Project Synopsis

on

Road Accident Analysis and Classification

Submitted as a part of course curriculum for

Bachelor of Technology

in

Computer Science

Submitted by

Anuj Jain (2000290120035)

Arth Srivastava (2000290120041)

Ayush Pratap Singh (2000290120054)

Under the Supervision of

Dr. Harsh Khatter

Assistant Professor

KIET Group of Institutions, Ghaziabad

Department of Computer Science

Dr. A.P.J. Abdul Kalam Technical University

2022-2023

DECLARATION

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature of Students
Name:
Roll No.:
Date:

CERTIFICATE

This is to certify that Project Report entitled "**Road Accident Analysis and Classification**" which is submitted by **Anuj Jain, Arth Srivastava, Ayush Pratap Singh** in partial fulfilment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date:

Supervisor Signature
Supervisor Name
(Designation)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the synopsis of the B.Tech Major Project undertaken during B.Tech. Third Year. We owe a special debt of gratitude to Dr. Harsh Khatter, Assistant Professor, Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Ajay Kumar Shrivastava, Head of the Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:
Date :
Name :
Roll No:

ABSTRACT

India is home to the second largest road network in the world with a total road length of approximately 62.1 lakh kilometers. This massive network serves as the nation’s lifeline transporting over 64.5% of all goods within the country in addition to being the preferred option for move of over 90% of India’s passenger traffic. While roads remain synonymous with development and growth in the country, they have also been a nemesis for users with India also carrying the dubious distinction of leading the global tally of annual deaths and injuries on account of road accidents. An asymmetry exists between number of vehicles and deaths due to road accidents with India’s one percent global share of number of vehicles accounting for almost 11% deaths due to road accidents.

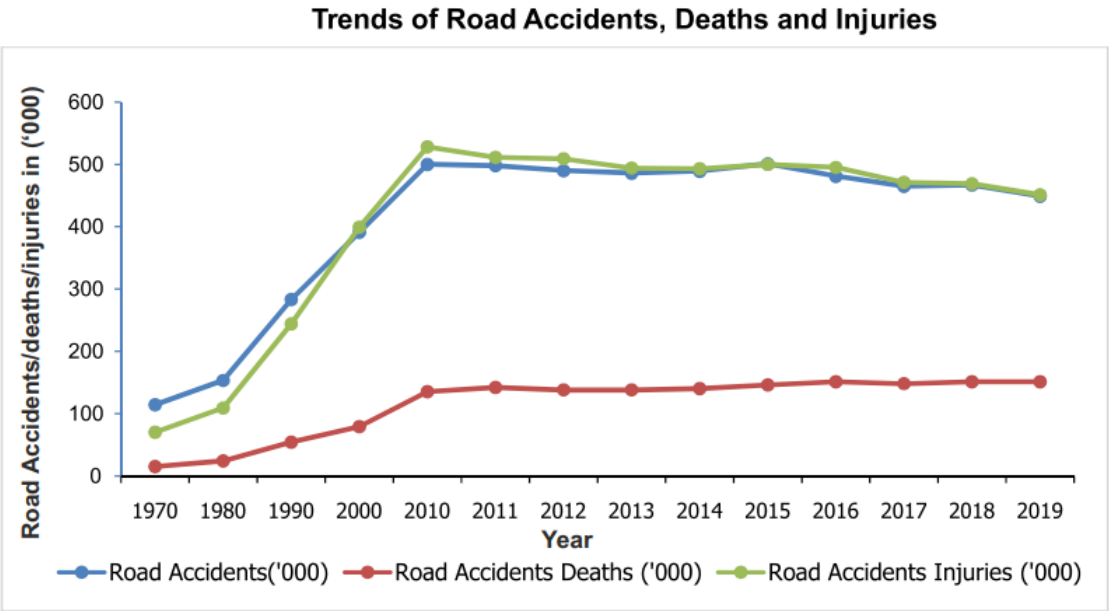
Road traffic accidents have been the world's leading cause of death and have increased significantly over the last three decades. The WHO Road Traffic Prevention Report lists road traffic accidents as the third leading cause of global disease burden, up from ninth place in 19901. India's contribution in this regard is one of the largest in the world, the country's share being second the most traffic accidents in the world and the most deaths.

TABLE OF CONTENTS

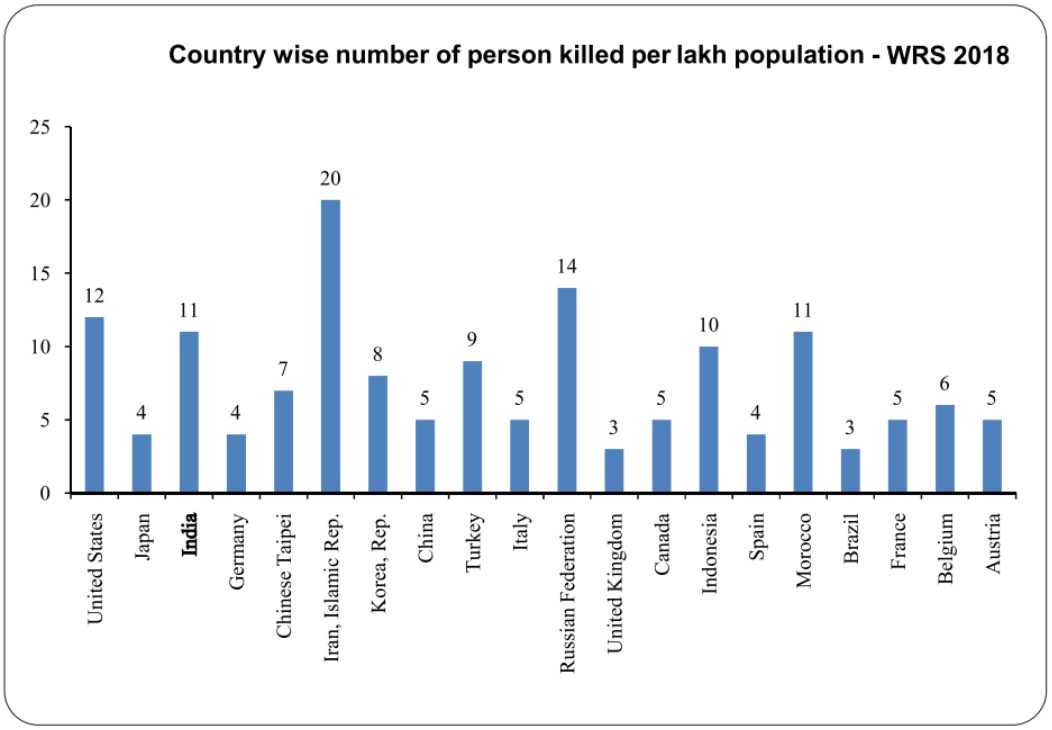
LIST OF FIGURES

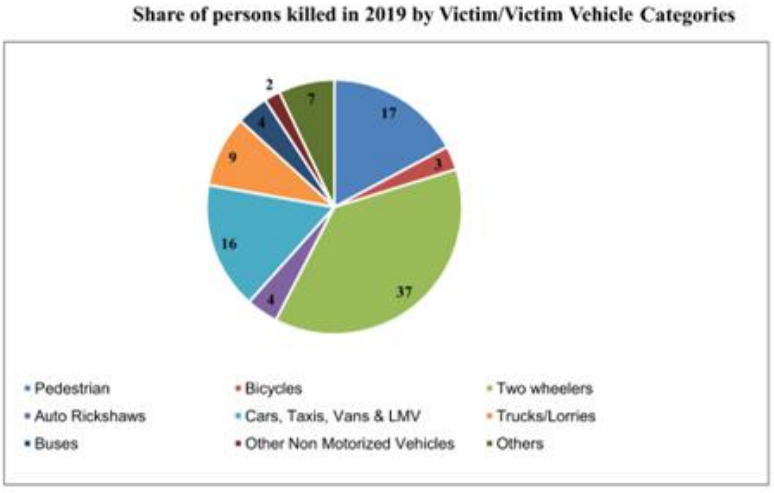
INTRODUCTION

Road accidents are a major cause of fatalities in India and other nations too. Fatality rate in developing nations is very high due to various aspects. According to a report, the number of deaths because of road accidents in India reached 1,51,000 in 2018. It results in loss of human life as well as capital. A WHO report stated that, almost 11% of deaths were due to road accidents in the year 2018. According to the Global status report on road safety outlined by World Health Organization in 2018, over 1.35 million people are killed each year and almost 3,700 people are killed every day globally in road-accidents involving cars, motorcycles, bicycles, buses, pedestrians, or tucks. Amongst 199 countries, India ranked number one in the number of deaths due to accidents. These are one of the most difficult real-world problems to tackle with, due to its high order of unpredictability. The persistence as well as existence of this problem may be prevalent to a different degree for each & every place. The only approach that can help decrease the number of road accidents, is to analyze the reasons that lead to these accidents.
Apart from being the lifeblood of the country and an enabler of socio-economic growth and development, India's road network is also the reason for the highest number of accidents in the country with road accidents accounting for 36-38% (average of 1,50,000 each year) deaths due to other causes during the period from 2015-19.



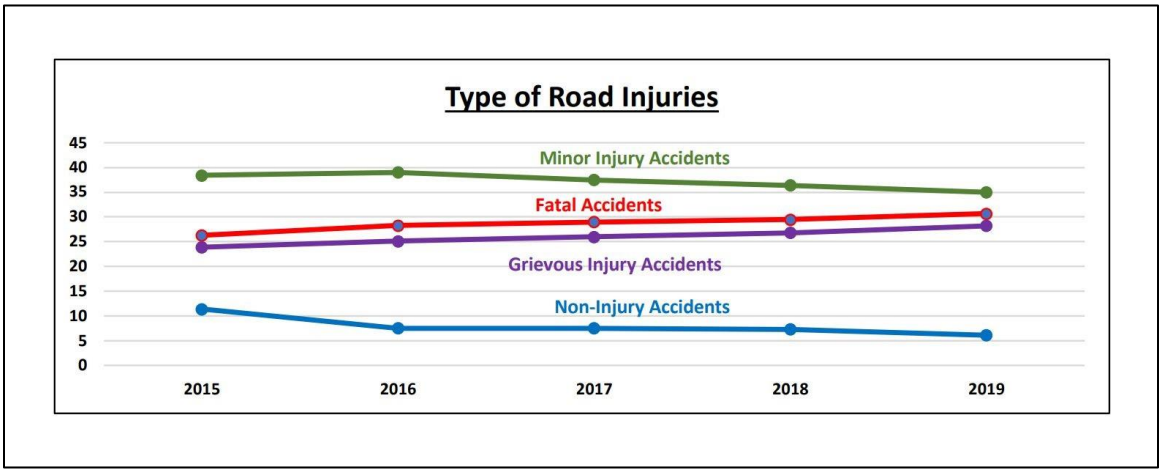
The only approach that can help decrease the number of road accidents, is to analyze the reasons that lead to these accidents. For safe driving advice, careful analysis of traffic data is crucial to identify variables closely related to fatal crashes. Machine learning (ML) is used to analyze different algorithms with experience and improve results. Concepts of data analysis, data visualization, and machine learning help solve real-world problems by exploring and gaining valuable insights, which in turn help act to solve a targeted problem and make predictions accordingly.


Country wise number of person killed per lakh population - WRS 2018
25
20
15
10
5
0
12
4
11
4
7
20
8
5
9
5
14
3
5
10
4
11
3
5
6
5
United States
Japan
India
Germany
Chinese Taipei
Iran, Islamic Rep.
Korea, Rep.
China
Turkey
Italy
Russian Federation
United Kingdom
Canada
Indonesia
Spain
Morocco
Brazil
France
Belgium
Austria


Share of persons killed in 2019 by Victim/Victim Vehicle Categories
17
3
37
4
16
9
4
2
7
Pedestrian
Bicycles
Two wheelers
Auto Rickshaws
Cars, Taxis, Vans & LMV
Trucks/Lorries
Buses
Other Non Motorized Vehicles
Others

PROBLEM STATEMENT

Accurate analysis is required to deal with massive traffic accidents on the ground. This analysis is done more deeply to determine the severity of traffic accidents using supervised learning techniques. It classifies accidents as fatal, serious, minor injury and motor vehicle accidents. Discovering the associations among the traffic accidents and related injuries is the key factor in reducing the traffic accidents. Identification of injuries severity is a key factor for the proper treatment. As number of traffic accidents are increasing and injuries severity is a critical factor to identify. Public suffering from many major injuries even after many years of accidents.



There are various problems in real time for the prevention of chance of accidents in the location. Accidents are one of the serious issues faced by people in this modern era. The main reasons for this are the negligence and carelessness of people. Even pedestrians are also facing severe injuries and the present system is not effective. The present road conditions provide a major aid for accidents to occur. Moreover, people are unaware of the speed limit and the accident-prone areas while they are traveling. If they are made aware of these things the accidents and problems can be reduced to a certain limit.

OBJECTIVE

The main objective of this model is to visualize the data to understand and detect which regions have the most accident-prone area, in what type of weather the accidents are occurring and at what hour/day/week/month/year accident records are more and analyze the data with help of machine-learning algorithms and predict the accuracy of accidents that might occur in the future.

LITERATURE REVIEW

Road Accident Analysis using Machine Learning

Authors: Jayesh Patil; Mandar Prabhu; Dhaval Walavalkar; Vivian Brian Lobo

Accidents through roadways have been a great threat to developed as well as underdeveloped countries. Road accidents and their safety have been a major concern of the world and over the years everyone has tried to deal with it. Road traffic and careless driving happen all over the world. It also affects many pedestrians. With no fault, they become victims. Many road accidents occur because of numerous factors like atmospheric changes, sharp curves, and human faults. Injuries caused by road accidents are major but sometimes imperceptible, which later affect health too. The purpose of this study is to analyze traffic accidents in a popular metropolitan city, i.e., Bengaluru, using k-means algorithm and machine learning looking at accident-prone areas and their root causes. k-means is an arrangement of vector quantization that targets to divide n instances into k groups wherein each instance is a part of a cluster with a closest average functioning as an architype for the cluster. It is popular for cluster analysis. k-means curtails cluster variances but not regular Euclidean distances, which would be a difficult Fermat–Weber problem, i.e., mean enhances squared errors, whereas only a geometric median decreases Euclidean distances. For instance, better Euclidean solutions can be determined using k-medians and k-medoids. ML has been helping us to solve many problems in our day-to-day life. It has helped to analyze data provided and provide appropriate solutions to problems that occur. Due to which, the study uses k-means algorithm. This study aimed to determine the reason behind the major cause of the increase in the number of road accidents happening around. For the past few years, it was noticed that the rate of road accidents had been increasing at an alarming rate due to various factors like drunk driving, problems related to climate, human error, etc. Considering this, the study of road accidents can play an important role to prevent road accidents that would have happened soon.

Overview of use of decision tree algorithms in machine learning

Authors: Arundhati Navada; Aamir Nizam Ansari; Siddharth Patil; Balwant Sonkamble

A decision tree is a tree whose internal nodes can be taken as tests (for patterns in the input data) and whose leaf nodes can be taken as categories (for their patterns). These tests are filtered through the tree to get the correct results for the input pattern. Decision tree algorithms can be applied and used in various fields. It can be used to replace statistical functions for information serving, text extraction, finding missing information in class, improving search engines, and finding various applications in medical fields. Several decision tree algorithms have been developed. They have different accuracy and cost effectiveness. It is also very important for us to know which algorithm is best to use. ID3 is one of the oldest decision tree algorithms. It is very useful for making simple decision trees, but as complexity increases, its accuracy for making good decision trees decreases. Therefore, IDA (Intelligent Decision Tree Algorithm) and C4.5 algorithms were prepared.

A State of Art ML Based Clustering Algorithms for Data Mining

Authors: Amjad Ali; Zaid Bin Faheem; Muhammad Waseem; Umar Draz; Zanab Safdar; Shafiq Hussain; Sana Yaseen

Data mining is an unsupervised learning technique for extracting data and hidden relationships. Data mining has greater importance in data science and machine learning because through data mining all the hidden information that defines different aspects of a data set is revealed. Clustering is a data mining technique for grouping data based on measures of similarity. Objects or data points in a cluster are similar. Correspondingly, the objects or data points of the other cluster are also similar. But when you compare these clusters, they differ. Clustering is considered the most important unsupervised learning method because it deals with finding structure in unlabeled data collection. Various approaches can be used for clustering, such as partition clustering, hierarchical clustering, density-based clustering, and network-based clustering. These clustering methods can be done using several algorithms such as K-means clustering, fuzzy C-means clustering. K-means clustering is a cluster analysis technique that aims to divide n observations into groups where each observation is closely related. The algorithm is called k-means because it creates the desired number of similar clusters, with the mean placed in the center of the cluster. The goal of the algorithm is to find the k-means of the desired data that we want to manage in clusters.

Analysis of road traffic fatal accidents using data mining techniques

Authors: Liling Li Department of Computer Science, Central Michigan University, USA Sharad Shrestha Department of Computer Science, Central Michigan University, USA Gongzhu Hu Department of Computer Science, Central Michigan University, USA

Road safety is of great concern to both transport authorities and ordinary citizens. For safe driving advice, careful analysis of traffic data is crucial to identify variables closely related to fatal crashes. This paper applies statistical analysis and data algorithms to the FARS fatal crash data set to address this issue. The association of fatalities with other characteristics such as crash mode, weather, ground conditions, light and drunk driving was examined. Association rules were found with the Apriori algorithm, the classification model was built with the Naive Bayes classifier, and clusters were formed with a simple K-means clustering algorithm. Certain safety driving suggestions were made based on statistics, association rules, classification model and resulting clusters. The cluster result showed that some states/territories have higher death rates while others have lower death rates. We may be more careful when driving in high-risk states/areas. Through the completed task, the data never seems strong enough to make a strong decision. When more data is available (e.g., non-fatal accidents, weather data, mileage data, etc.), more experiments can be conducted and more recommendations can be made based on the data.

Accuracy vs. Cost in Decision Trees: A Survey

Authors: Mona Al Hamad Department of Information Systems, University of Bahrain, Manama, Bahrain Ahmed M. Zeki Department of Information Systems, University of Bahrain, Manama, Bahrain

Decision trees have been widely used for classification in many fields such as finance, marketing, engineering, and medicine. The expanded scope required a thorough understanding of various aspects of decision-making disabilities. In addition, it is important to understand the different costs associated with the classification task of a decision tree classifier and their relation to the accuracy of the classifier, because balancing the two is a major problem in many fields such as medical diagnosis today. The focus of the work is to look at the different costs of building a DT, their calculation, different accuracy measures to evaluate the efficiency of a classifier, and the relationship between classification accuracy and cost. Based on the analysis, the relationship between classification accuracy and DT costs are found to be proportional. In addition, most researchers have focused on either test costs or misclassification costs between different types of costs when constructing a cost-sensitive DT.

Road Accident Analysis and Hotspot Prediction using Clustering

Authors: Hui Xu, Shunyu Yao, Qianyun Li, Zhiwei Ye, Hubei University of Technology, No. 28, Nanli Road, Hong-shan District, Wuhan, China

Among the dominant clustering algorithms, K-means has emerged as one of the most widely used technologies, mainly due to its simplicity and efficiency. However, the choice of the original algorithm centers and the sensitivity to noise can reduce the clustering effect. To solve these problems, this paper proposes an improved K-means algorithm. The idea of the CLIQUE network is used to remove noise and obtain spatial density. The original center is then selected according to the Fast Search and Find of Density Peaks (CFSFDP) method. Also, the concept of granularity mitigates the effect of grid density error on initial center selection by avoiding cluster center selection by manually participating in the peak density algorithm. Compared with the original K-Means algorithm, the improved algorithm proposed in this paper has higher accuracy, less difference in clustering effects of different data, and less dependence on parameters.

Traffic Accident Detection Using Random Forest Classifier

Authors: Nejdet Dogru International Burch University, Faculty of Engineering and Natural Sciences, Sarajevo, Bosnia and Herzegovina Abdulhamit Subasi Effat University, College of Engineering, Jeddah, Saudi Arabia

The Internet of Things (IoT) has grown in recent years with many completely different applications in the fields of military, maritime, smart transportation, smart healthcare, and smart cities. Although the Internet of Things has significant advantages over traditional information and communication technologies in Intelligent Transportation Systems (ITS), these applications are still rare. Although the safety of roads and vehicles is constantly improving, with the improvement of the Internet of Things, the number of traffic accidents has increased in recent decades. Therefore, it is necessary to look for an effective idea to reduce the frequency and severity of traffic accidents. Therefore, this paper presents an intelligent traffic accident detection system in which vehicles exchange their microscopic vehicles with each other. The designed system uses simulated data collected from traffic ad-hoc networks (VANET), which supports vehicle speeds and coordinates and thus sends traffic messages to drivers. In addition, it shows how machine learning strategies are used for ITS traffic accident detection. It was shown that by providing vehicle position and speed values, vehicle behavior can be analyzed and accidents can be easily detected. Supervised machine learning algorithms such as Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Random Forests (RF) are forced on traffic data to develop a model that separates accidents from normal cases. The performance of the RF algorithm program was found to be better than the ANN and SVM algorithms in terms of accuracy. The RF algorithm program showed better performance with 91.56 percent accuracy than SVM with 88.71 percent accuracy and ANN with 90.02 percent accuracy.

Road Traffic Accidents Injury Data Analytics

Authors: Mohamed K Nour College of Computer and Information Systems Umm Al-Qura University, Atif Naseer Science and Technology Unit Umm Al-Qura University, Basem Alkazemi College of Computer and Information Systems Umm Al-Qura University, Muhammad Abid Jamil College of Computer and Information Systems Umm Al-Qura University

Traffic safety researchers working with traffic accident data have seen success in analyzing traffic crashes using applied data analysis techniques, although little progress has been made in predicting traffic injuries. This paper uses advanced analysis methods to predict injury severity and evaluate their performance. The study uses predictive modeling techniques to identify key factors influencing risk and accident. The study uses publicly available data from the UK Department for Transport covering the period 2005-2019. The paper presents an approach that is general enough to be applied to datasets from various other countries. The results showed that tree-based techniques such as XGBoost are more efficient than regression-based techniques such as ANN. In addition to work, identify interesting relationships and recognized information about quality problems.

An Implementation of Naive Bayes Classifier

Authors: Feng-Jen Yang Department of Computer Science, Florida Polytechnic University, Lakeland, Florida, USA

Classification is a commonly used machine learning and data mining approach. Depending on the number of object classifications used to classify the dataset, different approaches can be chosen to perform the classification work. Binary classifiers often use decision trees and support vector machines, but those two approaches have the limitation that the number of target classifiers cannot exceed two. This rigid limitation makes it difficult to generalize them to fit a broad understanding of real classification, where there are usually more than two target classifications. In terms of general tool acquisition, the Naive Bayes classifier is more suitable for general classification expectations. As a mathematical classification method, the Naive Bayes classifier contains a series of probability calculations whose purpose is to find the most appropriate classification for the information in the given problem area. This article describes the implementation of the Naive Bayes classifier. Bayesian probabilistic calculations have been widely used to predict outcomes under uncertainty. Using the powerful built-in programming constructs of the Python programming language, the implementation of the Naive Bayes classifier is done without much intensive coding. The main contribution of this application is to provide a general set of tools that can be applied to many different classification fields. This classifier can be used as a general tool and is suitable for many classification domains. A sample data set is selected to test this classifier to ensure that all the probability calculations involved are correct.

K-Means Clustering Algorithms: Implementation and Comparison

Authors: Gregory A. Wilkin, Xiuzhen Huang, Arkansas State University, AR, USA

The relationship between big biological data has become a hot research topic. It is recommended to use clustering methods to group similar data so that when a large amount of data is needed, all the data can be easily found close to the search results. Here we explore a popular method for clustering data, k-means clustering. Here, they implemented two versions of the k-means clustering algorithm. First, this algorithm is called Lloyd's k-means clustering algorithm. It is a relatively faster and understandable algorithm. Another version of k-means clustering we implemented is called the progressive greedy k-means clustering algorithm. This is a more conservative approach and can take much longer, but can sometimes produce better results than the previous one. These results are based on the running time and mean square error distortion of the algorithms and are compared to analyze the complexity and efficiency.

PROPOSED METHODOLOGY

The models are created using accident data to help understand the characteristics of many features such as driver behavior, driving conditions, lighting, weather conditions, etc. It can help users calculate safety measures that are useful to avoid accidents. How the statistical method is based on directed graphs can be illustrated by comparing two scenarios based on out-of-sample forecasts. The model identifies statistically significant factors that can be used to predict the probability of accidents and injuries that can be implemented and reduce the risk factor.
Here, the investigation of the traffic accident is carried out by analyzing some data, presenting some important questions from the point of view of the investigation. Questions like when is the most dangerous time to drive, how many accidents happen in the countryside, in the city and in other places. What is the trend in the number of accidents every year, are there more deaths in restricted areas and so on. The purpose of this analysis is to show the most important information from the point of view of the traffic accident and enable prediction. Following steps are involved while building the model:

I. Data Gathering

The dataset can have the following attributes –

- Latitude
- Longitude
- Age of Driver
- Weather Conditions
- Vehicle type/model
- Condition of Vehicle
- Time of accident
- Gender of Casualty
- Speed of colliding vehicle

II. Data Initialization

Data initialization consists of the following phases:

- Association: Data combination from various means is done.
- Altering: k-means clustering is used to alter data and classify into various clusters having similarities.
- Clipping: Data is selected based on requirements, and rest is used for analysis purpose.
- Categorizing: After successful application of the algorithm, clusters are categorized into various parts to process it further.

III. Data Training and Analysis

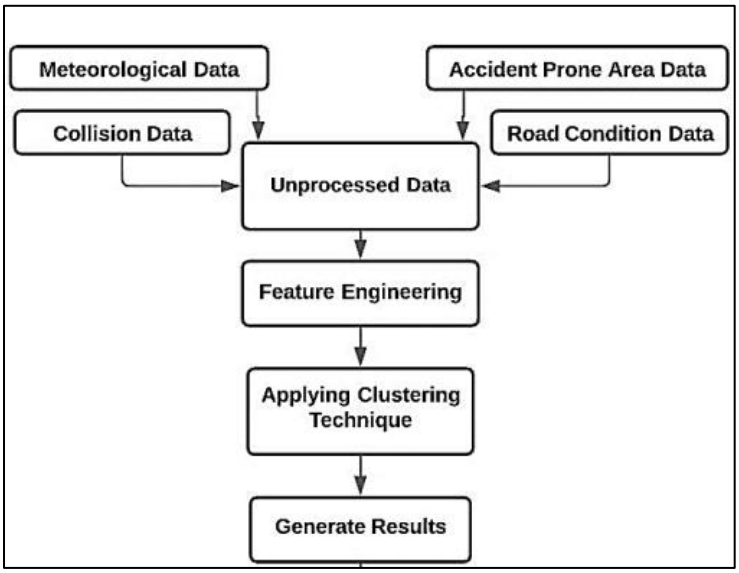
Initialized data is then used for training and analysis where 70% of data is used for training purpose and 30% for testing purpose. The model is compatible to changes in any circumstances which make it feasible to use for a very long duration.

IV. Testing

Raw data is taken into consideration and preprocessed with the help various machine learning algorithms like k-means clustering, decision trees, random forests etc. for acquiring a prediction model.

V. Output

After training and analysis, our model can effectively predict the severity of accident-prone areas based on past dataset of certain locations and classify them.



TECHNOLOGY USED

- Python
- Jupyter Notebook
- scikit-learn library
- XGBoost
- NumPy Library
- Pandas
- Matplotlib
- Seaborn Library
- HTML
- CSS
- JavaScript
- Flask

CONCLUSION

Using python, jupyter notebook and Scikit learn, pandas and matplot data science libraries, a work-flow is developed for processing the dataset and generate the corresponding accident severity prediction models. It is composed of several nodes, namely:

1) Dataset: contains the pre-processed data for the experiment.
2) Explore Data: is an optional node to help in data exploration and viewing some statistics about the data before modelling.
3) Model: contains the algorithms that will be used for model generation.
4) Apply: where the model is applied to the predictors to generate the required results.
5) Predictors: sample dataset for testing the prediction.
6) Prediction: the resulted table after applying the model on the predictors.

ML is a rapidly growing technique in the field of research and analytics. It helps to study and learn various data patterns and make accurate predictions. Generalization in ML is the key factor that is used in clustering to make a dataset function faster. In this study, we developed a prediction model to solve the problem of road accidents in India by taking all factors into consideration and minimum chance of deviation. It has been observed that the accident rate in India is quite high despite various strict measures. Our study can in turn benefit the society, which will help in analyzing hotspots and the cause of accidents so that they will not do the same thing which led to an accident at that place. The developed system is so simple and reliable that any user can easily follow and avoid mishaps.

REFERENCES

1. Dhaval Walavalkar, Jayesh Patil, Mandar Prabhu, Vivian Brian Lobo - Road Accident Analysis using Machine Learning
2. Arundhati Navada, Aamir Nizam Ansari, Balwant Sonkamble, Siddharth Patil - Overview of use of decision tree algorithms in machine learning
3. Amjad Ali, Muhammad Waseem , Sana Yaseen, Shafiq Hussain ,Umar Draz, Zaid Bin Faheem, Zanab Safdar.- A State of Art ML Based Clustering Algorithms for Data Mining
4. Liling Li Department of Computer Science, Central Michigan University, USA Gongzhu Hu Department of Computer Science, Central Michigan University, USA Sharad Shrestha Department of Computer Science, Central Michigan University, USA - Analysis of road traffic fatal accidents using data mining techniques
5. Ahmed M. Zeki Department of Information Systems, University of Bahrain, Manama, Bahrain Mona Al Hamad Department of Information Systems, University of Bahrain, Manama, Bahrain - Accuracy vs. Cost in Decision Trees: A Survey
6. Hui Xu, Shunyu Yao, Qianyun Li, Zhiwei Ye, Hubei University of Technology, No. 28, Nanli Road, Hong-shan District, Wuhan, China - Road Accident Analysis and Hotspot Prediction using Clustering
7. Nejdet Dogru International Burch University, Faculty of Engineering and Natural Sciences, Sarajevo, Bosnia and Herzegovina Abdulhamit Subasi Effat University, College of Engineering, Jeddah, Saudi Arabia - Traffic Accident Detection Using Random Forest Classifier
8. Mohamed K Nour College of Computer and Information Systems Umm Al-Qura University, Atif Naseer Science and Technology Unit Umm Al-Qura University, Basem Alkazemi College of Computer and Information Systems Umm Al-Qura University, Muhammad Abid Jamil College of Computer and Information Systems Umm Al-Qura University - Road Traffic Accidents Injury Data Analytics
9. Feng-Jen Yang Department of Computer Science, Florida Polytechnic University, Lakeland, Florida, USA - An Implementation of Naive Bayes Classifier
10. Gregory A. Wilkin, Xiuzhen Huang, Arkansas State University, AR, USA - K-Means Clustering Algorithms: Implementation and Comparison