Comparative Analysis of Classification Algorithms for Road Accident Prediction

Tejaswini Durge
Department of Information Technology
Vishwakarma Institute of Information
Technology, Pune
tejaswinidurge41@gmail.com

Soham Deshmukh
Department of Information Technology
Vishwakarma Institute of Information
Technology, Pune
sohamatuldeshmukh2003@gmail.com

Aniket Bansod
Department of Information Technology
Vishwakarma Institute of Information
Technology, Pune
aniketbansod2004@gmail.com

Sanika Boke
Department of Information Technology
Vishwakarma Institute of Information
Technology, Pune
sanikaboke@gmail.com

Dr. Pravin Futane
Department of Information Technology
Vishwakarma Institute of Information
Technology, Pune
pravin.futane@viit.ac.in

***Abstract*—In recent years, the rising number of road accidents has become a significant public concern, calling for intelligent and data-driven approaches to predict and prevent such incidents. This study presents a comparative analysis of multiple classification algorithms aimed at predicting the likelihood of road accidents based on historical and real-time traffic data. By evaluating models such as Decision Trees, Random Forest, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN), the research aims to identify the most accurate and efficient algorithm for practical deployment. Our findings indicate that ensemble methods, particularly Random Forest, consistently outperform other algorithms in terms of accuracy, precision, and recall. However, the performance of models can vary depending on the quality and granularity of input features such as weather conditions, time of day, road type, and traffic volume. This analysis emphasizes the importance of feature selection and data preprocessing in enhancing model accuracy. Ultimately, the research not only guides the selection of suitable classification models for accident prediction but also paves the way for integrating such systems into smart city infrastructures to enhance road safety and response planning. issues.**

***Index Terms*—Machine Learning, KNN, SVM, Decision Trees, Random Forest, Classification, Prediction**

I. Introduction

Road accidents kill thousands of people every year and injure countless others, and are among the most critical world wide public health concern. Consider a family driving home from a party, when a tragic crash is precipitated by a mix of bad visibility, overspeeding, and delayed driver response. Such sad accidents are not unusual. Actually, car crashes frequently strike unheralded, inflicting lasting trauma in place of jubilation. Governments and transportation authorities globally are persistently on the lookout for technologies that will make these happen less, and prediction models based on data have emerged as one of the most potential fronts.

Under the umbrella of smart cities , AI being able to foresee car crashes ahead of time could significantly revolutionize how we conceptualize road safety. Machine learning, especially classification algorithms, has shown its promise in detecting patterns and correlations from past accident records. These algorithms can examine many features like time, weather conditions, traffic, road types, and human actions to forecast the probability of accidents. This transformation from reactive to proactive safety management can be the difference between saving lives and dollars.

This research work explores a comparative analysis of some of the widely used classification algorithms, namely Decision Trees, Random Forest, Support Vector Machines, and K-Nearest Neighbors, in order to determine their performance towards predicting road accidents. Utilizing past accident datasets, we compare each model with respect to crucial performance measures such as accuracy, precision, recall, and F1-score. The final goal is to select the most appropriate algorithm that best trades off between performance and interpretability, as well as computational costs.

A practical instance may be considered in urban centers such as Mumbai or Los Angeles, where driving patterns are typically prone to irregularity and jamming. Were a model that could predict accident risks identify vulnerable zones or hours—e.g., Saturday night in areas close to entertainment clusters—the authorities can station preventive systems such as patrolling, alarm systems, or even automated re-routing. Technology's integration into civic administration shows a glimpse into a future more secure.

In this paper, we not only compare the effectiveness of these algorithms but also explore the influence of data quality, feature selection, and model tuning on their predictive power. This research aims to contribute to the ongoing efforts in intelligent transportation systems and automated road safety management.

TABLE I
COMPARATIVE LITERATURE REVIEW

Sr. No.	Paper Title	ML Techniques Used	Contributions / Literature Review Summary
1	*Road Accident Prediction Using Machine Learning*, Sharma et al. (2020) [6]	Decision Trees (DT), Random Forest (RF), Logistic Regression (LR)	This study highlights the application of Random Forest in accident prediction, achieving the highest accuracy. The study underscores the importance of ensemble learning models in predicting road accidents and their reliability. It emphasizes the potential of machine learning for enhancing traffic safety.
2	*A Machine Learning Approach for Road Accident Prediction*, Ahmed et al. (2021) [5]	K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machines (SVM)	The paper investigates the effect of feature engineering and data preprocessing in accident prediction models. The use of feature selection improved model performance, and the study demonstrates how preprocessing can significantly impact the predictive power of machine learning models.
3	*Traffic Big Data Accident Risk Prediction*, Chen et al. (2020) [10]	Support Vector Machines (SVM), Deep Learning (DL)	This study explores the role of real-time data such as weather, time, and sensor information in enhancing the prediction of road accidents. The research emphasizes the importance of incorporating contextual data for improving model accuracy and reliability in accident risk prediction.
4	*Comparative Study on Accident Severity Prediction Using Machine Learning Models*, Patel et al. (2022) [15]	Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR)	Patel et al. compare the performance of various machine learning models in predicting accident severity. They found that Random Forest consistently outperforms other models across multiple datasets. The paper also highlights the strengths of ensemble models in dealing with large and complex datasets.
5	*Review on Traffic Accident Models Using Machine Learning Techniques*, Gupta et al. (2021) [7]	Ensemble Learning, Deep Learning (DL), Tree-based Models	Gupta et al. provide a comprehensive review of different machine learning models used in traffic accident prediction. The paper identifies the strengths and weaknesses of various algorithms and advocates for ensemble methods and deep learning due to their robustness in handling traffic accident data.
6	*Road Accident Prediction Using Machine Learning and Deep Learning*, Srivastava et al. (2023) [8]	Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Random Forest (RF)	This paper explores hybrid models combining image data (e.g., from cameras) with tabular data for accident prediction. The study demonstrates how deep learning can improve the predictive capabilities of road accident models and provides insights into integrating smart city technologies.
7	*Real-Time Intelligent Accident Prediction System Using GPS and Weather Data*, Jain et al. (2022) [4]	Decision Trees (DT), GPS-based Data, Feature Engineering	Jain et al. focus on real-time data for accident prediction, particularly GPS and weather data. The study demonstrates how incorporating real-time and spatial data can improve the prediction accuracy of road accident models, especially for dynamic environments.

II. LITERATURE REVIEW

1) Identification of crash propensity based on traffic speed conditions Abdel-Aty and Pande were among the pioneers in conducting a study of traffic accident forecasting, where they investigated the potential for crashes based on the analysis of certain traffic speed regimes. In their study, they used comprehensive traffic flow and speed data, collected using loop detectors, to predict the probability of real-time crashes on urban highways. They showed that below-average and highly varying traffic speeds are good crash predictors.

In the study, logistic regression models were used to provide an estimate of the likelihood of crash events occurring, thus forming the basis of binary classification modeling in traffic safety analysis. This is in alignment with our endeavor, where an accident's existence or nonexistence is considered as a binary classification problem. Their approach in correlating traffic-specific and environmental variables with crash events justifies using supervised learning in such predictions. [6]

2) Learning from Imbalanced Data He and Garcia also solved an essential problem in predictive modeling of rare occurrences like accidents: class imbalance. In nearly all real-world datasets, cases that are not accidents far outnumber accident cases, and this can significantly impair the performance of machine learning classifiers. Their paper summarized different methods for addressing this imbalance, such as oversampling, undersampling, and cost-sensitive learning.

They especially highlighted that standard classifiers are prone to majority class bias, which results in poor recall on the minority (accident) class. This observation was important for our model construction, which resulted in the utilization of SMOTE (Synthetic Minority Oversampling Technique) and weighted classifiers to facilitate balanced learning. Their evaluation metrics, such as precision, recall, and F1-score, informed our model performance measurement. [12]

III. METHODOLOGY

This section delineates the exhaustive methodology employed for developing the traffic accident prediction model, encompassing all phases from data preparation through model evaluation. The methodology is organized into three main phases:

Data Preprocessing: The initial phase consists of meticulous data cleaning and preparation procedures aimed at establishing raw data quality and consistency for subsequent analytical processes.

Feature Engineering: The selection and transformation of relevant features occurs to boost model predictive capabilities. This research indicates to investigate various models to determine the most effective method for accurately predicting accidents.

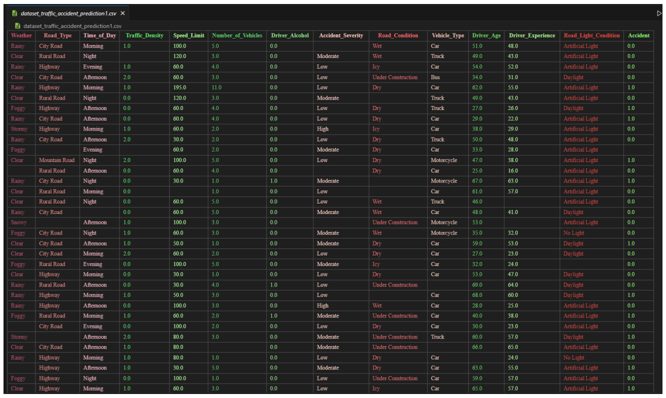
A. Data Collection and Preprocessing

This section describes the project initiation which includes data collection from different reliable sources followed by preprocessing to achieve model-readiness with accurate and consistent results.

1) Dataset Description: For this research, a traffic accident dataset is selected from kaggle (`dataset_traffic_accident_prediction1.csv`) is used. Dataset includes a variety of features related to road conditions, environmental factors, driver attributes and vehicle characteristics. The dataset's target variable is `Accident` which is binary (0 for no accident and 1 for accident) hence this a binary classification problem.

The dataset contains the following feature categories:

- **Environmental Factors:** Weather, Time of Day, Road Light Condition
- **Road Characteristics:** Road Type, Road Condition, Speed Limit
- **Driver Attributes:** Driver Age, Driver Experience, Driver Alcohol Consumption
- **Traffic Conditions:** Traffic Density, Number of Vehicles
- **Vehicle Information:** Vehicle Type
- **Outcome Variables:** Accident (target), Accident Severity



dataset_traffic_accident_prediction1.csv

dataset_traffic_accident_prediction1.csv

Weather	Road_Type	Time_of_Day	Traffic_Density	Speed_Limit	Number_of_Vehicles	Driver_Alcohol	Accident_Severity	Road_Condition	Vehicle_Type	Driver_Age	Driver_Experience	Road_Light_Condition	Accident
Rainy	City Road	Morning	1.0	100.0	5.0	0.0		Wet	Car	51.0	48.0	Artificial Light	0.0
Clear	Rural Road	Night		120.0	3.0	0.0	Moderate	Wet	Truck	49.0	43.0	Artificial Light	0.0
Rainy	Highway	Evening	1.0	60.0	4.0	0.0	Low	Icy	Car	54.0	52.0	Artificial Light	0.0
Clear	City Road	Afternoon	2.0	60.0	3.0	0.0	Low	Under Construction	Bus	34.0	31.0	Daylight	0.0
Rainy	Highway	Morning	1.0	195.0	11.0	0.0	Low	Dry	Car	62.0	55.0	Artificial Light	1.0
Clear	Rural Road	Night	0.0	120.0	3.0	0.0	Moderate		Truck	49.0	43.0	Artificial Light	0.0
Foggy	Highway	Afternoon	0.0	60.0	4.0	0.0	Low	Dry	Truck	27.0	26.0	Daylight	1.0
Rainy	City Road	Afternoon	0.0	60.0	4.0	0.0	Low	Dry	Car	29.0	22.0	Artificial Light	1.0
Stormy	Highway	Morning	1.0	60.0	2.0	0.0	High	Icy	Car	38.0	29.0	Artificial Light	0.0
Rainy	City Road	Afternoon	2.0	30.0	2.0	0.0	Low	Dry	Truck	50.0	48.0	Artificial Light	0.0
Foggy		Evening		60.0	2.0	0.0	Moderate	Dry	Car	33.0	28.0	Artificial Light	
Clear	Mountain Road	Night	2.0	100.0	5.0	0.0	Low	Dry	Motorcycle	47.0	38.0	Artificial Light	1.0
	Rural Road	Afternoon	0.0	60.0	4.0	0.0		Dry	Car	25.0	16.0	Artificial Light	0.0
Rainy	City Road	Night	0.0	30.0	1.0	1.0	Moderate		Motorcycle	67.0	63.0	Artificial Light	1.0
Clear	Rural Road	Morning	0.0		1.0	0.0	Low		Car	61.0	57.0	Artificial Light	0.0
Clear	Rural Road	Night	0.0	60.0	5.0	0.0	Low	Wet	Truck	46.0		Artificial Light	0.0
Rainy	City Road		0.0	60.0	5.0	0.0	Moderate	Wet	Car	48.0	41.0	Daylight	0.0
Snowy		Afternoon	1.0	100.0	3.0	0.0		Under Construction	Motorcycle	53.0		Artificial Light	0.0
Foggy	City Road	Night	1.0	60.0	3.0	0.0	Moderate	Wet	Motorcycle	35.0	32.0	No Light	0.0
Clear	City Road	Afternoon	1.0	50.0	1.0	0.0	Moderate	Dry	Car	59.0	53.0	Daylight	1.0
Clear	City Road	Morning	2.0	60.0	2.0	0.0	Low	Dry	Car	27.0	23.0	Daylight	0.0
Foggy	Rural Road	Evening	0.0	100.0	5.0	0.0	Moderate	Icy	Car	32.0	24.0		0.0
Clear	Highway	Morning	0.0	30.0	1.0	0.0	Low	Dry	Car	53.0	47.0	Daylight	0.0
Rainy	Rural Road	Afternoon	0.0	30.0	4.0	1.0	Low	Under Construction		69.0	64.0	Daylight	0.0
Rainy	Highway	Morning	1.0	50.0	3.0	0.0	Low		Car	68.0	60.0	Daylight	1.0
Rainy	Highway	Afternoon	0.0	100.0	3.0	0.0	High	Wet	Car	28.0	25.0	Artificial Light	0.0
Foggy	Rural Road	Morning	1.0	60.0	2.0	1.0	Moderate	Under Construction	Car	40.0	38.0	Artificial Light	1.0
	City Road	Evening	0.0	100.0	2.0	0.0	Low	Dry	Car	30.0	23.0	Artificial Light	0.0
Stormy		Afternoon	2.0	80.0	3.0	0.0	Moderate	Under Construction	Truck	60.0	57.0	Daylight	1.0
Clear	City Road	Afternoon	1.0	80.0		0.0	Moderate	Under Construction		66.0	65.0	Artificial Light	0.0
Rainy	Highway	Morning	1.0	80.0	1.0	0.0	Low	Dry	Car		24.0	No Light	0.0
	Highway	Afternoon	1.0	30.0	5.0	0.0	Moderate	Dry	Car	63.0	55.0	Artificial Light	1.0
Foggy	Highway	Night	0.0	100.0	1.0	0.0	Low	Under Construction	Car	59.0	57.0	Artificial Light	0.0
Clear	Highway	Morning	1.0	60.0	3.0	0.0	Low	Icy	Car	65.0	57.0	Artificial Light	1.0

Fig. 1. Raw Traffic Accident Dataset Before Preprocessing

2) Data Preprocessing: The machine learning pipeline depends on data preprocessing as its key step when working with classification tasks such as accident prediction. The processing of raw data ensures the correction of statistical errors in addition to value inconsistencies and data format issues to deliver features that ready for modeling. The algorithm needs proper preprocessing because it allows pattern discovery without being influenced by data irregularities or measurement disparities.

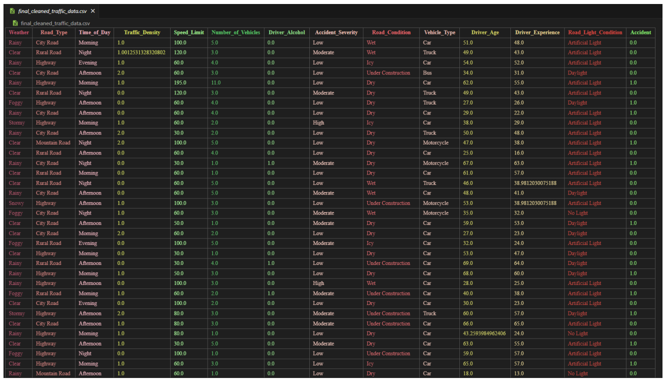
The first stage of preprocessing consisted of evaluating missing data elements in every feature. In processing categorical data we filled in absent values of Weather, Road Type and Vehicle Type with the most commonly occurring category according to mode imputation. The procedure maintained the original distribution patterns of categorical data. When dealing with numerical variables the analysis incorporated distribution-based imputation techniques. The skews of each feature were assessed before deciding to apply mean imputation for normal variables and median imputation for variables with skew distributions. This approach yielded a methodology which protected against outliers but kept intact the statistical elements of the dataset [1].

Special care was given to handle the target variable known as Accident. The model excluded target variables with missing values because predictive modeling required their actual values while avoidance of bias and disrupted prediction accuracy became concerns [2].

The evaluation of imputation method effectiveness took place during distribution analysis. The distribution patterns of numerical features were evaluated through KDE kernels along with Q-Q plots found in their histograms. The visualizations displayed feature distributions to help scientists determine normal patterns and skewed distributions which guided their selection of imputation techniques [3]. The feature distribution analysis consists of Fig 3 which shows the graphical representation of both histograms and Q-Q plots.

The researcher performed data verification of types prior to statistical analysis that produced numerical feature statistics as mean values together with standard deviations and interquartile range statistics. The feature engineering process obtained input for potential outlier detection from this analysis according to [4].

The standardized CSV file Fig 2 served as the basis for model development after the cleaning process had been completed. A strong preprocessing system provided an effective method to let the machine learning model understand traffic data through consistent data handling.



final_cleaned_traffic_data.csv

final_cleaned_traffic_data.csv

Weather	Road_Type	Time_of_Day	Traffic_Density	Speed_Limit	Number_of_Vehicles	Driver_Alcohol	Accident_Severity	Road_Condition	Vehicle_Type	Driver_Age	Driver_Experience	Road_Light_Condition	Accident
Rainy	City Road	Morning	1.0	100.0	5.0	0.0	Low	Wet	Car	51.0	48.0	Artificial Light	0.0
Clear	Rural Road	Night	1.0012531328320802	120.0	3.0	0.0	Moderate	Wet	Truck	49.0	43.0	Artificial Light	0.0
Rainy	Highway	Evening	1.0	60.0	4.0	0.0	Low	Icy	Car	54.0	52.0	Artificial Light	0.0
Clear	City Road	Afternoon	2.0	60.0	3.0	0.0	Low	Under Construction	Bus	34.0	31.0	Daylight	0.0
Rainy	Highway	Morning	1.0	195.0	11.0	0.0	Low	Dry	Car	62.0	55.0	Artificial Light	1.0
Clear	Rural Road	Night	0.0	120.0	3.0	0.0	Moderate	Dry	Truck	49.0	43.0	Artificial Light	0.0
Foggy	Highway	Afternoon	0.0	60.0	4.0	0.0	Low	Dry	Truck	27.0	26.0	Daylight	1.0
Rainy	City Road	Afternoon	0.0	60.0	4.0	0.0	Low	Dry	Car	29.0	22.0	Artificial Light	1.0
Stormy	Highway	Morning	1.0	60.0	2.0	0.0	High	Icy	Car	38.0	29.0	Artificial Light	0.0
Rainy	City Road	Afternoon	2.0	30.0	2.0	0.0	Low	Dry	Truck	50.0	48.0	Artificial Light	0.0
Clear	Mountain Road	Night	2.0	100.0	5.0	0.0	Low	Dry	Motorcycle	47.0	38.0	Artificial Light	1.0
Clear	Rural Road	Afternoon	0.0	60.0	4.0	0.0	Low	Dry	Car	25.0	16.0	Artificial Light	0.0
Rainy	City Road	Night	0.0	30.0	1.0	1.0	Moderate	Dry	Motorcycle	67.0	63.0	Artificial Light	1.0
Clear	Rural Road	Morning	0.0	60.0	1.0	0.0	Low	Dry	Car	61.0	57.0	Artificial Light	0.0
Clear	Rural Road	Night	0.0	60.0	5.0	0.0	Low	Wet	Truck	46.0	38.9812030075188	Artificial Light	0.0
Rainy	City Road	Afternoon	0.0	60.0	5.0	0.0	Moderate	Wet	Car	48.0	41.0	Daylight	0.0
Snowy	Highway	Afternoon	1.0	100.0	3.0	0.0	Low	Under Construction	Motorcycle	53.0	38.9812030075188	Artificial Light	0.0
Foggy	City Road	Night	1.0	60.0	3.0	0.0	Moderate	Wet	Motorcycle	35.0	32.0	No Light	0.0
Clear	City Road	Afternoon	1.0	50.0	1.0	0.0	Moderate	Dry	Car	59.0	53.0	Daylight	1.0
Clear	City Road	Morning	2.0	60.0	2.0	0.0	Low	Dry	Car	27.0	23.0	Daylight	0.0
Foggy	Rural Road	Evening	0.0	100.0	5.0	0.0	Moderate	Icy	Car	32.0	24.0	Artificial Light	0.0
Clear	Highway	Morning	0.0	30.0	1.0	0.0	Low	Dry	Car	53.0	47.0	Daylight	0.0
Rainy	Rural Road	Afternoon	0.0	30.0	4.0	1.0	Low	Under Construction	Car	69.0	64.0	Daylight	0.0
Rainy	Highway	Morning	1.0	50.0	3.0	0.0	Low	Dry	Car	68.0	60.0	Daylight	1.0
Rainy	Highway	Afternoon	0.0	100.0	3.0	0.0	High	Wet	Car	28.0	25.0	Artificial Light	0.0
Foggy	Rural Road	Morning	1.0	60.0	2.0	1.0	Moderate	Under Construction	Car	40.0	38.0	Artificial Light	1.0
Clear	City Road	Evening	0.0	100.0	2.0	0.0	Low	Dry	Car	30.0	23.0	Artificial Light	0.0
Stormy	Highway	Afternoon	2.0	80.0	3.0	0.0	Moderate	Under Construction	Truck	60.0	57.0	Daylight	1.0
Clear	City Road	Afternoon	1.0	80.0	3.0	0.0	Moderate	Under Construction	Car	66.0	65.0	Artificial Light	0.0
Rainy	Highway	Morning	1.0	80.0	1.0	0.0	Low	Dry	Car	43.2593984962406	24.0	No Light	0.0
Clear	Highway	Afternoon	1.0	30.0	5.0	0.0	Moderate	Dry	Car	63.0	55.0	Artificial Light	1.0
Foggy	Highway	Night	0.0	100.0	1.0	0.0	Low	Under Construction	Car	59.0	57.0	Artificial Light	0.0
Clear	Highway	Morning	1.0	60.0	3.0	0.0	Low	Icy	Car	65.0	57.0	Artificial Light	1.0
Rainy	Mountain Road	Afternoon	1.0	60.0	1.0	0.0	Low	Dry	Car	18.0	13.0	No Light	0.0

Fig. 2. overview of Cleaned Traffic Accident Dataset

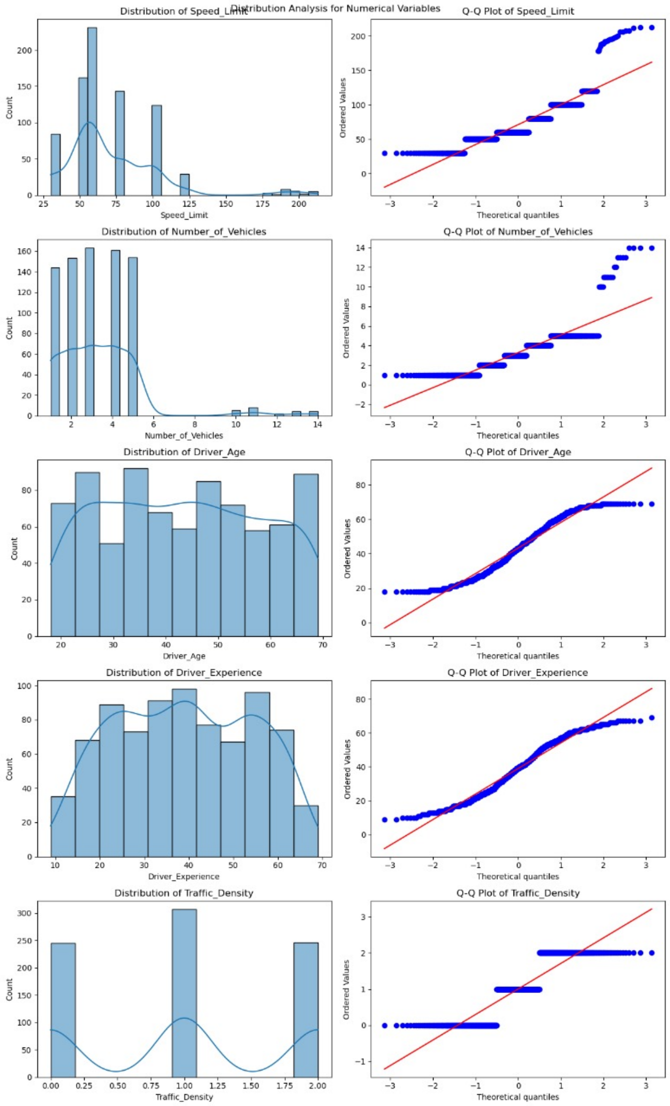


Fig. 3. Distribution analysis for numerical features using histograms and Q-Q plots.

B. Feature Engineering

Feature engineering is a crucial phase in the development of machine learning models, enabling the transformation of raw inputs into structured, algorithm-ready formats while embedding domain expertise into the modeling pipeline [5]. For traffic accident prediction, it significantly enhances model performance by capturing complex, nonlinear interactions among risk determinants [6].

This study's feature engineering approach was guided by three theoretical considerations:

- **Feature Representation Theory:** Variables were encoded in ways that preserved their intrinsic structure [7]. Nominal features were encoded to prevent artificial ordering, whereas ordinal variables retained their natural sequence.
- **Curse of Dimensionality Mitigation:** To combat overfitting and computational inefficiency, careful feature selection and derivation ensured a parsimonious yet expressive representation [8].
- **Domain-Knowledge Incorporation:** Based on traffic safety literature, derived features were engineered to model synergistic effects, where combinations of factors significantly increase accident risk beyond their individual contributions [15].

1) Feature Categorization: Features were grouped as follows:

- **Nominal:** Weather, Road Type, Road Condition, Vehicle Type, Road Light Condition
- **Ordinal:** Time of Day (Morning, Afternoon, Evening, Night)
- **Numerical:** Traffic Density, Speed Limit, Number of Vehicles, Driver Alcohol, Driver Age

2) Transformation and Encoding: Transformations were applied using a `ColumnTransformer` pipeline in *scikit-learn*:

- **Standardization:** Numerical features were scaled to zero mean and unit variance.
- **Ordinal Encoding:** Time of Day was encoded to reflect its temporal ordering.
- **One-Hot Encoding:** Nominal variables were encoded with `drop=′first′` to avoid multicollinearity.

3) Domain-Informed Feature Derivation: Domain knowledge inspired the creation of new interaction features:

- **Alcohol Risk Amplification:** Driver Alcohol values were scaled to reflect their disproportionate impact on accident risk [10].
- **Multicollinearity Management:** Due to high correlation between Driver Age and Driver Experience, the latter was excluded, following best practices in traffic data modeling [11].

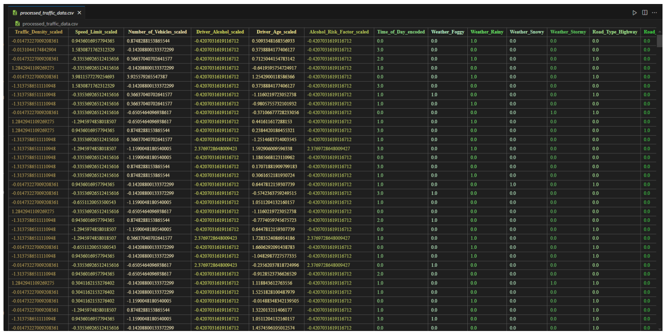
4) Data Partitioning: A stratified 75-25 train-test split preserved class proportions in both sets, addressing the common issue of class imbalance in accident datasets [12].

In total, the final feature set included 6 standardized numerical features, 1 ordinally encoded feature, and 14 one-hot encoded features, resulting in 21 model-ready predictors mentioned in Fig4.

C. Model Training and Evaluation

In the context of accident prediction, selecting an appropriate machine learning model necessitates the evaluation of diverse algorithms to ensure robust and generalizable performance [14]. Transportation safety modeling is inherently complex due to challenges such as class imbalance, heterogeneity in features, and non-linear relationships among predictors [15]. Accordingly, our model selection and evaluation process was guided by three foundational theoretical principles:

1) **Algorithm Diversity Principle:** We evaluated a spectrum of algorithms across linear models, non-parametric techniques, kernel-based approaches, and ensemble



processed_traffic_data.csv

Traffic_Density_scaled	Speed_Limit_scaled	Number_of_Vehicles_scaled	Driver_Alcohol_scaled	Driver_Age_scaled	Alcohol_Risk_Factor_scaled	Time_of_Day_encoded	Weather_Foggy	Weather_Rainy	Weather_Snowy	Weather_Stormy	Road_Type_Highway	Road_
-0.01473227009208361	0.9436016957794365	0.8748288153865544	-0.4207031619116712	0.5093348168356933	-0.4207031619116712	0.0	0.0	1.0	0.0	0.0	0.0	0.0
-0.0131044174842904	1.5830871762312329	-0.14208800133372299	-0.4207031619116712	0.3738884177406127	-0.4207031619116712	3.0	0.0	0.0	0.0	0.0	0.0	0.0
-0.01473227009208361	-0.33536926512415616	0.36637040702641577	-0.4207031619116712	0.7125044154783142	-0.4207031619116712	2.0	0.0	1.0	0.0	0.0	1.0	0.0
1.2842941109269275	-0.33536926512415616	-0.14208800133372299	-0.4207031619116712	-0.6419595754724917	-0.4207031619116712	1.0	0.0	0.0	0.0	0.0	0.0	0.0
-0.01473227009208361	3.9811577279254693	3.925579265547387	-0.4207031619116712	1.2542900118586366	-0.4207031619116712	0.0	0.0	1.0	0.0	0.0	1.0	0.0
-1.3137586511110948	1.5830871762312329	-0.14208800133372299	-0.4207031619116712	0.3738884177406127	-0.4207031619116712	3.0	0.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-0.33536926512415616	0.36637040702641577	-0.4207031619116712	-1.1160219723052738	-0.4207031619116712	1.0	1.0	0.0	0.0	0.0	1.0	0.0
-1.3137586511110948	-0.33536926512415616	0.36637040702641577	-0.4207031619116712	-0.9805755732101932	-0.4207031619116712	1.0	0.0	1.0	0.0	0.0	0.0	0.0
-0.01473227009208361	-0.33536926512415616	-0.6505464096938617	-0.4207031619116712	-0.37106677728233056	-0.4207031619116712	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1.2842941109269275	-1.2945974858018507	-0.6505464096938617	-0.4207031619116712	0.441611617288153	-0.4207031619116712	1.0	0.0	1.0	0.0	0.0	0.0	0.0
1.2842941109269275	0.9436016957794365	0.8748288153865544	-0.4207031619116712	0.2384420186455321	-0.4207031619116712	3.0	0.0	0.0	0.0	0.0	0.0	1.0
-1.3137586511110948	-0.33536926512415616	0.36637040702641577	-0.4207031619116712	-1.2514683714003545	-0.4207031619116712	1.0	0.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-1.2945974858018507	-1.1590048180540005	2.3769728648009423	1.592906009596338	2.3769728648009427	3.0	0.0	1.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-0.33536926512415616	-1.1590048180540005	-0.4207031619116712	1.1865668123110962	-0.4207031619116712	0.0	0.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-0.33536926512415616	0.8748288153865544	-0.4207031619116712	0.17071881909799183	-0.4207031619116712	3.0	0.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-0.33536926512415616	0.8748288153865544	-0.4207031619116712	0.3061652181930724	-0.4207031619116712	1.0	0.0	1.0	0.0	0.0	0.0	0.0
-0.01473227009208361	0.9436016957794365	-0.14208800133372299	-0.4207031619116712	0.6447812159307739	-0.4207031619116712	1.0	0.0	0.0	1.0	0.0	1.0	0.0
-0.01473227009208361	-0.33536926512415616	-0.14208800133372299	-0.4207031619116712	-0.5742363759249515	-0.4207031619116712	3.0	1.0	0.0	0.0	0.0	0.0	0.0
-0.01473227009208361	-0.6551120053500543	-1.1590048180540005	-0.4207031619116712	1.0511204132160157	-0.4207031619116712	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1.2842941109269275	-0.33536926512415616	-0.6505464096938617	-0.4207031619116712	-1.1160219723052738	-0.4207031619116712	0.0	0.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	0.9436016957794365	0.8748288153865544	-0.4207031619116712	-0.7774059745675723	-0.4207031619116712	2.0	1.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	-1.2945974858018507	-1.1590048180540005	-0.4207031619116712	0.6447812159307739	-0.4207031619116712	0.0	0.0	0.0	0.0	0.0	1.0	0.0
-1.3137586511110948	-1.2945974858018507	0.36637040702641577	2.3769728648009423	1.7283524086914186	2.3769728648009427	1.0	0.0	1.0	0.0	0.0	0.0	0.0
-0.01473227009208361	-0.6551120053500543	-0.14208800133372299	-0.4207031619116712	1.6606292091438783	-0.4207031619116712	0.0	0.0	1.0	0.0	0.0	1.0	0.0
-1.3137586511110948	0.9436016957794365	-0.14208800133372299	-0.4207031619116712	-1.0482987727577335	-0.4207031619116712	1.0	0.0	1.0	0.0	0.0	1.0	0.0
-0.01473227009208361	-0.33536926512415616	-0.6505464096938617	2.3769728648009423	-0.23562037818724996	2.3769728648009427	0.0	1.0	0.0	0.0	0.0	0.0	0.0
-1.3137586511110948	0.9436016957794365	-0.6505464096938617	-0.4207031619116712	-0.9128523736626529	-0.4207031619116712	2.0	0.0	0.0	0.0	0.0	0.0	0.0
1.2842941109269275	0.3041162153276402	-0.14208800133372299	-0.4207031619116712	1.118843612763556	-0.4207031619116712	1.0	0.0	0.0	0.0	1.0	1.0	0.0
-0.01473227009208361	0.3041162153276402	-0.14208800133372299	-0.4207031619116712	1.5251828100487979	-0.4207031619116712	1.0	0.0	0.0	0.0	0.0	0.0	0.0
-0.01473227009208361	0.3041162153276402	-1.1590048180540005	-0.4207031619116712	-0.01488348342139505	-0.4207031619116712	0.0	0.0	1.0	0.0	0.0	1.0	0.0
-0.01473227009208361	-1.2945974858018507	0.8748288153865544	-0.4207031619116712	1.322013211406177	-0.4207031619116712	1.0	0.0	0.0	0.0	0.0	1.0	0.0
-1.3137586511110948	0.9436016957794365	-1.1590048180540005	-0.4207031619116712	1.0511204132160157	-0.4207031619116712	3.0	1.0	0.0	0.0	0.0	1.0	0.0
-0.01473227009208361	-0.33536926512415616	-0.14208800133372299	-0.4207031619116712	1.4574596105012574	-0.4207031619116712	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Fig. 4. Traffic Accident Dataset after preprocessing

methods, to increase the likelihood of capturing underlying data patterns effectively [17].

2) **Performance-Interpretability Trade-off:** In safety-critical applications, model interpretability is vital. While ensemble methods often offer superior accuracy, they may reduce transparency. Hence, model selection considered both performance and interpretability [18].
3) **Comprehensive Evaluation Framework:** We employed multiple evaluation metrics beyond accuracy, including F1-score and ROC-AUC, to address potential class imbalance and better capture real-world predictive performance [19].

1) Model Selection Strategy: Eight classification algorithms were implemented, each representing a distinct methodological family:

- *Logistic Regression:* A baseline linear model with high interpretability.
- *K-Nearest Neighbors (KNN):* A distance-based non-parametric approach suitable for local pattern recognition.
- *Decision Tree:* A rule-based model inherently capable of handling categorical variables.
- *Random Forest:* A bagging ensemble technique known for reducing variance and enhancing stability.
- *AdaBoost and Gradient Boosting:* Sequential ensemble methods that emphasize difficult-to-classify samples.
- *XGBoost:* A scalable gradient boosting algorithm with built-in regularization [20].
- *Support Vector Machine (SVM):* A kernel-based method effective in high-dimensional spaces.

The Random Forest classifier was further optimized using domain knowledge with parameters such as $n\_estimators = 200$, $max\_depth = 8$, and $class\_weight =$ "balanced" to address class imbalance.

2) Evaluation Framework: A multi-metric evaluation framework was designed to assess each model's performance across several dimensions:

- **Accuracy:** Overall prediction correctness.
- **F1-Score:** The harmonic mean of precision and recall, crucial in imbalanced datasets.
- **ROC-AUC:** A threshold-independent metric evaluating class separability.
- **Training Time:** Indicator of computational efficiency.
- **Confusion Matrix and Classification Report:** Provided class-wise performance insights.

Models were trained and evaluated using a stratified train-test split to preserve class distribution:

```
train_test_split(X, y,
test_size=0.2, stratify=y,
random_state=42)
```

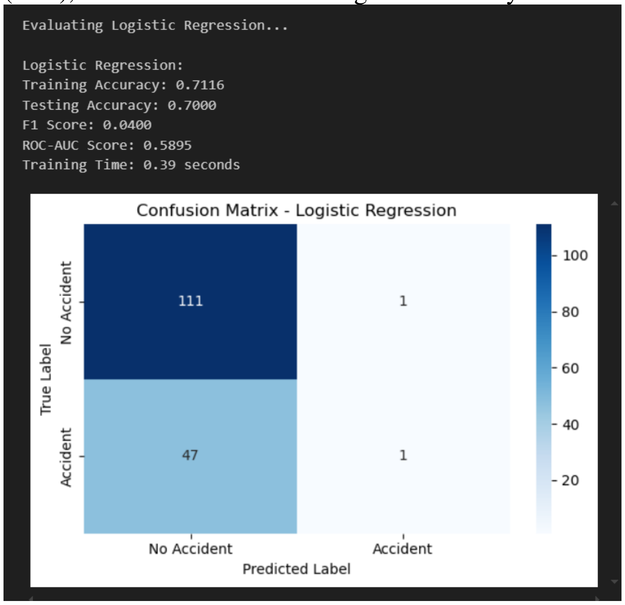
3) Feature Engineering and Interaction Terms: Feature engineering included the creation of domain-informed interaction variables that captured compounded effects such as:

- `Alcohol_Night`: Interaction of alcohol involvement and night-time driving.
- `Speed_Wet_Road`: Speed limit adjusted by adverse road conditions (wet/icy).
- `Motorcycle_Speed`: Speed limit contextualized for motorcycles.
- `Low_Visibility_Speed`: Speed under poor visibility conditions.

These features enhanced the model's capability to learn complex contextual patterns relevant to accident risk.

4) Comparative Analysis: Each model's performance was visualized through confusion matrices and tabular summaries to enable straightforward comparison. Ensemble models, particularly Random Forest, Gradient Boosting, and XGBoost, showed competitive F1-scores, while Logistic Regression emerged as the best in overall accuracy and interpretability.

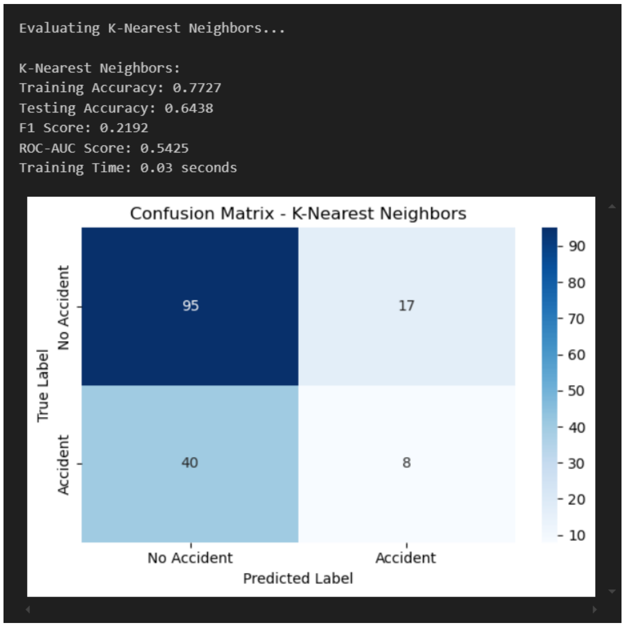
- **Logistic Regression:** Achieved 70% testing accuracy and ROC-AUC of 0.5895. While the F1 score was low (0.04), the model demonstrated generalizability.



```
Evaluating Logistic Regression...

Logistic Regression:
Training Accuracy: 0.7116
Testing Accuracy: 0.7000
F1 Score: 0.0400
ROC-AUC Score: 0.5895
Training Time: 0.39 seconds
```

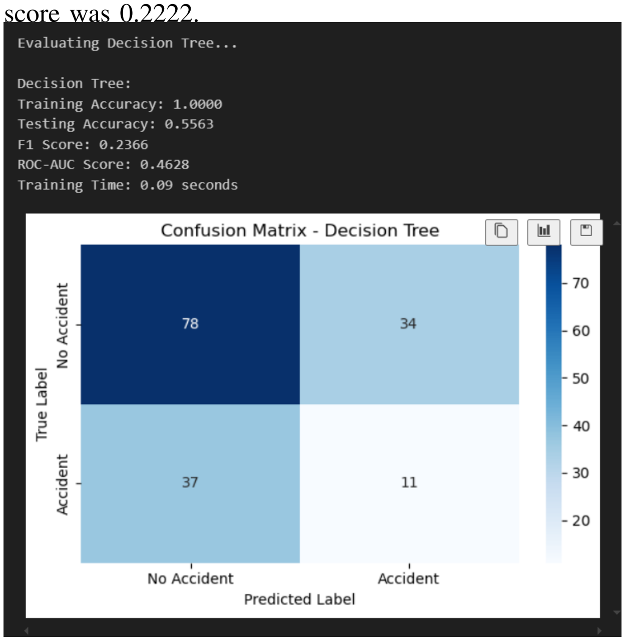
- **K-Nearest Neighbors (KNN):** Achieved 64.4% testing accuracy with a moderate F1 score of 0.2192, showing sensitivity to local decision boundaries.



```
Evaluating K-Nearest Neighbors...

K-Nearest Neighbors:
Training Accuracy: 0.7727
Testing Accuracy: 0.6438
F1 Score: 0.2192
ROC-AUC Score: 0.5425
Training Time: 0.03 seconds
```

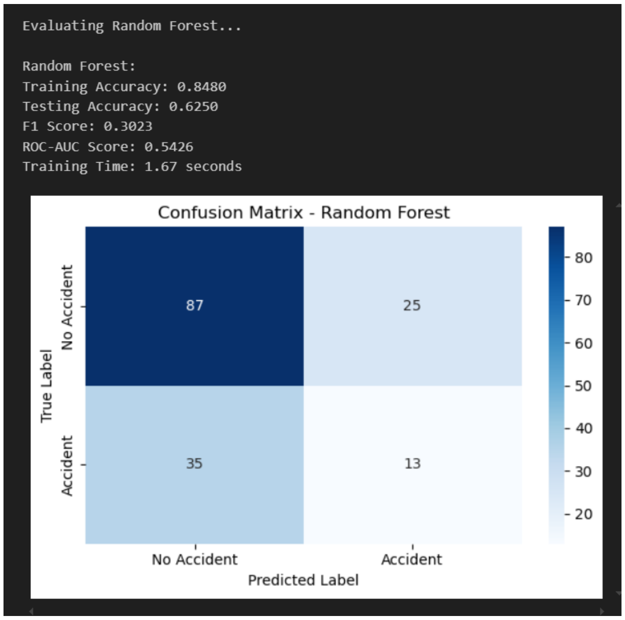
- **Decision Tree:** While it achieved perfect training accuracy, it overfit with just 56.2% test accuracy. F1 score was 0.2222.



```
Evaluating Decision Tree...

Decision Tree:
Training Accuracy: 1.0000
Testing Accuracy: 0.5563
F1 Score: 0.2366
ROC-AUC Score: 0.4628
Training Time: 0.09 seconds
```

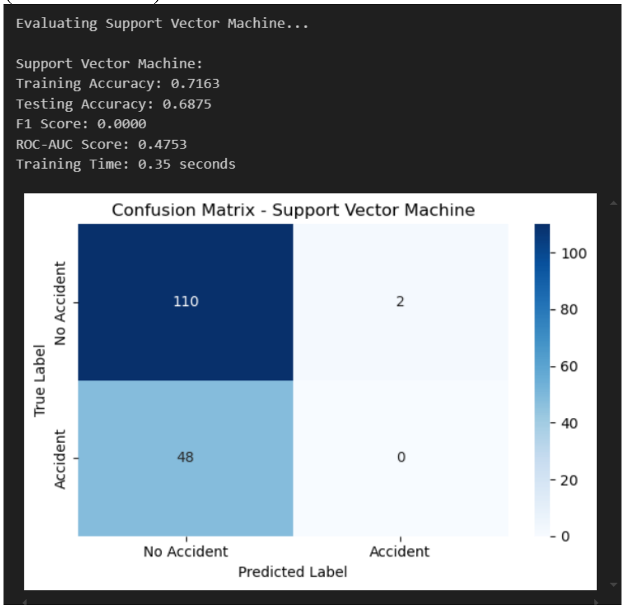
- **Random Forest:** Delivered the highest F1 score (0.3023) and a decent accuracy of 62.5%, showing its strength in handling feature interactions.



```
Evaluating Random Forest...

Random Forest:
Training Accuracy: 0.8480
Testing Accuracy: 0.6250
F1 Score: 0.3023
ROC-AUC Score: 0.5426
Training Time: 1.67 seconds
```

- **Support Vector Machine (SVM):** Balanced performance with 68.7% accuracy but failed to classify positive class (F1 score = 0).



```
Evaluating Support Vector Machine...

Support Vector Machine:
Training Accuracy: 0.7163
Testing Accuracy: 0.6875
F1 Score: 0.0000
ROC-AUC Score: 0.4753
Training Time: 0.35 seconds
```

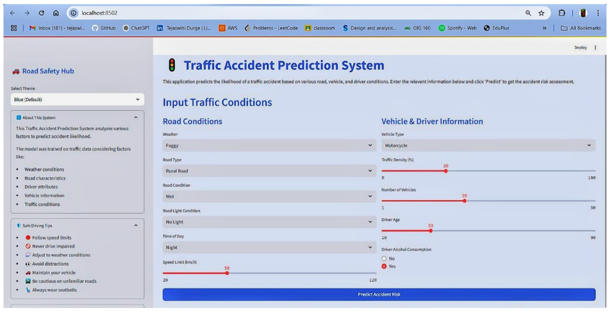
A full comparison is summarized in Table II:

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Accuracy	F1 Score	ROC-AUC	Time (s)
Logistic Regression	0.7000	0.0400	0.5895	1.02
K-Nearest Neighbors	0.6438	0.2192	0.5425	0.04
Decision Tree	0.5625	0.2222	0.4613	0.19
Random Forest	0.6250	0.3023	0.5426	2.87
AdaBoost	0.7000	0.0000	0.6303	0.71
Gradient Boosting	0.6875	0.1071	0.4851	0.82
XGBoost	0.6188	0.2078	0.4494	1.03
SVM	0.6875	0.0000	0.4753	0.43

IV. Results

To analyze the performance of the suggested Traffic Accident Prediction System, a test scenario was executed using the Streamlit-based web interface. The system was subjected to a high-risk scenario to show its ability to measure accident probability and give actionable safety recommendations. Test scenario Inputs shown in fig 5



🚗 Road Safety Hub

Select Theme

Blue (Default)

ℹ About This System

This Traffic Accident Prediction System analyzes various factors to predict accident likelihood.

The model was trained on traffic data considering factors like:

- Weather conditions
- Road characteristics
- Driver attributes
- Vehicle information
- Traffic conditions

Safe Driving Tips

- 🔴 Follow speed limits
- 🚫 Never drive impaired
- Adjust to weather conditions
- 👀 Avoid distractions
- 🚗 Maintain your vehicle
- Be cautious on unfamiliar roads
- Always wear seatbelts

🚦 Traffic Accident Prediction System

This application predicts the likelihood of a traffic accident based on various road, vehicle, and driver conditions. Enter the relevant information below and click 'Predict' to get the accident risk assessment.

Input Traffic Conditions

Road Conditions

Weather: Foggy

Road Type: Rural Road

Road Condition: Wet

Road Light Condition: No Light

Time of Day: Night

Speed Limit (km/h): 58 (20 – 120)

Vehicle & Driver Information

Vehicle Type: Motorcycle

Traffic Density (%): 30 (0 – 100)

Number of Vehicles: 20 (1 – 50)

Driver Age: 33 (18 – 90)

Driver Alcohol Consumption

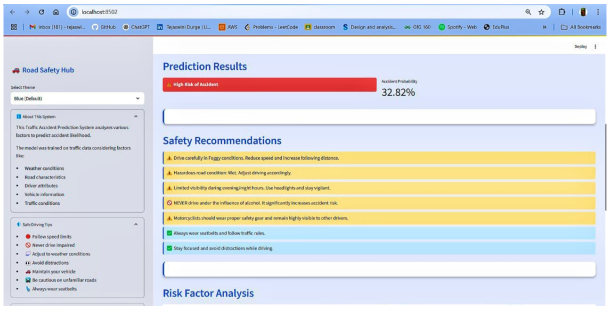
○ No

● Yes

Predict Accident Risk

Fig. 5. Input Features

According to the input feature, the model calculated an accident probability of 32.82%. Since the system's pre-defined safety risk limit is 25%, this case was identified as high risk. As a result, a visible alert message was displayed as shown in fig 6.



🚗 Road Safety Hub

Select Theme

Blue (Default)

ℹ About This System

This Traffic Accident Prediction System analyzes various factors to predict accident likelihood.

The model was trained on traffic data considering factors like:

- Weather conditions
- Road characteristics
- Driver attributes
- Vehicle information
- Traffic conditions

Safe Driving Tips

- 🔴 Follow speed limits
- 🚫 Never drive impaired
- Adjust to weather conditions
- 👀 Avoid distractions
- 🚗 Maintain your vehicle
- Be cautious on unfamiliar roads
- Always wear seatbelts

Prediction Results

⚠ High Risk of Accident

Accident Probability

32.82%

Safety Recommendations

⚠ Drive carefully in Foggy conditions. Reduce speed and increase following distance.

⚠ Hazardous road condition: Wet. Adjust driving accordingly.

⚠ Limited visibility during evening/night hours. Use headlights and stay vigilant.

🚫 NEVER drive under the influence of alcohol. It significantly increases accident risk.

⚠ Motorcyclists should wear proper safety gear and remain highly visible to other drivers.

✅ Always wear seatbelts and follow traffic rules.

✅ Stay focused and avoid distractions while driving.

Risk Factor Analysis

Fig. 6. Predicted Output

"High Risk of Accident"

Proposed system provides customized safety recommendations based on contextual information and expert-based conditional logic. The recommendations were:

- Take extreme care in conditions of fog by slowing down and increasing the distance from the vehicle in front.
- Adjust driving behavior to compensate for wet road surfaces.
- Because of night driving and poor visibility, make sure headlights are on and remain vigilant.
- Avoid driving while intoxicated, as it increases the risk of accidents.
- Motorcyclists should dress in safety gear and remain conspicuous to others.
- Always wear seatbelts and strictly follow traffic rules.
- Stay attentive and do not get distracted while driving.

These guidelines were dynamically adjusted to the input conditions to help users.

To enhance interpretability and encourage user recognition, the system provided a horizontal bar chart showing the relative contribution of each input feature to the total accident risk. The risk factor analysis was based on a custom-weighted scoring algorithm that focused on significant interaction effects, especially with alcohol consumption and poor weather. The overriding factors that controlled risk shown in Fig. 7

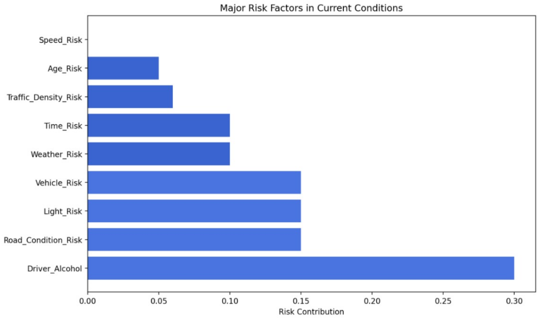


Fig. 7. Risk Factors

The system user interface was structured with three separate components in order to make interaction and comprehension easier:

- **Input Panel** – to input driver, vehicle, and environmental parameters 5
- **Prediction Panel** – showing the calculated accident probability and risk classification 6
- **Risk Breakdown Panel** – demonstrating the contribution of each factor to the forecasted risk through an interpretive visualization 7

V. Comparison with Other Methodologies:

In the context of accident detection, this section offers a comparison of modern machine learning techniques with more conventional approaches.

A. Machine Learning Techniques vs. Conventional Statistical Methods

For the modeling of crash-frequency data, traditional statistical models such as Poisson and negative binomial regressions have been used extensively. These models often grapple with over dispersion and unobserved heterogeneity in the data and impose a specific distribution [15]. Additionally, they may not be able to capture the complex nonlinear interactions present in data on traffic accidents.

Conversely, machine learning methods independent of strong distributional assumptions like Random Forests and ensemble methods offer model flexibility in representing complex patterns. These methods are adept at modeling complex relationships among factors and can cope with high-dimensional data. [17], [20].

B. Managing Unbalanced and Missing Data

Missing data is a common issue in accident data sets that, if not controlled, can lead to skewed results. Traditional methods often resort to simple imputation methods or case removal. However, sophisticated imputation methods considering the data structure of the data are employed in sophisticated machine learning methods, which generate more accurate and reliable models [1].

Also, accident data may sometimes be class imbalanced, with fewer non-severe incidents than severe ones. Traditional models might bias towards the majority class. Resampling, cost-sensitive learning, and ensemble approaches are some of the strategies used in machine learning techniques to counteract the impacts of class imbalance and improve the model's capacity to identify infrequent but crucial events. [12], [14].

C. Selection and Feature Engineering

For the model to function well, feature engineering and selection must be done well. For feature selection, traditional methods mostly rely on domain expertise, which can be laborious and may miss little patterns. To effectively find the most predictive variables, machine learning approaches use automated feature selection techniques including regularization methods and recursive feature removal [4], [7].

D. Data Preprocessing and Scaling Techniques

The performance of the model is greatly impacted by data preprocessing, such as scaling and normalization. Traditional methods may employ standard scaling methods without considering the specific requirements of different algorithms. Current research emphasizes the importance of choosing proper scaling methods tailored to the given machine learning models in order to achieve optimal performance [2].

E. Interpretability and Model Transparency

Although they are interpretable, traditional statistical models were not able to be as predictive as advanced machine learning models. Transparency is an issue, however, since some machine learning algorithms are black-box. By balancing interpretability and accuracy, advances in interpretable machine learning provide techniques and tools to explain model conclusions [18].

F. Summary of Comparative Analysis

When it comes to detecting accidents, both traditional methods and machine learning (ML) methods offer their own strengths and challenges. Traditional models cannot always process complex, high-dimensional data, although they are often easier to interpret and require less computational power. Machine learning methods, on the other hand, are well-suited to process large, unbalanced data sets and expose concealed patterns.

Table III contains a comprehensive comparison of these two paradigms in several aspects, such as interpretability, scalability, feature engineering, and data handling. This comparison study gives the basis to select appropriate methods according to the specific goals and data characteristics of an accident detection system.

TABLE III
COMPARING MACHINE LEARNING TECHNIQUES WITH CONVENTIONAL METHODS FOR ACCIDENT DETECTION

Aspect	**Traditional Methods**	**Machine Learning Approaches**
Model Assumptions	Require specific distributional assumptions	Flexible, non-parametric models
Handling Missing Data	Simple imputation or case deletion	Advanced imputation techniques
Class Imbalance	Often biased towards majority class	Incorporate resampling and cost-sensitive methods
Feature Selection	Manual, based on domain knowledge	Automated, data-driven techniques
Data Scaling	Standard techniques applied uniformly	Tailored scaling based on algorithm requirements
Interpretability	High, but limited complexity	Varies; enhanced by interpretable ML tools

VI. CONCLUSION

Forecasting road accidents isn't simply a question of numbers and code, it's about avoiding heartbreak, saving lives, and turning our daily drives safer. Through this study, we investigated how various classification models could be employed as pre-incident warning systems to identify accident-prone states early on. Although each model had its strengths to offer, some were more dependable and precise than others based on the situation. True power resides in selecting the appropriate model for the correct setting—urban mayhem or country roads.

While roads keep filling up and information keeps pouring in, the possibility of creating a smarter and more adaptive safety systems becomes not only possible but also obligatory. Our comparison isn't the last word, it's a leap ahead in a very much bigger movement toward predictive road safety. The vision is uncomplicated: a world where there are fewer sirens wailing at night, fewer families who receive life-changing phone calls, and every road trip one does is one where they arrive safe and sound. With the correct application of data and smart systems, that vision doesn't sound so far out.

REFERENCES

[1] T. Emmanuel, T. Maupong, D. Mpoeleng, B. Webala, and T. Ntsala, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–37, 2021, doi: 10.1186/s40537-021-00516-9.

[2] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *arXiv preprint arXiv:2212.12343*, 2022.

[3] C. Li, "Preprocessing methods and pipelines of data mining: An overview," *arXiv preprint arXiv:1906.08510*, 2019.

[4] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv preprint arXiv:1601.07996*, 2016.

[5] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2018.

[6] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *Journal of Safety Research*, vol. 38, no. 1, pp. 97–108, 2007.

[7] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

[8] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[9] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[10] J. C. Fell and R. B. Voas, "The effectiveness of reducing illegal blood alcohol concentration (BAC) limits for driving: Evidence for lowering the limit to .05 BAC," *Journal of Safety Research*, vol. 49, pp. 11–17, 2014.

[11] F. L. Mannering, V. Shankar, and C. R. Bhat, "Unobserved heterogeneity and the statistical analysis of highway accident data," *Analytic Methods in Accident Research*, vol. 11, pp. 1–16, 2016.

[12] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[13] R. Elvik, "Risk of road accident associated with the use of drugs: A systematic review and meta-analysis of evidence from epidemiological studies," *Accident Analysis & Prevention*, vol. 60, pp. 254–267, 2013.

[14] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018.

[15] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[16] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: Methodological frontier and future directions," *Analytic Methods in Accident Research*, vol. 1, pp. 1–22, 2014.

[17] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[18] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.

[19] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] R. Elvik, "Risk of road accident associated with the use of drugs: A systematic review and meta-analysis of evidence from epidemiological studies," *Accident Analysis & Prevention*, vol. 60, pp. 254–267, 2013.