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.

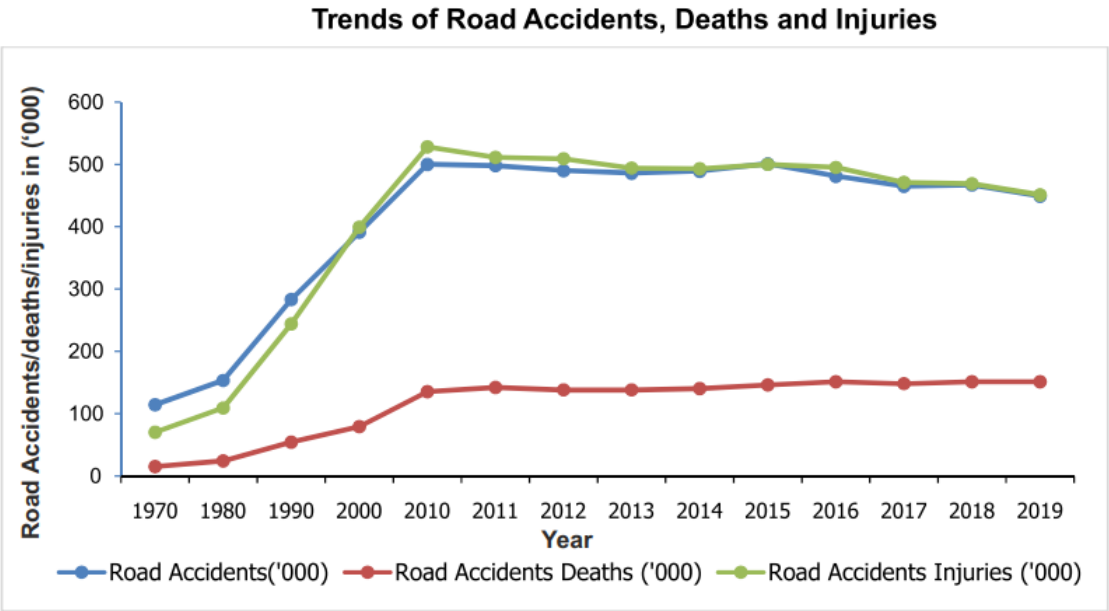
Road accidents have been the leading cause of deaths worldwide with the last three decades seeing a substantial increase in this regard. The WHO's World Report on Road Traffic Injury Prevention lists Road Accidents as the third leading contributor to the global burden of disease, up from ninth position in 19901. India's contribution in this regard is amongst the highest in the world with the country accounting for the second highest number of road accidents globally and the highest number of deaths. A total of 1,51,113 people were killed in India in 4,80,652 road accidents as against China whose figures of 63,093 deaths from 2,12,846 place it a distant second.

TABLE OF CONTENTS

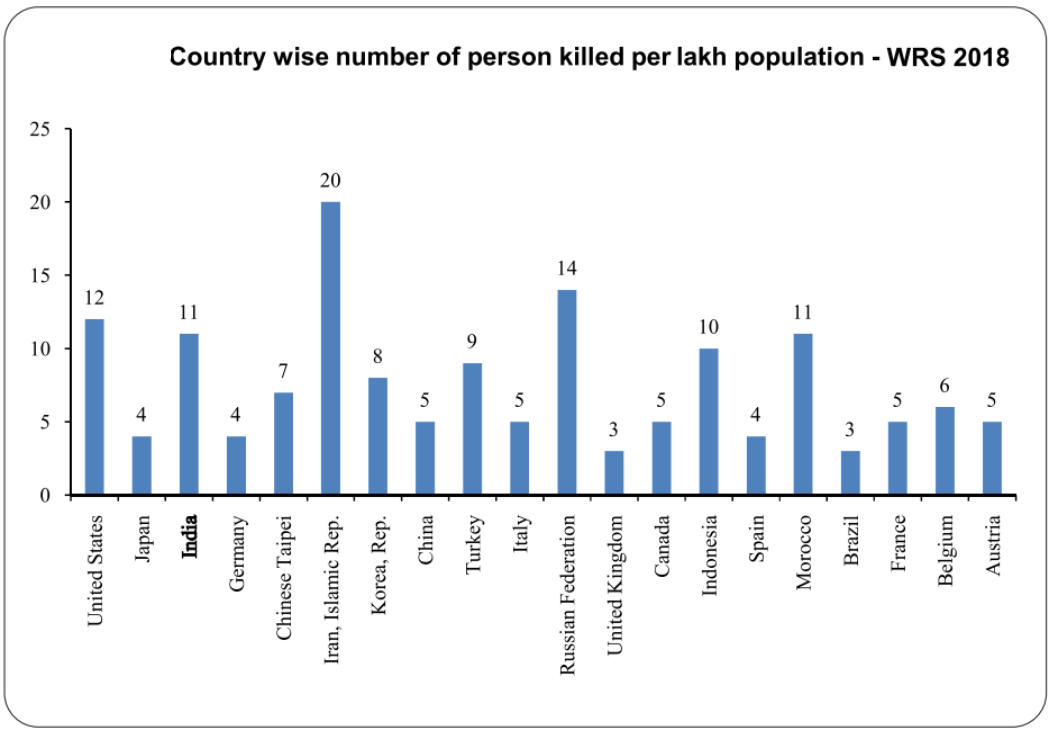
LIST OF FIGURES

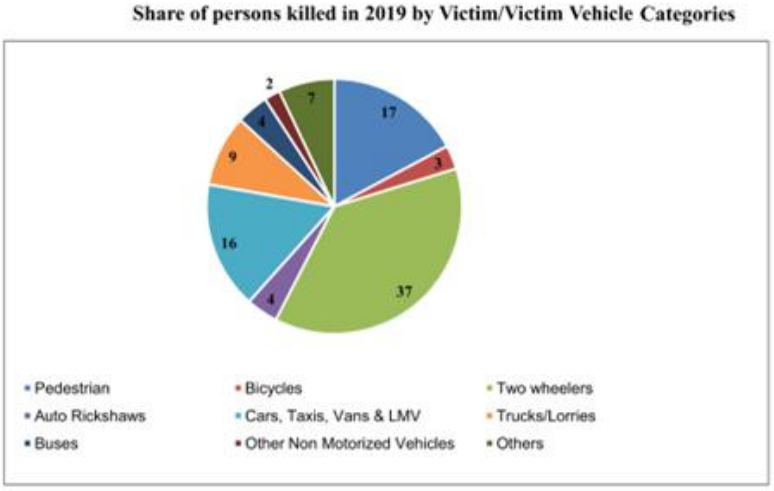
INTRODUCTION

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.
India's road network, besides being the lifeline of the Nation and a major contributor to socio-economic growth and development, also has the largest contribution to accidental deaths 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.



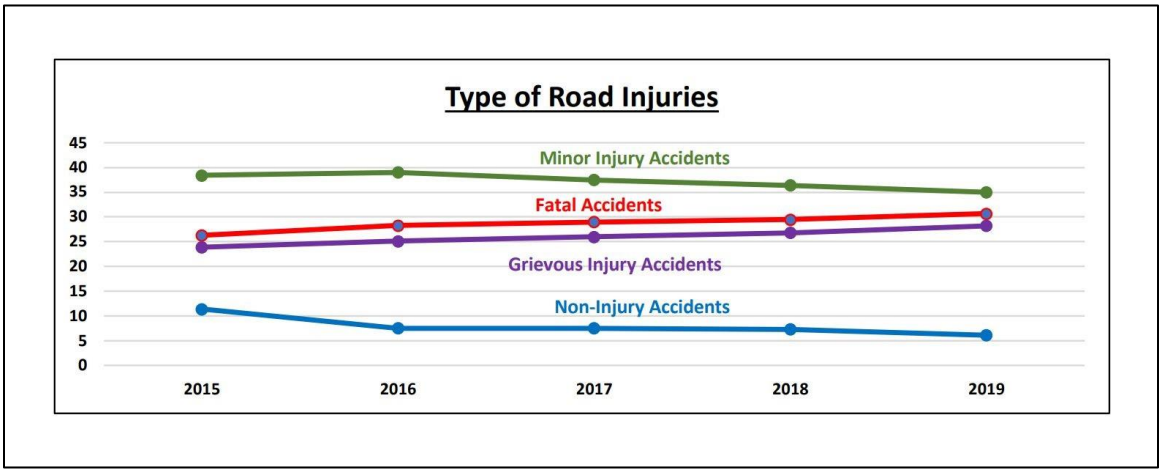
The only approach that can help decrease the number of road accidents, is to analyze the reasons that lead to these accidents. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. Machine learning (ML) is used to analyze various algorithms through experience and improve results. The concepts of Data Analysis, Data Visualization & Machine Learning help to tackle real world problems, by exploring & deriving valuable insights, which in turn help in taking measures to solve the targeted problem & predict 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

To handle the enormous number of road accidents in a locality a precise analysis is required. This analysis will be done more deeply to determine the intensity of the road accidents by using supervised learning techniques. This will classify the severity of the accidents as fatal, grievous, simple injury and motor collision. 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 its safety have been a major concern for the world, and everyone is trying to handle this since years. Road traffic and reckless driving occur in every part of the world. Because of this, many pedestrians are affected too. 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. This study aims to analyze road accidents in one of the popular metropolitan cities, i.e., Bengaluru, through k-means algorithm and machine learning by scrutinizing accident-prone or hotspot 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 (on input data patterns) and whose leaf nodes can be taken as categories (of these patterns). These tests are filtered down through the tree to get the right output to the input pattern. Decision Tree algorithms can be applied and used in various fields. It can be used as a replacement for statistical procedures to find data, to extract text, to find missing data in a class, to improve search engines and it also finds various applications in medical fields. Many Decision tree algorithms have been formulated. They have different accuracy and cost effectiveness. It is also very important for us to know which algorithm is best to use. The ID3 is one of the oldest Decision tree algorithms. It is very useful while making simple decision trees but as the complications increases its accuracy to make good Decision trees decreases. Hence IDA (intelligent decision tree algorithm) and C4.5 algorithms have been formulated.

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 to extract the insights and hidden relationships among data. Data mining has more importance in data science and machine learning because through data mining all hidden information is shown to determine various aspects of the data set. Clustering is a data mining technique to group the data, on the basis of similarity measures. The objects or data points in a cluster are similar. Similarly, objects or data points in another cluster will also be similar. But when these clusters are compared, they are dissimilar to each other. Clustering is considered the most important unsupervised learning technique because it deals with finding a structure in a collection of unlabeled data. Clustering can be done by the different approaches like partitioning clustering, hierarchical clustering, density-based clustering, and grid-based clustering. These clustering approaches can be done by the numbers of algorithms, such as K-means clustering, Fuzzy C-means clustering. K-means clustering is a cluster analysis method aimed at dividing n observation into groups in which each observation is closely related. The algorithm is called the k-means because it creates the number of similar clusters that we want, where the mean-value is placed at the center of the cluster. The algorithm is about finding the k-means of the desired data, which 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

Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In this paper, statistical analysis and data mining algorithms are applied on the FARS Fatal Accident dataset as an attempt to address this problem. The relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver were investigated. Association rules were discovered by Apriori algorithm, classification model was built by Naive Bayes classifier, and clusters were formed by simple K-means clustering algorithm. Certain safety driving suggestions were made based on statistics, association rules, classification model, and clusters obtained. From the clustering result, it was observed that some states/regions have higher fatal rate, while some others lower. We may pay more attention when driving within those risky states/regions. Through the task performed, data seems never to be enough to make a strong decision. If more data, like non-fatal accident data, weather data, mileage data, and so on, are available, more test could be performed thus more suggestion could be made from 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 applied widely for classification in many fields such as finance, marketing, engineering, and medicine. The increased field of application, made the requirement for understanding various aspects of decision trees in deep. In addition, it is crucial to understand the different type of costs associated with the classification task in a decision tree classifier and their relationship with the classifier's accuracy, as balancing the two is a major concern these days in many fields such as medical diagnosis. The focus of the paper is to review different type of costs consumed in building a DT, how to calculate them, different accuracy measures for evaluating the performance of a classifier, and the relationship between the classification accuracy and cost. Based on the survey made, the relationship between the classification accuracy and cost in DTs is found to be proportional. In addition, among different type of costs, most researchers focused on either test cost or on misclassification cost for constructing 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 prevailing clustering algorithms, K-means has become one among the foremost wide used technologies, principally as a result of its simplicity and effectiveness. However, the choice of the initial algorithmic centers and also the sensitivity to noise can scale back the clustering effect. To resolve these issues, this paper proposes an improved K-means algorithm. The idea of CLIQUE grid is employed to get rid of the noise and procure the regional density. Then the initial center point is selected according to the method of Fast Search and Find of Density Peaks (CFSFDP). Also, the impact of grid density error on initial center selection is mitigated by the granularity concept, while 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 article has higher accuracy, less difference in clustering effect on different data, and less parameter dependence

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 been growing in recent years with the enhancements in many totally different applications within the military, marine, intelligent transportation, smart health, and smart city domains. Though IoT brings vital advantages over traditional information and communication (ICT) technologies for Intelligent Transportation Systems (ITS), these applications are still rare. Although there is a continual improvement in road and vehicle safety, further as enhancements in IoT, road traffic accidents are increasing over the last decades. Therefore, it is necessary to search an effective idea to scale back the frequency and severity of traffic accidents. Hence, this paper presents an intelligent traffic accident detection system in which vehicles exchange their microscopic vehicle variables with each other. The planned system uses simulated knowledge collected from transport ad-hoc networks (VANETs) supported the speeds and coordinates of the vehicles and so, it sends traffic alerts to the drivers. Moreover, it shows how machine learning strategies are exploited to discover accidents on freeways in ITS. It is shown that if position and rate values of vehicle are given, vehicles' behavior can be analyzed and accidents can be detected easily. Supervised machine learning algorithms like Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Random Forests (RF) are enforced on traffic data to develop a model to differentiate accident cases from normal cases. The performance of RF algorithmic program, in terms of its accuracy, was found superior to ANN and SVM algorithms. RF algorithmic program has showed higher performance with 91.56% accuracy than SVM with 88.71% and ANN with 90.02% 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

Road safety researchers working on road accident data have witnessed success in road traffic accidents analysis through the application data analytic techniques, though, little progress was made into the prediction of road injury. This paper applies advanced data analytics methods to predict injury severity levels and evaluates their performance. The study uses predictive modelling techniques to identify risk and key factors that contributes to accident severity. The study uses publicly available data from UK department of transport that covers the period from 2005 to 2019. The paper presents an approach which is general enough so that can be applied to different data sets from other countries. The results identified that tree-based techniques such as XGBoost outperform regression-based ones, such as ANN. In addition to the paper, identifies interesting relationships and acknowledged issues related to quality of data.

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 target classifications that are used to classify a data set, different approaches might be chosen to perform the classification work. For binary classifications usually decision trees and support vector machines are commonly adopted, but these two approaches are under a constraint that the number of target classifications cannot go beyond two. This rigid constraint makes them hard to be generalized to fit a broad sense of real-life classification works in which the number of target classifications are usually more than two. From the standing point of acquiring a general tool kit, the Naive Bayes Classifier is more suitable for general classification expectations. As a mathematical classification approach, the Naive Bayes classifier involves a series of probabilistic computations for the purpose of finding the best-fitted classification for a given piece of data within a problem domain. In this paper, an implementation of Naive Bayes classifier is described. Bayesian-based probabilistic computations have been widely adopted to predict outcomes under uncertainties. By taking advantage of the powerful built-in programming constructs in Python programming language, this implementation of Naive Bayes classifier is done without much intensive coding. The main contribution of this implementation is to provide a general tool kit that can be applied in a big variety of classification domains. This classifier can be used as a general tool kit and applicable to various domains of classifications. To ensure the correctness of all probabilistic computations involved, a sample data set is selected to test this classifier

K-Means Clustering Algorithms: Implementation and Comparison

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

The relationship among the large amount of biological data has become a hot research topic. It is desirable to have clustering methods to group similar data together so that, when a lot of data is needed, all data are easily found near some search result. Here we study a popular method, k-means clustering, for data clustering. Here they have implemented two versions of the k-means clustering algorithm. First, an algorithm is called the Lloyd's k-means clustering algorithm. This is a relatively faster algorithm and is straight forward. The other version of k-means clustering that 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 yield better results than the former. These results will be based on the running time for the algorithms and the mean squared error distortion and will be compared for analysis of complexity and efficiency.

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 … These data can be accessed using Microsoft excel sheet and the required answer can be obtained. This analysis aims to highlight the data with the most importance in a road traffic accident and allow predictions to be made.

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

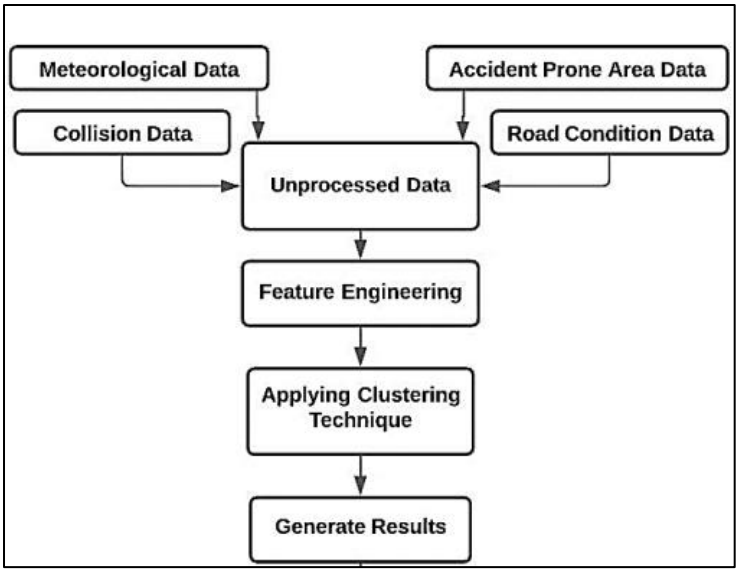
Initialized data is then used for training and analysis where 70% of data is used for training purpose and 30% for testing purpose. The model is compatible to changes in any circumstances which make it feasible to use for a very long duration.

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