Instructions for homework submission

1. Complete two parts in this homework:
   - Math questions: Include your solution in LaTeX document. Show your work. Submission with embedded photos of handwritten work will not be graded.
   - Programming questions: Complete the given skeleton `Python` code. **For questions requiring visualization, analysis, and discussion, please include your solution in the same LaTeX document.**
2. Submit your work to Gradescope including:
   - A PDF document for written parts: `FirstName_LastName_HW2.pdf`. LaTeX source code is not required.
   - A completed `Python` code: `FirstName_LastName_HW2.py`.
   - **There are two separate submission portals on `Gradescope`:** one for code and one for the report. Submitting your work to the wrong portal will result in a loss of marks.
   - Please assign your answer in PDF report to its corresponding question when submitting to Gradescope. Submitting your work without assigning correspoinding question will result in a loss of marks.
3. Start early!
4. Total: 100 points.

This project utilizes the Bank Marketing dataset[1], provided by UCI Machine Learning Repository. We acknowledge and appreciate their efforts in collecting and maintaining this data.

Math Questions

Problem 1: Information Gain

Note: This is not a programming assignment, so you may **not** use programming tools to solve this problem. Show your work.

Suppose you are given 6 training points as shown below for a classification problem with two binary attributes X_1 and X_2 and three classes $Y \in \{1, 2, 3\}$. You will use a decision tree learner based on information gain.

1. Calculate the information gain for both X_1 and X_2.
2. Report which attribute is used for the first split. Draw the decision tree using this split.
3. Conduct classification for the test example $X_1 = 0$ and $X_2 = 1$.

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

[1] S. Moro, P. Rita, and P. Cortez. "Bank Marketing," UCI Machine Learning Repository, 2014. [Online]. Available: https://doi.org/10.24432/C5K306.

Problem 2: Entropy

Note: This is not a programming assignment, so you may **not** use programming tools to solve this problem. Show your work.

Suppose you are given the same training points as above for a classification problem with two binary attributes X_1 and X_2 and three classes $Y \in \{1, 2, 3\}$. You will use a decision tree learner based on entropy.

1. Calculate the conditional entropy for both X_1 and X_2.
2. Report which attribute is used for the first split. Draw the decision tree using this split.
3. Conduct classification for the test example $X_1 = 0$ and $X_2 = 1$.

Programming Questions

Part A: Classification Tree (50 points)

In this problem, you will code up a classification tree from scratch. Trees are a special class of graphs with only directed edges and no cycles, also known as directed acyclic graphs (DAGs). Each child node has only one parent node.

1. **Data Processing and EDA:**
   (a) Split the dataset into train and validation sets. Please use a 80:20 split.
   (b) Read and print the training data. Add a short description in your pdf report about the data.
   (c) Extract the features and the label. The label is "`y`".
   (d) Plot histograms of all variables and provide a brief discussion.
2. **Implementation:**
   (a) Implement a classification tree from scratch. Do not use ML libraries.
   (b) Train the model using training data and validate it with the validation data.
   (c) Aim for `F1 score > 0.4`.

Part B: Boosting (20 points)

Use a decision-tree-based ensemble algorithm for the `Bank` Dataset.

1. Define a function `train_XGBoost` to train an XGBoost model with L2 regularization. Use bootstrapping (100 iterations) and evaluate different values of α in `alpha_vals = [1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3]`. Report the optimal α.
2. Initialize `my_best_model` given in the skeleton code with a `XGBClassifier()` on your best hyperparameters. You need to achieve `F1 score > 0.55`.
3. Plot the ROC curve and report the area under the curve (AUC). Include axes labels, legend, and title.