Phase-2 Submission

Student Name: Naveenraj P

Register Number: 712523205043

Institution: PPG Institute of Technology

Department: B. Tech Information Technology

Date of Submission: 09/05/2025

Github Repository Link: **https://github.com/Naveen-Raj12/nm_naveenraj_ds**

---

1. Problem Statement

- *Road accidents pose a significant threat to public safety, resulting in thousands of fatalities and injuries annually. By analyzing historical traffic data, AI can be used to identify accident-prone areas and predict the likelihood of accidents, helping authorities take preventive actions.*
- *This is a **classification problem**, where the goal is to predict whether an accident will occur based on input features like weather, time, location, and vehicle conditions.*
- *Solving this problem can greatly enhance road safety, assist traffic departments in preventive planning, and ultimately reduce road fatalities.*

2. Project Objectives

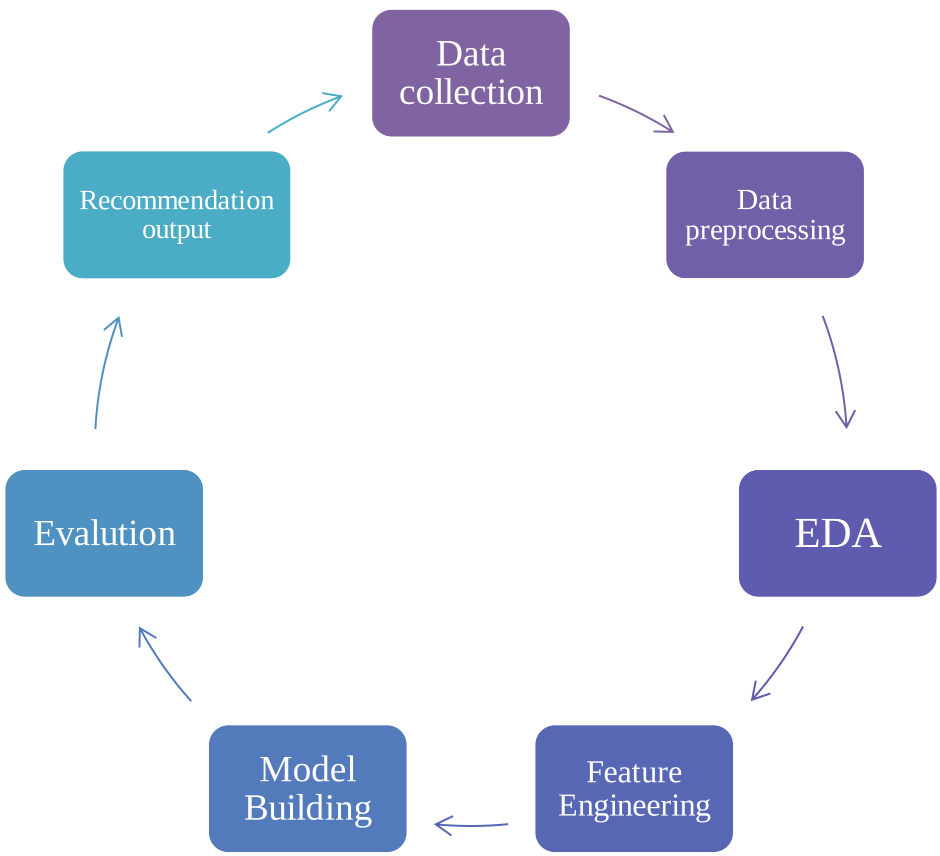
As we transition from planning to implementation, the project goals are:

- *To analyze and preprocess traffic accident data.*
- *To build **predictive models** that can classify whether an accident will happen or not.*

- *To **identify key risk factors** using feature importance and correlation analysis.*
- *To make predictions with high accuracy and interpretability.*
- ***Model Goal:** Achieve a balance between **performance metrics** (like F1-score) and **interpretability** (to inform decision-makers).*

***Evolution:** After exploring the data, we realized that certain features (e.g., weather, light conditions) had a stronger influence than initially thought, shifting focus to include these in more detail.*

3. Flowchart of the Project Workflow



4. Data Description

 ***Source:** [Example: Kaggle's US Accident Dataset or Government Open Traffic Data Portals]*

 ***Type:** Structured tabular data (CSV or SQL database)*

 Attributes:

- *Weather condition*
- *Light condition*
- *Road surface*
- *Time and date of accident*
- *Vehicle count*
- *Location (latitude/longitude or city/state)*

 Size: *~100,000 rows × ~20 columns*

 Nature: *Static (snapshot of past data)*

 Target Variable: *Accident Severity or Binary (Accident: Yes/No)*

5. Data Preprocessing

 Missing Values:

Replaced missing weather values with mode.

Time fields cleaned using datetime parsing.

 Duplicates: *Removed ~3% duplicated entries.*

 Outliers:

Speed values over 300 km/hr considered outliers and dropped.

Extreme visibility (0 or >100 miles) corrected.

 Data Type Conversion:

Converted date strings into datetime objects.

Latitude and longitude preserved as float.

 Encoding:

One-hot encoded categorical features like weather and light conditions.

Label encoded severity levels.

 Normalization:

MinMaxScaler used on speed, temperature, visibility to bring values between 0 and 1.

6. Exploratory Data Analysis (EDA)

 Univariate Analysis:

- *Most accidents occur during rush hours (8-10 AM, 5-7 PM).*
- *Weekends showed a spike in high-speed collisions.*
- *Accidents more frequent in foggy or rainy weather.*

 Bivariate/Multivariate Analysis:

- *Strong correlation between accident severity and weather, lighting.*
- *Pairplot showed overlapping regions for accidents in early morning low-light.*

 Insights:

- *Time of day, road type, and weather are strong predictors.*
- *Poor visibility and wet roads significantly increase accident probability*

7. Feature Engineering

New Features Created:

Extracted 'Hour', 'Day of Week', and 'Month' from timestamp.

Calculated is_weekend from day.

Feature Transformation:

Combined 'Weather Description' into broader categories (e.g., Clear, Rainy, Foggy).

Binned speed into categories: Low, Medium, High.

Interaction Features:

Created visibility × weather interaction to capture poor conditions.

Dimensionality Reduction (optional):

PCA used to reduce 20+ features to 10 principal components (trial phase only).

8. Model Building

- ***Models Used:***

Random Forest: *Good for classification and feature importance.*

Logistic Regression: *Baseline model for comparison.*

- ***Train/Test Split:*** *80/20 with stratified sampling.*

- ***Evaluation Metrics:***

Accuracy: *% of correct predictions.*

Precision: *% of correct positive predictions.*

Recall: *% of actual positives captured.*

F1-score: *Harmonic mean of precision and recall.*

 Results:

- *Random Forest achieved **F1-score of 0.87**, outperforming logistic regression (0.72).*

9. Visualization of Results & Model Insights

 Confusion Matrix: *Showed high true positive rate.*

 Feature Importance Plot:

Top 5 features: weather, hour, visibility, road condition, light condition.

 ROC Curve: *AUC = 0.92 indicating strong classification.*

 Interpretation:

- *Model identifies dangerous combinations (e.g., low light + rain + high speed).*
- *Can assist traffic authorities in issuing alerts or adjusting signals.*

10. Tools and Technologies Used

Language: *Python*

IDE/Environment: *Google Colab / Jupyter Notebook*

Libraries:

Data: pandas, numpy

Visualization: seaborn, matplotlib, plotly

Modeling: scikit-learn, xgboost

Metrics: sklearn.metrics

Optional Visualization Tools: *Tableau / Power BI (for dashboarding)*

11. Team Members and Contributions

Team Member Name	**Contribution**
Vijaya Varma R	Data cleaning, EDA, Feature Engineering
Naveenraj P	Model building and evaluation
Sri Hari Krishna R	Documentation and visualizations
Sanjeev R	Documentation and visualizations
Sneha Jenifer J	Project coordination and Github upload