**ML Assignment on Decision Tree**

**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.

**THEORY:**

**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 target, label or categories.

**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

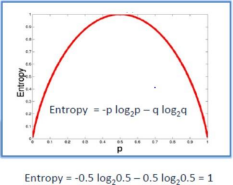
**Leaf Nodes** – the nodes where further splitting is not possible are called leaf nodes or terminal nodes

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

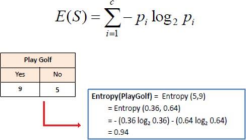
Pruning – It is cutting down some nodes to stop overfitting

**Entropy:**

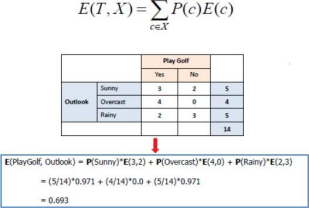
Entropy is used to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided it has entropy of one.



a) Entropy using the frequency table of one attribute:



b) Entropy using the frequency table of two attributes:



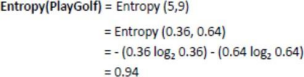
**Information Gain**

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

**Constructing a decision tree**

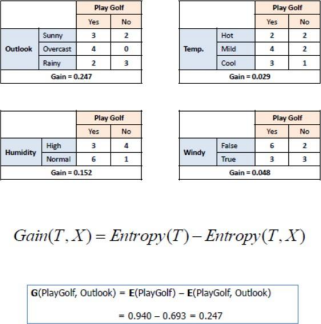
IS all about finding attributes that return the highest information gain (i.e., the most homogeneous branches)

Step 1: Calculate entropy of the target.

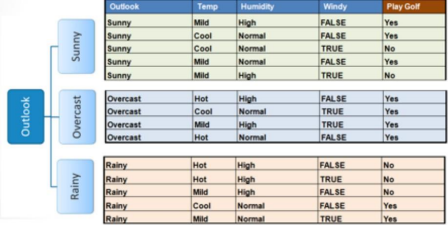


Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated.

Then it is added proportionally, to get 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.

**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 another 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 for pruning:

(i) 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. (ii) Post-pruning – once our tree is built to its depth, we can start pruning the nodes based on their significance.

**CONCLUSION:**

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