Probability

Temp - Cool, Hot, Mild

Cricket - Yes, No

Total Rows = 20

Temp --- Yes --- No

Cool --- 5 --- 2 --- 7/20

Hot --- 3 --- 4 --- 7/20

Mild --- 3 --- 3 --- 6/20

| |

Total - 11/20 - 9/20

Decision Tree

Entropy - To check the purity of column values (Categories)

Information Gain - To check how much information a column is gaining

Gini index - Same as Entropy, will check the purity of column values (Categories)

In Dataset -

Independent column, input column, input features are the same - Can be many

Silimarly, Target Column, output column, dependent feature are the same. - Only one

Dataset - Play Tennis -

Input Column - Outlook, Temperature, Humidity, Wind

Output Column - Play Tennis

Prob(Yes) - 9/14

Prob(No) - 5/14

For complete Dataset

Prob(Sunny) - 5/14 - Prob(Sunny|Yes) - 2/5 - Prob(Sunny|No) - 3/5

Prob(Overcast) - 4/14 - Prob(Over|Yes) - 4/4 - Prob(Over|No) - 0/4

Prob(Rain) - 5/14 - Prob(Rain|Yes) - 3/5 - Prob(Rain|No) - 2/5

We will calculate the entropy and information gain for every column

To create the tree we have to start from root node,

To decide which column should be root node we will consider the max information gain of column

Entropy range from 0 to 1

0 means pure and 1 means impure

Entropy = -∑ p x $\log_2(p)$

Entropy = $-P_{yes} \log_2(P_{yes}) - P_{no} \log_2(P_{no})$

= $-(9/14)*\log_2(9/14) - (5/14)*\log_2(5/14)$

= 0.94

Entropy(sunny) = $-P_{yes} \log_2(P_{yes}) - P_{no} \log_2(P_{no})$

= $-(2/5)*\log_2(2/5) - (3/5)*\log_2(3/5)$

= 0.97

Entropy(over.) = $-P_{yes} \log_2(P_{yes}) - P_{no} \log_2(P_{no})$

= $-(4/4)*\log_2(4/4) - (0/4)*\log_2(0/4)$

= 0

Entropy(Rain) = $-P_{yes} \log_2(P_{yes}) - P_{no} \log_2(P_{no})$

= $-(3/5)*\log_2(3/5) - (2/5)*\log_2(2/5)$

= 0.97

Information Gain =

Overall Entropy – Weighted Entropy of column values

= E(table) – [(sunny/total)*E(sunny) + (over/total)*E(over) + (rain/total)*E(rain)]

= 0.94 – [(5/14)*0.97 + (4/14)*0 + (5/14)*0.97]

= 0.94 - 0.69

IG(Outlook) = 0.25