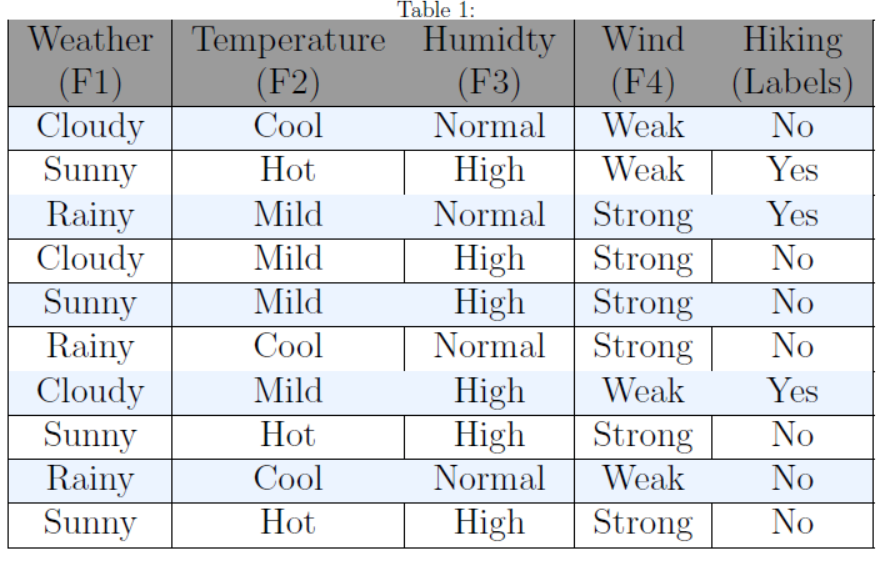
**ELG5255 Applied Machine Learning**

REPORT of: Group Assignment 4 (Group-18)

**Part 1: Calculations**



1. **Build a decision tree by using Gini Index (i.e., Gini = 1 - , where is the number of class).**

Hiking (labels) 🡪 P (Yes) = , P (No) =

**We will calculate probabilities of classes in F1, F2, F3, and F4 in this table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** |
| P(F1 = Cloudy) = | P(F2 = Cool) = | P(F3 = Normal) = | P(F4 = Weak) = |
| P(F1 = Sunny) = | P(F2 = Hot) = | P(F3 = High) = | P(F4 = Strong) = |
| P(F1 = Rainy) = | P(F2 = Mild) = |  |  |

**We will calculate the Gini Index for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Cloudy and Hiking = Yes) = | P(F1 = Cloudy and Hiking = No) = |
| P(F1 = Sunny and Hiking = Yes) = | P(F1 = Sunny and Hiking = No) = |
| P(F1 = Rainy and Hiking = Yes) = | P(F1 = Rainy and Hiking = No) = |

Gini Index of Cloudy = 1-(+) = 0.44

Gini Index of Sunny = 1-(+) = 0.375

Gini Index of Rainy = 1-(+) = 0.44

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Weather (F1) = \* 0.44 + \* 0.375 + \* 0.44 = 0.414

**We will calculate the Gini Index for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Cool and Hiking = Yes) = | P(F2 = Cool and Hiking = No) = |
| P(F2 = Hot and Hiking = Yes) = | P(F2 = Hot and Hiking = No) = |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |

Gini Index of Cool = 1-(+) = 0

Gini Index of Hot = 1-(+) = 0.44

Gini Index of Mild = 1-(+) = 0.5

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Temperature (F2) = \* 0 + \* 0.44 + \* 0.5 = 0.332

**We will calculate the Gini Index for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = Normal and Hiking = Yes) = | P(F3 = Normal and Hiking = No) = |
| P(F3 = High and Hiking = Yes) = | P(F3 = High and Hiking = No) = |

Gini Index of Normal = 1-(+) = 0.375

Gini Index of High = 1-(+) = 0.44

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Humidty (F3) = \* 0.375 + \* 0.44 = 0.414

**We will calculate the Gini Index for Wind (F4)**

|  |  |
| --- | --- |
| Wind(F4) | |
| P(F4 = Weak and Hiking = Yes) = | P(F4 = Weak and Hiking = No) = |
| P(F4 = Strong and Hiking = Yes) = | P(F4 = Strong and Hiking = No) = |

Gini Index of Weak = 1-(+) = 0.5

Gini Index of Strong = 1-(+) = 0.278

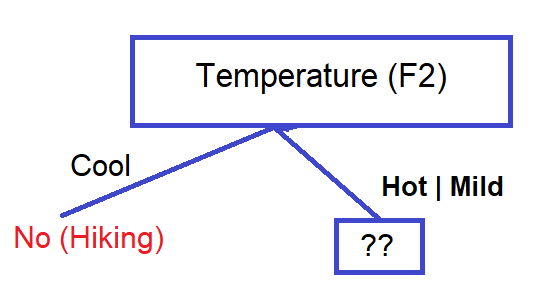
Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Wind (F4) = \* 0.5 + \* 0.278 = 0.367

**Gini Index attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | 0.414 |
| Temperature (F2) | **0.332** |
| Humidty (F3) | 0.414 |
| Wind (F4) | 0.367 |

From the above table, we observe that ‘Temperature (F2)’ has the lowest Gini Index and hence it will be chosen as the root node for how decision tree works.



We will repeat the same procedure to determine the sub-nodes or branches of the decision tree.

We will calculate the Gini Index for the ‘Hot | Mild’ branch of Temperature (F2) as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** | **Hiking** |
| Sunny | Hot | High | Weak | Yes |
| Rainy | Mild | Normal | Strong | Yes |
| Cloudy | Mild | High | Strong | No |
| Sunny | Mild | High | Strong | No |
| Cloudy | Mild | High | Weak | Yes |
| Sunny | Hot | High | Strong | No |
| Sunny | Hot | High | Strong | No |

**We will calculate probabilities of classes in F1, F2, F3, and F4 of this table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** |
| P(F1 = Sunny) = | P(F2 = Hot) = | P(F3 = High) = | P(F4 = Weak) = |
| P(F1 = Sunny) = | P(F2 = Mild) = | P(F3 = Normal) = | P(F4 = Strong) = |
| P(F1 = Cloudy) = |  |  |  |

**We will calculate the Gini Index for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Sunny and Hiking = Yes) = | P(F1 = Sunny and Hiking = No) = |
| P(F1 = Rainy and Hiking = Yes) = |  |
| P(F1 = Cloudy and Hiking = Yes) = | P(F1 = Cloudy and Hiking = No) = |

Gini Index of Sunny = 1-(+) = 0.375

Gini Index of Rainy = 1-() = 0

Gini Index of Sunny = 1-(+) = 0.5

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Weather (F1) = \* 0.375 + \* 0 + \* 0.5 = 0.357

**We will calculate the Gini Index for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Hot and Hiking = Yes) = | P(F2 = Hot and Hiking = No) = |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |

Gini Index of Hot = 1-(+) = 0.44

Gini Index of Mild = 1-(+) = 0.5

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Temperature (F2) = \* 0.44 + \* 0.5 = 0.474

**We will calculate the Gini Index for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = High and Hiking = Yes) = | P(F3 = High and Hiking = No) = |
| P(F3 = Normal and Hiking = Yes) = |  |

Gini Index of High = 1-(+) = 0.44

Gini Index of High = 1-() = 0

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Humidty (F3) = \* 0.44 + \* 0 = 0.377

**We will calculate the Gini Index for Wind (F4)**

|  |  |
| --- | --- |
| Wind (F4) | |
| P(F4 = Weak and Hiking = Yes) = |  |
| P(F4 = Strong and Hiking = Yes) = | P(F4 = Strong and Hiking = No) = |

Gini Index of Weak = 1- () = 0

Gini Index of High = 1- () = 0.32

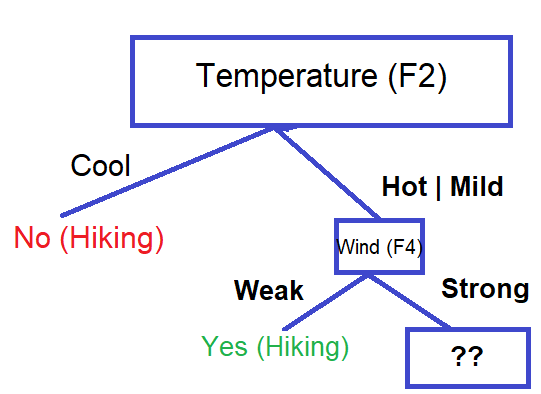
Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Wind (F4) = \* 0 + \* 0.32 = 0.229

**Gini Index attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | 0.357 |
| Temperature (F2) | 0.474 |
| Humidty (F3) | 0.377 |
| Wind (F4) | **0.229** |

From the above table, we observe that ‘Wind (F4)’ has the lowest Gini Index and hence it will be chosen as the child node for the ‘Hot | Mild’ branch of Temperature (F2).



We will calculate the Gini Index for the ‘Strong’ branch of Wind (F4) as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** | **Hiking** |
| Rainy | Mild | Normal | Strong | Yes |
| Cloudy | Mild | High | Strong | No |
| Sunny | Mild | High | Strong | No |
| Sunny | Hot | High | Strong | No |
| Sunny | Hot | High | Strong | No |

**We will calculate probabilities of classes in F1, F2, and F3 of the above table:**

|  |  |  |
| --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** |
| P(F1 = Rainy) = | P(F2 = Mild) = | P(F3 = Normal) = |
| P(F1 = Cloudy) = | P(F2 = Hot) = | P(F3 = High) = |
| P(F1 = Sunny) = |  |  |

**We will calculate the Gini Index for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Rainy and Hiking = Yes) = |  |
|  | P(F1 = Cloudy and Hiking = No) = |
|  | P(F1 = Sunny and Hiking = Yes) = |

Gini Index of Rainy = 1-() = 0

Gini Index of Cloudy = 1-() = 0

Gini Index of Sunny = 1-() = 0

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Weather (F1) = \* 0 + \* 0 + \* 0 = 0

**We will calculate the Gini Index for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |
|  | P(F2 = Hot and Hiking = No) = |

Gini Index of Mild = 1-(+) = 0.44

Gini Index of Hot = 1-() = 0

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Temperature (F2) = \* 0.44 + \* 0 = 0.27

**We will calculate the Gini Index for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = Normal and Hiking = Yes) = |  |
|  | P(F3 = High and Hiking = No) = |

Gini Index of Normal = 1-() = 0

Gini Index of High = 1-() = 0

Weighted sum of the Gini Indices can be calculated as follows:

Gini Index of Humidty (F3) = \* 0 + \* 0 = 0

**Gini Index attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | **0** |
| Temperature (F2) | 0.27 |
| Humidty (F3) | **0** |

From the above table, we observe that ‘Weather (F1)’ or Humidty (F3) have the lowest Gini Index and hence they will be chosen as the child node for the ‘Strong’ branch of Wind (F4).

|  |  |
| --- | --- |
|  |  |

1. **Build a decision tree by using Information Gain (i.e., IG (T, a) = Entropy (T) – Entropy (T |a), More information about IG).**

The first thing that we need to do is work out which feature to use as the root node. We start by computing the entropy of hiking (labels):

🡪 P (Yes) = , P (No) =

Entropy (Hiking) = - - = 0.881

**We will calculate probabilities of classes in F1, F2, F3, and F4 in this table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** |
| P(F1 = Cloudy) = | P(F2 = Cool) = | P(F3 = Normal) = | P(F4 = Weak) = |
| P(F1 = Sunny) = | P(F2 = Hot) = | P(F3 = High) = | P(F4 = Strong) = |
| P(F1 = Rainy) = | P(F2 = Mild) = |  |  |

**We will calculate the Information Gain for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Cloudy and Hiking = Yes) = | P(F1 = Cloudy and Hiking = No) = |
| P(F1 = Sunny and Hiking = Yes) = | P(F1 = Sunny and Hiking = No) = |
| P(F1 = Rainy and Hiking = Yes) = | P(F1 = Rainy and Hiking = No) = |

GAIN (Hiking, Weather (F1)) = 0.881 - Entropy ()

* Entropy ()
* Entropy ()

GAIN (Hiking, Weather (F1)) = 0.881 - (- - )

* (- - )
* (- - )

= 0.881 - 0.275 - 0.234 - 0.275 = 0.097

**We will calculate the Information Gain for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Cool and Hiking = Yes) = | P(F2 = Cool and Hiking = No) = |
| P(F2 = Hot and Hiking = Yes) = | P(F2 = Hot and Hiking = No) = |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |

GAIN (Hiking, Temperature (F2)) = 0.881 - Entropy ()

* Entropy ()
* Entropy ()

GAIN (Hiking, Temperature (F2)) = 0.881 - (- - )

* (- - )
* (- - )

= 0.881 - 0 - 0.275 - 0.4 = 0.206

**We will calculate the Information Gain for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = Normal and Hiking = Yes) = | P(F3 = Normal and Hiking = No) = |
| P(F3 = High and Hiking = Yes) = | P(F3 = High and Hiking = No) = |

GAIN (Hiking, Humidty (F3)) = 0.881 - Entropy ()

* Entropy ()

GAIN (Hiking, Humidty (F3)) = 0.881 - (- - )

* (- - )

= 0.881 - 0.324 - 0.551 = 0.006

**We will calculate the Information Gain for Wind (F4)**

|  |  |
| --- | --- |
| Wind(F4) | |
| P(F4 = Weak and Hiking = Yes) = | P(F4 = Weak and Hiking = No) = |
| P(F4 = Strong and Hiking = Yes) = | P(F4 = Strong and Hiking = No) = |

GAIN (Hiking, Wind (F4)) = 0.881 - Entropy ()

* Entropy ()

GAIN (Hiking, Wind (F4)) = 0.881 - (- - )

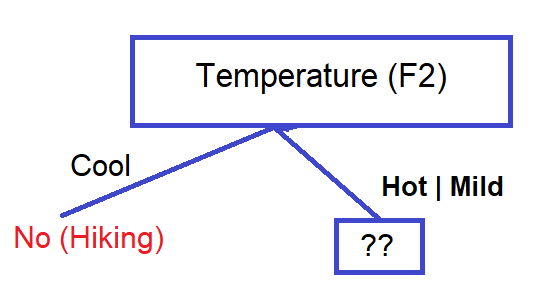
* (- - )

= 0.881 - 0.4 - 0.39 = 0.091

**Information Gain** **attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | 0.097 |
| Temperature (F2) | **0.206** |
| Humidty (F3) | 0.006 |
| Wind (F4) | 0.091 |

From the above table, we observe that ‘Temperature (F2)’ has the highest Information Gain and hence it will be chosen as the root node for how decision tree works.



We will repeat the same procedure to determine the sub-nodes or branches of the decision tree.

We will calculate the Information Gain for the ‘Hot | Mild’ branch of Temperature (F2) as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** | **Hiking** |
| Sunny | Hot | High | Weak | Yes |
| Rainy | Mild | Normal | Strong | Yes |
| Cloudy | Mild | High | Strong | No |
| Sunny | Mild | High | Strong | No |
| Cloudy | Mild | High | Weak | Yes |
| Sunny | Hot | High | Strong | No |
| Sunny | Hot | High | Strong | No |

We start by computing the entropy of hiking (labels) in the above table:

🡪 P (Yes) = , P (No) =

Entropy (Hiking) = - - = 0.985

**We will calculate probabilities of classes in F1, F2, F3, and F4 of this table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** |
| P(F1 = Sunny) = | P(F2 = Hot) = | P(F3 = High) = | P(F4 = Weak) = |
| P(F1 = Sunny) = | P(F2 = Mild) = | P(F3 = Normal) = | P(F4 = Strong) = |
| P(F1 = Cloudy) = |  |  |  |

**We will calculate the Information Gain for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Sunny and Hiking = Yes) = | P(F1 = Sunny and Hiking = No) = |
| P(F1 = Rainy and Hiking = Yes) = |  |
| P(F1 = Cloudy and Hiking = Yes) = | P(F1 = Cloudy and Hiking = No) = |

GAIN (Hiking, Weather (F1)) = 0.985 - Entropy ()

- Entropy ()

- Entropy ()

GAIN (Hiking, Weather (F1)) = 0.985- (- - )

- (- ) - (- )

= 0.985 - 0.46 - 0 - 0.286 = 0.239

**We will calculate the Gini Index for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Hot and Hiking = Yes) = | P(F2 = Hot and Hiking = No) = |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |

GAIN (Hiking, Temperature (F2)) = 0.985 - Entropy ()

- Entropy ()

GAIN (Hiking, Temperature (F2)) = 0.985- (- - )

- (-

= 0.985 - 0.393 - 0.571 = 0.021

**We will calculate the Information Gain for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = High and Hiking = Yes) = | P(F3 = High and Hiking = No) = |
| P(F3 = Normal and Hiking = Yes) = |  |

GAIN (Hiking, Humidty (F3)) = 0. 985 - Entropy ()

- Entropy ()

GAIN (Hiking, Humidty (F3)) = 0. 985 - (- - ) - (-

= 0. 985 - 0.787 - 0 = 0.198

**We will calculate the Information Gain for Wind (F4)**

|  |  |
| --- | --- |
| Wind (F4) | |
| P(F4 = Weak and Hiking = Yes) = |  |
| P(F4 = Strong and Hiking = Yes) = | P(F4 = Strong and Hiking = No) = |

GAIN (Hiking, Wind (F4)) = 0. 985 - Entropy ()

- Entropy ()

GAIN (Hiking, Wind (F4)) = 0. 985 - (- )

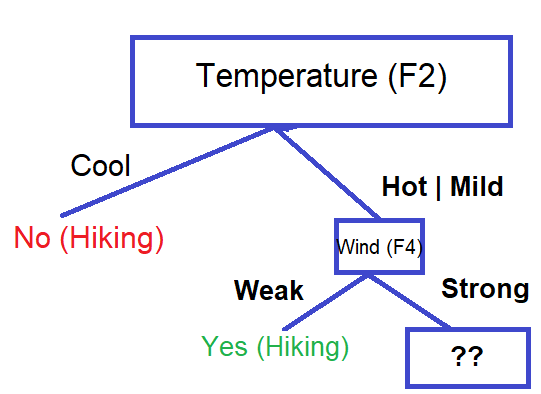
- (- - )

= 0. 985 - 0 - 0.516 = 0.469

**Information Gain attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | 0.239 |
| Temperature (F2) | 0.021 |
| Humidty (F3) | 0.198 |
| Wind (F4) | **0.469** |

From the above table, we observe that ‘Wind (F4)’ has the highest Information Gain and hence it will be chosen as the child node for the ‘Hot’ branch of Temperature (F2).



We will calculate the Information Gain for the ‘Strong’ branch of Wind (F4) as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** | **Wind(F4)** | **Hiking** |
| Rainy | Mild | Normal | Strong | Yes |
| Cloudy | Mild | High | Strong | No |
| Sunny | Mild | High | Strong | No |
| Sunny | Hot | High | Strong | No |
| Sunny | Hot | High | Strong | No |

We start by computing the entropy of hiking (labels) in the above table:

🡪 P (Yes) = , P (No) =

Entropy (Hiking) = - - = 0.722

**We will calculate probabilities of classes in F1, F2, and F3 of the above table:**

|  |  |  |
| --- | --- | --- |
| **Weather (F1)** | **Temperature (F2)** | **Humidty (F3)** |
| P(F1 = Rainy) = | P(F2 = Mild) = | P(F3 = Normal) = |
| P(F1 = Cloudy) = | P(F2 = Hot) = | P(F3 = High) = |
| P(F1 = Sunny) = |  |  |

**We will calculate the Information Gain for Weather (F1)**

|  |  |
| --- | --- |
| Weather (F1) | |
| P(F1 = Rainy and Hiking = Yes) = |  |
|  | P(F1 = Cloudy and Hiking = No) = |
|  | P(F1 = Sunny and Hiking = Yes) = |

GAIN (Hiking, Weather (F1)) = 0.722- Entropy ()

- Entropy ()

- Entropy ()

GAIN (Hiking, Weather (F1)) = 0.722 - (- ) - (- ) - (- )

= 0.722 - 0 - 0 - 0 = 0.722

**We will calculate the Information Gain for Temperature (F2)**

|  |  |
| --- | --- |
| Temperature (F2) | |
| P(F2 = Mild and Hiking = Yes) = | P(F2 = Mild and Hiking = No) = |
|  | P(F2 = Hot and Hiking = No) = |

GAIN (Hiking, Temperature (F2)) = 0.722- Entropy ()

- Entropy ()

GAIN (Hiking, Temperature (F2)) = 0.722 - (- ) - (- )

= 0.722 - 0.551 - 0 = 0.171

**We will calculate the Information Gain for Humidty (F3)**

|  |  |
| --- | --- |
| Humidty (F3) | |
| P(F3 = Normal and Hiking = Yes) = |  |
|  | P(F3 = High and Hiking = No) = |

GAIN (Hiking, Humidty (F3)) = 0.722- Entropy ()

- Entropy ()

GAIN (Hiking, Humidty (F3)) = 0.722 - (- ) - (- )

= 0.722 - 0 - 0 = 0.722

**Information Gain attributes or features**

|  |  |
| --- | --- |
| Weather (F1) | **0.722** |
| Temperature (F2) | 0.171 |
| Humidty (F3) | **0.722** |

From the above table, we observe that ‘Weather (F1)’ or Humidty (F3) have the highest Information Gain and hence it will be chosen as the child node the ‘Strong’ branch of Wind (F4).

|  |  |
| --- | --- |
|  |  |

1. **Compare the advantages and disadvantages between Gini Index and Information Gain.**

|  |  |  |
| --- | --- | --- |
|  | **Gini Index** | **Information Gain** |
| **The advantages** | * It favors larger partitions (distributions) and is very easy to implement. * It can handle the values that are non-negative because it is measured by subtracting the sum of squared probabilities of each class from one. * It computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. | * It favors partitions that have small counts but many distinct values. * It measures the entropy differences before and after splitting and depicts the impurity in class variables. * It use Entropy as the base calculation, you have a wider range of results. * It computes the difference between entropy before and after split and specifies the impurity in class elements. * It determines the reduction of the uncertainty after splitting the dataset on a particular feature such that if the value of information gain increases. * The feature having the highest value of information gain is accounted for as the best feature to be chosen for split. |
| **The disadvantages** | * The Gini Index doesn’t have a wider range of results, but it caps at one. * While working on categorical data variables, gini index gives results either in “success” or “failure” and performs binary splitting only. * It is prone to systematic and random data errors. Therefore, inaccurate data can distort the validity of the coefficient. | * It is not preferred as it involves ‘log’ function that results in the computational complexity. * It can’t handle the values that are non-negative. * It supports smaller partitions (distributions) with various distinct values; there is a need to perform an experiment with data and splitting criterion. |