International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

ACCIDENT INFORMATION MINING AND INSURANCE DISPUTE RESOLUTION

Lubnaah Jaleel[1], Sneha G[2], Sneha Susan Thomas[3], Mrs. Anitha Moses[4]

[1,2,3]Student, Dept. of Computer Science and Engineering, Panimalar Engineering College, Tamil Nadu, India
[4]Associate Professor, Dept. of Computer Science and Engineering, Panimalar Engineering College, Tamil Nadu, India

---

Abstract - *Road accidents have been estimated as the largest cause of death for children and young adults between the age of 5 and 25, with a minimum of 1 death every 24 seconds and an estimated 1.35 billion each year. The World Health Organization's Global Status Report on Road Safety, based on data from 2016, showed that the situation is worsening. In the traditional road accident incidents, the police record the statements of the victim, law-breaker, and eyewitness, procures the medical certificate from the hospital, detain the vehicle for inspection, all of which is stored as a written document. The proposed model aims to provide a centralized database for a victim or a relative to access the information stored by the police, the medico-legal form and other relevant documents used to claim the insurance. K-means Clustering algorithm is applied to this cumulated data to acquire valuable statistical information such as identifying high risk locations, the ratio of accidents in each location, the cause of the catastrophe and make it available to both users and police authorities. The association rules are determined using Apriori Algorithm.*

Key Words: **K-means clustering- RapidMiner tool- unsupervised learning- pattern mining-Apriori Algorithm- association rule learning**

1. INTRODUCTION

1.1 Survey on Accident and Traffic Based Data Mining System

In recent years, the vehicle density has increased in magnitude. To add to the agony of the people, the road conditions have diminished ensuing in a number of road accidents.

The accident and traffic based mining systems at present, use Big Data and its analytic tools in order to predict the traffic risk pattern. This is quite efficient if the datasets are large, thus speculating an approximate likelihood of a casualty. However, despite large datasets, the accident and traffic information are to a great degree heterogeneous in nature. The heterogeneity of data and the need for large datasets hinders the data mining process. Further, the absence and isolated nature of the accident information poses a challenge in the prediction of traffic accident risk.

However, with the growing technology, methodology to predict and complete the information by comparing with other similar or nearby locations has evolved and explored in [1]. It has solved two major problems 1) absence of a framework to combine the use of data from heterogeneous sources 2) a method to report mining results obtained for real-time datasets [1].

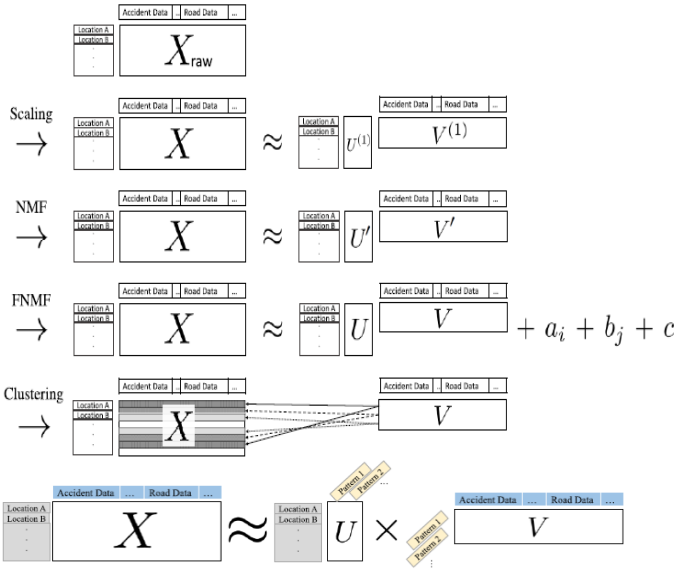

Fig -1: Existing model

1.2 Survey on Insurance Claim

In India, insurance companies provide insurance policies under different terms and conditions. The insurance industry is dependent on multiple processes between transacting parties for initiating, maintaining and closing diverse kind of policies.

In order to apply for insurance, the following documents are needed

- Copy of the insurance policy
- FIR
- The filled claim form
- Registration copy for the vehicle
- Copy of the driving license
- The original estimate of the repairs
- Medical receipts in case of any bodily injuries

Gathering the documents independently is a laborious task.

1.3 Need for Consolidated System

The existing system uses an elaborate unified algorithm that extends the non-negative matrix factorization by including a multiplicative update algorithm. Though this has been proved effective, it has overlooked the ranking of risky locations. Another shortcoming of the system is it does not encompass features to report the results of data mining to the corresponding authorities.

Thus there is a need for a system that accommodates all the requirements to cater to the needs of the users and the authorities. In this paper, we emphasize the need to publish the analyzed data. Broadcasting this information allows the authorities to make vital decisions to prevent mishaps in the future.

1.4 Need for Consolidated System

The significance and necessity of this paper can be summarized as follows:

1.4.1 **A Centralized Framework:** In this paper, we propose a portal with a centralized database to upload and access the above information for the purpose of an insurance claim. Significantly reducing the manual collection and maintenance of large amounts of data.

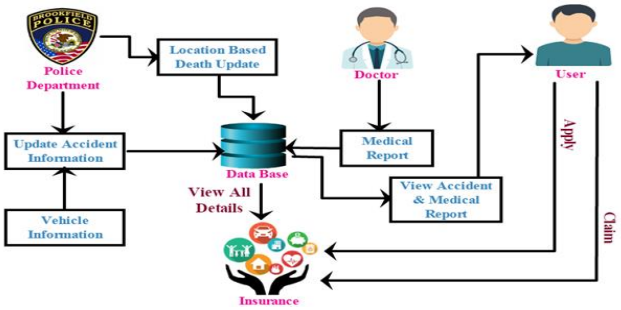

Fig -2: Proposed System Architecture

1.4.2 **A Novel Scheme for Accident Data Mining:** The collected data is primarily used for the purpose of data mining and clustering. The proposed system provides 3 main features: a) it clusters the various number of accidents occurred in each region based on a variety of factors – severity, type of vehicle, etc.; b) it predicts the number of accidents possible in those areas in the future; c) it can identify other regions with similar conditions- road conditions, traffic, etc.; where the probability of accidents is high.

1.4.3 **A State-Of-The-Art Reporting Technique:** The information obtained from the data clustering and predictive analysis algorithms is displayed on a portal. Personalized view of the number of accidents in a specific area is visible to a number of people without the need to create an account. Additionally email notifications are sent to users who have created the accounts. These users will receive cautionary emails to take discretion when they are in the regions with high accident ratios. Thus, the proposed model aims to create awareness and enlighten people about the mishaps.

Every year over 1 lakh people die in India due to road accidents. Some of these deaths are unavoidable, spur of the moment occurrences. But a majority of these accidents could have been predicted and avoided. The proposed model predominantly intends to solve this problem by bringing to light the various causes of accidents in each zone and the steps to be taken to prevent them. An additional objective of the proposed model is to predict a count of accidents and areas where their ratio is likely to increase. The proposed model resolves to bring down the ratio by at least 24% (nearly 35,000 deaths).

The remainder of the paper is organized as follows. Section II includes a review of the existing work on the accident data mining techniques and algorithms for clustering and predictive analysis. Section III will introduce the methodology used for analyzing integrated accident data and derive statistical information and the structure of the centralized system for document access. Section IV demonstrates the effectiveness of the proposed methodology using real accident data collected and maintained by the police and available for access by the users. Finally, Section V includes the conclusion of the paper.

2. PREVIOUS WORK

Li Zhu et al [2], has proposed framework of conducting big data analytics in Intelligent Transport System, where the data source and collection methods, data analytics methods and platforms, and big data analytics application categories are summarized.

Sheena Angra et al [3], focuses on the application of data mining in the field of education and implementation of three widely used data mining techniques using Rapid Miner on the data collected through a survey.

José María Luna et al [4], proposed new efficient pattern mining algorithms to work in big data. To this aim, a series of algorithms based on the MapReduce framework and the Hadoop open-source implementation have been proposed.

Eleonora D'Andrea et al [5], proposed a system that fetches tweets from Twitter according to several search criteria; processes tweets, by applying text mining techniques; and finally performs the classification of tweets. The aim is to assign the appropriate class label to each tweet, as related to a traffic event or not.

Xindong Wu et al [6], proposed a paper that presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining

and analysis, user interest modeling, and security and privacy considerations.

Stefan Atev et al [7], present a vision-based system addressing this problem and describe the practical adaptations necessary to achieve real-time performance. Innovative low-overhead collision-prediction algorithms (such as the one using the time-as-axis paradigm) are presented. The proposed system was able to perform successfully in real time on videos of quarter-video graphics array (VGA) (320 × 240) resolution under various weather conditions.

Sachin Kumar and Durga Toshniwal [8], proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining is used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm.

3. METHODOLOGY

The proposed model focuses on 2 principal modules:

- The centralized database for document upload and access, its security and a front-end for user access.
- The data mining component consisting of data clustering, predictive analysis and association rules.

3.1 Front-End, Centralized Database And Its Encryption

The proposed system uses JSP programming language and Java Servlets through NetBeans IDE for the front-end development. JSP solves the need for a responsive dynamic front-end that has been designed with a language capable of separating markup language and programming language code involving database actions and other tasks. The proposed system requires efficient server processing which has been achieved by combining JSP with Java Servlets.

The police collect the essential documents from the site of an accident and the hospital which are then encrypted using AES and consolidated onto a portal. The user can then access these documents via a key sent by the police upon identity verification.

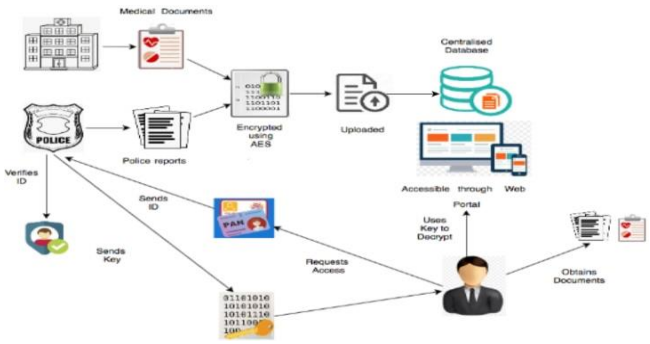

Fig -3: Proposed System Flow Diagram

The reasons for having chosen AES encryption among other encryption techniques as follows.

AES is exponentially stronger than DES and is more efficient. The Figures shown demonstrate the performance levels of AES:

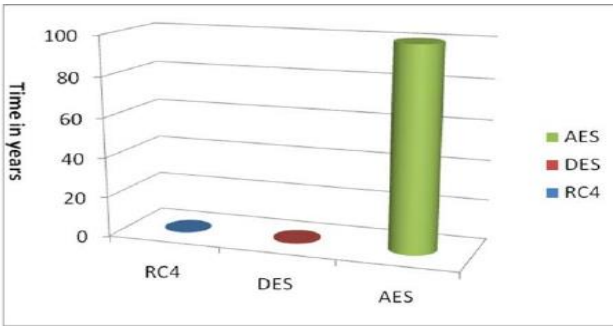

Chart -1: Encryption techniques safe time for AES, DES and RC4 encryption algorithms

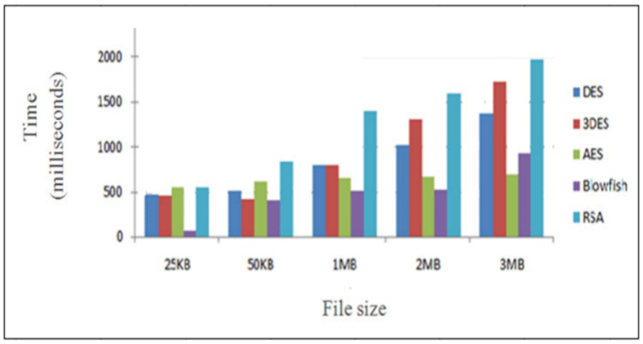

Chart -2: Encryption time vs. File size for DES, 3DES, AES, Blowfish, and RSA

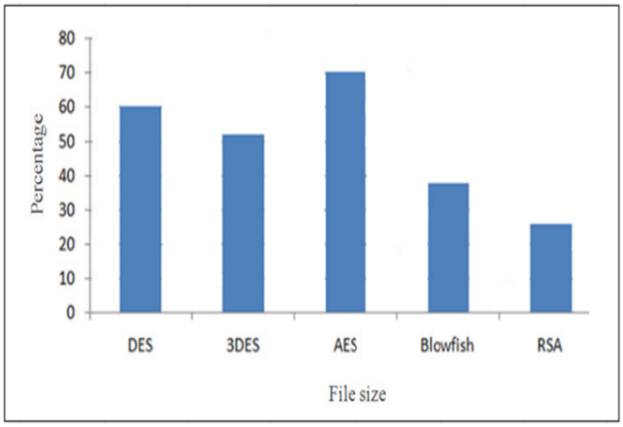

Chart -3: AES manifests the highest avalanche effect, whereas RSA manifests the least avalanche effect.

3.2 Data Mining - Cluster Analysis

The proposed system uses the K-means clustering algorithm. K-means is an iterative algorithm that partitions or divides the datasets into K clusters, used predominantly for classifying.

Each cluster is a non-overlapping subgroup. Also, each data point belongs to one group. This assignment is such that *"the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum."* [9]

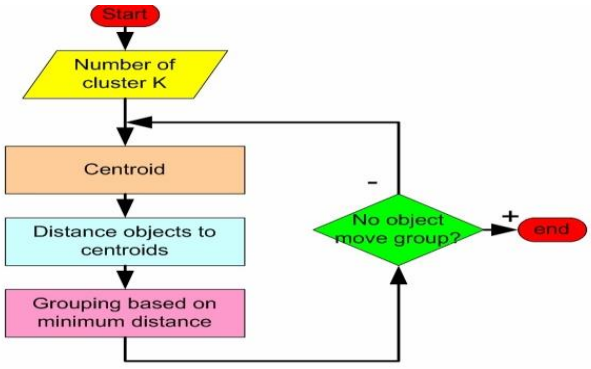

Fig -4: Flowchart of K-means clustering

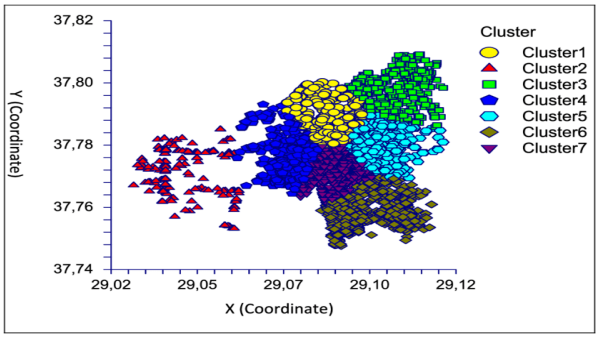

Fig -5: Sample clusters obtained by K-means clustering approach.

3.3 Data Mining- Predictive Analysis

Apriori Algorithm which is based on prior knowledge of frequent datasets is used in our proposed model.

It works as follows: Apriori assumes that *"All subsets of a frequent itemset must be frequent (Apriori property). If an itemset is infrequent all its supersets will be infrequent." [10]*

Apriori is used to mine the relevant association rules. Association rules are rule-based machine learning method, created by searching the databases using if-then criteria to uncover important relationships among variables. For n items in the set *I*, the total number of possible association rules is $3^n - 2^{n+1} + 1$.

The 3 defined relationships are

$$Rule: X \Rightarrow Y \quad Support = \frac{frq(X,Y)}{N} \quad Confidence = \frac{frq(X,Y)}{frq(X)} \quad Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Fig -6: Relationships of Apriori Algorithm

3.4 Collecting Datasets

The proposed system aims to perform clustering and predictive analysis using the above algorithms over 10000-15000 datasets.

These datasets are derived from the information collected and combined by Open Government Data (OGD) Platform India [11]. The various statistical tables include details like number of accidents in each area, their time of occurrence and month and year of occurrence. The severity of the accident, fatalities occurred if any, number of people injured and/or number of deaths occurred are also tabulated. Distinct tables include type of vehicles, weather condition, road condition, vehicular defects (if any), type of road in relation to the location in terms of latitude and longitude and proximity to prominent places like schools, residential areas, markets etc. All of which is combined into a single consolidated matrix, where each row of matrix represents a single place/area.

ui.R | server.R | only_road_accidents_data3

Filter

	STATE/UT	YEAR	0-3 hrs. (Night)	3-6 hrs. (Night)	6-9 hrs (Day)	9-12 hrs (Day)	12-15 hrs (Day)	15-18 hrs (Day)	18-21 hrs (Night)	21-24 hrs (Night)	To
1	A & N Islands	2001	2	6	29	40	39	40	18	7	
2	A & N Islands	2002	2	6	22	41	33	33	23	8	
3	A & N Islands	2003	2	8	31	35	28	36	25	15	
4	A & N Islands	2004	2	5	29	42	43	43	37	14	
5	A & N Islands	2005	0	8	27	28	38	42	50	13	
6	A & N Islands	2006	1	3	17	33	33	23	38	7	
7	A & N Islands	2007	2	5	20	30	30	27	31	7	
8	A & N Islands	2008	3	7	33	24	40	31	40	13	
9	A & N Islands	2009	2	6	35	41	64	54	50	19	
10	A & N Islands	2010	2	10	36	45	64	57	53	18	

Fig -7: Sample datasets collected.

3.5 Clustering And Prediction Mechanism Using Rapid Miner

In order to obtain distinct clusters based on severity of accidents, the proposed system intends to use RapidMiner as follows. The clusters are chosen based on number of distinct locations.

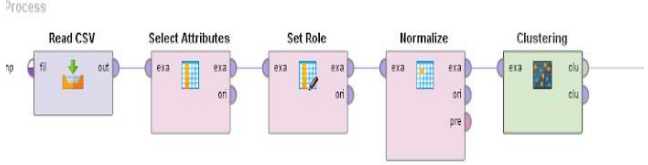

Fig -8: K-means using RapidMiner

The datasets are selected and attributes are set. By setting the role as the area where the accident took place, the 3 clusters are obtained.

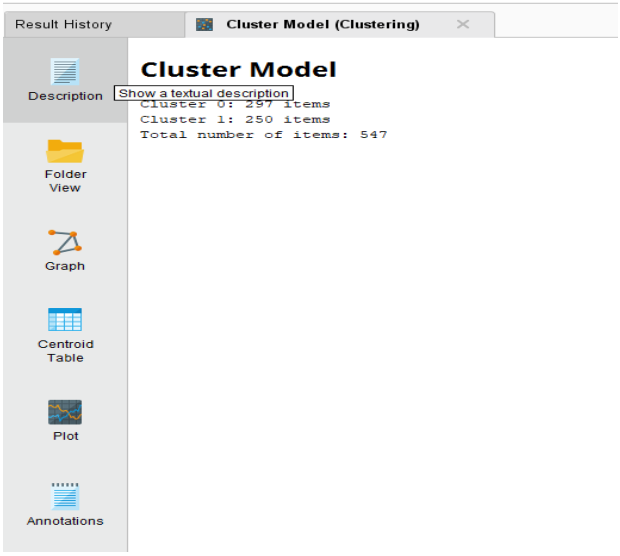

Fig -9: K-means Cluster Model

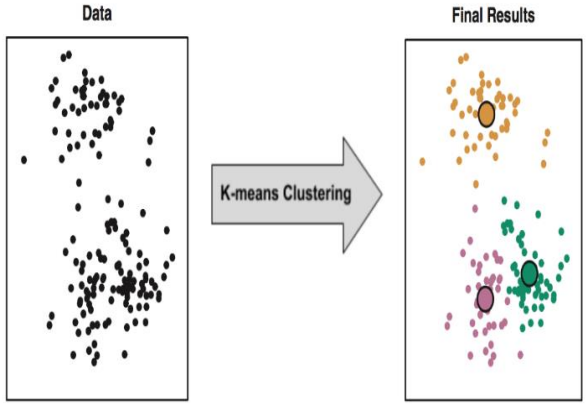

Fig -10: Demonstration of K-means Cluster

For the purpose of prediction, we require association rules. These rules are obtained by

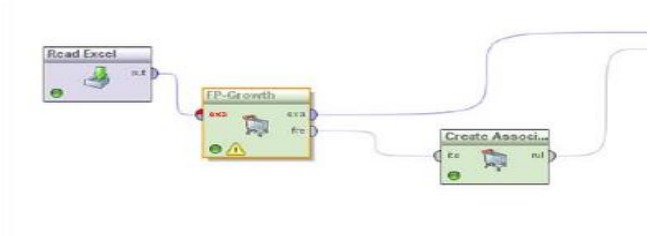

Fig -11: Association Rule creation using RapidMiner

Some of the rules obtained are as follows.

Table -1: Sample rules extracted using Apriori algorithm

ASSOCIATION RULES			
RULES	SUPPORT	CONFIDENCE	LIFT
Vehicle Type=Car → Accident Severity=Slight	0.278	0.84	1.03
Conditions=Fine no high winds	0.306	0.98	1.03
Sex of Driver=Male→ No of Casualties=1	0.306	0.81	1.01
Age of Driver =18-20 →No. Of Accidents >10	0.229	0.91	1.17

4. RESULT AND EVALUATION OF PROPOSED MODEL

In this section, the proposed model has been executed and its characteristics such as accuracy have been analyzed. The results of clustering and prediction have been obtained as graphical cluster and a decision tree respectively.

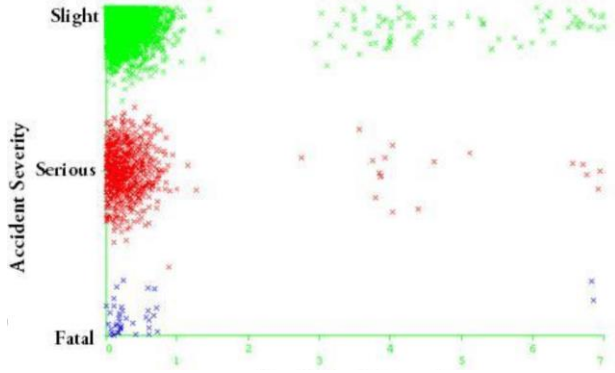

Chart -4: Sample Accident data analysis based of severity of Accidents

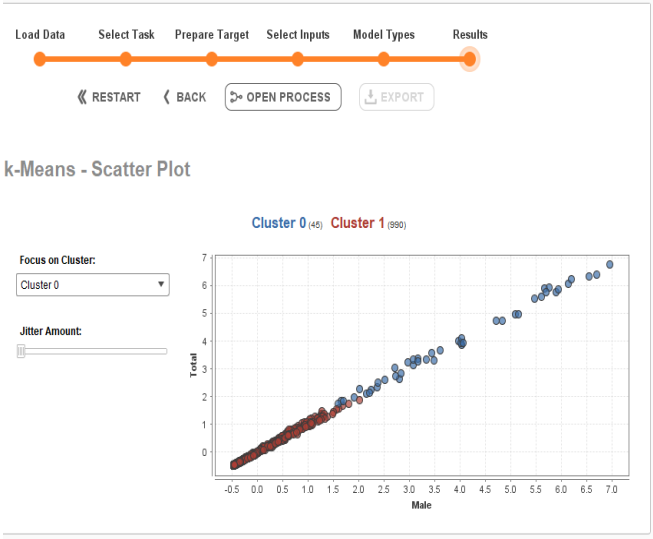

Chart -5: Cluster Analysis using K-means

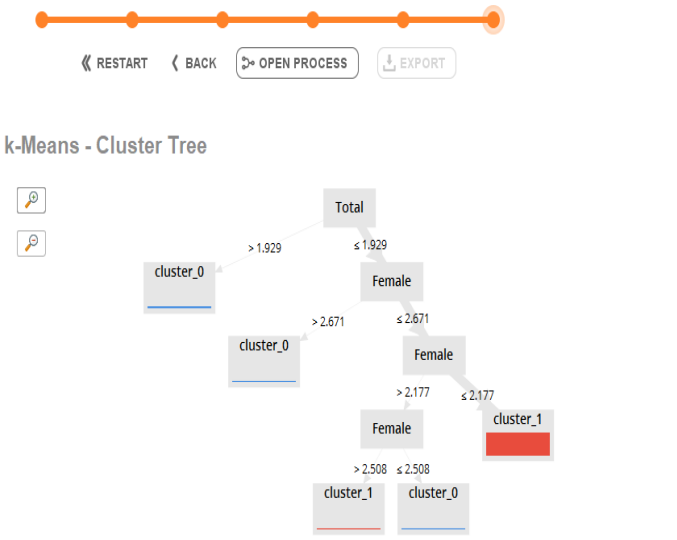

Fig -12: Results of Clustering

Model	Accuracy	Standard Deviation	Runtime
Naive Bayes	1.0	0.0	445.0
Generalized Linear Model	1.0	0.0	676.0
Logistic Regression	1.0	0.0	318.0
Fast Large Margin	1.0	0.0	785.0
Deep Learning	1.0	0.0	2622.0
Decision Tree	1.0	0.0	466.0
Random Forest	1.0	0.0	8155.0
Gradient Boosted Trees			
Support Vector Machine			

Fig -13: Results of Predictive Analysis

5. CONCLUSION

Road traffic accidents take a significant toll on road users and their life. The proposed system has been expected to perform with high accuracy and efficiency. The main goal of the project to alert the authorities and the people about the accident-prone areas and to take the necessary steps to prevent them in the future is successfully achieved. Data mining has been efficiently applied to derive the essential information and prediction is done with utmost accuracy, to estimate the future incidents. The proposed system is estimated to raise awareness and thereby reduce future catastrophes.

REFERENCES

1) Koichi Moriya, Shin Matsushima, and Kenji Yamanishi "Traffic Risk Mining From Heterogeneous Road Statistics" IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 19, NO. 11, NOVEMBER 2018

2) Li Zhu, Fei Richard Yu, Fellow, IEEE, Yige Wang, Bin Ning, Fellow, IEEE, and Tao Tang "Big Data Analytics in Intelligent Transportation Systems: A Survey" IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 20, NO. 1, JANUARY 2019

3) Sheena Angra, Sachin Ahuja "Implementation of Data Mining Algorithms on Students Data using Rapid Miner", 2017 International Conference On Big Data Analytics and Computational Intelligence (ICBDACI)

4) José María Luna, Member, IEEE, Francisco Padillo, Mykola Pechenizkiy, Member, IEEE, and Sebastián Ventura, Senior Member, IEEE"Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data", IEEE TRANSACTIONS ON CYBERNETICS

5) Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, Member, IEEE, and Francesco Marcelloni, Member, IEEE "Real-Time Detection of Traffic From Twitter Stream Analysis", IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

6) Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE," Data Mining with Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

7) Stefan Atev, Hemanth Arumugam, Osama Masoud, Ravi Janardan, Senior Member, IEEE, and Nikolaos P. Papanikolopoulos, Senior Member, IEEE "A Vision-Based Approach to Collision Prediction at Traffic Intersections", IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 6, NO. 4, DECEMBER 2005

8) Sachin Kumar and Durga Toshniwal, "A data mining framework to analyse road accident data" in Journal of Big Data (2015) 2:26 DOI 10.1186/s40537-015-0035-y

9) Imad Dabbura "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks" Sep 17, 2018: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

10) https://www.geeksforgeeks.org/apriori-algorithm/

11) Open Government Data (OGD) Platform India, https://data.gov.in/