Birzeit University
Department of Electrical & Computer Engineering
First Semester, 2023/2024
ENCS5141 Intelligent Systems Lab
Assignment 2

1. Background

Ensemble methods in machine learning involve combining multiple models to improve overall predictive performance. Bagging (Bootstrap Aggregating) and Boosting are two common techniques. In Bagging, like Random Forest, multiple decision trees are trained independently on different subsets of the data, and their predictions are averaged or voted upon. This helps reduce overfitting and enhances model robustness. In Boosting, like XGBoost, models are trained sequentially, with each subsequent model focusing on correcting errors made by the previous ones. XGBoost employs gradient boosting techniques, regularization, and custom loss functions to optimize model performance, making it a powerful and efficient boosting algorithm in ensemble methods.

2. Objective:

The objective of this assignment is to conduct a comprehensive comparative study between Random Forest and XGBoost, two popular ensemble learning techniques, in different scenarios. Students will explore their strengths, weaknesses, and applicability across diverse datasets to gain a deeper understanding of when and why one might outperform the other.

3. Scenario Design:

Identify and create distinct scenarios representing real-world situations. Consider datasets that cover three issues from the following:

1. Imbalanced classes,
2. Noisy data or features,
3. Varying degrees of dimensionality,
4. and large datasets.

Train and evaluate both Random Forest and XGBoost models on each scenario. Tune their hyperparameters using grid search or other optimization techniques.

There are many websites that contain datasets, each with its own focus and strengths. Here are some popular options:

- Kaggle: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu/
- Google Dataset Search: https://datasetsearch.research.google.com/
- Microsoft Research Open Data: https://www.microsoft.com/en-us/research/project/microsoft-research-open-data/

4. Report

Write a comprehensive report that include the following:

1. **Literature Review:** Briefly explain the core principles of Random Forest and XGBoost, including their training methodologies (ensemble learning vs. boosting) and key hyperparameters. Discuss the advantages and limitations of each algorithm, highlight key differences between the two algorithms, such as the use of decision trees, regularization techniques, and learning rate optimization.
2. **Scenarios designed and analysis:** for each scenario briefly describe the objectives, and the data set selected. Define and justify appropriate evaluation metrics based on the nature of the datasets (Common metrics include accuracy, precision, recall, and F1 score). Draw and compare the results obtained. Compare the computational efficiency of Random Forest and XGBoost, considering training time and memory usage.
3. **Conclusion and recommendations:** Based on your findings, summarize the strengths and weaknesses of Random Forest and XGBoost in each analyzed scenario. Provide recommendations on which algorithm might be preferred for different types of problems, considering factors like data characteristics, desired interpretability, available computational resources, and others.