NAME: HEER

ROLL-NUMBER:22I-2371

SECTION : A

SUBMITTED TO: DR BASHARAT HUSSAIN

ANO: 1

MACHINE LEARNING…

LINK TO VIDEO: https://youtu.be/EpOd0LIo_4c

1. Dataset Overview

Domain of the Dataset

The dataset belongs to the transportation domain, specifically focusing on traffic accidents. It includes detailed information about accidents such as severity, location, number of vehicles involved, casualties, and road conditions. This dataset is crucial for analyzing road safety and improving traffic management strategies.

Target Variable

The target variable is Accident_Severity, which categorizes accidents into:

- 1: Minor accident
- 2: Serious accident
- 3: Fatal accident

Since the target variable is categorical, this is a classification problem.

Dataset Characteristics

- **Number of Features**: 33 (Mix of numeric, categorical, and datetime variables)
- **Number of Records**: 1,504,150
- **Type of Problem**: **Classification**

READING DATASET

thus is the link:
https://drive.google.com/file/d/1sTiyFDyWmj66paV1mWhrMgXfJMZLugMp/view?usp=sharing

Task 2: Exploratory Data Analysis (EDA) (15 Marks)

1. Perform data visualization: o Histograms for numeric features. o Scatter plots & correlation matrix. o Boxplots to identify outliers.
2. Identify missing values and handle them appropriately.
3. Identify Outliers as well
4. Identify and discuss important features.

2. Exploratory Data Analysis (EDA)

Data Visualization & Key Insights

Histograms for Numeric Features:

- **Number of Vehicles & Casualties:** Highly skewed, with most accidents involving fewer vehicles and casualties.
- **Accident Severity:** Severity 3 (Fatal Accidents) appears more frequent than Severity 1 (Minor Accidents).
- **Speed Limit:** Most accidents occurred at 30 km/h, followed by 50 km/h (common speed limits in urban areas).

Scatter Plots & Correlation Analysis:

- **Number of Vehicles vs. Number of Casualties:** Weak positive correlation. Some outliers indicate severe pile-ups.
- **Speed Limit vs. Number of Casualties:** Weak correlation. Higher speeds sometimes lead to more casualties, but other factors (road conditions, weather) play a role.

- **Year vs. Number of Vehicles:** No clear trend over time, but outliers indicate significant accidents in certain years (e.g., 2006, 2012, 2014).
- **Accident Severity vs. Number of Vehicles:** Minor and serious accidents mostly involve 1-20 vehicles. Fatal accidents (Severity 3) involve more vehicles but with high variance.
- **Latitude vs. Longitude (Geographic Distribution of Accidents):** Higher concentration in urban areas, particularly around London, indicating a relationship between population density and accident frequency.

Correlation Matrix (Heatmap):

- **Latitude & Longitude:** Strong correlation (0.96) as expected.
- **Accident Severity & Number of Casualties:** Weak correlation (0.24).
- **Number of Vehicles & Number of Casualties:** Moderate correlation (0.24), indicating that multi-vehicle accidents tend to have more casualties.

Missing Value Treatment

Column Name	Missing Values (%)	Treatment
Junction Control	40.08%	Dropped (High missing % and potential bias)
Special Conditions at Site	97.57%	Dropped (Too many missing values)

Carriageway Hazards	98.19%	Dropped (Irrelevant due to extreme missing values)
Longitude & Location Easting OSGR	0.67%	Imputed with median
Time	0.77%	Imputed with mode
Pedestrian Crossing Features	<1%	Imputed with mode
LSOA of Accident Location	7.2%	Imputed with mode

Outlier Identification & Treatment

- **Number of Casualties**: Outliers (values exceeding 60) were detected. These represent major accidents and are retained for model learning.
- **Number of Vehicles**: Most accidents involve <10 vehicles, but some exceed **40-60** (outliers retained).
- **Speed Limit**: Common values (30, 50, 70), but extreme values checked for errors.

Feature Importance Analysis

- **Highly Important Features:**

    - **Accident Severity:** Directly influenced by casualties, speed limits, number of vehicles, road type.
    - **Number of Casualties:** Major indicator of accident impact.
    - **Speed Limit & Urban/Rural Area:** Determines accident severity risk.
    - **Geographical Features (Latitude, Longitude):** Identifies high-risk accident zones.

3. Data Preprocessing & Feature Engineering

Missing Value Handling

- High-missing columns dropped.
- Low-missing values imputed using mode or median.

Encoding Categorical Features

- One-Hot Encoding: Applied to nominal categorical features.
- Label Encoding: Applied to ordinal categorical features.

Feature Scaling

- **MinMax Scaling** applied to numerical features.

Feature Engineering

- **No additional features were created**.

4. Model Selection & Training

Models Trained:

- **Decision Tree Classifier**
- **Logistic Regression**
- **SGD Classifier**

Evaluation Metrics Used:

- **Classification Metrics:** Accuracy, Precision, Recall, F1-score.
- **Cross-validation (`cv=5`)** applied for better evaluation.

5. Hyperparameter Tuning & Model Performance

Post-Tuning Model Performance:

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.6170	0.7916	0.6170	0.6886
SGD Classifier	0.0656	0.7438	0.0656	0.0997

Logistic Regression	0.8510	0.7725	0.8510	0.7828

LINK TO VIDEO: https://youtu.be/EpOd0LIo_4c