Project Documentation — Road Accidents: -

Project Design:

Problem Definition

- This project focuses on analyzing the UK Road Traffic Accidents dataset, sourced from Kaggle, which contains detailed accident records collected through the STATS19 reporting system. Each record provides information including accident location (coordinates and district), time and date, road characteristics (type, speed limit, junction details), environmental conditions (weather, lighting, road surface), vehicles involved, and casualties. The aim is to clean, preprocess, and explore this dataset to prepare it for predictive modeling, identifying patterns and relationships in the data. The analysis will support subsequent objectives such as severity prediction, hotspot identification, and understanding environmental and temporal factors affecting accidents.

Objectives

- Predict accident severity (Slight, Serious, Fatal) using historical data.
- Identify key contributing factors (speed limit, road type, weather, lighting, time).
- Detect high-risk accident hotspots.
- Analyze time-based risk patterns (day of week, time of day).
- Assess environmental impact on accident severity.
- Provide actionable recommendations for traffic safety improvement

System Components

1. Data Preprocessing
2. Exploratory Data Analysis (EDA)
3. Model Training
4. Model Evaluation
5. User Interface

Input fields as Tabular data are:

Dataset Features (Accident Road Prediction):

- Accident_Index (Text)
- Location_Easting_OSGR (Numeric)
- Location_Northing_OSGR (Numeric)
- Longitude (Float)
- Latitude (Float)
- Police_Force (Categorical)
- Accident_Severity (Categorical)
- Number_of_Casualties (Numeric)
- Date (Date)
- Day_of_Week (Categorical)
- Time (Time)
- Local_Authority_(District) (Categorical)
- Local_Authority_(Highway) (Categorical)
- 1st_Road_Number (Numeric)
- Road_Type (Categorical)
- Speed_limit (Numeric)
- Junction_Detail (Categorical)
- Junction_Control (Categorical)
- 2nd_Road_Class (Categorical)
- 2nd_Road_Number (Numeric)
- Pedestrian_Crossing-Human_Control (Categorical)
- Pedestrian_Crossing-Physical_Facilities (Categorical)
- Light_Conditions (Categorical)
- Weather_Conditions (Categorical)
- Road_Surface_Conditions (Categorical)
- Special_Conditions_at_Site (Categorical)
- Carriageway_Hazards (Categorical)
- Urban_or_Rural_Area (Categorical)
- Did_Police_Officer_Attend_Scene_of_Accident (Categorical)
- LSOA_of_Accident_Location (Categorical)
- Year (Numeric)

Exploratory Data Analysis (EDA)

- Data Cleaning
  - Handling missing values
  - Drop duplicate
  - Formatting
  - Handling outlier
- Data Analysis
  - Information about data
  - Descriptive statistics
- Data Visualization
  - Understand questions to make answers using charts
  - Charts --> heatmaps, histogram, pie chart, bar charts
  - Using Power BI to make Dashboard

Machine Learning Model

Data Preprocessing

- Handling missing values
- Standardization of numerical features
- Encoding of categorical features
- Train-test split: 80% training / 20% testing

Algorithms Tested

- **LightGBM**
- **XGBoost**

Tutorials / Walkthrough How to Run the Project

Download the project from GitHub.

Install dependencies: pip install NumPy pandas scikit-learn matplotlib seaborn

Open the Jupiter Notebook or Python script.

Enter values.

Receive prediction: Accident Servetary

User Interface Explanation

User enters Accidents parameters.

Clicks “Predict.”

System shows probability of road Accident with simple output.

Final Report Summary

Results

Accuracy: 0.89

F1-score: strong performance for Road Accident

ROC AUC: ~

Challenges

- The dataset contains 1.4 million records, which made it difficult to apply certain imbalance-handling techniques such as SMOTE or oversampling methods due to high computational cost.
- Our severity target includes three classes (Slight, Serious, Fatal), but the distribution is strongly unbalanced. This large difference between classes made it challenging for the model to

learn the minority classes effectively.

- We attempted multiple solutions — including class weights, sampling techniques, and algorithm adjustments — but none of them produced stable results because of the huge data size and extreme imbalance.
- The imbalance directly affected the model's accuracy and its ability to generalize, especially for predicting Serious and Fatal accidents.

Project link:

https://github.com/mariammohamed-cell/Graduation-Project-DEPI

Team Members & roles:

Name	Role
Mariam Mohamed Ibrahim	Milestone (1,2,3, 4) Team Leader
Mariam Reda ElSayed	Milestone (1,2,4)
Amal Nageh Esmail	Milestone (1,2,3)
Karim Hamada Mohamed	Milestone (1,2,5)
Ahmed Walied Ahmed	Milestone (1,2,5)
Rahaf Mohamed Adel	