Assignment 1

Write jupyter notebook scripts for the questions.

1. Distance measure(50 points)

Compute the distance measure between every pair of flowers in the iris dataset. Your script will

- take a parameter that specifies what distance measure to use: Euclidean or cosine similarity.
- produce a .csv file that contains the distance. The indices for the rows and columns are the flowers.

2. Decision Tree (50 points)

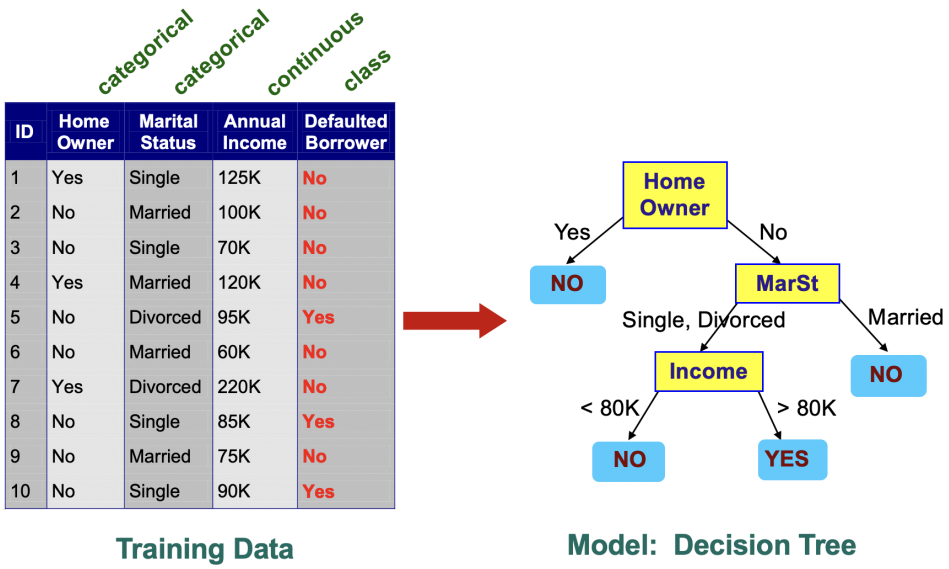

Figure 1: An example of decision tree.

Compute the impurity at each node in Figure 1 using Gini impurity index, Entropy, and maxclassification error.

- Present your results as a small table with rows corresponding to nodes and columns corresponding to types of impurity measures.
- Compute the gain at each node defined as:

  $$\Delta = \text{impurity of parent node - weighted impurity of the child nodes}$$

  The weights are the proportions of samples/records in the corresponding child nodes.