German International University of Applied Sciences
Informatics and Computer Science
Dr. Caroline Sabty
Eng. Aya Abdalla

Advanced Machine Learning, Spring 2024
Assignment 1
Due date is February 13, 2024 at 11:59 PM

In this project, you will explore and compare the performance of three popular ensemble learning models: Random Forest, AdaBoost, and Gradient Boost. The goal is to fine-tune these models on three distinct datasets, perform necessary preprocessing steps, and analyze the results to draw valuable insights.

Tasks

a) **Dataset Selection:** Choose three diverse datasets suitable for classification or regression tasks from those provided on cms.

b) **Data Preprocessing:** Perform thorough data preprocessing steps, including handling missing values, encoding categorical variables, handling null values, scaling numerical features, and any other necessary data cleaning procedures.(you have to apply at least 2 in each dataset)

c) **Model Implementation:** Implement Random Forest, AdaBoost, and Gradient Boost models using a suitable programming language (e.g., Python with scikit-learn). Ensure that the models are properly initialized.

d) **Hyperparameter Tuning:** Fine-tune the hyperparameters of each model using techniques such as grid search or random search. Optimize the models for the best performance on each dataset. (you have to fine tune at least two parameters)

e) **Training and Evaluation:** Train the tuned models on the respective datasets and evaluate their performance using appropriate metrics (accuracy, precision, recall, F1-score).

f) **Comparison Table:** Create a table that summarizes the accuracy for each ensemble model on each dataset.

g) **Visualization:**

Accuracy Comparison Bar Chart: Create a bar chart showing the accuracy of each ensemble model on different datasets. This should provide a quick visual comparison of their overall performance.

Grid Search Heatmap: Create a heatmap where the x-axis represents the hyperparameter values of one parameter, the y-axis represents the hyperparameter values of another parameter, and the color in each cell represents the performance metric.(select the highest accuracy model in each dataset and create the plot)

h) **Insights and Conclusions:** Report the results and draw insights from the comparison. Discuss the strengths and weaknesses of each ensemble model and how they perform on different types of datasets.

Deliverables

a) **Codebase:** Provide well-documented code implementing the ensemble models, including preprocessing and hyperparameter tuning.

b) **Report:** Submit a comprehensive report detailing for each dataset the preprocessing steps, hyperparameter tuning process, and results. Include visualizations, comparison table that summarizes the accuracy for each ensemble model on each dataset, and insights obtained from the comparison.

Evaluation Criteria

Students will be assessed based on the completeness of the implementation, the quality of the analysis, each team should be able to address every single detail of their implementations on the three models, note that the evaluation will be individual