ASSIGNMENT 3

Title : Classification using Machine Learning

Problem Statement :
Perform following operations on given dataset.

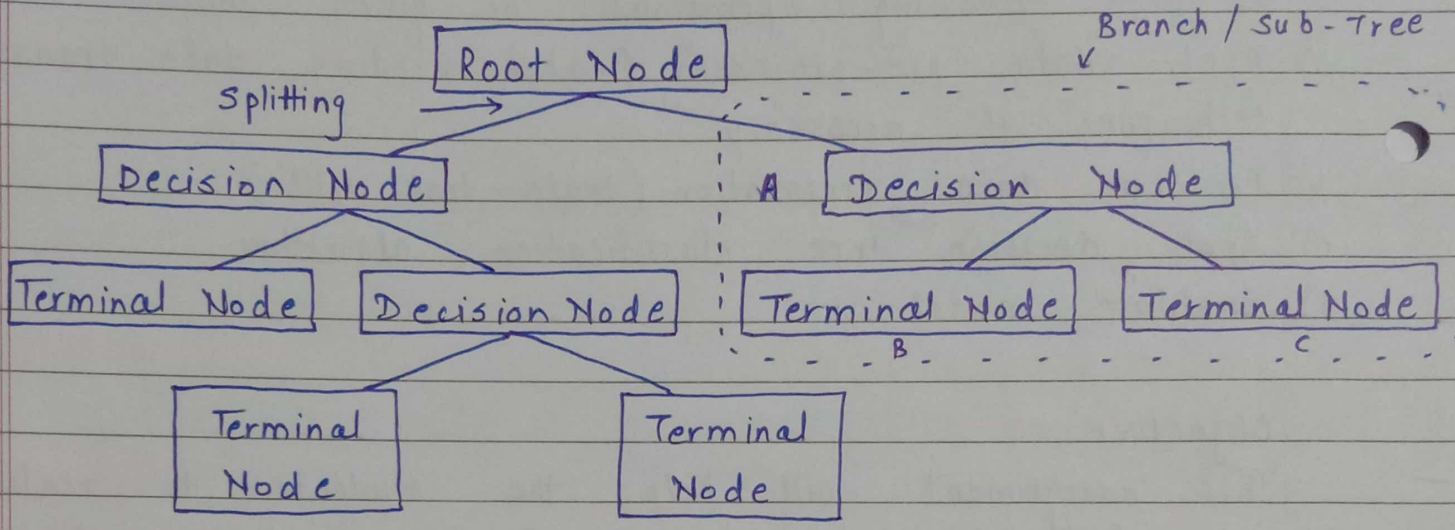
a) Apply data pre-processing (Label encoding, data transformation) techniques if necessary.
b) Perform data preparation (train-test split).
c) Apply decision tree classification algorithm.
d) Evaluate model.

Objective :
This assignment will help the students to realize how the decision tree classifier can be used and predictions using the same can be performed.

Theory :
Classification :
It is a process of categorizing a given set of data into classes. It can be performed on both structured or unstructured data. The process starts with predicting the class of the given data points. The classes are often referred to as target, label or categories.

Decision Tree :

It uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.



Root nodes : It is the node present at the beginning of a decision tree. From this node, the population starts dividing according to various features.

Decision nodes : The nodes we get after splitting the root are called decision nodes.

Leaf nodes : The nodes where further splitting is not possible are called leaf nodes / terminal nodes.

Sub-tree : A subsection of decision tree is called a subtree.

Pruning : It is cutting down some nodes to stop overfitting.

Entropy :
It's used to calculate the homogeneity of a sample. If the sample is completely homogenous, the entropy is zero and if the sample is equally divided, it has entropy of one.

(a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

(b) Entropy using the frequency table of two attributes :

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Information gain :
The info. gain is based on decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attributes that return the highest info. gain (i.e. the most homogenous branch).

Step 1 : Calculate entropy of the target :

Step 2 : The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it's added proportionally to get total entropy for the split. The resulting entropy is subtracted from entropy before the split. The result is

the info. gain or decrease in entropy.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Step 3 : Choose attribute with largest information gain as the decision node, divide the dataset by it's branches and repeat the same process on every branch.

Step 4a : A branch with entropy of 0 is a leaf node.

Step 4b : A branch with entropy more than 0 needs further splitting.

Step 5 : The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

Decision tree to decision rules :
A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

Pruning : It is a method that can help us avoid overfitting. It helps in improving the performance of the tree by cutting the nodes or sub-nodes which are not significant. It removes the branches which have very low importance. There are mainly 2 ways of pruning :

① Pre Pruning :
We can stop growing the tree earlier which means we can prune / remove / cut a node if it has low importance while growing the tree.

② Post Pruning :
Once our tree is built to its depth, we can start pruning the nodes based on their significance.

Application :
Helpful in solving classification problems.

Conclusion :
Decision tree is used for classification and predictions are made using the decision tree classifier.