Assignment 7

Problem statement

Assignment on Classification Technique Every year many students take the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is thetargetvariable.DataSet: https://www.kaggle.com/mohansacharya/graduate-admissions The counselor of the firm is supposed to check whether the student will get an admission or not based on his/her GRE score and Academic Score. So to help the counselor to make appropriate decisions build a machine learning model classifier using a Decision tree to predict whether a student will get admission or not. a) Apply Data pre-processing (Label Encoding, Data Transformation….) techniques if necessary. b) Perform data preparation (Train-Test Split) c) Apply Machine Learning Algorithm d) Evaluate Model..

S/W Packages and H/W apparatus used:

OS: **Ubuntu/Windows,**
Tool: Google Colab
Packages: **Numpy, Pandas, Matplotlib and SkLearn**

Theory :

1. Classification:

Classification 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 given data points. The classes are often referred to as targets, label or categories.

2. What is a 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 nodes are called Decision Node

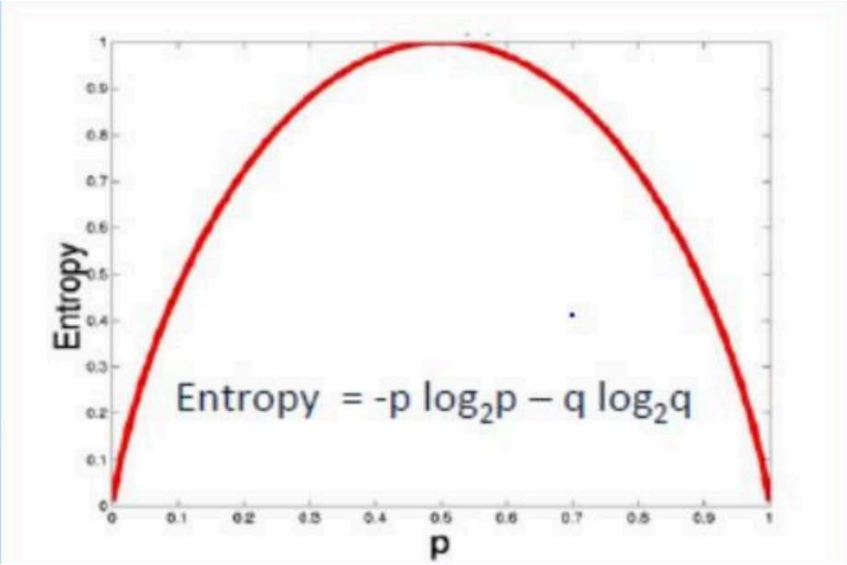
Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes

Sub-tree – just like a small portion of a graph is called a sub-graph similarly a subsection of this decision tree is called a sub-tree.

Pruning – It is cutting down some nodes to stop overfitting

3. Entropy:

Entropy is used to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided it has an entropy of one.



4. Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute.

5. Constructing a decision tree :

Is all about finding attributes that return the highest information gain (i.e., the most homogeneous branches)

Step 1: Calculate the entropy of the target.

Step 2: The dataset is then split into different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get the total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

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

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

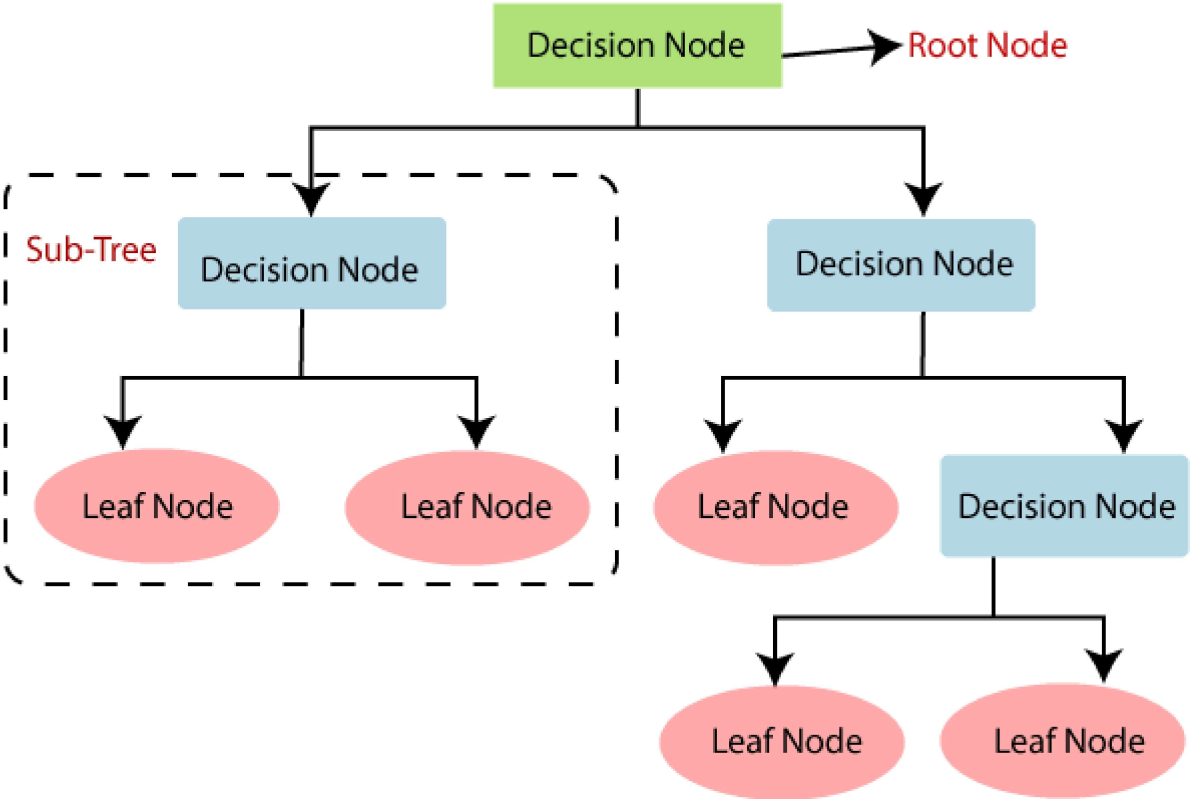
Step 4b: A branch with entropy more than 0 needs further splitting.

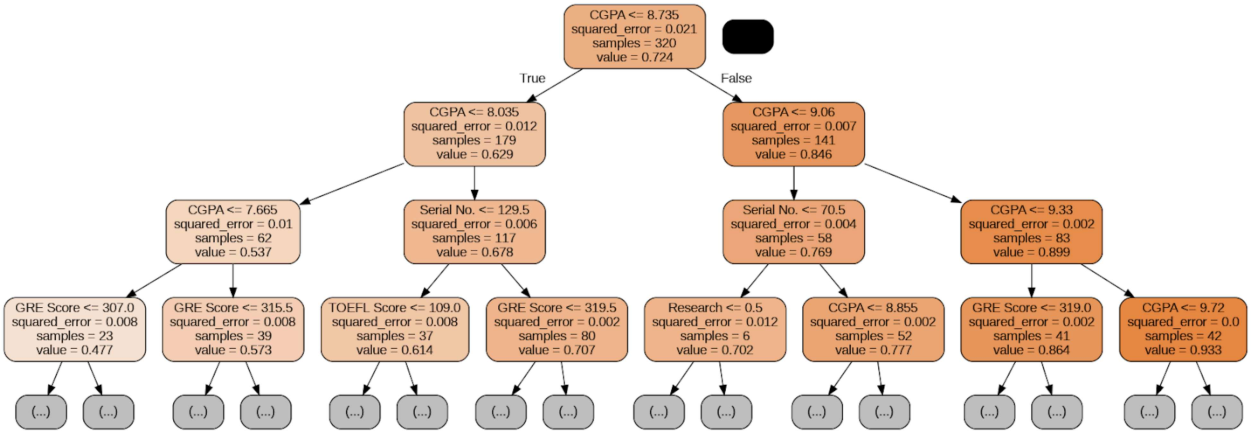
Step 5: The ID3 algorithm is run recursively on the non-leaf branches until all data is classified.

6. Decision Tree to Decision Rules

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

DIAGRAM :


Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node


CGPA <= 8.735
squared_error = 0.021
samples = 320
value = 0.724
True
False
CGPA <= 8.035
squared_error = 0.012
samples = 179
value = 0.629
CGPA <= 9.06
squared_error = 0.007
samples = 141
value = 0.846
CGPA <= 7.665
squared_error = 0.01
samples = 62
value = 0.537
Serial No. <= 129.5
squared_error = 0.006
samples = 117
value = 0.678
Serial No. <= 70.5
squared_error = 0.004
samples = 58
value = 0.769
CGPA <= 9.33
squared_error = 0.002
samples = 83
value = 0.899
GRE Score <= 307.0
squared_error = 0.008
samples = 23
value = 0.477
GRE Score <= 315.5
squared_error = 0.008
samples = 39
value = 0.573
TOEFL Score <= 109.0
squared_error = 0.008
samples = 37
value = 0.614
GRE Score <= 319.5
squared_error = 0.002
samples = 80
value = 0.707
Research <= 0.5
squared_error = 0.012
samples = 6
value = 0.702
CGPA <= 8.855
squared_error = 0.002
samples = 52
value = 0.777
GRE Score <= 319.0
squared_error = 0.002
samples = 41
value = 0.864
CGPA <= 9.72
squared_error = 0.0
samples = 42
value = 0.933
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

CONCLUSION:

Classification techniques help in classifying problems and help figure out the relationship between the various variables.