iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** III **Month of publication:** March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33280

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

Road Traffic Accident Analysis and Visualization of Accident Prone Areas

Megna T Sreedhar[1], Karthika K P[2], Linda Sara Mathew[3]
[1, 2]*M. Tech Students,* [3]*Assistant Professor, Department of Computer Science & Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala*

Abstract: ***Road accident has a huge impact on society where there is a great cost for fatalities and injuries. Analysing the road accidents and thereby finding a solution is an emerging research area. The project mainly deals with the analysis of road accidents that happened in UK. Initially, the attributes such as sex, age, weather conditions, light conditions etc. were analysed and its effect over the road accident was analysed and found out their relationship according to certain conditions. This can be implemented using heatmap, boxplot, histogram etc. At the second level, a logistic regression model was used to explain the effect of weather and road surface condition towards the accident severity, through variables retrieved from the database. Also, the major accident-prone areas can be identified and plot it using Base map. Apart from this, we extend the functionality of existing services of the Google maps using Geojson to visualize accident prone areas.***
Keywords: Data Mining, Machine Learning, Road Accident Analysis, Correlation, Logistic Regression

I. INTRODUCTION

Road safety becomes a significant public health concern once the statistics shows that over 3,000 people in the world dies daily because of road traffic injury. The collection and use of correct and comprehensive information associated with road accidents is extremely necessary to road safety management. The road accident data are for in-depth study in identifying the contributory factors to have a better understanding of the chain-of-events.

The identification of factors affecting road crashes [1], [2] obtained from the crash investigation and reconstruction has not been conducted in practice in the almost all the countries. The goal of this study was to analyse the road accident in UK by using the dataset on road accidents that happened in UK from the year 2005 to 2015. This can lead us to several conclusion on the trend of road accidents happened there.

Information about the driver, vehicle, weather, roadway, and environment were obtained from the dataset. Today, one of the main works of government is the traffic safety. Pointing out the vitalness of topic, identifying the reasons of road accidents has become the main focus to reduce the harm caused by traffic accidents [3]. The data mining outcomes will help the various organizations such as transportation, to inquire the accidents data recorded by the police information system, discover hidden patterns and trends thus predict the future consequences.

Efficient and effective decisions are taken to lessen accidents. We all know that, the exponentially increasing amounts of data being generated each year make getting useful information from that data more and more critical. Analysis of such data become necessary to understand the core information. The process to make this happen includes simple query and reporting, statistical analysis, more complex multidimensional analysis, and data mining. In the case of road accident analysis, we used a dataset that includes more than 42 lakhs tuples with more than 100 attributes.

Categorizing the major aspect of traffic collisions and their severity assists the highway safety development in improving the road conditions according to the variation in vehicle occupancy need of specific sector of the population. Correlation analysis examines many of the attributes that help to derive the relation between two numerical variables. In this, we have done many correlation analysis in order to reach several conclusions over the road accidents.

In this initially, the attributes such as sex, age, weather conditions, light conditions etc. were analysed and its effect over the road accident was analysed and found out their relationship according to certain conditions. This has been implemented using heatmap, boxplot etc. At the second level, a logistic regression model was used to explain the effect of weather and road surface condition towards the accident severity through variables retrieved from the database. Logistic regressor used here can predict the accident severity for a particular weather and road surface condition. In that last section, we have identified the major accident-prone areas in UK and plotted it in the Base map and Google map using Geojson.

II. AREA OF RESEARCH

A. Data Mining

Data mining [4] is that the method of finding anomalies, patterns and correlations among massive knowledge sets to predict outcomes. Employing a broad vary of techniques, you'll be able to use this data to extend revenues, cut costs, improve client relationships, scale back risks and additional. It is the process of excavation through knowledge to find hidden connections and predict future trends features. Generally, it is brought up as "knowledge discovery in databases", the term "data mining" wasn't coined till the Nineteen Nineties. However, its foundation contains 3 tangled scientific disciplines: statistics (the numeric study of knowledge relationships), artificial intelligence (human-like intelligence displayed by package and/or machines) and machine learning (algorithms which will learn from knowledge to form predictions). What was previous is new once more, as data processing technology keeps evolving to stay pace with the limitless potential of huge knowledge and cheap computing power.

Over the last decade, advances in process power and speed have enabled North American nation to maneuver on the far side manual, tedious and long practices to fast, straightforward and automatic knowledge analysis. The additional complicated the information sets collected, the additional potential there is to uncover relevant insights. Retailers, banks, makers, telecommunications suppliers and insurers, among others, square measure victimisation data processing to find relationships among everything from valuation, promotions and demographics to however the economy, risk, competition and social media square measure moving their business models, revenues, operations and client relationships.

Data Mining may be a general class of techniques that may be applied to totally different sorts of datasets, similar to programming may be a general class of techniques that may be applied victimization totally different languages to try and do various things. Data processing rely upon analysing throughout time, during this case it rely upon your issues or goals that you just need to succeed in it. If your info was terribly huge additionally you engineered knowledge warehouse in right means you may get the various output over time.

The actual data processing task is that the semi-automatic or automatic analysis of huge quantities of information to extract previously unknown, fascinating patterns like groups of information records (cluster analysis), uncommon records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This typically involves exploitation info techniques like spatial indices. These patterns will then be seen as a sort of outline of the input file, and should be employed in more analysis or, for instance, in machine learning and prognostic analytics. For instance, the information mining step may determine multiple teams within the data, which may then be accustomed get additional correct prediction results by a choice web. Neither the information assortment, information preparation, nor result interpretation and news are an element of the information mining step, however do belong to the general KDD method as extra steps.

The distinction between information analysis and data processing is that information analysis is employed to check models and hypotheses on the dataset, e.g., analysing the effectiveness of a promoting campaign, despite the number of data; in distinction, data processing uses machine learning and applied mathematics models to uncover underground or hidden patterns in a very giant volume of information.

1) *Classification:* In the word of machine learning, classification is taken into account an instance of supervised learning, i.e., learning wherever a coaching set of properly known observations is obtainable. The corresponding unsupervised procedure is understood as clustering, and involves grouping information into classes supported some live of inherent similarity or distance. Often, the individual observations are analysed into a group of quantitative properties, well-known diversely as explanatory variables or features. These properties might diversely be categorical, ordinal, integer-valued (e.g., the number of occurrences of a selected word in an email) or real-valued (e.g., a measure of blood pressure). Alternative classifiers work by scrutiny observations to previous observations by suggests that of a similarity or distance perform. An algorithmic program that implements classification, particularly in a very concrete implementation, is understood as a classifier. The term "classifier" generally additionally refers to the mathematical relation, enforced by a classification algorithmic program, that maps computer file to a class.
2) *Clustering*: Cluster analysis or clustering is the task of grouping objects in such a method that objects within the similar group (called a cluster) are a lot of similar to every aside from to those in alternative teams (clusters). It's a main task of beta data processing, and a standard technique for applied mathematics information analysis, utilized in several fields, together with pattern recognition, image analysis, info retrieval, bioinformatics, information compression, computer graphics and machine learning. Cluster analysis itself isn't one specific algorithmic program, however the final task to be solved. It may be achieved by numerous algorithms that dissent considerably in their understanding of what constitutes a cluster and the way to with efficiency notice them. In style notions of clusters embody teams with little distances between cluster members, dense areas of

the info house, intervals or explicit applied mathematics distributions. The clump will so be developed as a multi-objective improvement downside. The acceptable clump algorithmic program and parameter settings (including parameters like the space operate to use, a density threshold or the number of expected clusters) rely upon the individual information set and meant use of the results. Cluster analysis intrinsically isn't an automatic task, however associate degree repetitious method of data discovery or interactive multi-objective improvement that involves trial and failure. It's typically necessary to change information pre-processing and model parameters till the result achieves the specified properties.

B. Machine Learning

Machine learning (ML) is a technique of data analysis that is used for analytical model building. It is a branch of artificial intelligence which is based on the concept that systems can learn from information that is being available from the data, determine patterns and reach at decisions with less or no human intervention. Because of new computing technologies, machine learning these days isn't like machine learning of the past. It absolutely was born from pattern recognition and also the theory that computers will learn while not being programmed to perform specific tasks; researchers fascinated by computer science wished to check if computers may learn from information. The iterative aspect of machine learning is very important as a result of as models are exposed to new information, they're able to severally adapt. They learn from previous computations to supply reliable, repeatable choices and results. It's a science that's not new – however one that has gained contemporary momentum. Essentially, it is a method of teaching the computers to take a good decision and improve predictions or behaviours based on some data. The data is something which is entirely on the problem. It might be readings from a robot's sensors because it learns to steer, or the proper output of a program surely input. Otherwise, to consider machine learning is that it's "pattern recognition" - the act of teaching a program to react to or acknowledge patterns. In computer science, Artificial intelligence (AI), is sometimes called as machine intelligence, is intelligence shown by the machines, in contrast to the natural intelligence displayed by humans and other creatures. Computer science defines as any device that takes information from its environment and takes necessary actions that maximize its chance of successfully achieving its goals. But Machine Learning [5] can be a subset of AI. If some behaviour exists in past, then you will predict if or it will happen once more. It suggests that if there aren't any past cases then there's no prediction. The things like Image Recognition and Natural Language Processing (NLP) are great examples of ML things like Image Recognition and tongue process is nice samples of metric capacity unit.

III.RELATED WORKS

The driver's route unfamiliarity and therefore the interactions between familiar and unacquainted with drivers might have an effect on each the driving performances and therefore the probability of road crashes was explained in [1]. The familiarity was confirmed as an influential issue on the accident risk, probably because of distraction and dangerous behaviours, whereas the influence of being unknown on the accident disposition has some unclarified aspects. However, crashes to unknown drivers could cluster at sites showing high summer traffic variation and in summer months.

With the fast development of urbanization, the boom of car numbers has resulted in serious traffic accidents, that crystal rectifier to casualties and large economic losses. The flexibility to predict the chance of accident is vital within the interference of the prevalence of accidents and to cut back the damages caused by accidents in an exceedingly proactive manner. However, traffic accident risk prediction with high spatiotemporal resolution is difficult, in the main because of the advanced setting, human behaviour, and lack of period information [4]. The analysis between driver gender and age as associated with the injury crash frequency and road state of affairs was proposed in [5]. The main objective is to develop safety performance functions (SPFs) on 2-lane rural roads to predict the no. of injury crashes annually per 10(8) vehicles/km on the road section employing a study on the influence of the human factors (gender, age, range of drivers) and road situation (combination of infrastructure and environmental conditions found at the positioning at the time of the crash) on the consequences of a crash by varied the dynamic. Countermeasures are prompt to cut back the injury crash rate and embrace completely different awareness campaigns and structural measures on the segments of road. Authors in [6] explained about an identification model for road traffic accident black spot. With the fast development of the social economy and fast urbanization, the overall variety of motorized vehicles continues to grow at a high rate. Roads in massive and medium sized cities have become progressively full, that ends up in frequent traffic accidents. To reinforce road traffic safety and scale back the traffic accident rate, effectively characteristic accident black spots is of nice importance. The improved K-means bunch formula was projected to unravel the shortcomings of the normal formula, that is vulnerable to outliers and initial bunch centres. By using this formula, the traffic accidents within the dataset are divided into 2 two categories such as black spots and non-black spots. Then, mistreatment the updated dataset, they utilized a Bayesian network to construct a plant disease identification model, and applied different wide used for comparison.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

IV. PROPOSED SYSTEM

In the case of road accident analysis, we used a dataset that includes more than 42 lakhs tuples with more than 100 attributes. Information about the driver, vehicle, weather, roadway, and environment were obtained from the dataset. Categorizing the major aspect of traffic collisions and their severity assists the highway safety development in improving the road conditions according to the variation in vehicle occupancy need of specific sector of the population. Correlation analysis examines many of the attributes that help to derive the relation between two numerical variables. In this, we have done much correlation analysis in order to reach several conclusions over the road accidents. In this initially, the attributes such as sex, age, weather conditions, light conditions etc. were analysed and its effect over the road accident was analysed and found out their relationship according to certain conditions. This has been implemented using heatmap, boxplot etc. At the second level, a logistic regression model was used to explain the effect of weather and road surface condition towards the accident severity, through variables retrieved from the database. Logistic regressor used here can predict the accident severity for a particular weather and road surface condition. In the last section, we have identified the major accident-prone areas in UK and plotted it in the Base map and Google map using Geojson.

Many simple factors can lead to severe road accidents. So, it is highly important to identify which all are the ones that are leading to such a situation. Analysis over a large dataset can be made simple by means of visualization using heatmaps [7], boxplots [8], histogram [9] etc. Several factors such as weather condition, light conditions, age of driver, sex of driver etc. have some dependencies within themselves. Visualization can make it easy to understand about it.

Here logistic regression [10] is used to classify the accident severity with respect to the weather condition and road surface condition. It is clear that these two attributes are the major factors that affect the accident severity. The logistic regression is chosen here in such a way that we have removed the data related to the accident severity 3, which are slight accidents. So that we can have binary classification. Most severe cases and serious cases of accident are classified. Accuracy rate of this classification is also very high. That is almost 87% of classification is correct. This is a good accuracy rate when we compare logistic regression with the linear regression.

In the base map section, the accident-prone areas in UK according to the accident severity can be displayed. The accident-prone area is selected with only considering the serious and fatal accidents. The location where more than 50 accidents take place is chosen as the accident-prone area, thereby we can get a correct idea about the location which is dangerous for driving. The base map instance is often accustomed calculate positions on the map and also the inverse operation, changing positions on the map to geographical coordinates [11].

For the google map visualization of accident-prone area [12] in UK, Geojson is used. For that, a csv file is to be created from the observations acquired. Additional parameters such as marker size, marker symbol, marker colour etc are added. So, the attributes are Latitude, Longitude, Accident-Severity, Accident-Count, Lat-Long, Location, marker-colour, marker-size, marker-symbol, Title. Latitude and Longitude includes the exact location details. Location is the name of location according to the latitude and longitude correspondingly. The conversion is done by the library Geolocator. The conversion is quite time consuming. The two-accident severity is given with 2 colours, red for fatal accidents and orange for serious accidents. The colour is given in the csv file as hexadecimal values accordingly. The size of symbol is setup as small for accident count between 50 and 100, medium for accident count between 100 and 150, and large for accident count greater than 150. Title is given for viewing the name of location while hovering over the map. In Geojson, the process is done by means of scrolling down the csv file created accordingly to the page of Geojson or loading the csv file into it. Many functions such as zoom in, zoom out, search, viewing the location details while clicking on the symbol are provided. Differences in the size give the correct idea about the most dangerous location.

V. ANALYSIS & RESULTS

A. Correlation Analysis

Initially, we had done analysis by means of visualizing the relationship between several attributes. The first analysis is the correlation between age and sex of driver. The heatmap corresponding to this correlation is plotted in the Fig. 1. From this, the conclusion we can arrive at is that the correlation between age and sex of driver who met with the accident is too low. From Fig. 2 which is the correlation between hour, accident severity, light and weather condition, we can see that there is a higher correlation between light condition and hour while considering the correlation between light and weather condition in aspect to the accident severity.

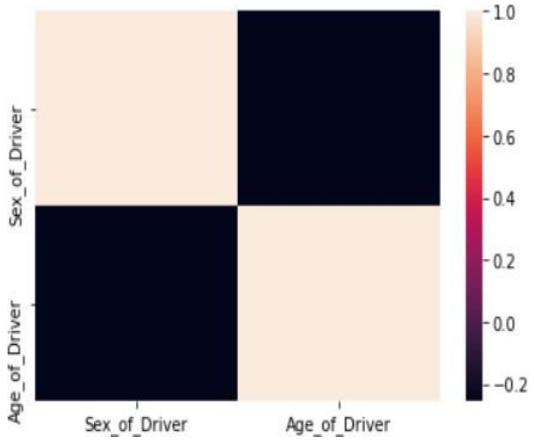


Fig. 1 Correlation between Age and Sex of Driver

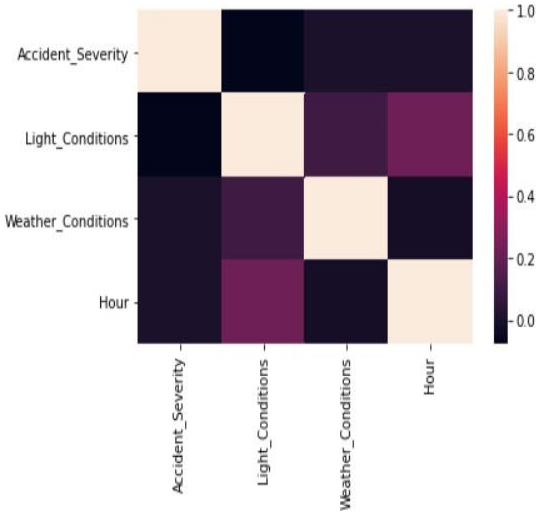


Fig. 2 Correlation between Hour, Weather, Light and Accident Severity

The correlation between hour and day of week with respect to accident count is shown in Fig. 3. From this we get the hour at which more accident takes place. That is more accidents takes place at the time approximately morning 8-9 and evening 4-5. Similarly, the correlation between month and date with respect to accident count is shown in Fig. 4 and it shows that the accident count is higher in the second half of the year.

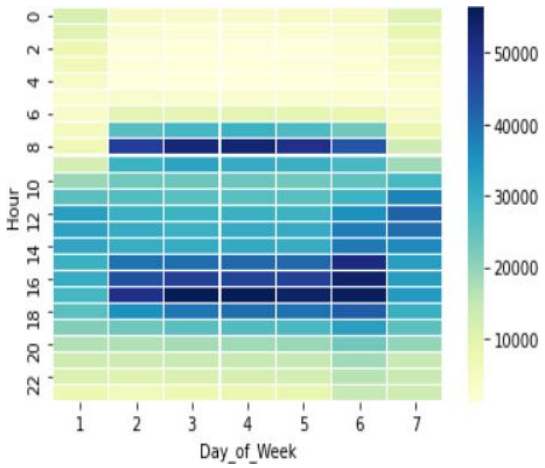


Fig. 3 Correlation between Hour and Day of Week

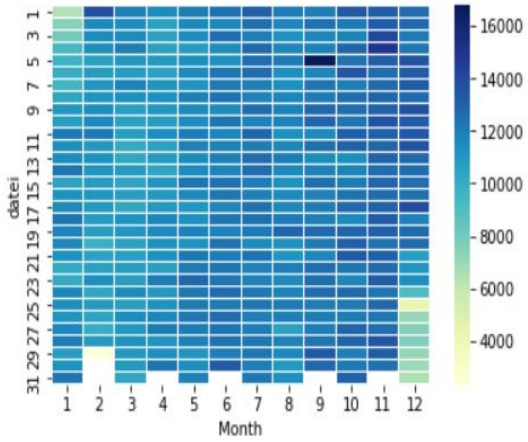


Fig. 4 Correlation between Date and Month

The boxplot in Fig. 5 shows the relation between age and home area type of driver. The driver home area type includes urban, small and rural. From this we can find that most of the drivers who met with the accident are of an age in between 25 and 55.

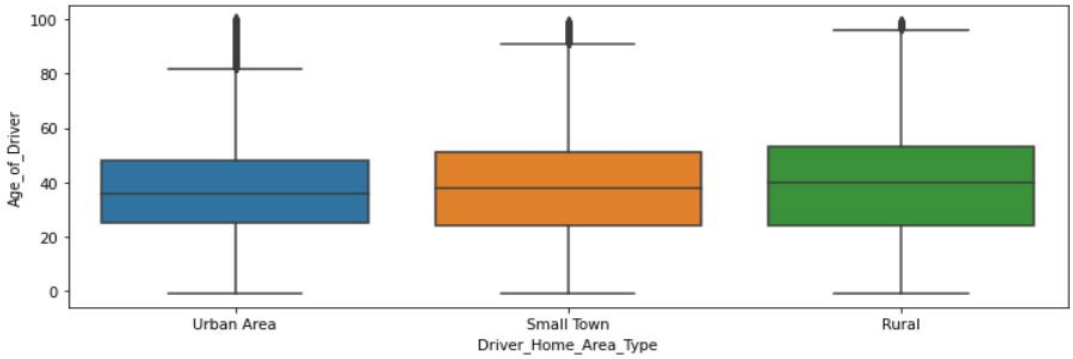


Fig. 5 Boxplot: Age vs Home Area Type of Driver

B. Visualization of Analysis

The histogram corresponding to the age and journey purpose of driver with respect to accident severity is shown in Figure 6. From this it is very clear that most of the fatal and serious accidents occurs in the journey of taking pupil to/from school. The heatmap shown in Fig. 3 and this histogram are providing us a common conclusion. From the Fig. 7 we can see that major road accidents with all severities happens approximately between the time 12 and 17. The next histogram plotted in Fig. 8 depicts the accident severity and count with respect to weather conditions. We can see that that majority of road accidents with all severities occurred during fine and no high wind weather condition.

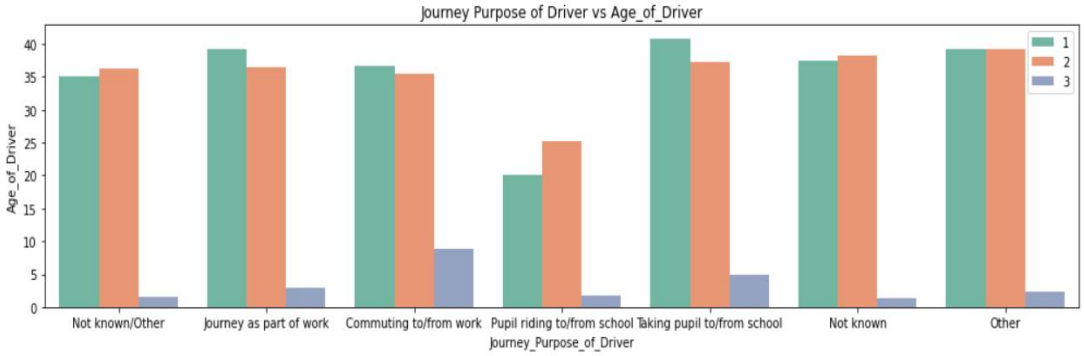


Fig. 6 Histogram: Age vs Journey Purpose of Driver w. r. t. Accident Severity

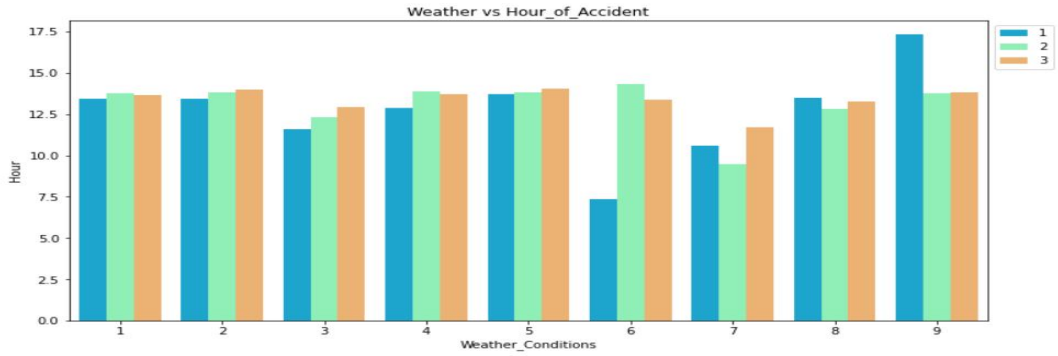


Fig. 7 Histogram: Hour vs Weather Conditions w. r. t. Accident Severity

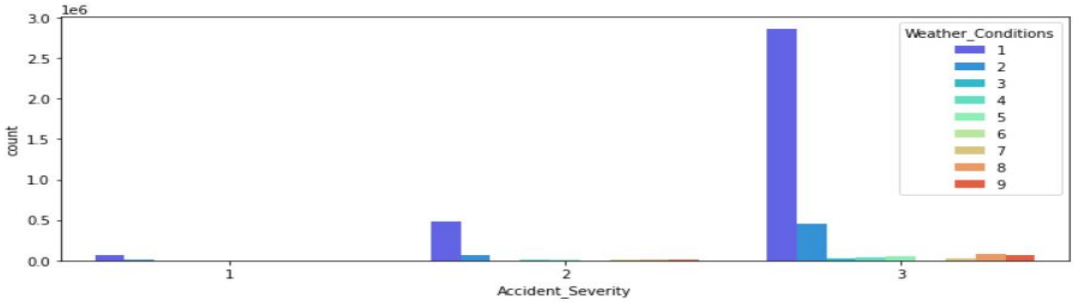


Fig. 8 Histogram: Count vs Accident Severity w. r. t. Weather Conditions

C. Performance Evaluation using Logistic Regression

Since the main agenda is to model a system using logistic regression specially to classify accident severity using weather and road surface conditions. The Fig. 9 shows the accuracy that we obtained from this model and it is about 87%.

```
from sklearn.metrics import confusion_matrix, accuracy_score
accuracy_score(y_test, y_pred)
```

```
0.8757197724960152
```

Fig. 9 Accuracy

D. Visualization of Accident Prone Areas using Base Map

The Fig. 10 shows the visualization of accident-prone areas in UK using Base map. The major accident-prone areas are at the southern side of the country.

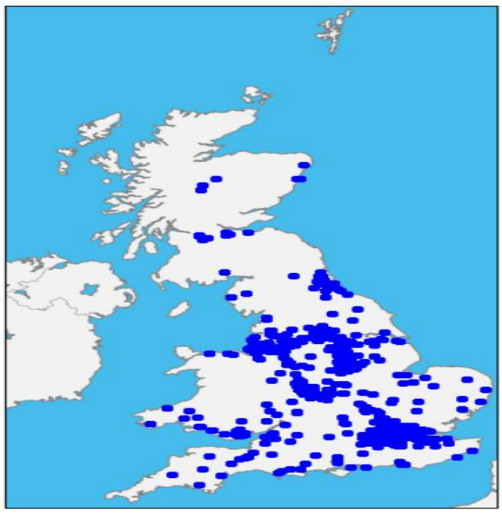


Fig. 10 Visualization of Accident Prone Areas using Base Map

E. Visualization of Accident Prone Areas using Google Map

The visualization of accident-prone areas in UK using Google map is shown in Fig. 11 and it is done using Geojson. The two accident severities can be seen with 2 colors, red for fatal accidents and orange for serious accidents with different sizes for different accident counts. We can also see the name of location while hovering over the map. If we click on any of the locations, we can see the details about that location as shown in Fig. 12.

Fig. 11 Visualization of Accident Prone Areas using Google Map

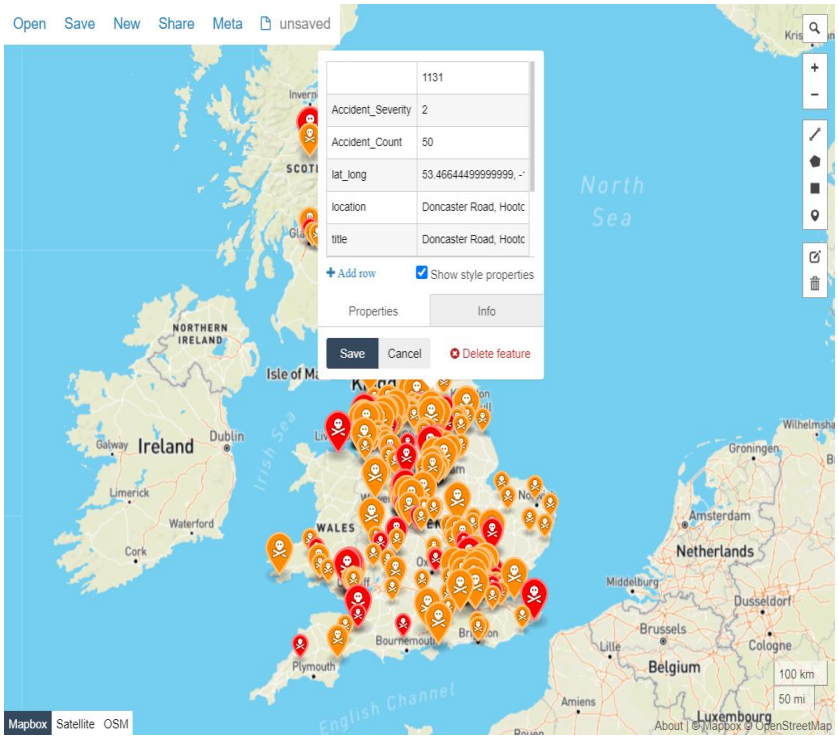


Fig. 12 Location Details of Accident Prone Areas

VI. CONCLUSIONS

In the present situation, road accident analysis has a great role in identifying the major factors and its features that causes road accidents. Data mining platform is very helpful in performing analysis of an improper dynamic data. This work mainly deals with the analysis of road accidents that happened in UK from the year 2005 to 2015. We used a dataset that includes about 43 lakhs tuples with more than 100 attributes. We have analyzed several features and found its correlation using heatmap and visualizes its dependencies using heatmap, boxplot and bar graph. We have used logistic regression to find out relationship between several attributes with respect to accident severity. Also, the major accident-prone areas are identified and plotted it using Base map. Apart from this, we have also extended the functionality of existing services of the Google maps to visualize the accident-prone areas using Geojson. We had reached at several conclusion over this road accident analysis and found out that weather conditions and road surface conditions are the major reasons behind the accident. We had also found out that the time and reason for the major road accidents. The accident prone-areas thus obtained can be used in the future analysis purpose as well as we can extend this by giving warning to the driver during real time navigation using Google map.

VII. ACKNOWLEDGMENT

We express our heartfelt gratitude and thanks to our project guide Prof. Linda Sara Mathew and project coordinator Dr. Jisha P Abraham for their corrections, suggestions and sincere efforts in helping us to work on this project. We owe special thanks to our principal Dr. Mathew K and Head of the Department Prof. Joby George for providing the necessary facilities and their encouragement and support. We also express our sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping us to work on this project.

REFERENCES

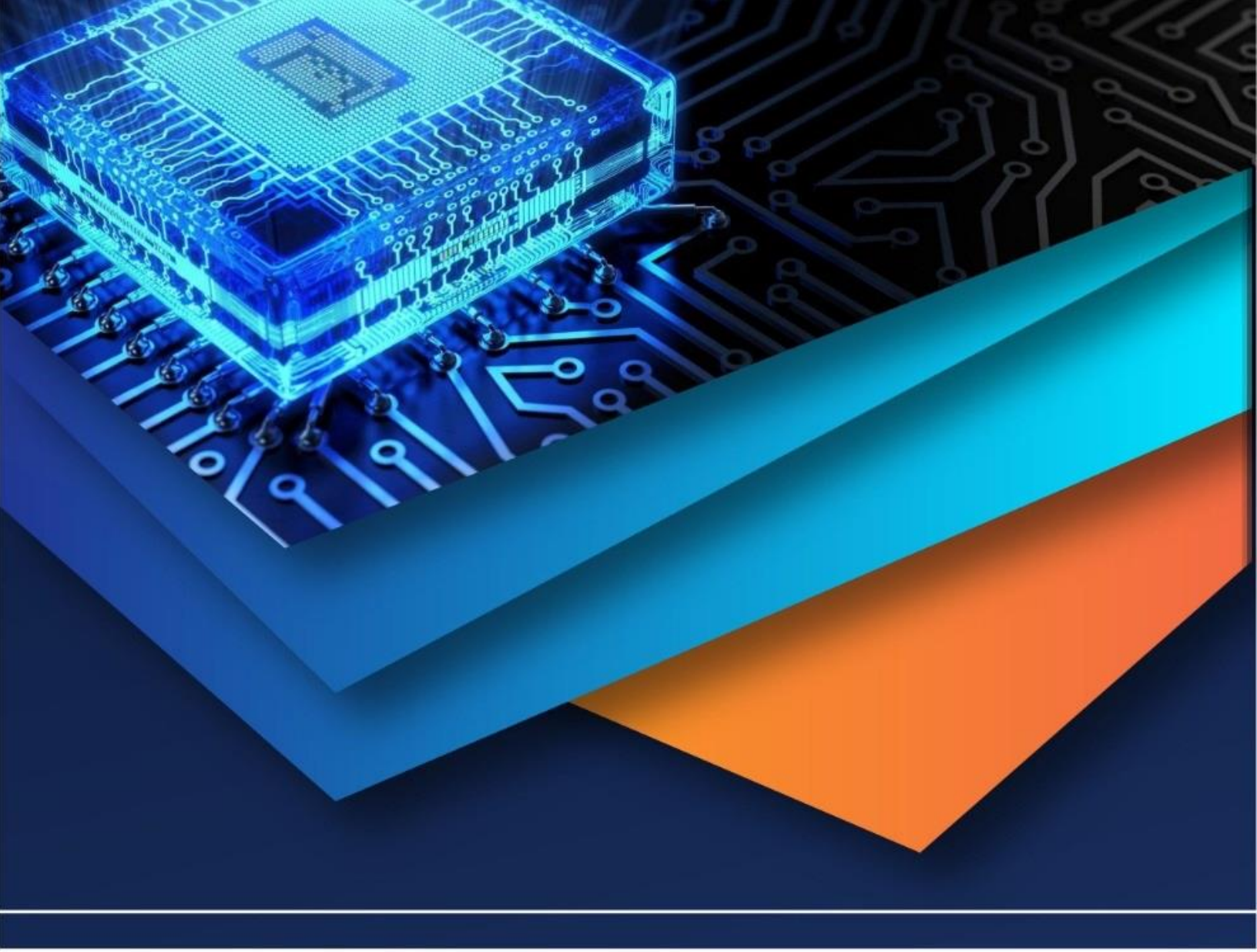
[1] Paolo Intini, Nicola Berloco and Pasquale Colonna, Exploring the Relationships Between Drivers' Familiarity and Two-Lane Rural Road Accidents-A Multi-Level Study, in Accident Analysis & Prevention, Vol. 111, Pages 280-296, Feb. 2018.
[2] Tibebe Beshah, Shawndra Hill, Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia, in World Congress on Information and Communication Technologies, Pages 1127-1134, 2012.
[3] N. Sklyarov, Analysis of Problems of Road Traffic Accidents Motor Vehicle Examination Improvement, in Avtomobilnyi Transp., Vol. 29, No. 3, Pages 339-357, Mar. 2011.
[4] Honglei Ren, You Song, JingXin Liu, and Yucheng Hu, A Deep Learning Approach to the Prediction of Short-Term Traffic Accident Risk, in Research Gate, Oct. 2017.
[5] Francesca Russo, Salvatore Antonio Biancardo and Gianluca Dell'Acqua, Road Safety from the Perspective of Driver Gender and Age as Related to the Injury Crash Frequency and Road Scenario, 2015.
[6] Cheng Zhang, Yue Shu and Lixin Yan, A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China, in IEEE Access, Vol. 7, Pages 140197 - 140205, Sept. 2019.
[7] Sandor Szenasi, Imre Felde, Road Accident Black Spot Localisation Using Morphological Image Processing Methods on Heatmap, in IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI), Dec. 2019.
[8] Georgy Shevlyakov, Lakshminarayan Choudur and Pavel Smirnov, Robust Versions of the Tukey Boxplot with their Application to Detection of Outliers, in IEEE International Conference on Acoustics, Speech and Signal Processing, Oct. 2013.
[9] Zhao Geng, ZhenMin Peng and Robert S.Laramee, Angular Histograms: Frequency-Based Visualizations for Large, High Dimensional Data, in IEEE Transactions on Visualization and Computer Graphics, Vol. 17, Issue 12, Dec. 2011.
[10] Tao Lu, Zhu Dunyao and Yan Lixin, The Traffic Accident Hotspot Prediction: Based on the Logistic Regression Method, in International Conference on Transportation Information and Safety (ICTIS), Sept. 2015.
[11] Anik Vega Vitianingsih, Dwi Cahyono, Geographical Information System for Mapping Accident-Prone Roads and Development of New Road Using Multi-Attribute Utility Method, in 2nd International Conference on Science and Technology-Computer (ICST), Mar. 2017.
[12] Gagandeep Kaur, Harpreet Kaur, Prediction of the Cause of Accident- and Accident-Prone Location on Roads Using Data Mining Techniques, in 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Dec. 2017.

AUTHORS PROFILE

Megna T Sreedhar received her B. Tech Degree in Computer Science and Engineering from MES College of Engineering, Kuttippuram, Kerala affiliated to Calicut University in 2017 and currently pursuing M. Tech in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam. Kerala affiliated to APJ Abdul Kalam Technological University. Her research interest is in Machine Learning and Data Science.

Karthika K P received her B. Tech Degree in Computer Science and Engineering from Government Engineering College, Palakkad, Kerala affiliated to Calicut University in 2018 and currently pursuing M. Tech in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam, Kerala affiliated to APJ Abdul Kalam Technological University. Her research interest is in Machine Learning and Data Science.

Prof. Linda Sara Mathew is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala. She received her B.Tech Degree in Computer Science and Engineering from Mahatma Gandhi University in 2002 and ME in Computer Science and Engineering from Anna university in 2011. She is interested in the areas Data Mining, Soft Computing, Neural Network and Image Processing.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY