BANK NOTES CLASSIFICATION

Title of the Task: The banknote dataset involves predicting whether a given banknote is authentic given a number of measures taken from a photograph.

Description: The dataset contains 1,372 rows with 5 numeric variables. It is a classification problem with two classes (binary classification).

Below provides a list of the five variables in the dataset.variance of Wavelet Transformed image (continuous).

1. skewness of Wavelet Transformed image (continuous).
2. kurtosis of Wavelet Transformed image (continuous).
3. entropy of image (continuous).
4. class (integer).

Link to the dataset: http://archive.ics.uci.edu/ml/datasets/banknote+authentication

Objective: Implementing the CART (Classification and Regression Trees) Algorithm.

Description: Classification and Regression Trees or CART for short is an acronym introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems. The representation of the CART model is a binary tree. This is the same binary tree from algorithms and data structures, nothing too fancy (each node can have zero, one or two child nodes). A node represents a single input variable (X) and a split point on that variable, assuming the variable is numeric. The leaf nodes (also called terminal nodes) of the tree contain an output variable (y) which is used to make a prediction. Once created, a tree can be navigated with a new row of data following each branch with the splits until a final prediction is made. Creating a binary decision tree is actually a process of dividing up the input space. A greedy approach is used to divide the space called recursive binary splitting. This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function.

The split with the best cost (lowest cost because we minimize cost) is selected. All input variables and all possible split points are evaluated and chosen in a greedy manner based on the cost function.

- Regression: The cost function that is minimized to choose split points is the sum squared error across all training samples that fall within the rectangle.
- Classification: The Gini cost function is used which provides an indication of how pure the nodes are, where node purity refers to how mixed the training data assigned to each node is.

Splitting continues until nodes contain a minimum number of training examples or a maximum tree depth is reached.

Mandatory extensions to the task: This can be a theoretical assumption regarding the data but any attempt shall be valued in your favor. No need to be concerned about accurate results. We need to see one's effort here.
Cross Entropy. Another cost function for evaluating splits is cross entropy (logloss). You could implement and experiment with this alternative cost function.

1. **Tree Pruning**. An important technique for reducing overfitting of the training dataset is to prune the trees. Investigate and implement tree pruning methods.
2. **Categorical Dataset**. Experiment with categorical input data and splits that may use equality instead of ranking.
3. **Regression**. Adapt the tree for regression using a different cost function and method for creating terminal nodes.

Submission: Key things to keep in mind

1. The submission has to be in the form of a Jupyter Notebook along with any necessary data files or models.
2. The submitted files could be shared via a google drive link or by sharing a github link for the same.
3. Feel free to refer to any online codes or references but, your submission shall be evaluated for plagiarism as well.
4. Time for the task is 2 days, and it starts as soon as you confirm that you have received the mail. We expect honesty and loyalty in this part.
5. No clarification shall be provided from our end, this task is quite crisp, clear and concise.
6. Once you've completed the task, kindly share the link to the submission folder or github url with hroperations@alphaai.biz.

All the best for the task!