The given dataset is

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp. | Humidity | Wind | Play |
| D1 | Sunny | 85 | 85 | Weak | No |
| D2 | Sunny | 80 | 90 | Strong | No |
| D3 | Overcast | 83 | 86 | Weak | Yes |
| D4 | Rain | 70 | 96 | Weak | Yes |
| D5 | Rain | 68 | 80 | Weak | Yes |
| D6 | Rain | 65 | 70 | Strong | No |
| D7 | Overcast | 64 | 65 | Strong | Yes |
| D8 | Sunny | 72 | 95 | Weak | No |
| D9 | Sunny | 69 | 70 | Weak | Yes |
| D10 | Rain | 75 | 80 | Weak | Yes |
| D11 | Sunny | 75 | 70 | Strong | Yes |
| D12 | Overcast | 72 | 90 | Strong | Yes |
| D13 | Overcast | 81 | 75 | Weak | Yes |
| D14 | Rain | 71 | 91 | Strong | No |

In the given dataset, day, outlook, temp, humidity and wind are the independent variables and play is the target variable. To build a decision tree, we start splitting the datapoints from the root node and reach to a leaf node that has high homogeneity in the target variable. The splitting of the data points occur based on the Gain ratio.

Gain ratio is the ratio between Information Gain and Split Information.

Information Gain of a variable = Entropy of the target variable-Conditional Entropy of a Variable at each split.

Gain Ratio = Information Gain/ Split Information

Target variable play (Tennis) has 9 Yes and 5 No classes.

Entropy(Play) = -(9/14\*log(9/14))-(5/14\*log(5/14)) = 0.94

Outlook has 3 classes which are 5 sunny, 5 rain and 4 overcast.

Entropy(Outlook) = -(5/14\*log(5/14))-(4/14\*log(4/14))-(5/14\*log(5/14)) = 1.577 (Also called as split Information)

Information Gain(Outlook) = 0.94-(5/14)\*(-(2/5\*log(2/5))-(3/5\*log(3/5)))-(5/14)\*(-(2/5\*log(2/5))-(3/5\*log(3/5))) = 0.247

Gain Ratio = Information Gain/Split Information = 0.247/1.577 = 0.156.

Wind variable has 2 classes which are 8 Weak and 6 Strong.

Entropy(Wind) = -(8/14\*log(8/14))-(6/14\*log(6/14)) = 0.985

Information Gain(Wind) = 0.94-(8/14)\*(-(2/8\*log(2/8))-(6/8\*log(6/8)))-(6/14)\*(-(3/6\*log(3/6))-(3/6\*log(3/6))) = 0.048

Gain Ratio = Information Gain/Split Information = 0.048/0.985 = 0.049

Temperature variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 64 | 65 | 68 | 69 | 70 | 71 | 72 | 72 | 75 | 75 | 80 | 81 | 83 | 85 |
| Yes | No | Yes | Yes | Yes | No | No | Yes | Yes | Yes | No | Yes | Yes | No |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 64.5,66.5,70.5,72,77.5,80.5, and 84.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 64.5 | [1+,0-] | 0 | [8+,5-] | -(8/13\*log(8/13))-( 5/13\*log(5/13))=0.96 | 0.94-(1/14)(0)-(13/14)(0.96)=0.04 |
| 66.5 | [1+,1-] | -(1/2\*log(1/2))-( 1/2\*log(1/2))=1 | [8+,4-] | -(8/12\*log(8/12))-( 4/12\*log(4/12))=0.92 | 0.94-(2/14)(1)-(12/14)(0.92)=0.008 |
| 70.5 | [4+,1-] | -(1/5\*log (1/5))-(4/5\*log(4/5))=0.72 | [5+,4-] | -(5/9\*log(5/9))-( 4/9\*log(4/9))=0.99 | 0.94-(5/14)(0.72)-(9/14)(0.99)=0.046 |
| 72.5 | [4+,2-] | -(2/6\*log (2/6))-(4/6\*log(4/6))=0.92 | [5+,3-] | -(5/8\*log(5/8))-( 3/8\*log(3/8))=0.95 | 0.94-(6/14)(0.92)-(8/14)(0.95)=0.004 |
| 77.5 | [7+,3-] | -(7/10\*log (7/10))-(3/10\*log(3/10))=0.88 | [2+,2-] | -(2/4\*log(2/4))-( 2/4\*log(2/4))=1 | 0.94-(10/14)(0.88)-(4/14)(1)=0.027 |
| 80.5 | [7+4-] | -(7/11\*log (7/11))-(4/11\*log(4/11))=0.95 | [2+,1-] | -(2/3\*log(2/3))-( 1/3\*log(1/3))=0.92 | 0.94-(11/14)(0.95)-(3/14)(0.92)=0.003 |
| 84 | [9+,4-] | -(9/13\*log (9/13))-(4/13\*log(4/13))=0.89 | [0+,1-] | 0 | 0.94-(13/14)(0.89)-(1/14)(0)=0.113 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain Ratio |
| 64.5 | -(1/14\*log(1/14)+13/14\*log(13/14))=0.371 | 0.04/0.371=0.107 |
| 66.5 | -(2/14\*log(2/14)+12/14\*log(12/14))=0.591 | 0.008/0.591=0.013 |
| 70.5 | -(5/14\*log(5/14)+9/14\*log(9/14))=0.94 | 0.046/0.94 = 0.05 |
| 72.5 | -(8/14\*log(8/14)+6/14\*log(6/14))=0.985 | 0.004/0.985=0.004 |
| 77.5 | -(10/14\*log(10/14)+4/14\*log(4/14))=0.863 | 0.027/0.863=0.031 |
| 80.5 | -(11/14\*log(11/14)+3/14\*log(3/14))=0.749 | 0.003/0.749=0.004 |
| 84 | -(13/14\*log(13/14)+1/14\*log(1/14))=0.371 | 0.113/0.371=0.304 |

Humidity variable is also a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 70 | 70 | 70 | 75 | 80 | 80 | 85 | 86 | 90 | 90 | 91 | 95 | 96 |
| Yes | Yes | Yes | No | Yes | Yes | Yes | No | Yes | Yes | No | No | No | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 70,72.5,82.5,85.5,90 and 95.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 70 | [1+,0-] | 0 | [8+,5-] | -(8/13\*log(8/13))-( 5/13\*log(5/13))=0.96 | 0.94-(1/14)(0)-(13/14)(0.96)=0.048 |
| 72 | [3+,1-] | -(3/4\*log(3/4))-( 1/4\*log(1/4))=0.81 | [6+,4-] | -(6/10\*log(6/10))-( 4/10\*log(4/10))=0.97 | 0.94-(4/14)(0.81)-(10/14)(0.97)=0.015 |
| 82.5 | [6+,1-] | -(6/7\*log (6/7))-(1/7\*log(1/7))=0.59 | [3+,4-] | -(3/7\*log(3/7))-( 4/7\*log(4/7))=0.985 | 0.94-(7/14)(0.59)-(7/14)(0.985)=0.153 |
| 85.5 | [6+,2-] | -(6/8\*log (6/8))-(2/8\*log(2/8))=0.81 | [3+,3-] | -(3/6\*log(3/6))-( 3/6\*log(3/6))=1 | 0.94-(8/14)(0.81)-(4/14)(1)=0.191 |
| 90 | [8+,3-] | -(8/11\*log (8/11))-(3/11\*log(3/11))=0.84 | [1+,2-] | -(2/3\*log(2/3))-( 1/3\*log(1/3))=0.92 | 0.94-(11/14)(0.84)-(3/14)(0.92)=0.08 |
| 95.5 | [8+,5-] | -(8/13\*log (8/13))-(5/13\*log(5/13))=0.96 | [1+,0-] | 0 | 0.94-(13/14)(0.96)-(1/14)(0)=0.048 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain ratio |
| 70 | -(1/14\*log(1/14)+ 13/14\*log(13/14))=0.371 | 0.017/0.371=0.045 |
| 72 | -(4/14\*log(4/14)+ 10/14\*log(10/14))=0.863 | 0.015/0.863=0.017 |
| 82.5 | -(7/14\*log(7/14)+ 7/14\*log(7/14))=1 | 0.153/1=0.153 |
| 85.5 | -(8/14\*log(8/14)+6/14\*log(6/14))=0.985 | 0.191/0.985=0.194 |
| 90 | -(11/14\*log(11/14)+3/14\*log(3/14))=0.749 | 0.08/0.749=0.106 |
| 95.5 | -(13/14\*log(13/14)+1/14\*log(1/14))=0.371 | 0.048/0.371=0.129 |

The Gain ratio is highest for Temperature at threshold 84. So the first split of the node happens at Temperature with threshold=84.

Right Branch Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp. | Humidity | Wind | Play |
| D1 | Sunny | 85 | 85 | Weak | No |

Since it is only one record, this node will become a leaf node.

Temperature

<84

>=84

No

Left Branch Table of root node Temperature

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp. | Humidity | Wind | Play |
| D2 | Sunny | 80 | 90 | Strong | No |
| D3 | Overcast | 83 | 86 | Weak | Yes |
| D4 | Rain | 70 | 96 | Weak | Yes |
| D5 | Rain | 68 | 80 | Weak | Yes |
| D6 | Rain | 65 | 70 | Strong | No |
| D7 | Overcast | 64 | 65 | Strong | Yes |
| D8 | Sunny | 72 | 95 | Weak | No |
| D9 | Sunny | 69 | 70 | Weak | Yes |
| D10 | Rain | 75 | 80 | Weak | Yes |
| D11 | Sunny | 75 | 70 | Strong | Yes |
| D12 | Overcast | 72 | 90 | Strong | Yes |
| D13 | Overcast | 81 | 75 | Weak | Yes |
| D14 | Rain | 71 | 91 | Strong | No |

Entropy(Play) = -(9/13\*log(9/13))-( 4/13 \*log(4/13)) = 0.89

Entropy(Wind) = -(7/13\*log(7/13))-( 6/13 \*log(6/13)) = 0.99

Information Gain (Wind) = 0.89-(7/13)\*(0.591)-(6/13)\*(1) = 0.11

Gain Ratio (Wind) = 0.11/0.99=0.11

Entropy(Outlook) = -(4/13\*log(4/13))-( 4/13\*log(4/13))-( 5/13\*log(5/13)) = 1.576

Information Gain (Outlook) = 0.89-4/13\*(1)-5/13\*(0.971)=0.21

Gain Ratio (Outlook) = 0.21/1.576 = 0.133

Temperature variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 64 | 65 | 68 | 69 | 70 | 71 | 72 | 72 | 75 | 75 | 80 | 81 | 83 |
| Yes | No | Yes | Yes | Yes | No | No | Yes | Yes | Yes | No | Yes | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 64.5,66.5,70.5,72,77.5 and 80.5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 64.5 | [1+,0-] | 0 | [8+,4-] | -(8/12\*log(8/12))-( 4/12\*log(4/12))=0.918 | 0.89-(1/13)(0)-(12/13)(0.918)=0.042 |
| 66.5 | [1+,1-] | -(1/2\*log(1/2))-( 1/2\*log(1/2))=1 | [8+,3-] | -(8/11\*log(8/11))-( 3/11\*log(3/11))=0.845 | 0.89-(2/13)(1)-(11/13)(0.845)=0.021 |
| 70.5 | [4+,1-] | -(1/5\*log (1/5))-(4/5\*log(4/5))=0.72 | [5+,3-] | -(5/8\*log(5/8))-( 43/8\*log(3/8))=0.954 | 0.89-(5/13)(0.72)-(8/13)(0.954)=0.026 |
| 72 | [4+,2-] | -(2/6\*log (2/6))-(4/6\*log(4/6))=0.92 | [5+,2-] | -(5/7\*log(5/7))-( 2/7\*log(2/7))=0.863 | 0.89-(6/13)(0.92)-(7/13)(0.863)=0 |
| 77.5 | [7+,3-] | -(7/10\*log (7/10))-(3/10\*log(3/10))=0.88 | [2+,1-] | -(2/3\*log(2/3))-( 1/3\*log(1/3))=0.918 | 0.89-(10/13)(0.88)-(3/13)(0.918)=0.001 |
| 80.5 | [7+4-] | -(7/11\*log (7/11))-(4/11\*log(4/11))=0.95 | [2+,0-] | 0 | 0.89-(11/13)(0.95)-(2/13)(0)=0.086 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain Ratio |
| 64.5 | -(1/13\*log(1/13)+12/13\*log(12/13))=0.391 | 0.042/0.391=0.107 |
| 66.5 | -(2/13\*log(2/13)+11/13\*log(11/13))=0.619 | 0.021/0.619=0.034 |
| 70.5 | -(5/13\*log(5/13)+8/13\*log(8/13))=0.961 | 0.026/0.961 = 0.027 |
| 72.5 | -(8/13\*log(8/13)+5/13\*log(5/13))=0.961 | 0/0.961=0 |
| 77.5 | -(10/13\*log(10/13)+3/13\*log(3/13))=0.78 | 0.001/0.78=0.01 |
| 80.5 | -(11/13\*log(11/13)+2/13\*log(2/13))=0.619 | 0.086/0.619=0.139 |

Humidity variable is also a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 70 | 70 | 70 | 75 | 80 | 80 | 86 | 90 | 90 | 91 | 95 | 96 |
| Yes | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | No | No | No | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 70,72.5,90,95.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 70 | [1+,0-] | 0 | [8+,4-] | -(8/12\*log (8/12))-(4/12\*log(4/12))=0.918 | 0.89-(1/13)(0)-(12/13)(0.918)=0.043 |
| 72.5 | [3+,1-] | -(3/4\*log(3/4))-( 1/4\*log(1/4))=0.81 | [6+,3-] | -(6/9\*log(6/9))-( 3/9\*log(3/9))=0.918 | 0.89-(4/13)(0.81)-(9/13)(0.918)=0.005 |
| 90 | [7+,1-] | -(7/8\*log (7/8))-( 1/8\*log(1/8))=0.54 | [2+,3-] | -(2/5\*log(2/5))-( 3/5\*log(3/5))=0.971 | 0.89-(8/13)(0.54)-(5/13)(0.971)=0.184 |
| 95.5 | [8+,4-] | -(8/12\*log (8/12))-(4/12\*log(4/12))=0.918 | [1+,0-] | 0 | 0.89-(12/13)(0.918)-(1/13)(0)=0.042 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain ratio |
| 70 | -(1/13\*log(1/13)+ 12/13\*log(12/13))=0.391 | 0.043/0.391=0.109 |
| 72.5 | -(4/13\*log(4/13)+ 9/13\*log(9/13))=0.89 | 0.005/0.89=0.005 |
| 90 | -(8/13\*log(8/13)+ 5/13\*log(5/13))=0.961 | 0.184/0.961 =0.191 |
| 95.5 | -(1/13\*log(1/13)+ 12/13\*log(12/13))=0.391 | 0.042/0.391=0.107 |

After comparing the gain ratios of al the variables, humidity having gain ratio of 0.184 would be the node for splitting branch.

Temperature

<84

>=84

Humidity

No

Left Branch of Humidity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp. | Humidity | Wind | Play |
| D3 | Overcast | 83 | 86 | Weak | Yes |
| D5 | Rain | 68 | 80 | Weak | Yes |
| D6 | Rain | 65 | 70 | Strong | No |
| D7 | Overcast | 64 | 65 | Strong | Yes |
| D9 | Sunny | 69 | 70 | Weak | Yes |
| D10 | Rain | 75 | 80 | Weak | Yes |
| D11 | Sunny | 75 | 70 | Strong | Yes |
| D13 | Overcast | 81 | 75 | Weak | Yes |

Entropy(Play) = -(7/8\*log(7/8))-( 1/8 \*log(1/8)) = 0.543

Entropy(Wind) = -(5/8\*log(5/8))-(3/8 \*log(3/8)) = 0.95

Information Gain (Wind) = 0.543-(5/8)\*(0)-(3/8)\*(0.918) = 0.199

Gain Ratio (Wind) = 0.199/0.95=0.209

Entropy(Outlook) = -(2/8\*log(2/8))-( 3/8\*log(3/8))-( 3/8\*log(3/8)) = 1.561

Information Gain (Outlook) = 0.543-3/8\*(0.918)=0.199

Gain Ratio (Outlook) = 0.199/1.561 = 0.127

Temperature variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 64 | 65 | 68 | 69 | 75 | 75 | 81 | 83 |
| Yes | No | Yes | Yes | Yes | Yes | Yes | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 64.5 and 66.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 64.5 | [1+,0-] | 0 | [6+,1-] | -(6/7\*log (6/7))-( 1/7\*log(1/7))=0.591 | 0.543-(1/8)(0)-(7/8)(0.591)=0.025 |
| 66.5 | [1+,1-] | -(1/2\*log(1/2))-( 1/2\*log(1/2))=1 | [6+,0-] | 0 | 0.543-(2/8)(1) =0.64 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain Ratio |
| 64.5 | -(1/8\*log(1/8)+7/8\*log(7/8))=0.543 | 0.025/0.543=0.046 |
| 66.5 | -(2/8\*log(2/8)+6/8\*log(6/8))=0.811 | 0.64/0.811=0.789 |

Humidity variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 70 | 70 | 70 | 75 | 80 | 80 | 86 |
| Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 70 and 72.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 70 | [1+,0-] | 0 | [6+,1-] | -(6/7\*log (6/7))-( 1/7\*log(1/7))=0.591 | 0.543-(1/8)(0)-(7/8)(0.591)=0.025 |
| 72.5 | [3+,1-] | -(3/4\*log(3/4))-( 1/4\*log(1/4))=0.81 | [4+,0-] | 0 | 0.543-(4/8)(0.81)-(4/8)(0)=0.138 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain ratio |
| 70 | -(1/8\*log(1/8)+ 7/8\*log(7/8))=0.543 | 0.025/0.543=0.046 |
| 72.5 | -(4/8\*log(4/8)+ 4/8\*log(4/8))=1 | 0.138/1=0.138 |

After comparing the gain ratios of all the variables, temperature having gain ratio of 0.789 would be the node for splitting branch.

Temperature

Temperature

<84

>=84

1

Humidity

No

>=90

<90

Temperature

6

Yes

Right Branch of Humidity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp. | Humidity | Wind | Play |
| D2 | Sunny | 80 | 90 | Strong | No |
| D4 | Rain | 70 | 96 | Weak | Yes |
| D8 | Sunny | 72 | 95 | Weak | No |
| D12 | Overcast | 72 | 90 | Strong | Yes |
| D14 | Rain | 71 | 91 | Strong | No |

Entropy(Play) = -(2/5\*log(2/5))-( 3/5 \*log(3/5)) = 0.971

Entropy(Wind) = -(2/5\*log(2/5))-( 3/5 \*log(3/5)) = 0.971

Information Gain (Wind) = 0.971-(2/5)\*(1)-(3/5)\*(0.918) = 0.201

Gain Ratio (Wind) = 0.201/0.971=0.207

Entropy(Outlook) = -(2/5\*log(2/5))-(2/5\*log(2/5))-( 1/5\*log(1/5)) = 1.521

Information Gain (Outlook) = 0.971-2/5\*(1)=0.571

Gain Ratio (Outlook) = 0.571/1.521 = 0.375

Temperature variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 70 | 71 | 72 | 72 | 80 |
| Yes | No | No | Yes | No |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 70.5,72,76.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 70.5 | [1+,0-] | 0 | [1+,3-] | -(1/4\*log (1/4))-( 3/4\*log(3/4))=0.811 | 0.971-(1/5)(0)-(4/5)(0.811)=0.322 |
| 72 | [1+,1-] | -(1/2\*log(1/2))-( 1/2\*log(1/2))=1 | [1+,2-] | -(2/3\*log(2/3))-( 1/3\*log(1/3))=0.918 | 0.971-(2/5)(1)-(3/5)(0.918)=0.02 |
| 76 | [2+,2-] | -(2/4\*log(2/4))-( 2/4\*log(2/4))=1 | [0+,1-] | 0 | 0.971-(4/5)(1)-(1/5)(0)=0.171 |

|  |  |  |
| --- | --- | --- |
| Threshold | Split Information | Gain ratio |
| 70.5 | -(1/5\*log (1/5))-(4/5\*log(4/5))=0.72 | 0.322/0.72=0.44 |
| 72 | -(1/5\*log (1/5))-(4/5\*log(4/5))=0.72 | 0.02/0.72=0.02 |
| 76 | -(1/5\*log (1/5))-(4/5\*log(4/5))=0.72 | 0.171/0.72=0.238 |

Humidity variable is a continuous variable. So we split the variable into different thresholds and calculate the Entropy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 90 | 90 | 91 | 95 | 96 |
| Yes | No | No | No | Yes |

Thresholds are calculated at the mid points of change in the levels of the target variable. So here the thresholds are 90 and 95.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Threshold | < | | >= | | Information Gain |
| Combo | Entropy | Combo | Entropy |
| 90 | [0+,0-] | 0 | [2+,3-] | -(2/5\*log(2/5))-( 3/5\*log(3/5))=0.970 | 0.971-0.970=0.001 |
| 95.5 | [1+,3-] | -(1/4\*log(1/4))-( 3/4\*log(3/4))=0.811 | [1+,0-] | 1 | 0.971-(4/5)(0.811)-(1/5)(1)=0.122 |

After comparing the gain ratios of all the variables, outlook having gain ratio of 0.375 would be the node for splitting branch.

After getting the

Temperature

Temperature

<84

>=84

1

Humidity

No

>=90

<90

Overcast

Temperature

1

Outlook

<66.5

<66.5

6

Yes

Rain

Sunny

Yes

2

No

After carefully observing the records belonging to the following groups, on outlook node split = Rain and Temperature < 66.5, we can conclude that the complete tree structure would be in this way.

Temperature

Temperature

<84

>=84

1

Humidity

No

>=90

<90

Overcast

Temperature

1

Outlook

<66.5

<66.5

6

Yes

Rain

Sunny

Yes

Humidity

2

>=67.5

Humidity

No

<67.5

>=93

1

1

No

<93

Yes

1

No

1

Yes