Advanced ML II (MESIIN595624)

Homework Assignment:

Comparative Study Between Ensemble Learning and Evolutionary Learning to solve the Higgs Bozon Detection

Objective

The purpose of this homework is to:

1. Implement and compare **Ensemble Learning** techniques (Bagging and Boosting) with **Evolutionary Learning** using DEAP.
2. Evaluate and analyze the performance of these approaches on the **Higgs Boson Detection** dataset.
3. Understand the strengths, weaknesses, and trade-offs between these methods.

Dataset

Use the Higgs Boson dataset from Kaggle:

- **Link**: Higgs Boson Machine Learning Data
- **Description**: The dataset contains features derived from particle collision events to classify whether an event is a signal (Higgs boson) or background noise.

I. Instructions for Students

1. Dataset Preprocessing

- Download and explore the Higgs Boson dataset.
- Perform necessary preprocessing steps (see TD2_Bozon Higgs detection NoteBook):
  - Handle missing values (if any).
  - Normalize or standardize the features.
  - Split the data into training and testing sets.
  - Optionally, use feature selection to reduce dimensionality.

2. Ensemble Learning Implementation

You must implement at least two ensemble learning techniques:

1. **Bagging**: You can use the **Random Forest** classifier from sklearn or an other technique.
2. **Boosting**: You can use the **XGBoost** or **Gradient Boosting** classifier from sklearn or another technique.

3. Evolutionary Learning Implementation with DEAP

- Define the problem as a **classification task** with DEAP.
- Implement a Genetic Programming-based classifier:

  - **Primitive Set**: Include basic operations (e.g., addition, subtraction, multiplication).
  - **Terminal Set**: Use the features of the dataset as inputs.
  - **Fitness Function**: Maximize classification accuracy or F1-score.

- Train two evolutionary models (*eaSimple* and *eaMuPlusLambda*) and evaluate them on the testing set.

4. Performance Metrics

Compare all methods using: Accuracy, Precision, Recall, F1-Score, Training Time, etc.

5. Comparative Analysis

Provide a detailed comparison between the methods:

1. **Performance**:
   - Which method achieved the highest accuracy and F1-score?
   - How do precision and recall vary across methods?
2. **Efficiency**:
   - Compare the training and testing times of the three approaches.
   - Discuss the computational cost of evolutionary learning compared to ensemble learning.
3. **Model Complexity**:
   - Analyze the interpretability of the solutions.
   - Discuss the complexity of the evolved models versus ensemble methods.
4. **Advantages and Disadvantages**:
   - Highlight the strengths and weaknesses of each approach based on the results.

I. Deliverables

1. **Code Implementation**:
   - Python scripts or notebooks for Bagging, Boosting, and Evolutionary Learning.
2. **Report**: Includes:
   - Performance metrics in a comparative table.
   - Plots of training/testing times and performance metrics.
   - A written analysis discussing the observations and findings.

Grading Criteria

Criteria	Weight
Data Preprocessing and Analysis	10%
Implementation of Methods	40%
Comparative Analysis and Report	40%
Code Clarity and Organization	10%