**PHASE-2**

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

**Student Name:** Suraj .K

**Register Number:** 511723104011

**Institution:** Palllavan college of Engineering

**Department:** Computer science Engineering

**Date of Submission:**

**Github Link**: [**https://github.com/Thulasimathi26/Data-Science.git**](https://github.com/Thulasimathi26/Data-Science.git)

# Problem Statement

Road traffic accidents cause significant loss of life, injuries, and economic damage annually. Traditional analysis methods are often limited to historical assessments and do not efficiently predict future incidents. With the rise of AI and the availability of traffic, weather, and road condition data, we can now build predictive models that identify risk patterns and forecast accident-prone zones.

This project aims to develop an AI-driven predictive system capable of analyzing multiple factors (weather, time, road type, traffic patterns) to predict accident likelihoods, thus aiding authorities in taking preventive actions to enhance road safety.

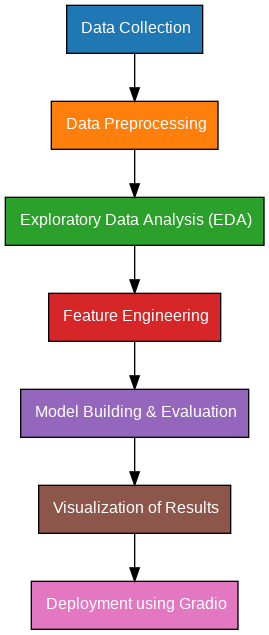
.

# Project Objectives

* Develop a machine learning-based system to predict road traffic accidents.
* Identify key risk factors influencing accident occurrence and severity.
* Analyze accident-prone zones and generate actionable insights.
* Provide a web-based visualization/dashboard showcasing accident hotspots.

Recommend strategies for proactive road safety improvements

# Flowchart of the Project Workflow



# Data Description

* **Datasets Used:**
  + UK Road Safety Data
  + US NHTSA Fatality Analysis Reporting System (FARS)
  + Kaggle Datasets (UK and US accidents)
  + OpenWeatherMap / NOAA Weather Data
* **Attributes Covered**: Demographics (age, address, parents’ education), academics (G1,

G2, study time), and behavior (alcohol consumption, absences)

* Dataset Link: [Student Performance - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/320/student%2Bperformance)
* **Types of Data:**
  + Temporal Data (Date, time)
  + Geospatial Data (Location, road type)
  + Environmental Conditions (Weather, road surface, lighting)
  + Traffic Data (Vehicle count, congestion)
  + Accident Details (Severity, casualties, vehicles involved)

Dataset Links:

* [UK Road Accidents (Kaggle)](https://www.kaggle.com/datasets/sukhbirrahil/uk-road-safety-accidents-and-vehicles)
* [US Accidents Dataset (Kaggle)](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

# Data Preprocessing

* Handling missing values, duplicates.
* Encoding categorical variables.
* Normalizing numerical features.
* Feature engineering (time buckets, weather severity index).
* Merging datasets on common attributes like location and time.

# Exploratory Data Analysis (EDA)

* Analyzing accident distribution across time, weather conditions, road types.
* Identifying high-frequency accident zones.
* Correlation analysis between features (e.g., weather severity and accidents).
* Visualizing accident patterns through histograms, bar charts, and heatmaps.

**Key Insights:**

* Bad weather and nighttime significantly increase accident risks.
* Certain intersections and highways are more accident-prone.

# Feature Engineering

* Creating derived features (e.g., peak hours, weather categories).
* Dimensionality reduction (PCA if necessary).
* Feature selection based on correlation and feature importance.

# Model Building

* **Algorithms Used:**
  + Logistic Regression (for binary classification)
  + Random Forest Classifier (for feature importance and prediction)
  + XGBoost (for enhanced performance)
* **Model Selection Rationale:**
  + Logistic Regression: Simple baseline model.
  + Random Forest: Handles non-linear relationships and feature ranking.
  + XGBoost: Highly accurate gradient boosting method.
* **Train-Test Split:**
  + 80% training, 20% testing.
* **Evaluation Metrics:**
  + Accuracy, Precision, Recall, F1-Score.
  + ROC-AUC for classification tasks.

# Visualization of Results & Model Insights

* Heatmaps of accident-prone zones using Folium.
* Feature importance visualization from Random Forest.
* Confusion matrix and ROC curve for model evaluation.
* Dashboard for displaying accident trends and predictive insights.

# Tools and Technologies Used

● **Programming Language**: Python 3 ● **Notebook Environment**: Google Colab ● **Key Libraries**:

○ pandas, numpy for data handling

○ matplotlib, seaborn, plotly for visualizations

○ scikit-learn for preprocessing and modeling

○ Gradio for interface deployment

# Team Members and Contributions

# 

|  |  |
| --- | --- |
| ***Team members*** | ***contribution*** |
| *Suraj .k* | *Data cleaning* |
| *Owalraj .B* | *EDA* |
| *Mohamed shahil .S* | *Feature engineering* |
| *Gokul .D* | *Model development* |
| *Suresh .P.B* | *Documentation and reporting]* |