iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61540

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: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Road Accident Analysis and Classification Using Machine Learning Algorithms

Meet Gorasia[1], Shivraj Ghorpade[2], Dhruv Sakariya[3], Parth Sadigale[4], Prapti Deshmukh[5], Rashmi Rane[6]
[1, 2, 3, 4, 5]Student, [6]Professor, Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Abstract: ***India is facing a severe public health crisis characterized by a significant number of road accidents resulting in fatalities and life-altering injuries. This study undertakes a detailed analysis of road accidents in India using multi-year accident data to unveil recurring patterns and identify high-risk demographics. The research meticulously examines the primary causes of these accidents, including driver behavior such as speeding, and inadequate road infrastructure such as poor lighting, signage, potholes, and vehicle malfunctions.***
Moreover, the paper proposes a comprehensive approach to prevent these tragedies. It suggests strategies for enhancing road design and maintenance, enforcing traffic regulations more rigorously, and conducting targeted public awareness campaigns to promote responsible driving practices. By highlighting the alarming increase in fatalities and the substantial socio-economic costs associated with these accidents, this research emphasizes the critical need for immediate and effective interventions to establish a safer road environment for all citizens in India.
Keywords: Road Accident Analysis, Machine Learning, Classification, Prevention

I. INTRODUCTION

India has the world's second-largest network of roads, totaling about 66.71 lakh km out of which National Highways and State Highways comprise just 2% and 3% respectively while the rest of 95% are under the category of "Other Roads" according to the Ministry of Road Transport and Highways, Government of India [1]. This vast network serves as the lifeblood of India's economic growth, facilitating the transport of goods and travel of people to every part of the nation. However, India still struggles to achieve a desirable safety standard on its roads.

While the government pushes ambitious infrastructure development plans, a significant portion of the existing network suffers from shortcomings. A recent road quality index report says India's road quality index is 4.50 out of 7.00 being the highest [2], highlighting the need for significant improvement. Many roads are dangerous to drive due to a lack of proper lighting and signboards upon which governing bodies should work and maintain their safety guidelines. Although India has the second largest network of roads, they are still lower quality less maintained, and narrower due to poor project management and lack of funds.

The consequence of these shortcomings is tragically evident in the staggering number of road accidents. According to the latest report by the World Health Organization (WHO) released in 2023, India accounts for a shocking 10% of global road fatalities [3]. This translates to a devastating 1,68,491 lives lost and 4,43,366 injured out of a total of 4,61,312 accidents in 2022 alone [4], which approximates 462 lives lost in a single day alone and 20 lives lost every single hour. Recent reports show that the age group of 18-45 accounted for about 66.5% of total road accidents [4].

The situation becomes even more concerning when we dive deeper into city-specific data. A recent study by the Ministry of Road Transport and Highways (MoRTH) revealed that Tamil Nadu has the highest number of road accidents in the country, with 64,105 accidents reported in 2022 accounting for 13.9% of total accidents across the nation. But we can see that there's decline of approximately 15.1% as compared to that of 2021[4]. The nation's capital Delhi ranked 19th. This points towards a crucial need for targeted interventions and stricter enforcement of traffic regulations in such high-risk zones.

Despite the rapid economic development India still experiences a critical gap in public awareness about road safety. The majority of people lack knowledge about road safety and safe driving according to a recent survey. Moreover, rural people lack knowledge about traffic rules and other safety measures. This lack of awareness, coupled with the existing infrastructure challenges, creates a perfect storm for preventable tragedies.

©IJRASET: All Rights are Reserved

The purpose of this research is to investigate the reasons behind road accidents in India, analyze their impact on individuals, families, and society, and suggest solutions to create a safer road environment for everyone. By addressing these issues, India can effectively transform its extensive network into a secure and efficient path for progress.

TABLE I
ACCIDENTAL STATISTICS FOR THE YEARS 2021 AND 2022

Parameters	**2021**	**2022**	**%age change**
Total number of accidents	4,12,432	4,61,312	11.8
Total number of people killed	1,53,972	1,68,491	9.4
Total number of people injured	3,84,448	4,43,366	15.3

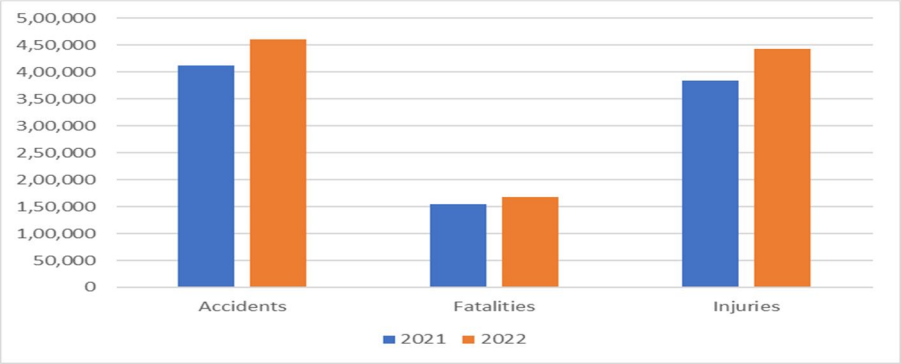


Fig 1: Accidental Statistics for the years 2021 and 2022

II. PROPOSED MODEL

Classification algorithms, as a form of supervised learning, aim to organize data into predefined classes or categories using input features. In the realm of road safety, these models can be utilized to forecast if a particular combination of factors—such as weather conditions, road type, time of day, lighting conditions, and vehicle attributes—might lead to an accident. Through the examination of historical accident data and the discernment of patterns within diverse contributing factors, classification algorithms offer valuable insights into the severity associated with various types of road accidents.

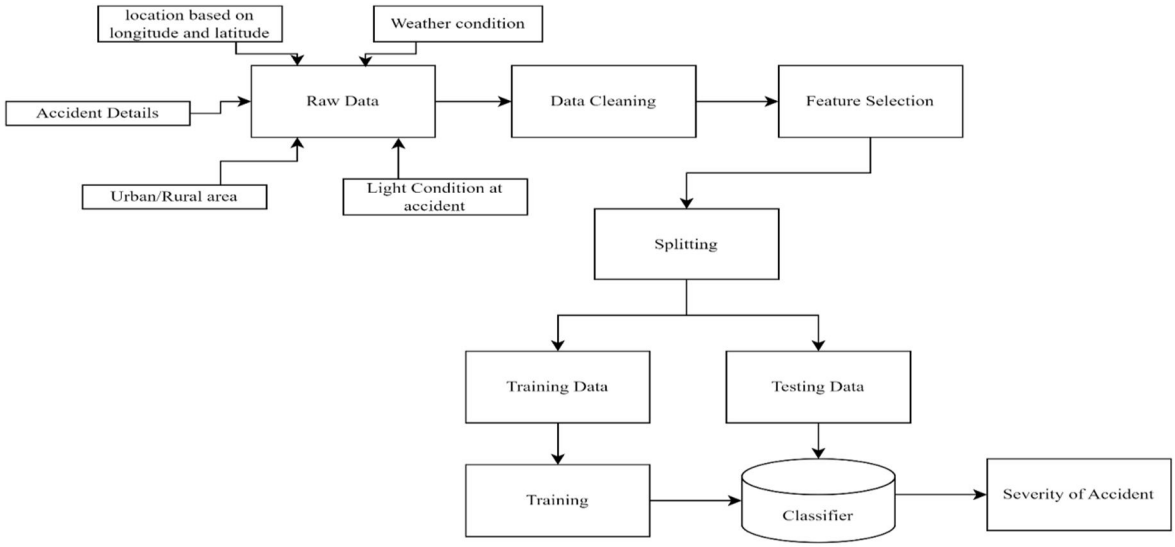


Fig 2: System architecture diagram for proposed models

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

In our study, we intend to utilize both Decision Tree, K-Nearest Neighbour (kNN), Logistic Regression, and Random Forest algorithms for classifying road accidents based on the severity of the accident, with 0 being the slight injury/no injury and 2 being fatal). We will employ a comprehensive dataset comprising historical accident records along with environmental, infrastructural, and vehicular attributes. By assessing the performance of these classification models and examining their classification, our goal is to pinpoint the primary risk factors linked to road accidents and the severity of the accident so that they can be worked upon as a matter of precautionary measure.

A. Data Collection(Raw Data):

The road accident dataset from Kaggle provides useful insights on traffic incidents, including details such as accident date and time, location (latitude and longitude), severity of the accident, number of vehicles involved, and weather conditions at the time of the incident. This dataset enables analysis of factors contributing to accidents and can be used to develop classification models for accident severity based on various parameters. Respected authorities can utilize this dataset to implement targeted interventions aimed at reducing road accidents and improving overall road safety.

B. Data Preprocessing

Preprocessing the dataset involves addressing missing values through imputation techniques and handling outliers to prevent skewed analysis results. Categorical variables are encoded into numerical format, employing techniques like one-hot encoding or label encoding, while numerical variables are standardized or normalized to ensure equal contribution to the analysis. Overall, preprocessing ensures the dataset is clean, consistent, and optimized for subsequent analysis and modeling tasks.

C. Exploratory data analysis (EDA):

During the exploratory data analysis (EDA) of the road accident dataset, we dive into its composition to grasp its underlying characteristics. Descriptive statistics offer insights into numerical attributes, highlighting their distributions. Visualizations such as histograms and density plots provide a graphical representation of these distributions. Bar charts are employed to unveil patterns within categorical variables, while correlation matrices aid in identifying potential predictors of accident occurrence or severity. Through EDA, we gain crucial insights into data distribution, interrelationships among variables, and potential predictors, which serve as a foundation for subsequent analyses.

D. Model Selection:

In road accident severity classification, amidst a range of algorithms, Random Forest emerges as notable for their interpretability and efficacy in unveiling complex relationships within the dataset. Random Forest partitions the feature space into hierarchical rules, offering transparent insights into the factors impacting accident severity. Although Decision Tree, K-Nearest Neighbour (kNN), and Logistic Regression also play crucial roles in accident analysis, our focus lies on Random Forest for classifying accident severity.

E. Model Training and Evaluation:

To train and evaluate predictive models for road accident classification, the dataset is divided into two subsets: a training set and a testing set. The training set is used to train the models, while the testing set is kept separate to evaluate their performance. Cross-validation techniques, such as k-fold cross-validation, are then applied to assess how well the trained models generalize to unseen data.

F. Model Interpretation:

Analyzing the trained models is critical for extracting insights into the determinants of road accident severity. Through examination of feature importance scores (such as accuracy, F1-score, precision, and recall) and decision boundaries, we can find the underlying mechanisms driving accident severity prediction. These scores highlight the variables that influence the most on classifying accident severity, helping respected authorities to prioritize necessary infrastructural changes effectively.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

G. Validation and Deployment:

After training, it's crucial to thoroughly validate these models using suitable methods to ensure their effectiveness on unseen data and their capacity to generalize beyond the initial training data. Throughout the validation process, a range of performance metrics is computed to assess the models' predictive accuracy. These metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), offer valuable insights into the models' capability to accurately distinguish between slight/no harm and severe and fatal injuries. Additionally, they provide a comprehensive overview of the models' overall performance compared to baseline or benchmark models.

III. METHODOLOGY

A. Decision Tree:

The Gini index is used in a decision tree for classification to identify the optimal feature and threshold for data splitting at each node. To produce pure leaf nodes that reflect particular class labels, the method recursively chooses splits that minimize the impurity of a node's class distribution, as measured by the Gini index. Data is predicted by moving through the tree according to learned rules until it reaches a leaf node, which is then classified.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

B. k-Nearest Neighbour (kNN):

kNN is a classification technique that relies on the similarity of features. The data are analyzed, their similarities and distances from one another are measured, and K values are used to cluster the data. There are several methods for calculating distance; in our study, we employed the Euclidean distance measure. By measuring the distance between the clusters and allocating the class of fresh input data to the closest one, it is classed.[6]

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

C. Random Forest:

Random Forest is an ensemble learning technique that combines predictions from multiple individual decision trees to enhance predictive accuracy and reliability. This method involves training numerous decision trees and then aggregating their outputs to determine the final prediction.

IV. RESULTS AND DISCUSSIONS

The performance of different machine learning models, including Random Forest, k-Nearest Neighbors (kNN), and Logistic Regression, was evaluated using various evaluation metrics such as precision, recall, F1 score, and accuracy. The experiments were conducted on [describe dataset or datasets used for evaluation] to assess the effectiveness of these models in [briefly mention the task or problem being addressed, e.g., binary classification, regression].

Random Forest achieved an 87% accuracy, 79% precision, 87% recall, and 81% F1 score. The ensemble nature of Random Forest enabled it to capture complex patterns in the data and achieve robust performance, particularly in handling imbalanced datasets.

k-Nearest Neighbors (kNN) demonstrated 85% accuracy, 79% precision, 85% recall, and 81% F1 score. kNN's performance varied based on the choice of (k) and the characteristics of the dataset, with better results observed for datasets with smooth decision boundaries and well-separated classes.

Decision Tree yielded an accuracy of 86%, precision of 78%, recall of 86%, and F1 score of 81%. Decision Trees demonstrated clear interpretability and performed well on datasets with discrete and hierarchical features

Overall, the choice of the most suitable model depends on the specific characteristics of the dataset and the desired trade-offs between precision, recall, F1-score, and accuracy. The results suggest that each model has its strengths and weaknesses, and understanding these nuances is essential for selecting the appropriate model for a given task or application.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

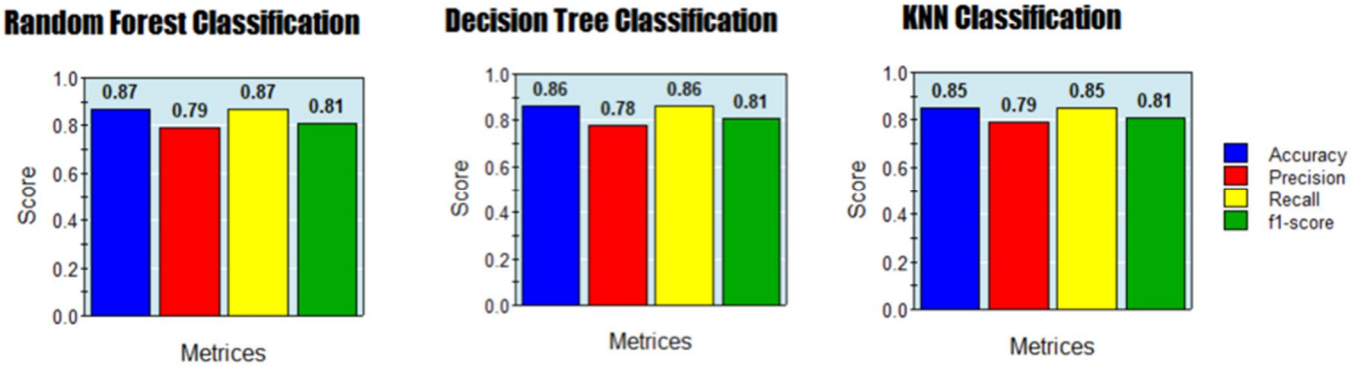


Fig 3: Scores of various classification algorithms

V. CONCLUSION

Worldwide, road accidents are a major concern as the ninth leading cause of death. These figures are frighteningly high even in India. This is an unacceptable situation that needs to be dealt with urgently. Preventing such incidents requires concerted efforts across different areas. One approach stresses engineering controls like road improvements as opined by MoRTH.[1] They suggested that upgrading infrastructure would greatly reduce road traffic accidents. Another way includes administrative measures which entail the enactment of stricter traffic laws. Additionally, behavioral interventions play an essential part in this process.

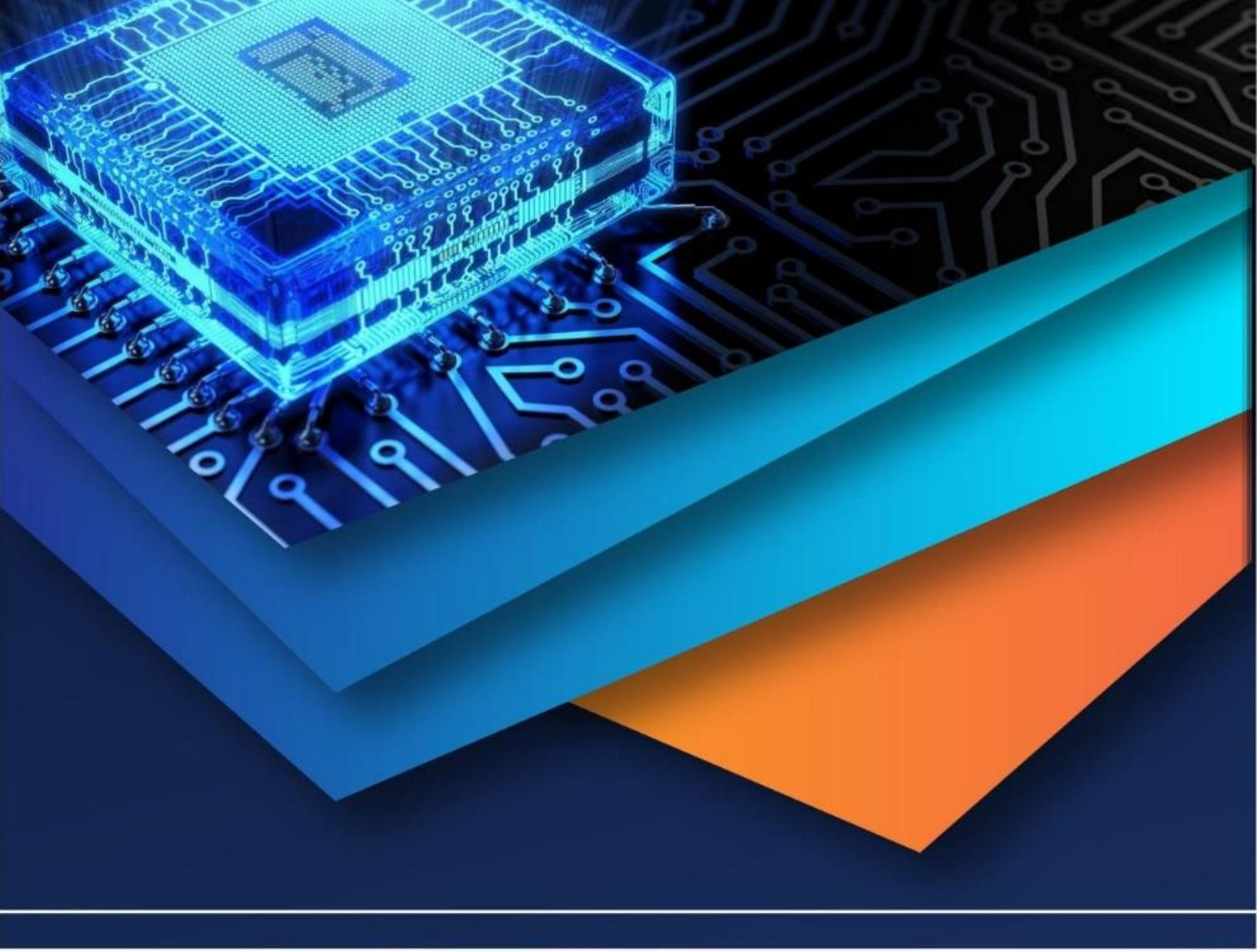
In this study, we performed an extensive analysis and classification of road accidents using machine learning models: Random Forest, k-Nearest Neighbors (kNN), and Decision Tree (DT). Our goal was to predict and categorize road accidents based on multiple factors and assess the effectiveness of each model. Each of these models demonstrated unique strengths and performance metrics, highlighting the significance of employing machine learning methods to support road safety initiatives and guide targeted interventions aimed at accident prevention. This research underscores the importance of utilizing advanced analytical techniques to enhance our understanding of road safety dynamics and inform evidence-based strategies for accident reduction.

VI. ACKNOWLEDGEMENT

We would like to express our sincere appreciation to Prof. Rashmi Rane for her excellent advice, knowledge, and support at every stage of our research effort. Her constructive criticism have greatly influenced the direction and quality of this work. We also want to express our gratitude to Dr. Vishwanath Karad of MIT World Peace University in Pune, India, for giving us access to all of the resources we needed to do this work. Additionally, We would want to express our gratitude to our family and friends for their patience, support, and understanding throughout this academic journey.

REFERENCES

[1] Ministry of Road Transport and Highways, Government of India: https://morth.nic.in/road-accident-in-india
[2] https://worldpopulationreview.com/country-rankings/road-quality-by-country
[3] World Health Organization: https://www.who.int/publications-detail-redirect/9789240086517
[4] Ministry of Road Transport and Highways, Government of India: https://morth.nic.in/sites/default/files/RA_2022_30_Oct.pdf
[5] Z. Zhang, “Introduction to machine learning: k-nearest neighbors,” Annals of Translational Medicine (2016), vol 4, issue no. 11, pp.218- 218.
[6] H. İ. Bülbül, T. Kaya and Y. Tulgar, "Analysis for Status of the Road Accident Occurrence and Determination of the Risk of Accident by Machine Learning in Istanbul," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 426-430.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)