A Framework for Analysis of Road Accidents

Shristi Sonal
Sathyabama University ,Chennai
shristisonal3012@gmail.com

Saumya Suman
SathyabamaUniversity,Chennai
saumyasuman96@gmail.com

***Abstract*- The road accident data analysis use data mining and machine learning techniques, focusing on identifying factors that affect the severity of an accident. There are a variety of reasons that contribute to accidents. Some of them are internal to the driver but many are external. For example, adverse weather conditions like fog, rainfall or snowfall cause partial visibility and it may become difficult as well as risky to drive on such roads. It is expected that the findings from this paper would help civic authorities to take proactive actions on likely crash prone weather and traffic conditions.**

Keywords- Data analysis, Machine learning, data, accident, vehicle, regression analysis, python

I. INTRODUCTION

Road safety becomes a major public health concern when the statistics show that more than 3000 people around the world succumb to death daily due to road traffic injury. In addition, road crashes lead to the global economic losses as estimated in road traffic injury costs to US$518 billion per year. The huge economic losses are an economic burden for developing countries. The road data are necessary not only for statistical analysis in setting priority targets but also for in-depth study in identifying the contributory factors to have a better understanding of the chain of events. There are a lot of Data Mining algorithms which are available to find out the association between independent variables in a huge data. The most popular and commonly used algorithm is Association rule mining. This can be used to detect the significant associations between the data stored in the large database. Apriori, predictive Apriori and FP-growth algorithm are the most common association rule mining methods which are used. The results obtained from these
data mining approach can help understand the most significant factors or often Repeating patterns. The generated pattern identifies the most dangerous roads in terms of road accidents and
necessary measures can be taken to avoid accidents in those roads.

II. RELATED WORK

- Hasan et al (2016) has published a paper with the title of Factors Contributing to Motorcycle Fatal Crashes on National Highways in India. It focus is only on the motorcycles.

- The objective of this work was to identify the factors that influence the motorcycle fatal crashes. It uses LR models to find its result.

- This algorithm helped to explore the various factors involved in the accidents. It ended up in finding the collision type, number of vehicles, number of lanes, and time of the crash were also observed to have a significant effect on motorcycle fatal crash.

- According to Sachin Kumar and Durga Toshniwal in "Analyzing Road Accident Data Using Association Rule Mining" (2015), data mining techniques can been used to analyze the data provided by EMRI (Emergency Management research Institute) in which the accident data is first clustered using K-modes clustering algorithm and further association rule mining technique is applied to identify circumstances in which an accident may occur for each cluster. Here, an Apriori algorithm has been applied on every cluster using WEKA3.6 to generate association rules

- In the further study by Sachin Kumar a new work was been proposed in the paper "Data mining approach to characterize road accident locations" (Sachin et al, 2016).The objective of this research was to identify various factors that affect road accidents. In this Association Rule Mining and k means clustering techniques were used. Both the techniques helped to identify locations at which the accident occurs frequently. The result was identified as high, moderate and low-frequency accident locations.

- In "Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States" (Wei et al, 2016), the Ordered probit model was used to analyze the effect of various factors on injury severity. It also focused on the weather conditions. It concluded that bad weather and road visibility found to increase the probability of injury severity.

- As most of the studies did not pay enough attention on the age of the drivers as a factor or cause for any kind of accidents, research titled "The relationship between age with driving attitudes &behaviors among older Americans "(Alexander *et al*, 2015) had an objective to identify the relationship between age with driving attitudes &behaviors. It used LR model that helped to identify the association between dependent and independent variables. It concluded that younger drivers engage more in unsafe traffic safety compared to older drivers

- "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques" (Liling*et al*, 2017) has used Apriori algorithms, Naive Bayes and k Means clustering algorithm to find out variables that are closely related to

fatal accidents. The performance of the algorithm was only 67.9%. It concluded that the environmental factors do not strongly influence the fatal rate, while the human factors have been a stronger influence on the fatal rate.

- In "Investigating injury severity risk factors in automobile crashes" by (Dursun et.al, 2017), the focus was been given to identify the person, vehicle, and accident related risk factors in automobile crashes. It used several methods which includes ANN, SVM, C5 and LR. The performance of ANN was 85.77% whereas SVM was 90.41%. C5 reached a level of 86.61% and LR was 76.96%. It concluded that use of seat belt, the manner of collision and drug usage are the top predictors of the injury severity.

- In Applying association rules mining algorithms for traffic accidents in Dubai" (El Tayebet.al [20]) uses Apriori and Predictive Apriori association rule mining algorithms to discover the links between accident factors & accident severity.The rule generated by Apriori algorithm was more effective than Predictive Apriori algorithms. Hence they concluded that the Apriori rule mining algorithm generates better rules than Predictive Apriori rules mining algorithms.

III. PROPOSED WORK

A. ***Data collection and preparation:***

For any data analysis, the most important aspect is the data. Collecting the right kind of data is very important. Analyzing and understanding the content and structure of the data needs special attention. The data used here for the analysis is been taken from https://data.gov.in. This website is also known as Open Government Data (OGD) Platform India. This portal is a single-point access to datasets, documents, services, tools and applications published by ministries, departments and
Organizations of the Government of India. This portal has a lot of information about many sectors.
Recorded details include the time, date, day, and location of the accident. It also includes the age and other details of the driver. It includes the severity of the accident and type of the vehicle. The weather conditions were also present in the data.
The presence of all this information made it easy to analyze the data and to derive conclusions from it.

B. ***Analysis and Implementation:***

Once the data is been collected, the task of analyzing it comes further. To analyze the data we need some tool which simplifies the work. We had a clear idea of using Python for coding. The reason to choose Python is the beauty of this language which simplifies a lot of work and makes things easier. Python has a lot of built-in packages which helps a lot in analysis. After a lot of research we came to know about Anaconda. Anaconda is a freemium open source distribution of Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. There are a lot of reasons to choose Anaconda. It has the most useful packages for Mathematics, Science and Engineering already installed for you. It contains porting for all the popular python libraries that can be used indata science. The most important being scikit-learn, numpy, pandas, scipy etc. Plus it also comes with the jupyter notebook and Ipython distribution. So, it saves you from importing numerous libraries separately.
We used Jupyter notebook for our analysis. The documents produced by the notebook contains both computer code (e.g. Python) and rich text elements like figures and graphs as well as executable documents which can be run to perform data analysis. The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access or can be installed on a remote server and accessed through the internet. In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a "Dashboard" (Notebook Dashboard), a "control panel" showing local files and allowing to open notebook documents or shutting down their kernels.
The packages which played a major role in the analysis are pandas and numpy. Pandas is used for data manipulation and
analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. .It provides high- performance, easy to use structures and data analysis tools.
NumPy stands for 'Numerical Python' or 'Numeric python'. It is open source module which provides fast computation on arrays and matrices. NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Now talking about the algorithm we used. There are a lot of algorithms present which help us in analyzing data. Machine learning and data analytics techniques are a boon in this field. The algorithm we opted for is Regression Analysis.

Regression Analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between dependent variable and one or more independent variables.

More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. There are many types of regression analysis. The one we used is Linear Regression.

Linear regression is a linear approach for modelling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X*. The case of one explanatory variable is called simple linear regression. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Finally talking about the flow of work. Once the data was collected the data was been imported through pandas and the preprocessing was done. The processed data was been analyzed and on the basis of the analysis the graphs were plotted. These graphs depicted the conclusions which can be derived from the analysis.

IV. Result and Discussion

The analysis was done considering various attributes. These attributes played important role in the occurrence of any accident. Based on the analysis the graphs are been plotted. These graphs gives a clear idea about the conclusions drawn from the analysis.

Table 1: Road accidents attributes used for analysis

ATTRIBUTE	TYPE	VALUES
No. of victims	Nominal	1,2,>2
Date of accident	Nominal	1,2,…31
Time of accident	Nominal	1,…12
Road type	Nominal	Highway/other
Weather condition	Nominal	Cloudy/moderate
Lighting on road	Binary	Daylight/Streetlight
Accident severity	Nominal	Critical/non-critical
Age group	Nominal	Children,Young,Adult,Senior

Table 2: Set of attributes and their values

Accident Index	Police force	Accident Severity	No.of victims	No. of casulties	Date	Time	Road Type	Weather condition
0	1	2	1	1	4-01-2005	17:42	2	1
1	1	3	1	1	5-01-2005	17:36	1	4
2	1	3	2	1	6-01-2005	0:15	1	4
3	1	3	1	1	7-01-2005	10:35	1	1
4	1	3	1	1	10-01-2005	21:13	1	7

The various attributes and the graphs are as follows:

1. *Based on type of region :*

According to the data we had, we could classify the accidents broadly in two types based on its region – Urban and Rural. The analysis showed the obvious results that the number of accidents occurring in the urban regions are more than the accidents occurring in rural regions.

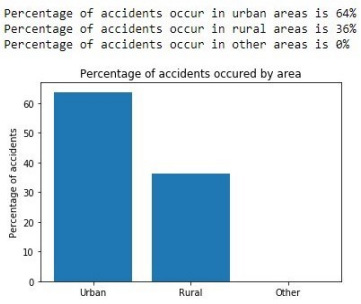

Fig: Analysis based on Type of region

Hence, the percentage of accidents occurred in urban areas consist of 64% whereas the rural areas consist of 36%.
Hence, we can conclude that urban areas are more prone to accidents than rural areas.

2. *Time :*

In today's world there is no time when we can't see vehicles on the road and accidents can occur any time. So we can't say that any particular time is dangerous or safe for driving

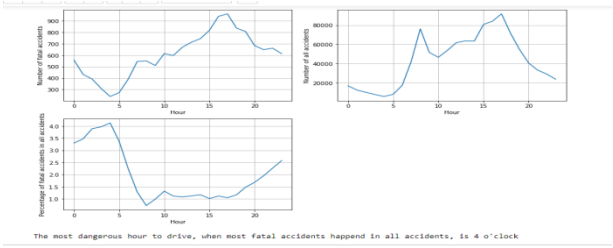

Fig: Analysis based on Time

In our data, morning 4 o'clock has proven to be one of the most dangerous time

3. *Gender of the driver :*

In the modern world where equality between men and women has taken place in almost everywhere, women are into driving as well.

Though the number of women drivers are still less than that of men.

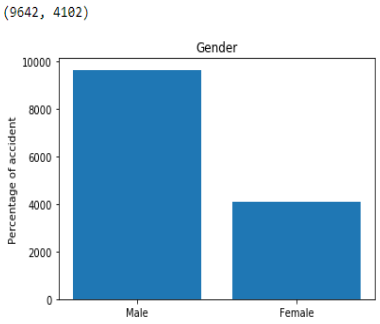

Fig: Analysis based on Gender of the driver

The analysis shows that the number of male drivers involved in accidents are more than that of female.

4. *Number of accidents :*

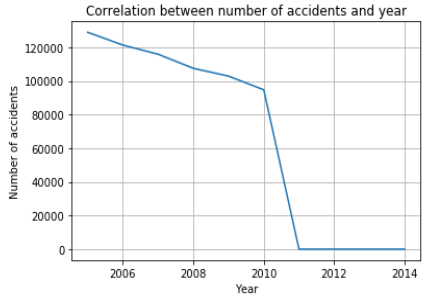

Fig: Analysis based on Number of accidents

Here, graph is plotted in order to view the result for the trend of the accident over the years.

The above result shows that the number of accidents occurring has been reduced over time.

5. *Speed-limit :*

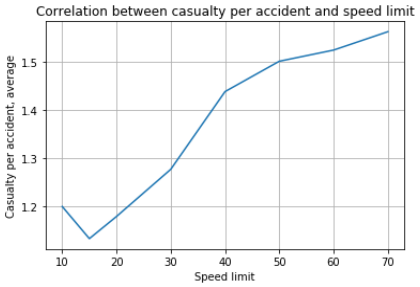

Fig: Analysis based on Speed-limit of vehicle

Here, the graph is plotted in order to view the result for the speed- limit of vehicles facing accidents.

The areas that allow high speed limits have faced more number of accidents than the areas with less speed limits.

6. *Age :*

We can see a diversity in terms of age of the drivers. People of almost every eligible age drives.

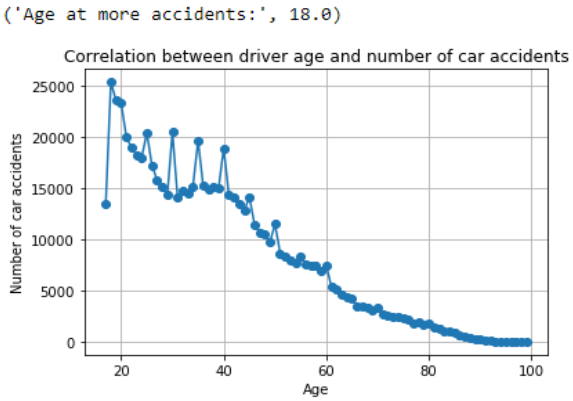

Fig: Analysis based on Age

Here, the graph is plotted in order to view the result for the driver's age which is more prone to road accidents.

The above graph shows that the young drivers of 18 are more prone to accidents.

V. Conclusion

In this study, the technique of regression with a large set of accident's data to identify the reasons of road accidents were used. Analysis is done for the identification of factors involved in the accident that occur together which is then plotted in a graph form. This shares a lot in understanding the circumstances and causes of accident. This ultimately helps the Government to adapt the traffic safety policies with different types of accidents and situations. The main result for this study is that although the characteristics of humanity and behavior are very important in occurrence of all road accidents. But we can understand that spatial features and infrastructure play a major role in the accident. In this study, it is tried to choose various number of attributes to provide a lot of valuable information for the government to provide better safety policies.

This article can be a step forward towards providing useful information for highway engineers and transportation designers to design safer roads.

References

[1] Naqvi, H. M., & Tiwari, G., (2017), Factors Contributing to Motorcycle Fatal Crashes on National Highways in India, Transportation Research Procedia, 25, 2089-2102.

[2] Sachin Kumar , Durga Toshniwal ,"Analyzing Road Accident Data Using Association Rule Mining", IEEE 2015 International Conference on Computing, Communication and Security (ICCCS)

[3] Kumar, S., & Toshniwal, D, (2016), A data mining approach to characterize road accident locations, Journal of Modern Transportation, 24(1), 62-72.

[4] Hao, W., Kamga, C., Yang, X., Ma, J., Thorson, E., Zhong, M., & Wu, C., (2016), Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States, Transportation research part F: traffic psychology and behavior, 43, 379-386.

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

[6] El Tayeb, A. A., Pareek, V., & Araar, A. (2015). Applying association rules mining algorithms for traffic accidents in Dubai. International Journal of Soft Computing and Engineering.

[7] Bahram Sadeghi Bigham ,(2014),ROAD ACCIDENT DATA ANALYSIS: A DATA MINING APPROACH, Indian Journal Of Scientific Research 3(3):437-443.

[8] Divya Bansal, Lekha Bhambhu, "Execution of Apriori algorithm of data mining directed towards tumultuouscrimes concerningwomen", *International Journal of AdvancedResearch in Computer Science and Software Engineering*, vol. 3, no. 9, September 2013.

[9] K Jayasudha, C Chandrasekar, "An overview of data mining in road traffic and accident analysis", Journal ofComputer Applicatio*ns*, vol. 2, no. 4, pp. 32-37, 2009.

[10] S. Krishnaveni, M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques", *International Journal of Computer Applications*, vol. 23, no. 7, pp. 40-48, June 2011.