Assignment: End-to-End Machine Learning Pipeline

Objective

Apply everything you have learned so far to build a complete machine learning pipeline — from raw data to model evaluation.

Dataset

Download the dataset from Kaggle. Don't use any built-in library datasets, such as those from scikit-learn or seaborn.

◇ Assignment Tasks

1. Data Handling (NumPy & Pandas)

- Load dataset into a Pandas DataFrame.
- Perform initial checks.
- Handle missing values and duplicates.
- Convert categorical features into numerical form if needed.

2. Exploratory Data Analysis (EDA)

- Use NumPy & Pandas for basic statistics.
- Visualize data using:
  - Matplotlib / Seaborn
  - Plotly: at least one interactive plot (e.g., scatter or bar chart).

3. Feature Engineering

- Split dataset into features (X) and target (y).
- Normalize/scale data if necessary.
- Perform train-test split.

4. Model Training

- Train the following models:
- - KNN Classifier
- - Decision Tree Classifier
- - Random Forest Classifier
- Compare baseline results.

5. Feature Importance

- Extract and visualize feature importance from Random Forest.
- Discuss which features contribute most to predictions.

6. Hyperparameter Tuning

- Use RandomizedSearchCV to optimize hyperparameters:
- - KNN → n_neighbors, weights, metric
- - Decision Tree → max_depth, min_samples_split
- - Random Forest → n_estimators, max_depth, min_samples_split
- Compare default vs tuned models.

7. Model Evaluation

- Evaluate models using:
- - Accuracy
- - Precision, Recall, F1-score
- - Confusion Matrix
- Plot ROC Curve for the best-performing model.

8. Conclusion

- Which model performed best and why?
- Which features were most important?
- How did hyperparameter tuning improve results?

Deliverables

1. Jupyter Notebook with well-commented code and results.
2. Report (2–3 pages) summarizing:
- Dataset insights
- Visualization findings
- Model comparison table
- Key conclusions