Social Impact of Efficient Data Analytics to Prevent the Rate of Accidents

Guided by Sowmya BJ, *Assistant Professor, MS Ramaiah Institute of Technology, Bangalore, India*

Abhinav Shikhar, Aishwarya Kamraj, Gowri Ramaprasad, Naveen JS

Students, MS Ramaiah Institute of Technology , Bangalore, India

Abstract: **The International Traffic Safety Data & Analysis Group (IRTAD)[19] conducts surveys generating massive datasets with millions of data points. The rich dataset contains detailed information of approximately 4.8 million accidents occurred in a span of 5 years in Australia. It consists of the city names, the type of accident, condition of light, severity, speed zone, consumption of alcohol etc. The paper focuses on the application of data analytics to minimize the rate of accidents by studying, testing and training the past data. The objective of the project is to predict certain dependent data by using the independent factors on which it depends, uncover the pattern followed by the data collected over years and to draw helpful conclusions which would bring about some solutions to current crash related problems. Using some of the relevant attributes, demographic graphs [15] are plotted to bring out better insights.**

Keywords: Rule of Thumb, K-means clustering, Decision Trees, Data Analysis, R Analysis, Pruned model

I. Introduction

"The survey by NSW [14] shows that over the period from 2005 to 2013, about 30 people are hospitalized by crashes on roads each day". Taking into account, the effects of accidents, this paper aims at predicting the nature of the future accidents. However, several challenges are presented in this analysis due to the dearth of accident data analysis [1] and examination of the cause and effects of this data. Therefore, researchers use common data mining and prediction models [6] on the wealth of accident related information available. Performing trend analysis and recognizing particular patterns might evolve the responsiveness to accidents in metropolitan areas. The dataset has close to 4.8 million data points and consists of attributes such as the city names, the type of accident, the condition of light, severity, speed zone, whether it was caused due to the consumption of alcohol, whether it is a hit and run case, whether it is attended by the police and so on. The idea is to use statistical analysis using algorithms such as k-means and decision trees to arrive at a substantial value of information. In this study, we aim to find common causes of accidents by mining the past accident data. The approach to this study places a focus on both quantitative and qualitative research. A case-by-case approach [18] attempts to examine the accidents as isolated events and examines the possible causes of the accident. This, approach is useful in collecting the data and creating the database and metadata. The second approach called as Statistical approach [18] is used to analyze and observe patterns and common trends in accidents. In our research, we aim to focus heavily on the second approach since we are examining pre-recorded data. Statistical Methods is dealing with the analysis of causes of accidents in cities. Since, the causes and empirical analysis [1] can lead to prevention of accidents; the concept of accident involvement risk suggests itself as a methodical framework for empirical accident causation studies.

The general objective was to observe a common order of events which lead to severe accidents and analysis pertaining to high-density or "accident prone" areas. We further want to examine the levels of severity in an accident to examine how accidents can be mitigated. We also examine, all the parties affected by the project such as pedestrians, pillion riders, car drivers and wildlife. We also examine the extent to which each party is affected and whether there is a correlation between cause and effect in the cases of their accidents. The intent of this paper is to examine and ponder on the solution to accidents and accident response in urban areas. We do so using common statistics and analytical tools. The R language [6] has been used to mine the data for statistics. R is an open source tool and is highly extensible. The prepackaged library is readily available. Modules have been created to deal with the huge datasets and to bring out unique insights into the results. The module involving the determination of the severity of the accidents, on the basis of certain factors such as the city, consumption of alcohol, speed zone, the condition of the light etc. The above prediction is done using Decision Tree analysis which generates a decision tree as a result of prediction. Because it's a programmable environment that uses command-line scripting, you can store a series of complex data-analysis steps in R.

II. DESIGN

The dataset is in .csv format and is read in MS-Excel, which is the input here. The R system [6] is used to perform normalizing, pre-processing, distribution and other major analytic data steps. We then prepared a hypothetical analytic model plan using sample data, which is referred to testing in the system architecture. Some inferences are drawn from the results obtained from the testing data and are further generalized for the testing data. The statistical analysis [2] are first performed on the training data and then on the testing data[12]. The optimization is performed later in order to minimize the deviation of the obtained results. This deviation is measured in the terms of the mean values. After pruning of decision trees is performed, the optimal prediction and the decision tree is generated.

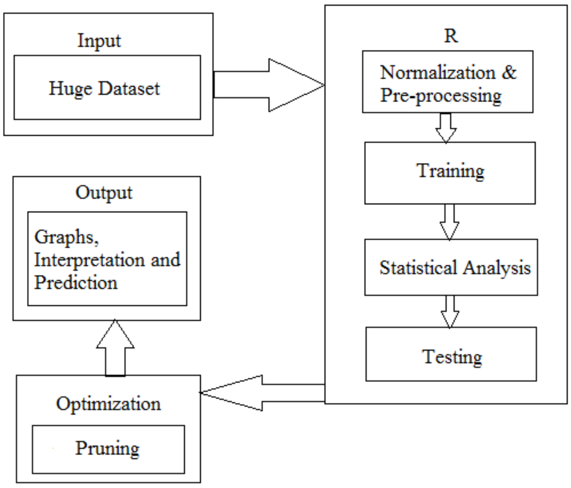


Fig 1. Architecture and Design of the system

III. Implementation

The reading of data, normalization[1] of data, training[12] the machine, implementation of the algorithms, statistical Analysis, testing of the accuracy of the predicted result, further optimization etc is done using the analytical tool, R. It has exhaustive libraries which support the prediction, analysis and implementation of the algorithms. The Comprehensive R Archive Network (CRAN) [6] is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.

The packages [13] used for the implementation are -

A. **Plotrix**: A large number of specialized plots and accessory functions like color scaling, text placement and legends. The plotrix package[6]] is intended to allow users to get many sorts of specialized plots quickly, yet allow easy customization of those plots without learning a great deal of specialized syntax.
B. **Tree**: It runs a K-fold cross-validation experiment[13] to find the deviance or number of misclassifications as a function of the cost-complexity parameter k. The overall deviance or a vector of contributions from the cases at each node. The overall deviance is the sum over leaves in the latter case.
C. **MASS**: Computes confidence intervals for one or more parameters in a fitted mode
D. **Cluster:** It computes agglomerative hierarchical clustering [19] of the dataset m constructs a hierarchy of clustering. At first, each observation is a small cluster by itself. Clusters are merged until only one large cluster remains which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster.
E. **gdata**: package provides various R programming tools[6] for data manipulation
F. **ggplot2:** An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics conditioning [13] and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources. It also implements a sophisticated multidimensional conditioning system and a consistent interface to map data to aesthetic attributes.
G. **ISLR:** It is the most commonly used package to plot the decision trees [5] based on certain parameters and to prune the tree based on the tree generated leading to the better accuracy of the prediction.

Decision tree builds classification or regression models [16] in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes. The algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous, then the entropy would be zero and if the sample is an equally divided it has entropy of one. The entropy [5] of the sample is obtained using the formula ,where H(S) is the entropy, p(x) is the number of successful cases.

$$\boldsymbol{H(S) = -\sum_{x\in X} p(x)\, log_2\, p(x)}$$

The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. This is obtained using the formula below, where IG is the Information Gain [16] H is the entropy and p is the number of successful cases present for each case.

$$\boldsymbol{IG(A,S) = H(S) - \sum_{t\in T} p(t) H(t)}$$

We choose the attribute with the largest information gain as the decision node. A branch with entropy of 0 is a leaf node. A branch with entropy more than 0 needs further splitting. The algorithm runs recursively on the non-leaf branches, until the data gets classified. *K-means* clustering is a method of vector quantization [7] originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering [11] aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name.

Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k such that k is less than or equal to n ($k \leq n$) and here, $\boldsymbol{\mu}_i$ is the mean of points in S_i.

$$arg_{\ s}\ min \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2$$

The determination of the number of clusters [17] in k-means is difficult. So, we used the Rule of thumb which sets the number of clusters as below

$$K \approx \sqrt{n/2}$$

Here n is the number of the objects or the data points.

IV. MODULES

A. *Day wise comparison of alcohol consumption*

In this module, we focus primarily on finding the days of the week when the accidents caused due to the consumption of alcohol are more. Usually, the alcohol consumption is more at the weekends than at the weekdays. The radial pie chart generated from the data depicts the same. The result of this analysis is depicted using a radial pie chart [15] in Fig 2.

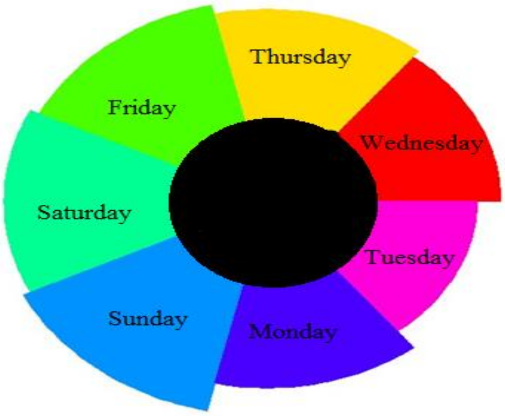


Fig 2. Days and alcohol consumption

B. *K-means clustering [7] for the city and the number of total number of deaths occurred due to the accidents*

This module uses K-means algorithm [11] to determine the high-density accident zones in the country. The x-axis represents twenty-five cities and the y-axis counted the total number of persons who were in accidents. Each of the cities in the country is associated with an integer value between 1 to 25. It was concluded that the region of cities between (20-25) experienced a higher density of accidents as the clusters are more here and the cities numbered as 4,9,13 and 16 experienced comparatively lesser number of accidents leading to death of people. The k-means algorithm was conclusive in determining the accidents density however; this method did not describe the quality or severity of the accident. Further analysis on the severity of the accident was required. The result of this analysis [5] is depicted in Fig 3.

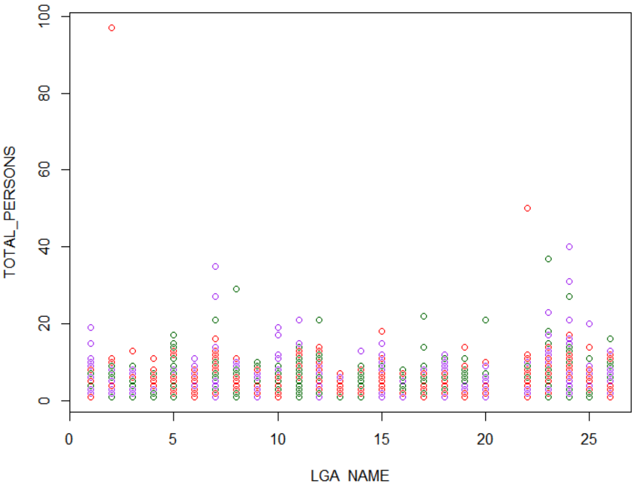


Fig 3. Clustering the cities and deaths occurred

C. *Decision tree [5] to find the severity of the accident using consumption of alcohol and light condition as the independent factors*

In this module, the decision tree algorithm [12] was implemented to discern the cause of accidents in the metropolitan areas. The alcohol time was used as the root node, light condition and age of pedestrian were used as the right and left nodes, and the number of injuries was the child of the light condition. We were able to discern, that there was a high incidence of injuries when a particular path was high-lighted. When the person was under the influence of alcohol and the light conditions were low, the number of deaths averaged at 3.5. Furthermore, the age of the pedestrian played an important role in the cause of injuries. The result of this analysis is depicted in the form of decision tree in Fig 4.

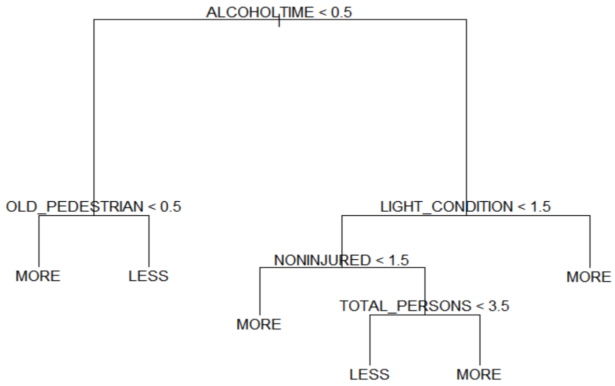


Fig 4. Decision Tree to find the severity of the accident based on status of Alcohol Consumption, Age of Pedestrians and the Light conditions.

D. *Decision tree [16] to find the severity of the accident using light condition and day of week as the independent factors*

In this module, the decision tree algorithm [16] was implemented to predict the severity of accidents using the light condition based decision tree which indicates the number of accidents according to the light condition and the day of the week. It was observed that the optimal light condition was during dawn and night time with street lights, where accidents were of a moderate quantity. However, the number of accidents during total darkness and harsh sun were highly severe than the average. The result of this analysis is depicted in Fig 5.

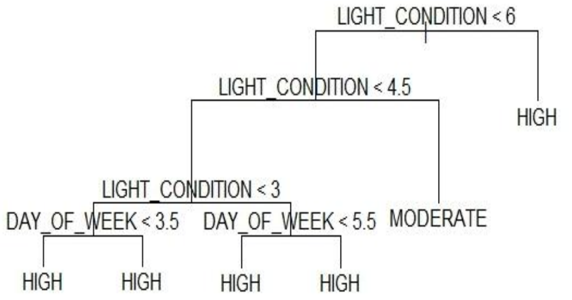


Fig 5. Decision Tree based on Light conditions and day of week.

E. *Severity Analysis based on Alcohol Consumption, Light Condition and Speed Zone*

In this module for the severity analysis, we examined the severity of the accidents based on alcohol consumption, light condition and the speed-zone. The results gave a deeper insight into the factors that caused accidents. The severity of alcohol related accidents averaged around 40,000 and it drastically reduced when there was no alcohol consumed by the injured party. This leads to the conclusion that most of the cases when the driver of the vehicle has consumed alcohol, the severity was more, which would have caused an accident. The bar graph [15] for this analysis is depicted in Fig 6.

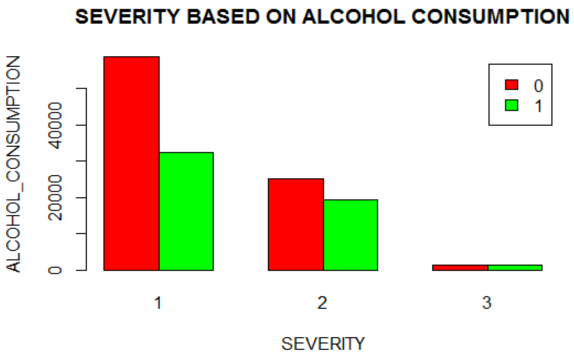


Fig 6. Severity Analysis based on the alcohol consumption

Subsequently, accidents were higher in number and were more severe, when there was pitch-darkness and street lights were absent (greater than 40,000 in number). In stark contrast, night-drives with street lights reduced the number of severe accidents to less than 10,000. The result of this analysis is depicted in Fig 7. Lastly, the numbers of accidents were higher when the speed limit was around 60 and reduced to less than 10,000 when the speed was less than 40 kilometers per hour. An aberrant observation was the number of accident when the speed limit was 30kmph. There were, an increasing severity when the number of deaths were more than 20,000. The result of this analysis in the form of bar graph [15] is depicted in Fig 8.

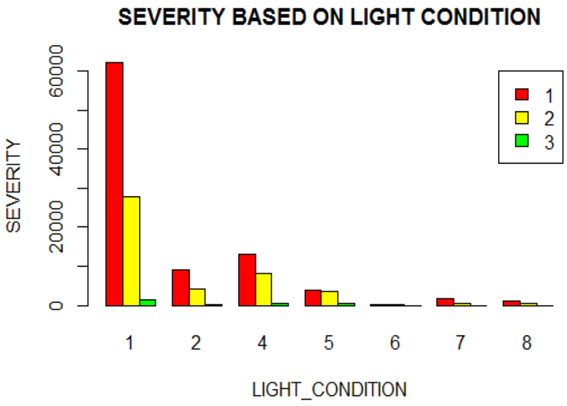


Fig 7. Severity Analysis based on the light condition

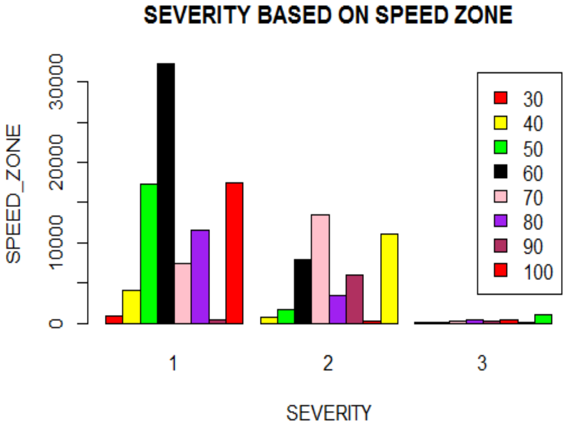


Fig 8. Severity analysis based on speed zone in which the accident occurred

F. *The cause – analysis* [2] *for the accidents in the weekends*

In this module, we examined the contribution of each of the reasons towards the occurrence of the accidents. The usual reasons behind the accidents are consumption of alcohol, struck by street animals, street lights off, heavy vehicles moving etc. We could easily analyze that most of the accidents on the weekends are caused due to the consumption of alcohol, followed by the absence of street lights and collision with the heavy vehicles. The result of this analysis in the form of 3-D Pie Chart [15] is depicted in Fig 9.

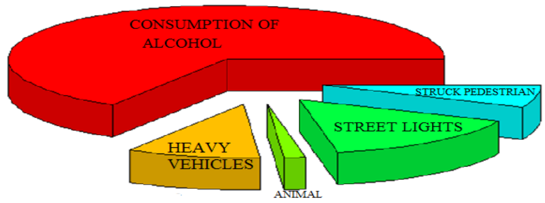


Fig 9. Comparison of accident types on weekends

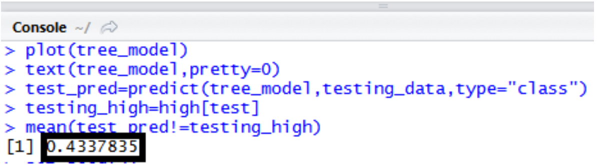
V. OPTIMISATION AND COMPARISON

The decision tree algorithm makes use of entropy [5], which is calculated using a formula, specified earlier. Using the value of the entropy, we find out the largest information gain value. The information gain [16] represents nothing but the extent to which the prediction of the event happens correctly using the method or algorithm used for training the machine. The comparison of the values of the means of the deviations for the module C, when the pruning was done and when it was done is shown in Table 1.

Module	Number of Data points	Pruning Done	Mean of Deviations
C	4.28 million	No	0.4337385
C	4.28 million	Yes	0.4247297

Table 1. Comparison of means of deviations before and after pruning

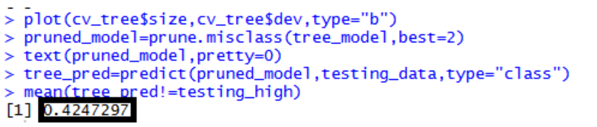
Initially, when there was no pruning [10] introduced to the decision tree, the mean of the deviations was almost 0.433 as in Fig 10.



```
> plot(tree_model)
> text(tree_model,pretty=0)
> test_pred=predict(tree_model,testing_data,type="class")
> testing_high=high[test]
> mean(test_pred!=testing_high)
[1] 0.4337835
```

Fig 10. Before pruning

The pruning basically shows us a graph using which we can decide the numbers of levels to be used for the decision making. Usually, the number of levels for the pruning lies in the middle of the range. Here, we use the value of 2 for the pruning, which means that it would generate a new tree based on the new level value.



```
> plot(cv_tree$size,cv_tree$dev,type="b")
> pruned_model=prune.misclass(tree_model,best=2)
> text(pruned_model,pretty=0)
> tree_pred=predict(pruned_model,testing_data,type="class")
> mean(tree_pred!=testing_high)
[1] 0.4247297
```

Fig 11. After pruning

After pruning is performed, the value of mean of the deviations has fallen down to 0.424. Pruning causes a reduction in the value of the mean of the deviations [10], which means that it adds accuracy to the prediction and makes the model fit to the data better. Hence, the concept of pruning must be done taking into consideration the correct value of best variable which would cause a substantial reduction to the value of the mean of the deviations.

VI. FUTURE WORK

This system aims to improve the approach to solving the method of collection and analysis of accident related information. This research attempts to reduce the average of time required by an ambulance to arrive at the spot of the accident and turnaround time with further data analysis. This predictive model can be implemented in several mobile applications which can minimize risk of accidents. This data can be used to implement protective measure to ensure road-safety of elderly citizens, wildlife and school children etc. We believe that this data can also be used, to build hospitals in locations so as to optimize response time and ensure timely arrival of ambulances, provide cab service to the people when they have consumed alcohol etc. The government can use this analysis to install more street lights along the positions where accidents are occurring in higher numbers during dawn, dusk or dark nights. This data can be used by policy-makers to design policies which ensure road safety. The optimal speed limit, alcohol consumption patterns and time of injury could be decided based on the analysis. This data can be used, by the department of forestry, wildlife etc., for deeper analysis and scrutiny.

VII. CONCLUSION

The analysis of the accident data is performed quantitatively and qualitatively. The sequences of conditions which lead to severe accidents are examined using the decision trees. Also, regions of high density accidents, the severity of the accident and the attributing factors were examined using the k-means clustering algorithm. The project in its early iteration can glean insight into locations which are accident prone and the causes for it. The secondary, goal of this research is to examine, how wildlife gets affected by road traffic. This predictive model can be implemented in several mobile applications which can minimize risk of accidents. It also attempts to research the possible causes of accident and the possible solutions to them. Further work can be done to build upon the insights that have been uncovered in this paper.

VIII. ACKNOWLEDGMENT

We are deeply indebted to Dr. K G Srinivas, Head of the Department, Computer Science & Engineering and Sowmya B J, Assistant Professor, Dept of CSE, MSRIT, Bangalore for their tremendous help in initiating the project, developing some of the ideas for the project, ceaseless guidance and support, analyzing the performance of the project at a conceptual level and to provide very valuable comments on the analysis reports prepared for the project.

IX. References

[1] Praticò Filippo.G (Reggio Calabria Mediterranean University), Vaiana Rosolino (Researcher, DIPITER Department Of Territorial Planning, University of Calabria-Arcavacata Campus (Cosenza)), Accident Data Analysis: An experimental Investigation for a rural, multilane, median separated, Italian Road.

[2] Heinz Hautzinger, Claus Pastor, Manfred Pfeiffer, Jochen Schmidt, Analysis Methods for Accident and Injury Risk Studies.

[3] Chen Lei, Xu Nuo, Analyzing Method of Traffic Accident Causation through Experts Method and Statistical Analysis .

[4] A.K. Jain (Michigan State University),M.N. Murty (Indian Institute of Science) and P.J. Flynn (The Ohio State University), Data Clustering: A Review:

[5] Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal, The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree.

[6] A Programming Environment for Data Analysis and Graphics Version 3.2.2 (2015-08-14), An Introduction to R Notes on R.

[7] S S Dimov, and C D Nguyen (Manufacturing Engineering Centre, Cardiff University, Cardiff, UK), Selection of K in K-means clustering D T Pharm.

[8] Kardi Teknomo,PhD, K-Means Clustering Tutorial.

[9] MaxCameron Monash (University Accident Research Centre), Accident Data Analysis To Develop Target Groups Countermeasures Volume 1 :Methods and Conclusions.

[10] Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man Machine Studies 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.R.

[11] https://en.wikipedia.org/wiki/K-means_clustering

[12] Utgoff, P. E. (1989). Incremental induction of decision trees. Machine learning, 4(2), 161-186.

[13] Institute for Statistics and Mathematics Available: https://cran.r-project.org/

[15] Robert. I. Kabakoff,, Available http://www.statmethods.net/graphs(2014)

[16] Deng,H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)

[17] Dasgupta, S. and Freund, Y. (July 2009). "Random Projection Trees for Vector Quantization" .Information Theory, IEEE Transactions on **55**: 3229–3242

[18] YING ZHAO ,GEORGE KARYPIS ,Uni of Minnesota, Department of Computer Science and Engineering And Digital Technology Center and Army HPC Research Center, “ Hierarchical Clustering Algorithms for Document Datasets” (2003)

[19] International Transport Forum, “Road Safety Report” 2014