Our dataset consists of past days with information about the outlook (Sunny, Overcast, Rainy), temperature (Hot, Mild, Cool), humidity (High, Normal), wind (Weak, Strong), and whether or not we went to the picnic (Yes/No).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlook** | **Temperature** | **Humidity** | **Wind** | **Picnic** |
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Overcast | Hot | High | Weak | Yes |
| Rainy | Mild | High | Weak | Yes |
| Rainy | Cool | Normal | Weak | Yes |
| Rainy | Cool | Normal | Strong | No |
| Overcast | Cool | Normal | Strong | Yes |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Rainy | Mild | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |
| Rainy | Mild | High | Strong | No |

Export to Sheets

**1. Entropy**

Entropy measures the impurity or disorder in a dataset. In our picnic example, let's calculate the entropy of the 'Picnic' decision.

Total instances = 14

Number of 'Yes' = 9

Number of 'No' = 5

The probability of 'Yes' is P(Yes)=9/14

The probability of 'No' is P(No)=5/14

Entropy (S) = −P(Yes)log2​(P(Yes))−P(No)log2​(P(No))

Entropy (S) = −(9/14)log2​(9/14)−(5/14)log2​(5/14)≈0.94

This value indicates a moderate level of disorder in our target variable 'Picnic'.

**2. Information Gain**

Information Gain measures the reduction in entropy achieved by splitting the dataset on a particular attribute. Let's calculate the information gain for the 'Outlook' attribute.

Entropy of the original dataset (as calculated above) ≈0.94

Now, let's look at the entropy after splitting on 'Outlook':

* **Outlook = Sunny (5 instances: Yes=2, No=3):** Entropy(Sunny) = −(2/5)log2​(2/5)−(3/5)log2​(3/5)≈0.971
* **Outlook = Overcast (4 instances: Yes=4, No=0):** Entropy(Overcast) = −(4/4)log2​(4/4)−(0/4)log2​(0/4)=0 (perfectly classified)
* **Outlook = Rainy (5 instances: Yes=3, No=2):** Entropy(Rainy) = −(3/5)log2​(3/5)−(2/5)log2​(2/5)≈0.971

The weighted average entropy after splitting on 'Outlook' is: (5/14)∗0.971+(4/14)∗0+(5/14)∗0.971≈0.693

Information Gain (Outlook) = Entropy(S) - Weighted Average Entropy(Outlook)

Information Gain (Outlook) ≈0.94−0.693=0.247

This positive information gain suggests that 'Outlook' is a useful attribute for splitting our data to predict whether or not to go for a picnic.

**3. Gini Index**

The Gini Index measures the impurity of a dataset based on the probability distribution of the classes. A Gini Index of 0 indicates perfect purity. Let's calculate the Gini Index for the original 'Picnic' decision.

P(Yes)=9/14

P(No)=5/14

Gini(S) = 1−(P(Yes)2+P(No)2) Gini(S) = 1−((9/14)2+(5/14)2)≈1−(0.413+0.128)=0.459

Now, let's calculate the Gini Index for the 'Outlook' attribute:

* **Outlook = Sunny:** Gini(Sunny) = 1−((2/5)2+(3/5)2)=1−(0.16+0.36)=0.48
* **Outlook = Overcast:** Gini(Overcast) = 1−((4/4)2+(0/4)2)=1−(1+0)=0
* **Outlook = Rainy:** Gini(Rainy) = 1−((3/5)2+(2/5)2)=1−(0.36+0.16)=0.48

The weighted average Gini Index after splitting on 'Outlook' is: (5/14)∗0.48+(4/14)∗0+(5/14)∗0.48≈0.343

A lower Gini Index after the split indicates a reduction in impurity.