Homework 2 (100 points): This homework is due on January 27ᵗʰ at 11:59 pm.

Ensemble Learning

In this homework, you will solve the same problem as homework 1 using a different machine learning technique. Upload a .txt file with a link to your file as your submission on Submitty.

You need to perform the following tasks for this homework:

Task 1 (30 points): Implement a Decision Tree Classifier for your classification problem. You may use a built-in package to implement your classifier. Additionally, do the following:

- Visualize the decision tree structure for at least three different parameter settings. Comment on how the depth and complexity change the tree.
- Do some research on what sensitivity analysis is and how it is performed (include citations). Perform a sensitivity analysis to measure the impact of at least two input features on your model's decision boundary.

Task 2 (30 points): From the Bagging and Boosting ensemble methods pick any one algorithm from each category. Implement both the algorithms using the same data.

- Use stratified k-fold cross-validation with at least three different folds (e.g., 5, 10, 15). You may do your own research on this technique (include citations).
- Evaluate the models using any three-evaluation metrics of your choice (e.g. accuracy, Precision, F1-score etc.).
- Comment on the behavior of each algorithm under the metrics. Does the performance ranking change based on the metric used? Why?

Task 3 (40 points): Compare the effectiveness of the three models implemented above. Analyze the results using the following:

- A confusion matrix for one selected test fold.
- A statistical test (e.g., paired t-test) to determine if differences between models are significant.
- A discussion on the trade-off between bias and variance for each model.

The following task is for Graduate level only (6000 level): This task is more open ended and emphasizes the research aspect of implementing a model. You will be exploring the impact of hyperparameter tuning which we haven't discussed in detail so far.

Task (50 points): For the same classification problem solved above, implement the XGBoost algorithm. If you picked XGBoost as one of the boosting algorithms in task 2, you may use the same implementation. Implement and evaluate XGBoost with the following requirements:

1. Perform a **grid search or random search** over at least 3 hyperparameters, such as learning rate, max depth, and subsample.

2. Analyze the sensitivity of your model to changes in these parameters.
3. **Optional (no points taken off if not done)** - Create plots to show the effect of each parameter on accuracy and another metric.

Note: An experiment can be defined as a systematic way of picking parameter values. This could be something that you come up with yourself or you may refer to the exiting literature on design of experiments for hyperparameter tuning. This task will require you to do some research into this open-source library and hyper-parameter tuning yourself. A good place to start is here: https://www.jeremyjordan.me/hyperparameter-tuning/

https://xgboost.readthedocs.io/en/latest/parameter.html#general-parameters