International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

ANALYSIS OF ROAD ACCIDENTS USING APRIORI, NAIVE-BAYES AND K-MEANS

Dnymish Patil, Rohit Franklin, Sahil Deshmukh, Sarath Pillai and Prof. Madhu Nashipudimath

Department of Information Technology, PCE, Navi Mumbai, India - 410206

Road accidents are the main cause of death as well as serious injuries in the world. As a human being, everyone wants to avoid traffic accidents and stay safe. In order to stay safe, careful analysis of roadway traffic accident data is important to find out factors that are related to fatal, grievous injury, minor injuries, and non-injuries. The relationship between fatal rate and other attributes include combining weather conditions, road type, sunlight conditions, speed limit, drunk driver and so on are considered. Here, three data mining algorithms namely Apriori, Naïve-Bayes and K-Means are applied on the accident dataset to predict the accident severity. Apriori Algorithm is used for finding association rules between the attributes. Naïve-Bayes based approach is used for classifying how attributes are conditionally independent. K-means is used to form clusters and analyze them based on attributes. Comparison based on parameters is done to prove the efficiency of the various road accident detection techniques and approaches. By using this analysis, government/private agencies can take decisions in developing new roads and taking additional safety measures for the general public and awakening a sense of responsibility among road users.

Keywords—Road Accidents, Road conditions, Algorithms, Road accident detection systems.

1. Introduction

Road accident detection is considered to be the contemporary ever growing process focused primarily to reduce death. This study provides road accident detection techniques by analyzing the novel ideas. The analysis of these methods provides a better understanding of the steps involved in each process in a way of consequently increasing the scope for finding the efficient techniques to achieve maximum accurate performance. The comparison of the techniques used here, that is Apriori, Naive-Bayes and K-Means is carried out in terms of precision and recall. Environmental factors like roadway surface, weather, and light conditions do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have a stronger effect on the fatality rate. From the clustering result we can see the states/regions which have a higher fatality rate, while some others lower. We should pay more attention when driving within these risky states/regions. Current system is manual where government sector make use of this data and analyze it manually. Based on the analysis, they will take precautionary measures to reduce the number of accidents.

2. Literature Survey

This section reviews the research works carried out by different researchers that are related to the proposed work and provides a review of recent trends in motor vehicular accidents, factors influencing motor traffic safety and various methodologies commonly used in traffic safety studies. The literature review focuses primarily on the type of analysis available and the importance of determining and examining risk factors in general.

Liling Li, Sharad Shrestha and Gongzhu Hu [1] applied data processing algorithms on an outsized dataset for the analysis of road accidents. The link between fatal rate and different attributes together with collision manner, weather, surface condition, lightweight condition, and drunk driver were investigated. Association rules were discovered by Apriori rule, classification model was designed by Naive Bayes classifier and clusters were fashioned by straightforward K-means agglomeration rule.

P.D.S.S.Lakshmi Kumari & S.Suresh Kumar [2] applied mining algorithms on critical accident dataset to address this problem. The relationship between critical rate and other attributes combining weather conditions, road type, sunlight condition, speed limit, and drunk drivers are considered. Apriori Algorithm for finding an association between attributes. Naive based approach for classifying how attributes are conditionally independent.K-means is used to form clusters and to analyze them based on attributes. By using these Statistics government agencies can take decisions in developing new roads and taking additional safety measures.

Poojitha Shetty, S. P., Kashyap, S. V., & Madi, V [3] describes the way to mine frequent patterns inflicting road accidents from collected information. It finds associations among road accidents and predict the kind of accidents for existing also as for brand new roads. use of association and classification rules to find the patterns between road accidents and also as predicting road accidents for brand spanking new roads. Descriptive mining is applied on previous road accidents to mine frequent patterns in combos with alternative factors. within the planned system, apriori rule is employed to predict the patterns of road accidents by analyzing previous road accidents information. The results obtained from data processing approach will facilitate perceive the foremost important factors or usually continuance patterns. The generated pattern identifies the foremost dangerous roads in terms of road accidents and necessary measures will be taken to avoid accidents in those roads.

Shahsitha Siddique, V., & Ramakrishnan, N [4] discusses the algorithms that show higher performance within the previous studies and additionally the survey examines the foremost widely used data processing tools. The planned model implements by mistreatment the algorithmic rule that shows higher performance throughout the experiment to beat the shortcomings of previous studies on accident severity prediction. Road traffic accident historical information is obtained from National main road Authority of India (NHAI). Algorithms used are unit random tree,Naive-Bayes,Apriori,FP-Growth.This study identifies the appropriate algorithms, tools, review of recent studies and models on accident severity analysis and prediction,which helps to extract hidden road traffic accident patterns for future.

Atnafu, B., & Kaur, G [5] considered Deep Belief Network, supervised Latent Dirichlet Allocation, Support Vector Machine , Hybrid cluster, association rule mining, Random Forest, AdaBoostM1, Naïve mathematician, J48, PART, Preliminary real time autonomous accident detection system , Naive mathematician, C4.5, C&RT, RndTree,call list,rule induction, random tree, multi-class Support Vector Machine , Naive mathematician, J48, Random Forest algorithmic program, Apriori association rule mining, Decision Trees,Neural Networks,call tree and Support Vector Machines for analysis. The comparison is finished by the experimental results of the ways in terms of accuracy, precision, recall and F-measure. Attributes used square measure driver age, driver sex, vehicle category for result analysis.

Prasath, A., & Punithavalli, M [6] states that road transport is one of the most vital forms of transportation, connecting both long and short distances in our country. There are several attributes, which affect the intensity of a road accident like speed of the vehicle, road conditions, time of the accident etc. The accuracy of the algorithms are compared and it is found that KNN performs better than the other two algorithms employed. The major aim of the work is to find the accident severity. Also the work aims to reduce road accidents by giving awareness to public using the above method.

Beshah, T., & Hill, S.[7] proposed that road traffic accidents (RTAs) are a significant public health concern, leading to associate calculable one.2 million deaths and fifty million injuries worldwide every year. Within the developing world, RTAs are among the leading reasons behind death and injury; Abyssinia above all experiences the very best rate of such accidents. The results of this study might be utilized by the individual stakeholders to market road safety. Whereas the ways are easy, the results of this work may have an incredible impact on the well-being of Ethiopian civilians. The algorithms used are Decision Tree (J48), Naive Bayes, K-Nearest Neighbors.

Solanke, N. A., & Gotmare, A. D. [8] states that roadway traffic safety could be a major concern for transportation governing agencies also as standard voters.Data Mining is putting off hidden patterns from Brobdingnagian databases. it's usually utilized in selling, police investigation, fraud detection and scientific discovery. The algorithms used are Apriori Association, Naive Bayes Classification and K-Means Clustering. In data processing, machine learning is especially centered as analysis that is mechanically learnt to acknowledge complicated patterns and build intelligent selections of supported information.

Durga Karthik, P. Karthikeyan, S.Kalaivani & K.Vijayarekha [9] proposed that road accidents are a major cause of death and were analysed based on data mining algorithms such as disabilities.It uses Naive Bayes, Random Forest and J48. The aim of the traffic accident analysis for a region is to investigate the cause for accidents and to determine dangerous locations in a region. Multivariate analysis of traffic accidents data is critical to identify major causes for fatal accidents.

Siddthan, R., & Nagarajan, [10] proposed a new prediction system based on the clustering and rule mining algorithms. Here, the dataset obtained from the UCI repository is taken as the input, which is preprocessed for eliminating the irrelevant attributes and filling the missing

values in the dataset. Then, the noise free dataset is considered for further analysis, in which the clustering algorithm is applied for grouping the attributes of the data into a cluster. It uses Apriori and Density based clustering.From the evaluation, it is observed that the proposed Apriori algorithm provides better results compared than the existing technique.

Janani, G., & Devi, N. R. [11] states that traffic accident data of Coimbatore is collected and cleaned in order to use it to test the predictive model.In this paper algorithms used are Association rule, Naive Bayes and K-Means Clustering. The assessment of the Classification model showed that Naive Bayes algorithm outper- forms with an accuracy of 92.45 % when compared with other algorithms. In contrast with the previ- ously published work of authors, which focused on driver characteristics and dietary habits, this paper focused on the contribution of various road-related factors such as the role of environment, place where the accident occurred and cause of the accident.

3. Proposed Work

In order to predict the pattern of recent road accidents, associate association and classification data processing techniques area unit used, namely, Apriori, Naïve Thomas Bayes classifier and Kmeans, that area unit extremely scalable . although we have a tendency to area unit functioning on an information set with various records with some attributes, this classifier will yield best results. There area unit models that assign category labels to downside instances, that area unit depicted as vectors of feature values, and also the category labels area unit drawn from some finite set. the info is collected from police stations that area unit restricted to a region.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

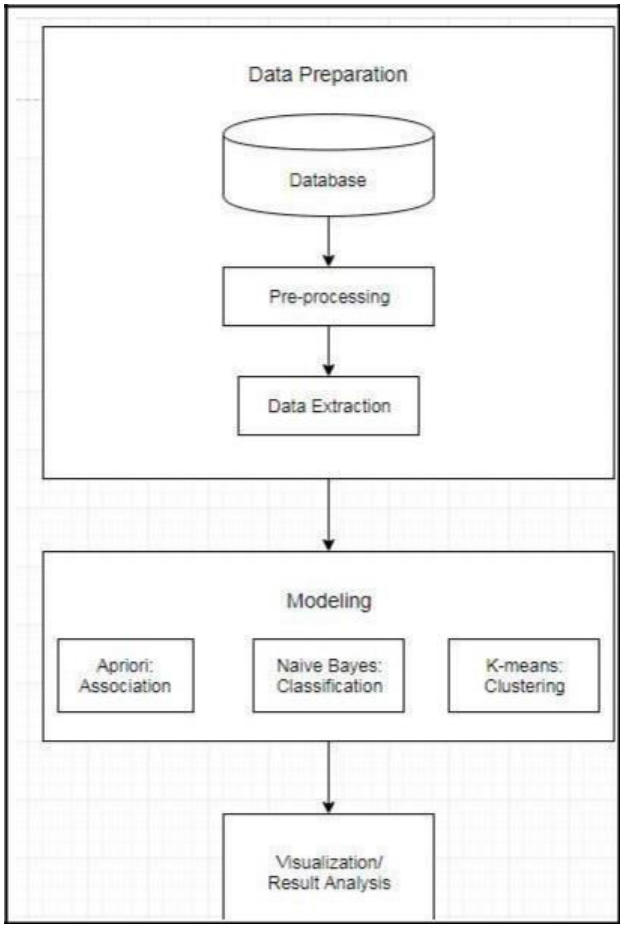

Fig. 1 Proposed system architecture

A. Data Preparation : Data preparation is performed before each model construction. All records with missing value in the chosen attributes are removed. All numerical values are converted to nominal value according to the data dictionary in the attached user guide. Fatal rate is calculated and binned into two categories: High and Low.Preprocessing is done to remove unnecessary words from the dataset and making the raw data for implementing algorithms on it.

Steps:-

1.Gather Data : The data preparation process begins with finding the right data.This can come from an existing data catalog or can be added when needed.

2.Discover and assess data : When grouping the info, it's vital to get every dataset. This step is concerning planning to recognize the info and understanding what needs to be done before the info becomes helpful in an exceedingly explicit context.

3.Cleanse and validate data : Improvement up the info is historically the foremost time intense a part of the info preparation method, however it's crucial for removing faulty knowledge and filling in gaps.Once knowledge has been cleaned, it should be valid by testing for errors within the knowledge preparation method up to the current purpose. Typically times, a blunder within the system can become apparent throughout this step and can get to be resolved before moving forward.

4.Transform and enrich information : Remodeling data is that the method of change the format or worth entries so as to succeed in a well-defined outcome, or to create the info additional simply understood by a wider audience. Enriching information refers to adding and connecting information with alternative connected info to supply deeper insights.

5.Store data : Once prepared, the data can be stored or channeled into a third party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

B. Modeling : Modeling is the process of making a knowledge model for the information to be stored in a very Database. This data model may be a conceptual representation of knowledge objects, the associations between different data objects and therefore the rules. Data modeling helps within the visual representation of knowledge and enforces business rules, regulatory compliance , and the government policies on the information. Data Models ensure consistency in naming conventions, default values, semantics, security while ensuring quality of the information.Data model emphasizes on what data is required and the way it should be organized rather than what operations have to be performed on the information.Data Model is like architect's building plan which helps to create a conceptual model and set the connection between data items.We first calculate several statistics from the dataset to indicate the essential characteristics of the fatal accidents. We then apply Apriori, Naïve-Bayes and K-Means to search out relationships among the attributes and therefore, the patterns.

Apriori Algorithm : Descriptive or predictive mining applied on previous road accidents data in combination with other important information such as weather, speed limit or road conditions creates an interesting alternative with potentially useful and helpful outcome for all involved stakeholders. Association rule mining is used to analyze the previous data and obtain the patterns between road accidents. The two criteria used for association rule mining are support and confidence. Apriori algorithm is one of the techniques to implement association rule mining. In the proposed system, we use apriori algorithm to predict the patterns of road accidents by analyzing previous road accidents data.

The steps for the Apriori Algorithm:-

- Scan the data set and find the support(s) of each item.
- Generate L1 (Frequent one item set). Use Lk-1, join Lk-1 to generate the set of candidate

k-item set.

- Scan the candidate k item set and generate the support of each candidate k – item set.
- Add to frequent item set, until C=Null Set.
- For each item in the frequent item set generate all non empty subsets.
- For each non empty subset determine the confidence. If confidence is greater than or equal to this specified confidence .Then add to Strong Association Rule.

Naive-Bayes Classification : Naive Bayes is a probabilistic classifier based on Bayes theorem. It assumes variables are independent of each other. The algorithm is easy to build and works well with huge data sets. It has been used because it makes use of small training data to estimate the parameters important for classification. Bayes Theorem states the following:-

P(attribute value ai/subject value vj)=(n_c + mp)/(n+m) where:-

- o n = the number of training examples for which v = vj
- o nc = number of examples for which v = vj and a = ai
- o p = a prior estimate for P(aijvj)
- o m = the equivalent sample size

The steps for the Naive-Bayes Classifier:-

- Scan the dataset.
- Calculate the probability of each attribute value. [n, n_c, m, p]
- Apply the formulae.
- Multiply the probabilities by p.
- Compare the values and classify the attribute

values to one of the predefined set of class.

K-Means Clustering : K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k-centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early age group is done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After we have these k new centroids , a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.Finally, this algorithm aims at minimizing an objective function know as squared error function given by where:-

$$J(V) = \sum_{i=1}^{c} \sum_{i} (\|x_i - v_j\|)^2$$

- '||xi - vj||' is the Euclidean distance between xi and vj.
- 'ci' is the number of data points in the ith cluster.
- 'c' is the number of cluster centers.

The steps for the K-Means Clustering:-

- Let X = {x1,x2,x3,....., xn} be the set of data points and V = {v1,v2,.....,vc} be the

set of centers.

- Randomly select 'c' cluster centers.
- Calculate the distance between each data point and cluster centers.
- Assign the data point to the cluster center whose distance from the cluster center is

minimum of all the cluster centers.

- Recalculate the new cluster center using:

where, 'ci' represents the number of data points in ith cluster.

15

- Recalculate the distance between each data point and new obtained cluster centers.
- If no data point was reassigned then stop, otherwise repeat from step 3.

C. Result Analysis : The results of the analysis include association rules among the variables, classification based on TP rate, FP rate, Precision, Recall, FMeasure, ROC area and clustering of number of fatal accidents as being high or low risks. This will be represented in visual tools like graphs.

4. Requirement Analysis

The implementation detail is given in this section.

4.1 Hardware and software details

The experiment setup is carried out on a computer system which has different hardware and software specifications.

The hardware used are Processor Intel Core i7, HDD 1 TB, RAM 8 GB.

The software used are Operating Language - Windows 10, ProgrammingLanguage - Java, Front-end - JTattoo, JFrame, JFreeChart, Database - MySQL and Back-end - Weka Tool, SQL.

4.2 Evaluation Parameters

Fatal Rate : Fatal Rate denotes the class of fatality in a fatal accident and is classified into two classes namely low and high. It is computed as:-

$$\text{Fatal Rate} = \frac{\text{Fatals}}{\text{Persons}}$$

Where Fatals is the number of fatalities and Persons is the number of persons involved in the accident.

4.3 Performance Evaluation

4.3.1Apriori Algorithm

After applying Apriori Algorithm to the dataset, the output is generated in the form of rules. In this case, it generates 16 rules which is shown in Table 2.

Table 2 Apriori Algorithm Rules

Rules	Fatal Rate	Support
REL_JUNC = Non_Junction	High	0.572
REL_JUNC = Non_Junction, PAVE_TYP = Blacktop	High	0.552
REL_JUNC = Non_Junction, C_M_ZONE = None	High	0.558
REL_JUNC = Non_Junction, TRA_CONT = No_Controls	High	0.558
TRA_CONT = No_Controls	High	0.611
TRA_CONT = No_Controls , WEATHER = Clear_Cloud	High	0.56
PAVE_TYP = Blacktop, TRA_CONT = No_Controls	High	0.59
PAVE_TYP = Blacktop, TRA_CONT = No_Controls , C_M_ZONE = None	High	0.576
TRA_CONT = No_Controls, C_M_ZONE = None	High	0.596
WEATHER = Clear_Cloud	High	0.634
PAVE_TYP = Blacktop, WEATHER = Clear_Cloud	High	0.614
PAVE_TYP = Blacktop, C_M_ZONE = None, WEATHER = Clear_Cloud	High	0.596
C_M_ZONE = None, WEATHER = Clear_Cloud	High	0.615
PAVE_TYP = Blacktop	High	0.667
PAVE_TYP = Blacktop, C_M_ZONE = None	High	0.649
C_M_ZONE = None	High	0.67

From the above table, it can be observed that attributes such as REL_JUNC, TRA_CONT, PAVE_TYP, WEATHER and C_M_ZONE have a strong influence on the fatal rate being high. The above mentioned attributes are represented graphically in Figure 2

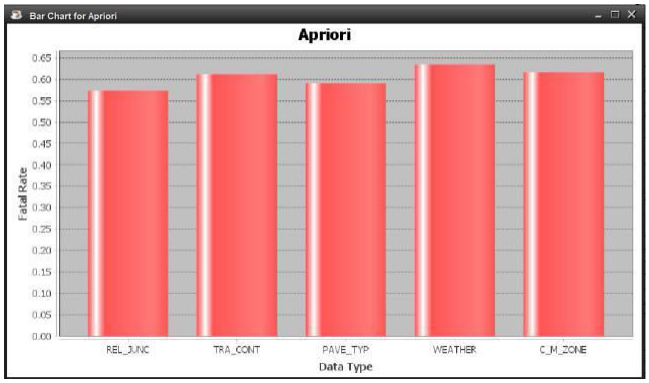

Fig 2 Performance Evaluation of Apriori

4.3.2 Naïve-Bayes Classification

Naïve-Bayes is then applied to the dataset and generates an output in the form of five measures, namely TP Rate, FP Rate, Precision, Recall, F-Measure and ROC Area. These five measures correspond to the two classes of Fatal Rate, that is low and high. Along with the two classes, there is also an average measure of the two classes. A table of the above measures is shown in Table 3

Table 3 Naïve-Bayes Measures

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
High	0.761	0.334	0.835	0.761	0.796	0.773
Low	0.666	0.239	0.556	0.666	0.606	0.773
Average	0.731	0.305	0.748	0.731	0.737	0.773

A graphical representation of the measures are shown below in Figure 3

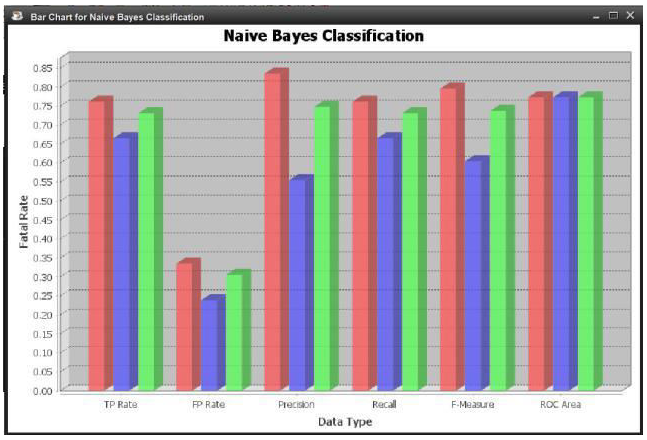

Fig 3Performance Evaluation of Naïve-Bayes

4.3.3 K-Means Clustering

K-Means Clustering is applied to the dataset to produce an output in the form of clusters. Here, there are three such clusters formed and are divided among the two classes of Fatal Rate. The possibility of Fatal Rate being high is 66% as compared to the Fatal Rate being low which is 34%. A table representing the three classes are shown in Table 4

Table 4 K-Means Clusters

Attribute	Cluster 1	Cluster 2	Cluster 3
STATE	Alabama	Alabama	Alabama
MONTH	Jun	Feb	Jul
HARM_EV	Motor_Vehicle_in_Transport_on_Same_Roadway	Overturn_Rollover	Motor_Vehicle_in Transport_on_Same_Roadway
LGT_COND	Daylight	Dark	Daylight
MAN_COLL	Not_Collision_with_Motor_Vehicle	Not_Collision_with_Motor_Vehicle	Angle_Front_to_Side_Right_Angle
REL_JUNC	Non_Junction	Non_Junction	Non_Junction
PAVE_TYP	Blacktop	Blacktop	Blacktop
TRA_CONT	No_controls	No_controls	No_controls
C_M_ZONE	None	None	None
WEATHER	Clear_Cloud	Clear_Cloud	Clear_Cloud
FATAL_Rate	High	High	Low

A graphical representation of the classes are shown in Figure 4

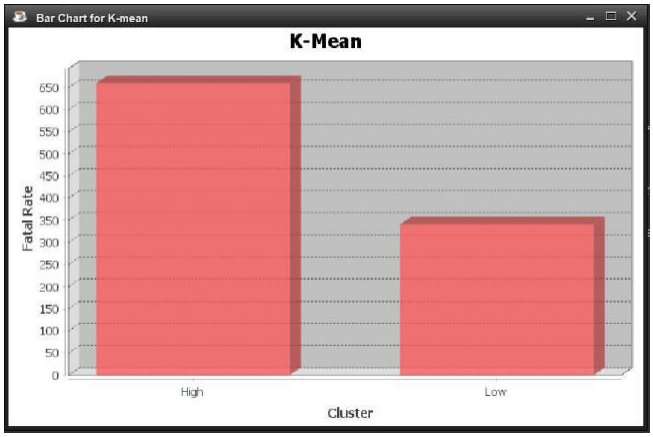

Fig 4Performance Evaluation of K-Means

5. Conclusion

Road accident analysis is considered to be the contemporary ever growing process focused primarily to reduce death. This study provides road accident detection techniques by analyzing the novel ideas. The analysis of these methods provides a better understanding of the steps involved in each process in a way of consequently increasing the scope for finding the efficient techniques to achieve maximum accurate performance. The comparison of the techniques used here, that is Apriori, Naive-Bayes and K-Means is carried out. Environmental factors like roadway surface, weather, and light conditions do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have a stronger

effect on the fatality rate. From the clustering result we can see the states/regions which have a higher fatality rate, while others have a lower fatality rate. We should pay more attention when driving within these risky states/regions. The current detection system is manual where the government sector makes use of this data by analyzing it manually. Based on this analysis, they will take precautionary measures to reduce the number of accidents.

References

[1] Li, L., Shrestha, S., & Hu, G. (2017, June). Analysis of road traffic fatal accidents using data mining techniques. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*(pp. 363-370). IEEE.

[2] P.D.S.S.Lakshmi Kumari & S.Suresh Kumar (2018). Mining Techniques for Analysing Road Accidents.

[3] Poojitha Shetty, S. P., Kashyap, S. V., & Madi, V. (2017). Analysis of road accidents using data mining techniques.

[4] Shahsitha Siddique, V., & Ramakrishnan, N. (2019). Analysing Road Accident Criticality using Data mining.

[5] Atnafu, B., & Kaur, G. (2017). Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques in Maharashtra, India. *International Journal of Current Engineering and Technology*,*7*,1974-1978.

[6] Prasath, A., & Punithavalli, M. (2018). A review on road accident detection using data mining techniques. *International Journal of Advanced Research in Computer Science*,*9*(2).

[7] Beshah, T., & Hill, S. (2010, March). Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. In *2010 AAAI Spring Symposium Series*.

[8] Solanke, N. A., & Gotmare, A. D. Roadway Traffic Analysis using Data Mining Techniques for Providing Safety Measures to Avoid Fatal Accidents.

[9] Durga Karthik, P. Karthikeyan, S.Kalaivani & K.Vijayarekha (2019). Identifying Efficient Road Safety Prediction Model Using Data Mining Classifiers.

[10] Siddthan, R., & Nagarajan, A. An Analysis of Road Accidental Data Using Clustering and Itemset Mining Algorithms.

[11] Janani, G., & Devi, N. R. (2018). Road traffic accidentsanalysisusingdatamining techniques. *JITA-JOURNAL OF INFORMATION TECHNOLOGY AND APPLICATIONS*,*14*(2).