ROAD ACCIDENT PREDICTION USING MACHINE LEARNING

Viswanadhuni Reshma Priya[1], Shaik BB Ayesha[2], Vennapusa Venkata Vamsi[3],

Ms. Ramya Palli[4]

Department of Electronics and Communication Engineering,

SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.

ABSTRACT:

Due to the escalating number of vehicles on roads, the incidence of daily traffic accidents is rising dramatically. This surge in accidents and fatalities underscores the urgency of accurately forecasting accident rates over time to facilitate informed decisions by transportation authorities. Analyzing accident occurrences can provide valuable insights for developing effective strategies to mitigate accidents. Although accidents often exhibit inherent uncertainty, patterns of regularity emerge over time in specific areas. Leveraging these patterns can enable us to develop robust accident prediction models. Our study focuses on examining the correlations between road accidents, road conditions, and environmental factors to enhance accident prediction accuracy.

Utilizing machine learning techniques, including the K-Nearest Neighbors (K-NN) Algorithm, Support Vector Machines (SVM), Random Forest Algorithm, Naive Bayes Algorithm, and Decision Tree Algorithm, this paper aims to construct a predictive model for accident occurrences. By harnessing the power of these algorithms, this paper improves understanding of accident dynamics and contributes to the development of proactive measures to reduce accidents on roads.

Keywords: Road Accident, Machine Learning (ML), Dataset.

INTRODUCTION:

Road accidents are a major hazard to public safety, causing injuries, deaths, and economic losses around the world. In recent times, advances in machine learning have created new opportunities for accident prevention and reduction. This paper use data-driven methodologies to anticipate the likelihood of traffic accidents grounded on a variety of contributing factors. Our exploration focuses on examining historical accident data, meteorological conditions, road infrastructure, and vehicle characteristics. Machine learning methods are employed in this project to construct accurate models capable of estimating accident probabilities. These predictive technologies hold the potential to assist traffic authorities, emergency services, and policymakers in devising visionary strategies to reduce accident rates.

This paper investigates five established machine learning techniques: Random Forest,

Decision Tree, Naive Bayes, K-nearest neighbors (KNN), and Support Vector Machine. The aim is to discern the most effective accident prediction algorithm through trial and review. The findings of this paper contribute to the enhancement of road safety initiatives, ultimately fostering a safer transportation environment for all.

LITERATURE REVIEW:

The field of road safety and accident prediction has witnessed significant advancements with the application of machine learning algorithms. Various studies have explored different methodologies and models to analyze factors influencing road accidents and predict accident severity.

Chen et al. [1], highways emerge as the predominant location for the occurrence of a significant portion of reported accidents.

Williams et al. [2], revealed that the age and driving experience of individuals significantly influence the frequency of accidents.

Sarkar et al. [3] conducted a comparative analysis investigating the prevalence of accidents across various road types. In addition to examining the factors contributing to accidents, their study revealed a higher incidence of accidents on highways compared to standard roads similar to [1].

Zheng et al. [4] conducted research focusing on the spectrum of injuries resulting from motor vehicle accidents. Additionally, they investigated the emotional state of the drivers involved, exploring its potential influence on the occurrence of accidents.

Tessa K. Anderson et al. [5] proposed a method aimed at identifying high-density casualty zones. This strategy involves implementing a clustering procedure to identify clusters where stochastic events are likely to occur. Consequently, it allows for the assessment of their occurrence within these clusters.

Briz-Redon et al. [6] took into account several factors including the log-straight model, driver characteristics, pedestrian traits, road traffic, and vehicle classifications. This comprehensive approach provides clear insights into the factors influencing casualties in school zones.

De Ona et al. [7] employed Latent Class Clustering and Bayesian methodologies in their examination of automobile collisions to identify the key determinants of casualty severity. The simultaneous application of these two techniques is particularly noteworthy as it unveils supplementary insights into the data.

Mahendra G et al. [8] devised a tool for detecting reckless driving on roadways and promptly notifying traffic authorities in the event of any speed infractions.

Jamal Raiyn et al. [9] present a model that identifies traffic incidents by analyzing the speed fluctuations of vehicles located both upstream and downstream of a specific point on the highway.

X. Gao et al. [10], a novel algorithm termed Weighted Quantitative Random Forest (WQRF) was introduced. This algorithm aims to forecast employee salary turnover within various industries. The predictive model incorporates several key features, including overtime hours, age, monthly

income, distance from home, and tenure within the company. These factors collectively contribute to the model's ability to anticipate shifts in employee salary turnover rates.

Our paper builds upon these research findings by employing a comparative study of machine learning algorithms including Random Forest, Decision Tree, SVM, KNN, and Naive Bayes. Through rigorous data preprocessing, feature selection, and model training, this project aims to develop a robust predictive model for road accident severity prediction. By synthesizing insights from existing literature and leveraging state-of-the-art machine learning techniques, this project seeksto contribute to the advancement of road safety measures and accident prevention strategies. This paper aims to provide valuable insights for stakeholders in traffic management, law enforcement agencies, and policymakers to make informed decisions and improve road safety for all.

PROPOSED METHODOLOGY:

In the context of road accident prediction, the process involves several key steps. First, relevant accident data is collected and pre-processed, addressing issues like missing values and outliers. Feature engineering follows, where meaningful features are extracted from the data. The dataset is then split into training and testing subsets. Next, machine learning models (including Random Forest, Decision Tree, Naive Bayes, KNN, and SVM) are selected and trained using the training data. Model evaluation metrics, such as accuracy and precision, guide the choice of the best- performing model. Applying the chosen model(s) to the testing data yields accident probability predictions. Finally, result analysis provides insights for decision-makers and safety improvements. Overall, this process aims to enhance road safety and prevent accidents**.** The architecture of accident prediction is shown in Fig.1.

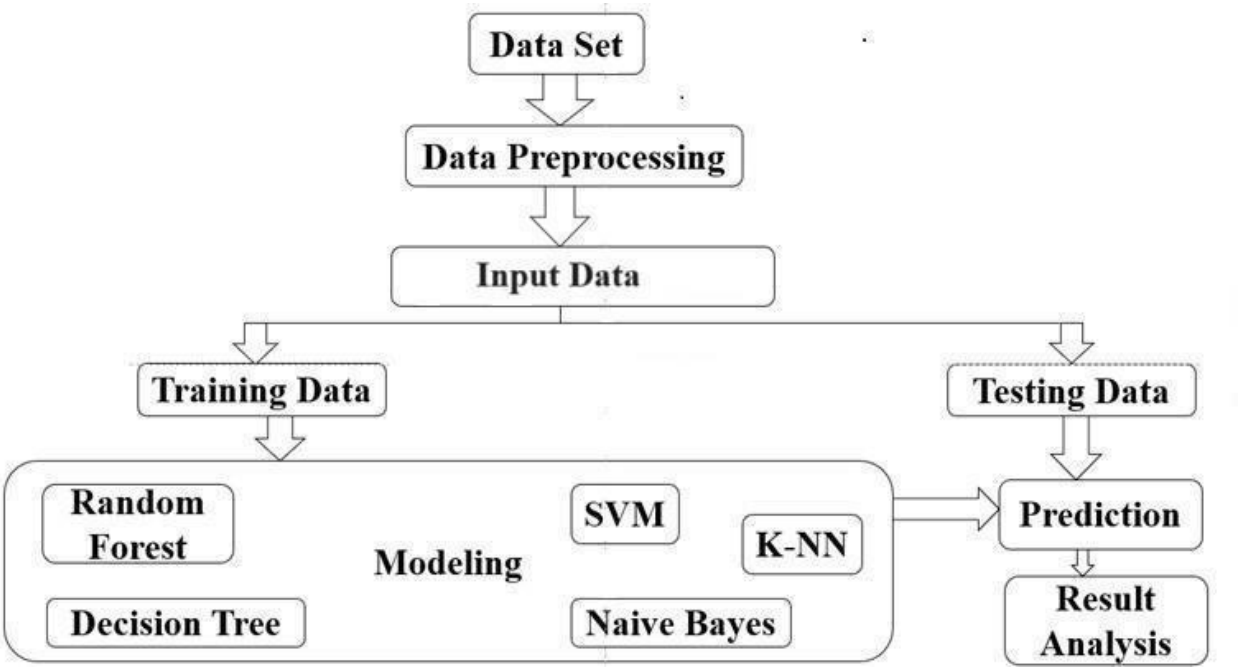

Fig.1: Accident Prediction Architecture

Dataset: The dataset on road accidents plays a pivotal role in our paper's capacity to estimate accident probability accurately. Through an examination of this dataset, patterns, correlations, and risk factors associated with accidents can be discerned. The data furnishes crucial insights into variables such as weather conditions, road types, vehicle characteristics, and light conditions. By Leveraging machine learning algorithms, this paper aims to develop predictive models that can assessthe likelihood of accidents. These models hold promise for enhancing road safety measures and contributing to the reduction of accidents.

Data Preprocessing: The code starts by importing necessary libraries. pandas(pd) is a powerful tool for data manipulation and analysis in Python. It allows us to load data from various sources (like CSV files in this case) and perform operations like viewing summaries, cleaning, and transforming the data. NumPy (np) is another essential library for scientific computing. It provides efficient mathematical functions and data structures that are often used in machine learning algorithms. Finally, the warnings library helps suppress warnings that might clutter the output during data loading.

i. **Data Acquisition:** Load the road accident information. contain numerous characteristics such as vehicle type, road conditions, and accident severity.
ii. **Identifying Categorical Features:** Datasets include categorical attributes that indicate non-numerical attributes. In this case, features such as vehicle type, road conditions, daylight, and weather conditions are most likely classified. This allows us to better understand the distribution of categorical data and plan for appropriate encoding methods.
iii. **Transforming Categorical Data (Label Encoding):** Since machine learning algorithms typically work best with numerical data, the code employs label encoding to transform categorical features. This method replaces the original category labels (e.g., "sunny" and "rainy") with numerical values (e.g., 0, 1). This allows machine learning models to understand the relationships between these features and the target variable more effectively.

TRAINING AND TESTING DATA:

Splitting the data into two sets: training and testing. The training set (typically 70% of the data) is used to train the machine learning models, while the testing set (the remaining 30%) is used to evaluate their performance on unseen data. This helps prevent overfitting, where the models become too good at fitting the training data but perform poorly on new data.

Random Forest:

Random Forest is a powerful ensemble learning technique for classification and regression tasks. It combines the predictions of multiple decision trees to create a more robust and accurate

model.

i. **Bootstrap Aggregation (Bagging):** Instead of using the entire dataset to train a single decision tree, Random Forest draws multiple random samples (with replacement) from the original data. These samples are called bootstrap replicates. This injects randomness and helps prevent the trees from overfitting the specific training data.

ii. **Decision Tree Building:** For each bootstrap replicate, a decision tree is constructed. These trees can have different depths and may use different features at each split. This diversity helps the forest to capture a wider range of patterns in the data.

iii. **Random Feature Selection:** At each node of a decision tree in the forest, instead of considering all features for splitting, a random subset of features is chosen as candidates for the split. This further increases diversity among the trees and reduces the chance of overfitting to irrelevant features.

iv. **Making Predictions:** In classification tasks (like predicting accident severity) the most frequent class predicted by the trees becomes the final prediction for the forest. In regression tasks, the average of the individual tree predictions is used as the final prediction.

Decision Tree:

A decision tree is a supervised learning algorithm that resembles a tree structure for classification and regression tasks. It works by recursively splitting the data based on features that best distinguish between different classes or predict a continuous value.

i. **Building the Tree:** The algorithm starts with the entire dataset as the root node. It chooses the most informative feature (the one that best separates the data) to split the nodeinto child nodes. This selection can be based on various measures like information gain (classification) or variance reduction (regression). The splitting process continues recursively on each child node, using the best remaining feature for further separation. The process stops when a stopping criterion is met, such as reaching a certain depth in the tree, having pure classes (classification) or minimal variance (regression) in a node, or having no more informative features to split on.

ii. **Making Predictions:** Once the tree is built, a new data point (whose class or value needs to be predicted) is passed through the tree. At each node, the value of the corresponding feature in the data point is compared to the splitting threshold. The data point is directed to the left or right child node based on the comparison. This process continues until the data point reaches a leaf node, which represents the predicted class (classification) or predicted continuous value (regression).

iii. **Information Gain (Classification):** This measure calculates how much "purity"

(separation between classes) is gained by splitting on a particular feature. Higher information gain indicates a more informative split.

Information (Feature A)=Entropy (Parent) - Σ [Entropy (Child)* Proportion (Child)] (1)

iv. **Gini Impurity (Classification):** Another common measure for calculating impurity, particularly for imbalanced datasets.

Impurity (Parent) = 1 - Σ [Proportion (Class i) ^2] (2)

K-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) is a non-parametric, supervised machine learning algorithm used for both classification and regression tasks. It works by classifying a data point based on the similarity of its k-nearest neighbors in the training data.

i. **Data Representation:** Each data point is represented as a feature vector, containing values for all features (e.g., vehicle type, speed, weather).

ii. **Distance Metric:** A distance metric is chosen to measure the similarity between data points. Common choices include Euclidean distance, Manhattan distance, or Mahala Nobis distance.

iii. **K Selection:** K, the number of nearest neighbors to consider, is a crucial parameter. A higher k value leads to smoother decision boundaries but might be more susceptible to noise. A lower k value can capture local variations but might be prone to overfitting.

iv. **Classification (for accident severity prediction):** Calculate the distance between the new data point and all points in the training data. Identify the k nearest neighbors based onthe chosen distance metric. Determine the most frequent accident severity class (e.g., high, low) among these k neighbors. Assign the new data point to the most frequent class, predicting its accident severity.

v. **Euclidean Distance:**

distance (x1, x2) = sqrt (sum ((x1_i – x2_i)^2 for i in features)) (3)

Naive Bayes Classifier: A Probabilistic Approach

The Naive Bayes classifier is a popular supervised learning algorithm for classification tasks. It works based on Bayes' theorem, a fundamental concept in probability theory that allows us to calculate the conditional probability of an event (accident severity in this case) occurring given the presence of certain features (vehicle type, road conditions, etc.).

The core formula for Naive Bayes classification is:

P (Class | Features) = (P (Features | Class) * P (Class)) / P(Features) (4)

Support Vector Machine (SVM):

SVM is a powerful machine-learning algorithm for classification tasks. It aims to find a hyperplane in the feature space that best separates the data points belonging to different classes. Here's a breakdown of the algorithm with formulas and its application in your road accident severity prediction project.

i. **Hyperplane Equation:** A hyperplane in n-dimensional space can be represented by the equation:

$$w^T * x + b = 0 \quad (5)$$

ii. **Margin:** The margin between the hyperplane and a support vector is calculated as:

$$|w^T * x_s + b| / ||w|| \quad (6)$$

The objective of SVM is to maximize this margin, which intuitively translates to finding the best separation between the classes. In summary, SVM aims to find the optimal hyperplane $w^T*x+b=0$ that separates the data points with the maximum margin while penalizing misclassifications based on the regularization parameter C. The use of kernel functions allows SVM to handle non-linearly separable data by mapping it to a higher-dimensional space where linear separation is possible

RESULTS:

Table 1: Evaluation Metrics of Classification Algorithms

Algorithm	Accuracy	Precision	f1-score	Recall
Random Forest	0.85	0.76	0.78	0.85
Decision Tree	0.84	0.75	0.78	0.84
SVM	0.85	0.72	0.78	0.85
KNN	0.84	0.74	0.78	0.84
Naive Bayes	0.84	0.72	0.78	0.84

Table 1. represents the evaluation metrics of classification algorithms. The Random Forest algorithm emerged as the top-performing model in our road accident prediction project, showcasing the highest accuracy among the algorithms tested. So, in our project, the Random Forest algorithm is used in web development to predict the accident probability.

Web Development

A web page is designed to improve the user experience in this project. Offering a user- friendly platform, this web interface allows users to quickly and efficiently obtain predictive outputs without complexity. Users can easily input parameters such as light conditions, vehicle type, weather

conditions, and road type to receive accurate predictions on the likelihood of road accidents occurring. we will get the URL http://127.0.0.1:5000/ for the web page and we can run this URL in any browser. The web page is shown in Fig. 2.

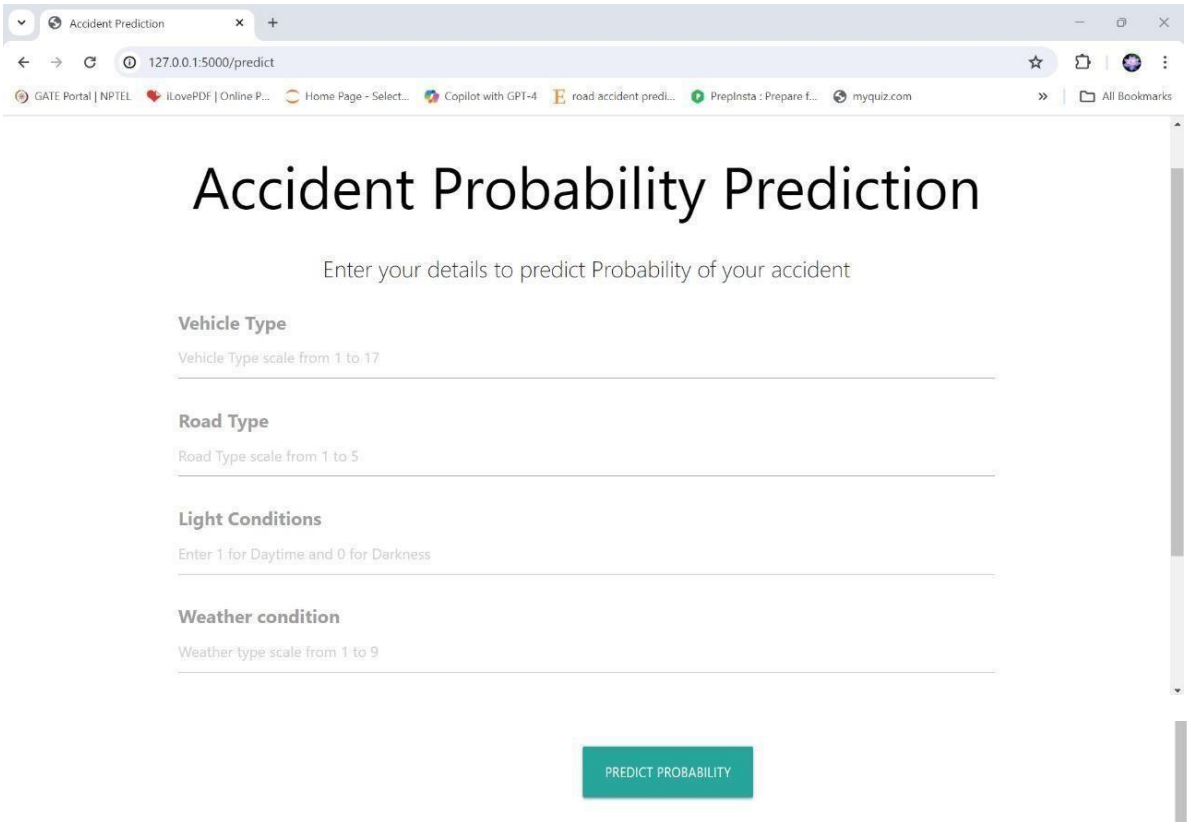

Fig.2: web page

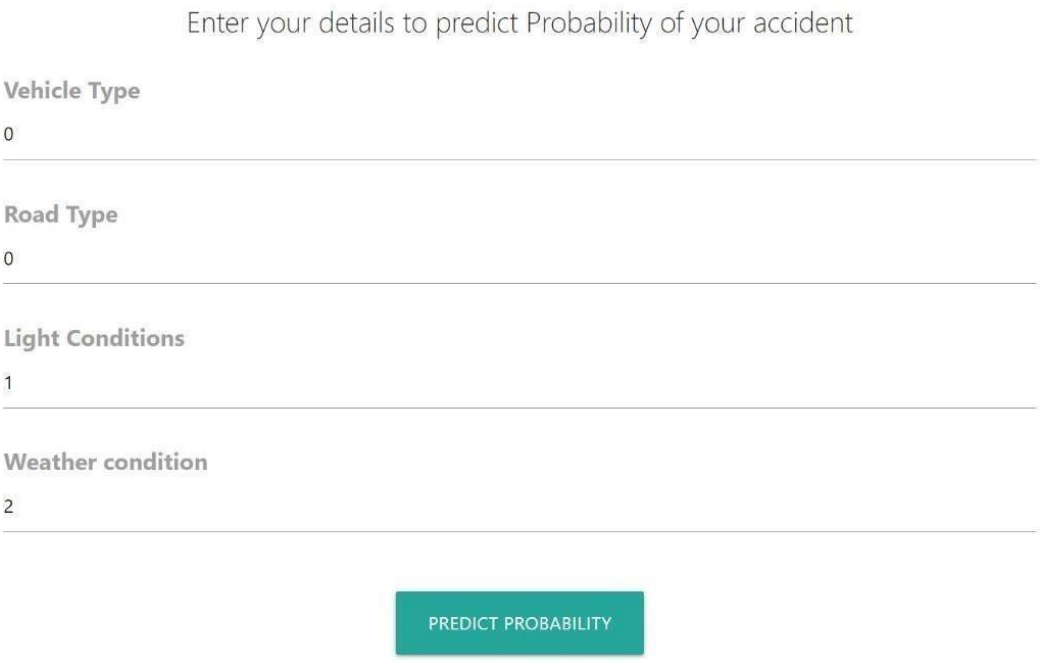

Fig.3: Output for Low Probability of Accident

Fig. 3 consists of parameters Vehicle type – Automobile, Road Type – Asphalt roads, Light Conditions – Daylight, and Weather conditions – Normal.

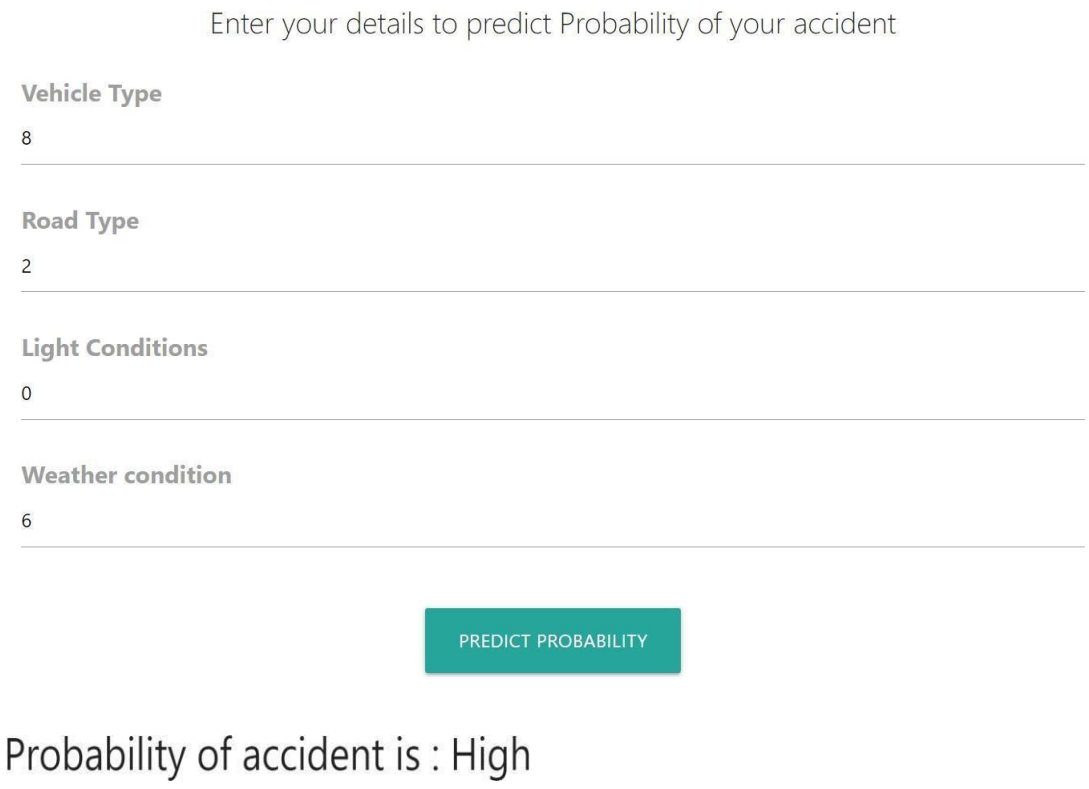

Fig.4: Output for High Probability of Accident

Fig.4 consists of parameters Vehicle type – Public (> 45 seats), Road Type – Earth roads, Light Conditions – Darkness, and Weather conditions – Snow.

CONCLUSION

Road accidents have a profound impact on individuals and communities, highlighting the importance of proactive measures to reduce their occurrence. While it is each individual's responsibility to adopt safe driving practices, the role of road development authorities and automobile industries in creating safer infrastructure and vehicles cannot be understated. However, addressing the multifaceted nature of accidents requires a comprehensive approach that includes predictive modeling based on historical data and regulatory compliance. Our project successfully developed an application capable of efficiently predicting road accidents by leveraging machine learning algorithms. By analyzing factors such as vehicle types, light and weather conditions, and road types, our model can assess the risk probability of accidents in different areas with a high degreeof accuracy.

The application of machine learning algorithms, including Random Forest, Decision Tree, SVM, KNN, and Naive Bayes, over a carefully curated dataset enabled us to create a robust predictive model. Moving forward, the integration of such predictive models into road safety initiatives can significantly contribute to accident reduction efforts. By utilizing data-driven

approaches, authorities can prioritize resources and interventions effectively, leading to safer roads and reduced casualties. In conclusion, our project showcases the potential of machine learning in enhancing road safety and accident prediction.

REFERENCES

[1] Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. J. Inf. Sci. Eng 31(2): 573-595. http://dx.doi.org/10.1504/IJASM.2015.068609.

[2] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009 V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01.

[3] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. http://dx.doi.org/10.14299/ijser.2013.01.

[4] Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short clustering. Neurocomputing http://dx.doi.org/10.1504/IJIT.2018.090859.

[5] Tessa KA. Kernel density estimation and K-means clustering to profile road accident hotspots. Elsevier. Accident Analysis and Prevention. 2009; 41:359–364.

[6] A. Briz-Redon, F. Martinez-Ruiz, and F. Montes. Estimating the occurrence of traffic accidents near school locations: A case study from Valencia (Spain) including several approaches. Elsevier. Accident Analysis & Prevention. 2019;132.

[7] De Ona. J, Lopez. G, Mujalli. R, Calvo.F. J. Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks. Elsevier. Accident Analysis and Prevention. 2013;51:1-10.

[8] Mahendra G, Dayananda R B. Vehicle rash drive control system. International Journal for Research in Engineering Application and Management. 2018;04(03):676-681.

[9] Jamal Raiyn, Tomer Toledo. Real-time road traffic anomaly detection. Journal of Transportation Technologies. 2014; 4(3):256-266.

[10] X. Gao, J. Wen, C Zhang. An improved random forest algorithm for predicting employee turnover. Hindawi. 2019; 4140707:12 p.

[11] World Health Organization. Road Traffic Injuries. Available online: https://www.who.int/news-room/fact-sheets/detail/road traffic-injuries (accessed on 20 June 2021).