accidents and highway geometric design relationships. Roadway and truck accident data from the Highway Safety Information System (HSIS) have been employed to illustrate the use and the limitations of these models. It was demonstrated that the conventional linear regression models lack the distributional property to describe adequately random,

discrete, nonnegative, and typically sporadic vehicle accident events on the road. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships. Abdelwahab et al. studied the 1997 accident data for the Central Florida area [2]. The analysis focused on vehicle accidents that occurred at signalized intersections. The injury severity was divided into three classes: no injury, possible injury and disabling injury.

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

A recent study illustrated that the residential and shopping sites are more hazardous than village areas.as might have been predicted , the frequencies of the casualties were higher near the zones of residence possibly because of the higher exposure.A study revealed that the casualty rates among the residential areas are classified as relatively deprived and significantly higher than those from relatively affluent area.

2. LITERATURE SURVEY

2.LITERATURE SURVEY

2.1 RELATED WORK

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

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

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

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

There have been works in the prediction of accident severity that have used

algorithms such as Random Forest, Naive Bayes, linear regression and other methods to predict the severity of accidents. These methods of road traffic accidents have played a major role in setting up precautionary measures along areas that have been classified as danger zones or potential accident sites.

Road Accident Prediction has been done in various countries using a number of algorithms but one of the biggest issues is the fact that there lies a data imbalance. As all the data collected is of the occurrence of an accident but no record of the absenceof an accident. Therefore various methods have been used to perform negative sampling. Another issue is that it is difficult to perform road accident analysis for larger areas. All papers have utilised datasets consisting of only a small area or restricted themselves to a few road segments. Accident Risk Prediction based on Driving behaviour Feature using XGboost and Cart uses various parameters of driving behavior and are evaluated using which key features depending on correlation to the occurence of the accident is selected. This ensures that only the required features based on contribution to the accident plays a role in prediction and leaves out the redundant measures that have an indirect role to play in the collision.

Using XGBoost to predict the crash using characteristics of collision, time of the accident and the location of the accident and environmental factors showed to have the most accurate results. For usage of Naive Bayes algorithm it was found that grouping of characteristics into elements such as vehicles, road, human and environment helped get a more accurate result.

They compared the performance of Multi-layered Perceptron (MLP) and Fuzzy ARTMAP, and found that the MLP classification accuracy is higher than the Fuzzy ARTMAP. Levenberg-Marquardt algorithm was used for the MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively. The Fuzzy ARTMAP achieved a classification accuracy of 56.1 percent. Yang et al. used neural network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs. They

performed the Cramer's V Coefficient test to identify significant variables that cause injury to reduce the dimensions of the data. Then, they applied data transformation method with a frequency-based scheme to transform categorical codes into numerical values. They used the Critical Analysis Reporting Environment (CARE) system, which was developed at the University of Alabama, using a Backpropagation (BP) neural network. They used the 1997 Alabama interstate alcohol-related data, and further studied the weights on the trained network to obtain a set of controllable cause variables that are likely causing the injury during a crash. The target variable in their study had two classes: injury and non-injury, in which injury class included fatalities. They found that by controlling a single variable (such as the driving speed, or the light conditions) they potentially could reduce fatalities and injuries by up to 40%. Sohn et al. applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accidents.

The individual classifiers used were neural network and decision tree. They applied a clustering algorithm to the dataset to divide it into subsets, and then used each subset of data to train the classifiers. They found that classification based on clustering works better if the variation in observations is relatively large as in Korean road traffic accident data. Mussone et al. used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. They chose feed-forward MLP using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual conflict points, number of realconflict points, typeof intersection, accident type, road surface condition, and weather conditions). The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at nighttime. Dia et al. used real-world data for developing a multi-layered MLP neural network freeway incident detection model. They compared the performance of the neural network model and the incident detection model in operation on Melbourne's freeways. Results showed that neural network model could provide faster and more reliable incident detection over the model that was in operation.

They also found that failure to provide speed data at a station could significantly deteriorate model performance within that section of the freeway. Shankar et al. applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident.

They found that there is a greater probability of evident injury or disabling injury/fatality relative to no evident injury if at least one driver did not use a restraint system at the time of the accident. Kim et al. developed a log-linear model to clarify the role of driver characteristics and behaviors in the causal sequence leading to more severe injuries. They found that alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries. Abdel-Aty et al. used the Fatality Analysis Reporting System (FARS) crash databases covering the period of 1975-2000 to analyze the effect of the increasing number of Light Truck Vehicle (LTV) registrations on fatal angle collision trends in the US [1]. They investigated the number of annual fatalities that resulted from angle collisions as well as collision configuration.

Time series modeling results showed that fatalities in angle collisions will increase in the next 10 years, and that they are affected by the expected overall increase of the percentage of LTVs in traffic. Bedard et al. applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk [3]. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Evanco conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times [6]. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways. Ossiander et al. used Poisson regression to analyze the association between the fatal crash rate (fatal crashes per vehicle mile traveled) and the speed limit increase [13]. They found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State. Finally, researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities and the results of their work can be found in. This paper investigates

application of neural networks, decision trees and a hybrid combination of decision tree and neural network to build models that could predict injury severity. The remaining parts of the paper are organized as follows. In Section 2, more details about the problem and the pre-processing of data to be used are presented, followed, in Section 3, by a short description the different machine learning paradigms used. Performance analysis is presented in Section 4 and finally some discussions and conclusions are given towards the end.

Traffic accidents have become one of the leading causes of death and injuries worldwide, with more than 1.25 million of deaths per year, between 20 and 50 million injured and a global proportion of 18 deaths per 100,000 inhabitants (World Health Organization, 2015). Most of these deaths occur in undeveloped countries (16%) and in developing countries (74%), which converts mortality from traffic accidents, not only into a public health problem, but also into a socio-economic development issue.

To address this public health and socio-economic problem, the United Nations proposes in the initiative The Global Goals for sustainable development (The Global Goals, 2017), in the numeral three, to reduce in a half the world number of deaths and injuries resulting from traffic accidents by 2020. Therefore, investigation on prediction and prevention of traffic accidents are highly relevant to fulfill this goal for sustainable development and to reduce the mortality on roads.

Several studies have been carried out to propose road accident prediction and analysis models, each one of them framed within the socioeconomic, cultural and development conditions of the country where it was proposed, and therefore, making evident the difficulty of proposing a single predictive or analytical model that works in all contexts. The main areas of interest that these models study are i) detection of problematic areas for circulation (Cao et al., 2015; Fawcett et al., 2017; Kumar and Toshniwal, 2016a, Kumar and Toshniwal, 2016b); ii) real time detection of traffic incidents (D'Andrea et al., 2015; Gu et al., 2016); iii) road accident forecasting (Chen et al., 2016; Lin et al., 2015; Lu et al., 2017; Park et al., 2016; Shi and Abdel-Aty, 2015; You et al., 2017); and iv) prediction the severity of the consequences suffered by involved in a road accident (Kaplan and Prato, 2013; Mohamed et al., 2013; Zheng

et al., 2019). Government data refers to those data sets that are generated, collected, preserved, stored and made available to the public by government entities or those that are delegated to exercise functions of control, execution or reporting of information concerning road accidents (Cao et al., 2015; Kumar and Toshniwal, 2015a; Taamnehet al., 2017). Among these agencies can be included police bodies, traffic police and road concessionaires. Government data can be characterized as historical, since it contains data spanning several decades, and can be considered as reliable, because it is supported by the custody process of the entities responsible for the data. Government data is usually the technical support for the generation of public policy in each country regarding aspects like road infrastructure design and road security plans. One aspect to consider is that not all the variables of the data set may be available for public access, such as specific demographic information, sex and ethnicity, incompliance with the information privacy laws current in each country.

2.2 Python

Python is a programming language, which means it'a a language both people and computers can understand. Python was developed by a Dutch software engineer named Guido van Rossum, who created the language to solve some problems he saw in computer languages of the time. Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, and a syntax that allows programmers to express concepts in fewer lines of code, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation.

One significant advantage of learning Python is that it's a general-purpose language that can be applied in a large variety of projects. Below are just some of the most common fields where Python has found its use:

- Data science
- Scientific and mathematical computing
- Web development
- Computer graphics
- Basic game development
- Mapping and geography (GIS software)

Python's ecosystem is growing over the years and it's more and more capable of the statistical analysis. It's the best compromise between scale and sophistication (in terms od data processing). Python emphasizes productivity and readability. Python is used by programmers that want to delve into data analysis or apply statistical techniques (and by devs that turn to data science) There are plenty of Python scientific packages for data visualization, machine learning, natural language processing, complex data analysis and more. All of these factors make Python a great tool for scientific computing and a solid alternative for commercial packages such as MatLab. The most popular libraries and tools for data science are:

Pandas: a library for data manipulation and analysis. The library provides data structures and operations for manipulating numerical tables and time series.

NumPy: the fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

SciPy: a library used by scientists, analysts, and engineers doing scientific computing and technical computing.

Being a free, cross-platform, general-purpose and high-level programming language, Python has been widely adopted by the scientific community. Scientists value Python for its precise and efficient syntax, relatively flat learning curve and the fact that it integrates well with other languages (e.g. C/C++). As a result of this popularity there are plenty of Python scientific packages for data visualization, machine learning, natural language processing, complex data analysis and more. All of these factors

make Python a great tool for scientific computing and a solid alternative for commercial packages such as MatLab.

Matplotlib : Matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib allows you to generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, and more.

NumPy : NumPy is the fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

NetworkX : NetworkX is a library for studying graphs which helps you create, manipulate, and study the structure, dynamics, and functions of complex networks.

TomoPy : TomoPy is an open-sourced Python toolbox to perform tomographic data processing and image reconstruction tasks. TomoPy provides a collaborative framework for the analysis of synchrotron tomographic data with the goal to unify the effort of different facilities and beamlines performing similar tasks.

Theano : Theano is a numerical computation Python library. Theano allows you to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently.

SymPy : SymPy is a library for symbolic computation and includes features ranging from basic symbolic arithmetic to calculus, algebra, discrete mathematics andquantum physics. It provides computer algebra capabilities either as a standalone application, as a library to other applications, or live on the web.

SciPy : SciPy is a library used by scientists, analysts, and engineers doing scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

Scikit-learn : Scikit-learn is a machine learning library. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-image : Scikit-image is a image processing library. It includes algorithms for segmentation, geometric transformations, color space manipulation, analysis, filtering, morphology, feature detection, and more.

ScientificPython : ScientificPython is a collection of modules for scientific computing. It contains support for geometry, mathematical functions, statistics, physical units, IO, visualization, and parallelization.

SageMath : SageMath is mathematical software with features covering many aspects of mathematics, including algebra, combinatorics, numerical mathematics, number theory, and calculus. SageMath uses the Python, supporting procedural, functional and object-oriented constructs.

Veusz : Veusz is a scientific plotting and graphing package designed to produce publication-quality plots in popular vector formats, including PDF, PostScript and SVG.

Graph-tool : Graph-tool is a module for the manipulation and statistical analysis of graphs.

SunPy : SunPy is a data-analysis environment specializing in providing the software necessary to analyze solar and heliospheric data in Python.

Bokeh : Bokeh is a Python interactive visualization library that targets modern web browsers for presentation. Bokeh can help anyone who would like to quickly and easily create interactive plots, dashboards, and data applications. Its goal is to provide elegant, concise construction of novel graphics in the style of D3.js, but also deliver this capability with high-performance interactivity over very large or streaming datasets.

TensorFlow : TensorFlow is an open source software library for machine learning across a range of tasks, developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use. It is currently used forboth research and production at Google products, often replacing the role of its closed-source predecessor, DistBelief.

Nilearn : Nilearn is a Python module for fast and easy statistical learning on NeuroImaging data. Nilearn makes it easy to use many advanced machine learning, pattern recognition and multivariate statistical techniques on neuroimaging data for applications such as MVPA (Mutli-Voxel Pattern Analysis), decoding, predictive modelling, functional connectivity, brain parcellations, connectomes.

Dmelt : DataMelt, or DMelt, is a software for numeric computation, statistics, analysis of large data volumes ("big data") and scientific visualization. The program

can be used in many areas, such as natural sciences, engineering, modeling and analysis of financial markets. DMelt can be used with several scripting languages including Python/Jython, BeanShell, Groovy, Ruby, as well as with Java.

Python-weka-wrapper : Weka is a suite of machine learning software written inJava, developed at the University of Waikato, New Zealand. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The python-weka-wrapper package makes it easy to run Weka algorithms and filters from within Python.

Dask : Dask is a flexible parallel computing library for analytic computing composed of two components: 1) dynamic task scheduling optimized for computation, optimized for interactive computational workloads, and 2) Big Data collections like parallel arrays, dataframes, and lists that extend common interfaces like NumPy, Pandas, or Python iterators to larger-than-memory or distributed environments.

Python Saves Time : Even the classic “Hello, world” program illustrates this point:

```
print("Hello, world")
```

For comparison, this is what the same program looks like in Java:

```
public class HelloWorld {
    public static void main(String[] args)
        { System.out.println("Hello, world");
    }
}
```

3.SYSTEM ANALYSIS

3.SYSTEM ANALYSIS

SYSTEM ANALYSIS

The prediction, analysis and cause of flight delays have been a major problem for air traffic control, decision making by airlines and ground delay response programs. Studies are conducted on the delay propagation of the sequence. Also, studying the predictive model of arrival delay and departure delay with meteorological features is encouraged.

3.1 PROBLEM DEFINITION

Models are created using accident data records which can help to understand the characteristics of many features like drivers behavior, roadway conditions, light condition, weather conditions and so on. This can help the users to compute the safety measures which is useful to avoid accidents. It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. the model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it.

Here the road accident study is done by analyzing some data by giving some queries which is relevant to the study. The queries like what is the most dangerous time to drive , what fractions of accidents occur in rural, urban and other areas. What is the trend in the number of accidents that occur each year,do accidents in high speed limit areas have more casualties and so on.

3.2 EXISTING SYSTEM

Traffic accideny data with Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data. Road accidents are one of the most relevant causes of injuries and death worldwide, and therefore, they constitute a significant field of research on the use of advanced algorithms and techniques to

analyze and predict traffic accidents and determine the most relevant elements that contribute to road accidents.

3.2.1 LIMITATIONS OF EXISTING SYSTEM

- Data set if heavy data mining concepts doeskin have a goof scopes for the analysis.
- Prediction model always wont let us the accurate values.
- No proper resolution of how the accidents delay has been happening.
- Impacts of the delay on next location and time schedules

3.3 PROPOSED 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, pollution, road structure, etc. We propose to use this dataset to predict the severity of the accident caused due to the various factors that cause it and the conditions prevailing at the time of its occurrence. This will be done by training the data on algorithms such as logistic regression, classification to see which model performs the best.

3.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

- This model helps in predicting the delays of the flight under different circumstances like weather and technical problems.
- In this model analysis helps us to overcome other flight delays andmanagement issues.
- Prevention or avoidance of accidents and can know the major accident prone areas.
- Accident prediction final result is to find the percentage of accident in particular area

3.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis are

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

3.4.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it wouldrequire.
The following are some of the important financial questions asked during preliminary investigation:

- The costs conduct a full system investigation.
- The cost of the hardware and software.
- The benefits in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it give an indication of the system is economically possible for development.

3.4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3.4.3 BEHAVIORAL FEASIBILITY

This includes the following questions:

- Is there sufficient support for the users?
- Will the proposed system cause harm?

The project would be beneficial because it satisfies the objectives when developed

and installed. All behavioral aspects are considered carefully and conclude that the project is behaviorally feasible.

3.5 HARDWARE & SOFTWARE REQUIREMENTS

3.5.1 HARDWARE REQUIREMENTS:

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

- Processor : Intel Dual Core@ CPU 2.90GHz.
- Hard disk : 500GB and Above.
- RAM : 4GB and Above.
- Monitor : 5 inches or above.

3.5.2 SOFTWARE REQUIREMENTS:

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

Operating system : Windows 8, 10
Languages : Python
Backend : Machine Learning
IDE : Jupyter

4. ARCHITECTURE

4. ARCHITECTURE

4.1 PROJECT ARCITECTURE

This project architecture shows the procedure followed for flight delay using machine learning, starting from input to final prediction.

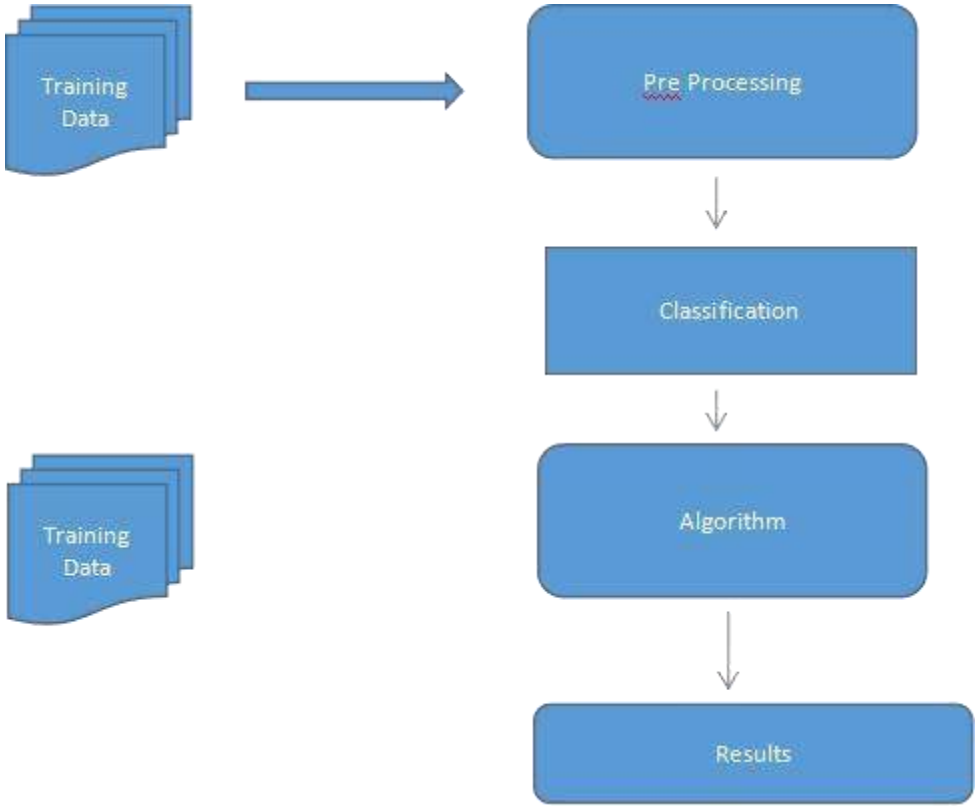


Figure 4.1: Project Architecture

4.2 DESCRIPTION

Input Data: Input data is generally in csv format where the data is fetched and mapped in the data framed from the source columns.

Reading Data: pandas sklearn library is used to read the data into the data frame.

Separating Features: In this following step we are going to separate the features which we take to train the model by giving the target value i.e. 1/0 for the particular

of features.

Normalization: Normalization is a very important step while we are dealing with the large values in the features as the higher bit integers will cost high computational power and time. To achieve the efficiency in computation we are going to normalize the data values.

Training and test data: Training data is passed to the MLP classifier to train the model. Test data is used to test the trained model whether it is making correct predictions or not.

4.3 USE CASE DIAGRAM

The use case graph is for demonstrating the direct of the structure. This chart contains the course of action of use cases, performing pros and their relationship. This chart might be utilized to address the static perspective of the structure.

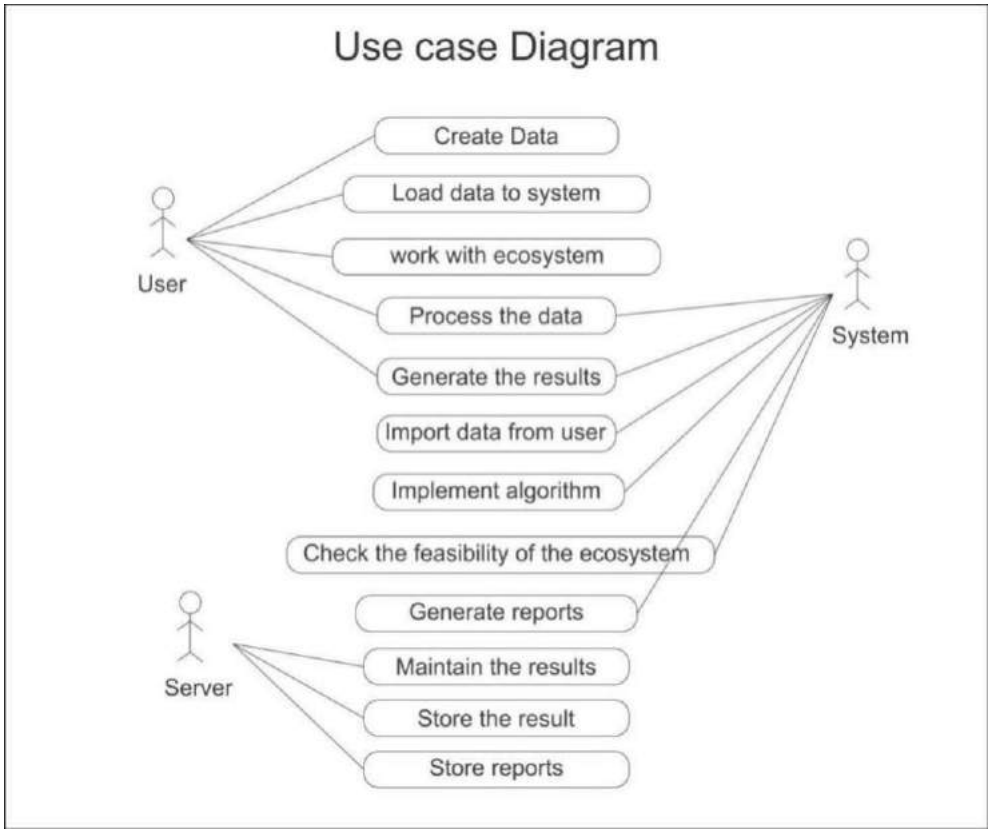


Figure 4.2: Use Case Diagram

4.4 CLASS DIAGRAM

The class graph is the most normally pulled in layout UML. It addresses the static course of action perspective of the structure. It solidifies the strategy of classes, interfaces, joint attempts and their affiliations.

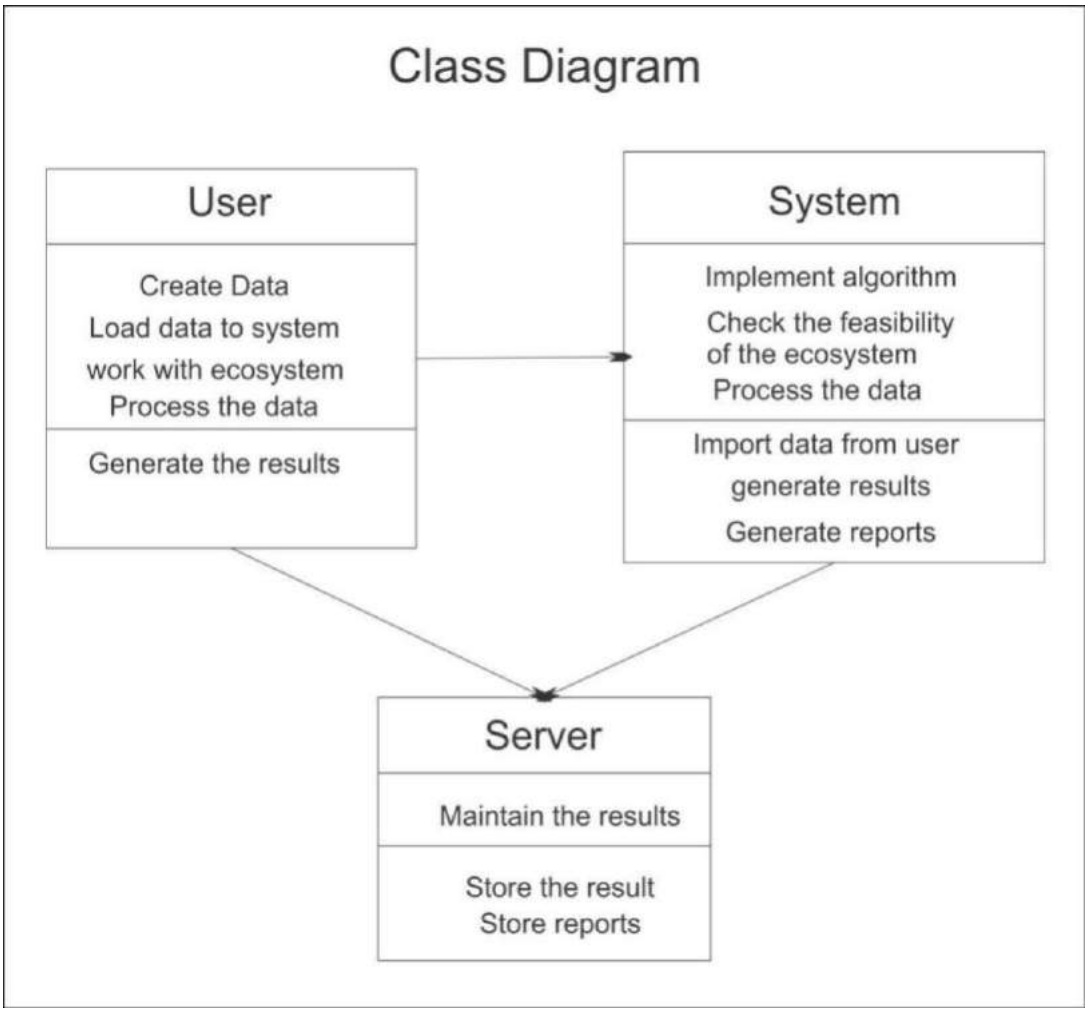


Figure 4.3: Class Diagram

4.5 SEQUENCE DIAGRAM

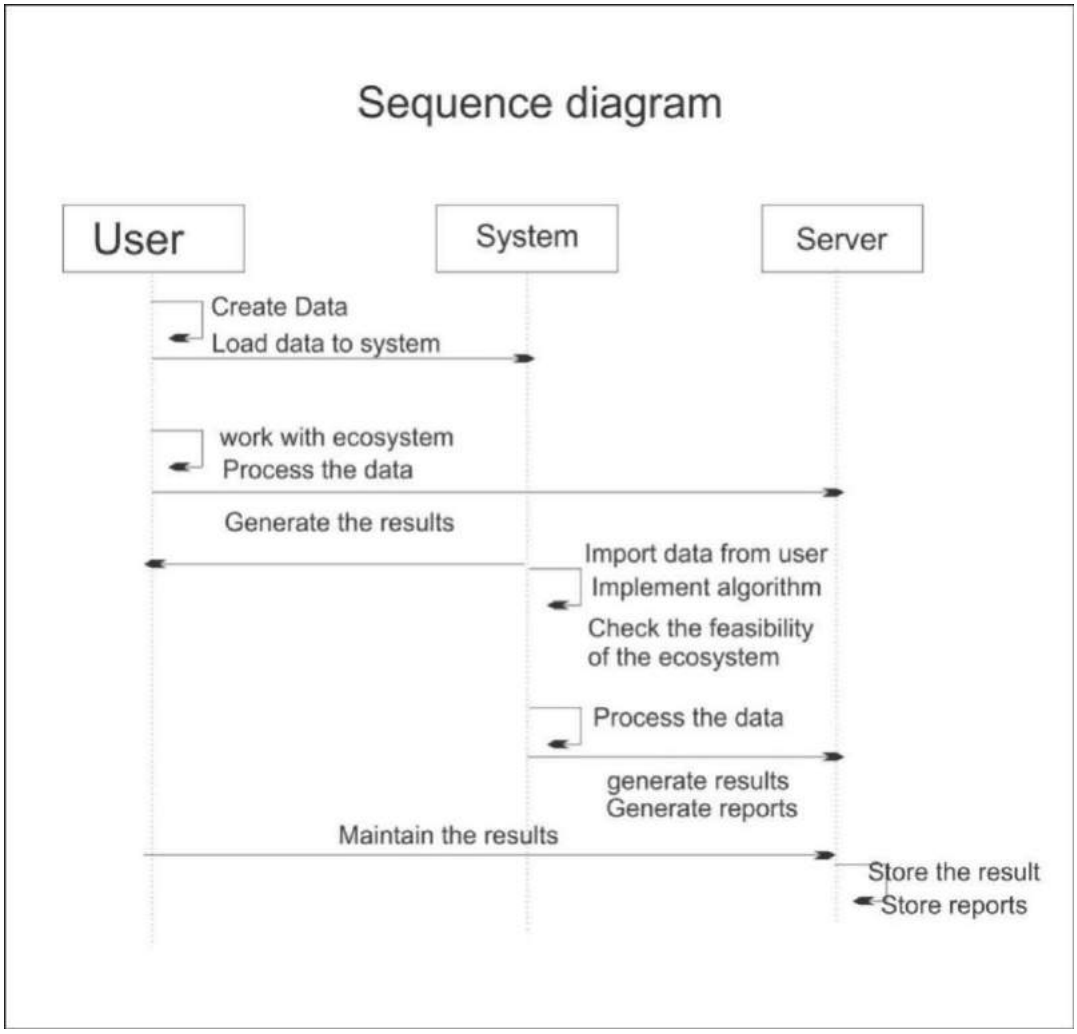


Figure 4.4: Sequence Diagram

4.6 ACTIVITY DIAGRAM

It describes about flow of activity states.

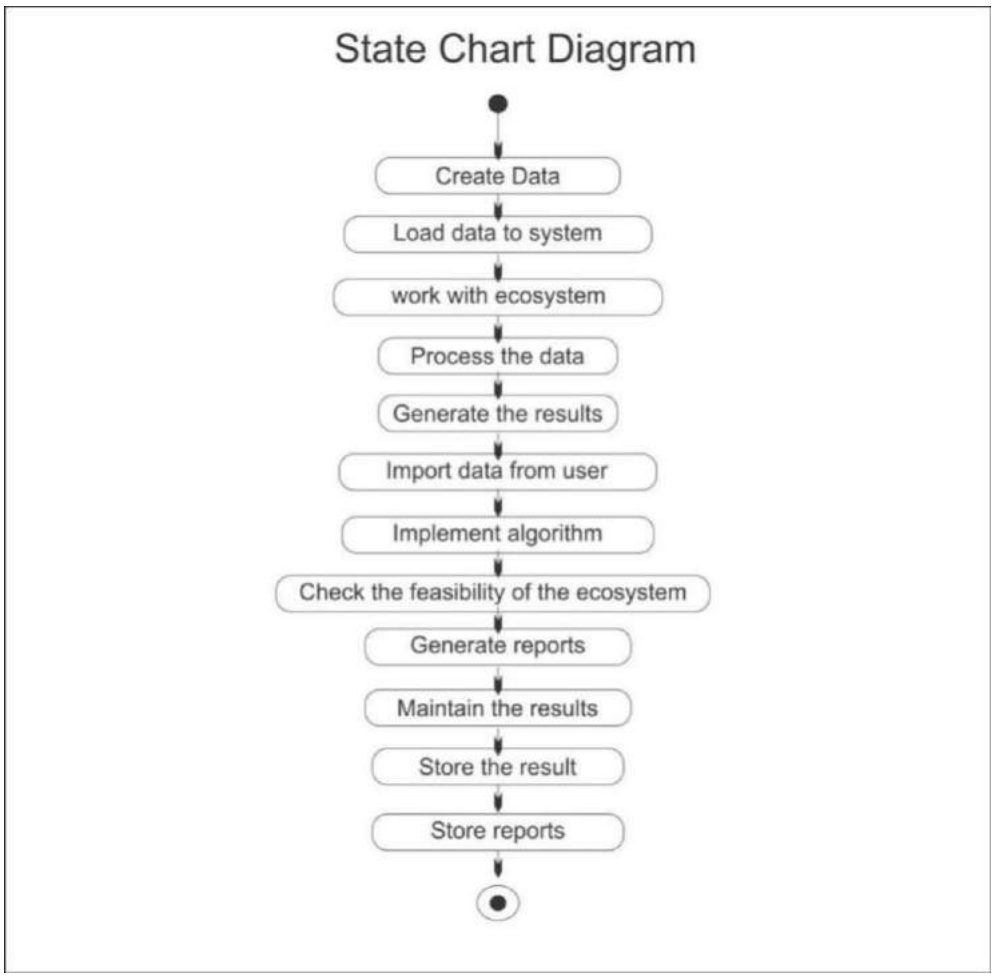


Figure 4.5: Activity Diagram

5. IMPLEMENTATION

5.IMPLEMENTATION

5.1 SAMPLE CODE

```
# import pandas as pd
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns



df = pd.read_csv( "C:/Users/sravan/Desktop/2021/chandu/project-3
traffic/accidents.csv" )
print ("The dataset has %d rows and %d columns" % (df.shape[0] , df.shape[1]))
df.head()

sns.countplot(y = "severity" , data = df )
plt.tight_layout()
plt.show()

pd.DataFrame( {"count": df["severity"].value_counts().values } , index =
df["severity"].value_counts().index )
df = df.loc[df["severity"] >     1].loc[df["severity"] < 4]

df["month"] = df["time"].apply(lambda x:int(x[:2]))
df["day"] = df["time"].apply(lambda x:int(x[3:5]))
df["year"] = df["time"].apply(lambda x:int(x[6:8]))
df["hour"] =    df["time"].apply(lambda x: int(x[9:11]) if str(x)[15] == 'A' else 12+
int(x[9:11])    )
df["lon"] = df["lon"].apply(lambda x:abs(x)) #so that multinomialNB works (only
```

```
with positive features)
#creating the date at the datetime format (easier to deal with)
df[ "date" ]= df[["month" , "day" ,"year"]].apply(lambda x:pd.datetime(month=
x['month'] , day = x['day']     , year = 2000+x["year"]), axis = 1)
df["weekday"] =     df["date"].apply(lambda x:x.weekday())



df2 = df.loc[df["severity"] == 2]
df3 = df.loc[df["severity"] == 3]

xx2 , yy2 = df2["lat"] , -df2["lon"]
xx3 , yy3 = df3["lat"] , -df3["lon"]

pts2 = plt.scatter(xx2,yy2,color = 'b' )
pts3 = plt.scatter(xx3,yy3,color = 'r' )
plt.legend((pts2, pts3), ('Severity = 2', 'Severity = 3'),loc='lower left')
plt.title("Accident Severity Map")
plt.tight_layout()
plt.show()

#classifiction , hour vs counts

sns.countplot(y = "hour" , data = df , order = range(1,25))
plt.title("Number of accidents by hour")
plt.tight_layout()
plt.show()

#classification , month vs count

sns.countplot(y = "month" , data = df)
plt.title("Number of accidents by month")
plt.tight_layout()
plt.show()
```

```
#classification , weekdays vs count

sns.countplot(y = "weekday" , data = df)
plt.title("Number of accidents by weekday")
plt.tight_layout()
plt.show()

severity_by_hour = pd.crosstab(index = df["hour"] , columns = df["severity"] )
severity_by_hour = pd.DataFrame(severity_by_hour.values) severity_by_hour["ratio"]
= severity_by_hour.apply(lambda x:x[0]/float(x[1]) , axis = 1)
severity_by_hour.sort_values(by = "ratio")

X = df[["month" , "hour" , "year", "weekday" ,"lon" , "lat"]]
y = df["severity"].apply(lambda x:x-2) # shifting to 0-1 values instead of 2-3



from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
random_state=42)



from sklearn.metrics import *
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
def printScores(y_test, y_pred, classif_name):
    print ("----------     "    + classif_name + " ----------")
    print ("recall : %0.2f" %    recall_score(y_test, y_pred) )
    print ("precision : %0.2f" %    precision_score(y_test, y_pred))
    print ("f1 : %0.2f" %    f1_score(y_test, y_pred))
    print ("accuracy : %0.2f" %    accuracy_score(y_test, y_pred))
```

```
    print ("-----------------------------------------------------")

sev = y.value_counts()
pd.DataFrame(sev)

print ("worst accuracy: " , max(sev)/float(sum(sev)))

from sklearn.linear_model import LogisticRegression
clf = LogisticRegression()
clf.fit(X_train,y_train)
y_pred = pd.Series(clf.predict(X_test))
printScores(y_test, y_pred, "LogisticRegression")

from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
printScores(y_test, y_pred, "tree")

pd.DataFrame(100*clf.feature_importances_ , index = X_train.columns)
#Import modules
import numpy as np
import holidays
import pandas as pd
import seaborn as sns
import pickle
import time
import timeit


import matplotlib.pyplot as plt
plt.style.use('dark_background')
%matplotlib inline
```

```
import datetime
import math
from collections import Counter

#scipy
import scipy.stats as stats
from scipy import stats
from scipy.stats import chi2_contingency

#sklearn
import sklearn
from sklearn import ensemble
from sklearn import preprocessing
from sklearn.decomposition import PCA
from sklearn.ensemble import AdaBoostClassifier, BaggingClassifier,
ExtraTreesClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, f1_score, precision_score, log_loss,
recall_score
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score
from sklearn.model_selection import cross_val_score, GridSearchCV, train_test_split
from sklearn.preprocessing import LabelEncoder, OneHotEncoder, StandardScaler,
OrdinalEncoder
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.utils import resample

#for clustering
from sklearn.cluster import KMeans
```

```
from sklearn.preprocessing import normalize
from sklearn.decomposition import PCA
from sklearn.cluster import MiniBatchKMeans
from sklearn.metrics import silhouette_score

#other learners
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from kmodes.kmodes import KModes

#imblearn
from imblearn.ensemble import BalancedBaggingClassifier
from imblearn.ensemble import EasyEnsembleClassifier
from imblearn.ensemble import BalancedRandomForestClassifier

#webscraping
import requests
from bs4 import BeautifulSoup
import re
import urllib
from IPython.core.display import HTML

#time series
import statsmodels.api as sm
from pylab import rcParams
import itertools
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.stattools import acf, pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima_model import ARIMA


#warning ignorer
```

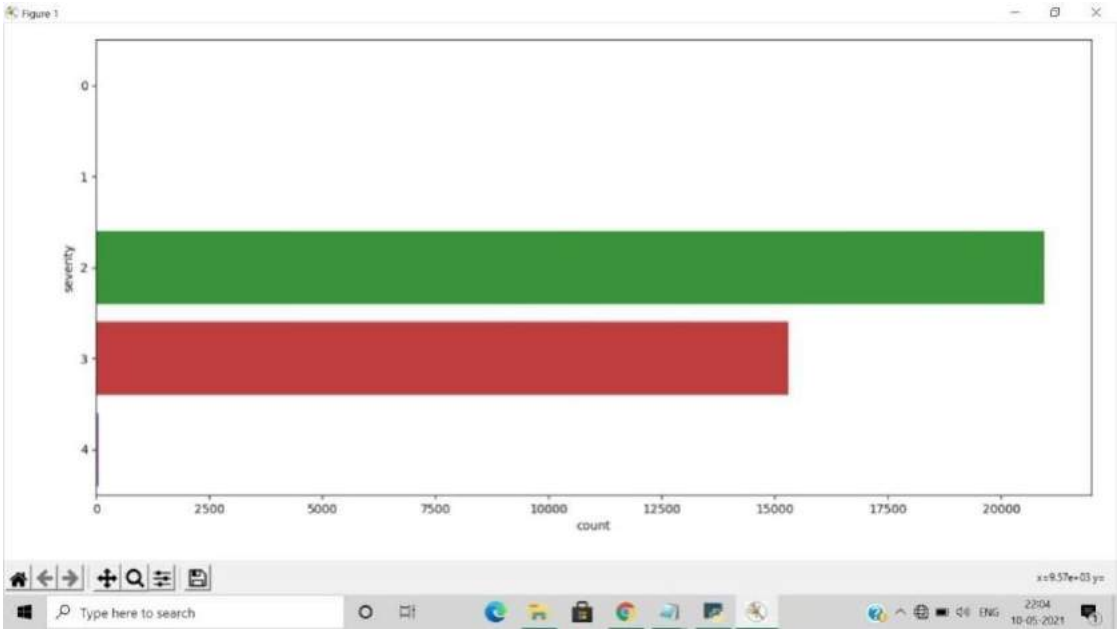
```
import warnings
warnings.filterwarnings("ignore")

#chunk cleaning and dataframing for accident column
acchunk = []
for chunk in ac:
    acchunk_filter =
        chunk[ (chunk.Year.astype(int) >=
        2010) & (chunk.Year.astype(int) <=
        2017) & (chunk['Road_Type'] !=
        "Unknown") &
        (chunk['Junction_Control'] != "Data missing or out of range") &
        (chunk['Carriageway_Hazards'] != "Data missing or out of range") &
        (chunk['Junction_Detail'] != "Data missing or out of range") &
        (chunk['Road_Surface_Conditions'] != "Data missing or out of range") &
        (chunk['Special_Conditions_at_Site'] != "Data missing or out of range") &
        (chunk['Weather_Conditions'] != "Data missing or out of range") &
        (chunk['Latitude'].notnull()) &
        (chunk['Longitude'].notnull())
    ]
    acchunk.append(acchunk_filter)
df1 = pd.concat(acchunk)


#chunk cleaning for vehicles column
vcchunk = []
for chunk2 in vc:
    vcchunk_filter = chunk2[
        (chunk2.Year.astype(int) >= 2010)&
        (chunk2.Year.astype(int) <= 2017) &
        (chunk2['Driver_Home_Area_Type'] != "Data missing or out of range") &
        (chunk2['Journey_Purpose_of_Driver'] != "Data missing or out of range")
&
        (chunk2['Junction_Location'] != "Data missing or out of range") &
```

```
            (chunk2['Was_Vehicle_Left_Hand_Drive'] != "Data missing or out of
range") &
            (chunk2['Hit_Object_in_Carriageway'] != "Data missing or out of range")
&
            (chunk2['Skidding_and_Overturning'] != "Data missing or out of range") &
            (chunk2['Towing_and_Articulation'] != "Data missing or out of range") &
            (chunk2['Vehicle_Leaving_Carriageway'] != "Data missing or out of range")
&
            (chunk2['Vehicle_Manoeuvre'] != "Data missing or out of range") &
            (chunk2['Vehicle_Type'] != "Data missing or out of range") &
            (chunk2['X1st_Point_of_Impact'] != "Data missing or out of range") &
            (chunk2['Sex_of_Driver'] != "Data missing or out of range") &
            (chunk2['Age_Band_of_Driver'] != "Data missing or out of range")

        ]
        vcchunk.append(vcchunk_filter)
df2 = pd.concat(vcchunk)



print("Accident's Columns:\n",df1.columns, "\n")

print("Vehicle's Columns:\n",df2.columns)

df = pd.merge(df1,df2)
print("Names of Combined Columns:\n",df.columns, "\n")
print("\nShape:\n",df.shape)
```

6. SCREENSHOTS

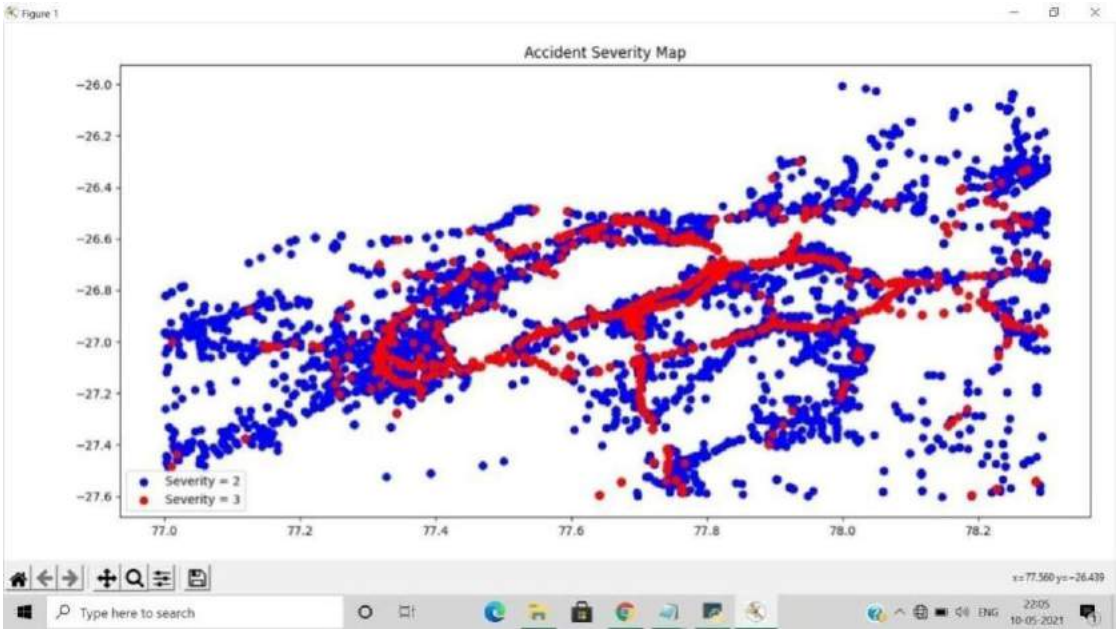
6.SCREEN SHOTS

6.1 Accidents by Server



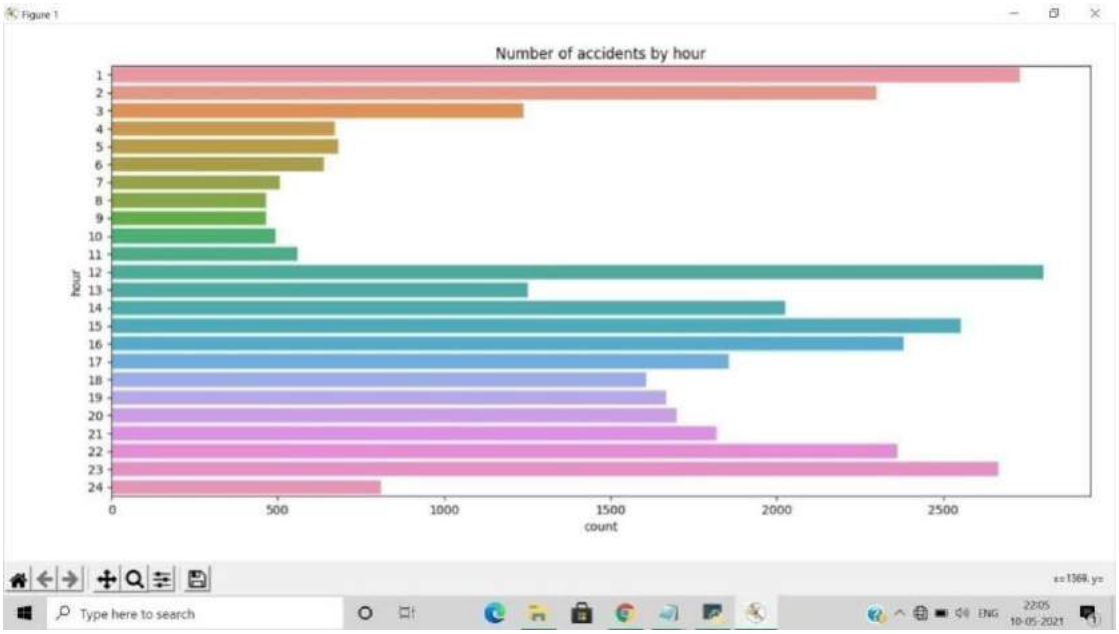
Screenshot 6.1: Severity of the accidents

6.2 Area wise Accidents



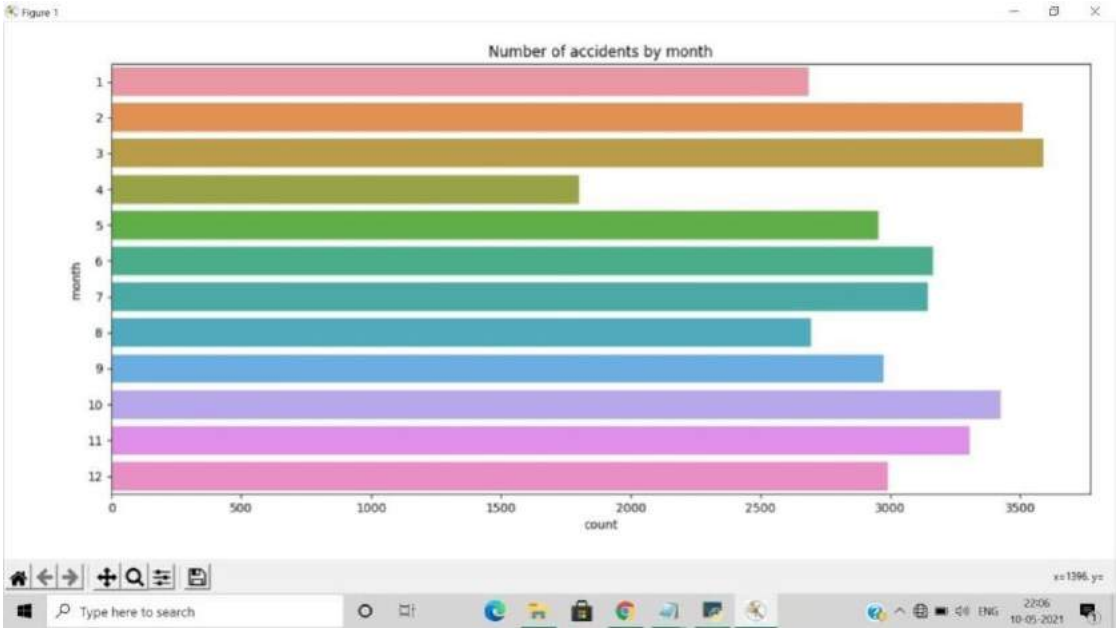
Screenshot 6.2: Location wise accidents of the data

6.3 Accidents by Hour



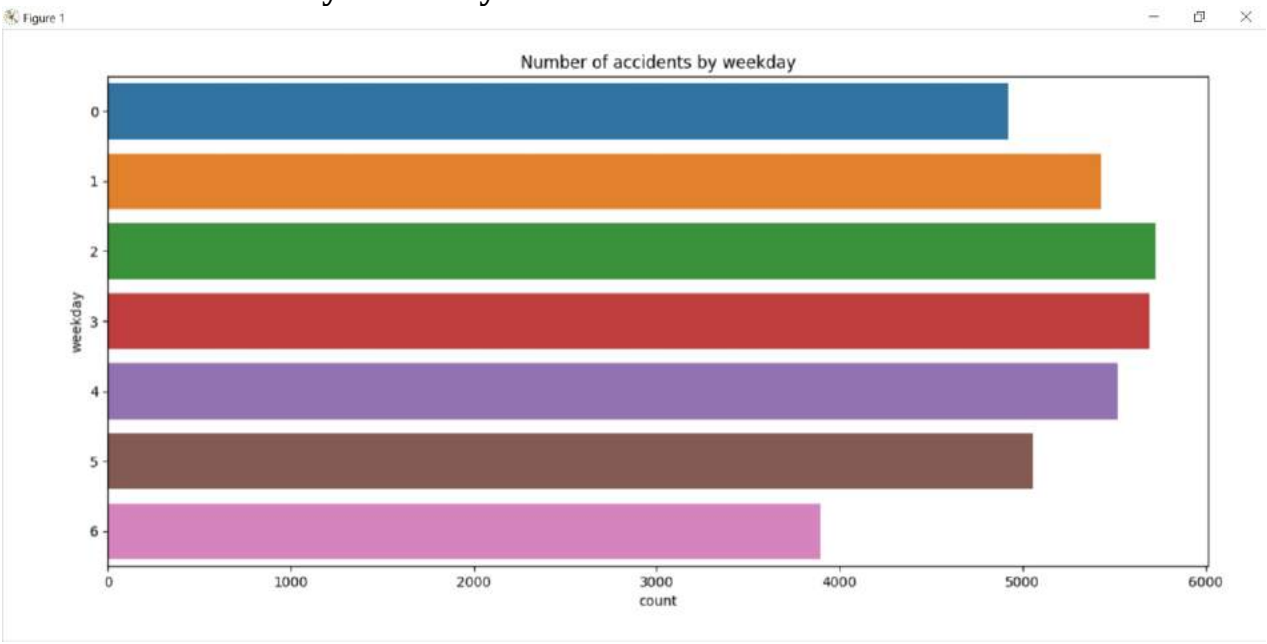
Screenshot 6.3: Accidents by hour

6.4 Accidents by Month



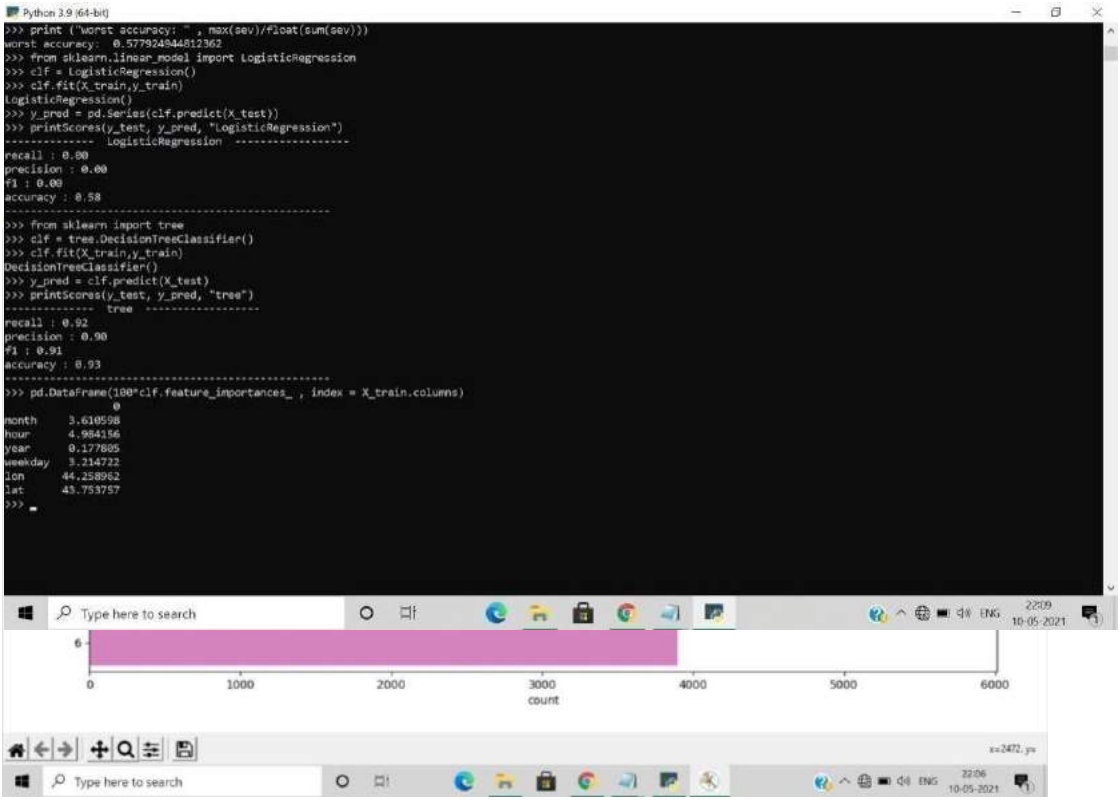
Screenshot 6.4: Accidents by Month

6.5 Accidents by Weekday



Screenshot 6.5: Accidents by Week

6.6 Regression Classifier



Screenshot 6.6: Linear model

7. TESTING

7.TESTING

7.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Testing is a procedure, which uncovers blunders in the program. Programming testing is a basic component of programming quality affirmation and speaks to a definitive audit of determination, outline and coding. The expanding perceivability of programming as a framework component and chaperon costs related with a product disappointment are propelling variables for we arranged, through testing. Testing is the way toward executing a program with the plan of finding a mistake. The plan of tests for programming and other built items can be as trying as the underlying outline of the item itself It is the significant quality measure utilized amid programming improvement. Amid testing, the program is executed with an arrangement of experiments and the yield of the program for the experiments is assessed to decide whether the program is executing as it is relied upon to perform.

A technique for programming testing coordinates the outline of programming experiments into an all around arranged arrangement of steps that outcome in fruitful improvement of the product. The procedure gives a guide that portrays the means to be taken, when, and how much exertion, time, and assets will be required. The procedure joins test arranging, experiment configuration, test execution, and test outcome gathering and assessment. The procedure gives direction to the specialist and an arrangement of points of reference for the chief. Due to time weights, advance must be quantifiable and issues must surface as ahead of schedule as would be prudent

Keeping in mind the end goal to ensure that the framework does not have blunders, the distinctive levels of testing techniques that are connected at varying periods of programming improvement are here.

Framework testing includes in-house testing of the whole framework before conveyance to the client. Its point is to fulfill the client the framework meets all necessities of the customer's determinations. This testing assesses working of framework from client perspective, with the assistance of particular report. It doesn't require any inward learning of framework like plan or structure of code.

It contains utilitarian and non-useful zones of utilization/item. Framework Testing is known as a super arrangement of a wide range of testing as all the significant sorts of testing are shrouded in it. In spite of the fact that attention on sorts of testing may differ on the premise of item, association procedures, course of events and necessities. Framework Testing is the start of genuine testing where you test an item all in all and not a module/highlight.

Acknowledgment testing, a testing method performed to decide if the product framework has met the prerequisite particulars. The principle motivation behind this test is to assess the framework's consistence with the business necessities and check in the event that it is has met the required criteria for conveyance to end clients. It is a pre-conveyance testing in which whole framework is tried at customer's site on genuine information to discover blunders. The acknowledgment test bodies of evidence are executed against the test information or utilizing an acknowledgment test content and afterward the outcomes are contrasted and the normal ones.

The acknowledgment test exercises are completed in stages. Right off the bat, the essential tests are executed, and if the test outcomes are palatable then the execution of more intricate situations are done

Bottom up Approach : Testing can be performed beginning from littlest and most reduced level modules and continuing each one in turn. In this approach testing is directed from sub module to primary module, if the fundamental module is not builtup a transitory program called DRIVERS is utilized to recreate the principle module.

At the point when base level modules are tried consideration swings to those on the following level that utilization the lower level ones they are tried exclusively and afterward connected with the already inspected bring down level modules.

Top down Approach : In this approach testing is directed from fundamental module to sub module. in the event that the sub module is not built up an impermanent program called STUB is utilized for mimic the sub module. This sort of testing begins from upper level modules. Since the nitty gritty exercises more often than not performed in the lower level schedules are not given stubs are composed. A stub is a module shell called by upper level module and that when achieved legitimately will restore a message to the calling module demonstrating that appropriate association happened.

7.2 TYPES OF TESTING

7.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results. Unit Testing is done on singular modules as they are finished and turned out to be executable. It is restricted just to the planner's prerequisites. It centers testing around the capacity or programming module. It Concentrates on the interior preparing rationale and information structures. It is rearranged when a module is composed with high union.

It is otherwise called Functional testing. A product testing strategy whereby the inward workings of the thing being tried are not known by the analyzer. For instance, in a discovery test on a product outline the analyzer just knows the information sources and what the normal results ought to be and not how the program touches base at those yields. The analyzer does not ever inspect the programming code and does not require any further learning of the program other than its determinations. In this system some experiments are produced as information conditions that completely execute every single practical prerequisite for the program. This testing has been utilizations to discover mistakes in the accompanying classifications.

7.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

It is otherwise called Glass box, Structural, Clear box and Open box testing . A product testing procedure whereby express learning of the inner workings of the thing being tried are utilized to choose the test information. Not at all like discovery testing, white box testing utilizes particular learning of programming code to inspect yields. The test is precise just if the analyzer comprehends what the program should do. He or she would then be able to check whether the program veers from its expected objective. White box testing does not represent blunders caused by oversight, and all obvious code should likewise be discernable. For an entire programming examination, both white box and discovery tests are required.

In this the experiments are produced on the rationale of every module by drawing stream diagrams of that module and sensible choices are tried on every one of the cases. It has been utilizations to produce the experiments in the accompanying cases

7.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must berejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs mustbe exercised.

Systems/Procedures: interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes.

7.3 TEST CASES

7.3.1 UPLOADING IMAGES

Test case ID	Test case name	Purpose	Test Case	Output
1	User uploads image	Use it for identification	The user uploads the a dog image	Uploaded successfully
2	User uploads 2nd image	Use it for identification	The user uploads the a non-dogimage	Uploaded successfully

7.3.2 CLASSIFICATION

Test case ID	Test case name	Purpose	Input	Output
1	Classification test 1	To check if the classifier performsits task	A dog image is given	Breed is predicted.
2	Classification test 2	To check if the classifier performsits task	A human image is given	It predicted as human.
3	Classification test 3	To check if the classifier performsits task	A frog imageis given	Predicted as not a dog.

8. CONCLUSION & FUTURE SCOPE

8.CONCLUSION & FUTURE SCOPE

8.1 PROJECT CONCLUSION

This paper provides a way to analyse the severity of road accidents and the factors that lead to them. It was observed that factors such as lighting conditions had a high effect on the severity of an accident. Factors like lighting and conditions can be improved upon to make roads safer which can then lead to lower rates of road accidents. Providing a database which contains such a large variety of data such as three classes of accident severity (slight, severe and fatal) and light conditions and details about the police officers at the scene, can be further analysed to provide useful insights and contribute to road safety. Although the occurrence of an accident cannot be controlled, the analysis of this data can enable the government and its citizens to take precautionary steps towards keeping themselves safer..

8.2 FUTURE SCOPE

Past research focused mainly on distinguishing between no-injury and injury (including fatality) classes. We extended the research to possible injury, non-incapacitating injury, incapacitating injury, and fatal injury classes. Our experiments showed that the model for fatal and non-fatal injury performed better than other classes. The ability of predicting fatal and non-fatal injury is very important since drivers' fatality has the highest cost to society economically and socially.

9. BIBLIOGRAPHY

9. BIBLIOGRAPHY

9.1 REFERENCES

[1] Abdel-Aty, M., and Abdelwahab, H., Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. Accident Analysis and Prevention, 2003.

[2] Abdelwahab, H. T. and Abdel-Aty, M. A., Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record 1746, Paper No. 01-2234.

[3] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002.

[4] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. Accident Analysis and Prevention, Vol. 30, No. 6, pp. 713-722, 1998.

[5] Dia, H., & Rose, G., Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. Transportation Research C, Vol. 5, No. 5, 1997, pp. 313-331.

[6] Evanco, W. M., The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462.

[7] Hand, D., Mannila, H., & Smyth, P., Principles of Data Mining. The MIT Press, 2001.

[8] Kim, K., Nitz, L., Richardson, J., & Li, L., Personal and Behavioral Predictors of Automobile Crash an Injury Severity. Accident Analysis and Prevention, Vol. 27, No. 4, 1995, pp. 469-481.

[9] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[10] Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.

[11] Mayhew, D. R., Ferguson, S. A., Desmond, K. J., & Simpson, G. M., Trends In Fatal Crashes Involving Female Drivers, 1975-1998. Accident Analysis and

Prevention, Vol. 35, 2003, pp. 407-415.

[12] Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol. 31, 1999, pp. 705-718..

9.2 WEBSITES

[1] https://web.stanford.edu/class/cs231a/prev_projects_2016/output%20(1).pdf [2] https://towardsdatascience.com/metrics-toevaluate-your-machine-learning-algorithm f10ba6e38234-.

ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM

A Prem kumar

UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

S Pranaya

UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

G Tharun Kumar

UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

B P Deepak Kumar

Assistant Professor, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Abstract:

The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

I Introduction

Road accidents have proved to be one of the leading causes of severe injury and has been on the increase over the years. With almost double the number of vehicles on the road compared to a few years ago, road accidents have been at an all time high; thus taking a huge toll on health, finance and property. Although various laws and safety measures have come into effect, there is always a probability of an accident occurring due to a variety of reasons. Driver neglect, driver recklessness, road conditions, weather conditions, driving skill and a number of other factors influence the safety of both the vehicle and the surroundings.Road accident reports in the UK suggest that driver error has been the leading cause of vehicle collision, with the driver failing to look at his surroundings properly. Driver misjudging distance and speed of both same side and oncoming traffic has

found to be a close second cause of accidents with about 80% of these collisions occurring on the same side of the road. Driving with poor maneuvering skills, low visibility, loss of control and driving on slippery surfaces also majorly contributed to the occurence of these accidents. With close to about 50000 cases having been reported in the year 2018, a vast majority of these accidents could have been avoided if the driver took the required precautions while on the road.

One of the most complicated and difficult daily needs is overland transportation. In India, more than 150,000 people are killed each year in traffic accidents. That's about 400 fatalities a day and far higher than developed auto markets like the US, which in 2016 logged about 40,000. Every year over 1 million vehicles are added to traffic averagely. 1.2 million People have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death- laceration ratio will increase.

Design and control of traffic by advanced systems come in view as the important need. Assumption on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous. Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data. Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using information from data and use statistical method.

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents.

II. LITERATURE SURVEY

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

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

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

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

There have been works in the prediction of accident severity that have used algorithms such as Random Forest, Naive Bayes, linear regression and other methods to predict the severity of accidents. These methods of road traffic accidents have played a major role in setting up precautionary measures along areas that have been classified as danger zones or potential accident sites.

Road Accident Prediction has been done in various countries using a number

of algorithms but one of the biggest issues is the fact that there lies a data imbalance. As all the data collected is of the occurrence of an accident but no record of the absence of an accident. Therefore various methods have been used to perform negative sampling. Another issue is that it is difficult to perform road accident analysis for larger areas. All

papers have utilised datasets consisting of only a small area or restricted themselves to a few road segments. Accident Risk Prediction based on Driving behaviour Feature using XGboost and Cart uses various parameters of driving behavior and are evaluated using which key features depending on correlation to the occurence of the accident is selected. This ensures that only the required features based on contribution to the accident plays a role in prediction and leaves out the redundant measures that have an indirect role to play in the collision.

Using XGBoost to predict the crash using characteristics of collision, time of the accident and the location of the accident and environmental factors showed to have the most accurate results. For usage of Naive Bayes algorithm it was found that grouping of characteristics into elements such as vehicles, road, human and environment helped get a more accurate result.

III. PROPOSED METHODOLOGY

Models are created using accident data records which can help to understand the characteristics of many features like drivers behavior, roadway conditions, light condition, weather conditions and so on. This can help the users to compute the safety measures which is useful to avoid accidents. It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. the model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it.

Here the road accident study is done by analyzing some data by giving some queries which is relevant to the study. The queries like what is the most dangerous time to drive , what fractions of accidents occur in rural, urban and other areas. What is the trend in the number of accidents that occur each year,do accidents in high speed limit areas have more casualties and so on.

Traffic accideny data with Data

mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data. Road accidents are one of the most relevant causes of injuries and death worldwide, and therefore, they constitute a significant field of research on the use of advanced algorithms and techniques to analyze and predict traffic accidents and determine the most relevant elements that contribute to road accidents.

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, pollution, road structure, etc. We propose to use this dataset to predict the severity of the accident caused due to the various factors that cause it and the conditions prevailing at the time of its occurrence. This will be done by training the data on algorithms such as logistic regression, classification to see which model performs the best.

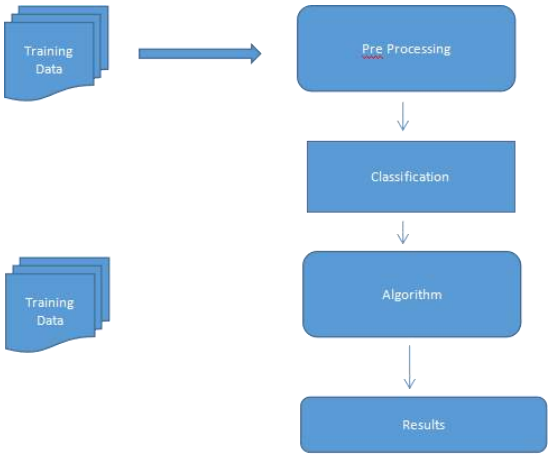


Fig 1: Project Architecture

IV RESULT ANALYSIS

After preprocessing and feature extraction of our dataset, 80% of the dataset was selected for training and 20% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrics.

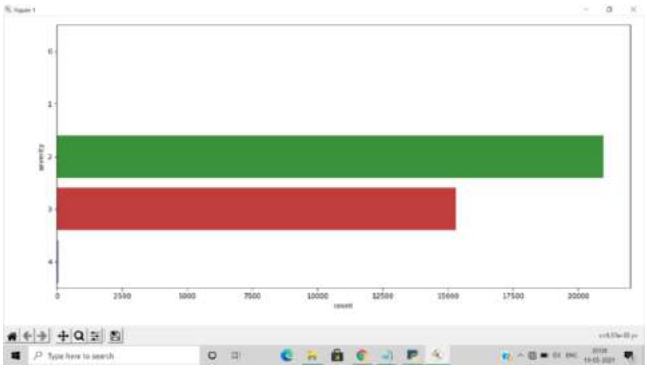


Fig 2 : Severity of the accidents

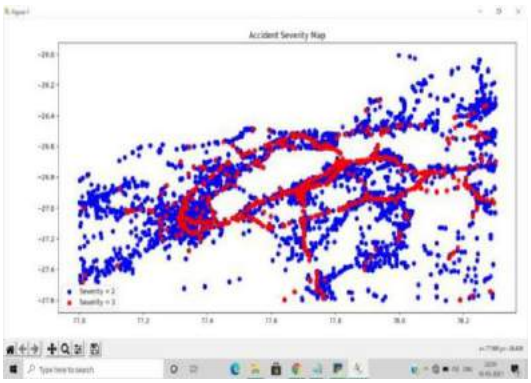


Fig 2: Location wise accidents of the data

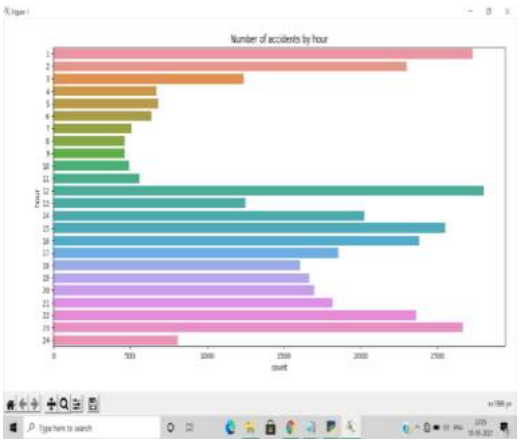


Fig 3: Accidents by hour

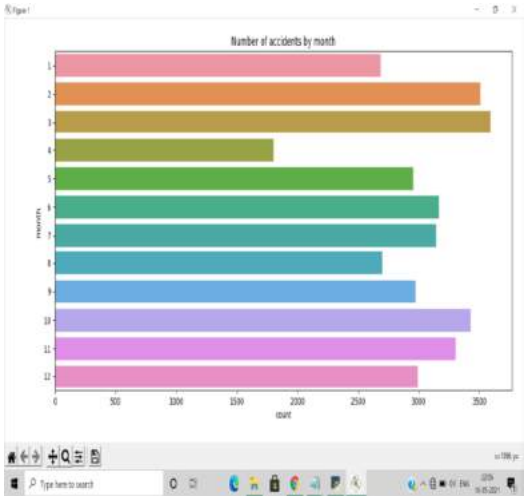


Fig 4 : Accidents by Month

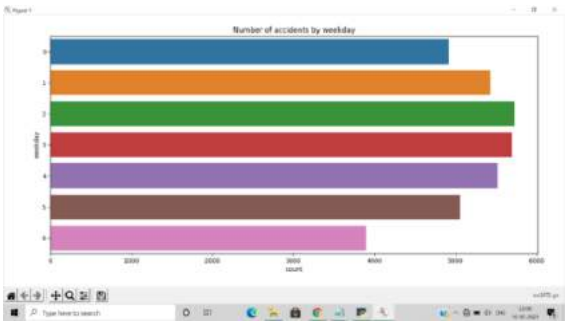


Fig 5 : Accidents by Week

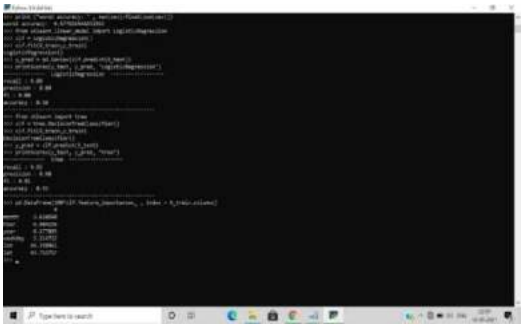


Fig6 :Linear model

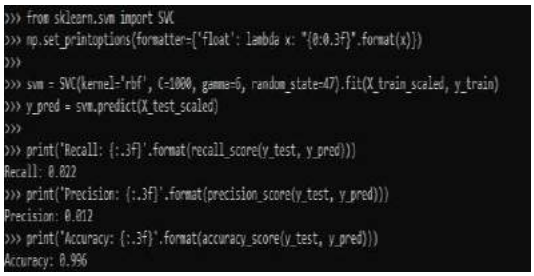


Fig 7 : SVC Classifier performance on the dataset

V CONCLUSION

This paper provides a way to analyse the severity of road accidents and the factors that lead to them. It was observed that factors such as lighting conditions had a high effect on the severity of an accident. Factors like lighting and conditions can be improved upon to make roads safer which can then lead to lower rates of road accidents. Providing a database which

contains such a large variety of data such as three classes of accident severity (slight, severe and fatal) and light conditions and details about the police officers at the scene, can be further analysed to provide useful insights and contribute to road safety. Although the occurrence of an accident cannot be controlled, the analysis of this data can enable the government and its citizens to take precautionary steps towards keeping themselves safer.

VI FUTURE SCOPE

Past research focused mainly on distinguishing between no-injury and injury (including fatality) classes. We extended the research to possible injury, non-incapacitating injury, incapacitating injury, and fatal injury classes. Our experiments showed that the model for fatal and non-fatal injury performed better than other classes. The ability of predicting fatal and non-fatal injury is very important since drivers' fatality has the highest cost to society economically and socially.

VII REFERENCES

[1] Abdel-Aty, M., and Abdelwahab, H., Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. Accident Analysis and Prevention, 2003.

[2] Abdelwahab, H. T. and Abdel-Aty, M. A., Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record 1746, Paper No. 01-2234.

[3] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002.

[4] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. Accident Analysis and Prevention, Vol. 30, No. 6, pp. 713-722, 1998.

[5] Dia, H., & Rose, G., Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. Transportation Research C, Vol. 5, No. 5, 1997, pp. 313-331.

[6] Evanco, W. M., The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462.

[7] Hand, D., Mannila, H., & Smyth, P., Principles of Data Mining. The MIT Press, 2001.

[8] Kim, K., Nitz, L., Richardson, J., & Li, L., Personal and Behavioral Predictors of Automobile Crash an Injury Severity. Accident Analysis and Prevention, Vol. 27, No. 4, 1995, pp. 469-481.

[9] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models.

Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[10] Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.

[11] Mayhew, D. R., Ferguson, S. A., Desmond, K. J., & Simpson, G. M., Trends In Fatal Crashes Involving Female Drivers, 1975-1998. Accident Analysis and Prevention, Vol. 35, 2003, pp. 407-415.

[12] Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol. 31, 1999, pp. 705-718.

[13] Prasadu Peddi (2017) "Design of Simulators for Job Group Resource Allocation Scheduling In Grid and Cloud Computing Environments", ISSN: 2319- 8753 volume 6 issue 8 pp: 17805-17811.

[14] Prasadu Peddi (2018), Data sharing Privacy in Mobile cloud using AES, ISSN 2319-1953, volume 7, issue 4.

The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

Certificate of Publication

This is to certify that the paper entitled

"ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM"

Authored by :

B P Deepak Kumar, Assistant Professor

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE V, MAY- 2021

Michal A. Olszewski Editor-In-Chief

IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/

The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

Certificate of Publication

This is to certify that the paper entitled

"ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM"

Authored by :

A. Prem kumar

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE V, MAY- 2021

Michal A. Olszewski Editor-In-Chief

IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/

The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

Certificate of Publication

This is to certify that the paper entitled

"ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM"

Authored by :

G. Tharun Kumar

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE V, MAY- 2021

Michal A. Olszewski Editor-In-Chief

IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/

The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

Certificate of Publication

This is to certify that the paper entitled

"ROAD ACCIDENT PREDICTION USING MACHINE LEARNING ALGORITHM"

Authored by :

S. Pranaya

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE V, MAY- 2021

Michal A. Olszewski Editor-In-Chief

IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/