NAME- RISHI KUMAR
CWID- 20015656
HOME WORK #3

Part 1: Prepare a 1-page slide/poster to review/summarize the concept of Decision Tree Learning

Decision Tree

A **decision tree** is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

- The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes
- Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes
- The leaf nodes represent all the possible outcomes within the dataset.

Types of the Decision tree:

- ID3:This algorithm leverages entropy and information gain as metrics to evaluate candidate splits
- C4.5: It can use information gain or gain ratios to evaluate split points within the decision trees.
- CART: This algorithm typically utilizes Gini impurity to identify the ideal attribute to split on. Gini impurity measures how often a randomly chosen attribute is misclassified. When evaluating using Gini impurity, a lower value is ideal.

Entropy is a concept that stems from information theory, which measures the impurity of the sample values. Entropy values can fall between 0 and 1

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

There are multiple ways to select the best attribute at each node, two methods, information gain, and Gini impurity, act as popular splitting criteria for decision tree models.

Information gain: Information gain represents the difference in entropy before and after a split on a given attribute. The attribute with the highest information gain will produce the best split as it's doing the best job at classifying the training data according to its target classification

$$\text{Information Gain}(S,a) = \text{Entropy}(S) - \sum_{v \in values(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Gini impurity: Gini impurity is the probability of incorrectly classifying a random data point in the dataset if it were labeled based on the class distribution of the dataset

$$\text{Gini Impurity} = 1 - \sum_i (p_i)^2$$

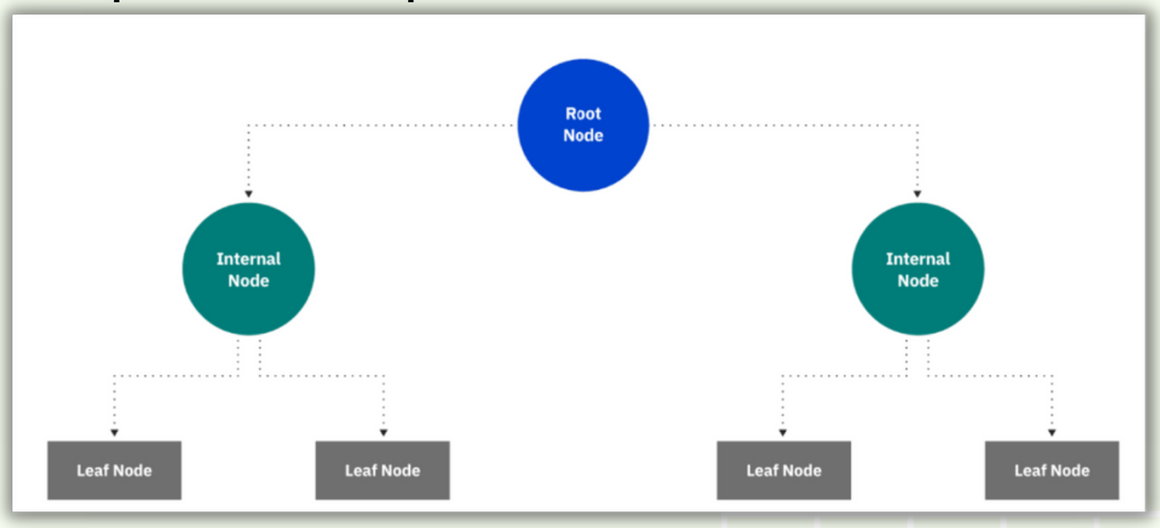
Advantages of the Decision tree:

- Easy to interpret:
- More flexible:
- Little to no data preparation required:

Disadvantages of the Decision tree:

- Prone to overfitting:
- High variance estimators:
- More costly:
- Not fully supported in sci-kit-learn:

Graphical depiction of Decision Tree



references:

1. *https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.*