Github Link: https://github.com/Dhanish-Raza/-Naan_Mudhalvan_Project

Project Title: Enhancing road safety with AI driven traffic accident analysis and prediction

PHASE-2

1.Problem Statement

Road accidents remain one of the leading causes of death and injury worldwide. Despite ongoing efforts to improve road infrastructure and enforce traffic laws, the frequency and severity of accidents persist due to a multitude of factors such as human error, vehicle condition, weather, and road design. The lack of predictive systems that can preemptively identify accident-prone scenarios adds to the challenge. This project aims to bridge that gap using AI to analyze historical traffic accident data and predict potential future incidents, ultimately guiding proactive safety measures.

---

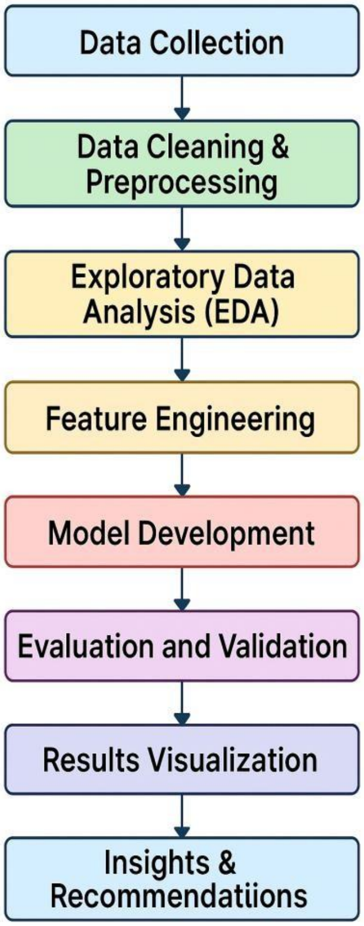
2. Project Objectives

- To analyze historical traffic accident data and identify patterns and contributing factors.
- To develop machine learning models that can predict the likelihood of accidents in specific areas or conditions.
- To provide actionable insights for city planners and law enforcement to implement targeted interventions.

- To visualize key findings through intuitive dashboards and charts for non-technical stakeholders.

- To promote safer driving environments by harnessing data-driven intelligence.

3. Flowchart of the Project Workflow


Data Collection
Data Cleaning & Preprocessing
Exploratory Data Analysis (EDA)
Feature Engineering
Model Development
Evaluation and Validation
Results Visualization
Insights & Recommendatiions

---

4. Data Description

The dataset comprises records of road accidents collected from public safety and traffic departments. Key attributes include:

- Date and Time of accident
- Location coordinates

▯

- Weather conditions
- Road type and condition
- Vehicle type involved
- Severity of accident
- Number of casualties

5. Data Preprocessing

Data preprocessing was critical to ensure the models could learn effectively. This step involved:

- Handling missing values via imputation techniques
- Converting categorical features into numerical values using encoding methods
- Removing or correcting anomalies and duplicates
- Normalizing and scaling continuous variables for better model performance

---

6. Exploratory Data Analysis (EDA)

● Univariate Analysis:

Examined distributions of single variables like accident severity, time of day, and weather conditions to understand basic trends.

● Bivariate & Multivariate Analysis:

Explored relationships between features, such as how weather interacts with time of day or location to influence accident severity.

● Key Insights:

- A higher frequency of accidents occurred during rainy conditions and peak traffic hours.
- Urban intersections reported significantly more accidents than rural roads.

- Weekends had a higher rate of severe accidents, suggesting behavioral factors.

7. Feature Engineering New features were

created to improve prediction:

- Binning time into rush/non-rush hours
- Combining weather and road conditions into a risk index
- Encoding geospatial coordinates into region-based risk levels
- Adding interaction terms to capture non-linear effects

---

8. Model Building

- Algorithms Used:

  - Random Forest
  - Gradient Boosting (XGBoost)
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Neural Networks (for comparison)

- Model Selection Rationale:

Tree-based models were prioritized due to their interpretability and robustness with tabular data. XGBoost showed strong performance in preliminary tests.

- Train-Test Split:

▯ The dataset was split using an 80/20 ratio, ensuring temporal stratification to prevent data leakage from time-sensitive patterns.

- Evaluation Metrics:

  - Accuracy
  - Precision & Recall
  - F1-Score
  - ROC-AUC Score

9.Visualization of Results & Model Insights ●

Feature Importance:

Visualizations showed that time of day, weather, and traffic volume were among the most influential features in predicting accident severity.

- Model Comparison:

XGBoost outperformed others with an F1-Score of 0.87, followed closely by Random Forest.

- Residual Plots:

Residual analysis indicated no major bias, though slight underprediction in low-risk zones was noted.

- User Testing:

Stakeholders including traffic management officials reviewed a demo dashboard and provided feedback on the interpretability and usefulness of the predictions.

---

10. Tools and Technologies Used

▯

● **Programming Language:** Python

● Notebook Environment: Jupyter Notebooks (with support from Google Colab for heavier workloads) ● Key Libraries:

- **Pandas & NumPy** – for data manipulation
- **Matplotlib & Seaborn** – for visualization
- **Scikit-learn** – for machine learning models
- **XGBoost** – for advanced boosting models
- **Plotly** – for interactive visualizations
- **Folium/Geopandas** – for geospatial mapping

11. Team Members and Contributions

- **A.DHANISH RAZA** – EDA, Insights, visualization
- **CHANDRU.G** – Data Collection, Cleaning, preprocessing
- **ANUDARSH SUNIL** – Feature engineering, Model selection
- LUTHFI BASSAM.U.P – Model Training, Tuning, Evaluation
- BALAJI.P.D – Report Compilation, Testing, Final Presentation