Assignment: Decision Trees

Course AI, Instructor Dr. Rauf Ahmed Shams Malick,

Start Date: 15/5/2021, End Date: 21/5/2021 Total Marks: 35

Note: There will be no acceptance for late submission, '0' on plagiarism

Decision trees are widely being used for classification. Entropy based information gain method is one of the most conventional and famous methods to measure the contribution of each feature. The contribution or information gain measure is used to decide the split of the tree. However, there are several parameters that can be used to build different types of trees. Respective tree will have its own interpretation. Measure of 'goodness' of a tree is tricky. Apart from the correctness of the test data, structural information is important which defining 'good' or 'bad' decision tree. For e.g. a few trees are highly imbalance, then a few are highly balance, some are purely random trees, some are depth defined trees, and so on. You are provided a very basic data set from UIC repository. Scikit library details are also provided to develop Decision Trees. Through a simple scikit library (as mentioned in following) you can choose your options for your decision trees. It means multiple decision trees are possible for a given data.

1. criterion{"gini", "entropy"}, default="gini"
2. splitter{"best", "random"}, default="best"
3. max_depth int, default=None
4. min_samples_split int or float, default=2
5. min_samples_leaf int or float, default=1
6. min_weight_fraction_leaf float, default=0.0
7. max_featuresint, float or {"auto", "sqrt", "log2"}, default=None
8. random_state int, RandomState instance or None, default=None
9. max_leaf_nodes int, default=None
10. min_impurity_decrease float, default=0.0
11. min_impurity_split float, default=0
12. class_weightdict, list of dict or "balanced", default=None
13. ccp_alphanon-negative float, default=0.0

You have to build '3' decision trees, by considering options from above list. You can only choose options based on following rules:

1. If your last digit of enrolment number is a prime number and its > 6 then select 3, 5, 11 options.
2. If your last digit of enrolment number is a prime number & an odd number & its < 6 then + 3 then divide by 2 (the resultant is 'a'). You have to solve a, a+2, a+3.
3. If your last digit of enrolment number is an odd number and its > 6 then select 4, 9, 13 options.
4. If your last digit of enrolment number is an odd number and its < 6 then select 1, 2, 8 options.

Dataset: https://archive.ics.uci.edu/ml/datasets/iris

Tutorial: https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Q1. You have to build multiple decision trees by at least three ways (as mentioned above, based on the instructed options). 15 Marks

Q2. Then visualize the three trees. 5 Marks

Q3 compare the differences among the three trees in case of given dataset. 5 Marks

Q4. Which particular type of tree is more suitable and which will not be? Share the reason. 10 Marks