Confusion Matrix

N = 165	Predicted NO	Predicted YES	
Actual NO	TN = 50	FP = 10	60
Actual YES	FN = 5	TP = 100	105
	55	110	

- 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.")

- Accuracy: How often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- Misclassification Rate: How often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- True Positive Rate: When it's actually yes, how often does it predict yes? (Recall)
  - TP/actual yes = 100/105 = 0.95
  - also known as "Sensitivity" or "Recall"
- False Positive Rate: When it's actually no, how often does it predict yes?
  - FP/actual no = 10/60 = 0.17
- True Negative Rate: When it's actually no, how often does it predict no?
  - TN/actual no = 50/60 = 0.83
  - equivalent to 1 minus False Positive Rate
  - also known as "Specificity"
- Precision: When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91
- Prevalence: How often does the yes condition actually occur in our sample?
  - actual yes/total = 105/165 = 0.64

F1 Score

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

where:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

In "macro" F1 a separate F1 score is calculated for each `species` value and then averaged.

Linear Regression

X (Independent)	Y (Dependent)	X - X_Bar	Y - Y_Bar	(X-X_Bar)^2	(X - X_Bar) * (Y - Y_Bar)	Yp	(Yp - Y_Bar)	(Yp - Y_Bar) ^ 2	(Y - Y_Bar)^2
1	3	-2	-0.6	4	1.2	2.8	-0.8	0.64	0.36
2	4	-1	0.4	1	-0.4	3.2	-0.4	0.16	0.16
3	2	0	-1.6	0	0	3.6	0	0	2.56
4	4	1	0.4	1	0.4	4.0	0.4	0.16	0.16
5	5	2	1.4	4	2.8	4.4	0.8	0.64	1.96
X_Bar 3	**Y_Bar. 3.6**			**Sum = 10**	**Sum = 4.0**			**Sum = 1.6**	**Sum = 5.2**

Equation of line : Y = mX + C

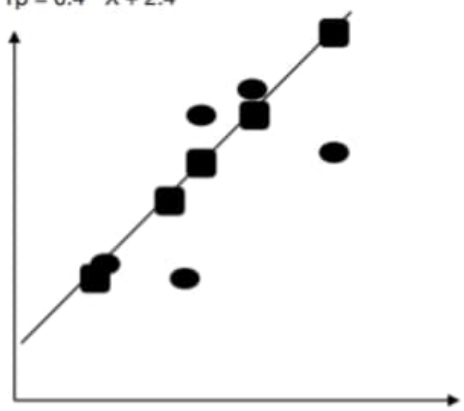
Y Dependent Variable
X Independent Variable
m Slope of line
c Y intercept of line

$$m = \frac{\Sigma (X - X\_Bar)(Y - Y\_Bar)}{\Sigma (X - X\_Bar)^{\wedge} 2}$$

m = 4/10
m = 0.4

Y = mX + C
3.6 = 0.4 * 3 + C
3.6 = 1.2 + C
C = 3.6 - 1.2
C = 2.4

Yp = 0.4 * X + 2.4



R Square method

Distance (predicted - mean)
VS
Distance (actual - mean)

R_Square formula

$$\frac{\Sigma (Yp - Y\_Bar)^{\wedge} 2}{\Sigma (Y - Y\_Bar)^{\wedge} 2}$$

R^2 = 1.6 / 5.2
R^2 = 0.3