ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Black Spot and Accidental Attributes Identification on State Highways and Ordinary District Roads Using Data Mining Techniques

Gagandeep Kaur
Department of Computer Engineering
Punjabi University
Patiala Punjab, India

Harpreet Kaur
Department of Computer Engineering
Punjabi University
Patiala Punjab, India

Abstract: Road traffic accident is a causative aspect and a particular instance of traumatic event that constitute major harm to life and property. Therefore vaticinating the cause of occurrence of accident related information on roads is necessary. Statistical and empirical analysis on State Highways (SH's) and Ordinary District Roads (ODR's) accidental datasets has been performed. The need of the study is to scrutinize the traffic related dataset through Exploratory Visualization Techniques, K-means and KNN Algorithms using RStudio. This paper present result by comparing the above said three mining techniques and predicts the cause of accident, accident prone location, analyze the time of accident, examine the reason of accident and anatomize the accused vehicle.

Keywords: Road traffic accident data, Black spot detection, Exploratory visualization techniques, K-means clustering, Nearest neighbor searching, Mining using RStudio.

I. Introduction

Road accidents cause complications which are increasing at alarming rate. Controlling the traffic accidents on roads is a crucial task. State Highways (SH's) and Ordinary District Roads (ODR's) form the economic background of the state. The importance of the study is to analyze the traffic accident data of State Highways (SH's) and Ordinary District Roads (ODR's) to predict the black spots and accidental attributes, factors to curb the menace caused by the accidents. The term accident black spot in management of road safety defines a place where accidents are been concentrated historically [2].The rationale behind investigation is to analyze the accidental data using exploratory visualization techniques and machine learning algorithms. These techniques and algorithms are applied on the Traffic accidental dataset to vaticinate the desired outcome in order to minimize the accident frequency.

The approaches are explained below:

1. Exploratory Visualization Technique: It is a technique to anatomize and examine the sets of data in order to abridge and encapsulate the crucial characteristics with visual and pictorial method. Exploratory Visualization analysis can be performed using scatter plot, correlation analysis, barplot, clustered barplot, histogram, pie chart etc.

2. Machine Learning: Machine learning concentrates on algorithm designing and make predictions on sets of data. It includes Supervised (KNN Algorithm) and Unsupervised learning (K-means Algorithm).

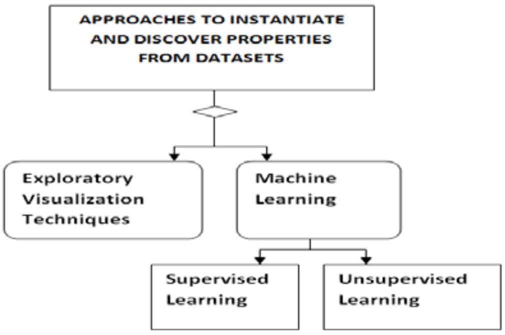

APPROACHES TO INSTANTIATE AND DISCOVER PROPERTIES FROM DATASETS
Exploratory Visualization Techniques
Machine Learning
Supervised Learning
Unsupervised Learning

Figure I. Representing approaches to discover properties from datasets

Fig. I shows the different approaches to extract the various parametric properties from bulk of dataset by using data harvesting techniques.

The purpose of the study is to predict the reason of accident, accident prone zone, time analyzation of accident, examine the accused vehicle using Exploratory visualization techniques,K-means which is famous clustering algorithm[5], KNN which is mostly used classification algorithm[7] and perform the comparative study of data mining techniques.

The paper has been organized into various sections. Section II describes the work related to the area of research. Section III illustrates the methodology used in performing the analysis. Section IV is related to the preprocessing of input file. Section V is related to the simulation and hence predicts the results. Section VI and VII deals with the conclusion and acknowledgement. Section VIII includes references.

II. Related Work

In order to group the data objects according to similarity reduce the mean squared distance of data point to its closest point. Clustering analysis is performed by K-means algorithm. In this study filtering algorithm called Lloyd's algorithm is studied. It is heuristic approach [4].

Classification is broad ranging research field in the data harvesting and mining which deals with many decision-making approaches for locating and identifying data. Typically the data is defined numerically through vector(x_1, x_2 ...xn) where *n* depicts number of features or attributes. Therefore, in an *n* dimensional space or area each piece of data can be treated as one point, and relates to one or more classes. Classification algorithms works in two steps: training and testing [8].

Cluster analysis is a depictive task and analytical approach in data harvesting that deals to identify homogenous group of object.K-mean is the most famous partitioned clustering approach [3].

Classification in data mining identifies the characteristics that indicate the group to which each case belongs. The pattern can be used for understanding the existing data and to predict how instances, features will behave in new space. Classification models in data mining analyze the already classified data (cases) and find a predictive and hidden pattern [7].

The study deals with the simulation of the K-Means algorithm. When dealing with huge volumes of set of data, algorithm tries to improve the running time by running through various iterations. The execution of K-Means consists of successive iterations on the datasets. Each iteration visit the whole data set in order to assign data objects to cluster on the basis of similarity [6].

Traffic safety management is the major task of the government. Pointing out the crucialness of the area, locating the reasons of road accidents has become the main aim to curb the harm caused by traffic accidents. On the other side, because of the alarming growth of data volume, investigation of the factors and attributes is required. Data mining outcome help various areas by exploring and predicting the future behavior and efficient decisions are taken to curb the road accidents [1].

The method adopted includes gathering the secondary data from concerned department, surveying the data physically and investigating the data through ranking method and severity index, accident density method, weighted severity index. Study depicts locating of black spots on national highway 4 (New Katraj Tunnel to Chandani Chowk) [2].

The aim and purpose of designing scientifically a road is to avoid the occurrence of accidents in order to curb severity. After the construction of a road, various measures should be taken to avoid the accidents. The possibility of accidents occurrence on road depends upon various attributes. In this paper, study related to the analysis of various factors affecting number of accidents for Guwahati city is done by gathering accidental data for various road stretches [10].

The problem of road accidents is enormously increasing in developing countries. The reason behind this is increased occupancy of vehicle. Accidents have been increasing over years. Controlling traffic on roads is a crucial task. Curbing accidents in accident prone locations is vital. The accident was increasing over a decade from 4% to 31%.The analysis and identification of accident prone areas is essential to curb the accidents which is a alarming issue [9].

Clustering is to unite data objects into different groups or clusters. It is most important task in data analysis, such as discovery of pattern, recognition of trends, summarization of data and image processing. K-means is a mostly used and well studied method in data mining for clustering the data objects on the basis of similarity [5].

III. Methodology Used

The Methodology is applied using R's Integrated Development Environment (Rstudio) which is graphical and statistical computing tool on road accidental dataset for analyzation. Accident Analysis through modeling is mostly used methodology adopted worldwide to examine the different reasons contributing to the increasing number of accidents [10].The Steps and Flow chart that describes the whole data mining process is as below:

1. Gathering of traffic accidental data.
2. Preprocessing of the traffic accidental dataset according to the algorithm requirement.
3. Required dataset is obtained on which desired computation are performed such as exploratory visualization techniques and machine learning approach.
4. After application of these approaches desired outcome is obtained.

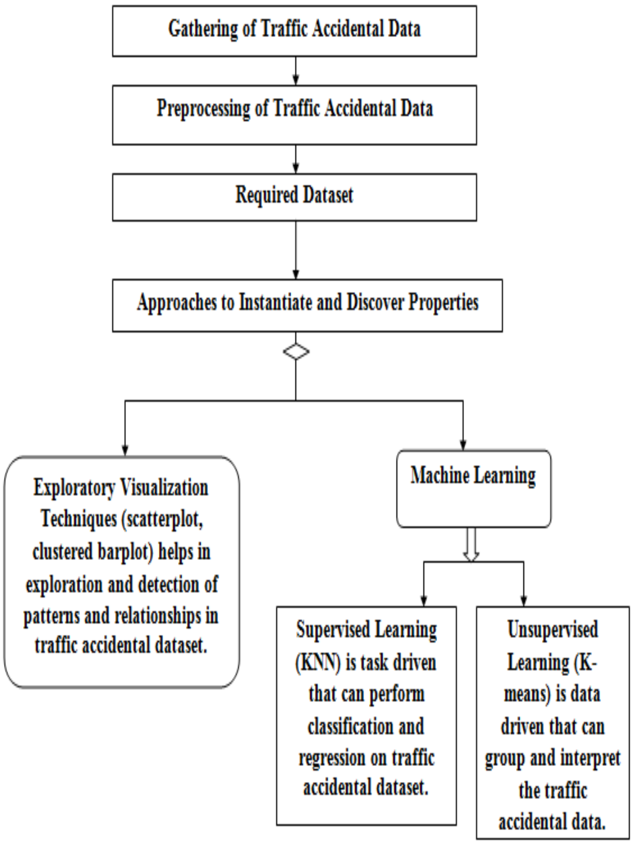
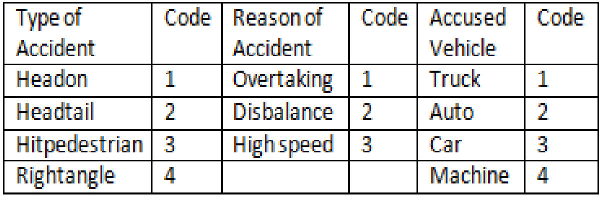

Gathering of Traffic Accidental Data
Preprocessing of Traffic Accidental Data
Required Dataset
Approaches to Instantiate and Discover Properties
Exploratory Visualization Techniques (scatterplot, clustered barplot) helps in exploration and detection of patterns and relationships in traffic accidental dataset.
Machine Learning
Supervised Learning (KNN) is task driven that can perform classification and regression on traffic accidental dataset.
Unsupervised Learning (K-means) is data driven that can group and interpret the traffic accidental data.

Figure II. Presenting steps of data mining of road and accidental data

Fig. II represents the data mining steps for the Road accidental data and analyzing using various data harvesting techniques and algorithms.

IV. Preprocessing of Input File

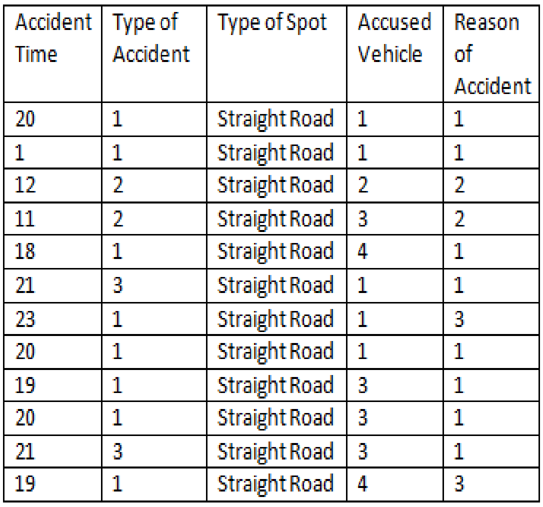
The input data file has been preprocessed. Data preprocessing is database driven application that prepares raw data for further processing. The alphanumeric data is converted into numerical form to get the desired output by using various data mining approaches.



Type of Accident	Code	Reason of Accident	Code	Accused Vehicle	Code
Headon	1	Overtaking	1	Truck	1
Headtail	2	Disbalance	2	Auto	2
Hitpedestrian	3	High speed	3	Car	3
Rightangle	4			Machine	4

Figure III. Displaying Code Assignment Notations

In Fig. III code has been assigned to the textual data which is used in Fig. IV. The data in textual form is converted in numerical form to get the required outcome.



Accident Time	Type of Accident	Type of Spot	Accused Vehicle	Reason of Accident
20	1	Straight Road	1	1
1	1	Straight Road	1	1
12	2	Straight Road	2	2
11	2	Straight Road	3	2
18	1	Straight Road	4	1
21	3	Straight Road	1	1
23	1	Straight Road	1	3
20	1	Straight Road	1	1
19	1	Straight Road	3	1
20	1	Straight Road	3	1
21	3	Straight Road	3	1
19	1	Straight Road	4	3

Figure IV. Presenting preprocessed data in the numerical form

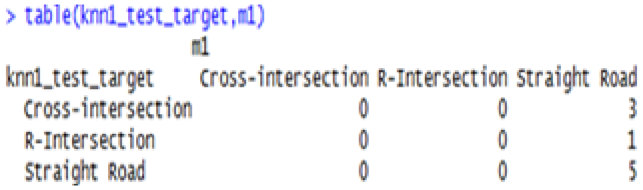
In Fig. IV the code has been assigned to the various parameters in the road accidental database and in the results and discussion section the various data mining techniques has been applied to get the desired result as well as outcome.

V. Results And Discussions

The simulation is performed by using RStudio which is an integrated development environment for R tool. Exploratory visualization techniques and data mining algorithms have been applied on the parameters of road accident data to analyze and predict the useful results which help to minimize the frequency of accidents. The total dataset was of 160 accidental data records. Simulation is performed on 40 records of each road.KNN algorithm provides the computation matrix and the outcome is given in the form of predictions. In K-means algorithm the dataset of 40 records has been divided into 3-clusters that are described with different color notation.

A. Simulation performed on State Highway Accidental data to identify and predict black spots

The simulation is performed on the State Highway namely Samana Patran Road and Patran Pehowa Road and the results of KNN Algorithm, K-means Algorithm and Exploratory visualization techniques are as below.



```
> table(knn1_test_target,m1)
                    m1
knn1_test_target     Cross-intersection R-Intersection Straight Road
  Cross-intersection                  0              0             3
  R-Intersection                      0              0             1
  Straight Road                       0              0             5
```

Figure V. Depicting Road I-SH Samana Patran Road-Prediction of Type of Spot through KNN algorithm

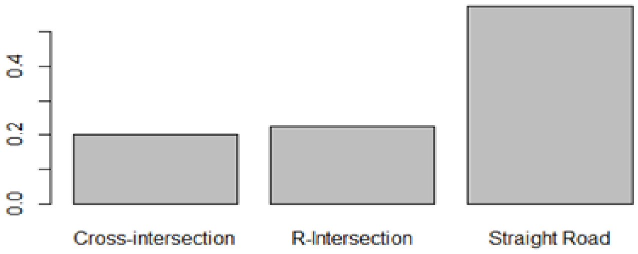
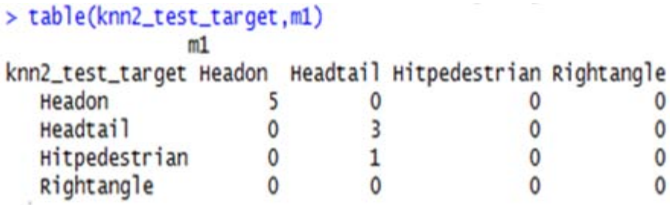


Figure VI. Representing Road I-SH Samana Patran Road-Prediction of Type of Spot through Exploratory Visualization Technique



```
> table(knn2_test_target,m1)
                 m1
knn2_test_target Headon  Headtail Hitpedestrian Rightangle
   Headon             5        0             0          0
   Headtail           0        3             0          0
   Hitpedestrian      0        1             0          0
   Rightangle         0        0             0          0
```

Figure VII. Showing Road II-SH Patiala Pehowa Road-Prediction of Accident Type through KNN algorithm

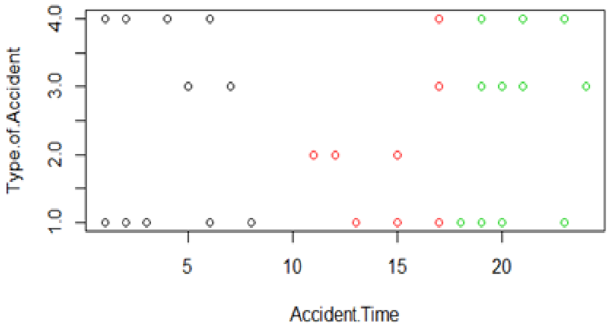
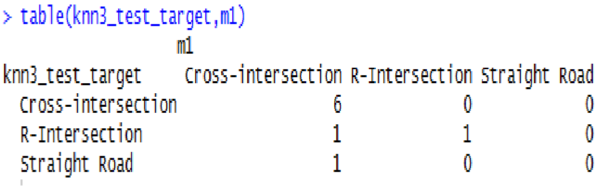


Figure VIII. Representing RoadI II-SH Patiala Pehowa Road-Prediction of Accident Type through K-means algorithm

In Fig. V, VI KNN algorithm and Exploratory visualization techniques are applied respectively, it is predicted that mostly on State Highways the accidents occur on Straight Roads. In Fig. VII, VIII KNN algorithm and K-means algorithm are applied respectively and it is predicted that mostly on State Highways the accident is of Head on collision type.

B. Simulation performed on Ordinary District Road Accidental data to identify and predict black spots

The simulation is performed on the Ordinary District Roads namely Patran Moonak Road , Khanouri Arno Road and the results of KNN Algorithm, K-means Algorithm and Exploratory visualization techniques are as below.



```
> table(knn3_test_target,m1)
                    m1
knn3_test_target     Cross-intersection R-Intersection Straight Road
  Cross-intersection                  6              0             0
  R-Intersection                      1              1             0
  Straight Road                       1              0             0
```

Figure IX. Depicting Road I-ODR Patran Moonak Road-Prediction of Type of Spot through KNN algorithm

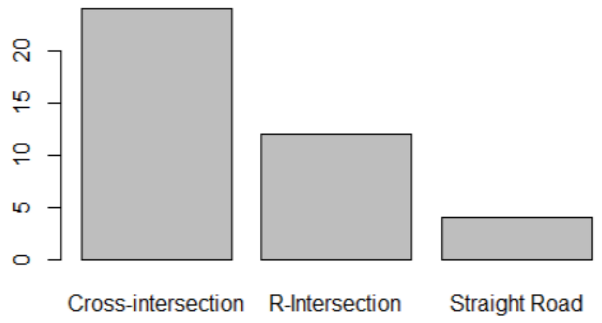


Figure X. Showing Road I-ODR Patran Moonak Road-Prediction of Type of Spot through Exploratory Visualization Technique

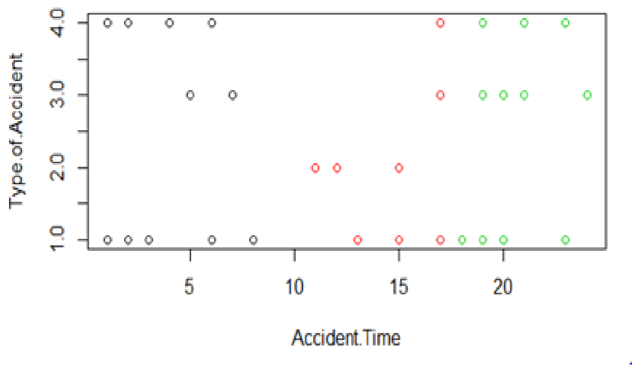


Figure XI. Depicting Road II-ODR Khanouri Arno Road- Prediction of Accident Type through K-means algorithm

In Fig. IX, X KNN algorithm and Exploratory visualization techniques are applied respectively, it is predicted that mostly on Ordinary District Roads the accidents occur majorly on Cross intersection. In Fig. XI K-means algorithm is applied and it is predicted that mostly on Ordinary District Roads the accident is of Head on collision type.

C. *Parametric analyzation using Exploratory visualization techniques on State Highways.*

The parametric analysis on various parameters are applied on road accidental data using Exploratory visualization techniques. The outcomes of the results are as below.

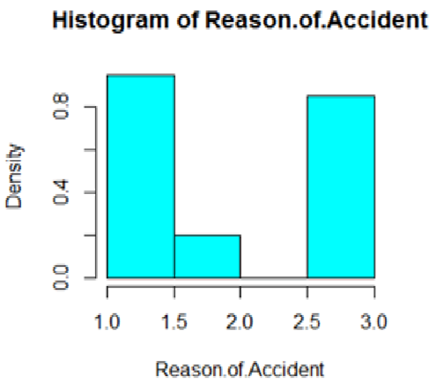


Figure XII. Depicting Road I-SH Samana Patran Road-Showing statistical information of Reason of Accident through Exploratory Visualization Technique

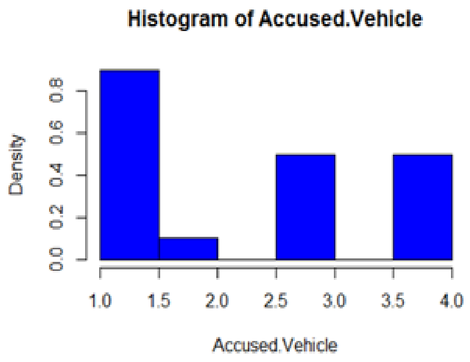
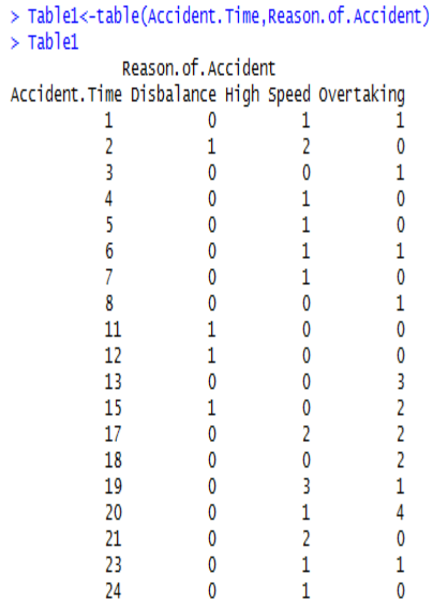


Figure XIII. Showing Road I-SH Samana Patran Road-Displaying frequency of Accused Vehicle through Exploratory Visualization Technique



```
> Table1<-table(Accident.Time,Reason.of.Accident)
> Table1
             Reason.of.Accident
Accident.Time Disbalance High Speed Overtaking
           1          0          1          1
           2          1          2          0
           3          0          0          1
           4          0          1          0
           5          0          1          0
           6          0          1          1
           7          0          1          0
           8          0          0          1
          11          1          0          0
          12          1          0          0
          13          0          0          3
          15          1          0          2
          17          0          2          2
          18          0          0          2
          19          0          3          1
          20          0          1          4
          21          0          2          0
          23          0          1          1
          24          0          1          0
```

Figure XIV. Representing Road I-SH Samana Patran Road-Representing Reason of Accident vs Accident Time through Exploratory Visualization Technique

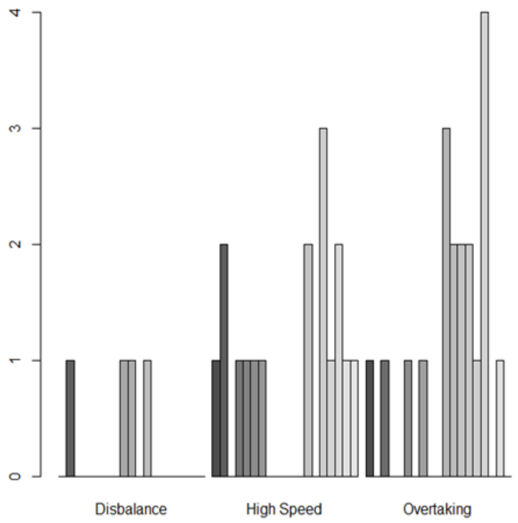


Figure XV. Displaying Output of (XIV) in the form of clustered barplot

In Fig. XII showing statistical information related to Reason of Accident through Exploratory Visualization Technique. Accidents occur frequently due to Overtaking, less frequently due to High speed and minimum due to disbalance. Fig. XIII displaying frequency of Accused Vehicle through Exploratory Visualization Technique that mainly accidents occur most constantly with trucks and less with other type of vehicles such as cars, autos, machines. Figure XIV, XV shows mostly accidents occur at night time due to overtaking of vehicles and automobiles.

D. Parametric analyzation using Exploratory visualization techniques on Ordinary District Roads.

The parametric analysis on various parameters are applied on Ordinary District Road accidental data using exploratory visualization technique. The outcomes of the results are as below.

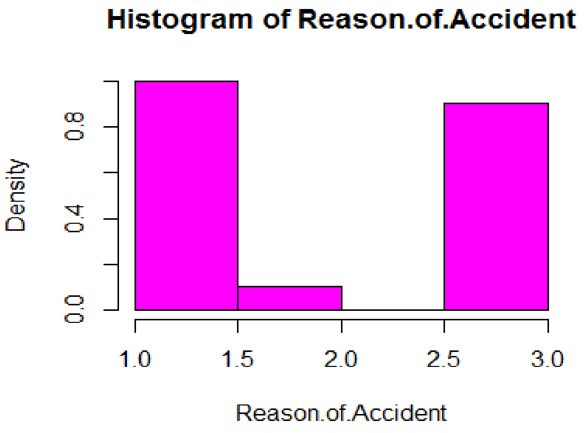


Figure XVI. Depicting Road I-ODR Patran Moonk Road-Showing statistical information of Reason of Accident through Exploratory Visualization Technique

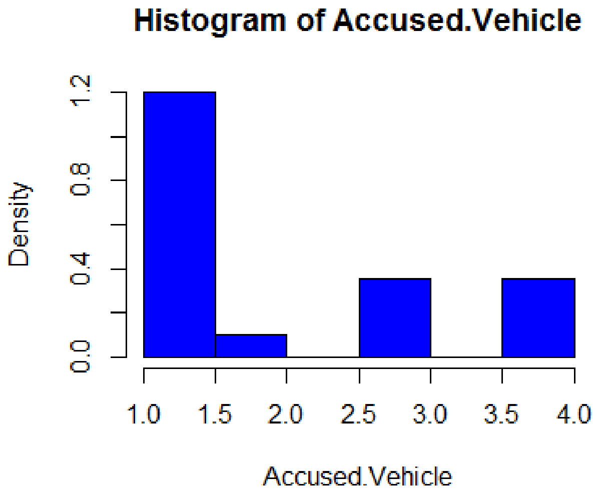
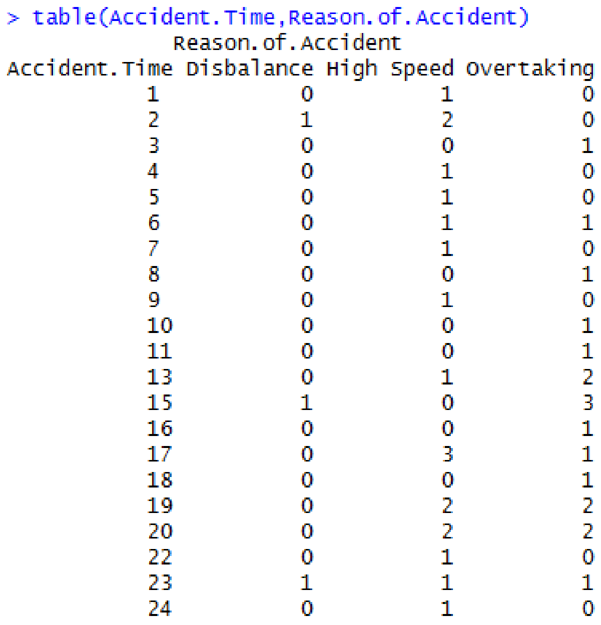


Figure XVII. Showing Road I-ODR Patran Moonak Road-Displaying frequency of Accused Vehicle through Exploratory Visualization Technique



```
> table(Accident.Time,Reason.of.Accident)
             Reason.of.Accident
Accident.Time Disbalance High Speed Overtaking
           1           0          1          0
           2           1          2          0
           3           0          0          1
           4           0          1          0
           5           0          1          0
           6           0          1          1
           7           0          1          0
           8           0          0          1
           9           0          1          0
           10          0          0          1
           11          0          0          1
           13          0          1          2
           15          1          0          3
           16          0          0          1
           17          0          3          1
           18          0          0          1
           19          0          2          2
           20          0          2          2
           22          0          1          0
           23          1          1          1
           24          0          1          0
```

Figure XVIII. Presenting Road I-ODR Patran Moonak Road-Representing Reason of Accident vs Accident Time through Exploratory Visualization Technique

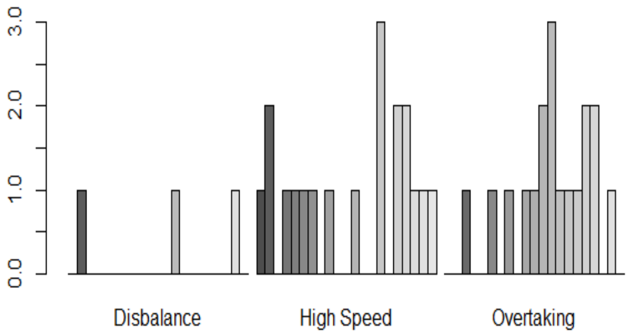
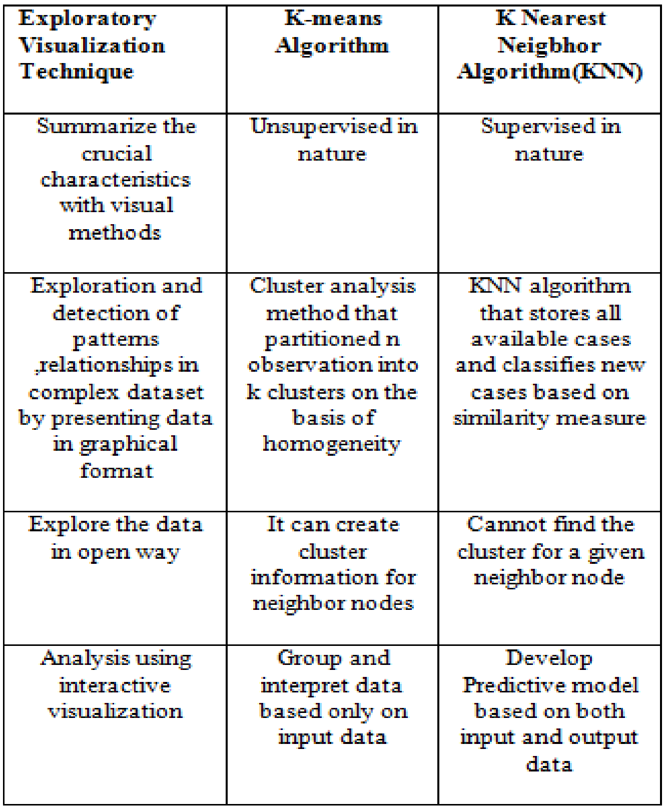


Figure XIX. Displaying Output of (XVIII) in the form of clustered barplot

In Fig. XVI showing statistical information of Reason of Accident through Exploratory visualization technique that accidents occur frequently due to Overtaking, less frequently due to High speed and minimum due to Disbalance.Fig. XVII displays frequency of Accused Vehicle through Exploratory visualization technique that mainly accidents occur with trucks and less with other type of vehicles such as cars, autos, machines. Fig. XVIII, XIX shows mostly accidents occur at night time due to overtaking of vehicles and high speed.

Table I. shows the comparison of different Data harvesting techniques. Exploratory visualization technique explores and summarizes the hidden patterns, trends, relationships in the graphical format. K-means algorithm creates clusters on the basis of homogeneity.KNN algorithm gather all the cases and classifies them.

Table I. Comparison of the various Data Mining Techniques on the basis of analysis



Exploratory Visualization Technique	K-means Algorithm	K Nearest Neigbhor Algorithm(KNN)
Summarize the crucial characteristics with visual methods	Unsupervised in nature	Supervised in nature
Exploration and detection of patterns ,relationships in complex dataset by presenting data in graphical format	Cluster analysis method that partitioned n observation into k clusters on the basis of homogeneity	KNN algorithm that stores all available cases and classifies new cases based on similarity measure
Explore the data in open way	It can create cluster information for neighbor nodes	Cannot find the cluster for a given neighbor node
Analysis using interactive visualization	Group and interpret data based only on input data	Develop Predictive model based on both input and output data

VI. CONCLUSION

The study helps us to derive the statistical model by using data mining algorithms and exploratory visualization techniques. From the study it has been concluded from all the techniques implemented that most of the accidents on State Highways occur mostly on Straight Road and Ordinary District Roads majorly occur on Cross intersection. The accidents are of Head of collision type.It has been observed that mostly accidents occur at night time, maximum due to overtaking of vehicles and high speed of vehicles and less due to disbalance of automobiles. The accused vehicle analyzed is truck. Future work will be to analyze and investigate the cause of severity of accidents by considering other parameters such as pavement design and condition, horizontal curves, shoulder width etc.

VII. ACKNOWLEDGMENT

We are really thankful to Executive Engineer Construction Division PWD B&R that help in gathering crucial dataset of time period (2012-2015) and Er.Harpreet Kaur, Assistant Professor Punjabi University, Patiala for providing assistance in dealing with Data mining using R tool as well as for comments that greatly improved the manuscript. We would also like to show our gratitude to them for sharing their pearls of wisdom with us during the course of this research.

VIII. REFERENCES

[1] Somayya Ebrahimkhani, Bahram Sadeghi Begham Farzaneh Moradkhani, "Road Accident Data Analysis: A Data Mining Approach," *Indian Journal of Scientific Research*, May 2014.

[2] R.P. Kulkarni, S.U. Bobade, M.S. Patil, A.M. Talathi, I.Y. Sayyad, S.V.Apte R.R.Sorate, "Identification of Accident Black Spots on National Highway 4 (New Katraj Tunnel to Chandani Chowk)," vol. 12, no. 3, pp. 61-67, May. - Jun. 2015.

[3] Monika Sharma Jyoti Yadav, "A Review of K-mean Algorithm," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 4, no. 7, pp. 2972-2976, July 2013.

[4] Nathan S. Netanyahu,Angela Y. Wu Tapas Kanungo, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation ," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, July 2002.

[5] Yanwei Yu, Lihong Wang,and Jinglei Liu Jianpeng Qi, "K*-Means: An Effective and Efficient K-means Clustering Algorithm," in *International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking*, 2016, pp. 242-249.

[6] Marian Cristian Mihˇaescu,Mihai Mocanu Cosmin Marian Poteras, "An Optimized Version of the K-Means Clustering," in *Federated Conference on Computer Science and Information Systems*, 2014, pp. 695–699.

[7] Mohammad Bolandraftar Sadegh Bafandeh Imandoust, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *Int. Journal of Engineering Research and Applications*, vol. 3, pp. 605-610, Sep-Oct 2013.

[8] Latifur Khan, Bhavani Thuraisingham Lei Wang, "An Effective Evidence Theory based K-nearest Neighbor (KNN) classification," in *International Conference on Web Intelligence and Intelligent Agent Technology*, 2008, pp. 797-801.

[9] R. Srinivasa Rao Dr. NSSR Murthy, "Development of model for road accidents based on intersection parameters using regression models," *International Journal of Scientific and Research Publications*, vol. 5, no. 1, pp. 1-8, January 2015.

[10] Sarbajit Bhattacharyya , Mrinal Roy , Pinak Paul Rupanjan Chakraborty, "Accident Analysis and the Suggestion of an Accident Prediction Model for Guwahati city," *International Journal of Innovative Research in Science*, vol. 4, no. 11, pp. 10774-10782, November 2015.