IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF TRAFFIC-VIOLATION USING MACHINE LEARNING

R.Sneha , Mrs.P.Jasmine Lois Ebenezer

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

Abstract: This project presents the prediction of traffic-violations using machine learning, more specifically, when most likely a traffic-violation may happen. Also, what are the contributing factors that may cause more damages (e.g., personal injury, property damage, etc.) are discussed in this work. The national database for traffic-violation wasconsideredfortheminingandanalyzedresultsindicatedthatafewspecific times are probable for traffic-violations. Moreover, most accidents happened on specific days and times. The findings of this work could help prevent some traffic-violations or reduce the chance of occurrence. These results can be used to increase cautions and traffic-safety tips. This work presents an in-depth analysis of road and traffic violations pattern using Data Analytics methods, aimed at improving road and traffic management, government planning and decision making. The study identified the road and traffic current management practice as basis of the design development and implementation of the road and traffic management system. The application managed all the road and traffic violation that will produce recorded set for analysis, which carried out from over of five years. Through data cleansing a total of twenty thousand six hundred forty record set was derived. It is important to find use of this record set, build analysis models, and use interactive tools to produce predictive data, understand the relevance, trends, and driving behaviors from the road and traffic violations data in terms of the following predictors: gender of the violator, vehicle owner address, location of violation, month and time the violation was committed and traffic enforcer who issued the citation. The study was able to establish a data analysis model by using a powerful classification and random forest which was executed using an open-source application named PyCharm. Finally, the developed application was evaluated by Python.

Index Terms **- Component,formatting,style,styling,insert.**

I. INTRODUCTION

When a law enforcement official issues a traffic summons (also known as traffic tickets), it is to inform the motorist, which includes anyone who drives a car, truck, or bus as well as anyone who rides a motorcycle, that they have been stopped . At some point or another, the majority of drivers are recommended for a moving infringement because they are speeding, running a red light, or committing some other type of criminal traffic infraction. The results of tickets are not calamitous but at the very least, dealing with a ticket requires an investment of time. The fact that many people do not consider street activity offenses to be wrongdoing, for some strange reason, gives the impression that they are not considered wrongdoing; however, nothing could be further from the truth. Because of the high number of fatalities that can result from traffic offenses around the world, they have become a major source of public concern. For some inexplicable reason, a significant proportion of people do not regard street activity crimes to be criminal offenses.

Data mining allows the processing of large amounts of historical data and the condensing of that information into valuable information that can be used to construct various models, such as prediction models, clustering models, and anomaly models. Data mining software such as the RapidMiner allows users to analyze data using various data mining approaches until knowledge is extracted as the information is accessible in the diverse organizations to make the best possible move.

According to the literature, several studies have been conducted on the demographic and socioeconomic features of criminal offenders from varied backgrounds. However, minimal studies characterize traffic offenders and drivers who receive citation tickets or warnings. Understanding traffic offences is critical because it will allow for the development of more effective prevention and enforcement strategies to reduce these offenses and, ultimately, road accidents on the road. Moreover, data mining can be applied in many industries to help improve or forecast many things. For traffic violations, prior research has primarily looked at how well drivers can predict the characteristics of other drivers who are ticketed for traffic violations.

This study aims to build a comparative model for classifying traffic violation types based on a data mining approach. Traffic violation types are categorized into Citation, Warning, and ESERO (Electronic Safety Equipment Repair Order) that referred to . The classification algorithms to be used include the Naïve Bayes, Gradient Boosted Trees, and Deep Learning algorithms. This research is scoped to traffic violation data from Montgomery County between the years 2013 to 2016, and the dataset was extracted from a website called data.world. The classification models developed in this paper will be measured for accuracy, recall, precision, and *f*-measure.

For the existing dataset from , the data used in this study came from two different sources: the Southwest City Police Department (SWCPD) and the United States Census Bureau in the year 2000. All traffic citation data was obtained from the

SWCPD and consisted of all traffic offenses committed between the dates of January 1, 1999, and October 10, 1999, totaling 87,792 traffic violations and 211,689 fines within that time period. In addition to driver demographic parameters (day, date, and time of the violation), the data on these violation occurrences includes information on the types of charges levied against the driver, his or her speed, and the legal speed limit in the area. The dataset consists of the accident day, year, variables, the vehicle involved, and people included. There are 39 qualities chosen from both datasets, and after data cleansing, 573 of accident information causes driver's casualty.

II. LITERATURE SURVEY

In [1], This paper presents the prediction of traffic-violations using data mining techniques, more specifically, when most likely a traffic-violation may happen. Also, the contributing factors that may cause more damages (e.g., personal injury, property damage, etc.) are discussed in this paper. The national database for traffic-violation was considered for the mining and analyzed results indicated that a few specific times are probable for traffic-violations. Moreover, most accidents happened on specific days and times. The findings of this work could help prevent some traffic-violations or reduce the chance of occurrence. These results can be used to increase cautions and traffic-safety tips.

In [2], Traffic summons, also known as traffic tickets, is a notice issued by a law enforcement official to a motorist, who is a person who drives a car, lorry, or bus, and a person who rides a motorcycle. This study is set to perform a comparative experiment to compare the performance of three classification algorithms (Naive Bayes, Gradient Boosted Trees, and Deep Learning algorithm) in classifying the traffic violation types. The performance of all the three classification models developed in this work is measured and compared. The results show that the Gradient Boosted Trees and Deep Learning algorithm have the best value in accuracy and recall but low precision. Naïve Bayes, on the other hand, has high recall since it is a picky classifier that only performs well in a dataset that is high in precision. This paper's results could serve as baseline results for investigations related to the classification of traffic violation types. It is also helpful for authorities to strategize and plan ways to reduce traffic violations among road users by studying the most common traffic violation types in an area, whether a citation, a warning, or an ESERO (Electronic Safety Equipment Repair Order).

In [3], In general, the criminalization has been significantly increased for past few years. Therefore, the use of technology is an essential for making the business work easier by providing several activities. Analysis of criminal investigation using data mining has created an important factor for understanding and predicting the activities of the criminals. The crime get classifies into various types but in this paper, the discussion is about traffic violation crime. The traffic violation crime occur while the driver breaks law which happen in vehicle driving on all kinds of roadways. The increase of light vehicles number in the cities have build high traffic volume and also signifies the crime of traffic violation has become more common that create severe damage to the property and high accident rate which have cause danger to the people's life. In order to resolve these issues, the technique of data mining have utilized in several machine learning algorithm for extracting awareness from large data volume and even discovered the trends and pattern for traffic violation crime. The K-means clustering technique is used for tracing the region of accident, where the activities of crime have occurred. In other hand, KNN classifier is utilized for identifying the criminal behavior by assisting of observation in past crimes and identifying same crime activity but in case of no previous information get discovered then this crime is consider as a new sample crime which get added to the crime dataset. Hence, this study discuss about the detection and investigation of crimes and the behavior of the criminals using K-means clustering and KNN classifier in the traffic violation crimes.

In [4], Presented in this paper is a comparative analysis of various Data Mining clustering methods for the grouping of roads, aimed at the estimation of Annual Average Daily Traffic (AADT). The analysis was carried out using data available from fifty-four Automatic Traffic Recorder (ATR) sites in the Province of Venice (Italy) and separated adjustment factors for passenger and truck vehicles in the grouping process. Errors in AADT estimation from 24-h sample counts indicate that model-based clustering methods give slightly better results compared to other tested methods, identifying a significant ATRs classification.

In [5], Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In this paper we apply statistics analysis and data mining algorithms on the FARS Fatal Accident dataset as an attempt to address this problem. The relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver were investigated. Association rules were discovered by Apriori algorithm, classification model was built by Naive Bayes classifier, and clusters were formed by simple K-means clustering algorithm. Certain safety driving suggestions were made based on statistics, association rules, classification model, and clusters obtained.

In [6], Traffic analysis has many purposes such as evaluating the performance and security of network operations and management. Therefore, network traffic analysis is considered vital for improving networks operation and security. This paper discusses different machine learning approaches for traffic analysis. Increased network traffic and the development of artificial intelligence require new ways to detect intrusions, analyze malware behavior, and categorize Internet traffic and other security aspects. Machine learning (ML) shows effective capabilities in solving network problems. A review of the techniques used in the traffic analysis is presented in this paper.

In [7], This paper presents a sample of mining algorithms in data represented in "One R, J48, Naïve Bayesian" to know its optimal which pertains to analysis of traffic accidents in Khartoum State occurring through years 2007 – 2016.It is important to note that 389931 record was analysed by the statistical reports structure to reach the mining stage in data in order to creating a mechanism that is capable of studying the elements which smartly play a significant part in traffic accidents for connection. The range of relation designation between them, and its significance in traffic accidents percentage is implemented on Weka program to apply algorithms in data and ,accordingly, the presentation of the results together with analysis since the results showed that the performance of J48 algorithm generally is of more qualifications and surpasses than the other algorithms in accidents data group. It spent 0.02 seconds and the rate of error in the sample was 0.02 through its implementation and assisted in the prediction of data. The paper concludes with the implementation of J48 classification algorithms for the production of the decision tree through Weka to the point of the existence of classification of cars' accidents which occurred according to time and harm. It also shows that the harm damage rate is of the highest reaching 301394 at the rate of 77.3%, and that 2012 and 2011 accidents were the highest of all years at the rate of 11.3%, and the rate of the lowest accidents in 2007 was 7.4%.

In [8], Road traffic accidents are very essential for common people, consequential an estimated 1.2 million deaths and 50 million injuries all over the world every year. In this emerging world, the road accidents are among the principal reason of fatality and injury. The concern of traffic safety has heaved immense alarms across the manageable enhancement of contemporary traffic and transportation. The analysis on road traffic accident grounds can detect the major aspects quickly, professionally and afford instructional techniques to the prevention of traffic accidents and reduction of road traffic accident, which might significantly decrease personal victim by means of road traffic accidents. The current research represents that the Data Mining techniques that have employed in the field of transport have been investigated. Through this comprehensive investigation, the techniques of Data Mining in the traffic study can enhance the administration level of road traffic safety productively.

In [9], Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. Data Mining is taking out of hidden patterns from huge database. It is commonly used in a marketing, surveillance, fraud detection and scientific discovery. In data mining, machine learning is mainly focused as research which is automatically learnt to recognize complex patterns and make intelligent decisions based on data. Globalization has affected many countries. There has been a drastic increase in the economic activities and consumption level, leading to expansion of travel and transportation. The increase in the vehicles, traffic lead to road accidents. Considering the importance of the road safety, government is trying to identify the causes of road accidents to reduce the accidents level. The exponential increase in the accidents data is making it difficult to analyses the constraints causing the road accidents. The paper describes how to mine frequent patterns causing road accidents from collected data set. We find associations among road accidents and predict the type of accidents for existing as well as for new roads. We make use of association and classification rules to discover the patterns between road accidents and as well as predict road accidents for new roads.

In [10], This paper is discussing about the road accident severity survey using data mining, where different approaches have been considered. We have collected research work carried out by different researchers based on road accidents. Article describing the review work in context of road accident case's using data mining approach. The article is consisting of collections of methods in different scenario with the aim to resolve the road accident. Every method is somewhere seeming to productive in some ways to decrease the no of causality. It will give a better edge to different country where the no of accidents is leading to fatality of life

III. METHODOLOGY

The Knowledge Discovery in Database (KDD) framework is used in this study to classify the different types of traffic violations. The KDD framework is a data mining system that seeks to uncover interesting patterns in the underlying data. KDD is beneficial for a large dataset, and it can process data from the database as indicated by client necessities. KDD also incorporates how information is prepared, what calculations can be applied to obtain a substantial measure of information proficiently, and how the results can be translated and visualized . KDD begins with data warehousing, in which a related field is coming from the database. Data warehousing helps set the phase in KDD in two important ways: data cleaning and data access .

There are five phases in KDD that need to be implemented to get the results from classification or prediction techniques. The five phases include selection, pre-processing, transformation, data mining, and evaluation . In this research, the data mining part will involve the classification process to predict traffic violation offenders based on the summons issued. The phases in these experiments are shown in Fig. 1 as adopted from the KDD methodology.

Fig.1.ExperimentalphasesadoptedfromtheKDDmethodology

Data Selection

According to the process of selecting a certain characteristic from an initial dataset that is most relevant to the data mining activities at hand is known as data selection. Because of the removal of irrelevant or repetitive features, the execution time for the data mining operation will be reduced, while the precision will be raised as a result of the process. When it comes to boosting the effectiveness of data mining algorithms, feature selection is crucial since it ensures that only meaningful and beneficial attributes are used. The final collection of features chosen meets key critical requirements for shrinking in terms of overall size .

Data gathering is one of the technical stages required and should be taken totally with the goal that it very well may be run and tested later to think about the execution of the order in expectation of criminal traffic offense . The primary reason for this stage is to get an appropriate dataset with proper credit and run the test. In this study, the dataset was extracted from data.world. The dataset was contained 35 attributes and 7,700 rows. After completing data preparation using TurboPrep, the dataset consists of 8 attributes and 7,700 rows. The eight attributes are date, description, vehicle type, year, make, model, violation type, and gender. Fig. 2 shows the dataset after data pre-processing.

	Date Of Stop	Description	VehicleType	Year	Make	Model	Violation Type	Gender
2	9/30/2014	DRIVER FAILUF	Automobile	2014	FORD	MUSTANG	Citation	M
3	3/31/2015	HEADLIGHTS (*	Automobile	2003	HONDA	2S	Warning	M
4	9/30/2014	FAILURE TO DI:	Automobile	2009	TOYOTA	CAMRY	Warning	F
5	3/31/2015	DRIVER FAILUF	Automobile	2007	ACURA	MDX	Warning	F
6	3/31/2015	STOP LIGHTS (*	Automobile	2003	NISSAN	MURANO	ESERO	M
7	3/31/2015	DRIVING MOT	Automobile	2007	HONDA	CIVIC	Citation	F
8	3/31/2015	DRIVING VEHI(	Automobile	2007	HONDA	CIVIC	Citation	F
9	3/31/2015	FAILURE OF IN	Automobile	2007	HONDA	CIVIC	Citation	F
10	3/31/2015	FAILURE TO DI:	Automobile	2007	HONDA	CIVIC	Citation	F
11	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
12	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
13	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
14	9/30/2014	FAILURE TO DI:	Automobile	2002	TOYOTA	CAMRY	Warning	F
15	9/30/2014	STOPLIGHT INC	Automobile	2002	TOYOTA	CAMRY	Warning	F

Data Pre-Processing

The improvement of data mining cannot be isolated from the fast advancement of data innovation that permits a comprehensive measure of information aggregated in line with the development of data innovation . Mining implies an endeavor to profit from a vast number of fundamental materials. Given the best practice, experts, talented individuals, and individuals who work to discover data in information mining propose some procedure with work process or approach well-ordered easy to expand odds of accomplishment in putting into utilization the examination.

Right off the bat, the dataset got from the site has any sections that are inadequate as missing information, invalid information, or even pointless information. Likewise, additional credits do not apply to the examination in data mining. The information is not significant it is additionally better evacuated because it is nearness can decrease the quality or precision of the data mining later. Data cleaning is essential in every research to detect and remove errors from the raw data [20]. TurboPrep processed the dataset in RapidMiner Tools in the pre-processing data phase. The dataset selected by the operator is read in the RapidMiner tool. Turbo Prep is designed to make data preparation less time-consuming anddifficult . It gives a user interface where a data is continuously visible front and center, so the data can make changes step-by-step and immediately see the results, with an exhaustive run of supporting capacities to get ready so the data for model-building or presentation.

During this process, firstly, data need to choose whether they want to do a prediction or clustering. After that, RapidMiner will display all the details in every attribute in the dataset. In TurboPrep, data can be transformed; for example, rename the attribute, change type, remove the column, and delete all the selected columns from the dataset . Moreover, one more thing is that TurboPrep can replace a missing value in the dataset. The best thing if using TurboPrep is that this tool provides quality measures. It means the user can see at a glance typical data quality problems. They can show the details about the quality measures are calculated in the dataset. The details will show missing value, infinite, IDs, stability, and valid. Users can check the details and then make a data transformation so that all the attributes with a high value of missing value and low stability will be removed from the dataset.

Classification Algorithms

Graduated boosted trees, Naïve Bayes, and Deep Learning were used in this classification experiment. According to , Based on the Bayes theorem, Naïve Bayes is a probabilistic classifier in which all variables or factors are presumed to be independently variable or factor from one another. The algorithm is straightforward to design and performs admirably when dealing with enormous datasets. According to the Bayes Theorem, the probability of $P(A|X) = P(X|A) \times P(A) / P(X) \times P(X)$, where $P(A)$ is the relative frequency of class A samples, and p is increased when $P(X|A)P(A)$ is increased, and p is increased when $P(X|A)P(A)$ is increased .

The second classification algorithm used in this experiment is the Gradient Boosted Trees . It is possible to train a boosted decision tree using an ensemble learning method, in which one independent tree corrects the errors of another independent tree. If the first tree makes a mistake, it is corrected by a second one, and so on. The second tree makes a mistake by the first and second trees, and so on. According to , boosting is one of the most effective learning concepts to be established in the last twenty years since it may combine a large number of poor learners into a single strong learner with little effort. A gradient boosted decision tree is a classification model that aggregates all tree-based classification models and uses estimations to gradually achieve its prediction outcomes. Boosting may be a nonlinear regression strategy that is adaptive and makes a difference in the precision of trees as they grow in complexity. Improved trees outperform normal trees in terms of accuracy but are slower and less interpretable by humans than standard trees. The Gradient boosting approach is designed to address these concerns.

The latest algorithm used in this research work is Deep Learning, which mimics human intelligence , and many recognition problems with huge training samples in numerous representations and high-speed streams benefit from the use of this technique. Deep learning, which is based on base learning technology (particularly, neural networks), can provide a cross-therapy information analysis to allow for better informed treatment decisions.

IV. ACKNOWLEDGMENT

RANDOM FOREST ALGORITHM (RFA)(Machine Language):

The random forest algorithm is a supervised organization algorithm. As the name proposes, the algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like, in the same way in the random forest classifier, the higher the amount of trees in the forest gives the high accuracy results. In RFA, instead of using the information gain or Gini index for calculating the root node, the process of finding the root node, and splitting the feature nodes will happen randomly.

RFA Advantages: The same random forest algorithm or the random forest classifier can use for both classification and the regression job. The random forest classifier will handle the missing values. When we have more trees in the forest, a random forest classifier won't overfit the model. The random forest algorithm can be used for characteristic engineering, which means the most important features out of the available features from the training dataset. Pseudocode for RFA: The Pseudocode for RFA can split into two stages:

(i) Random Forest creation Pseudocode. (ii) Pseudocode to perform prediction from the created random forest classifier.

Pseudocode for RFA:

1. Randomly select "k" features from total "m" features Where k<<m
2. Among the "k" features, calculate the node "d" Using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until "I" number of nodes has been Reached.
5. Build forest by Repeating steps 1 to 4 for "n" Number times to create an "n" number of trees.

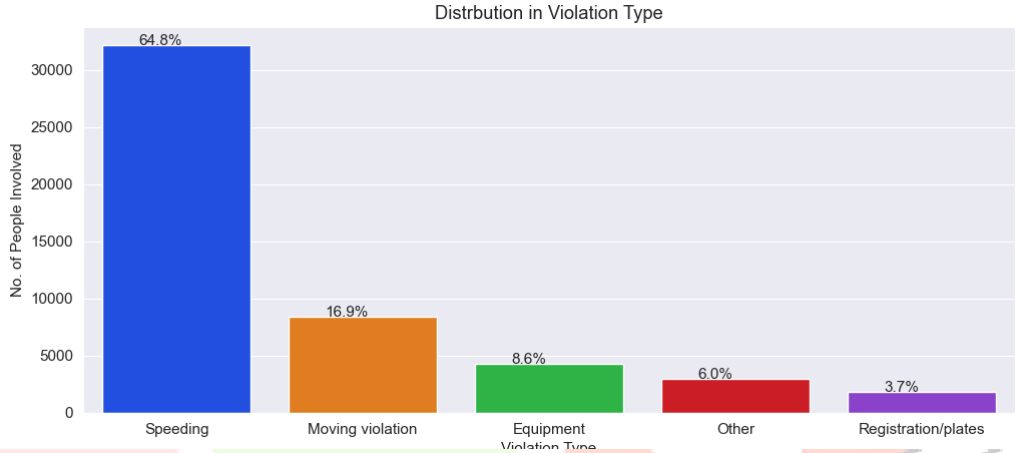

Figure 1 DISTRIBUTION OF VILATION

The above figure illustrates the distribution in violation type. The xaxis depicts the violation type and y-axis depicts number of people involved.

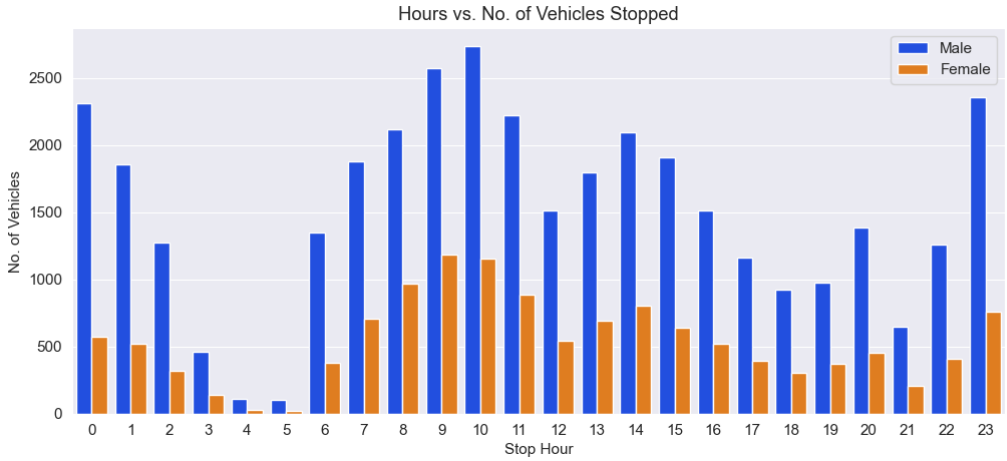

Figure 2hOURS VS nUMBER OF VEHICLES STOPPED

The above figure illustrates the hours versus number of vehicle stopped. The x-axis depicts the stop hour and y-axis depicts number of vehicles involved.

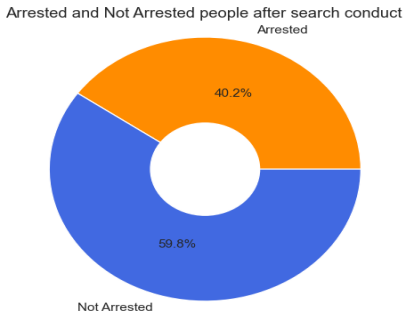

Figure 3 ARRESTED AND NOT ARRESTED PEOPLE AFTER SEARCH CONDUCT

The above figure illustrates the arrested and not arrested people after search conduct. The orange color depicts the arrested people and y-axis depicts number of not arrested people involved.

V. CONCLUSION

In this work, we have collected multiple researchers' works together in single document and discussed about the contribution towards impact of road and traffic accident on human life and society. This highlights the number of approaches used to avoid the accident happened in various countries and cities. The work also discussing about various data mining techniques which is proved supporting to resolve traffic accident severity problem and conclude which one could be optimal technique in road traffic accident scenario. The brief discussion will also help us to find better mining technique in this kind of problem. The paper has discussed the using of machine learning techniques in traffic analysis. It has given a brief overview and comparison among some existing ML approaches used in traffic analysis. Despite the big role of the machine learning, the work has shown that it still has some limitations. As future work, we plan to conduct a comprehensive study of the recent machine learning techniques, and provide a wide comparison among the most common approaches.

REFERENCES

1. Amiruzzaman, Md. "Prediction of Traffic-Violation Using Data Mining Techniques." *Proceedings of the Future Technologies Conference (FTC) 2018*, edited by Kohei Arai et al., vol. 880, Springer International Publishing, 2019, pp. 283–97. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-02686-8_23.
2. Amiruzzaman, Md. "Prediction of Traffic-Violation Using Data Mining Techniques." *Proceedings of the Future Technologies Conference (FTC) 2018*, edited by Kohei Arai et al., vol. 880, Springer International Publishing, 2019, pp. 283–97. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-02686-8_23.
3. S, Umadevi, and Nirmala Sugirtha Rajini S. "Detection of Traffic Violation Crime Using Data Mining Algorithms." *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, no. 0009-SPECIAL ISSUE, Sept. 2019, pp. 982–87. *DOI.org (Crossref)*, https://doi.org/10.5373/JARDCS/V11/20192660.
4. Gecchele, Gregorio, et al. "Data Mining Methods for Traffic Monitoring Data Analysis: A Case Study." *Procedia - Social and Behavioral Sciences*, vol. 20, 2011, pp. 455–64. *DOI.org (Crossref)*, https://doi.org/10.1016/j.sbspro.2011.08.052.
5. Li, Liling, et al. "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques." *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, IEEE, 2017, pp. 363–70. *DOI.org (Crossref)*, https://doi.org/10.1109/SERA.2017.7965753.
6. Alqudah, Nour, and Qussai Yaseen. "Machine Learning for Traffic Analysis: A Review." *Procedia Computer Science*, vol. 170, 2020, pp. 911–16. *DOI.org (Crossref)*, https://doi.org/10.1016/j.procs.2020.03.111.
7. Prince Sattam Bin Abdulaziz University, and Mozamel M. Saeed. "The Use of Data Mining Techniques in Analysing Traffic Accidents (An Application on Khartoum State)." *International Journal of Computer Trends and Technology*, vol. 62, no. 1, Aug. 2018, pp. 75–79. *DOI.org (Crossref)*, https://doi.org/10.14445/22312803/IJCTT-V62P110.
8. Chi, Seokho, et al. "Sustainable Road Management in Texas: Network-Level Flexible Pavement Structural Condition Analysis Using Data-Mining Techniques." *Journal of Computing in Civil Engineering*, vol. 28, no. 1, Jan. 2014, pp. 156–65. *DOI.org (Crossref)*, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000252.
9. Agarwal, Basant, and Namita Mittal. "Hybrid Approach for Detection of Anomaly Network Traffic Using Data Mining Techniques." *Procedia Technology*, vol. 6, 2012, pp. 996–1003. *DOI.org (Crossref)*, https://doi.org/10.1016/j.protcy.2012.10.121.