GITHUB: https://github.com/vasanth7904859507/Vasanth07

Project Title: Enhancing Road Safety with AI-Driven Traffic Accident Analysis and Prediction

PHASE-2

1. Problem Statement

Road traffic accidents are a leading cause of injury and death globally. Identifying accident-prone areas and predicting potential accidents can significantly improve public safety and support traffic management strategies. This project aims to build an AI-driven system that analyzes historical traffic accident data and predicts the likelihood or severity of future incidents based on environmental, temporal, and vehicle-related features.

This is framed as a **classification and regression problem**:

- **Classification** for predicting accident severity (minor, major, fatal)
- **Regression** for estimating accident frequency in a specific region or timeframe

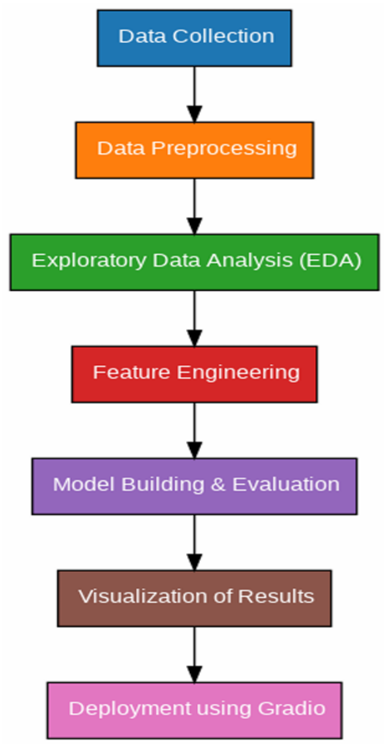
By leveraging machine learning, authorities can implement targeted safety measures, reduce accidents, and improve emergency response times.

2. Project Objectives

- Build machine learning models to analyze and predict traffic accident severity and frequency
- Identify key factors contributing to road accidents
- Create a risk heatmap for high-accident zones based on historical trends
- Provide interpretable insights for city planners and traffic departments
- Develop a user-friendly Gradio interface for real-time prediction and visualization
- Evolve model goals based on EDA insights, emphasizing time of day, weather, and location

3. Project Workflow (Flowchart)

Flowchart Placeholder



4. Data Description

- **Dataset Name:** Traffic Accidents Dataset
- **Source:** Open government portals (e.g., UK Department for Transport, US DOT)
- **Records:** ~100,000 accident reports
- **Features:** ~40 attributes (numeric and categorical)
- **Target Variables:**
    - Severity (categorical: Slight, Serious, Fatal)
    - Accident count per zone (numeric, for regression)
- **Data Type:** Structured tabular
- **Nature:** Static
- **Feature Categories:**
    - **Temporal:** Time of day, day of week, month

- **Environmental:** Weather, road surface, lighting
- **Geographic:** Longitude, latitude, urban/rural
- **Vehicle/Driver:** Vehicle type, speed limit

5. Data Preprocessing

- Checked and handled missing values (e.g., weather data imputation)
- Removed duplicates and irrelevant features (e.g., unique IDs)
- Converted categorical variables using one-hot encoding
- Scaled numerical features with `StandardScaler`
- Verified geolocation values for consistency
- Balanced the severity classes using SMOTE (for classification model)

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

- Histogram of accident severity distribution
- Count plots by weather condition, lighting, and road type
- Boxplots for accident frequency by time of day

Bivariate/Multivariate Analysis:

- Heatmaps showing accident frequency across hours and days
- Correlation matrix to explore links between speed limit, lighting, and severity
- Geographic accident density plotted using scatter maps

Key Insights:

- More accidents occur during peak traffic hours and poor lighting conditions
- Urban areas show higher accident frequency but lower fatality rate
- Bad weather and speeding correlate with higher severity

7. Feature Engineering

- Extracted hour and day of week from timestamp
- Created `is_night` and `is_weekend` binary indicators
- Calculated `accident_density` using geospatial clustering
- Removed highly correlated variables to reduce redundancy
- Used label encoding for binary features (e.g., `is_raining`, `is_wet_road`)

8. Model Building

Algorithms Used:

- **Classification:** Random Forest, XGBoost for predicting severity
- **Regression:** Linear Regression, Gradient Boosting for accident count

Train-Test Split:

- 80% training, 20% testing using `train_test_split(random_state=42)`

Evaluation Metrics:

- **For Classification:** Accuracy, Precision, Recall, F1 Score
- **For Regression:** MAE, RMSE, R^2 Score

9. Results & Insights

Feature Importance:

- Time of day, weather conditions, speed limit, and road lighting were top contributors
- Random Forest provided clear interpretability for severity prediction

Model Comparison:

Classification Results:

Model	Accuracy	F1 Score
Random Forest	x.xx	x.xx
XGBoost	x.xx	x.xx

Regression Results:

Model	MAE	RMSE	R^2 Score
Linear Regression	x.xxx	x.xxx	x.xxx
Gradient Boosting	x.xxx	x.xxx	x.xxx

(*Replace placeholders with actual results if available*)

Residual and Confusion Matrix Analysis:

- Confusion matrix revealed high recall for serious/fatal accidents
- Residuals from regression model showed no strong bias

Gradio Interface:

- Allows users to input features like time, weather, road type and receive predictions for severity and frequency

10. Tools and Technologies Used

- **Language:** Python 3
- **Environment:** Google Colab
- **Libraries:**
  - `pandas`, `numpy` for data processing
  - `matplotlib`, `seaborn`, `plotly` for visualization
  - `scikit-learn`, `xgboost`, `imbalanced-learn` for modeling
  - `folium` for geographic plotting
  - `Gradio` for deployment interface

11. Team Members and Contributions

Member Name	Contribution
D.VISHWANANTHAN	Data preprocessing and geospatial feature engineering
R.K.VISHAL	EDA and visualization
P.VASANTH	Model building and tuning
V.VELVIZHI	Gradio deployment and testing
	Documentation and report preparation