Example of using Entropy as a measure of node impurity in decision tree:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temperature | H | W | **Play Tennis** |
| D1 | Sunny | Hot | High | Weak | N |
| D2 | Sunny | Hot | High | Strong | N |
| D3 | Overcast | Hot | High | Weak | Y |
| D4 | Rain | Mild | High | Weak | Y |
| D5 | Rain | Cool | Normal | Weak | Y |
| D6 | Rain | Cool | Normal | Strong | N |
| D7 | Overcast | Cool | Normal | Strong | Y |
| D8 | Sunny | Mild | High | Weak | N |
| D9 | Sunny | Cool | Normal | Weak | Y |
| D10 | Rain | Mild | Normal | Weak | Y |
| D11 | Sunny | Mild | Normal | Strong | Y |
| D12 | Overcast | Mild | High | Strong | Y |
| D13 | Overcast | Hot | Normal | Weak | Y |
| D14 | Rain | Mild | High | Strong | N |

Entropy of the system: E(S) = E (9+,5-) = -9/14\*log2(9/14) -5/14\*log2(5/14) = 0.94

|  |  |  |
| --- | --- | --- |
| Outlook | | |
| Sunny | Overcast | Rain |
| [2+,3-] = 0.971 | [4+,0] = 0 | [3+,2-] = 0.971 |

|  |  |  |
| --- | --- | --- |
| Temperature | | |
| Hot | Mild | Cool |
| [2+,2-] = 1 | [4+,2-] = 0.918 | [3+,1-] = 0.811 |

|  |  |
| --- | --- |
| Humidity | |
| High | Normal |
| [3+,4-] = 0.985 | [6+,1-] = 0.592 |

|  |  |
| --- | --- |
| Wind | |
| Weak | Strong |
| [6+,2-] = 0.811 | [3+,3-] = 1 |

G (S, Outlook) = E(S) – 5/14 \* E(Sunny) – 4/14 \* E(Overcast) – 5/14 \* E(Rain) = **0.246**

G (S, Temperature) = E(S) – 4/14 \* E(Hot) – 6/14 \* E(Mild) – 4/14 \* E(Cool) = **0.029**

G (S, Humidity) = E(S) – 7/14 \* E(High) – 7/14 \* E(Normal) = **0.151**

G (S, Wind) = E(S) – 8/14 \* E(Weak) – 6/14 \* E(Strong) = **0.048**

We compare all the Gain values and the winner is Outlook having the greatest gain.

Outlook

?

?

Yes Play Tennis

Rain

Overcast

Sunny

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | T | H | W | Play Tennis |
| D1 | Sunny | Hot | High | Weak | N |
| D2 | Sunny | Hot | High | Strong | N |
| D8 | Sunny | Mild | High | Weak | N |
| D9 | Sunny | Cool | Normal | Weak | Y |
| D11 | Sunny | Mild | Normal | Strong | Y |

|  |  |  |
| --- | --- | --- |
| Temperature | | |
| Hot | Mild | Cool |
| [0+,2-] = 0 | [1+,1-] = 1 | [1+,0-] = 0 |

|  |  |
| --- | --- |
| Humidity | |
| High | Normal |
| [0+,3-] = 0 | [2+,0] = 0 |

|  |  |
| --- | --- |
| Wind | |
| Weak | Strong |
| [1+,2-] = 0.918 | [1+,1-] = 1 |

G (Sunny, Temperature) = 0.971 – 2/5 \* 1 = **0.571**

G (Sunny, Humidity) = 0.971 – 0 – 0 = **0.971**

G (Sunny, Wind) = 0.971 – 3/5 \* E(Weak) – 2/5 \* E(Strong) = **0.02**

Comparing the previous gains; Humidity is the winner having the greatest gain.

?

Yes Play Tennis

YES

NO

Normal

High

Humidity

Outlook

Sunny

Overcast

Rain

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | T | H | W | Play Tennis |
| D4 | Rain | Mild | High | Weak | Y |
| D5 | Rain | Cool | Normal | Weak | Y |
| D6 | Rain | Cool | Normal | Strong | N |
| D10 | Rain | Mild | Normal | Weak | Y |
| D14 | Rain | Mild | High | Strong | N |

|  |  |
| --- | --- |
| Temperature | |
| Mild | Cool |
| [2+,1-] = 0.918 | [1+,1-] = 1 |

|  |  |
| --- | --- |
| Humidity | |
| High | Normal |
| [1+,1-] = 1 | [2+,1-] = 0.918 |

|  |  |
| --- | --- |
| Wind | |
| Weak | Strong |
| [3+,0-] = 0 | [0+,2-] = 0 |

G (Rain, Temperature) = 0.971 – 3/5 \* E(Mild) – 2/5 \* E(Cool) = **0.02**

G (Rain, Humidity) = 0.971 – 3/5 \* E(High) – 2/5 \* E(Normal) = **0.02**

G (Rain, Wind) = 0.971 – 3/5 \* E(Weak) – 2/5 \* E(Strong) = **0.971**

Again, by comparing the G values we can see that Wind is the winner.

NO

YES

Strong

Weak

Wind

Yes Play Tennis

Overcast

Outlook

Sunny

Rain

Humidity

High

Normal

NO

YES