Table of Contents

1.Introduction

The project's main goal is to forecast traffic accidents by taking into account a number of variables, including time, location, and weather. Its goal is to create a model that can predict the probability of collisions so that preventative actions can be taken to increase traffic safety. The project's goals are to minimize damage, optimize traffic management, and lower the number of accidents. It makes use of data-driven insights to assist authorities in taking preventative measures and improving road safety in general.

2.Dataset Description

- **Source**
  -Link - https://www.kaggle.com/datasets/denkuznetz/traffic-accident-prediction/data
  -Reference - *Traffic Accident Prediction* 💥🚗. (2024, December 11). Kaggle. https://www.kaggle.com/datasets/denkuznetz/traffic-accident-prediction/data

- **Dataset Description**[**Traffic Accident Prediction 💥🚙 | Kaggle**]
  - This dataset have 840 instances and 14 features
  - This is a classification problem. Because it predicts categories. Specifically for this dataset it predicts accidents will happen or not based on the features the model is being provided in 1/0 format. That's why it is considered a classification problem
  - As this dataset contains 14 features it has 14 datapoints.
  - We have 14 features in this dataset. Among them 10 are categorical and 4 are quantitative.
    - Weather indicates the weather condition. Which is a categorical feature
    - Road_Type indicates the type of road. Which is a categorical feature
    - Time_of_day indicates the time of a particular day. Which is categorical feature
    - Traffic_Density indicates the density of traffic in a particular time. Which is a categorical feature

- Speed_Limit indicates the limit of speed at a particular time of a day. Which is a quantitative feature.
- Number_Of_Vehicles indicates the number of vehicles on a particular road at a particular time of day. It is a quantitative feature.
- Driver_Alcohol indicates if a driver is drunk or not. It is a categorical feature.
- Accident_Severity indicates the severity of an accident. It is a categorical feature.
- Road_Condition indicates the condition of a road. It is a categorical feature
- Vehicle_Type indicates the type of the vehicle. It is a categorical feature
- Driver_Age indicates the age of a driver. It is a quantitative feature
- Driver_Experience indicates the experience of a Driver. It is a quantitative feature
- Road_Light indicates the type of road. It is a categorical feature.
- Accident is the target variable indicating the accident is happening or not. It is a categorical feature.

- The correlation of all features can be understood by looking at the heatmap below:

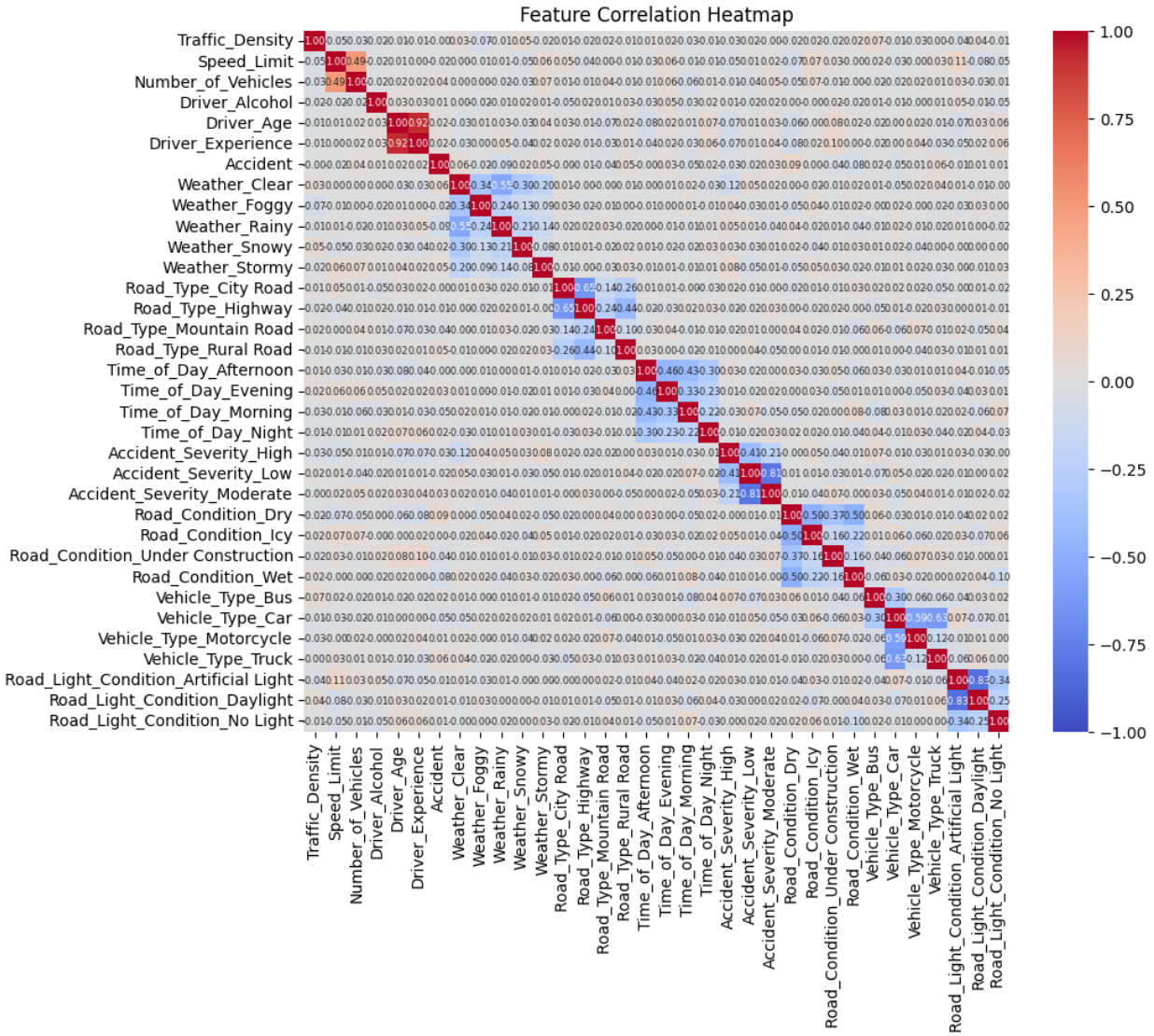


Figure 1: Heatmap of Traffic Accident Dataset

- **Imbalanced Dataset**

  - No. There are two classes in the output feature. One is 0, another one is 1. Class 0 has 575 instances and class 1 has 247 instances.

  - Number of instances in every class is shown below

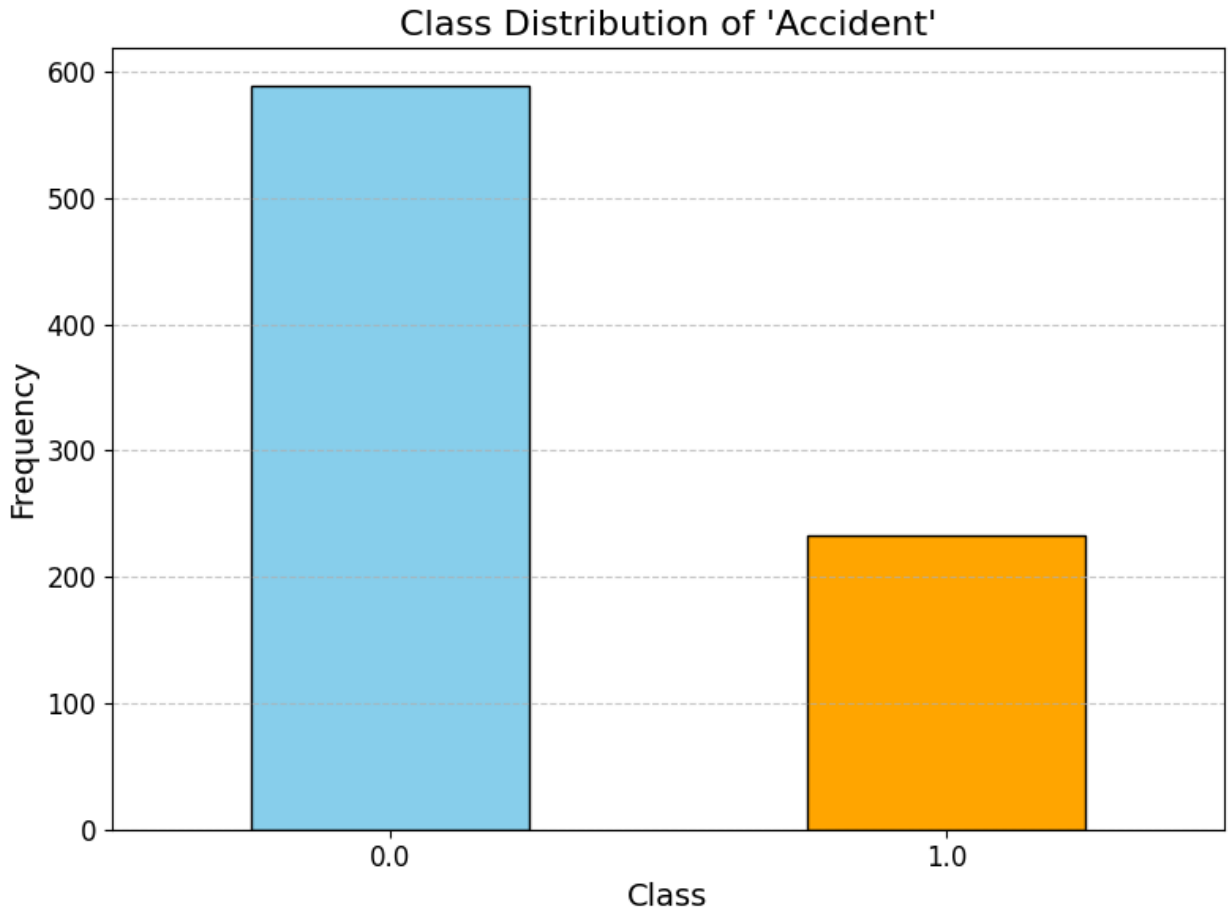


Figure 2 : Number of instances of category class

3.Dataset pre-processing

- **Faults**
  Null values- We have 42 null values in each feature
  Categorical values - We have 10 categorical feature which consists of categorical values

- **Solutions**

We have 42 null values in each of the features. If we drop each of the instances carrying null values will unnecessarily reduce the dataset size. So, we have used median imputation in those features carrying quantitative values. So, that any kind of outliers don't impact the values used to impute the null values with. Mode imputation is used in features carrying categorical values.

There are categorical value in 10 features most of the machine learning model can't work with the categorical variable we have to convert them into numeric or binary values. Except for the Driver_Alcohol each of the categorical features in this dataset have more 2 unique classes. So, we have to use One-Hot Encoding here.

4.Feature Scaling

In our dataset, we have to do feature scale in the Speed_Limit, Number_Of_Vehicles, Driver_age named feature to ensure that all numerical features in the dataset are on a comparable scale and to prevent bias in the model. And we have used Robust Scaler technique because Robust scaling uses the median and IQR instead of the mean and standard deviation, making it resistant to outliers. Outliers have little impact on the scaling process, so this method is ideal for datasets with significant outliers

5.Dataset Splitting

We have splitted the dataset into 70-30. Meaning for training we have used 70% of the data and for testing we have used 30% of the data.

6.Model Training and Testing

From the available models we have decided to use Naive Bayes, Decision Tree and Random Forest

Naive Bayes

1. Naive Bayes is computationally efficient and simple. It works well with lots of feature which we have

2. Naive Bayes assumes features are conditionally independent, in our dataset we have checked correlation and if we find 0.8 or high correlation among two feature we Simply drop one feature. So we can assume the conditional independence here
3. Naive Bayes works well with imbalanced data. As we have imbalanced data in the dataset , we can use here

Decision Tree

1. Decision Tree is highly interpretable. We can easily trace the logic behind the decision making process.
2. Decision Tree can capture nonlinear relationships between features.
3. A decision tree will inherently select the most important features for classification and split the data based on these features.

Random Forest

1. While a single decision tree is easy to interpret, it may not perform as well on its own. Random Forest improves upon decision trees by reducing variance and preventing overfitting, which typically leads to better accuracy and prediction.
2. Random Forest is robust to outliers and can handle missing data better than many other algorithms.
3. Random Forest is an ensemble method that combines multiple decision trees to make a more robust prediction

7. Model Selection/Comparison Analysis

The metrics used to evaluate the performance of the models are:

(i) Accuracy Score: This is the most common metric and represents the overall correctness of the model's predictions. It is the ratio of correctly predicted instances to the total number of instances

(ii) Precision Score: This metric focuses on the accuracy of positive predictions. It is the ratio of correctly predicted positive instances to the total predicted positive instances.

(iii) Recall Score: This metric measures the ability of the model to identify all positive instances. It is the ratio of correctly predicted positive instances to the total actual positive instances.

(iv) F1 Score: This is a harmonic mean of precision and recall, providing a balanced measure of both metrics. It is useful when both false positives and false negatives are to be considered.

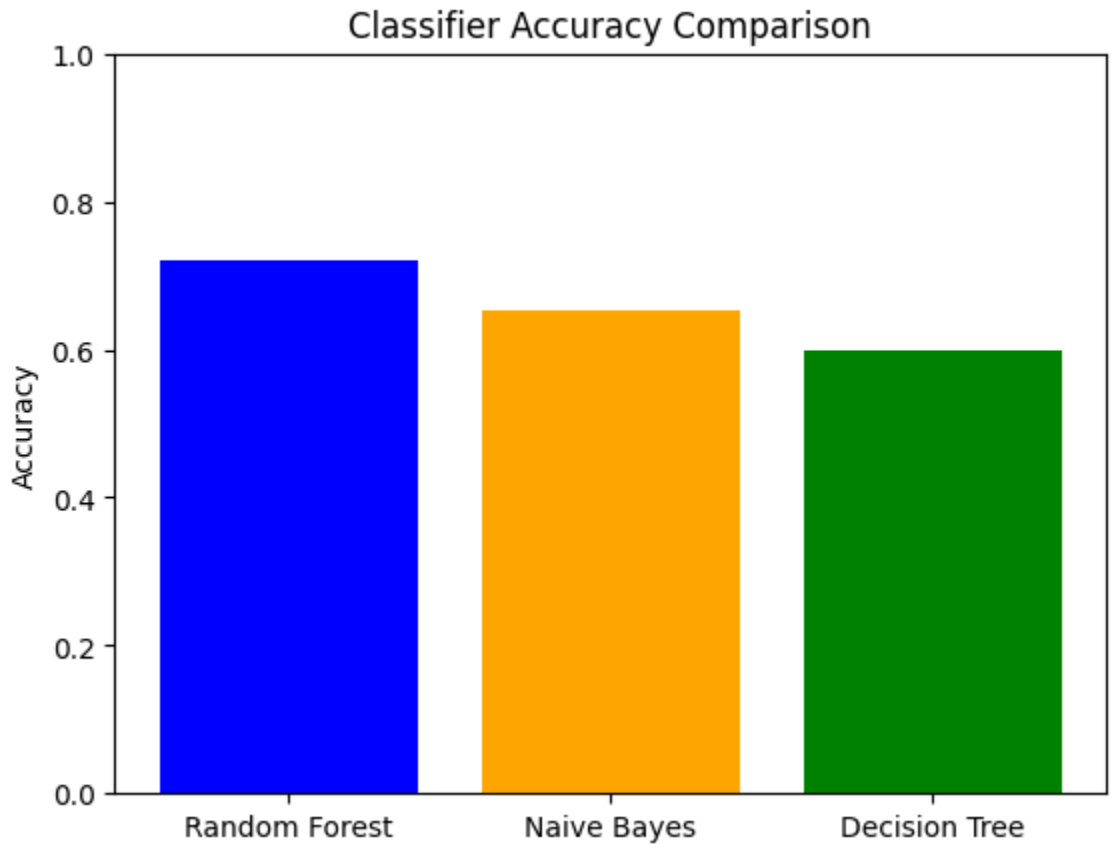


Figure 3: Accuracy

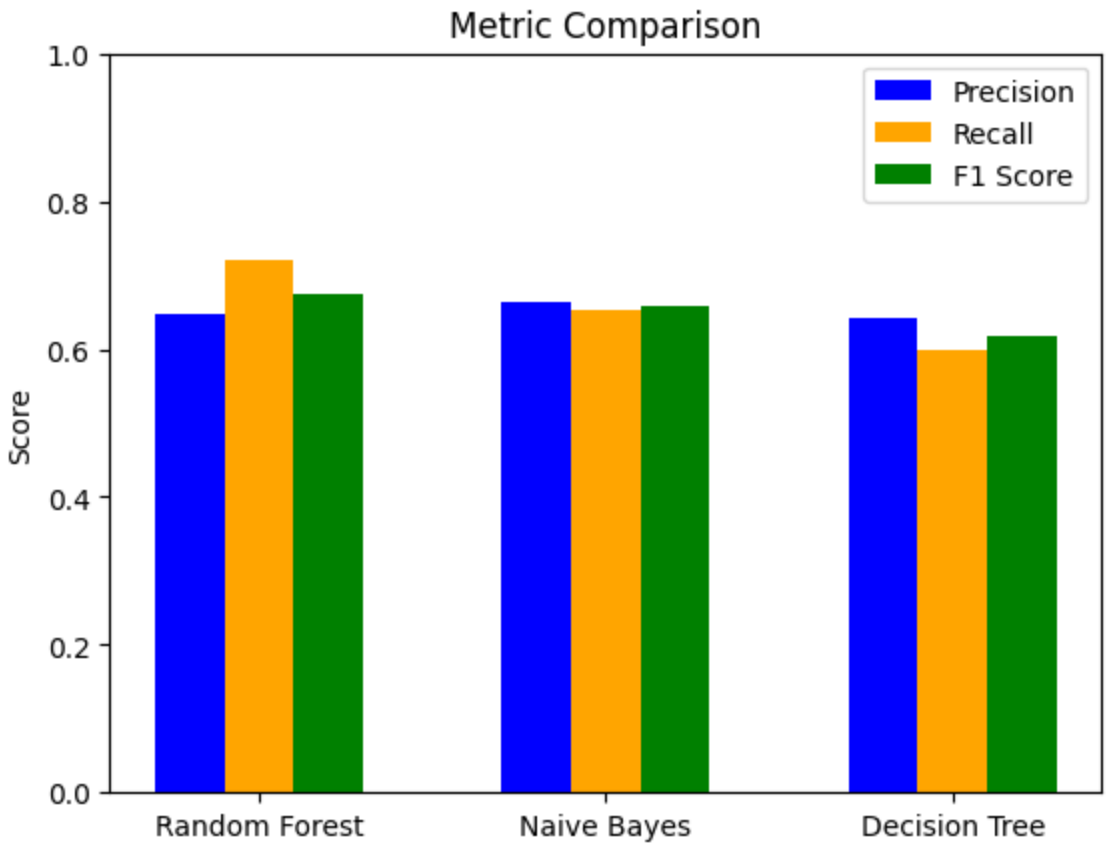
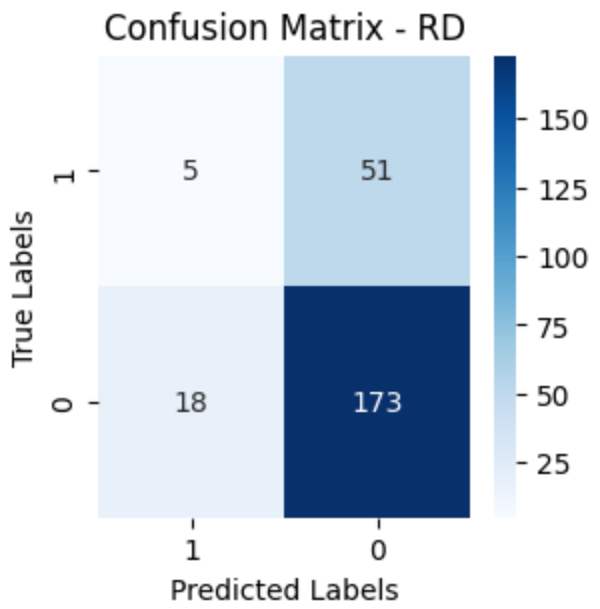
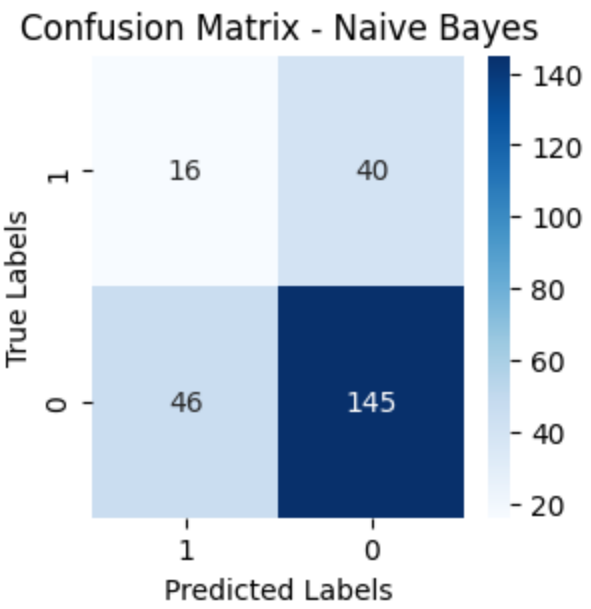
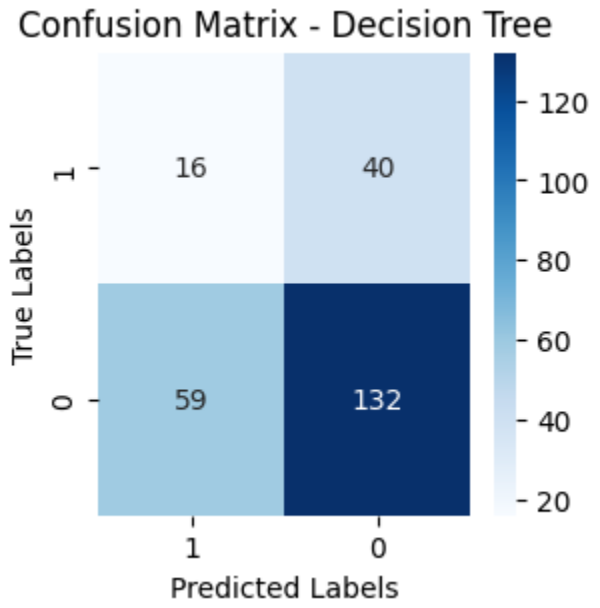


Figure 4: Recall,Precision and F1 Score







In our model, High recall is important when missing an accident (i.e., false negatives) could have serious consequences, such as when trying to predict potential accidents that need timely intervention or preventive measures. So, we will use Random Forest as it has the highest recall score 72%.

8.Conclusion

In this project, we aimed to develop a model that predicts traffic accidents based on various features such as weather conditions, time of day, traffic density, road type, and driver-related factors. The goal was to create a system that can forecast the likelihood of accidents, enabling authorities to take proactive measures to reduce traffic-related injuries and fatalities.

After preprocessing the dataset and addressing issues like missing values, outliers, and categorical variables, we proceeded to train and test three different machine learning models: Naive Bayes, Decision Tree, and Random Forest. Each model was evaluated based on key metrics including accuracy, precision, recall, and F1 score.

Our analysis showed that Random Forest outperformed the other models, particularly in terms of recall. Recall is a crucial metric in this context, as false negatives (failing to predict an accident) could lead to missed opportunities for preventive actions, thus increasing the risk of accidents and their severity. Since Random Forest achieved a recall score of 72%, it strikes a good balance between accuracy and the ability to identify potential accidents. This makes it the most suitable model for this problem, where minimizing the chances of missing an accident is critical for safety and timely interventions.

Overall, Random Forest stands out due to its robustness, ability to handle outliers, and higher recall, making it the best choice for predicting traffic accidents in this dataset. By leveraging this model, traffic management systems, emergency services, and public safety organizations can potentially reduce the occurrence and severity of accidents by enabling timely interventions.

In conclusion, our model demonstrates the potential of machine learning techniques in improving road safety and managing traffic accidents, emphasizing the importance of accurate predictions in high-stakes scenarios. Future work could focus on further fine-tuning the model, incorporating additional features, and exploring real-time prediction systems to enhance the effectiveness of traffic safety measures.