**Phase-2**

**Student Name:** NARESH P

**Register Number:** 510623106031

**Institution:** C ABDUL HAKEEM COLLEGE ENGINEERING AND TECHNOLOGY, MELVISHARAM

**Department:** ECE

**Date of Submission:** 08-05-2025

**Github Repository Link:** <https://github.com/naresh905/NARESH.git>

# 1. Problem Statement

Road traffic accidents remain a major global concern, causing significant loss of life and property. Using historical accident data, this project focuses on analysing patterns and building predictive models to anticipate accident-prone zones or conditions. This is a **classification problem**, where the objective is to predict the likelihood or severity of an accident occurring under specific conditions (weather, road type, time of day, etc.). Solving this problem can aid traffic authorities in proactively managing resources and improving road safety through data-driven decision-making.

# 2. Project Objectives

 to pre-process and analyse historical traffic accident data for trends and patterns.

 to develop and compare machine learning models that can predict the occurrence or severity of accidents.

 to identify key factors contributing to accidents and derive actionable insights.

 the primary goal is achieving a balance between model **accuracy** and **interpretability** to facilitate real-world applications in traffic management systems.

**3. Flowchart of the Project Workflow**

Data Collection

Accident Reports, Weather

| Traffic, Geo Data, etc.)

Data Pre-processing

-Handle Missing Values

Remove Duplicates |

| - Convert Date-Time Fields |

| - Feature Extraction

Exploratory Date Analysis

Visualize Hotspots

| - Analyze Time Patterns |

| - Identify Key Features

Feature Engineering

Extract Hour, Day, Month

| - Encode Categorical Data |

| - Normalize/Scale Data

# 4. Data Description

* **Dataset Source**: ( <https://www.kaggle.com/datasets/willianoliveiragibin/key-factors-traffic-accidents> )
* **Type of Data**: Structured (CSV format)
* **Records and Features**: ~300,000 rows with ~50 features (depending on dataset used)
* **Static or Dynamic**: Static
* **Target Variable**: Accident Severity (e.g., Minor, Serious, Fatal)

# 5. Data Preprocessing

 handled missing values using imputation (mean/mode) and dropped rows with excessive nulls.

 Removed duplicate records.

 Converted date-time fields and extracted useful features (hour, day, month).

 Applied label encoding for ordinal features and one-hot encoding for nominal categories.

 Normalized numerical features to ensure balanced input for models.

# 6. Exploratory Data Analysis (EDA)

 **Univariate Analysis**: Histogram of accident severity, time-of-day distribution, day-of-week trends.

 **Bivariate/Multivariate Analysis**: Correlation heatmaps, pairplots between accident severity and road/weather conditions.

** Insights:**

* Most accidents occur during rush hours.
* Poor lighting and wet road surfaces increase accident severity.
* Urban areas report more frequent but less severe accidents.

# 7. Feature Engineering

 Extracted “peak hours” and “weekend” from timestamp data.

 Created severity risk score from weather, road type, and vehicle count.

 Applied PCA (optional) to reduce dimensionality and noise.

 Removed redundant features with high correlation (>0.9) to avoid multicollinearity.

# 8. Model Building

 Models Used: **Random Forest**, **Logistic Regression**

 Justification: Random Forest for handling complex interactions; Logistic Regression for baseline performance and interpretability.

 Train/Test Split: 80/20 with stratification.

 Evaluation Metrics:

* **Accuracy**: Random Forest – 87%, Logistic Regression – 78%
* **F1-Score**: Random Forest – 0.84
* Feature importance showed weather and light conditions among top predictors.

# 9. Visualization of Results & Model Insights

** Confusion Matrix**: Displayed true vs predicted accident severity.

** ROC Curve**: AUC for Random Forest: 0.91

 Feature **Importance**: Plotted top 10 features (e.g., weather, speed limit, time)

 **Interpretation**: Weather and time of day significantly influence accident risk. Nighttime driving under poor visibility increases accident likelihood.

# 10. Tools and Technologies Used

** Programming Language**: Python

** IDE**: Google Colab

** Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost

** Visualization Tools**: Plotly, seaborn, matplotlib

# 11. Team Members and Contributions

|  |  |  |  |
| --- | --- | --- | --- |
| **TEAM MEMBERS** | | **ROLES** | **RESPONSIBILITY** |
| Shahid Mohammed A | | Data Finder and Organizer | Collected traffic accident datasets, organized the data properly, and made sure it was ready for analysis. |
| Pugazhenthi S | | Design and Visualization Handler | Created graphs, charts, and visuals to show accident predictions and made the project easy to understand. |
| Jashwanth A | | Accident Trend Analyzer | Came up with the main ideas for applying AI to traffic accident prediction and helped shape the overall direction of the project. |
| Mithesh R | | Project presenter and Complier | Compiled the full project report, arranged all the sections neatly, and prepared  it for submission and presentation. |
| NARESH P | Accident Trend Analyaer | | Studied accident patterns, traffic conditions, and helped find what factors mostly cause road accidents. |