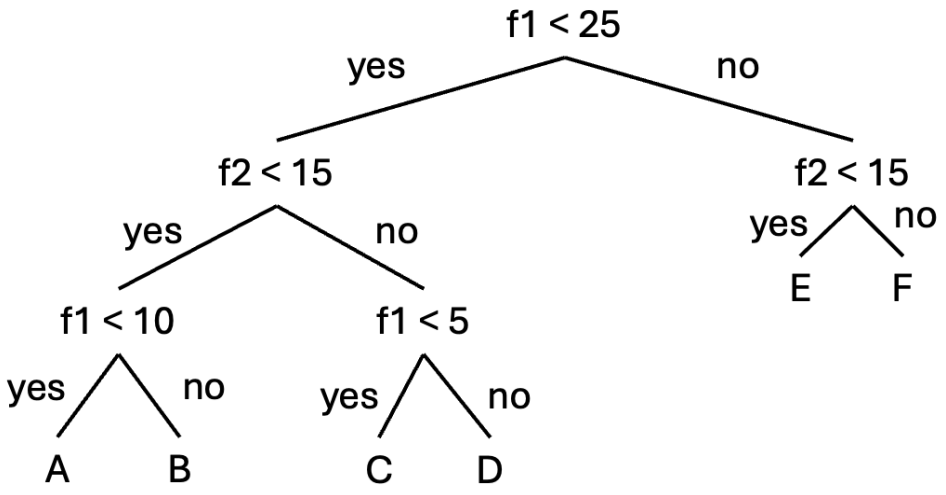
CS 5805: Machine Learning I
Spring 2025
Homework 3
Date Assigned: 2/17/2025
Date Due: 2/26/2025

1. (5x2=10 points) Answer True/False with a short 1-2 sentence explanation for each.
    a. Given a classification dataset it is always possible (even if undesirable) to grow a decision tree completely so that the entropy of the class label is zero.
    b. ID3 returns the best possible (optimal) decision tree due to its design of selecting the order of attributes.
    c. While selecting features to train a machine learning algorithm, it is imperative to select ones that provide no discrimination about the class so that the learning algorithm can be unbiased.
    d. In learning a decision tree from data, we must continue growing the decision tree till all instances are correctly classified and then can stop growing the tree.
    e. If we learn a decision tree, convert the tree into rules, prune some of the rules and/or delete some of the rules to form our model, the same model could have been obtained by simply pruning the tree in the first place.

2. (10+10=20 points; courtesy APD) Consider the following decision tree which has two features and six classes:
    a. Draw the decision boundaries defined by the tree on a 2D plane. The x-axis denotes the variable f1 and the y-axis denotes the variable f2. Each leaf is labeled with a letter (class). Write this letter in the corresponding region of your 2D plane.



b. Give another decision tree that is syntactically different but defines the same decision boundaries.

3. (15 points; courtesy APD) Consider a two-category classification task with the following training data. Manually, by hand, construct a complete (unpruned) decision tree for this data (with all of the rows below as training data) using information gain as your splitting criterion. For full credit, show i) the calculations for selecting the first condition (ie the root condition), and ii) the full tree with all levels and leaves clearly displayed. Calculations for levels below the root level will need to be done to obtain the full tree but need not be described in your solution (just the full tree is sufficient).

f1	f2	f3	f4	class
a	1	c	-1	c_1
b	0	c	-1	c_1
a	0	c	1	c_1
b	1	c	1	c_1
b	0	c	1	c_2
a	0	a	-1	c_2
a	1	a	-1	c_2
b	1	c	-1	c_2

4. (5+5+5=15 points)
   a. Prove that the entropy function is concave. A function f(x) is concave on an interval [a, b] if for any two points x1 and x2 in [a, b] and any λ, where 0 < λ < 1, the following holds:

   $$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$

   For instance, the value at the midpoint of every interval in the domain exceeds the average of its values at the ends of the interval.
   b. Is Information Gain also concave? If yes, prove it. If not, explain why not.
   c. What does concavity of entropy mean for the ID3 decision tree algorithm? Why or how is it relevant?

5. (20 points) Consider the Bank Marketing dataset available at (https://archive.ics.uci.edu/dataset/222/bank+marketing). The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution.

   DO NOT DOWNLOAD THE DATA FROM THE ABOVE WEBSITE! (We provided the link above so you can understand the context behind the dataset.) Instead, please download and use the cleaned up data supplied along with this assignment in Canvas, and build training and test sets from the dataset we provide.

   Your task is to build decision trees of various forms and then analyze the results. The classification goal is to predict if the client will subscribe to a term deposit (variable y). The value of y can be "yes" or "no" (2 classes). Use cross-validation with a suitable number of folds (which you must mention in your solution).

   Make sure to conduct basic data cleaning (starting from the dataset supplied with this assignment):

   - Cells with "unknown" are actually missing values. Decide by yourself how to handle missing values. Explain in your submitted solution how you handled them and any impact that might have had on your results.
   - Determine a strategy to encode categorical columns. Tree-based models can support categorical variables inherently but scikit-learn requires you to map them to numerical values. Thus, you can map, for instance, {blue, red, green} to {0, 1, 2}, or something similar. Alternatively, you can create three binary columns called {f1=blue, f1=red, f1=green} and fill in zero or 1 as appropriate. Pick one strategy and mention it in your solution.

   For each experiment, after fitting a tree model, evaluate its performance using various classification metrics such as accuracy, precision, recall, F1 score. Discuss what each metric measures, their relevance to classification in this domain, and any limitations they might have in this context.

   Use the following five algorithms: ID3 (Entropy), CART (Gini), Random Forests, Gradient Boosting, and XGBoost. Assess the performance of the models learnt by these algorithms and provide descriptive comparisons of their selective superiorities (e.g., algorithm _____ is better for ____, etc.).

6. (20 points) Consider the Forest Cover Type classification dataset available at (https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset). The dataset contains tree observations from four areas of the Roosevelt National Forest in Colorado, including information on tree type, shadow coverage, distance to nearby landmarks (roads etc.), soil type, and local topography.

   DO NOT DOWNLOAD THE DATA FROM THE ABOVE WEBSITE! (We provided the link above so you can understand the context behind the dataset.) Instead, please

download and use the data from Canvas. Note that in our version of the data, we have remapped the Cover_Type label from 1-7 to 0-6 for ease of implementation of all tree methods. We have also randomly sampled 10% of the entire dataset to yield a smaller dataset for this assignment.

Your task is to build decision trees of various forms and then analyze the results. The classification goal is to predict the cover type (variable "Cover_Type"). The value of y can be 0 to 6. Use cross-validation with a suitable number of folds (which you must mention in your solution).

For each experiment, after fitting a tree model, evaluate its performance using various classification metrics such as accuracy, precision, recall, F1 score. Discuss what each metric measures, their relevance to classification in this domain, and any limitations they might have in this context.

Use the following five algorithms: ID3 (Entropy), CART (Gini), Random Forests, Gradient Boosting, and XGBoost. Assess the performance of the models learnt by these algorithms and provide descriptive comparisons of their selective superiorities (e.g., algorithm _____ is better for ____, etc.).

Submission: Please include sufficient plots, graphs, and commentary to support your observations. Please organize your code (if any) into an IPYNB file and submit together with the main PDF work. Provide a link to your notebook on Google Colab (turn on the "Anyone with the link" sharing mode) that we can evaluate as necessary.