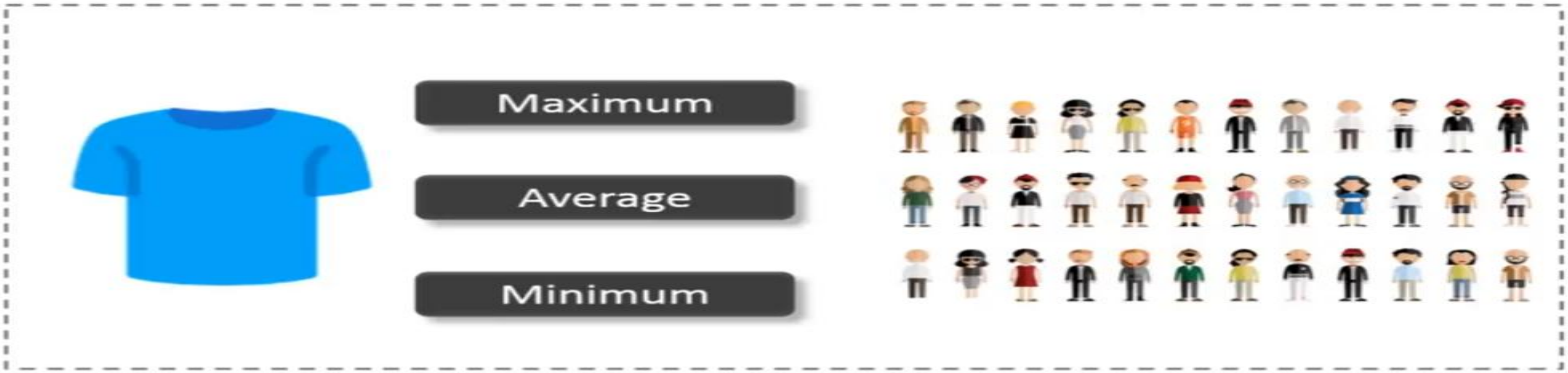
TYPES OF STATISTICS

Descriptive Statistics

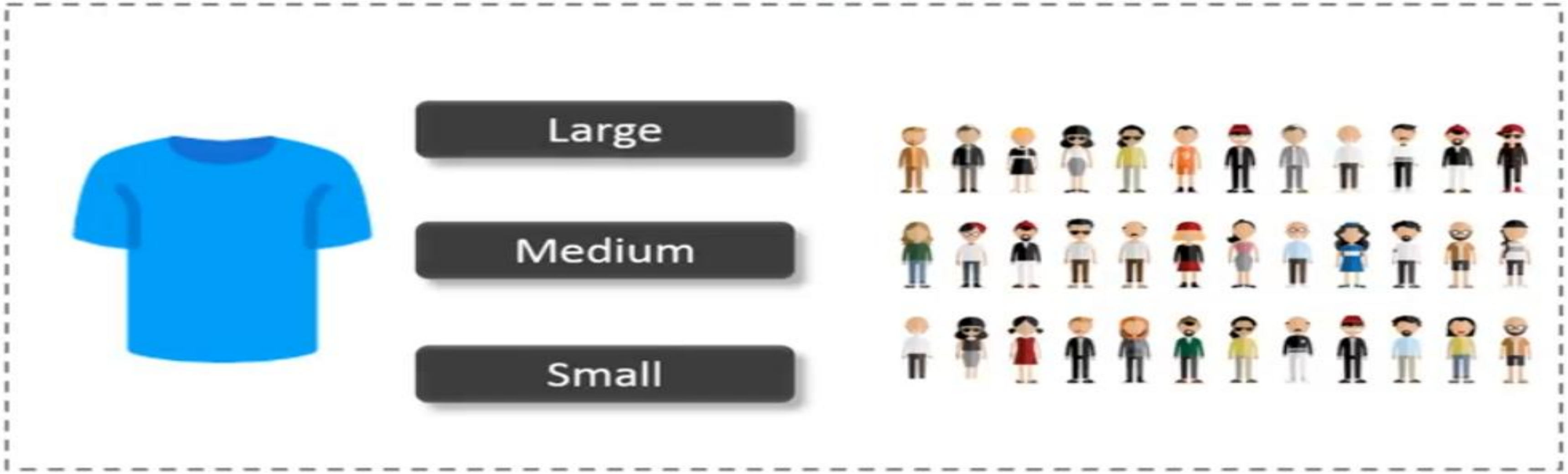
Descriptive statistics *uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.*



Descriptive Statistics is mainly focused upon the main characteristics of data. It provides graphical summary of the data.

Inferential Statistics

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.



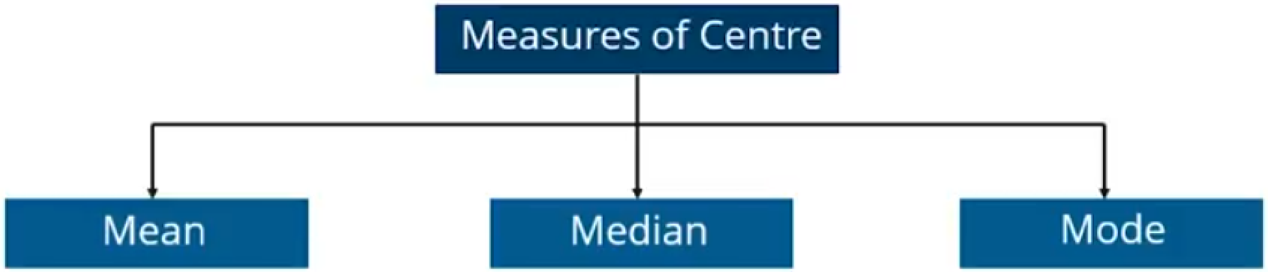
Inferential statistics, generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using a sample data.

Descriptive Statistics

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summari about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

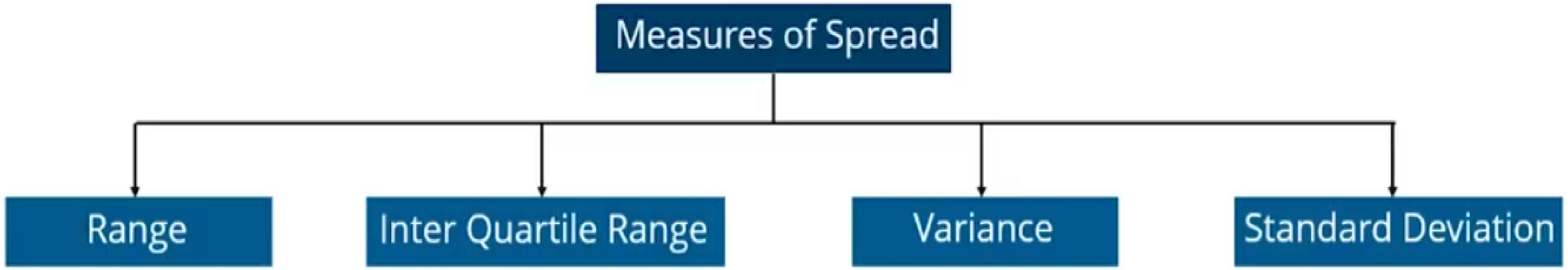
- **Measures of Central tendency**



Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Variability (spread)**

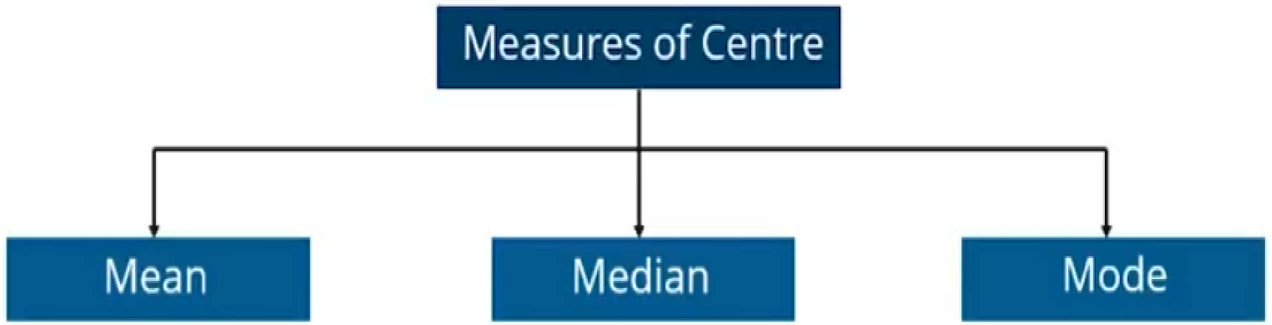


Measures of Centre

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summari about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Central tendency**

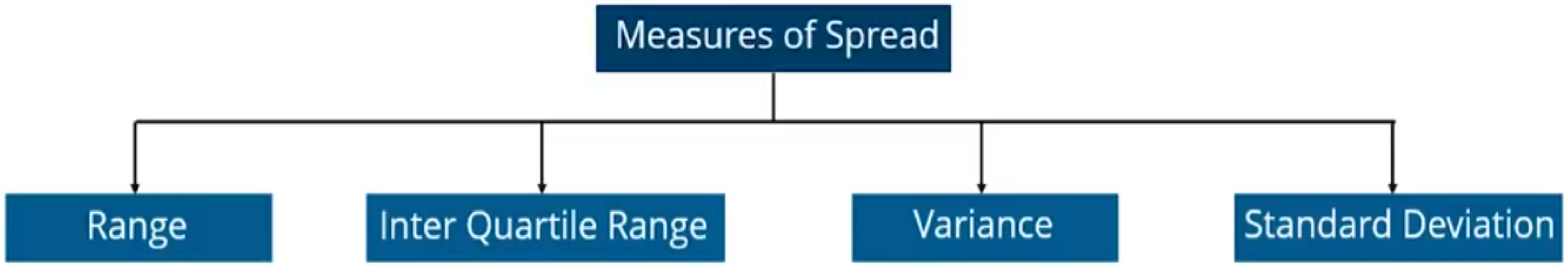


Measures of Spread

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Variability (spread)**



Mean

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

Measure of average of all the values in a sample is called Mean.

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

To find out the average horsepower of the cars among the population of cars, we will check and calculate the average of all values:

$$\frac{110+110+93+96+90+110+110+110}{8} = \mathbf{103.625}$$

Median

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

Measure of the central value of the sample set is called **Median.**

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

To find out the center value of mpg among the population of cars, arrange records in *Ascending order, i.e.,* ***21, 21, 21.3, 22.8, 23, 23, 23, 23***
In case of even entries, take average of the two middle values, i.e. (22.8+23)/2 = 22.9

Mode

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

The value most recurrent in the sample set is known as Mode.

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

To find the most common type of cylinder among the population of cars, check the value which is repeated most number of times, i.e., *cylinder type 6*

Measures of Spread

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Range is the given measure of how spread apart the values in a dataset are.

$$Range = Max(x_i) - Min(x_i)$$

Inter quartile Range

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

	Q1		Q2		Q3		
1	2	3	4	5	6	7	8

Example

Consider the marks of the 100 students below, ordered from the lowest to the highest scores

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

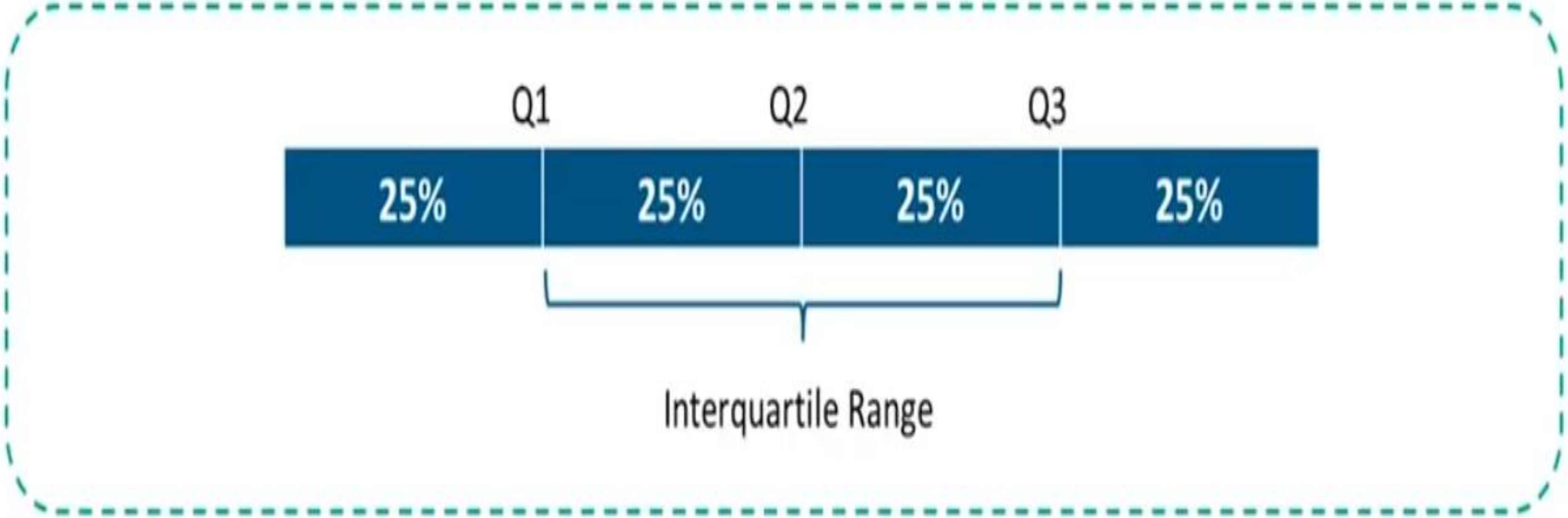
The first quartile (Q1) lies between the 25th and 26th.
Q1 = (45 + 45) ÷ 2 = 45

The second quartile (Q2) between the 50th and 51st.
Q2 = (58 + 59) ÷ 2 = 58.5

The third quartile (Q3) between the 75th and 76th.
Q3 = (71 + 71) ÷ 2 = 71

Inter Quartile Range(IQR) is the measure of variability, based on dividing a dataset into quartiles.

- *Quartiles divide a rank-ordered data set into four equal parts, denoted by Q1, Q2, and Q3, respectively*
- *The interquartile range is equal to Q3 minus Q1, i.e.. IQR = Q3 - Q1*


Q1
Q2
Q3
25%
25%
25%
25%
Interquartile Range

Variance

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Variance describes how much a random variable differs from its expected value.
It entails computing squares of deviations.

$$s^2 = \frac{\sum_{1}^{n=1} (x_i - \bar{x})^2}{n}$$

x : Individual data points
n : Total number of data points
$\bar{x}$ *: Mean of data points*

Standard Deviation

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | **Standard Deviation**

Deviation is the difference between each element from the mean.

Deviation = $(x_i - \mu)$

Population Variance is the average of squared deviations.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} = (x_i - \mu)^2$$

Sample Variance is the average of squared differences from the mean.

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{N} = (x_i - \bar{x})^2$$

Standard Deviation is the measure of the dispersion of a set of data from its mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Example

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 1

Find out the mean for your sample set.

The Mean is:

$$\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$$

μ=7

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 2

Then for each number, subtract the Mean and square the result.

$(x_i - \mu)^2$

$(9-7)^2$ $= 2^2=4$
$(2-7)^2$ $= (-5)^2=25$
$(5-7)^2$ $= (-2)^2=4$
And so on...

We get the following results:
4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 3

Then work out the mean of those squared differences.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}=(x_i-\mu)^2}$$

$$\frac{4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9}{20}$$

σ^2 = 8.9

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 4

Take square root of σ^2.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} = (x_i - \mu)^2}$$

□**σ = 2.983**

Information Gain and Entropy

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*

Use Case

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

Yes – 9
No – 5

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

Yes – 2
No – 3

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

Yes – 4
No – 0

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

Yes – 3
No – 2

From the total of 14 instances we have:

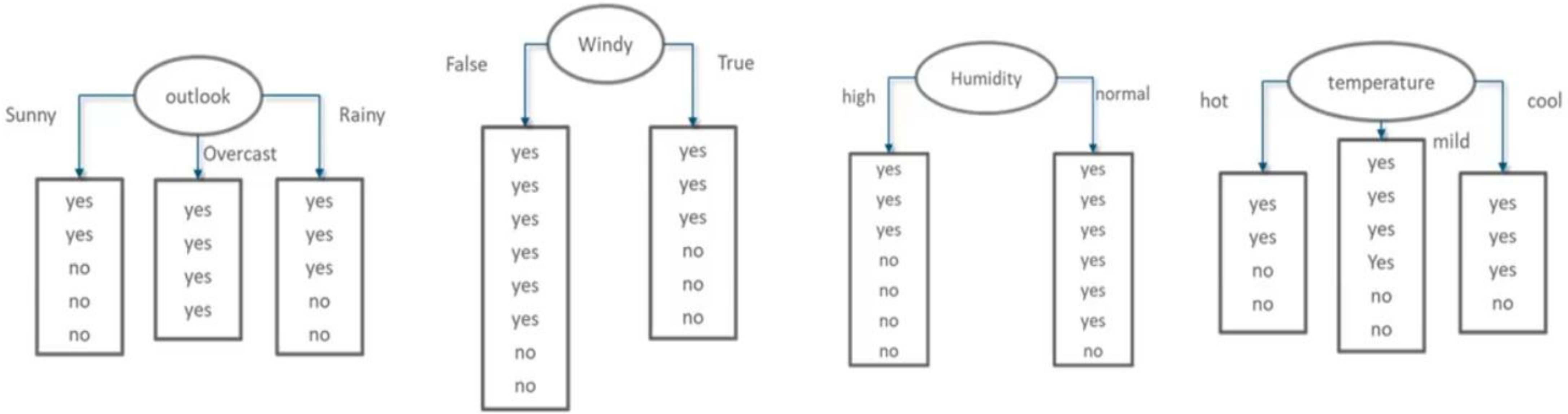
- *9 instances "yes"*
- *5 instances "no"*

The Entropy is:

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

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

Selecting the root variable



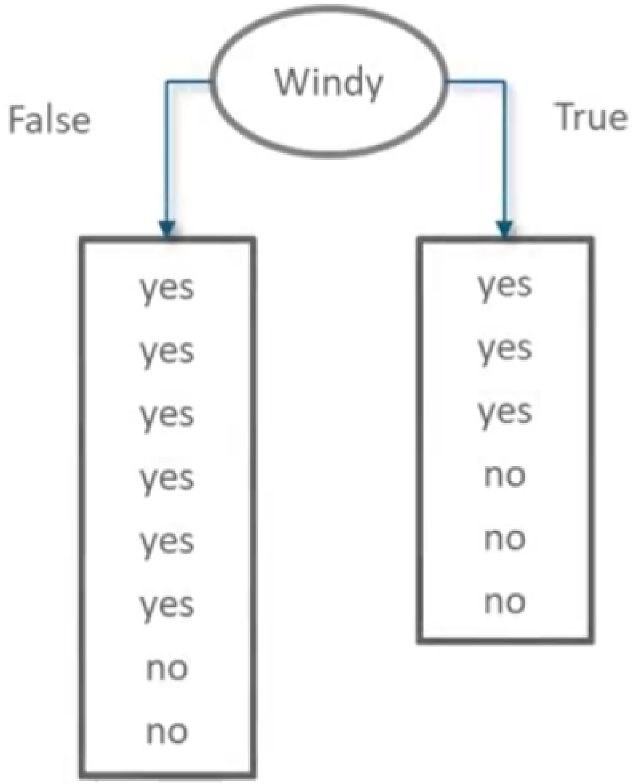
Information Gain of attribute "windy"

From the total of 14 instances we have:

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

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

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

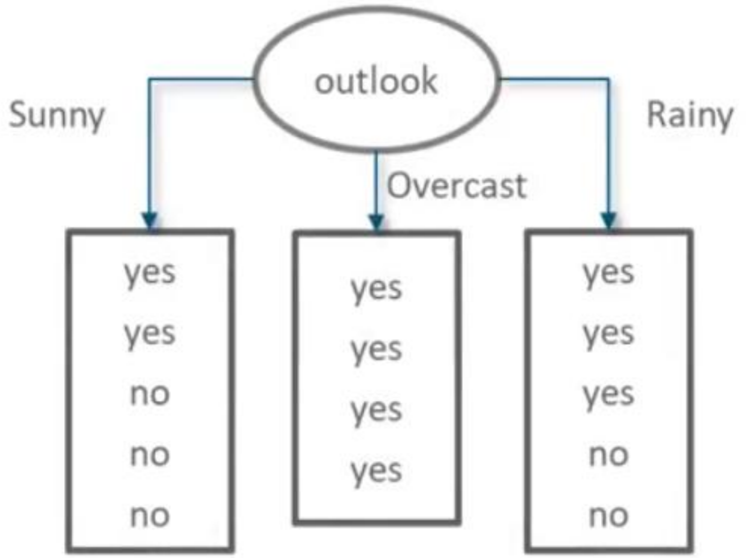


Information Gain of attribute "outlook"

From the total of 14 instances we have:

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

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

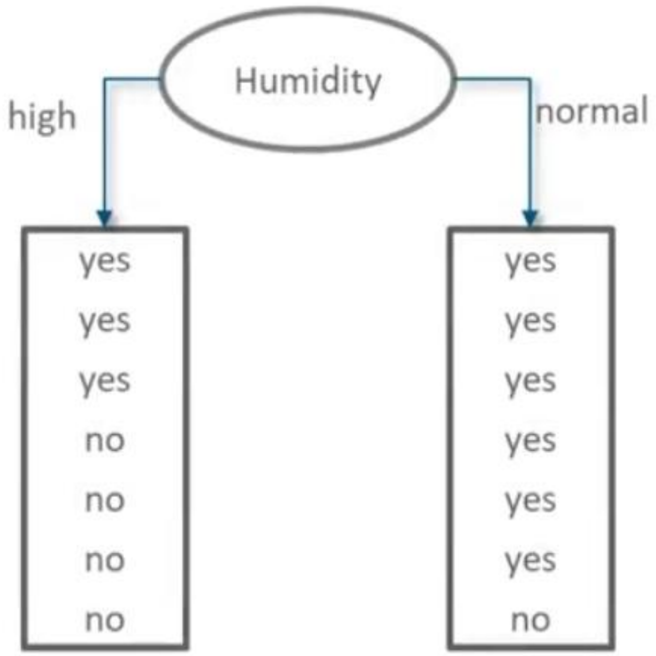


Information Gain of attribute "humidity"

From the total of 14 instances we have:

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

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

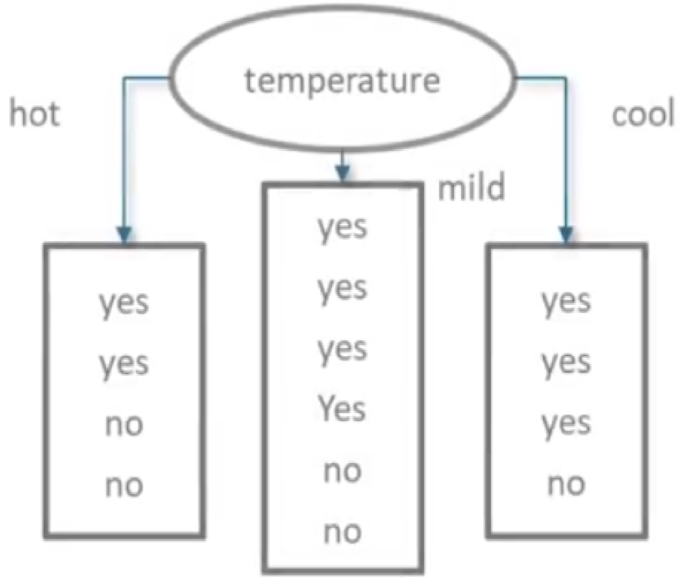


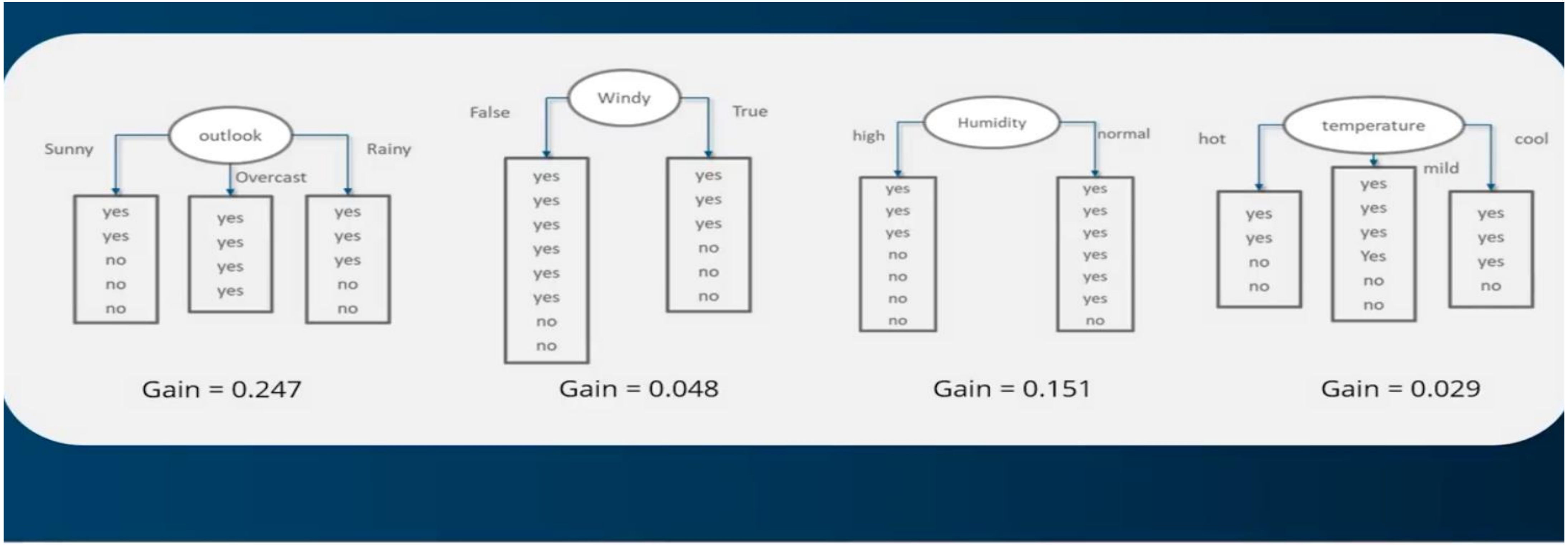
Information Gain of attribute "temperature"

From the total of 14 instances we have:

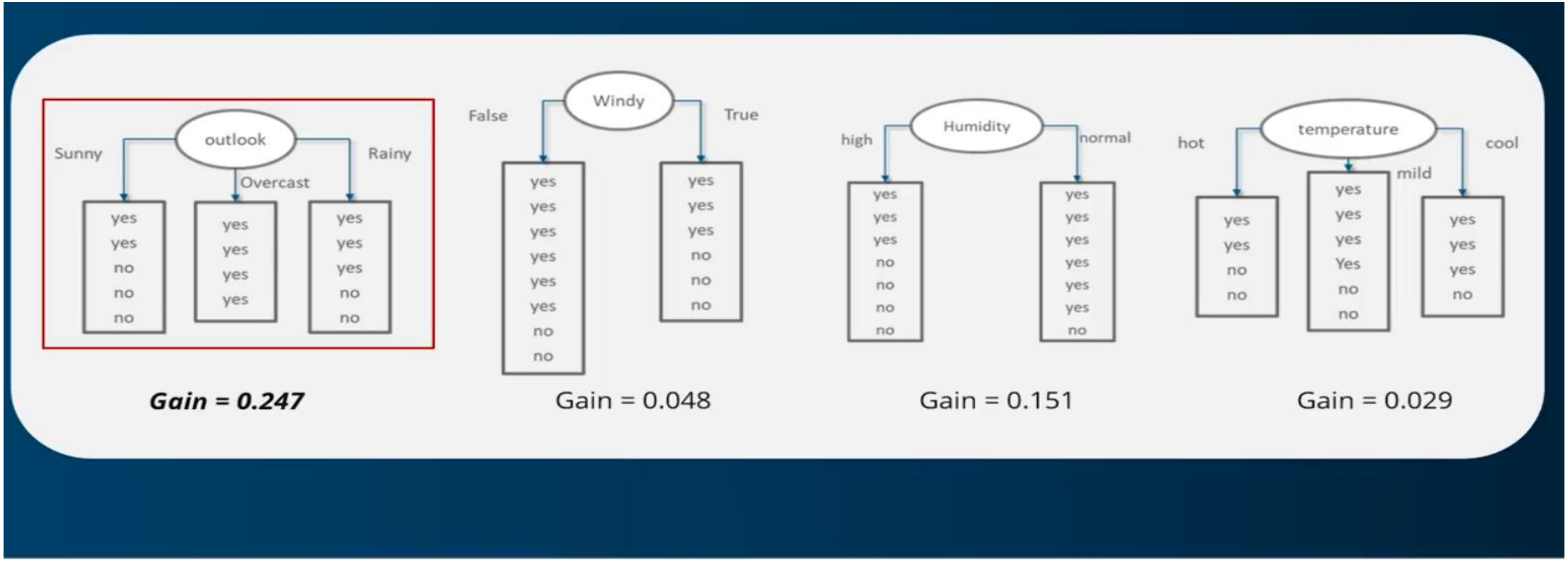
- *4 instances "hot"*
- *6 instances "mild"*
- *4 instances "cool"*

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





The variable with the highest IG is used to split the data at the root node.



The variable with the highest IG is used to split the data at the root node. The 'Outlook' variable has the highest IG, therefore it can be assigned to the root node.

Confusion Matrix

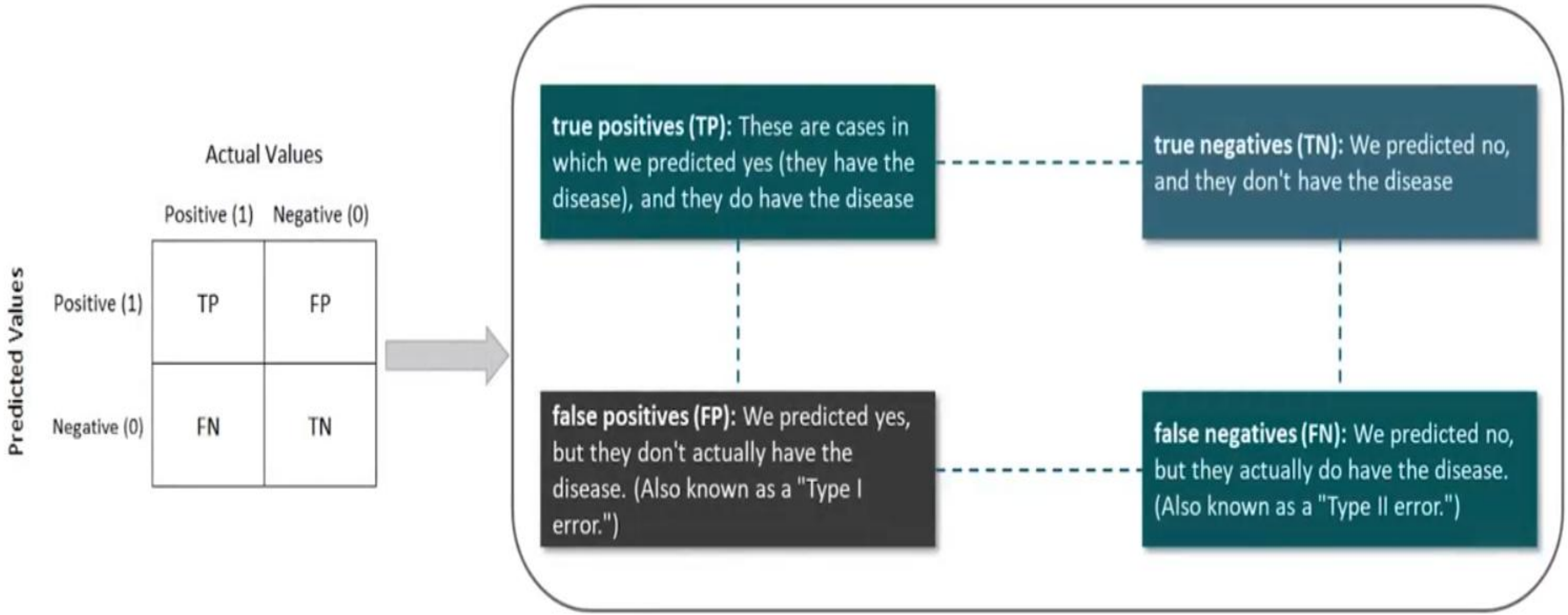
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

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

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

- 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


Actual Values
Positive (1)
Negative (0)
Predicted Values
Positive (1)
Negative (0)
TP
FP
FN
TN
true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease
true negatives (TN): We predicted no, and they don't have the disease
false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")