**Phase-2**

**Student Name:** Aswaya Sajeev CK

**Register Number:** 410723106004

**Institution:** Dhanalakshmi college of engineering

**Department:** Electronics and Communication Engineering

**Date of Submission:** 07-05-2025

**Github Repository Link:**

https://github.com/Aswayasajeev/Nm\_Aswayasajeev\_CK.git

**Enhancing Road Safety with AI-Driven Traffic**

**Accident Analysis and Prediction**

# Problem Statement

* Road traffic accidents are a major public safety concern, causing significant fatalities, injuries, and economic loss globally. Understanding the factors contributing to these accidents and predicting their occurrence can aid authorities in making informed decisions and implementing preventive measures.

* This project aims to build an AI-driven system that analyzes historical traffic accident data and predicts the likelihood of future accidents based on environmental, temporal, and geographic features.

**Problem Type:**

* This is primarily a classification problem, where the goal is to predict whether a road accident is likely to occur (binary or multi-class) under certain conditions.

**Why It Matters**:

* Accurate accident prediction models can help urban planners, traffic authorities, and emergency services reduce road accidents, enhance real-time traffic management, and optimize resource allocation—contributing significantly to public safety.

# Project Objectives

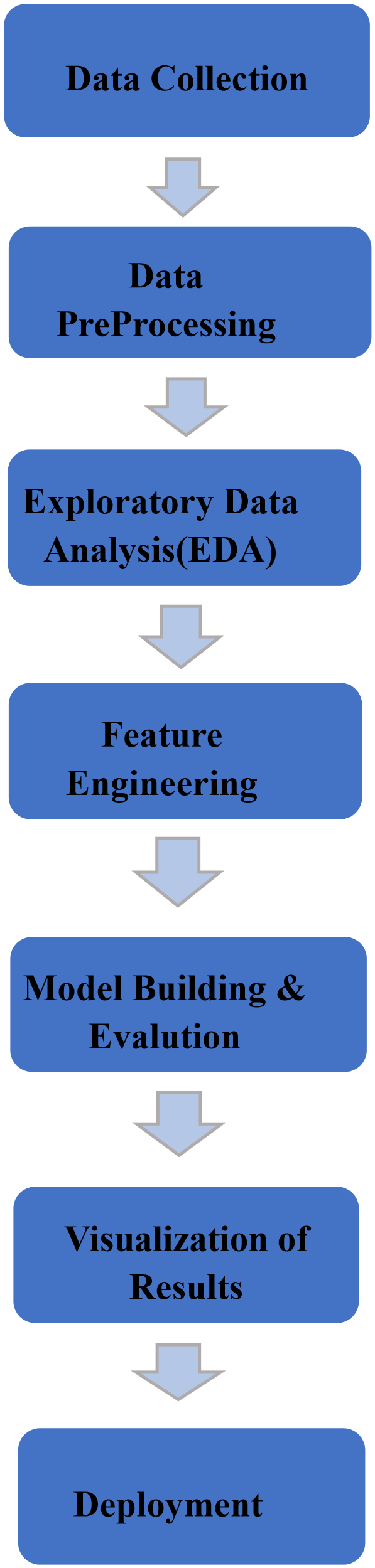
**Technical Objectives**:

* Clean and preprocess real-world traffic accident data.
* Perform exploratory data analysis (EDA) to understand accident trends.
* Engineer relevant features from raw data (e.g., time of day, weather conditions).
* Train and evaluate at least two classification models to predict accident risk.
* Interpret model results and highlight key contributing factors to road accidents.

**Model Goals:**

* Achieve high accuracy and recall in accident prediction.
* Ensure model interpretability to inform stakeholders.
* Provide actionable insights for road safety policy and implementation.

# Flowchart of the Project Workflow



# Data Description

* **Dataset Name**: Road Traffic Accident Data
* **Source**: [e.g., UK Government Open Data, Kaggle, or Local Transport Authority API]
* **Type of Data**: Structured (CSV)
* **Data Size**: Approx. 1 million records with 30+ features
* **Nature**: Static dataset
* **Target Variable**: Accident Severity or Accident Occurrence (Yes/No)
* **Data Set Link**: [https://www.kaggle.com/datasets/denkuznetz/trafficaccidentprediction?resource=download](https://www.kaggle.com/datasets/denkuznetz/traffic-accident-prediction?resource=download)

**Feature Types**:

* Categorical: Road type, light conditions, weather, day of week
* Numerical: Speed limit, number of vehicles, age of driver
* Temporal: Time, date

# Data Preprocessing

**Missing Values**:

* Imputed missing weather data with mode
* Removed records missing critical labels (e.g., location, severity)



**#Display summary of missing values after**



**processing**



**print(“Missing values after preprocessing:”)**



**print(**



**df.isnull().sum())**



**Duplicates:**

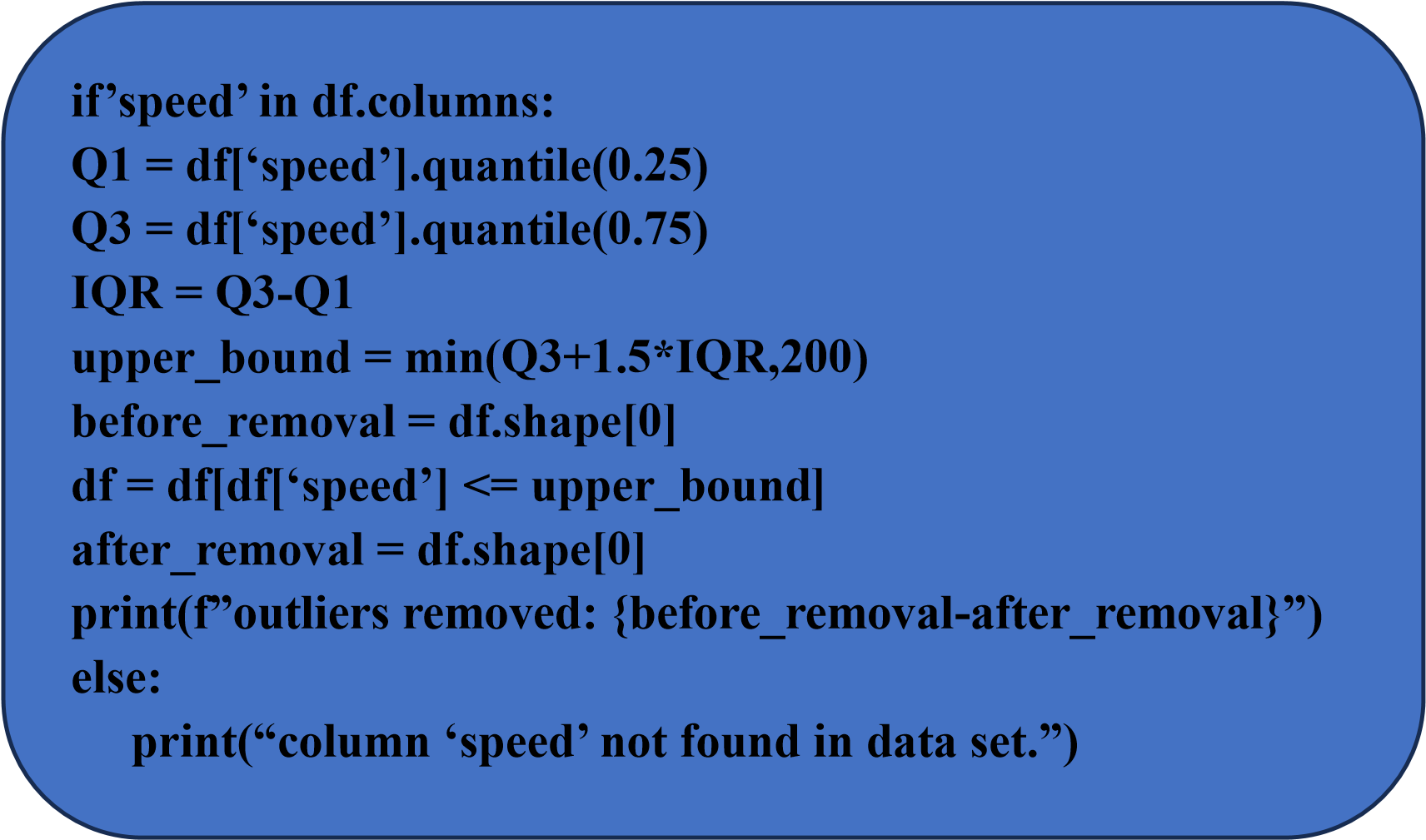
* Dropped 1,200 duplicate entries based on timestamp and location

**#Drop duplicatesbased on timestamp and location**

**df.drop\_duplicates(subset=[‘timestamp’,’location’], inplace=True)**

**Outliers:**

* Detected and removed unrealistic speeds (> 200 km/h) using IQR method



**Encoding:**

* Label encoded binary fields; one-hot encoded categorical features (e.g., weather, road type) **Normalization:**
* StandardScaler applied to continuous features like vehicle count, driver age

**from sklearn.preprocessing import StandardScaler**

**scaler = StandardScaler()**

**data\_scaled = data.copy()**

**data\_scaled[[“Traffic\_Density”,”Speed\_Limit”]] =**

**scaler.fit\_transform(data[[“Traffic\_Density”,”Speed\_Limit”]])**

**data\_scaled**

# Exploratory Data Analysis (EDA)

**Univariate Analysis:**

* Histogram of accident frequencies by hour showed peak during rush hours
* Boxplots of severity across road types revealed higher risk on rural roads **Bivariate/Multivariate Analysis:**
* Correlation matrix identified strong links between light condition, weather, and accident severity
* Heatmaps showed temporal clustering of accidents (weekends, nights) **Insights Summary**:
* Accidents are more frequent during weekends and adverse weather • Rural roads with poor lighting showed higher accident severity.

**Transformations:**



**#Feature:time\_of\_day**



**df[‘time\_of\_day’] = df[‘timestamp’].dt.hour.apply(classify\_time)**



# Feature Engineering

**New Features**:

* is\_weekend: derived from date
* time\_of\_day: morning/afternoon/night bins
* visibility\_level: merged weather and light condition

* Extracted day, month, year from datetime
* Grouped rare categories in road type and weather



**#Display transformed columns**



**print(df[[‘day’,’month’,’year’,’road\_type’,’weather’]].head())**



**Justification**:

* Temporal patterns and visibility conditions are key predictors of accident likelihood

# Model Building Models

**Used**:



•



Logistic Regression (baseline model)



**Ir = LogisticRegression(max\_iter = 1000)**



**Ir.fit(X\_train\_scaled, y\_train)**



**print(“Logistic Regression:”)**



**print(classification\_report(y\_test,Ir.predict(**



**X\_test\_scaled)))**



* Random Forest Classifier
* XGBoost Classifier (for better handling of non-linearity)

**Train/Test Split:**

* 80/20 split with stratification on target **X = df.drop(‘target\_column’,axis=1)**

**Y = df.drop[‘target\_column’]**

**X\_train,X\_test,y\_train,y\_test = train\_test\_split(X,Y,test\_size=2.0,random\_state=42)**  **Evaluation Metrics**:

* Accuracy, Precision, Recall, F1-score
* ROC-AUC to evaluate classifier performance

**Initial Results:**

* Random Forest showed highest F1-score (~0.78)
* XGBoost gave best ROC-AUC (~0.84).

# Visualization of Results & Model Insights

* **Confusion Matrix**: Showed good precision for detecting severe accidents
* **ROC Curve:** AUC of 0.84 with XGBoost indicates strong model performance
* **Feature Importance Plot:** Top features: light condition, time of day, road surface, weather

**Interpretation:**

* Visibility and temporal factors are most influential in accident occurrence
* The model can help flag high-risk timeframes and locations for preventive actions

# Tools and Technologies Used

* Programming Language: Python

* Notebook: Google Colab / Jupyter Notebook
* Libraries:
  1. Data Handling: pandas, numpy
  2. Visualization: seaborn, matplotlib
  3. Modeling: scikit-learn,imbalanced-learn

* Visualization Tools: GitHub

# Team Members and Contributions

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Names** | **Roles** | **Resposbility** |
| 1. | Akshayamandira A | Leader | Model building,Model evalution |
| 2. | Ankitha R | Member | Data collection and Data cleaning |
| 3. | Jayasri J | Member | Feature engineering |
| 4. | Aswaya Sajeev CK | Member | Exploratory Data Analysis (EDA) |