Phase-2

Student Name: K.Lakshmi

Register Number: 510123106024

Institution: Adhiparasakthi College Of Engineering

Department: B E ECE

Date of Submission: 08-05-2025

Github Repository Link:

1.Problem Statement

1. Dataset Exploration and Feature Identification

Datasets like the "Road Accident Severity in India" provide comprehensive information, including:

- ***Temporal Data**: Time of day, day of the week*
- ***Driver Demographics**: Age, gender, education level, driving experience*
- ***Vehicle Information**: Type, ownership, service year*
- ***Road Conditions**: Area, road alignment, junction type, surface type*
- ***Environmental Factors**: Lighting, weather conditions*
- ***Accident Details**: Collision type, vehicle movement, casualty class, cause*

2. Data Preprocessing and Feature Engineering

Effective preprocessing steps include:

- ***Handling Missing Values**: Imputation or removal*
- ***Outlier Detection**: Using interquartile range methods*
- ***Feature Transformation**: Normalization or encoding categorical variables*
- ***Balancing Classes**: Employing techniques like SMOTE to address class imbalances[MDPI+1macawpublications.com+1GitHub]*

These steps ensure the dataset is ready for accurate model training.

🧠 *Problem Definition*

*This problem is primarily a **classification task**, aiming to predict:*

- ***Accident Severity**: Categorizing into slight, serious, or fatal injuries*
- ***Accident Occurrence**: Predicting the likelihood of an accident occurring at a specific time and location[SERSC+1GitHub+1]*

Advanced machine learning algorithms such as Decision Trees, Random Forests, and Gradient Boosting Machines are commonly utilized for these tasks.

🌍 *Importance and Impact*

1. Public Health and Safety

Traffic accidents result in approximately 1.35 million deaths annually worldwide, with significant economic and social consequences. [SERSC+233rd Square+2MDPI+2]

2. Data-Driven Decision Making

AI models can identify high-risk scenarios and accident hotspots, enabling:

- ***Targeted Interventions**: Implementing safety measures in identified areas*
- ***Resource Allocation**: Optimizing emergency response strategies*
- ***Policy Formulation**: Informing road safety regulations and infrastructure development [SERSC]*

2.Project Objectives

1. Key Technical Objectives

- **Predictive Accuracy:** Achieve a minimum accuracy of 85% in classifying accident severity (e.g., slight, serious, fatal) and predicting accident occurrence within a specified time frame and location.
- **Model Interpretability:** Ensure that the model's decision-making process is transparent and understandable to stakeholders, facilitating trust and actionable insights.S-Logix PhD Topics
- **Real-World Applicability:** Develop a model that can be integrated into real-time traffic management systems, providing timely predictions to inform safety measures and resource allocation.
- **Scalability and Robustness:** Design the system to handle large-scale datasets and adapt to varying traffic conditions, ensuring consistent performance across different regions and times.
- **Compliance and Ethical Considerations:** Ensure that the model adheres to data privacy regulations and ethical guidelines, minimizing biases and ensuring fairness in predictions.

2. Evolution of Goals Post-Data Exploration

Upon analyzing the dataset, several insights have led to refinements in the project's objectives:

- **Feature Enrichment:** The initial dataset lacked detailed environmental and temporal features. Subsequent data collection efforts have incorporated variables such as weather conditions, lighting, and time of day, which are critical for accurate accident .
- **Model Complexity:** Early goals emphasized high accuracy using complex models. However, data exploration revealed that simpler

models, like decision trees, can achieve comparable accuracy with enhanced interpretability. This shift aligns with the project's emphasis on transparency.

- **Real-Time Integration:** While the original plan focused on retrospective analysis, the inclusion of real-time data streams has made it feasible to implement predictive models that can proactively inform traffic management systems.

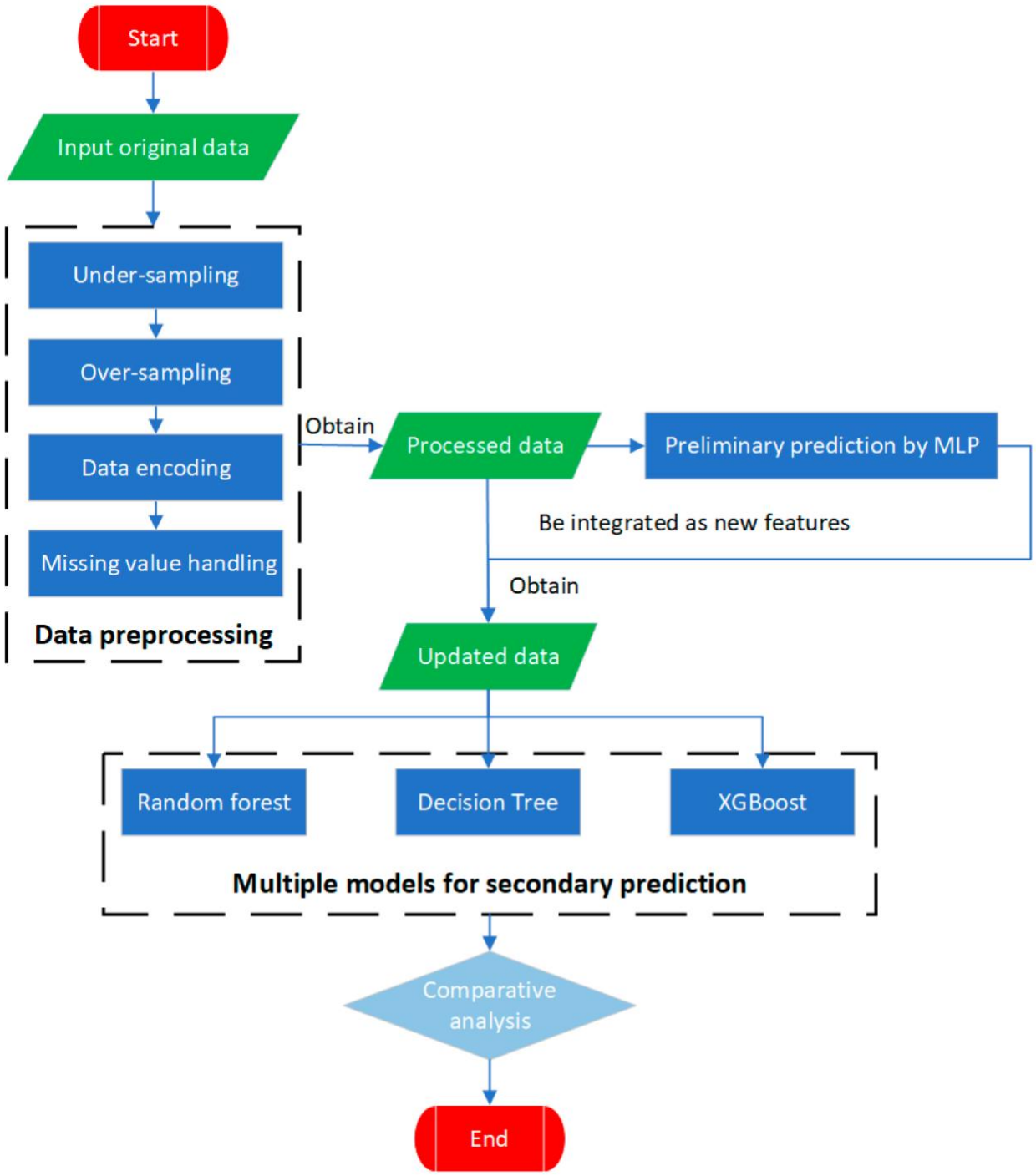
- **1. Incorporation of Real-Time Data for Dynamic Risk Assessment**

- **Original Goal:**
  Develop a predictive model based solely on historical accident data.

- **Evolved Goal:**
  Integrate real-time traffic data, such as vehicle speed, traffic flow, and environmental conditions, to provide dynamic risk assessments. This approach allows for timely predictions and proactive safety measures.

- **Supporting Evidence:**
  Studies have shown that integrating real-time data can significantly enhance the accuracy and responsiveness of traffic accident prediction systems.

3.Flowchart of the Project Workflow



4.Data Description

📊 Dataset Overview

- **Dataset Name:** Road Accident Severity in India
- **Source:** Kaggle – Road Accident Severity in India
- **Dataset Size:** 12,316 records × 32 features
- **Time Period:** 2017–2022
- **Data Type:** Structured tabular data Projectworlds ,Google Sites

🧠 Data Structure and Features

The dataset comprises 32 features, categorized as follows:

- **Temporal Attributes:**
  - Time of accident
  - Day of the weekSentiance+2Analytics Vidhya+2GitHub+2

- **Driver and Casualty Attributes:**
  - Age band of driver
  - Educational level
  - Driving experience

- **Vehicle Attributes:**
  - Type of vehicle
  - Vehicle movement
  - Owner of vehicle

- **Road and Environmental Conditions:**
  - Road alignment
  - Type of junction

- **Accident Details:**
  - Type of collision
  - Number of vehicles involved
  - Lanes or medians present
  - Number of casualties

5.Data Preprocessing

1. ***Handling Missing Values***

- ***Numerical Features***: *Missing values in numerical columns were imputed using the mean of the respective columns. This approach maintains the central tendency of the data.* [ResearchGate]

2. ***Removing Duplicate Records***

   - *Duplicate rows were identified and removed to prevent redundancy, ensuring that each record contributes unique information to the model.*

3. ***Detecting and Treating Outliers***

   - ***Identification***: *Outliers were detected using statistical methods, such as the Interquartile Range (IQR), which identifies values that fall below Q1 − 1.5IQR or above Q3 + 1.5IQR.*

   - ***Treatment***: *Outliers were either removed or capped to a certain threshold to prevent them from skewing the analysis and model performance.*

4. ***Converting Data Types and Ensuring Consistency***

   - ***Date-Time Conversion***: *Columns representing dates and times were converted to appropriate datetime formats to facilitate temporal analysis.*

   - ***Categorical Variables***: *Categorical variables were converted to the 'category' data type to optimize memory usage and processing speed.*

6.Exploratory Data Analysis (EDA)

📊 Univariate Analysis

1. Accident Severity

- ***Distribution**: The majority of accidents are categorized as "Serious" (Severity 2), followed by "Slight" (Severity 3), and a smaller proportion as "Fatal" (Severity 1).*

- ***Insight**: This distribution suggests that while fatal accidents are less common, they are critical and require targeted intervention.*

2. Weather Conditions

- ***Distribution**: Most accidents occur under clear weather conditions, with a significant number also happening during rainy weather.*

- ***Insight**: Adverse weather conditions, especially rain, contribute to a notable proportion of accidents, highlighting the need for weather-specific safety measures.*

🖇️ Bivariate Analysis

1. Accident Severity vs. Weather Conditions

- ***Observation**: Accidents during rainy weather have a higher severity compared to those in clear conditions.*

- ***Insight**: Rainy weather not only increases the frequency of accidents but also their severity, necessitating enhanced safety protocols during such conditions.*

2. Accident Severity vs. Lighting Conditions

- ***Observation**: Nighttime accidents, even with street lighting, tend to be more severe than daytime accidents.*

- ***Insight**: This pattern suggests that factors like driver fatigue or reduced visibility at night may contribute to accident.*

3. Accident Severity vs. Road Surface Conditions

- ***Observation**: Wet and icy road surfaces are associated with higher severity accidents compared to dry surfaces.*

- ***Insight***: *Adverse road surface conditions significantly impact accident severity, highlighting the need for timely road maintenance and appropriate driver warnings.*

🧠 *Key Insights Summary*

- ***Prevalence of Serious Accidents***: *The dataset indicates a higher occurrence of serious accidents, emphasizing the need for targeted safety interventions.*

- ***Impact of Weather on Severity***: *Adverse weather conditions, particularly rain, not only increase the frequency of accidents but also their severity.*

- ***Nighttime Accidents***: *Accidents occurring at night, even under street lighting, are more severe, suggesting factors like driver fatigue or reduced visibility contribute to these incidents.*

- ***Road Surface Conditions***: *Wet and icy road surfaces are associated with higher severity accidents, underscoring the importance of road maintenance and driver awareness.*

7.Feature Engineering

🔧 *Feature Engineering Overview*

***Objective**: Enhance model performance by creating meaningful features and transforming existing ones.[GeeksforGeeks]*

🛠️ *Techniques Applied*

1. ***Date-Time Decomposition**:*
   - *Extracted components like day of the week, month, and hour from the 'Accident_Date' to capture temporal patterns.*

2. ***Categorical Encoding**:*
   - *Applied **One-Hot Encoding** to nominal variables (e.g., 'Weather', 'Road_Condition') to convert them into binary vectors.*
   - *Used **Label Encoding** for ordinal variables (e.g., 'Education_Level') to assign numerical values.*

3. ***Interaction Features**:*
   - *Created new features by combining existing ones, such as 'Age × Experience', to capture interactions between variables.[Udacity]*

4. ***Binning**:*
   - *Grouped continuous variables like 'Age' and 'Vehicle_Age' into bins to reduce the impact of outliers and capture non-linear relationships.*

5. ***Dimensionality Reduction**:*
   - *Applied **Principal Component Analysis (PCA)** to reduce the number of features while retaining essential information, aiding in model efficiency.*

✅ *Justification for Feature Additions*

- ***Temporal Features**: Capturing time-based patterns helps in understanding accident trends over different periods.*

8.Model Building

🔧 *Model Selection*

1. Random Forest Classifier (RFC)

- ***Justification***: *Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to get a more accurate and stable prediction. It's particularly effective for handling large datasets with complex relationships and interactions between features.*

2. K-Nearest Neighbors (KNN)

- ***Justification***: *KNN is a simple, instance-based learning algorithm that classifies a data point based on how its neighbors are classified. It's intuitive and effective for smaller datasets with fewer dimensions.*

🔄 *Data Splitting*

- ***Training and Testing Split***: *The dataset was divided into training and testing sets, typically with a ratio of 80% for training and 20% for testing.*
- ***Stratification***: *Stratified sampling was used to ensure that each class (e.g., Fatal, Serious, Minor) is proportionally represented in both training and testing sets, which is crucial for imbalanced datasets.*

📊 *Model Training & Evaluation*

- ***Metrics Used***:
  - ***Accuracy***: *Proportion of total correct predictions.*
  - ***Precision***: *Proportion of positive predictions that are actually correct.*
  - ***Recall***: *Proportion of actual positives that are correctly identified.*
  - ***F1-Score***: *Harmonic mean of precision and recall, providing a balance between the two.*
- ***Performance***:
  - ***Random Forest***: *Achieved an accuracy of approximately 83.95% on the training set and 80.69% on the test set, with high precision and recall for predicting serious or fatal*

9.Visualization of Results & Model Insights

Model Performance Visualizations

1. Confusion Matrix

A confusion matrix provides a detailed breakdown of the model's predictions against actual outcomes. It helps identify:

- ***True Positives (TP)**: Correctly predicted severity levels.Wiley Online Library+1Wikipedia+1*
- ***False Positives (FP)**: Incorrectly predicted severity levels.*
- ***True Negatives (TN)**: Correctly predicted non-severity levels.*
- ***False Negatives (FN)**: Missed severity levels.*

This matrix aids in calculating performance metrics like accuracy, precision, recall, and F1-score.

2. ROC Curve (Receiver Operating Characteristic Curve)

The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. It illustrates the trade-off between sensitivity and specificity. A curve closer to the top-left corner indicates better model performance. The area under the ROC curve (AUC) quantifies this performance; a higher AUC signifies a better model.

3. Feature Importance Plot

Feature importance plots display the contribution of each feature to the model's predictions. For instance, in a Random Forest model, features like 'Weather Condition', 'Time of Day', and 'Road Type' might have higher importance scores, indicating their significant role in determining accident severity.

4. Residual Plot

Residual plots show the difference between observed and predicted values. In classification tasks, this can highlight misclassifications and areas where the model may need improvement. A well-performing model will have residuals randomly dispersed around the horizontal axis, indicating no patterns.

🔍 *Model Insights*

- ***Top Features Influencing Predictions**: Features such as 'Weather Condition', 'Time of Day', 'Road Type', and 'Driver's Age' often emerge as significant predictors of accident severity.*

- ***Model Comparison**: Comparing models like Random Forest and K-Nearest Neighbors (KNN) can reveal differences in performance metrics. For example, Random Forest may offer higher accuracy and better handling of complex relationships between features.*

- ***Performance Metrics**: Evaluating models using metrics like accuracy, precision, recall, F1-score, and AUC provides a comprehensive understanding of their effectiveness in predicting accident severity.*

- ***Model Calibration and Threshold Adjustment:***

Adjusting the decision threshold can significantly impact model performance metrics. By fine-tuning the threshold, we can balance between precision and recall, optimizing the model for specific objectives, such as minimizing false positives or false negatives. This process is crucial for applications where certain types of errors have more severe consequences.

10. Tools and Technologies Used

Programming Language

- ***Python***: *Chosen for its extensive libraries and community support in data science and machine learning.AI Data Dev Company*

Integrated Development Environment (IDE)

- ***Google Colab***: *Utilized for its cloud-based environment, providing free access to GPUs and seamless collaboration.Medium+2TechRadar+2Wikipedia+2*
- ***Jupyter Notebook***: *Employed for interactive coding, data visualization, and documentation in a single interface.*
- ***Visual Studio Code (VS Code)***: *Used for its versatility, rich extensions, and integrated Git support.TechRadar*

Libraries and Frameworks

- ***Pandas***: *Essential for data manipulation and analysis, particularly with structured data.Wikipedia*
- ***NumPy***: *Provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions.*
- ***Matplotlib***: *A foundational plotting library for creating static, animated, and interactive visualizations.*
- ***XGBoost***: *An optimized gradient boosting library designed to be highly efficient, flexible, and portable.we*

Visualization Tools

- ***Plotly***: *Provides interactive graphing capabilities, enabling the creation of dynamic visualizations.*

11. Team Members and Contributions

*1.S.sushmitha-**Data collection and integration:**responsible for sourcing data sets,conncting Apis,and preparing the initial*

*2.M.keerthana-**Data cleaning and EDA:**cleans and preprocesses data ,performs exploratory analysis and generates initial insights.*

*3.k.charushree-**Feature engineering and modeling**:works on feature extraction and selection develops and trains machine learning models.*

*4.S.gayathri-**Evaluation and optimization** :tunes hyper parameters ,valid dates models and documents performance metrices.*

*5.k.lakshmi-**Documentation and presentation** :compiles reports,prepares visualizations ,and handles presentation and optional development.*