**Name: A. Shiva Surya Saran**

**Roll No.: BE-B 20**

**Data Analytics**

**Problem 2**

**Decision Trees: ID3 Algorithm**

Dataset:

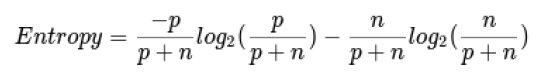
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 1 | No | No | No | No |
| 2 | Yes | Yes | Yes | Yes |
| 3 | Yes | Yes | No | No |
| 4 | Yes | No | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 6 | No | Yes | No | No |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |
| 9 | No | Yes | Yes | Yes |
| 10 | Yes | Yes | No | Yes |
| 11 | No | Yes | No | No |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |
| 14 | Yes | Yes | No | No |

Infected attribute will be used as Decision factor

P = 8

N = 6

Total = 14



Dataset (Entropy) = ((-P/(P+N)\*(LOG((P/(P+N)),2)))-((N/(P+N))\*(LOG((N/(P+N)),2))))

**Dataset (Entropy) = 0.985228**

Now calculating Entropy for Fever Attribute

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 3 | Yes | No | No | No |
| 14 | Yes | No | No | No |
| 2 | Yes | Yes | Yes | Yes |
| 4 | Yes | Yes | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 7 | Yes | Yes | Yes | Yes |
| 8 | Yes | Yes | Yes | Yes |
| 10 | Yes | No | No | Yes |
|  |  |  |  |  |
|  |  |  |  |  |
| ID | Fever | Cough | Breathing Issues | Infected |
| 1 | No | No | No | No |
| 6 | No | No | No | No |
| 11 | No | No | No | No |
| 13 | No | Yes | Yes | No |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Fever | p | n | Entropy |
| Yes | 6 | 2 | 0.811278 |
| No | 2 | 4 | 0.918296 |

Average Information Entropy for Fever = 0.8571

Gain (Fever) = 0.128085143

Now calculating Entropy for Cough Attribute

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 3 | Yes | Yes | No | No |
| 5 | Yes | Yes | Yes | Yes |
| 6 | No | Yes | No | No |
| 9 | No | Yes | Yes | Yes |
| 10 | Yes | Yes | No | Yes |
| 11 | No | Yes | No | No |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |
| 14 | Yes | Yes | No | No |
|  |  |  |  |  |
| ID | Fever | Cough | Breathing Issues | Infected |
| 1 | No | No | No | No |
| 4 | Yes | No | Yes | Yes |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Cough | p | n | Entropy |
| Yes | 5 | 5 | 1 |
| N | 3 | 1 | 0.811278 |

Average Information Entropy = 0.946079464

Gain = 0.039148536

Now calculating Entropy for Breathing Issues Attribute

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 4 | Yes | No | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |
|  |  |  |  |  |
| ID | Fever | Cough | Breathing Issues | Infected |
| 1 | No | No | No | No |
| 3 | Yes | Yes | No | No |
| 6 | No | Yes | No | No |
| 10 | Yes | Yes | No | Yes |
| 11 | No | Yes | No | No |
| 14 | Yes | Yes | No | No |

|  |  |  |  |
| --- | --- | --- | --- |
| Breathing Issues Attribute | p | n | Entropy |
| Yes | 7 | 1 | 0.543564 |
| N | 1 | 5 | 0.650022 |

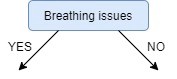
|  |  |
| --- | --- |
| Average Information Entropy | 0.589189 |

|  |  |
| --- | --- |
| Gain | 0.396039 |

Now, after comparing the Gains of each attribute:

|  |  |
| --- | --- |
| Attribute | Gain |
| Fever | 0.128085 |
| Cough | 0.039149 |
| Breathing Issues | 0.396039 |

Breathing has the highest gain. Therefore, the root node will be Breathing issues.



Next, from the remaining two unused features, namely, Fever and Cough, we decide which one is the best for the left branch of Breathing Issues. Since the left branch of Breathing Issues denotes YES, we will work with the subset of the original data i.e the set of rows having YES as the value in the Breathing Issues column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 4 | Yes | No | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |

New Entropy for above Subset is 0.543564443

Now Calculating Gain for Cough

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |

|  |  |  |  |
| --- | --- | --- | --- |
| Cough | p | n | Entropy |
| Yes | 4 | 1 | 0.721928 |
| N | 1 | 0 | 0 |

|  |  |
| --- | --- |
| Average Information Entropy | 0.451205 |
| Gain | 0.092359 |

Now Calculating Gain for Fever

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 4 | Yes | No | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |

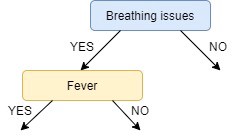
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |

|  |  |  |  |
| --- | --- | --- | --- |
| Fever | p | n | Entropy |
| Yes | 1 | 0 | 0 |
| N | 2 | 1 | 0.918296 |
| Average Information Entropy | | | 0.344361 |
| Gain | | | 0.199203 |

Now comparing gains of Fever and Cough

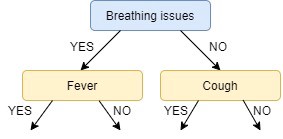
|  |  |
| --- | --- |
| Attribute | Gain |
| Fever | 0.199203 |
| Cough | 0.09 |
| Breathing Issues | 0.396039 |

Gain Fever is greater than that of Cough, so we select Fever as the left branch of Breathing Issues:



Next, we find the feature with the maximum IG for the right branch of Breathing Issues. But, since there is only one unused feature left we have no other choice but to make it the right branch of the root node.

So our tree now looks like this:



There are no more unused features, so we stop here and jump to the final step of creating the leaf nodes.

For the left leaf node of Fever, we see the subset of rows from the original data set that has Breathing Issues and Fever both values as YES.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 2 | Yes | Yes | Yes | Yes |
| 4 | Yes | No | Yes | Yes |
| 5 | Yes | Yes | Yes | Yes |
| 7 | Yes | No | Yes | Yes |
| 8 | Yes | No | Yes | Yes |

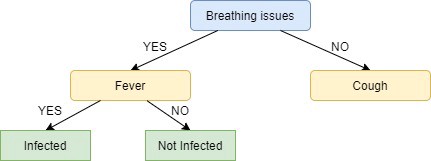
Since all the values in the target column are YES, we label the left leaf node as infected.

Similarly, for the right node of Fever we see the subset of rows from the original data set that have Breathing Issues value as YES and Fever as NO.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 9 | No | Yes | Yes | Yes |
| 12 | No | Yes | Yes | Yes |
| 13 | No | Yes | Yes | No |

Here not all but most of the values are NO, hence Not Infected becomes our right leaf node.

Tree looks like this:

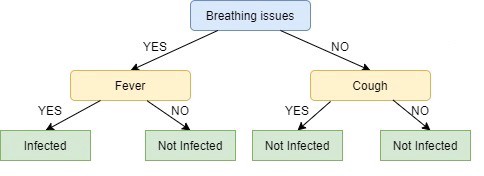


Now, we will repeat the same process for Cough to find out the leaf nodes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 1 | No | No | No | No |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Fever | Cough | Breathing Issues | Infected |
| 3 | Yes | Yes | No | No |
| 6 | No | Yes | No | No |
| 10 | Yes | Yes | No | Yes |
| 11 | No | Yes | No | No |
| 14 | Yes | Yes | No | No |

Since majority of the outcome is n ie. Not infected, the final tree looks like this.



***References***:

Link - <https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1>

Excel Formulae:

Entropy = = ((-H19/(H19+H20)\*(LOG((H19/(H19+H20)),2)))-((H20/(H19+H20))\*(LOG((H20/(H19+H20)),2))))

|  |  |
| --- | --- |
| P | H19 |
| N | H20 |

Average Information Entropy =((((M31+N31)/(J24+J25))\*O31)+(((M32+N32)/(J24+J25))\*O32))

|  |  |
| --- | --- |
| P | 7 |
| N | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Fever | p | n | Entropy |
| Yes | 1 | 0 | 0 |
| N | 2 | 1 | 0.918296 |