Road Accidents Analysis by Using Data Mining Technique

[1]Sheela Singh, [2]Prof. R.K. Singh
[1]M.Tech student, [2]Associate Professor
[1]Department of Computer Science and Engineering,
[1]K.N.I.T, Sultanpur, U.P, India.

Abstract: Traffic safety is an important concern for transportation agencies either it is any government agency or private one and citizens also. To make driving safe, careful analysis of traffic collision is necessary. To find out variables/features those are closely relatable to major fatal accidents. For doing this analysis in this project the datasheets which considered are like weather condition, driver's involvement, type of vehicle, road style, type of junction, age of vehicle and vehicular defect. By using number of accidents happen because of these factors the results will be calculated and analyzed. In this research association rule mining is used to get the relationship between one factor to another and some mathematical calculation to find out the factors which affect the collision most and name of states which are highly affected by these factors.

IndexTerms **– Road Accident Analysis, Data Mining, Association Rule.**

I. INTRODUCTION

Method of finding relationship/dependency in large information sheets, including strategies at common collection of machine learning, information systems and statistics is known as data mining. Data processing is knowledge domain of technology and statistics with a complete goal for retrieving info (with smart and capable methods) from knowledge set and remodel the knowledge into noticeable design for any useful work. Data modeling/ Data mining is an analysis procedure of "knowledge discovery in databases". Other than that, it additionally involves information and information management process, information pre-processing, model and illation concerns, power metrics, complexness concerns, post-processing of on-line change, image. The distinction between information analysis and information processing is the information analysis is employed for checking the models and hypotheses on the data sheets, e.g., analyzing the efficiency of a promoting work, no matter the number of data; in distinction, information processing uses machine-learning and applied mathematics models to uncover relationships in a very giant amount of information.

The term "data mining" is a name; as a result it is the retrieval (mining) of relationship and data from giant amounts of information, not the retrieval of information itself. It is also a buzzword and is commonly applicable to any style of large-amount data or scientific discipline (collection, extraction, reposition, analysis, and statistics) moreover as any use of computer science, as well as AI (e.g., machine learning) and business intelligence.

The real data processing is the analysis of enormous attributes of information to retrieve unknown, fascinating relations like teams of information records (cluster analysis), uncommon records (anomaly detection), and dependencies (consecutive pattern mining, association rule mining). These relations will then be seen as a sort of outline of the computer file, and will be employed in any analysis or, as an example, in artificial intelligence and prognostic analytics. As an example, the information mining process would possibly connect many groups within the data, which may be accustomed, acquire a lot of right prediction outcomes.

The data discovery in databases method is often outlined with the steps:

1. Selection process
2. Pre-processing process
3. Transformation process
4. Data mining process
5. Interpretation process

In this project, association rule mining is used. It is used here to find the pattern between the factors. Factors, who affect the collision most, may have some occurrence pattern because of them road accident happen. Association rule is if-then statement that facilitates to represent the likelihood of patterns between knowledge at intervals massive knowledge collection in many styles of databases. Association rule mining incorporates a variety of applications and is wide to facilitate discovering sales correlations in transactional knowledge, in medical knowledge sets etc. Association rule mining, at a basic level, incorporates the utilization of machine learning models to research knowledge for dependencies or relations, or co-occurrence, in an exceedingly info. It identifies if-then relations, which are mentioned as association rules.

Association rules are calculated from item sets, which are created from two or more features of data sets. If rules are designed from analysing all the potential item sets, there can be such a lot of rules that are made to find relationship between them. With that, association rules are generally created from rules, well-represented in knowledge.

II. LITERATURE SURVEY

The analysis on road accident is done by many people, some important results, which really pays role are included here. The analysis where authors Sachin Kumar and DurgaToshniwal used the dataset of FARS (Fatal Accident Reporting System) from California Polytechnic State University for analysing road accident fatality. In that process, author used Weka tool (a data mining tool) to get results and applied association rule, classification rule and clustering process in Weka tool. Apriori association algorithm to find frequent item sets, classification and clustering to find relationship among the values and the patterns.

As the result, author finds out that environment factor like road condition, light condition and weather do not strongly affect the collision rate , while the public reason like driver's condition of being affected by alcohol or not and the road crash type have strong effect on the death rate [1].

In alternative analysis author analysed the causes of accidents. In this one, the datasheet for the study carries road crashes results of the year 2008 created by the department of state of urban center and studied the work of Naive Bayes, J48 and AdaBoostM1 for predicting accuracy in classification result. The classification accuracy on test outcome shows for the subsequent 3 cases like road crash, vehicle and negligence. Random Forest outperforms than other classification algorithms rather than choosing all the values for classification rule. Genetic algorithmic program is employed for feature choice to scale back the spatiality of the data collection. During this work, they pushed the analysis to a few totally different types of cases like crash, Casualty and Vehicle for locating the reason behind collision and also the seriousness of collision [2].

The expansion of civilization and the vehicle population is causing more collisions in all civilized societies in our country. Some corporate administration measures are being implemented to alleviation congestion on the streets and make travel safer. According to MORTH-2015 Bharat has the foremost important variety of collision altogether over the globe. Collision fatality has been increasing at every step; later security on road may be a most crucial issue. Motorcar crashes have a remarkable result to our society and culture. The combination variety of collisions expanded by a pair of.5 % from 4, 89,400 of 2014 to 5, 01, 423 of 2015 thus there's a necessity of testing on road collisions causes and security measures should be taken to scale back road collisions. During this study Domino's Theory utilizes for investigation of road collisions and for traffic security uses Risk physiological state Theory moreover as collision hindrance model with essential live to scale back road collision [3].

According to some other analysis, the safety on road depends on driver, vehicles, and environment and these factors affect the road safety individually or in group. It is necessary to search and study the dependency between road crashes and road style parameters to avoid road crashes and to create a safe driving and travelling surrounding to society.

The road collision prediction model was designed by victimization multiple simple regression analysis. The collision percentages were significantly associated with the road style parameters studied, such as the dimensions of the pavement, the horizontal curvatures, the roughness index of the road and the junction (Entry, Exit). The developed model is useful for the emergence of safe highways. To boot, it'll contribute to distinguishing the possibly venturous locations on highways and to the treatment of safety enhancements [1].

A survey MORTH-2013 Bharat has the maximum number of road accidents in the whole World. Road collision fatality has been increasing every year. So for, Safety on road is a major matter of concern. A study results as, on NH-55 who connects to various major industrial organizations and mines. That results that maximum cases of deaths are due to trucks. The main reasons of collisions are due to high density, high speed, wrong parking, old girth trees on shoulder, visibility issues etc. [4]

Human factors in vehicle collisions embrace something associated with drivers and other people travelling on road that will contribute to a crash. Examples embrace driver behaviour, visual, audial capacity, decision-taking ability, and reaction speed. A 1985 report supported British and yank crash information found error because of driver, intoxication and different human factors contribute whole or part to concerning ninety three of crashes. Drivers' negligence because of mobile phones had nearly fourfold larger risk of bally their cars than people who weren't. Dialling a phone is that the most risky distraction, increasing a drivers' probability of collision by twelve times, followed by reading or writing, that accrued the danger by ten times [5].

Motorized vehicle speed The U.S. Department of Transportation's Federal route Administration review analysis on traffic speed in 1998 [6]. The outline says:

- The proof shows the chances of getting collide is inflated each for motor vehicles travelling slower or higher than the common speed.
- The chance of being slashed will exponentially increase with speeds a lot of quicker than the median speed.
- The fatality/ deadliness of a collision depend on the vehicle speed modification at impact.
- There is proscribed proof suggesting lower speed limits end in lower speeds on a system-wide basis.
- Most collisions associated with speed involve speed too quick for the conditions.

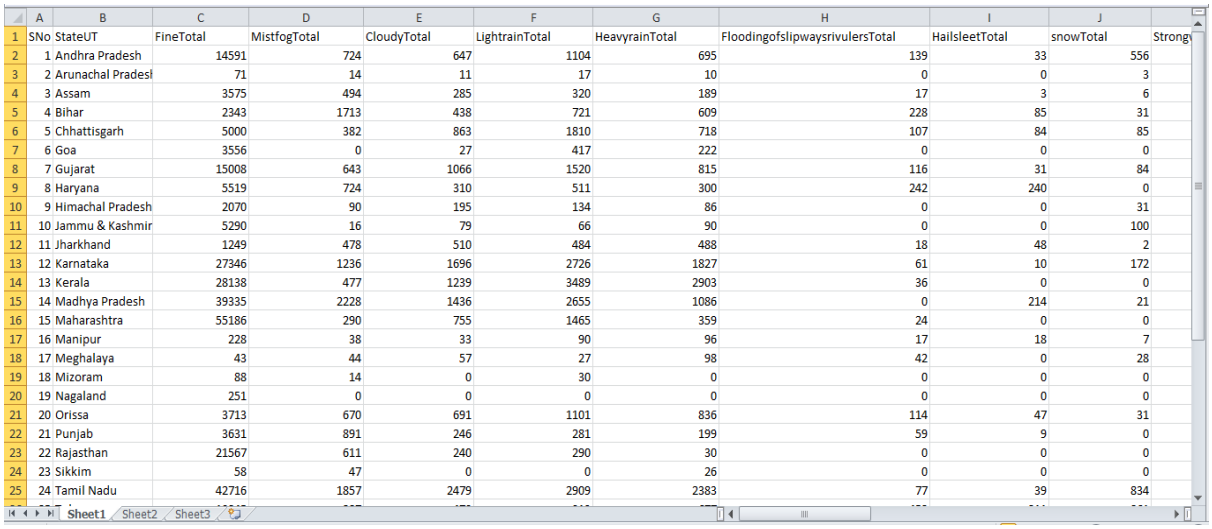
In Microscopic Traffic flow modelling they attempt to analyses the flow of traffic by modelling Driver- Driver and Driver-Road interaction within a traffic area. It also analyses the different situation like a case as driver meets a static problem (obstacle) or a dynamic obstacle. They analyses and focus on the minute aspect of traffic and analyses the interaction between vehicle-vehicle and individual vehicle behaviour. For this analysis optimal models and simulation models get into use.

III.METHODOLOGY

By using the records of road accidents in India, we are eager to find the features, which may be responsible for crashes on roads. These features have control over the road crashes. Analysing these features or combination of features can give us the result that can make people safe during travel or driving. Features which are always present during driving are mainly vehicle, driver, weather and road. As being a responsible and aware citizen, we should try to reduce the rate of accidents all over the country.

For analyzing road accidents, datasets are required, related to the whole country. Using quantitative method is beneficial if we have proper records related to the problem. We have used statistical method to retrieve the features (who affect the accidents most) and states (who faces the maximum number of accidents). We used road accident data from a portal provided by the Government of India.

Road accidents are random in nature and there is no clear pattern to predict and control the future mishap. Their analysis requires the knowledge of the features affecting them. By analysing the road accident data, we studied various permutations and combinations to find the measure reasons behind the collisions. So, we need to collect the dataset related to those features which become reason for the collision. These reasons are mainly like weather condition, driver's attention, age and type of vehicle, type of road and junction and defect in vehicle. Considering all these reason as features for this analysis, we are using datasets who have the record of accidents according to the states in India.



SNo	StateUT	FineTotal	MistfogTotal	CloudyTotal	LightrainTotal	HeavyrainTotal	FloodingofslipwaysrivulersTotal	HailsleetTotal	snowTotal	Strong
1	Andhra Pradesh	14591	724	647	1104	695	139	33	556	
2	Arunachal Pradesl	71	14	11	17	10	0	0	3	
3	Assam	3575	494	285	320	189	17	3	6	
4	Bihar	2343	1713	438	721	609	228	85	31	
5	Chhattisgarh	5000	382	863	1810	718	107	84	85	
6	Goa	3556	0	27	417	222	0	0	0	
7	Gujarat	15008	643	1066	1520	815	116	31	84	
8	Haryana	5519	724	310	511	300	242	240	0	
9	Himachal Pradesh	2070	90	195	134	86	0	0	31	
10	Jammu & Kashmir	5290	16	79	66	90	0	0	100	
11	Jharkhand	1249	478	510	484	488	18	48	2	
12	Karnataka	27346	1236	1696	2726	1827	61	10	172	
13	Kerala	28138	477	1239	3489	2903	36	0	0	
14	Madhya Pradesh	39335	2228	1436	2655	1086	0	214	21	
15	Maharashtra	55186	290	755	1465	359	24	0	0	
16	Manipur	228	38	33	90	96	17	18	7	
17	Meghalaya	43	44	57	27	98	42	0	28	
18	Mizoram	88	14	0	30	0	0	0	0	
19	Nagaland	251	0	0	0	0	0	0	0	
20	Orissa	3713	670	691	1101	836	114	47	31	
21	Punjab	3631	891	246	281	199	59	9	0	
22	Rajasthan	21567	611	240	290	30	0	0	0	
23	Sikkim	58	47	0	0	26	0	0	0	
24	Tamil Nadu	42716	1857	2479	2909	2383	77	39	834	

Figure 3.1: Accidents due to weather

Figure 3.1 shows the data sample of weather type. Weather includes the sub features like cloudy, foggy, fine, rainy and flood etc. Likewise, driver's negligence feature has sub-features like sleepy condition, intake of alcohol, speed, no/wrong pass, following other vehicle too closely, and cut in sharply after passing etc. All the other datasheets (type of road and junction, age and type of vehicle and defect) also contain the same type of data as the sub-features related to their feature.

Table 1 Description about datasets

Name of feature	Attributes of feature	Description	Outcome
Weather	Fine, Fog, Cloudy, Light rain, Heavy rain, Floody, Hail, Snow, Strong wind, Dust, Very hot and cold.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Driver's negligence	Speed, Intake of alcohol, wrong pass, Followed too closely, passed on Hill, Passed on curve, Cut in Sharply after passing, Improper passing, Wrong side of road, Failed to give signal.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Type of road	Single Lane, Double Lane, and Four Lanes with One median, Road with More Median.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Type of vehicle	Two-Wheelers, Three-Wheelers, Four-Wheelers, Bus, Trucks-Tempos-Tractors, Other Vehicles.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Type of junction	T-junction, Y-Junction, Four arms Junction, Junction with more than Four arms, Straggled Junction.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Age of vehicle	Less than 1 Year, Less than 5 year, 6-8 years,8-10 years, more than 10 year.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.
Vehicular defect	Defective brakes, Defective steering, punctured/Flat tyres, Bald tyres, Other Mechanical defects.	Collection of this dataset is done from records provided by Government of India during 2012-2016.	Calculate the total number of accidents with respect to each state and attributes.

For knowing the reasons who affect the road accident most, we have to analyses these datasheets collected from the Indian government portal data.gov.in; .This site is under observation of Indian Government.

This dataset can be used by other researchers for analysing some other features that affects the collisions. They may get some other conclusion, which will be beneficial in avoidance of the collisions.

In this analysis, Matlab is used for the statistical calculations and association rule implementation. Matlab is useful to store our growing patterns efficiently and matrix manipulation operations. It is used to find out the name of states and features and find the combination of features. Name of states that are facing maximum accidents, and features who affect them most. We have applied association rule mining to get the combination of feature which is most deadly for crash on road. Following are the steps which have been performed on data sets to get the objective.

IV. RESULT ANALYSIS

In this part result is shown, name of states which are facing maximum number of accidents and the features who affects the road accident most is explained by using bar graph. It also has the explanation on the combination of features in presence of which probability of collision increases. Accident means uncertain traffic collision so we can only analyse possibilities of reasons which may cause accident.

4.1 Top Three States Which Had Maximum Number Of Road Accidents

The features who affect the collision somehow are weather, vehicular defect, driver's involvement, type of vehicle, road defect, and type of junction and age of vehicle.

Every feature has its own record for number of accidents. These records are available by respective state names. As because of which feature, the state faces how much of accidents.

Figure 4.1 is plotted for showing the name of states by taking all the seven datasheets (age, type and defect in vehicle, type of road and junction, weather and driver's negligence) as input. Calculate percentage of accident state wise and find name of states where maximum accidents happen. The outcome of this process is shown in figure 4.1.

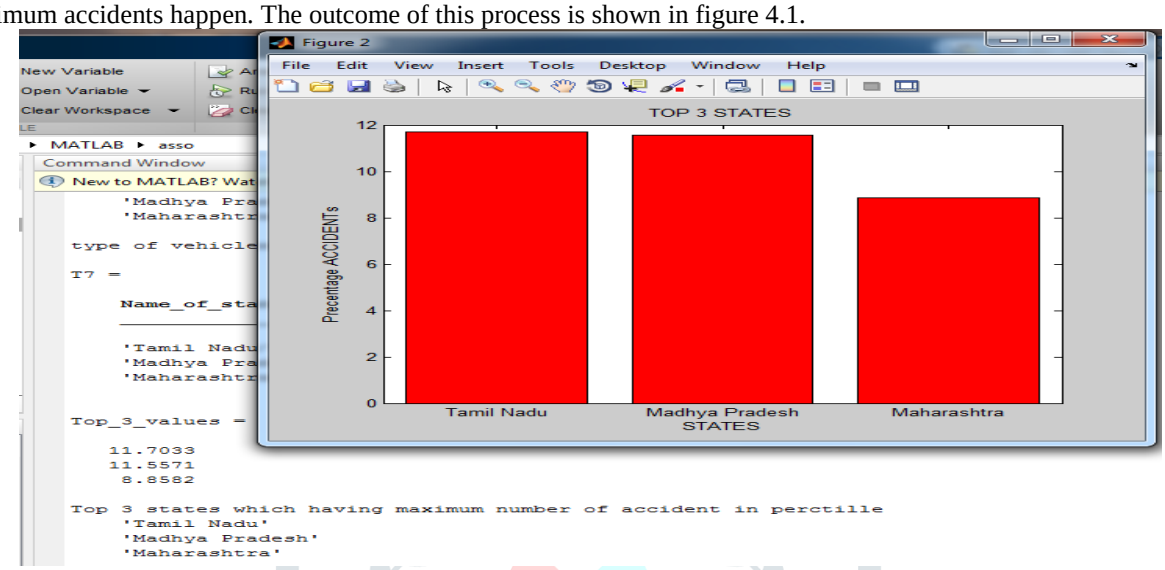


Figure 4.1: Bar Graph Showing States Having Maximum Number of Accidents

Figure 4.1 shows name of states which faces maximum number of accidents in India during 2012 to 2016. These states are Maharashtra, Madhya Pradesh and Tamil Nadu. For this calculation the features are considered like weather, road, vehicle and driver's condition. The percentage of accident faced by Tamil Nadu is 11.7033%, Madhya Pradesh is 11.5571% and Maharashtra 8.8582%.

Figure 4.2 is plotted for showing the name of states where most fatal accidents happen. By considering, accident data provided by Government of India as Input, we find out the percentage of deaths in each state of India. In this process, features like vehicle, driver and environment are considered.

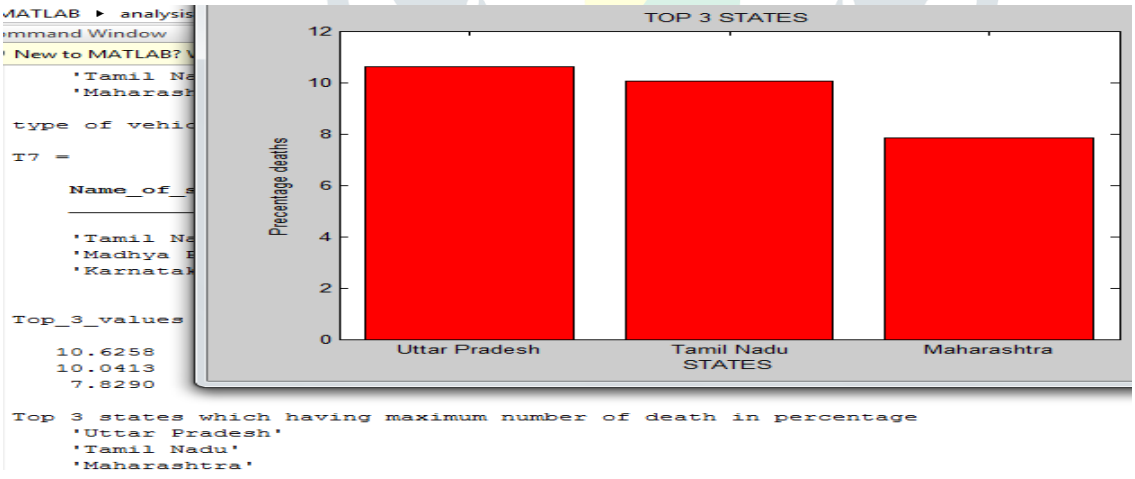


Figure 4.2: Bar Graph Showing States having fatal Accidents

Figure 4.2 shows the death percentage happens because of road accidents in India. Most deaths happen in Uttar Pradesh followed by Tamil Nadu and Maharashtra. Uttar Pradesh faces 10.6258%, Tamil Nadu faces 10.0413% and Maharashtra faces 7.8290% of deaths. That's shows that most fatal accidents happen in Uttar Pradesh.

By analysing figure: 4.1 and figure 4.2 we can say that maximum accidents happen in Tamil Nadu but most fatal accidents in which people die, happen in Uttar Pradesh.

4.2 Features Which Affect Road Accidents Most

Figure 4.3 is plotted by considering number of accidents happen because of sub-features as input. Sub-features belong to their particular features like vehicle, driver and environment. As driver's negligence is a feature and its sub-features are sleepy condition, intake of alcohol, speed, no/wrong pass, following other vehicle too closely, cut in sharply after passing etc.

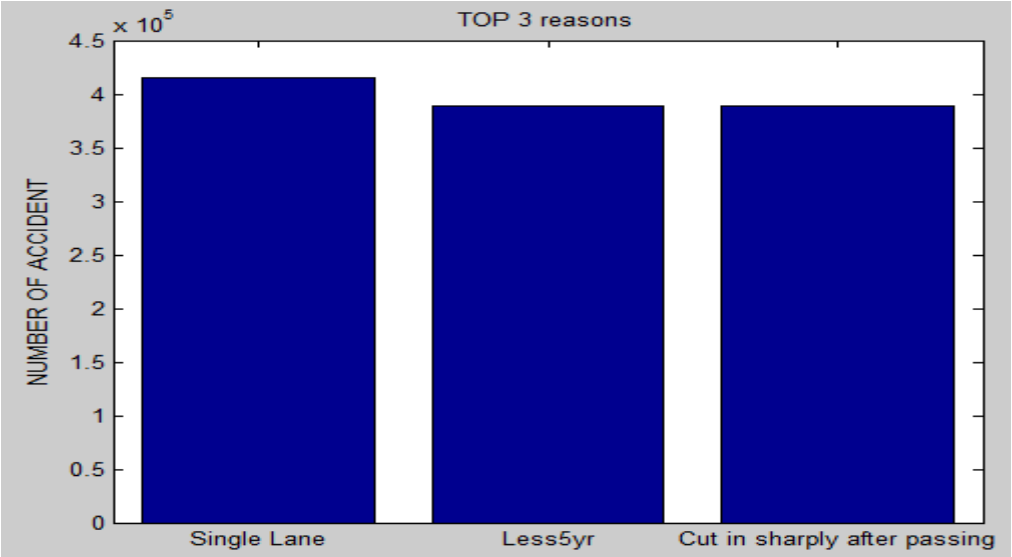


Figure 4.3: Bar Graph Showing Top three features

In figure 4.3 the bar graph is showing the name of factors which affect the road accidents most, these are single lane type of road, and vehicle age is less than 5 years and by driver's action when he/she cut in sharply after passing. This analysis presents that for being safe during travel, driver should avoid or should be extra careful while driving if it is single lane type of road, vehicles age is less than 5 years and driver should always be careful while passing another vehicle.

4.3 Finding Most Deadly Combination of Feature and Other Association

To find the combination of features, which causes more accident then other features, association between them is needed. Association means finding dependency, relationship or pattern between different data sets. Finding Association between features means relationship between features that are more accident prone.

By calculating number of accidents happen because these features with respect to one feature, relationship between them can be analysed. Accident because of any features in particular time period or with respect to one feature will represent the effect on accident.

By treating age of vehicle as a constant factor which will be present in every traffic collision, we have analysed the other factors whose probability is more in comparison to other factors.

We are considering when age of vehicle is less than 5 years then number of accidents happen in India most because of following reasons. In this analysis bar graph is plotted to analyse the most affecting factors.

```
Most deadly combination

R1all =

    'Less5yr'    ' T-Junction'
    'Less5yr'    'Exceeding lawful speed'
    'Less5yr'    'Single Lane'
    'Less5yr'    'Two-Wheelers'
    'Less5yr'    'Fine '
    'Less5yr'    'Other serious mechanical defect'
```

Figure 4.4: Result of Most Deadly Combination

Figure 4.4 showing us the association of factor with each other whose combination causes maximum accident. By analysing figure 4.4 we can say that if two-wheeler vehicle whose age is less than 5 years and if it is traveling through single lane T-junction road in fine weather and facing some mechanical defect than probability of facing road accident is maximum.

Figure 4.5 is plotted by taking the data of junction type associated with age of vehicle as input. Junction type is a feature and sub-features are T-junction, Y-junction, staggered junction etc. when number of accidents happen because of these will get associated with the data of age of vehicle, the outcome of this process is shown in figure 4.5.

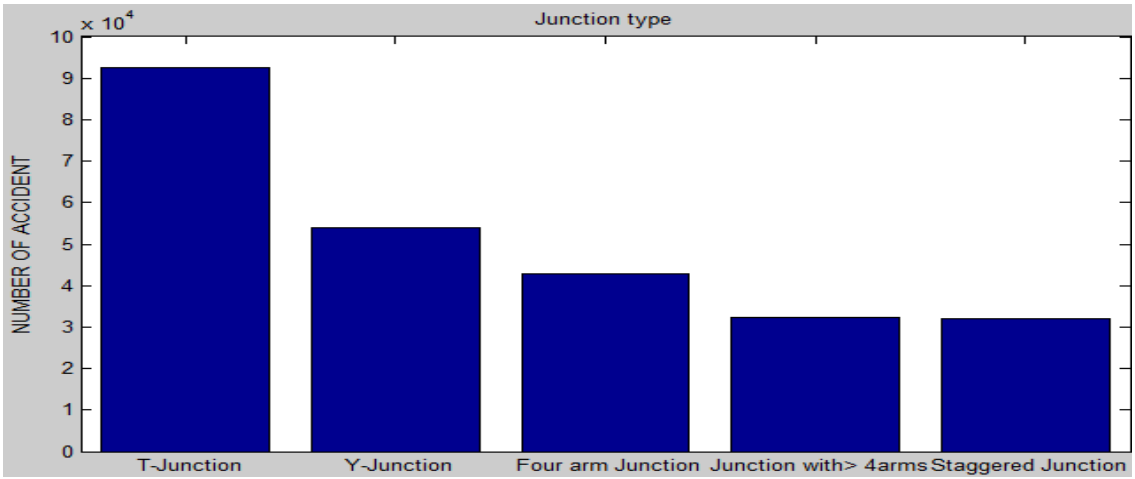


Figure 4.5: Bar Graph Showing Number of Accident Happen on Junction Type

Figure 4.5 is a bar graph plotted for accidents happen on junctions when age of vehicle is less than 5 years. That's shows T-junctions faces maximum number of accidents. In that series after T-junction, y junction, 4 arm junctions, junction with more than 4 arms and staggered junction pays role respectively in road accidents.

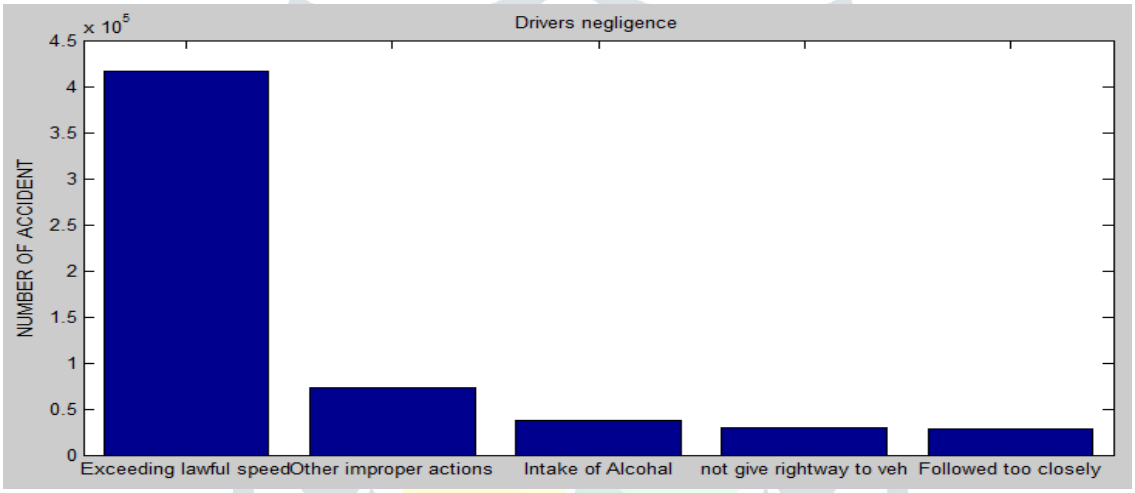


Figure 4.6: Bar graph Showing Number of Accident Happen because of Driver's Negligence

Figure 4.6 is plotted by considering driver's negligence feature associated with age of vehicle as input. The bar graph showing the numbers of accidents happen because of driver's negligence. In which exceeding lawful speed of vehicle causes maximum accident when age of vehicle is less than 5 years. After that accidents happen because of being drunk, no/ wrong pass given by driver to other driver and when driver follow some other vehicle too closely.

By considering, the number of accident because of road type associated with age of vehicle as input figure 4.7 gets plotted. Here road type is a feature and single lane, double lane, four lanes etc. are sub-features. When accidents happen on particular road type get associated with vehicle's age the outcome of this process is shown in the next bar graph.

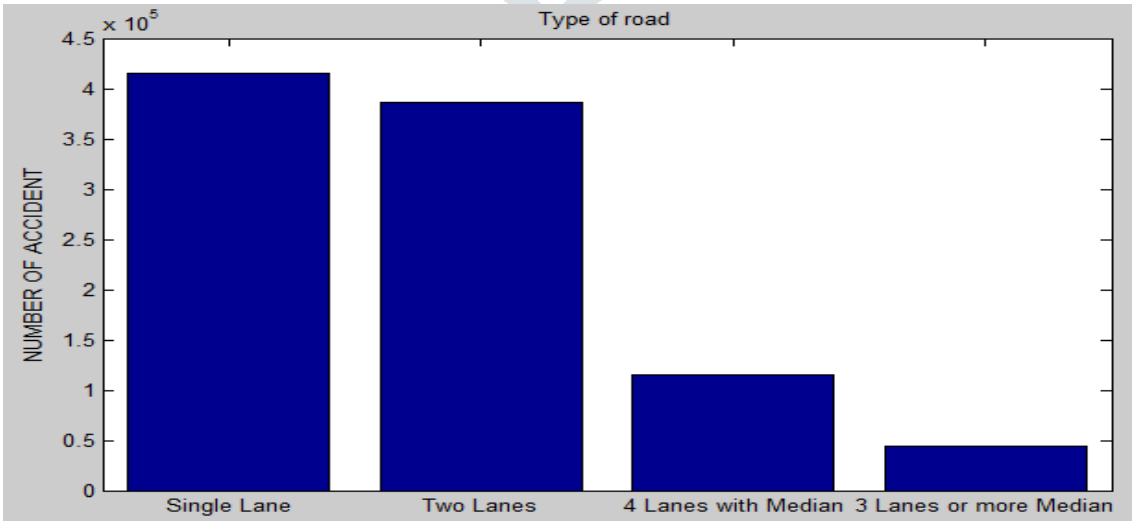


Figure4.7:Bar Graph Showing Type of Road Faces Number of Accidents

Figure 4.7 showing the number of accidents happen because of type of road. By analysing this bar graph we can say that most of the accidents happen on the single lane road after that two lane, four lanes with median and 3 lane of road structure plays role in counting of road accident.

Figure 4.8 is showing the outcome of number of accidents happen because of vehicle's type. For this, datasheet containing number of accidents happens because of type of vehicle is taken as input and plotted the graph with respect to age of vehicle when it is less than five years.

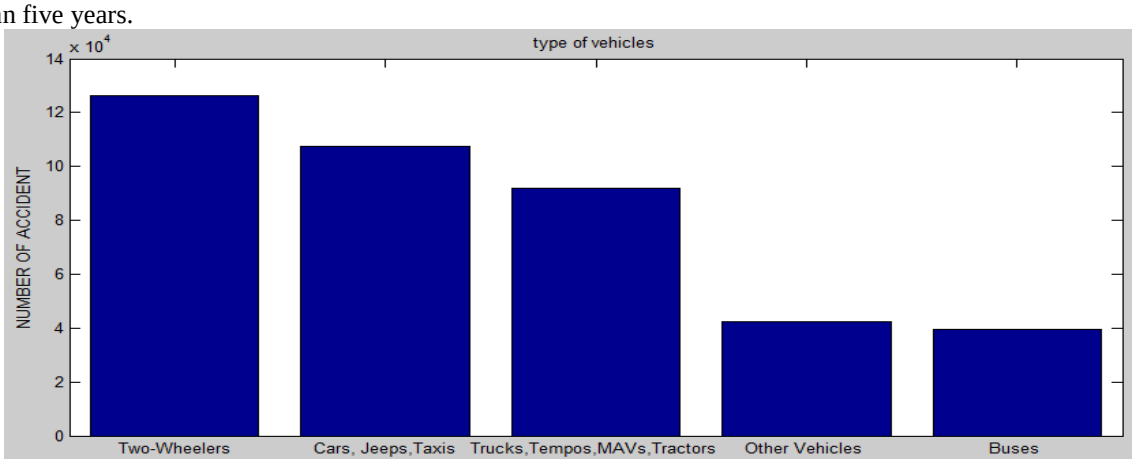


Figure 4.8: Bar Graph showing Number of Accidents Happen Because of Type of vehicle

Figure 4.8 shows the number of accidents happens when there were these specific types of vehicles were present when vehicles age was less than 5 years. This graph is showing that maximum accidents happen because of two-wheelers like bike, scooter etc. and after that 4 wheelers like cars, jeeps, taxis causes more accidents. Trucks, tempos, tractors and buses causes less accidents in comparison to others.

Figure 4.9 is plotted to analyse the effect of vehicle's defect on accidents. For that, datasheet who have record about accidents happen because of vehicle's defect is taken as input when age of vehicle is less than five years. Vehicle's good condition is very important while driving or traveling. If it is facing some kind of defect it may cause the accident. Defective Brake, steering, tyres or mechanical defect etc. are sub-features under vehicular defect feature.

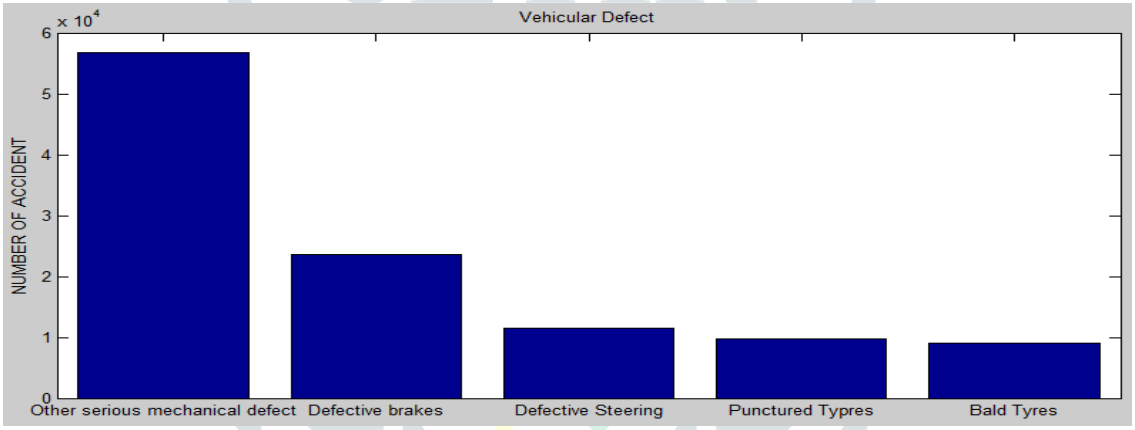


Figure 4.9: Bar Graph Showing Number of Accidents Happen Because of vehicular defect

Figure 4.9 showing the bar graph regarding vehicular defect causes number of accidents. By analysing this graph we can say that defective brake, defective steering, punctured tyres and bald tyres doesn't cause more accident in comparison to some mechanical defect.

Figure 4.10 is plotted by taking the accident record caused by weather condition as input. This analysis and plotting is done when age of vehicle is less than five years. Here weather is feature and foggy, rainy, fine, cloudy etc. are different sub-features. The outcome of this process will give us the name of sub-features who cause more accidents.

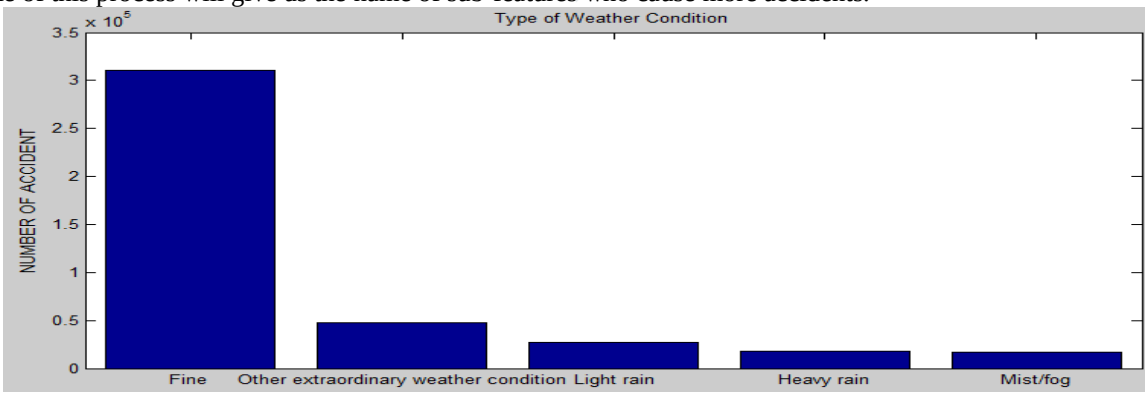


Figure 4.10: Bar Graph Showing Number of Accident Happen Because of Weather Condition

Figure 4.10 showing the weather condition in which maximum and minimum numbers of accidents happen. This bar graph is showing that maximum number of accident happen in fine weather because in that case sometimes driver become less attentive. Minimum accident happens when there is foggy weather out there. Heavy rain, light rain and some other extraordinary weather condition also cause considerable number of road accident.

V. CONCLUSION AND FUTURE SCOPE

By using association rule, the most deadly combination is two-wheeler vehicle whose age is less than 5 years and is traveling through single lane, T-junction road in fine weather and has some mechanical defect. By this research, the name of factors which affect the road accidents most are single lane type of road, vehicle age is less than 5 years and drivers' action when he/she cut in sharply after passing. States Tamil Nadu, Madhya Pradesh and Maharashtra face maximum number of accident with 11.7033%, 11.5571% and 8.8582% respectively in comparison to other states. And Uttar Pradesh, Tamil Nadu and Maharashtra face more fatal accidents and deaths with 10.6258%, 10.0413% and 7.8290% respectively in comparison to other stares.

We have achieved most deadly combination of features causing maximum number of accidents, likewise safest conditions can also be found by the analysis of the same data set. We have taken only few sub features of the selected feature. However there is a scope to consider other sub features for more generalized outcome.

REFERENCES

[1] Hermitte Thierry, "Review of Accident causation models used in Road Accident Research", January 2012.

[2] K Jayasudha and C Chandraseka, "An overview of data mining in road traffic and accident analysis", Journal of Computer Applications", 2(4):32–37, 2009.

[3]Sachin Kumar and DurgaToshniwal, "Analysing road accident data using association rule mining", In Proceedings of International Conference on Computing, Communication and Security, pages 1–6, 2015.

[4]Ravi Shenker, Arti Chowksey and Har Amrit singh sindhu, "Analysis of relationship between road safety and road design parameter of four lane national highway in India.", May 2015.

[5] A.N. Dehurey, A.K. Patnaik, A.K. Das et al., "Accident Analysis and Modelling on NH-55(India)", May 2013.

[6] U.S. Department of Transportation "Synthesis of Safety Research Related to Speed and Speed Limits", (PDF). Retrieved 5 March 2008.