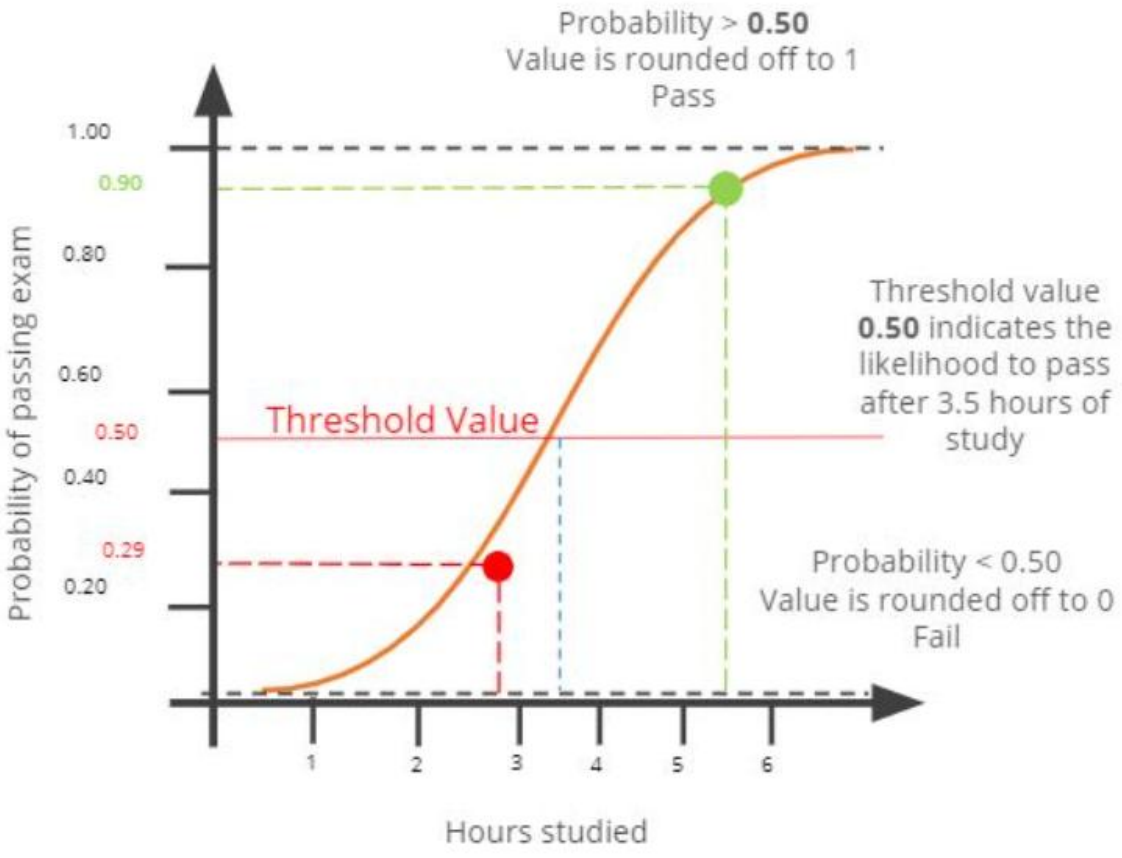
Logistic Regression

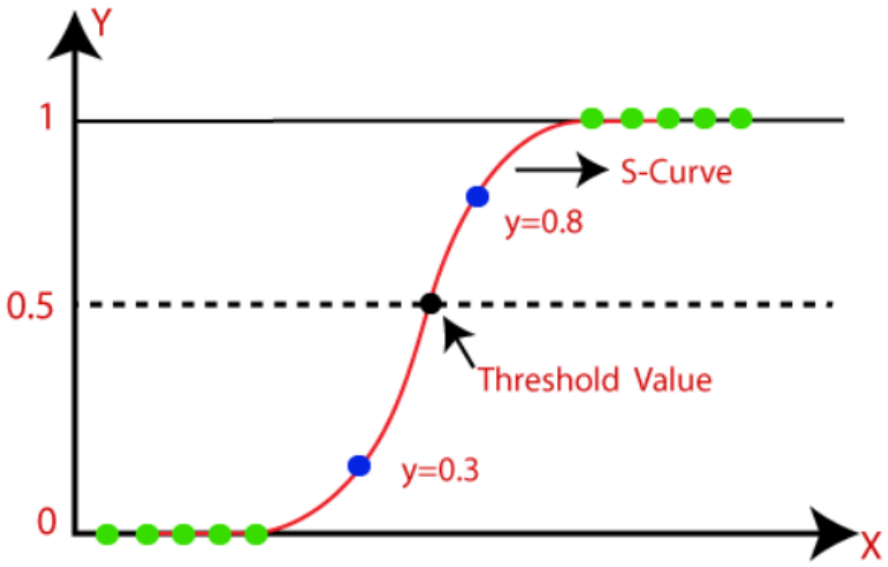
- Its one of the most popular ML algorithms, which comes in the category of Supervised Learning technique.
- It is used for predicting the categorical dependent variable
- It predicts the output of a categorical dependent variable. So the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

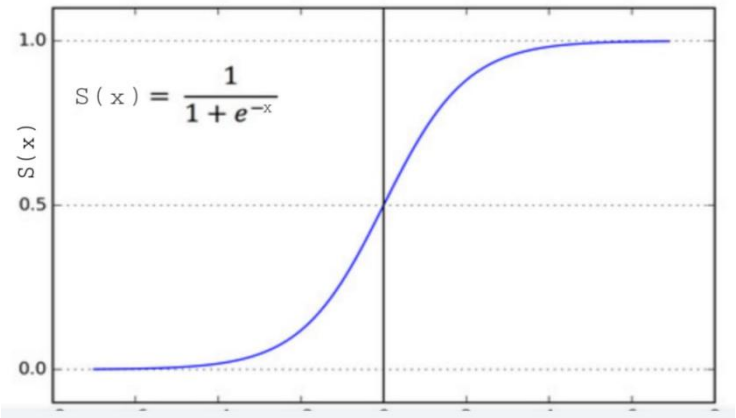


Sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

e (Eulers Constant) =2.71





For Good performance

1. The dependant variable in binary logistic regression **must be binary**.

2. Only the variables that are **relevant** should be included.

Performance Metrics

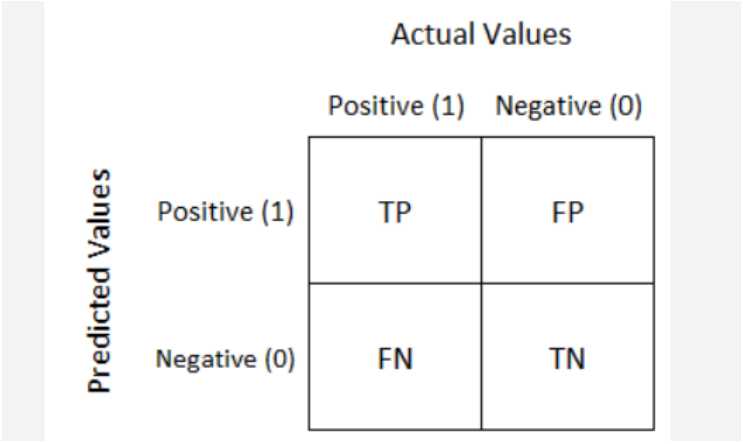
1. Accuracy :

Actuals : 0 0 1 1 0 1 0 0 1 1
Preds : 0 0 0 1 1 1 1 0 0 1

Total correct predictions/Total Predictions = 6/10 = 60%

Confusion Metrics:

Actual -> cancerous ; Pred -> cancerous (TP)
Actual -> Non cancerous ; Pred -> cancerous (FP)
Actual -> cancerous ; Pred -> Non cancerous (FN)
Actual -> Non cancerous ; Pred -> Non cancerous(TN)



$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP}
\end{aligned}
$$

Precision

- We use this metric when we want out **FP to be as low** as possible
- Out of total predicted positives, how many are actually positives.

Recall

- We use this metric when we want **FN to be low**
- Out of total actual positives, how many are predicted positives.

High precision and high recall are desirable,