ML Assignment 3

1. Use the dataset provided: Road Accident Data.csv
2. Each group must explore it thoroughly to understand and explain the domain context.
3. Have a well-documented **Jupyter Notebook** (.ipynb) containing:
    a. Clean and commented code
    b. Explanatory text (markdown cells) for each section
4. **TO BE SUBMITTED ON OR BEFORE 31/07/2025**

Assignment Tasks

Part 1: Data Understanding, Preprocessing and Feature Engineering

1. **Dataset Overview**

    - Load the dataset and give a clear summary.
    - Report the number of rows and columns.
    - Describe the data types and sample values.
    - Identify the target variable(s) for different tasks (e.g., accident severity, number of casualties, etc.)

2. **Data Cleaning**

    - Check and handle missing values (explain method used).
    - Detect and deal with duplicate records.
    - Convert categorical variables appropriately (label encoding, one-hot encoding, etc.).

3. **Exploratory Data Analysis (EDA)**

    - Plot distributions of numeric variables.
    - Visualize relationships between various features
    - Compute correlation matrix and visualize it.
    - Identify any anomalies or outliers.

4. **Feature Selection**

    - Use at least two different techniques (e.g., correlation, mutual information, tree-based importance) to select important features.

Part 2: Modeling

i. Develop **at least 5 different machine learning models**, covering different families of algorithms. You must include **one neural network model**.

ii. Use any relevant target variable (classification or regression, as appropriate). Apply 80/20 ratio in train/test split.
iii. Use **K-Fold Cross-Validation** and compare results from cross-validation vs. simple train/test split.

Part 3: Model Evaluation

5. **Classification Metrics (for classification tasks):**
   - Accuracy, Precision, Recall, F1-score, AUC-ROC
   - Confusion matrix (plot heatmap)

6. **Regression Metrics (if applicable)**

   - MAE, RMSE, R^2 score

7. **Visualization**

   - ROC curves (for classification models)
   - Predicted vs. actual plots
   - Feature importance bar charts (where applicable)

8. **Model Comparison**

   - Compare all five models based on performance and training time.
   - Summarize their strengths and weaknesses on this dataset.

9. Conclusion

   - Summarized findings and recommendations