**Phase-2 Submission**

**Student Name:** Nanthini M

**Register Number:** 712523104039

**Institution:** PPG Institute of Technology

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

**Date of Submission:** 30/04/2025

**Github Repository Link***:* [*https://github.com/M-NANTHINI-2004/NM\_NANTHINI\_DS*](https://github.com/M-NANTHINI-2004/NM_NANTHINI_DS)

**Project Title:Enhancing road safety with AI-driven traffic acciident analysis and prediction**

### **1.Problem Statement**

*Road accidents are a leading cause of injury and death globally. Identifying patterns in accident data can help predict high-risk scenarios and enable proactive safety measures. With the advent of AI, there's an opportunity to leverage historical traffic accident data to anticipate accidents and mitigate their impact.*

* ***Refined Problem****: Given a dataset of historical traffic accidents, predict the severity of future accidents or identify accident-prone zones/times based on environmental, temporal, and demographic factors.*
* ***Type of Problem****:* ***Classification*** *(e.g., predicting severity level) and* ***Regression*** *(e.g., predicting the number of injuries/fatalities).*
* ***Relevance****: This work supports smart city initiatives, reduces emergency response times, and can save lives by informing road safety policies and driver alerts.*

### **2. Project Objectives**

* ***Technical Objectives****:*
  + *Clean and preprocess real-world traffic accident data.*
  + *Perform EDA to uncover patterns and relationships.*
  + *Engineer features that improve model predictability.*
  + *Build and evaluate machine learning models for accident severity prediction.*
  + *Interpret and visualize model results for actionable insights.*
* ***Expected Outcomes****:*
  + *A predictive model with high classification accuracy or low regression error.*
  + *Clear identification of key features influencing accident severity.*
  + *Visual dashboard to interpret insights (if time permits).*
* ***Evolution****:*
  + *After data exploration, we expanded from severity prediction to also explore temporal and geographic accident patterns.*

### 

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

*Data collection*

*Data Preprocessing*

*Exploatory Dada Analysis(EDA)*

*Feature Engineering*

*Model Selection*

*Model training &Evaluation*

*Model Interpretation & Visualization*

*Final Reporting*

### **4. Data Description**

* ***Source****: Kaggle -* [*https://www.kaggle.com/datasets/ankushpanday1/global-road-accidents-dataset?resource=download*](https://www.kaggle.com/datasets/ankushpanday1/global-road-accidents-dataset?resource=download)
* ***Type****: Structured data (CSV format)*
* ***Records****: ~1.5 million records*
* ***Features****: 30+ features (weather conditions, timestamp, location, etc.)*
* ***Nature****: Static dataset*
* ***Target Variable****: Severity (for classification); Number\_of\_Casualties (for regression)*

### 

### **5. Data Preprocessing**

* ***Missing Values****:*
* *Columns like Weather\_Condition and Visibility had missing values; imputed using mode or mean.*
* ***Duplicates****: Identified and removed ~2% duplicate rows.*
* ***Outliers****: Detected via boxplots and removed/treated (e.g., extremely high casualty values).*
* ***Data Types****: Converted timestamp columns to datetime; ensured categorical fields used correct data types.*
* ***Encoding****:*
* *Label encoding for binary categories (e.g., Day/Night).*
* *One-hot encoding for multi-class features (e.g., Weather\_Condition, Road\_Type).*
* ***Normalization****:*
* *Applied StandardScaler to numeric features like Speed\_limit, Temperature.*

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

* ***Univariate Analysis****:*
  + *Severity distribution: Majority were minor injuries.*
  + *Histogram for time: Spike in accidents during peak hours.*
* ***Bivariate Analysis****:*
  + *Heatmap showed correlation between low visibility and higher severity.*
  + *Accidents increase in rain/snow (compared using grouped bar plots).*
* ***Key Insights****:*
  + *Time of day, weather, and road type significantly influence accident severity.*
  + *Urban areas have higher accident frequency but often lower severity due to reduced speed.*

### 

### **7. Feature Engineering**

* *Created new features:*
  + *Hour\_of\_Day, Day\_of\_Week, Is\_Weekend, Is\_Rush\_Hour*
  + *Weather\_Severity\_Score based on conditions*
* *Used binning:*
  + *Speed limits binned into categories: Low, Medium, High*
* *Applied PCA (optional):*
  + *Reduced high-dimensional weather features to top 2 components*

***Justification****: These features improved model performance and interpretability.*

### **8. Model Building**

* ***Models Used****:*
  + *Logistic Regression (baseline)*
  + *Random Forest Classifier (for classification)*
  + *Gradient Boosting / XGBoost (for performance comparison)*
* ***Data Split****:*
  + *80-20 training/testing split with stratification on severity*
* ***Metrics Used****:*
  + ***Classification****: Accuracy, Precision, Recall, F1-Score*
  + ***Regression****: RMSE, MAE, R² Score*

| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1-Score*** |
| --- | --- | --- | --- | --- |
| *Logistic Regression* | *68%* | *0.65* | *0.66* | *0.65* |
| *Random Forest* | *82%* | *0.81* | *0.80* | *0.80* |
| *XGBoost* | *85%* | *0.84* | *0.83* | *0.83* |

### 

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

* *[****Confusion Matrix****: Highlighted misclassifications in moderate severity class.*
* ***Feature Importance (Random Forest/XGBoost)****:*
  + *Top features: Weather\_Condition, Speed\_limit, Time\_of\_Day, Road\_Surface\_Conditions*
* ***ROC Curve****:*
  + *AUC: 0.88 (XGBoost)*
* ***Residual Plot*** *(for regression): Residuals showed no strong pattern, indicating low bias.*

### **10. Tools and Technologies Used**

* ***Language****: Python*
* ***IDE****: Jupyter Notebook (Google Colab)*
* ***Libraries****:*
  + *Data Handling: pandas, numpy*
  + *Visualization: matplotlib, seaborn, plotly*
  + *Modeling: scikit-learn, xgboost, lightgbm*
  + *Optional Dashboards: Tableau or Power BI*

### 

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

|  |  |  |
| --- | --- | --- |
| *Name* | *Role* | *Work description* |
| *Dhavaroshni A* | *Tean learder* | *Data Cleaning, Preprocessing, and EDA* |
| *Charumathi J* | *Team member 1* | *Feature Engineering and Model Training* |
| *Nanthini M* | *Team member 2* | *Visualization, Interpretation, and Reporting* |
| *Sharvesh S* | *Team member 3* | *Model Evaluation and Hyperparameter Tuning* |