Abstract

Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. The main objective of this project is to identify the clustering of the road accident based on the demographic and geographic. The scope of this project is limited to Trivandrum city only. In India the road safety is the major concern in any area. Recently the data published by City police Trivandrum is approximately 5200 road accident reported till Sept 2019. Based on historical data collected from government office as well as real time basis, we have analyzed it to identify the cluster of road accident (RT) cases which took place in Trivandrum City during Jan 2019-Dec 2019. It helps to implement good practices w.r.t traffic management which helped to reduce road accidents.

The relationship between fatal rate and other attributes including collision manner, accident type, surface condition, light condition, and types of vehicles involved in the accidents 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.

Contents

List of Figures

INTRODUCTION

There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time anywhere. Some accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out how to drive safer, data mining technique could be applied on the road accident dataset to find out some valuable information, thus give driving suggestion.

Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered as one of the most important tool in information technology in the previous decades. Association rule mining algorithm is a popular methodology to identify the significant relations between the data stored in large database and also plays a very important role in frequent item set mining. A classical association rule mining method is the Apriori algorithm whose main task is to find frequent item sets, which is the method we use to analyse the roadway traffic data. Classification in data mining methodology aims at constructing a model (classifier) from a training data set that can be used to classify records of unknown class labels. The Naïve Bayes technique is one of the very basic probability-based methods for classification that is based on the Bayes' hypothesis with the presumption of independence between each pair of variables.

We used the historical data collected from government office as well as real time basis and we have analyzed it to identify the cluster of road accident (RT) cases which took place in Trivandrum City during Jan 2019-Dec 2019. The Accidents Dataset contains all fatal accidents on public roads in 2019 reported by City police Trivandrum.

Names of Selected Classifiers:

1. Association rule mining
2. Naïve Bias
3. KNN

Methodology

The approach we took for our study follows the traditional data analysis steps, as shown in Fig 1.

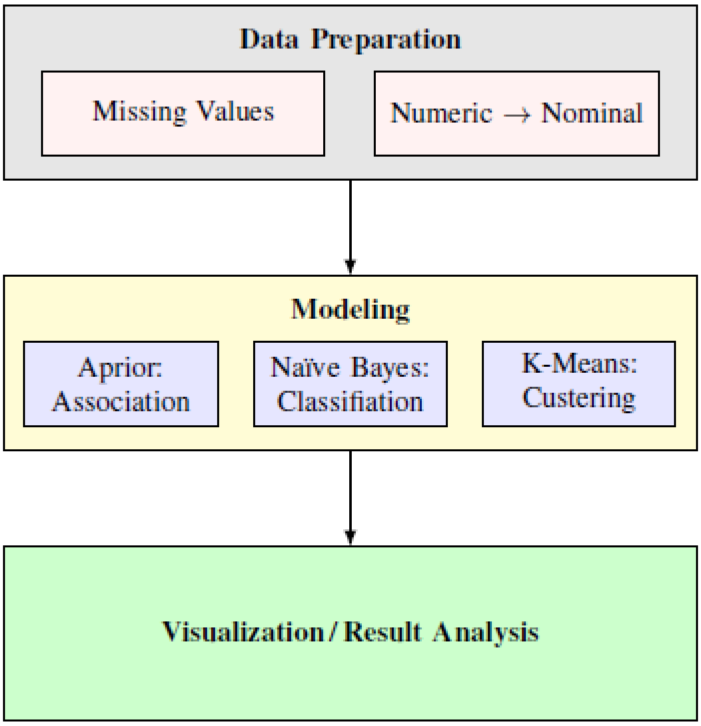


Fig. 1. Work flow

A. *Data Preparation*

Data preparation was performed before each model construction. All records with missing value in the chosen attributes were removed. All numerical values were converted to nominal values. Fatal rate were calculated and binned to two categories: high and low. Several variables are calculated from other independent variables.

B. *Modelling*

We first calculated several statistics from the dataset to show the basic characteristics of the fatal accidents. We then applied association rule mining, clustering, and Naïve Bayes classification to find relationships among the attributes and the patterns.

C. *Result Analysis*

The results of our analysis include association rules among the variables, clustering of regions in Trivandrum on the number of fatal accidents, and classification of the regions as being high or low risk of fatal accident. We used python to perform these analysis.

Experimental Results

A. Statistics results

The number of fatal accident in each region are shown in Fig 2. The most fatal accidents happened at Thiruvallam and the least at Nedumangad.

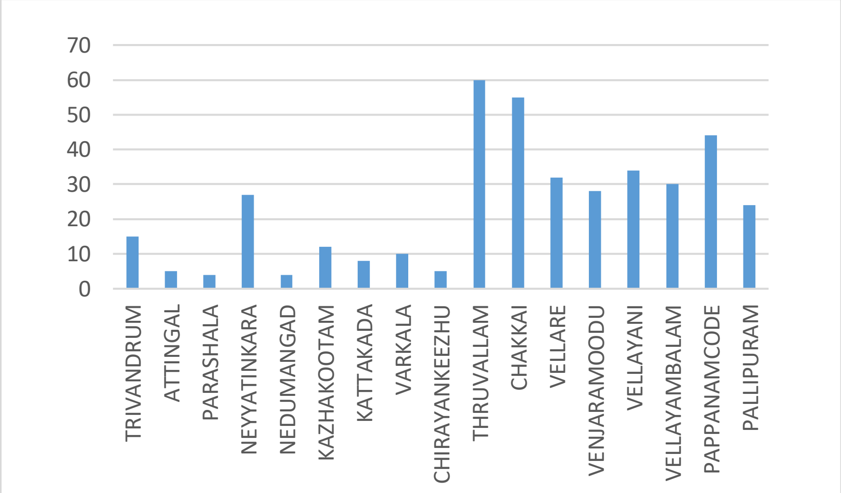
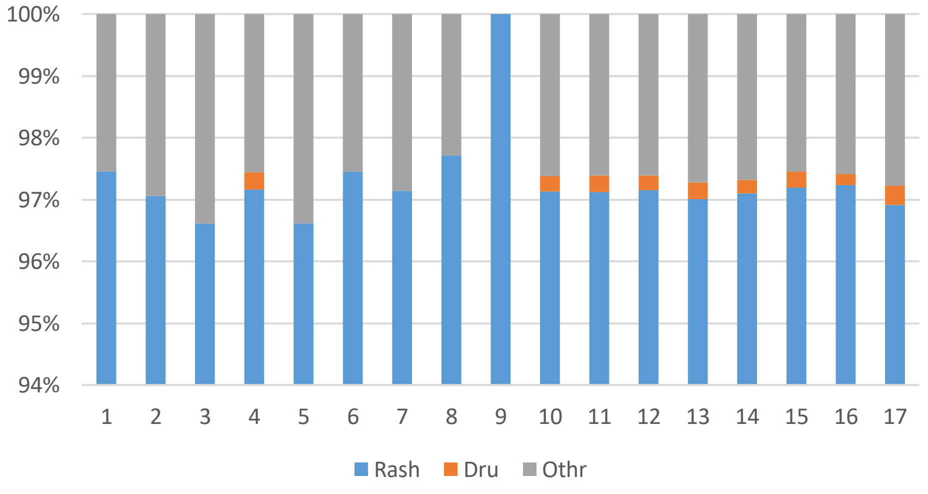


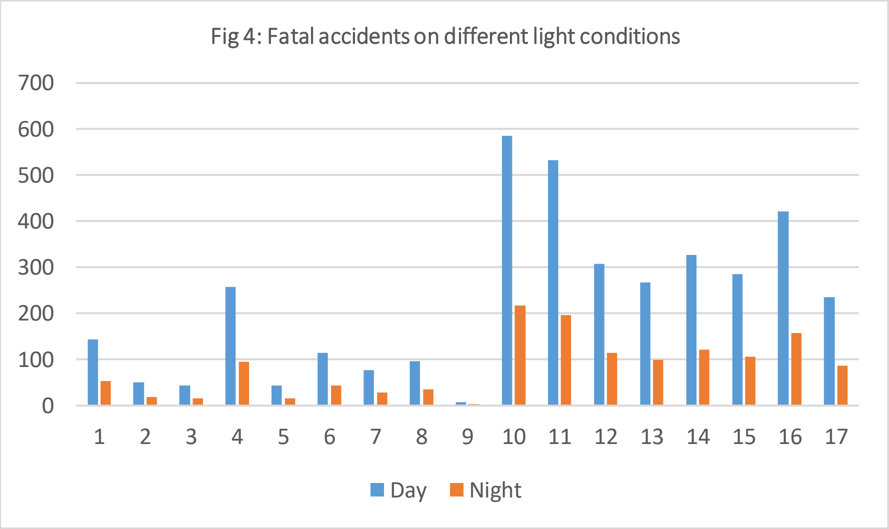
Fig 2: Fatal accidents per region

1) *<u>Collision Manner</u>:* The number of fatal accidents happened on different collision manners in comparison of districts and fatal involved are shown in Fig 3. Surprisingly, the most fatal accidents are not in collision while drunken drive. The percentage of accidents in different regions and fatal involved due to rash driving are much higher than the percentage of accident number due to other reasons, which reveals that Thiruvallam has higher fatal rate in a fatal accidents due to rash driving.

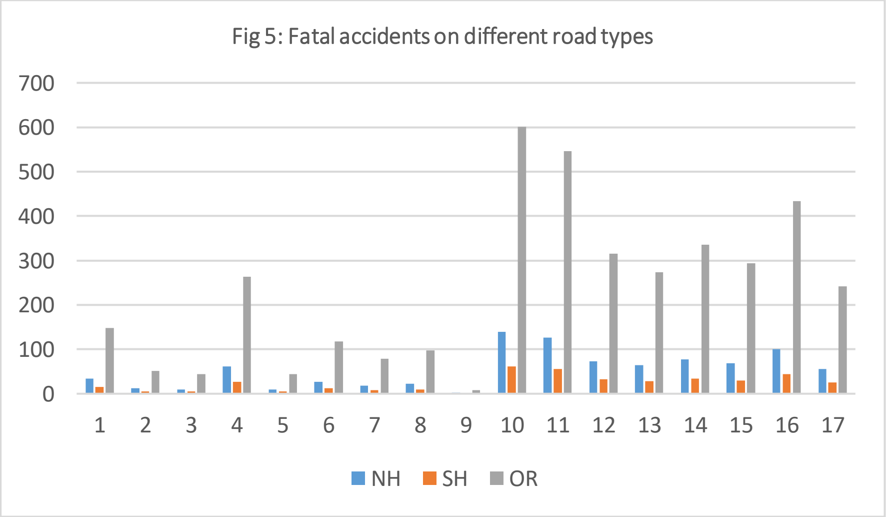
Fig 3: Fatal accidents on different collision manners



2) *Light Condition*: The number of fatal accidents happened on different light condition in comparison of regions and fatal involved are shown in Fig 4. Unsurprisingly, most fatal accidents happen in day light condition because much more roadway traffic happens in day time other than at night.



3) *Road Types*: The number of fatal accidents happened on different road types in comparison of districts and fatal involved are shown in Fig 5. Most fatal accidents happen in road types other than National highways and State Highways. This is understandable because the most usual case of road condition is that the road surfaces are terribly broken.



4) *Roadway Surface Condition:* The percentage of fatal accident happened on different types of vehicles involved is shown in Fig 6. Most fatal accidents happened between two-wheelers included. It is followed by car and auto included, as Thiruvallam the most and secondly Chakkai.

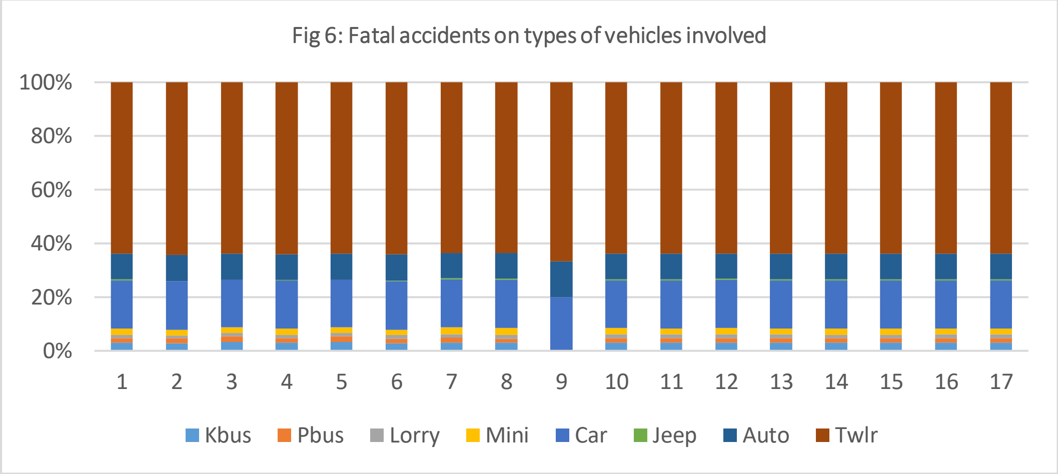


Fig 6: Fatal accidents on types of vehicles involved

Most fatal happened in short time and there is no significant relation between the rescue (like ambulance, police etc.) arrival time and fatal rate (correlation=0.1132231). The minimum arrival time is 0 minute, which either because the rescue services was at scene or data entry error. The average time take for the rescue arrival is 18.27 minutes, the median is 10 minutes, and the maximum time is over 18 hours.

We need to mention that all the data is about fatal accident, so no matter how long would it take for the rescue to arrive, there would always be fatal happening. Also, there is no variable recording at what time the death happened, and a lot of records are missing value at time, so very limited information could be inferred from the time relevant attributes. Similar statistics also performed on other attributes and the results are not significant so are not included here. After performing basic statistics and related work research, four attributes (collision manner, types of road, light condition, types of vehicles included) are considered and selected as affecting fatal rate.

B. Association Rule Mining

Before applying the algorithms, the tuples with missing value in chosen attributes were removed, the numerical values were converted to nominal values. The clean data was stored in CSV format and ready to be analysed.

A small partial sample of the dataset is shown in Table I. All values were converted to nominal values.

After applying Apriori algorithm with minimum support = 0.4 and minimum confidence = 0.6, association rules with fatal rate at the right side as decision were generated. The best 13 rules are shown in Table II.

We could see that fatal accidents involving rash driving have higher fatal rate, which means rash driving is much more dangerous than others. Also the two-wheelers in the day light has high fatal rate, this reveals that not only the accident percentage is higher, as shown in basic statistics, but also the fatal rate are high (with confidence level = 0.65).

TABLE I

CLEANED DATA FOR ASSOCIATION RULE MINING AND CLASSIFICATION

Light Type	Road Type	Types of Vehicles Involved	Collision Manner	Fatal Rate
day	NH	Twlr	Rash	high
night	OR	Auto	Dru	safe
night	NH	Kbus	Dru	low
day	SH	Twlr	Rash	high
night	OR	Auto	Othr	low
day	SH	Twlr	Rash	high
day	OR	Kbus	Othr	low
day	OR	Lorry	Dru	high
night	OR	Twlr	Rash	low
night	SH	Car	Othr	high
night	NH	Twlr	Rash	low
night	NH	Auto	Othr	safe

TABLE II

THIRTEEN ASSOCIATION RULES WITH HIGHEST CONFERENCE DISCOVERED BY APRIORI ALGORITHM

COL MAN=Rash	==⇒	Rate=high,	conf:(0.73)
LIGHT TYPE=day	==⇒	Rate=high,	conf:(0.68)
ROAD TYPE=OR	==⇒	Rate=high,	conf:(0.68)
LIGHT TYPE=day, VEH =Twlr	==⇒	Rate=high,	conf:(0.68)
LIGHT TYPE=day, COL MAN=Othr	==⇒	Rate=high,	conf:(0.66)
ROAD TYPE =OR, LIGHT TYPE =day, COL MAN =Dru	==⇒	Rate=high,	conf:(0.66)
COL MAN =Rash, VEH =Auto	==⇒	Rate=high,	conf:(0.66)
LIGHT TYPE =day	==⇒	Rate=high,	conf:(0.65)
LIGHT TYPE =day, ROAD TYPE=OR	==⇒	Rate=high,	conf:(0.65)
COL MAN=Rash , VEH =Twlr	==⇒	Rate=high,	conf:(0.65)
LIGHT TYPE =day, ROAD TYPE =OR , COL MAN =Rash	==⇒	Rate=high,	conf:(0.65)
VEH =Twlr	==⇒	Rate=high,	conf:(0.65)
ROAD TYPE =OR , COL MAN =Rash	==⇒	Rate=high,	conf:(0.63)

C. Classification

Naïve Bayes classifier was built on the cleaned data. Of the total records, the correctly classified are giving a 67.95% accuracy rate. The various evaluation measures are given in Table III.

The Naive Bayes Classifier shows that the fatal rate does not strongly depend on the given attributes, although they are considered as features in comparison to other attributes in the dataset.

TABLE III

RESULTS OF THE NA¨IVE BAYES CLASSIFICATION

	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area	Class
	0.996	0.996	0.681	0.996	0.809	0.561	High
	0.004	0.004	0.342	0.004	0.009	0.561	Low
Weighted Avg.	0.679	0.679	0.573	0.679	0.553	0.561	

D. Clustering of Regions

To find out which regions are similar to each other considering fatal rate, and which regions are safer or more risky to drive, clustering algorithm was performed on the fatal accident dataset.

To perform the clustering, total number of fatality per region was calculated. Also the population data for each region of Trivandrum district in 2019 was obtained from City police Trivandrum. With the fatal accident and the population dataset, fatalities per ten thousand people in the region was calculated. This allowed us to compare relative fatal rate in a region regardless of population of the region.

The K-means algorithm with Euclidean distance as the dissimilarity measure was applied to the data with two variables: population (in 10,000) and number of fatal accidents. The regions were grouped into 3 clusters as shown in Fig 7.

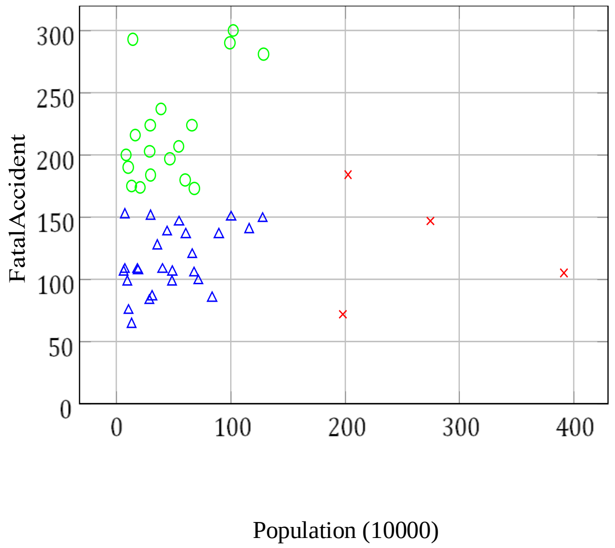


Fig. 7. Clusters of regions by fatality of per ten thousand people in a region

The three clusters are:

- Cluster A (blue): Those regions in cluster A represents the safe state with relatively lower fatal rate per ten thousand people.
- Cluster B (green): Cluster B which had relatively higher fatal rate.
- Cluster C (red): These regions have relatively large population and lower fatal rate. They are considered safe driving region and also outliers.

After careful observation it was found that none of the regions from northern part like Attingal, Nedumangad etc. lied on cluster A, and almost all the states from south were in cluster B. Only two states from the north were located in region of cluster A like Venjaramood, Vellayambalam etc., which is considered to be safe. Thiruvallam had highest fatal rate per ten thousand people, as well as had the highest number of per ten thousand people involved in fatal accidents. But Chakkai also had as much people involved in fatal accident as Thiruvallam.

This means that south (Thiruvallam) is much more risky compared to rest of the regions. North-east (Nedumangad) is the safest region and followed by north-west (Kattakada).

Algorithms Used

Naïve base classifier:

This classifier is a powerful probabilistic representation, and its use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real world datasets because it can give better predictive accuracy.

Association Rule Mining:

Association Rule Mining is one of the ways to find patterns in data. It finds features (dimensions) which occur together and features (dimensions) which are “correlated”. It is a rule-based machine learning method for discovering interesting associations and relationships among large sets of data items. This rule shows how frequently an item set occurs in a transaction. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Many algorithms for generating association rules have been proposed. Some well-known algorithms are Apriori, Eclat and FP-Growth.

K-Nearest Neighbour:

This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance.

CONCLUSION

As seen in statistics, association rule mining, and the classification, the environmental factors like roadway surface and light condition do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have stronger effect on the fatal rate. From the clustering result we could see that some regions have higher fatal rate, while some others lower. We may pay more attention when driving within those risky regions. Through the task performed, we realized that 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.

After analysing the quantitative data generated from the computer simulations, moreover their performance is closely competitive showing slight difference. So, more experiments on several other datasets need to be considered to draw a more general conclusion on the comparative performance of the classifiers.

REFERENCE

- https://www.researchgate.net/publication/318126696_Analysis_of_road_traffic_fatal_accidents_using_data_mining_techniques

- https://www.pantechsolutions.net/road-accident-analysis-using-machine-learning

- https://github.com/Small-Start/RoadSafety/blob/master/README.md

- https://pypi.python.org/pypi/numpy.html

- https://matplotlib.org/

- https://www.geeksforgeeks.org/

- http://pandas.pydata.org.html