**Phase-2**

**Student Name:** [S.Lakshmi Prasanna]

**Register Number:** [720323104082]

**Institution:** [Akshaya College of Engineering and Technology]

**Department:** [BE.Computer Science and Engineering]

**Date of Submission:** [Insert Date]

**Github Repository Link:** [Update the project source code to your Github Repository]

### **1. Problem Statement**

* ***Revisiting the initial problem statement:*** *Initially, the goal might have been broadly defined as analyzing traffic accidents.(e.g., weather, road conditions, time of day) and can focus specifically on predicting accident severity.*
* ***Refining the problem:*** *With this deeper understanding, you define the problem more precisely as a classification task—predicting the severity level of an accident based on various influencing factors.*
* *Accurately predicting accident severity enables faster emergency response and informed infrastructure planning. This leads to reduced fatalities, improved traffic safety, and more efficient resource allocation.*

### **2. Project Objectives**

* *The key technical objective is to develop a machine learning classification model that predicts the severity of traffic accidents using relevant features such as weather, road conditions, and time. The model aims to achieve high predictive accuracy and interpretability to support real-world decision-making for emergency response and urban planning. After exploring the dataset, the focus has evolved to emphasize feature selection and model explainability to ensure practical applicability in real-time scenarios.*

### **3. Flowchart of the Project Workflow**

### 

### **4. Data Description**

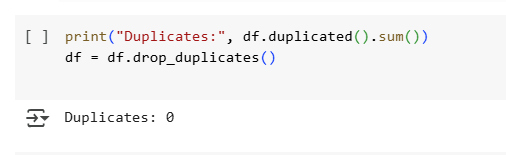
* ***Dataset name and origin: Dataset Name and Origin:*** *Traffic Accident Dataset" from Kaggle, or "Road Safety Dataset" from UCI Machine Learning Repository.*
* ***Type of data:***
* ***Structured:*** *This typically involves tabular data with columns and rows.*
* ***Time-series:*** *If your dataset includes timestamps or records over time for analysis of accident trends.*
* ***Unstructured:*** *If there are text descriptions of accidents or images (e.g., accident scene photos).*
* ***Image:*** *If you are using accident images to analyze vehicle damage, road conditions, or other aspects.*
* ***Text:*** *If there are textual descriptions of incidents or accident reports.*
* ***Number of records and features:*** *Example: 100,000 records with 20 features (e.g., accident type, location, weather conditions, time of day, etc.).*

*This number could vary depending on the dataset you're using.*

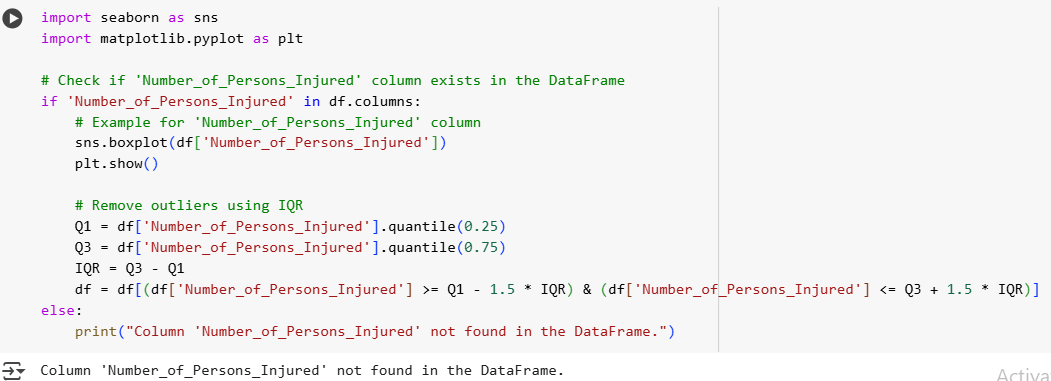
* ***Static or dynamic dataset:***
* ***Static:*** *If the dataset is fixed (e.g., historical data from a past year or specific period).*
* ***Dynamic:*** *If the dataset is continuously updated (e.g., real-time accident data from sensors or traffic cameras).*
* ***Target variable (if supervised learning):*** *Accident severity (minor, moderate, severe) Probability of an accident occurring in a certain area or time frame. Prediction of accident frequency. Risk of injury or fatality based on accident type and conditions*

### **5. Data Preprocessing**

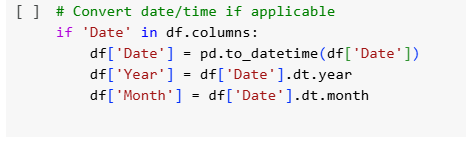
* ***Handle missing value:*****
* ***Remove or justify duplicate records:***

******

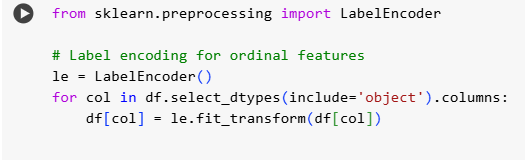
* ***Detect and treat outliers:***

**

* ***Convert data types and ensure consistency****:*

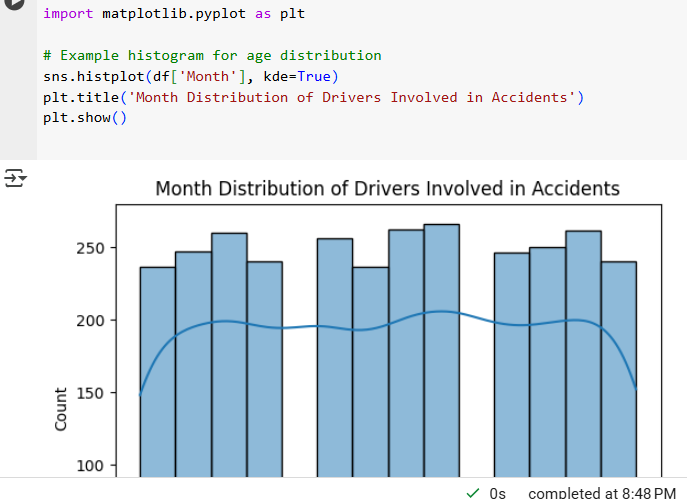
**

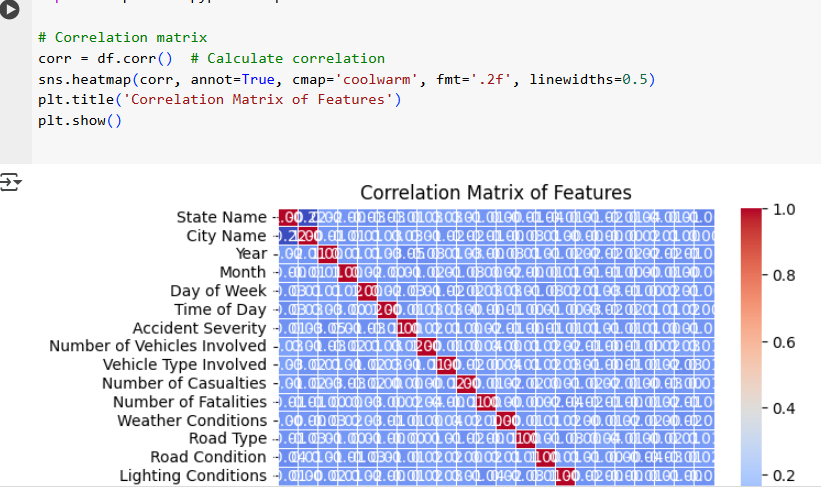
* ***Encode categorical variables (label encoding, one-hot encoding):***

**

### 

### **6. Exploratory Data Analysis (EDA)**

* ***Univariate Analysis:*****

***Bivariate/Multivariate Analysis:*** **

* ***Insights Summary:****Higher speed adverse weather conditions, and time of day strongly influence accident severity, with higher speeds and bad weather increasing risk. Age, road type, and traffic volume also play key roles in predicting accident occurrence and severity.*

### **7. Feature Engineering**

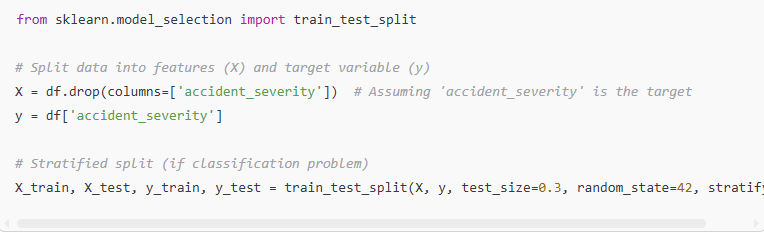
* ***Time of Day Features:*** *Feature: Extract hour of the day from timestamps to understand the time-specific patterns in accidents (e.g., rush hours or night-time accidents).*
* ***Irrelevant Features:*** *Removal: If certain features like accident ID or incident location (if irrelevant) are present, consider removing them as they do not directly impact the model's predictive power.*
* ***Principal Component Analysis (PCA):*** *Feature: Apply PCA on continuous features like speed, traffic volume, time of day, etc., to reduce the dimensionality of correlated features.*

### **8. Model Building**

### ***Logistic Regression (for Classification):*** *Logistic Regression is a simple yet effective model for binary classification problems (e.g., accident vs. no accident or severe vs. non-severe accidents). It provides probabilities, which can be useful in making decisions based on predicted risk.*

### ***Random Forest Classifier (for Classification):*** *Random Forest is a powerful ensemble method that works well for handling complex, non-linear relationships in data. It can capture interactions between features and is less prone to overfitting than individual decision trees.*

* ***Split data into training and testing sets:***

**

* ***Train models:*** *Accuracy: Measures the overall correctness of the model.*
* *Precision: The ability of the model to correctly identify positive outcomes.*
* *Recall: The model’s ability to identify all positive outcomes.*
* *F1-score: The harmonic mean of precision and recall, useful when dealing with imbalanced classes.*

### **9. Visualization of Results & Model Insights**

* ***Interpretation:*** *True Positives (TP) and True Negatives (TN) indicate correct predictions.False Positives (FP) and False Negatives (FN) show where the model misclassifies.*

*Plots True Positive Rate vs. False Positive Rate at various thresholds. AUC (Area Under Curve) measures overall classification performance.*

* ***Feature Importance Plot (Random Forest):*** *High-importance features (e.g., speed, weather, time of day) contribute significantly to accident predictions.*
* *Helps in understanding the driving factors behind the model’s decisi**cion*

### **10. Tools and Technologies Us**

### ***Programming Language:***

### ***Python –*** *Used for data preprocessing, feature engineering, model building, and evaluation.*

### ***IDE / Notebook Environment:***

### ***Google Colab –*** *Cloud-based Jupyter environment ideal for running notebooks without local setup.*

### ***Libraries Used:***

### ***pandas –*** *For data loading, cleaning, and manipulation.*

### ***numpy*** *– For numerical operations.*

### ***matplotlib & seaborn*** *– For plotting univariate, bivariate, and multivariate visualizations.*

### ***scikit-learn*** *– For model building, evaluation, and preprocessing tasks (e.g., train/test split, metrics).*

### **11. Team Members and Contributions**

* + *Data cleaning: Tamilmalar.R*
  + *EDA: Lakshmi Prasanna.S*
  + *Feature engineering: Lakshmi Prasanna.S*
  + *Model development: Sarojini devi*
  + *Visualization &map; Interpretation -Thisha Durga*