Phase-2

Student Name: R. Sowmiya

Register Number: 620123106108

Institution: AVS Engineering College

Department: ECE

Date of Submission: 10 – 05 - 2025

Github Repository Link:https://github.

com/sowmiyaraja0411/Webproject.git

---

1. Problem Statement

Road accidents are a major cause of death and injury globally. With the growth of urbanization and traffic, there is an urgent need to predict and analyze accidents using intelligent systems. This project aims to use AI to analyze traffic data, predict accident-prone zones, and identify risk factors contributing to accidents. The problem is a classification problem, where the model classifies accident severity or the likelihood of an accident occurring based on traffic and environmental data. Solving this problem can help prevent accidents, save lives, and support urban planning.

2. Project Objectives

- *Develop an AI model to predict traffic accident likelihood and severity.*
- *Analyze key factors contributing to accidents using historical datasets.*
- *Achieve high accuracy and interpretability in predictions.*
- *Translate predictions into actionable insights for traffic authorities.*
- *Evolve project goals through data exploration and EDA findings*

3. Flowchart of the Project Workflow

[Insert a diagram showing the workflow: Data Collection -> Preprocessing -> EDA -> Feature Engineering -> Model Building -> Evaluation -> Prediction]

4. Data Description

- Dataset Source: Kaggle - "US Accident Dataset"
- Type: Structured data
- Records and Features: ~3 million rows with 50+ features
- Target Variable: Accident Severity (1 to 4)
- Nature: Static dataset with time-series elements
- Features include: Location, time, weather conditions, road surface, visibility, etc.

5. Data Preprocessing

- Missing values handled through imputation and removal (e.g., missing weather info)
- Duplicate records removed.
- Outliers identified in distance, visibility using IQR method
- Categorical variables like weather condition encoded using one-hot encoding.
- Features standardized for model stability

6. Exploratory Data Analysis (EDA)

- Univariate Analysis: Distribution of accidents by severity, weather, time.
- Bivariate Analysis: Correlation between weather and accident severity.
- Insights:
- Most accidents occur in poor weather or during rush hours.
- Severity increases with reduced visibility and wet road surfaces.

7. Feature Engineering

- Created new feature: `is_rush_hour` from time.
- Converted date/time into weekday, hour.
- Binned distances into categories.
- Removed highly correlated redundant features.
- Applied PCA for dimensionality reduction (optional, based on results)

8. Model Building

- Models used: Random Forest Classifier, XGBoost Classifier
- Train-test split: 80/20 with stratification
- Metrics Used:
- Accuracy, Precision, Recall, F1-score
- Random Forest gave better interpretability, XGBoost better accuracy (~87%)

9. Visualization of Results & Model Insights

- Confusion Matrix: Showed balance of class predictions.
- Feature Importance: Weather condition, visibility, hour of day are top contributors.
- ROC Curve: AUC score of ~0.91 for XGBoost.
- Residual Plot: Used to evaluate misclassifications.

10. Tools and Technologies Used

- Language: Python
- IDE: Jupyter Notebook
- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost
- Visualization Tools: seaborn, matplotlib, plotly

11. Team Members and Contributions

- S. Pooja - Data Cleaning - EDA - Documentation
- B. Poornasri - Feature Engineering - Model Building
- R. Sowmiya - Model Evaluation-Result Visualization - Reporting