Department of Computer Science & Engineering

Final Year B. Tech. (CSE) – I: 2022-23

5CS462: PE5 - Data Mining Lab

Assignment No. 3

Name: Sanket Mote

PRN: 2019BTECS00113

Batch: B8

Title:

Design the data analysis tools (GUI) to perform the following pre-processing task.

1. Implement the decision tree classifier using the following attribute selection measures and graphically show/visualize the tree:

a. Information Gain b. Gain Ratio c. Gini Index

2. Tabulate the results in confusion matrix and evaluate the performance of above classifier using following metrics:

a. Recognition rate b. Misclassification rate c. Sensitivity d. Specificity e. Precision & Recall

3. Use the following categorical data sets from UCI machine learning repository:

a. Balance Scale data set b. Car evaluation data set c. Breast-cancer data set

Objective/Aim:

1. To implement data analysis tool using python programming language.
2. To implement decision tree classifier for selected attribute using information gain, gain ratio and Gini index.
3. To tabulate the results in confusion matrix.

Introduction:

Decision Trees are supervised machine learning algorithms that are best suited for classification and regression problems. These algorithms are constructed by implementing the particular splitting conditions at each node, breaking down the training data into subsets of output variables of the same class.

Theory:

Decision Tree Classifier:

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

Information Gain, Gain Ratio and Gini Index are the three fundamental criteria to measure the quality of a split in Decision Tree.

Entropy:

It is a measurement of the uncertainty in data. In the context of Classification Machine Learning, Entropy measures the diversification of the labels.

The formula for entropy is,

$$H = -(\sum p_i \log_2 p_i)$$

Information Gain:

The Information Gain of a split equals the original Entropy minus the weighted sum of the sub-entropies, with the weights equal to the proportion of data samples being moved to the sub-datasets.

$$IG_{split} = H - (\sum \frac{|D_j|}{|D|} * H_j)$$

Gain ratio:

Gain Ratio attempts to lessen the bias of Information Gain on highly branched predictors by introducing a normalizing term called the Intrinsic Information. The Intrinsic Information (II) is defined as the entropy of sub-dataset proportions. In other words, it is how hard for us to guess in which branch a randomly selected sample is put into. The formula of Intrinsic Information is:

$$II = -(\sum \frac{|D_j|}{|D|} * \log_2 \frac{|D_j|}{|D|})$$

The gain ratio is,

$$GainRatio = \frac{\text{Information Gain}}{\text{Intrinsic Information}}$$

Gini Index:

The Gini Index is given by,

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

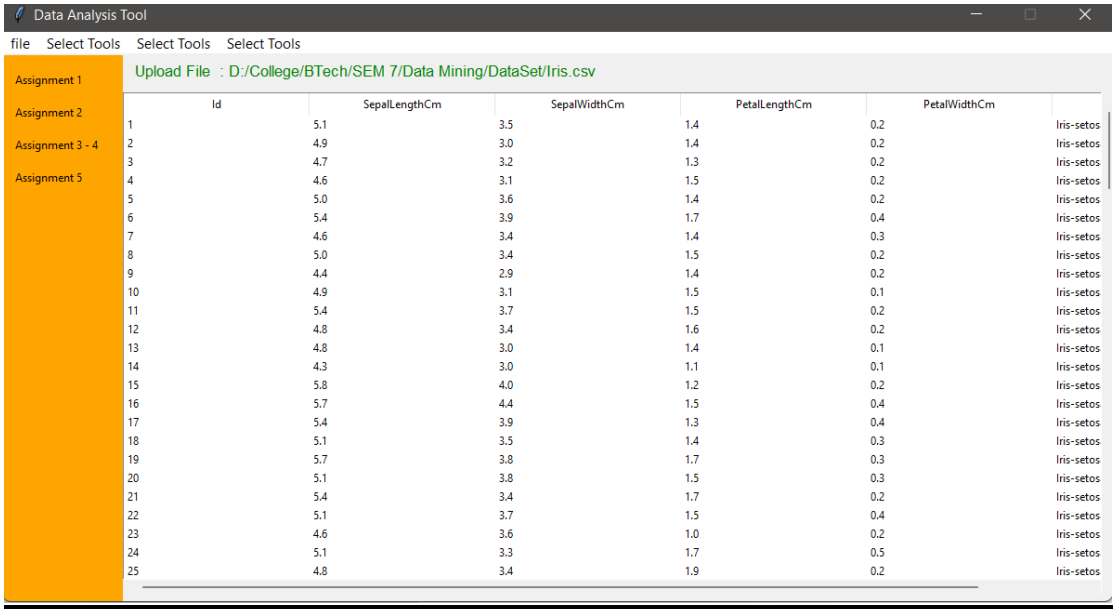
The Gini of split is given by,

$$\text{Gini}_{\text{split}} = \sum \frac{|D_j|}{|D|} \text{Gini}_j$$

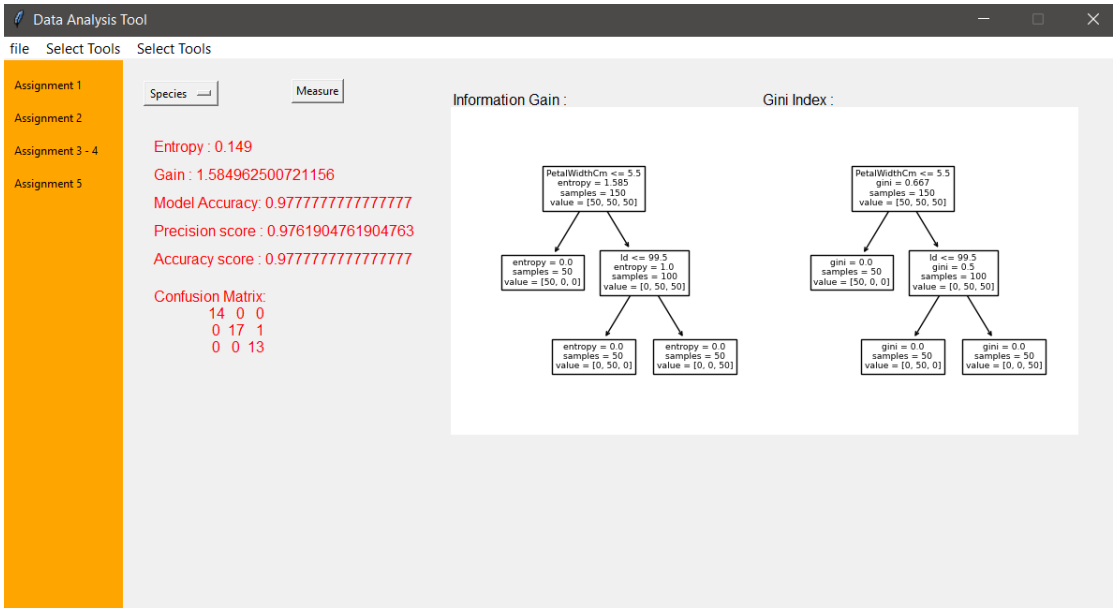
Procedure:

Given problem statement is solved using python programming language and specifically used tkinter module to implement GUI application and pandas module to load .csv file as dataset.

Results / Screen Shots :



Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	
1	5.1	3.5	1.4	0.2	Iris-setos
2	4.9	3.0	1.4	0.2	Iris-setos
3	4.7	3.2	1.3	0.2	Iris-setos
4	4.6	3.1	1.5	0.2	Iris-setos
5	5.0	3.6	1.4	0.2	Iris-setos
6	5.4	3.9	1.7	0.4	Iris-setos
7	4.6	3.4	1.4	0.3	Iris-setos
8	5.0	3.4	1.5	0.2	Iris-setos
9	4.4	2.9	1.4	0.2	Iris-setos
10	4.9	3.1	1.5	0.1	Iris-setos
11	5.4	3.7	1.5	0.2	Iris-setos
12	4.8	3.4	1.6	0.2	Iris-setos
13	4.8	3.0	1.4	0.1	Iris-setos
14	4.3	3.0	1.1	0.1	Iris-setos
15	5.8	4.0	1.2	0.2	Iris-setos
16	5.7	4.4	1.5	0.4	Iris-setos
17	5.4	3.9	1.3	0.4	Iris-setos
18	5.1	3.5	1.4	0.3	Iris-setos
19	5.7	3.8	1.7	0.3	Iris-setos
20	5.1	3.8	1.5	0.3	Iris-setos
21	5.4	3.4	1.7	0.2	Iris-setos
22	5.1	3.7	1.5	0.4	Iris-setos
23	4.6	3.6	1.0	0.2	Iris-setos
24	5.1	3.3	1.7	0.5	Iris-setos
25	4.8	3.4	1.9	0.2	Iris-setos



Conclusion:

Successfully implemented data analysis tool (GUI) for decision tree classifier based on information gain, gain ratio and gini index for selected attribute of given dataset.