Information Gain and Entropy

This topic is useful in decision tree and random forest. It is essential in building machine learning model.

Entropy

Entropy measures the impurity or uncertainty present in the data.

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where:

- ***S – set of all instances in the dataset***
- ***N – number of distinct class values***
- ***pi – event probability***

Information Gain (IG)

IG indicates how much "information" a particular feature/ variable gives us about the final outcome.

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A,S)$$

where:

H(S) – entropy of the whole dataset S

- ***|Sj| – number of instance with j value of an attribute A***
- ***|S| – total number of instances in dataset S***
- ***v – set of distinct values of an attribute A***
- ***H(Sj) – entropy of subset of instances for attribute A***
- ***H(A, S) – entropy of an attribute A***

Image Source: www.edureka.co/data-science

Information Gain and Entropy

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

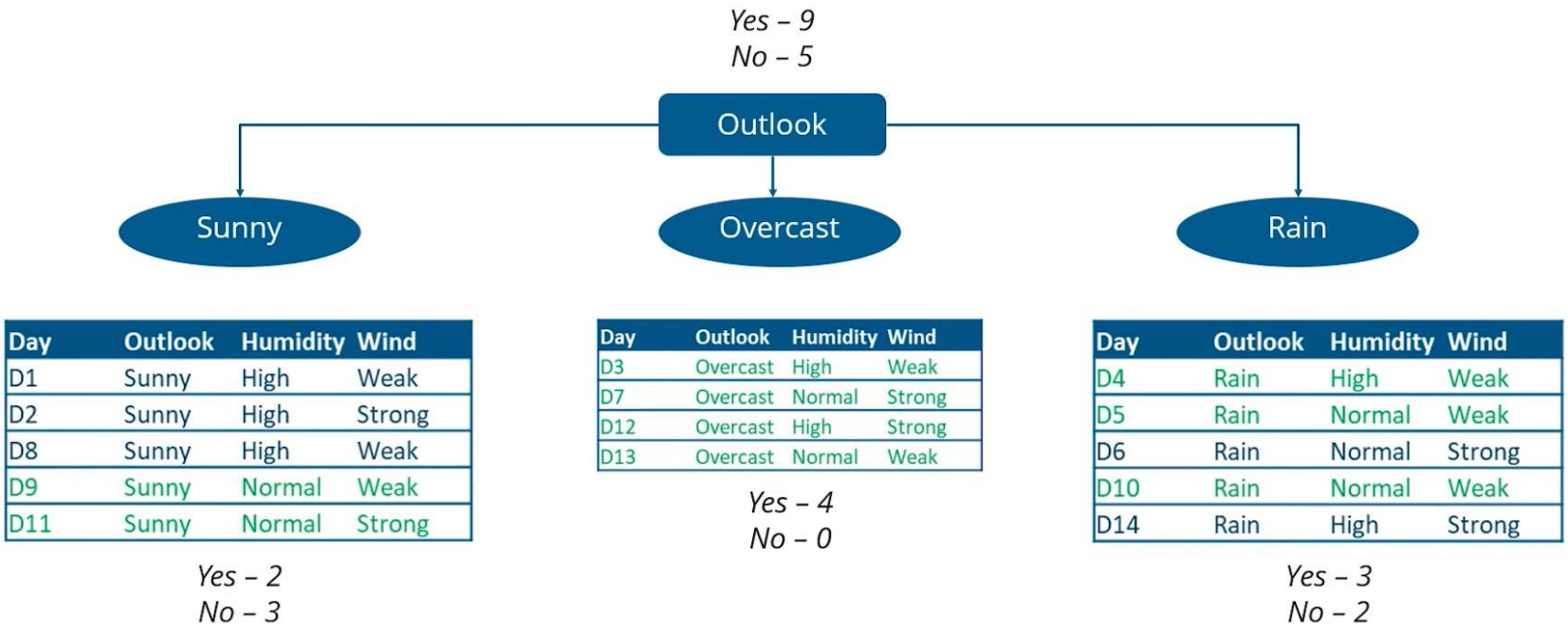
Target variable is "Play". Target variable is to be predicted.

Image Source: www.edureka.co/data-science

Information Gain and Entropy

Decision Tree

Image Source: www.edureka.co/data-science



Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

"Sunny" and "Rain" parameters have mixed outcome in form of "Yes" or "No" but for "Overcast" branch/variable, it results in a 100% pure subset. It shows that "Overcast" results into definite & certain output. This is exactly what entropy is used to measure. It calculates the impurities of uncertainty. The lesser the uncertainty of the entropy or variable, more significant is that variable. Hence, in "Overcast" there is literally no impurities. It is a 100% pure subset. So, we want variable like this to build our model.

Information Gain and Entropy

From the total of 14 instances we have:

- *9 instances "yes"*
- *5 instances "no"*

The Entropy is:

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Image Source: www.edureka.co/data-science

Information Gain and Entropy

Selecting the root variable

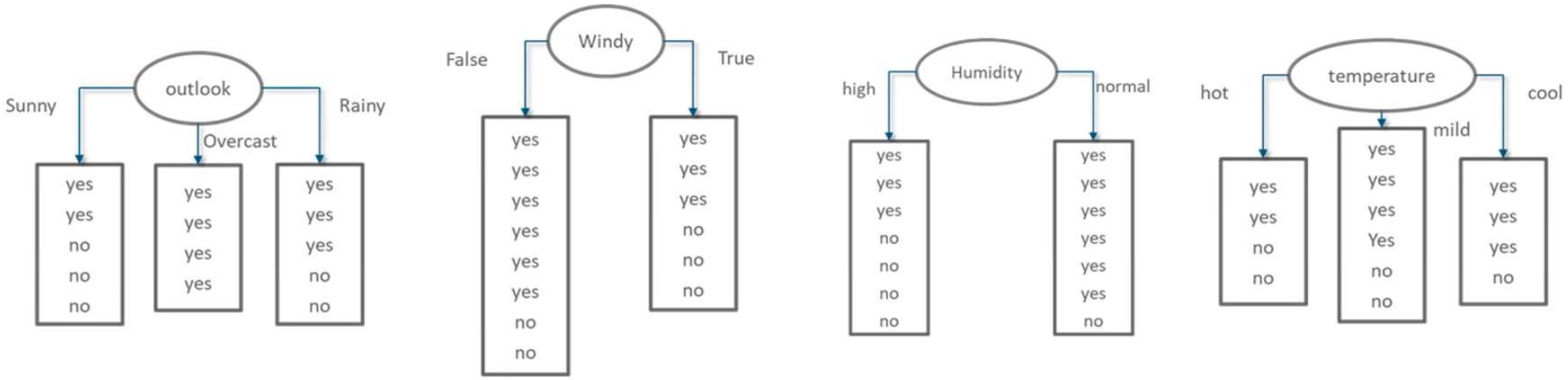


Image Source: www.edureka.co/data-science

Information Gain and Entropy

Information Gain of attribute "windy"

From the total of 14 instances we have:

- *6 instances "true"*
- *8 instances "false"*

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$Gain(A_{Windy}, S) = 0.940 - \frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048$$

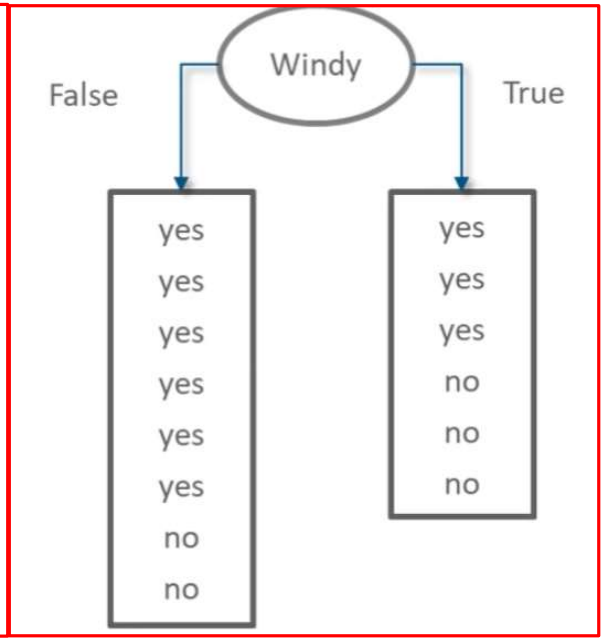


Image Source: www.edureka.co/data-science

Information Gain and Entropy

Information Gain of attribute "outlook"

From the total of 14 instances we have:

- *5 instances "sunny"*
- *4 instances "overcast"*
- *5 instances "rainy"*

$$Gain(A_{Outlook}, S) = 0.940 - \frac{5}{14} \cdot \left(-\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \right) + \frac{4}{14} \cdot \left(-\left(\frac{4}{4} \log_2 \frac{4}{4} \right) \right) + \frac{5}{14} \cdot \left(-\left(\frac{3}{5} \cdot \log_2 \frac{3}{5} + \frac{2}{5} \cdot \log_2 \frac{2}{5} \right) \right) = 0.247$$

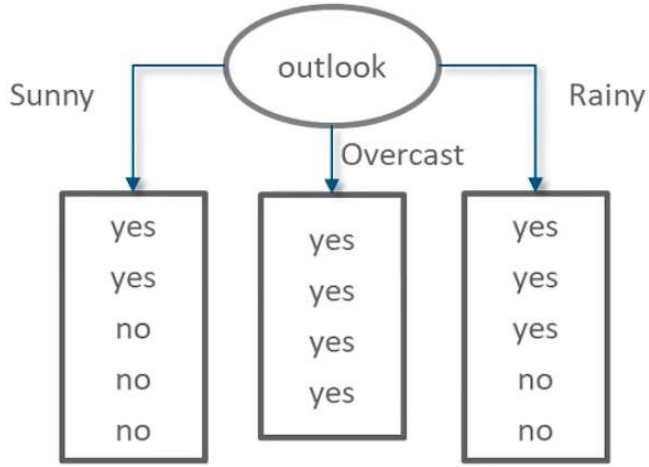


Image Source: www.edureka.co/data-science

Information Gain and Entropy

Information Gain of attribute "humidity"

From the total of 14 instances we have:

- *7 instances "high"*
- *7 instances "normal"*

$$Gain(A_{Humidity}, S) = 0.940 - \frac{7}{14} \cdot \left(-\left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7} \right) \right) + \frac{7}{14} \cdot \left(-\left(\frac{6}{7} \cdot \log_2 \frac{6}{7} + \frac{1}{7} \cdot \log_2 \frac{1}{7} \right) \right) = 0.151$$

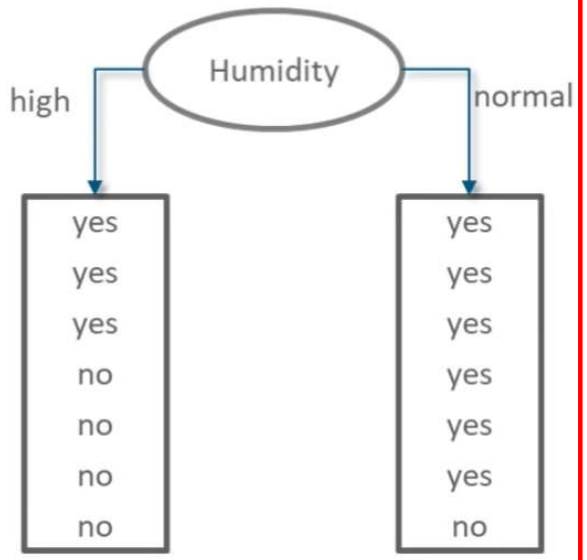


Image Source: www.edureka.co/data-science

Information Gain and Entropy

Information Gain of attribute "temperature"

From the total of 14 instances we have:

- *4 instances "hot"*
- *6 instances "mild"*
- *4 instances "cool"*

$$Gain(A_{Temperature}, S) = 0.940 -$$
$$\frac{4}{14} \cdot \left(-\left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4} \right) \right) +$$
$$\frac{6}{14} \cdot \left(-\left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6} \right) \right) +$$
$$\frac{4}{14} \cdot \left(-\left(\frac{3}{4} \cdot \log_2 \frac{3}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4} \right) \right) = 0.029$$

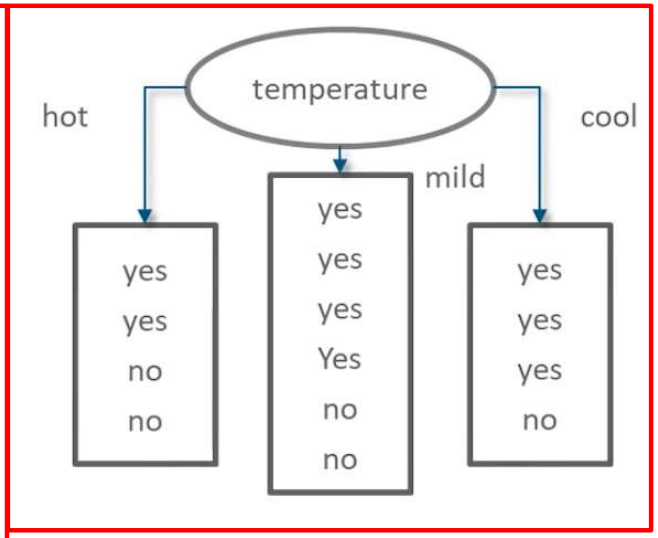
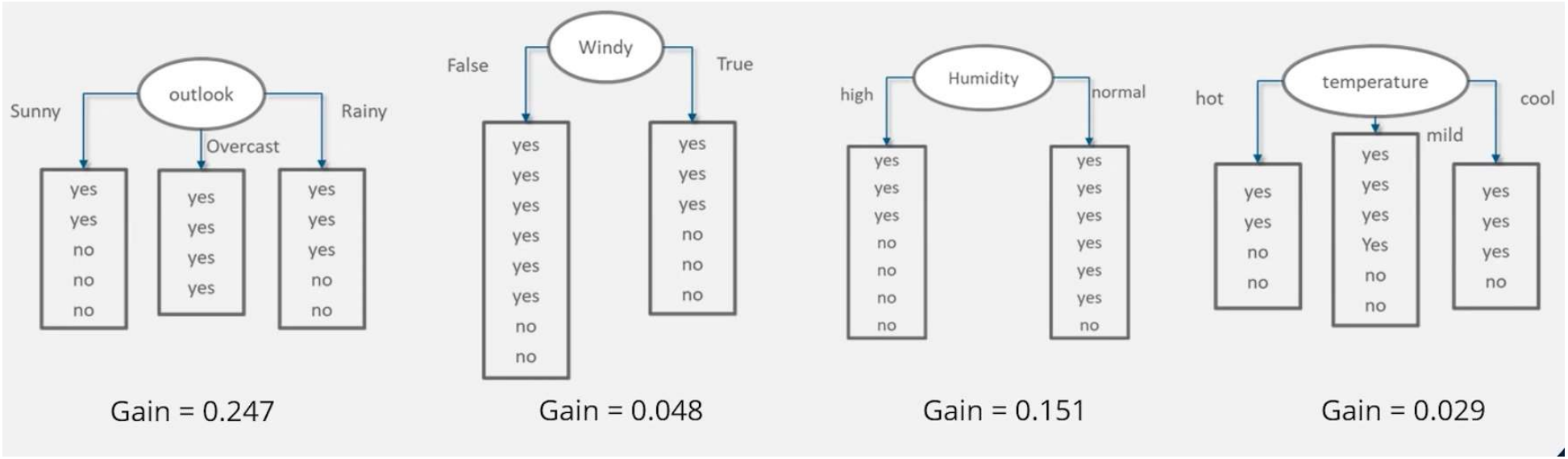


Image Source: www.edureka.co/data-science

Information Gain and Entropy



The variable with the highest information gain is used to split the data at the root node. That's we assign "outlook" as the root node, in this example.

Image Source: www.edureka.co/data-science

Confusion Matrix

It is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Confusion Matrix represents a tabular representation of Actual vs Predicted values
You can calculate the accuracy of your model with:

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

Image Source: www.edureka.co/data-science

Confusion Matrix

- There are two possible predicted classes: "yes" and "no"
- The classifier made a total of 165 predictions
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times
- In reality, 105 patients in the sample have the disease, and 60 patients do not

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Image Source: www.edureka.co/data-science

Information Gain and Entropy

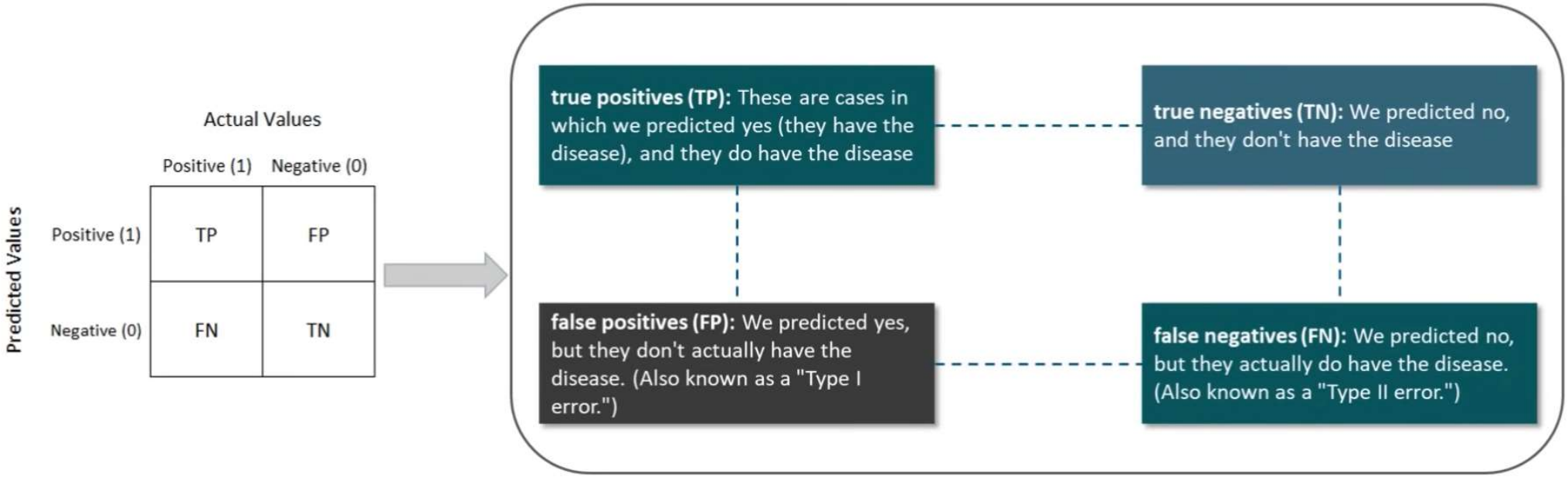


Image Source: www.edureka.co/data-science