IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Analysis and Prediction of Road Accident using machine learning techniques

Mr. Nagesh U B[1] Rekha Halli[2] Methish R[3] Roopashree J[4] Nagashree S[5]

Assistant Professor[1] UG Students[2, 3, 4, 5]
Department of Information Science and Engineering
Alva's Institute of Engineering and Technology

***Abstract* – Road accidents are one of the most relevant causes of injuries and death. This is also one of the serious issues, which can possibly cause disabilities, injuries and even fatalities. There are many of reasons that contribute to accidents. Some of them are internal to the driver but many are external. For example, adverse weather conditions like rain, cloudy, and fog cause partial visibility and it may become difficult as well as risky to drive on such roads. This paper aims to provide an Overview of the area of the art in the prediction of road accidents through clustering techniques and machine learning algorithms.**

***Index Terms* – road accident data, Machine learning, K-means Clustering, Analysis, Visualization, prediction.**

I. INTRODUCTION

According to the death statistics released by the World Health Organization, the number of traffic accidents occurring annually in the world is alarming. And Road and accidents are uncertain incidents. In now a day's traffic is increasing at a huge rate which leads to a large numbers of road accidents. So, Road accident prediction is one of the most important research areas in traffic safety. The occurrence of road traffic accidents is mainly affected by geometric characteristics of road, traffic flow, characteristics of drivers and environment of road. Many studies have been conducted to predict accident frequencies and analyze the characteristics of traffic accidents, including studies on hazardous location/hot spot identification, accident severities. Some studies focus on mechanism of accidents. Other factors include weather and light conditions of the road. No specific approach available for the traffic police to predict which area is accident prone. The road accident prediction play an important role in the integrated planning and management of traffic, the reason which with much randomness about the traffic accident include some nonlinear elements, such as people, car, road, climate and so on. Machine Learning algorithms can process large number of classification parameters and are able to obtain useful patterns. It can process huge amounts of data efficiently and can be scalable. And also the clustering technique is also helps in analysing and visualizing the road accident data.

II. RELATED WORK

Sachin Kumar et al. [1] , used data mining techniques to identify the locations where high frequency accidents are occurred and then analyze them to identify the factors that have an effect on road accidents at that locations. The first task is to divide the accident location into k groups using the k-means clustering algorithm based on road accident frequency counts.

Tessa K. Anderson et al. [2] proposed a method of identifying high-density accident hotspots, which creates a clustering technique that determines that stochastic indices are more likely to exist in some clusters, and can therefore be compared in time and space. The kernel density estimation tool enables the visualization and manipulation of density-based events as a whole, which in turn is used to create the basic spatial unit of the hotspot clustering method.

Shristi Sonal and Saumya Suman. [3]The road accident data analysis use data mining and machine learning techniques, focusing on identifying factors that affect the severity of an accident. 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.

E. Suganya,S. Vijayrani In "Analysis of road accidents in India using data mining classification algorithms" [4], Classification is a model finding process which is used for segmenting the data into different classes based on some constraints. This work analyzes the road accidents in India data set using classification algorithms namely linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. Performance measures used are accuracy, error rate and execution time.

III. PROPOSED SYSTEM

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 Kaggle and government websites.

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 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.

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.

Logistic 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.

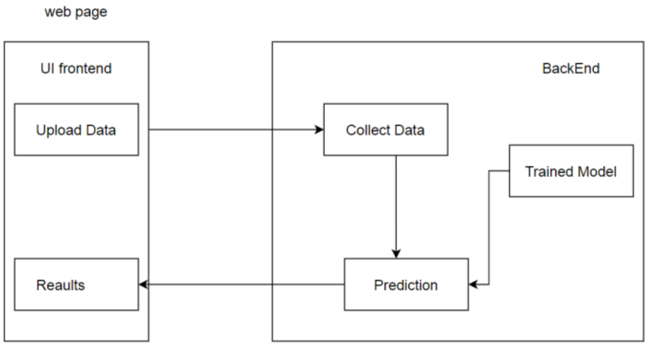


Fig 1 : Architecture

Scikits-learn

Scikit-learn is a free software machine learning library for the Python programming language.

- ➢Simple and efficient tools for data mining and data analysis
- ➢Accessible to everbody, and reusable in various contexts
- ➢Built on Numpy, SciPy, and matplotlib
- ➢Open source, commercially usable

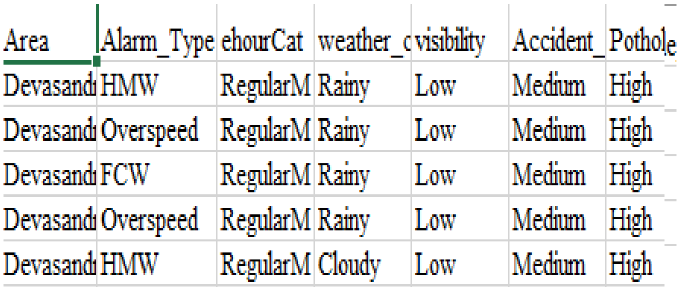
NUMPY

NumPy is an open source library available in Python that aids in mathematical, scientific, engineering, and data science programming. It works perfectly well for multi-dimensional arrays and matrices multiplication. NumPy is a programming language that deals with multi-dimensional arrays and matrices. On top of the arrays and matrices, NumPy supports a large number of mathematical operations.

NumPy is memory efficiency, meaning it can handle the vast amount of data more accessible than any other library. Besides, NumPy is very convenient to work with, especially for matrix multiplication and reshaping. On top of that, NumPy is fast. In fact, TensorFlow and Scikit learn to use NumPy array to compute the matrix multiplication in the back end.

ROAD ACCIDENT DATA SOURCES

We got the datasets from kaggle and other sources. In this step contains the attributes like area, alarm type, visibility, ecarttime, weather condition, accident severities and pothole severities. And alarm type data is collected by CAS (collision on avoidance system) device for particular area, time, weather condition and visibility. Accident severities and pothole severities and pothole severities taken from government sites Fig 2.



Area	Alarm_Type	ehourCat	weather_c	visibility	Accident_	Pothole
Devasand	HMW	RegularM	Rainy	Low	Medium	High
Devasand	Overspeed	RegularM	Rainy	Low	Medium	High
Devasand	FCW	RegularM	Rainy	Low	Medium	High
Devasand	Overspeed	RegularM	Rainy	Low	Medium	High
Devasand	HMW	RegularM	Cloudy	Low	Medium	High

Fig 2: screenshot of some of the datasets

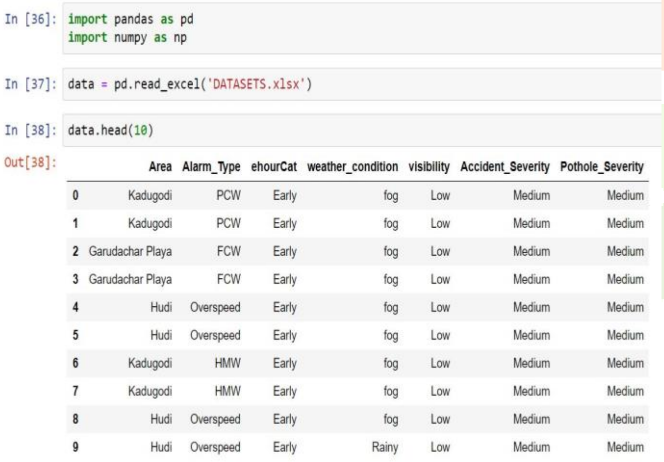
IV. IMPLIMENTATION AND RESULTS

Models are created using accident data records which can help to understand the characteristics of many features like, roadway conditions, weather conditions and so on. This can help the users to compute the safety measures which is useful to avoid accidents. It can be illustrated how statistical method based on directed graphs, by comparing two scenarios based on out-of-sample forecasts. the model is performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it.

Here the road accident study is done by analyzing some data by giving some queries which is relevant to the study. The queries like what is the most dangerous time to drive, what fractions of accidents occur in rural, urban and other areas. What is the trend in the number of accidents that occur each year, do accidents in high speed limit areas have more casualties and so on. These data can be accessed using Microsoft excel sheet and the required answer can be obtained. This analysis aims to highlight the data of the most importance in a road traffic accident and allow predictions to be made.

Data Importing:

Here, we are imported the data(fig 3) to perform analysis on this data. This data is consisted some attributes like area, alarm type, ehourcat, weather condition, visibility, accident severity and pothole severity. And data.head(10) views top 10 row of data frame.

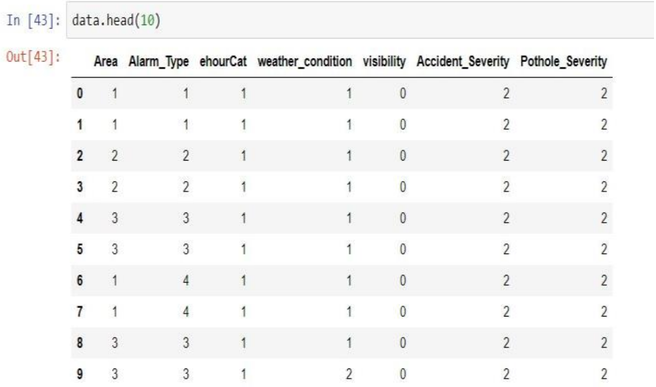


```
In [36]: import pandas as pd
         import numpy as np

In [37]: data = pd.read_excel('DATASETS.xlsx')

In [38]: data.head(10)

Out[38]:
```

	Area	Alarm_Type	ehourCat	weather_condition	visibility	Accident_Severity	Pothole_Severity
0	Kadugodi	PCW	Early	fog	Low	Medium	Medium
1	Kadugodi	PCW	Early	fog	Low	Medium	Medium
2	Garudachar Playa	FCW	Early	fog	Low	Medium	Medium
3	Garudachar Playa	FCW	Early	fog	Low	Medium	Medium
4	Hudi	Overspeed	Early	fog	Low	Medium	Medium
5	Hudi	Overspeed	Early	fog	Low	Medium	Medium
6	Kadugodi	HMW	Early	fog	Low	Medium	Medium
7	Kadugodi	HMW	Early	fog	Low	Medium	Medium
8	Hudi	Overspeed	Early	fog	Low	Medium	Medium
9	Hudi	Overspeed	Early	Rainy	Low	Medium	Medium

Fig 3 : Datasets

Converting the data into numerical form:



```
In [43]: data.head(10)

Out[43]:
```

	Area	Alarm_Type	ehourCat	weather_condition	visibility	Accident_Severity	Pothole_Severity
0	1	1	1	1	0	2	2
1	1	1	1	1	0	2	2
2	2	2	1	1	0	2	2
3	2	2	1	1	0	2	2
4	3	3	1	1	0	2	2
5	3	3	1	1	0	2	2
6	1	4	1	1	0	2	2
7	1	4	1	1	0	2	2
8	3	3	1	1	0	2	2
9	3	3	1	2	0	2	2

Fig 4 : Numerical Data

Here, we are converted the data(fig 4) into the numerical form because as we know when we train a machine learning model it should be in number. And this data.head(10) is views top 10 rows of data frame that which data is formed in number.

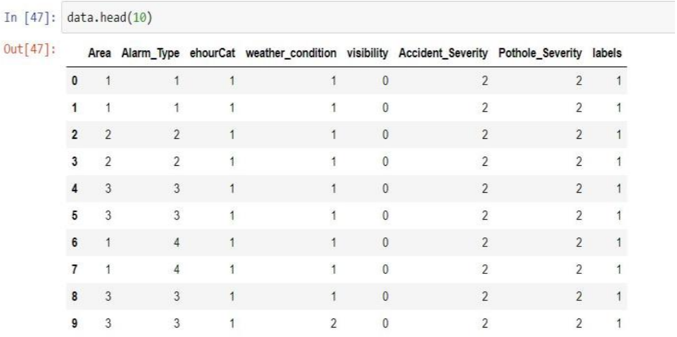
Used k-means clustering algorithm:

After converted the datasets into numerical form then we applied k-means cluster technique because our datasets are not consisted the labels or results like yes or no at the end. So, the datasets are unlabelled. Hence we are going with unsupervised learning approach. So, after using the unsupervised learning approach we predicted the results of road accident prediction zone at the end last column that's in high/low using k-means clustering technique.

Kmeans Algorithm :

1. Select K points as the intial centroids.
2. Repeat,
3. From K clusters by assigning all points to the closest centroid,
4. Re-compute the Centroid of each cluster,
5. Until the centroid don't change.

Clustering:



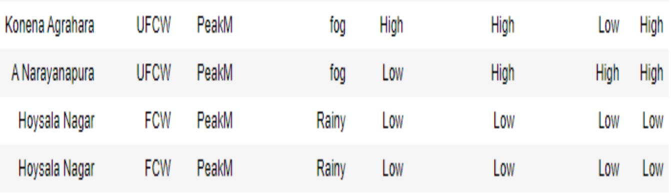
```
In [47]: data.head(10)

Out[47]:
```

	Area	Alarm_Type	ehourCat	weather_condition	visibility	Accident_Severity	Pothole_Severity	labels
0	1	1	1	1	0	2	2	1
1	1	1	1	1	0	2	2	1
2	2	2	1	1	0	2	2	1
3	2	2	1	1	0	2	2	1
4	3	3	1	1	0	2	2	1
5	3	3	1	1	0	2	2	1
6	1	4	1	1	0	2	2	1
7	1	4	1	1	0	2	2	1
8	3	3	1	1	0	2	2	1
9	3	3	1	2	0	2	2	1

Fig 5 : Numerical Labeled Data

Here, you can see the result that's after we applied k-means clustering we got the result of accident prediction zone is high or low(fig 5) at the and last column as labels. That's result in numerical form. Here we have taken High as (0) and Low as (1).

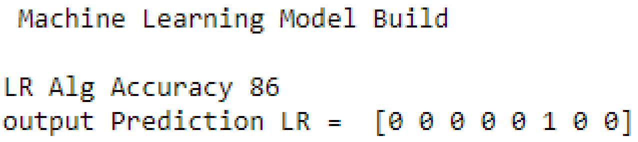
And also below table(fig 6), you can see the result in dataset form in String.



Konena Agrahara	UFCW	PeakM	fog	High	High	Low	High
A Narayanapura	UFCW	PeakM	fog	Low	High	High	High
Hoysala Nagar	FCW	PeakM	Rainy	Low	Low	Low	Low
Hoysala Nagar	FCW	PeakM	Rainy	Low	Low	Low	Low

Fig 6: Labeled datasets

After clustering we applied Logistic Regression algorithm for prediction which gave 86% accuracy(fig 7).



```
 Machine Learning Model Build

LR Alg Accuracy 86
output Prediction LR =  [0 0 0 0 0 1 0 0]
```

Fig 7 : Logistic Regression Accuracy

Visualization:

```
ehourCat  Early  PeakM  RegularM
labels
High         78     76       143
Low          21     43        38
```

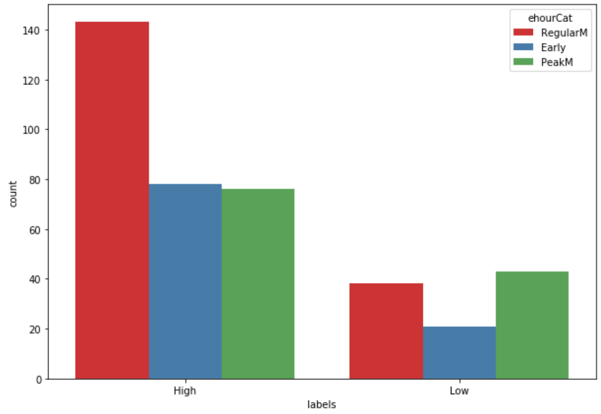


Fig 8: Graph of Labels vs Ehourcart

Here, this graph is formed based on the ehourCat(fig 8) it's shows the result of counts rate based on three attributes of ehourCat like Early, PeakM and RegularM. There is three colors red, blue and green. That's based on the result of low and high. And here red is indicated early, blue is indicate PeakM and green as RegularM.

```
weather_condition  Cloudy  Rainy  fog
labels
High                  101    100   96
Low                    44     29   29
```

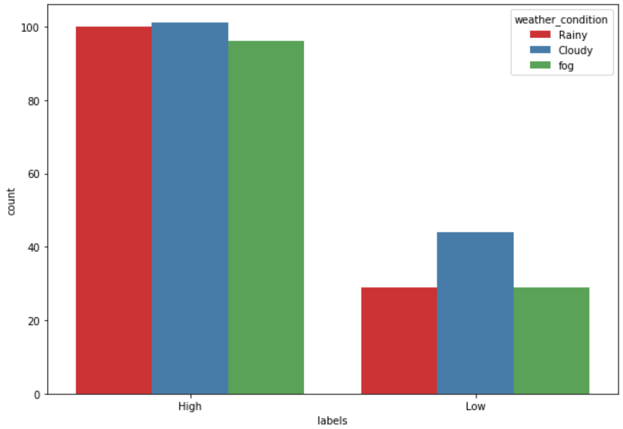


Fig 9: Graph of Labels vs weather_condition

Here, this graph(fig 9) is shows the results of counts and this plotted based on the weather condition. Here red colour indicating the fog, blue is indicating rainy, and green as cloudy.

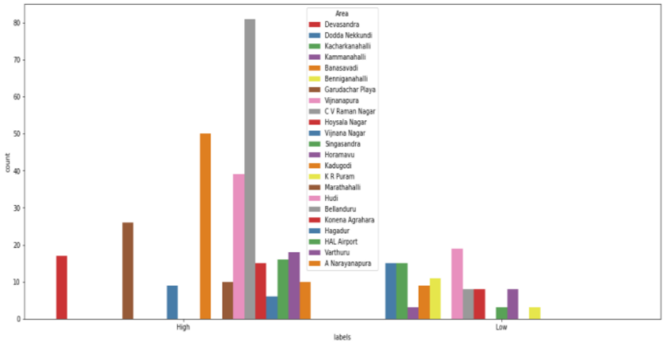


Fig 10: Graph of Labels vs Areas

Here, this graph(fig 10) is shows the results of counts based on areas.

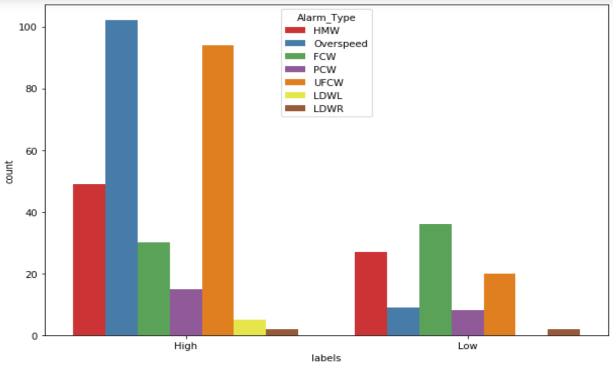
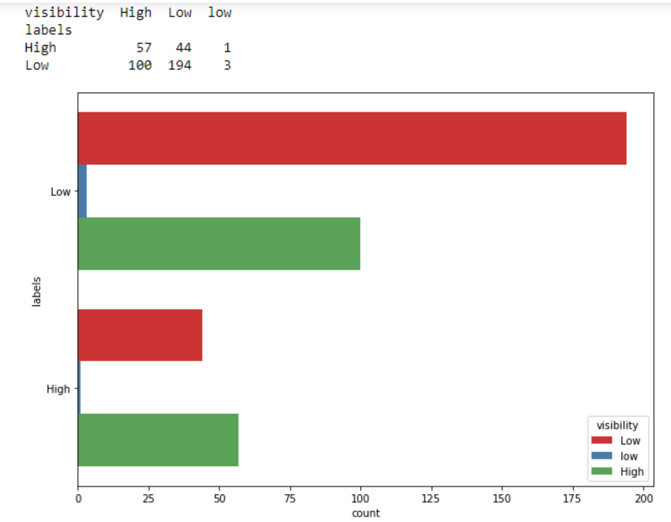


Fig 11: Graph of Labels vs Alarm_Type

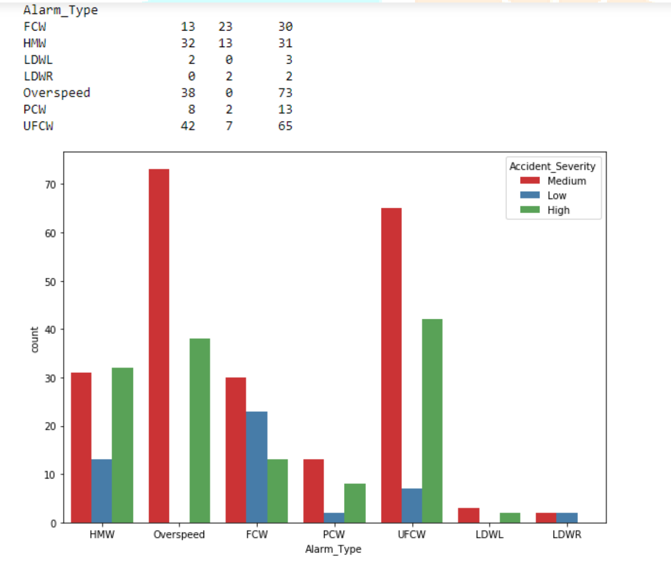
Here, this graph(fig 11) is plotted based on the Alarm types like FCW, HMW LDWL, LDWR, over speed, PCW, UFCW.



```
visibility  High  Low  low
labels
High          57   44    1
Low          100  194    3
```

Fig 12: Graph of Labels vs Visibility

This is the visibility graph(fig 12) and it's also shows the results of counts based on results of labels low and high. Here the red colour is indicating the results of low and blue is indicating the results of high.

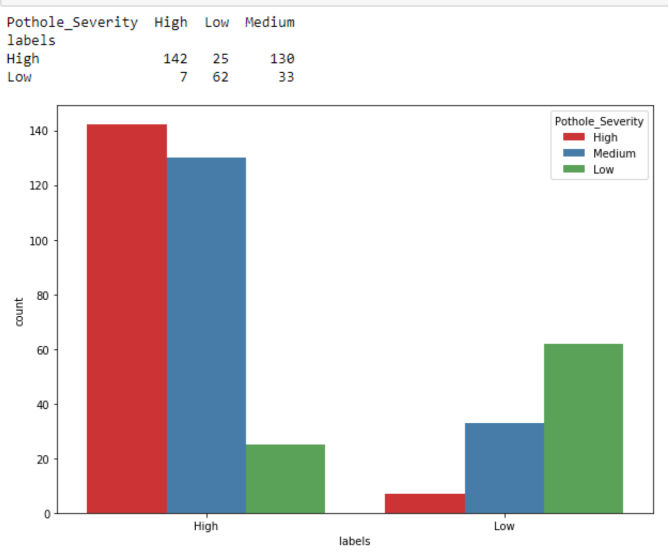


```
Alarm_Type
FCW              13   23    30
HMW              32   13    31
LDWL              2    0     3
LDWR              0    2     2
Overspeed        38    0    73
PCW               8    2    13
UFCW             42    7    65
```

Fig 13: Graph of Alarm_Type vs Accident_Severity

Here, in this graph(fig 13) the Accident Severity like medium, high and low. versus Alarm_type like FCW, HMW, LDWL, LDWR, Overspeed, PCW and UFCW

.

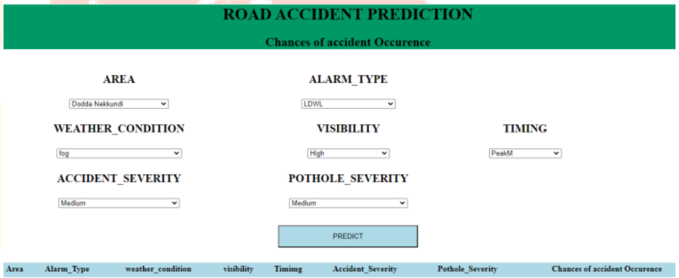


```
Pothole_Severity  High  Low  Medium
labels
High               142   25     130
Low                  7   62      33
```

Fig 13: Graph of Labels vs Pothole Severity

Here, in this graph(fig 13) the Pothole Severity like medium, high and low. And it's also shows the results of counts based on results of labels low and high. Here the red colour is indicating the result of medium and blue is indicating the results of high, and green as low.

Web page :



ROAD ACCIDENT PREDICTION

Chances of accident Occurence

AREA: Dodda Nekkundi

ALARM_TYPE: LDWL

WEATHER_CONDITION: fog

VISIBILITY: High

TIMING: PeakM

ACCIDENT_SEVERITY: Medium

POTHOLE_SEVERITY: Medium

Area	Alarm_Type	weather_condition	visibility	Timimg	Accident_Severity	Pothole_Severity	Chances of accident Occurence

Here, we have to give inputs and have to click the predict button ,once predict button is clicked we will get the chances of accident.

V. CONCLUSION

In this paper we have used the k-means clustering it's an unsupervised learning which is used for the unlabelled data therefore data are not labelled into any group of cluster. And also in this study, the techniques 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. And this ultimately helps the Government to adapt the traffic safety policies with different types of accidents and situations.

REFERENCES

[1] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations", J. Mod. Transport. 24(1):62–72..

[2] Tessa K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots", Accident Analysis and Prevention 41,359–364.

[3] Shristi Sonal and Saumya Suman "A Framework for Analysis of Road Accidents" Proceedings of International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)

[4] Analysis of road accidents in India using data mining classification algorithms- E. Suganya,S. Vijayrani.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[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.