**STUDENT PROJECT REPORT**

Student Name : M.VINOTHINI

Register Number : 623323104054

Institution : VETRI VINAYAHA COLLEGE OF ENGINEERING

AND TECHNOLOGY

Department : COMPUTER SCIENCE AND ENGINEERING

Date of Submission : **07 – 05 – 2025**

**Problem statement**

Road accidents are a major cause of death and injury, especially in countries like India. Every year, countless lives are lost due to preventable factors such as over-speeding, poor weather, or unsafe road conditions. While data exists, we still lack intelligent systems that can analyze and predict high-risk situations before they happen.

**In this project, we aim to answer the question:**

"Can we use past accident data and machine learning to predict whether certain road conditions are likely to cause accidents?"

**Why is this problem important?**

Because if we can successfully predict accident-prone scenarios, it opens up real-world applications such as:

* Alerting drivers in real-time using navigation systems,

* Helping city planners identify and fix accident hotspots,

* Guiding law enforcement in deploying resources more effectively,

* Ultimately saving lives by preventing accidents before they occur.

**Objectives**

The goal of the project is to build a machine learning model that can predict the likelihood of road accidents based on factors like location, time, weather, and road conditions.

**Our key technical objectives are:**

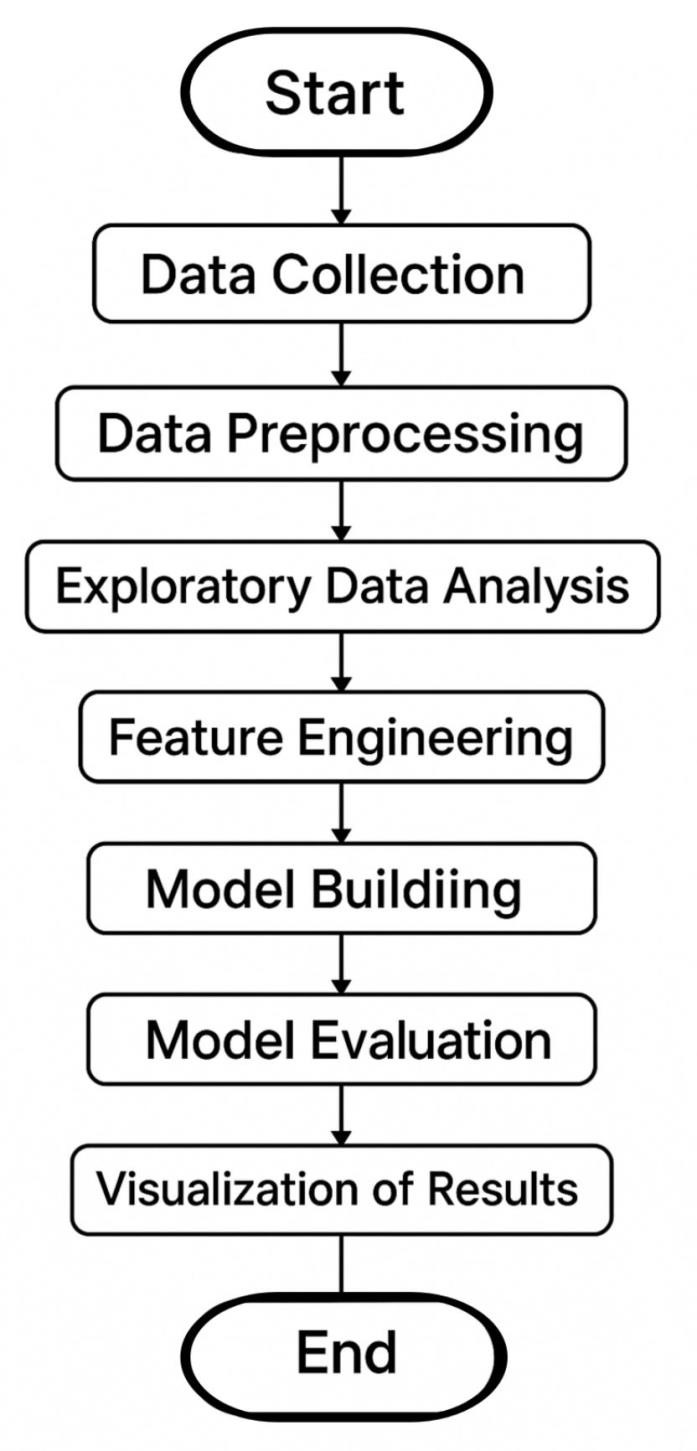
* To analyze patterns in historical accident data using EDA.

* To build and train a classification model that balances both accuracy and interpretability.

* To identify high-risk conditions or scenarios that contribute to accidents.

* To present insights in a clear, visual format that could be useful for decisionmakers.

**Flowchart of the Project Workflow :**



**Data description :**

For this project, we are using a publicly available Traffic Accident Dataset sourced from Kaggle. The data is structured and contains detailed records of road accidents, including time, location, weather, road surface conditions, and many other details.

Source: Kaggle (public dataset)

Type of Data: Structured, tabular

Number of Records: ~12,000+ accident cases

Number of Features: Around 15 columns (features)

Nature of Dataset: Static (downloaded once)

Target Variable: Accident occurrence

**Data Preprocessing :**

To prepare the dataset for modeling, several preprocessing steps were carried out to ensure clean and consistent data:

**Handling Missing Values:**

Columns with too many missing values were dropped. For others, we used mean/median imputation for numerical fields and mode imputation for categorical ones.

**Duplicate Records:**

Duplicate rows were identified and removed to avoid bias or redundancy in the model.

**Outlier Detection & Treatment:**

Used boxplots and statistical methods (like IQR) to detect outliers. Outliers were either capped or removed based on their impact.

**Data Type Conversion:**

Ensured each column had the correct data type

**Encoding Categorical Variables:**

Used Label Encoding for binary categories.

Applied One-Hot Encoding for multi-class categorical fields.

**Feature Scaling:**

Applied standardization (z-score) for continuous features to ensure uniformity across all model inputs.

**Exploratory Data Analysis (EDA) :**

We performed a detailed analysis to understand the structure, patterns, and relationships in the dataset.

**Univariate Analysis :**

Plotted histograms and boxplots to study the distribution of numerical features like speed, time, and weather conditions.

Used countplots for categorical variables such as accident severity, day of the week, and road type.

**Bivariate & Multivariate Analysis :**

Used correlation matrix to identify relationships between numerical features. Pairplots and scatterplots helped visualize interactions like time vs. severity or speed vs. impact.

Grouped bar plots revealed how weather or lighting conditions affect accident severity.

**Relationship with Target Variable :**

Severity levels were found to be strongly influenced by factors like weather, road surface, and time of day.

Accidents were more frequent during peak hours and under poor visibility conditions.

**Insights Summary :**

Weather, lighting, road type, and time are key predictors.

Some features showed clear separation in severity classes, which will help improve model accuracy.

**Feature Engineering :**

To enhance model performance, several new features were created and existing ones refined based on EDA insights and domain knowledge.

**Time Features:**

Extracted hour, day of week, and month from timestamp to better capture accident patterns during different times.

**Weather Categorization:**

Grouped similar weather conditions (e.g., light rain, heavy rain) into broader categories to reduce noise.

**Speed Binning:**

Converted continuous speed values into binned categories (e.g., low, moderate, high) to simplify interpretation.

**Day/Night Indicator:**

Created a binary feature to mark whether the accident occurred during day or night based on light conditions.

**Location Simplification:**

Simplified or grouped location-based fields (like urban vs. rural) to reduce sparsity.

Each feature was added with a clear goal , either to improve model learning, reduce complexity, or add context. Irrelevant or redundant columns were dropped to avoid overfitting.

**Model Building :**

To address the classification nature of our problem (predicting the severity or likelihood of an accident), we selected and implemented a machine learning model:

**Random Forest –** Selected for its ability to handle complex, non-linear relationships and provide better performance through ensemble learning.

**Data Splitting:**

The dataset was divided into training (80%) and testing (20%) sets using stratified sampling to maintain class balance.

**Training and Evaluation:**

The model is trained and evaluated using classification metrics:

* Accuracy
* Precision
* Recall
* F1-score

**Visualization of Results & Model Insights :**

To better understand model performance and interpretability, the following visualizations were used:

**Confusion Matrix:**

Showed the count of true positives, false positives, true negatives, and false negatives , helping us identify where the model performs well or struggles.

**ROC Curve:**

Visualized the trade-off between true positive rate and false positive rate. AUC scores indicated strong model performance.

**Feature Importance Plot (Random Forest):**

Helped us identify which features (e.g., weather condition, road type, time of day) had the most impact on predictions.

**Classification Report Bar Chart:**

Displayed precision, recall, and F1-score for each class visually for better comparison.

We used random forest model which offered clearer insights into which factors most influence accident outcomes.

**Tools and Technologies :**

**Programming Language:**

Python – chosen for its simplicity, rich ecosystem in data science and machine learning libraries.

**IDE/Notebook:**

Google Colab – used for its ease of collaboration and access to GPU resources.

**Libraries Used:**

* **pandas, numpy** – for data handling and numerical operations.
* **matplotlib, seaborn, plotly** – for visualizations.
* **scikit-learn, XGBoost** – for model building and evaluation.

**Team members and roles :**

* **S.V.SATHYA –** Project coordinator , EDA , machine learning model developer.

* **B. SWEDHA –** Data collector and preprocessing lead, Documentation & Presentation Lead.

* **M.VINOTHINI –** Data analyst , Visualization & Report Designer, feature engineer.