Phase-2

Student Name: Janani S

Register Number: 510823205016

Institution: Ganadipathy Tulsi's Jain Engineering College

Department: B.Tech-Information Technology

Date of Submission: 08-05-2025

Github Repository Link: github.com/Janani-hub01/phase-2

1. Problem Statement

- Upon revisiting the problem with access to traffic datasets-including historical accident data, weather conditions, road types, vehicle information, and driver behavior –the problem can be broken down into multiple machine learning tasks:

- Classification: Predict the severity level of an accident(minor,major,fatal)

- Impact: Proactively identifying high-risk scenarios or accident-prone zones can significantly reduce the number and severity of accidents through timely interventions (e.g., dynamic speed limits, alert systems)

2. Project Objectives

- Clean and preprocess traffic accident datasets.

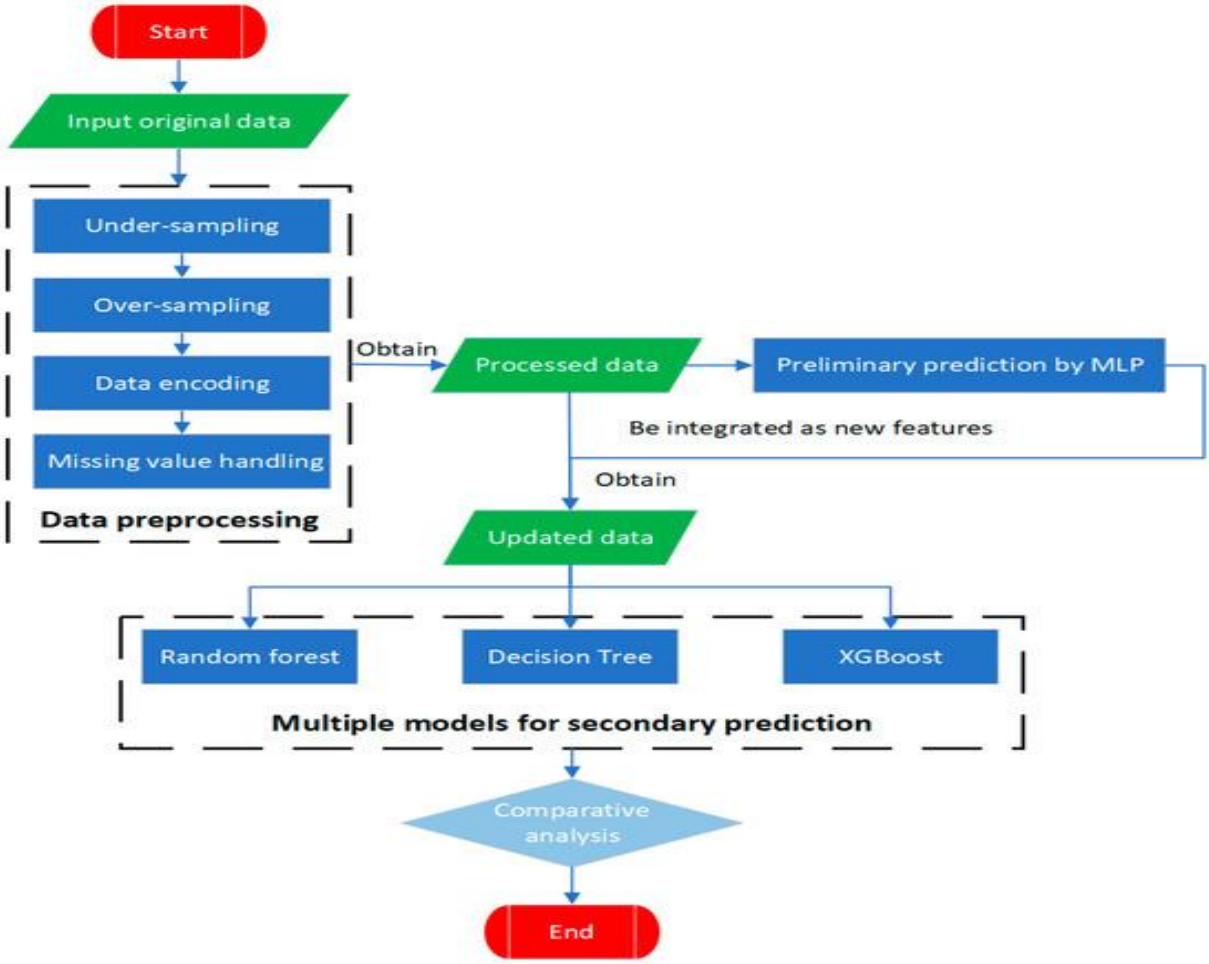
- Select and engineer relevant features such as time, location, weather, and traffic density.

- Train and evaluate classification models(e.g., Random forest, XGBoost)for accident likelihood.

- Model goals & Evolution:

    - Original Goal: Predict accidents using basic classification.
    - Evolved Goal: Integrate classification, regression, and clustering to provide a comprehensive traffic accident prediction and analysis solution with real world applicability.

3.Flowchart fort the project



4.Data Descritption

- *Dataset* : Datasets: Road Accidents in US

- source: kaggle (US Accidents (2016-2021

- Number of records: 10 of 51 columns

- Type of data: Structured data

5. Data Preprocessing

- We imputed numerical columns with the mean and filled categorical columns with 'Unknown'. Rows missing essential fields were removed.
- Duplicates were identified and dropped to maintain data integrity.
- Outliers were removed based on IQR to prevent skewed model training.
- Date and Time fields were converted properly; numerical types enforced for consistency.
- Label encoding was used for ordinal data, and one-hot encoding for nominal variables.

- Features were standardized to zero mean and unit variance for optimal model performance.

6. Exploratory Data Analysis (EDA)

- *Univariate Analysis:*
  - **Vehicle speed** tends to follow a normal distribution but with some outlier

  - **High severity accidents** are more frequent than expected.
  - **Rainy and Foggy** weather contributes to a notable number of accidents.
  - **Boxplot shows that high severity** accidents are associated with **higher average speed**s

    ○

- *Bivariate/Multivariate Analysis:*

    - Correlation matrix & heatmap
    - Grouped Bar Plot – Weather vs Severity
    - Pairplot

- *Insights Summary:*

Feature	Why it Matters
Vehicle_Speed	Strong predictor of severity and risk.
Weather	Influences visibility and road traction.
Road_Type	Dictates likely speed, traffic, and infrastructure.
Time of Day	Nighttime often correlates with higher accident risk.
Driver_Age (if available)	May influence reaction time and experience.

7. Feature Engineering

- Extract/Time Components.
- Binning speed into categories.
- Polynomial or Interaction Features(High speed in **rainy** weather may be a strong indicator of accident severity.)
- Dimensionality Reduction (PCA)-- Useful for simplifying feature space while retaining variance.

8. Model Building

- **Logistic Regression (Multinomial)** – A strong baseline for multiclass classification, interpretable, and fast.
- **Random Forest Classifier** – Handles non-linear relationships well, robust to outliers, and great for feature importance

These are suitable because:

- We're dealing with structured data.
- The target variable (Severity) is categorical with 3 classes.
- Data has both numerical and categorical features.

9. Visualization of Results & Model Insights

- Heatmaps of Accident Locations: show high risk zones on city maps.
- Time series graphs: show trends in accident occurrences across time(days,seasons).
- Confusion Matrix: For classification model performance(e.g.,predicting severity)

10. Tools and Technologies

Programming Language –Python: Main programming languag

- JavaScript (Node.js): Useful for real-time data dashboards and visualization in web apps.

- **Notebook/IDE** – Jupyter Notebook: Exploratory data analysis(EDA)

- **Libraries** – Pandas, Numpy –Data manipulation
  Matplotlib,seaborn,Plotly: Data visualization

Scikit-learn: ML algoriths(classification/regression)

- **Optional Tools for Deployment** – Tableu or Power BI: Data visualization

11. Team Members and Contributions

Janani S- Data collection & preprocessing, Tools and Technologies

Sanjusri – Tools and Technologies

Santhoshini – Deployment & documentation

Priyadharshini – Presentation