Machine Learning Assignment

Rajkumar Bhandari

Objective

To understand fundamental machine learning concepts and apply basic models to real-world data, focusing on interpretability and bias-variance behavior.

Part A – Theory (Short Answer Questions)

1. Define the following:
   - Machine Learning
   - Supervised vs Unsupervised Learning
   - Overfitting and Underfitting
   - Training, Validation, and Test Sets
   - Cross-Validation
2. Explain the **bias-variance tradeoff** with a diagram and example.
3. Describe how the following models work, with advantages and limitations:
   - Linear Regression
   - Logistic Regression
   - Decision Tree
   - K-Nearest Neighbors (KNN)
4. Discuss:
   - One feature selection method (e.g., correlation, chi-square)
   - One feature scaling technique (e.g., normalization, standardization)
5. What is **regularization**? How do L1 and L2 regularization reduce overfitting?

Part B – Practical (Python)

Dataset: Use one of the following or your own dataset:

- Iris Dataset
- Titanic Dataset

- UCI Soil Moisture Dataset
- Your Smart Irrigation Dataset

Tasks:

1. **Data Preprocessing**
    - Load the dataset using Pandas
    - Handle missing values
    - Encode categorical variables (if any)
    - Apply feature scaling
    - Split the dataset (e.g., 80% train, 20% test)
2. **Model Training**
    - Linear Regression (for regression)
    - Logistic Regression (for classification)
    - Decision Tree
    - K-Nearest Neighbors (KNN)
3. **Model Evaluation**
    - For regression: MSE, RMSE, R^2 Score
    - For classification: Accuracy, Precision, Recall, F1-Score, Confusion Matrix
    - Present comparison in a table
4. **Bias-Variance Behavior**
    - Vary complexity: e.g., max_depth for Decision Tree, k value for KNN
    - Plot training and testing error curves
5. **Bonus (Optional)**
    - Apply k-Fold Cross Validation
    - Tune hyperparameters manually using for-loops

Submission Guidelines

- Submit a Jupyter Notebook (.ipynb) with well-commented code.
- Submit a separate PDF with Part A answers.
- Clearly mention dataset source and preprocessing steps.
- Include a short conclusion summarizing the best model and justification.