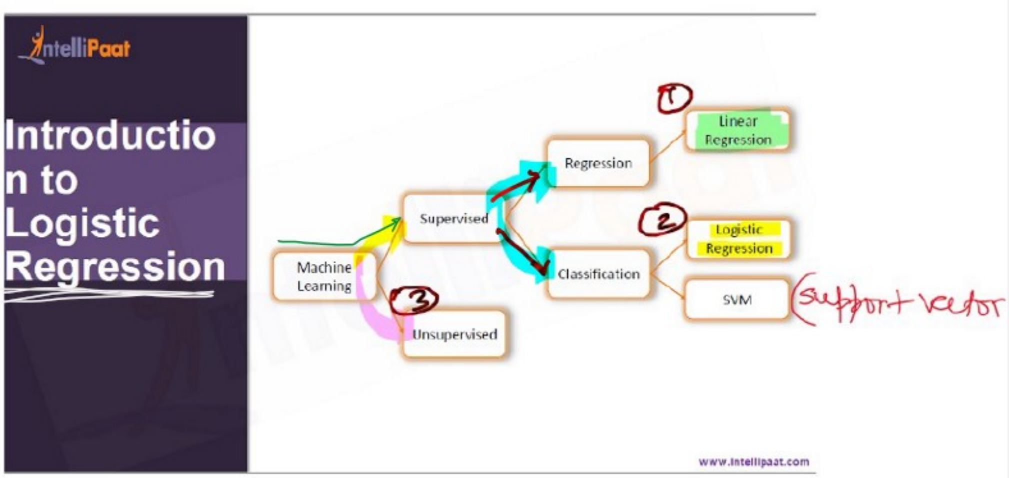
Logistic Regression

02 January 2024 07:02

Logistic vs Logistics

Logistic
- ML algorithm
- supervised learning
- classification

Logistics
- key component of supply chain Management
  - Logistics means transport which can be done in multiple modes:
    - a) Rail
    - b) Road
    - c) Air
    - d) Ocean.



Logistic regression is a statistical model/ machine learning algorithm that uses a logistic function (logit function) to model a binary dependent variable (y/target/response)

binary → two class

multiclass (more than two classes)

$$\underset{\downarrow}{Y} = f\underbrace{(x_1, x_2, x_3 \cdots x_n)}_{\downarrow}$$

Target variable — predictor variables

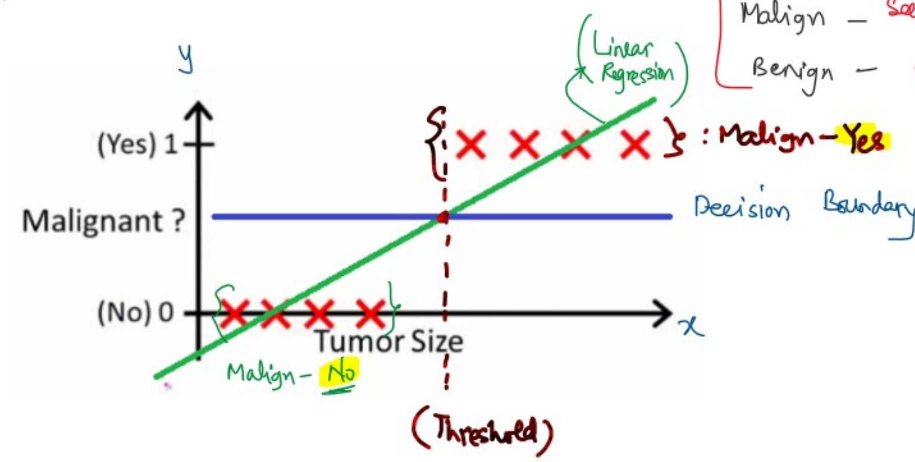
- categorical variable — Classification
- Continuous variable — Regression

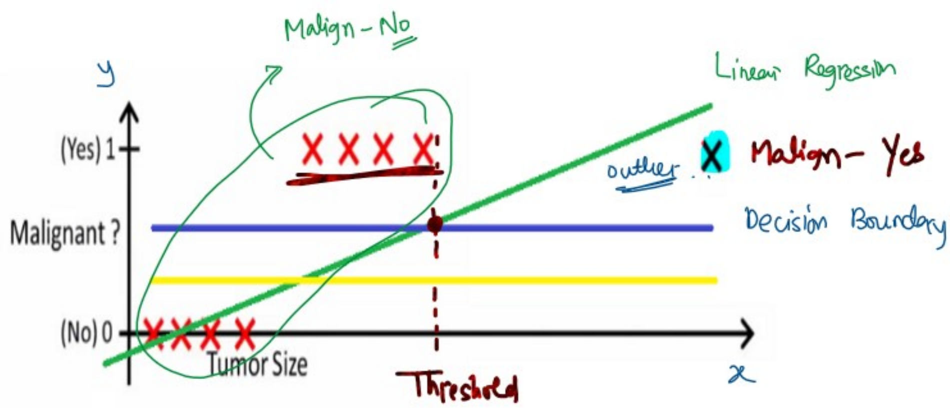
Spam mail Use-case

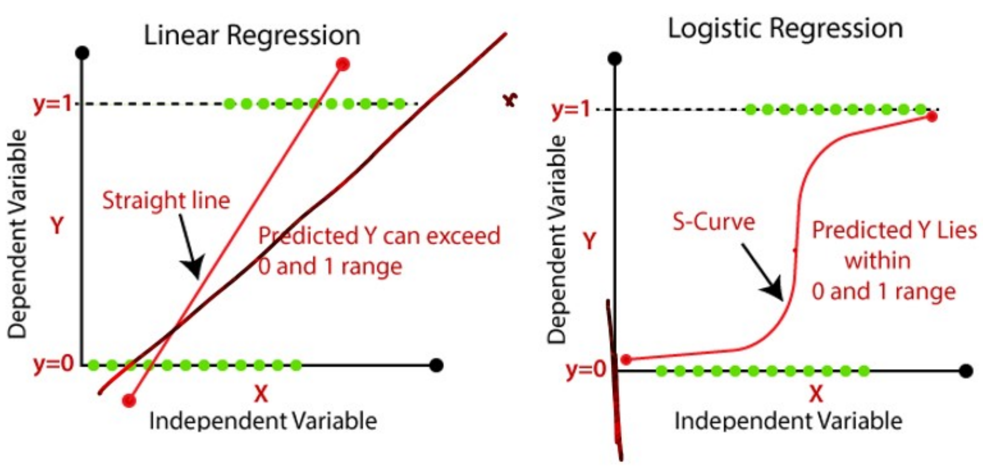
an incoming mail can be classified into either → a spam mail / → not a spam mail } 2 classes or 2 categories

Why do we need a classifier?

Malign – Seems to be a cancer
Benign – not a cancer.







Intuition

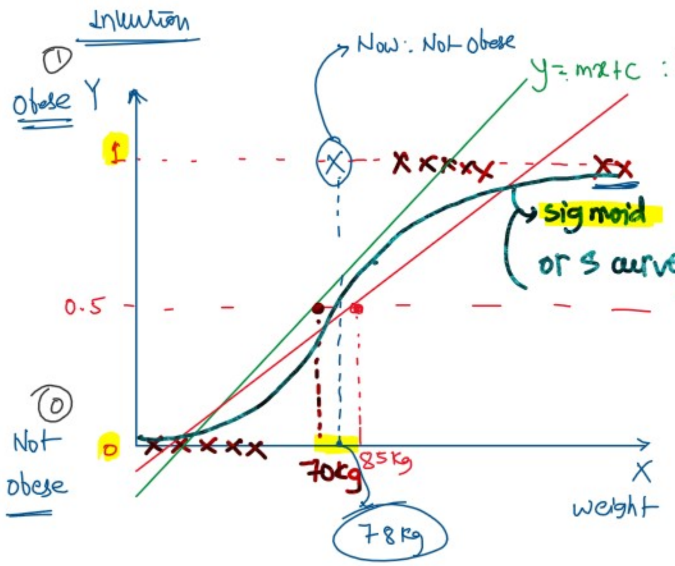
① obese Y ↑

Then: Obese

Now: Not-obese

$y = mx + c$: Linear Regression Line

Intuition



Condition

If $Y \geq 0.5 \Rightarrow 1$ (obese)

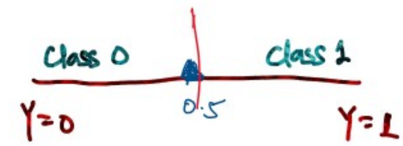
$Y < 0.5 \Rightarrow 0$ (Not obese)

Linear: $Y \rightarrow -\infty$ to ∞

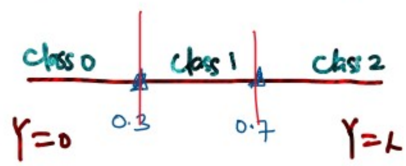
logistic: $Y \rightarrow (0$ to $1) \rightarrow$ Probabilities

$Y < 0.5$ | $Y \geq 0.5$

$Y =$ 0 to 1



$\Rightarrow$ 2 Classes

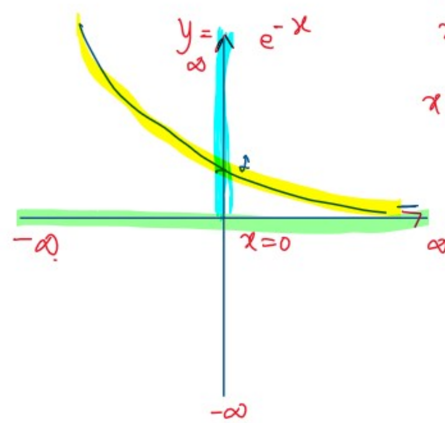


$\Rightarrow$ 3 Classes.

Mathematics for logistic regression

Sigmoid Function (S-curve)

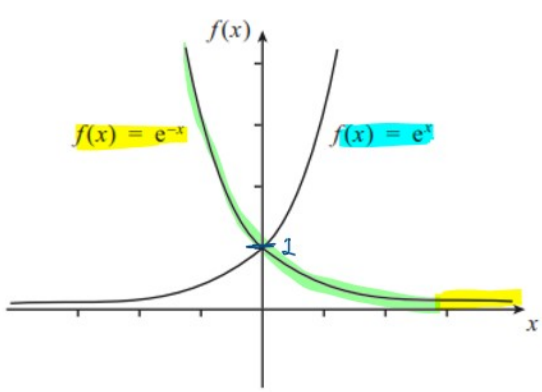
$$f(x) = \left(\frac{1}{1 + e^{-x}}\right) \rightarrow (-\infty < x < \infty)$$



$x = 0 \quad y = e^{-x} = e^{-0} = 1$

$x \rightarrow \infty \quad y = e^{-\infty} = \frac{1}{e^{\infty}} = \frac{1}{\infty} = 0$

$x \rightarrow -\infty \quad y = e^{-(-\infty)} = e^{\infty} = \infty$.



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

Domain: $x: (-\infty, \infty)$

Range: $y: (0, \infty)$

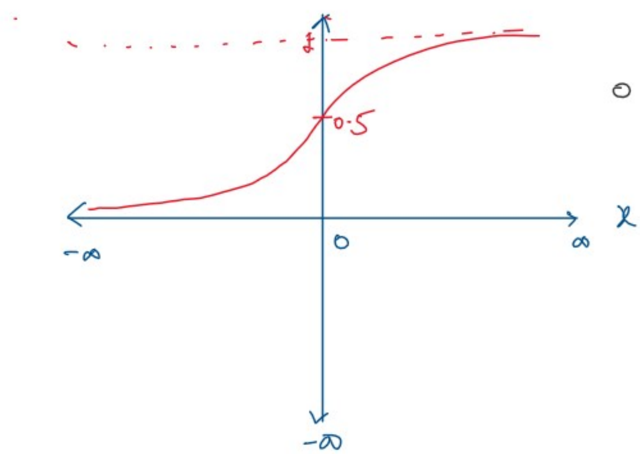
$x \to \infty$: $f(x) = \lim_{x\to\infty}\left(\frac{1}{1+e^{-x}}\right) = \left(\frac{1}{1+e^{-\infty}}\right) = \left(\frac{1}{1+\frac{1}{e^{\infty}}}\right) = \left(\frac{1}{1+\cancel{\frac{1}{\infty}}\,0}\right) = \frac{1}{1} = 1$

$\boxed{x = -\infty \text{ to } \infty}$

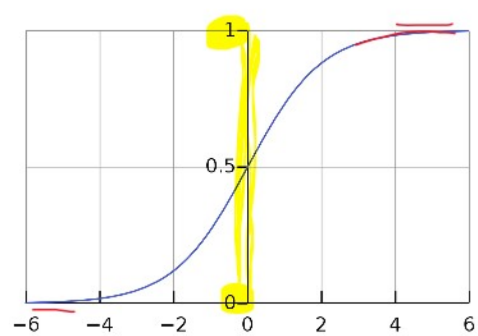
$x = 0$ $f(x) = \lim_{x\to 0}\left(\frac{1}{1+e^{-x}}\right) = \left(\frac{1}{1+e^{0}}\right) = \frac{1}{1+1} = \frac{1}{2} = 0.5$

$\boxed{y = 0 \text{ to } 1}$

$x \to -\infty$ $f(x) = \lim_{x\to-\infty}\left(\frac{1}{1+e^{-(-\infty)}}\right) = \frac{1}{1+e^{\infty}} = \frac{1}{1+\infty} = 0$



$0 < y < 1$



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

$$\boxed{P = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}$$

$$\frac{p}{1-p}$$

$$1 - p = \left(1 - \frac{1}{1+e^{-x}}\right)$$

$(x = \beta_0 + \beta_1 x)$

$$(1-p) = \left(\frac{\cancel{1}+e^{-x}-\cancel{1}}{\quad}\right) = \left(\frac{e^{-x}}{\quad}\right)$$

$$\boxed{P = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}$$

P → (0 to 1)

$\beta_0+\beta_1 x$ → (linear regression) $-\infty < x < \infty$

Logit Function
or
Logistic Function

$$(1-p) = \left(\frac{\cancel{1}+e^{-x}\cancel{-1}}{1+e^{-x}}\right) = \left(\frac{e^{-x}}{1+e^{-x}}\right)$$

$$\left(\frac{p}{1-p}\right) = \left(\frac{\frac{1}{\cancel{1+e^{-x}}}}{\frac{e^{-x}}{\cancel{1+e^{-x}}}}\right) = \frac{1}{e^{-x}} = e^{x}$$

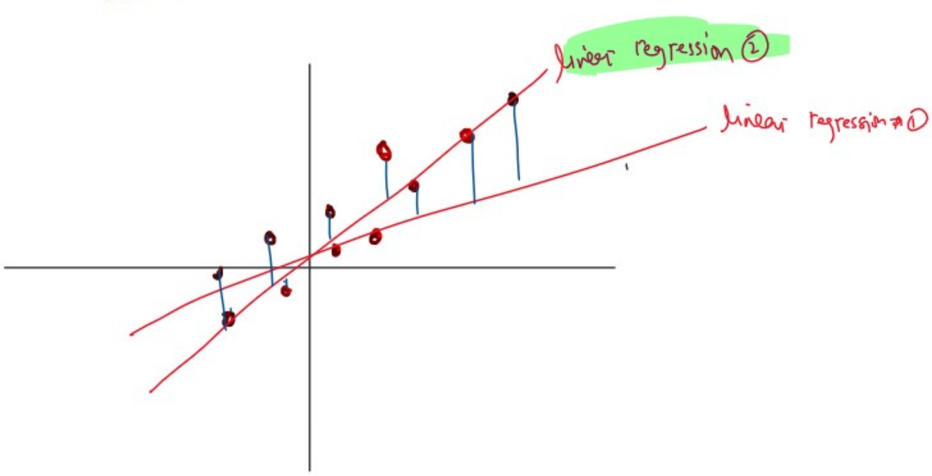
$$Y = \log_e\left(\frac{p}{1-p}\right) = \log_e e^{x}$$

$$\boxed{Y = x.}$$

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x}$$ #

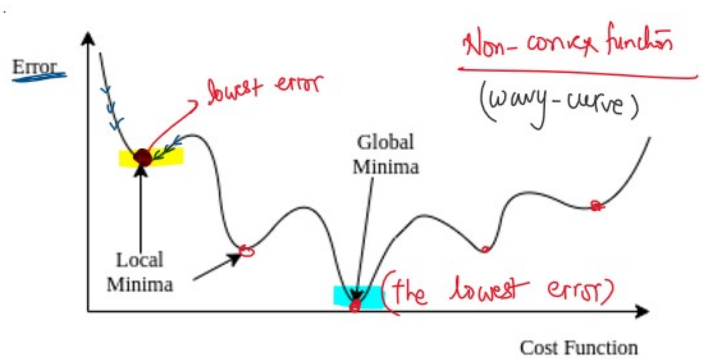
Cost Function

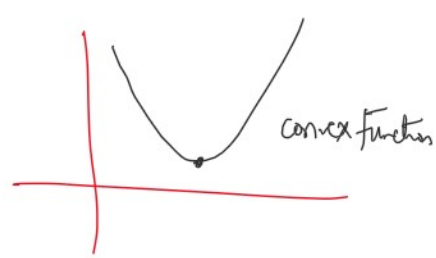
Cost function summarizes how well the model is behaving. In other words, we can say that the cost function helps to measure how close the model's predictions are to the actual outputs.



MSE: Mean Squared Error

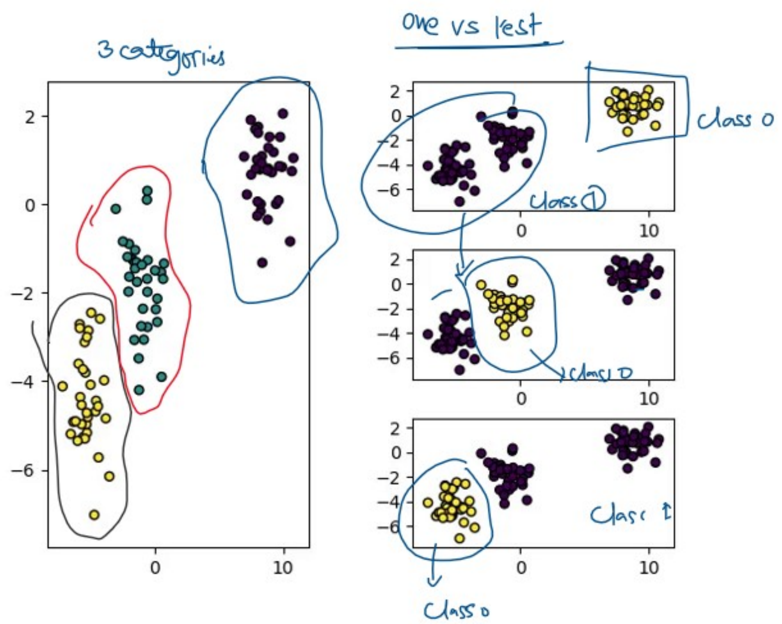
MSE in logistic





Maximum Likelihood Estimator as the cost function for logistic regression

Likelihood → Probability of occurrence — (0 to 1) logistic regression → Probability of occurrence: (0 to 1)



Assumptions of logistic Regression

Assumption #1 Response / target (Y) variable is binary (or multi-class)

$\underbrace{\text{binary (or multi-class)}}_{\text{categorical}}$

Assumption #2 Predictor variables are independent

$x_2 = x_1^2$

y	x_1	x_2	x_3
	1	1	
	2	4	
	3	9	

- observations should not come from repeated measurements of the same variable.

Assumption #3 No severe multicollinearity among the input variables (x)

± 0.7 or more

Assumption #4 No extreme outliers

Assumption #5 Sample size is sufficiently large.

Assumption #6 Linear relationship between input variables $(\beta_0 + \beta_1 x)$ and the logit of the response variable. $\log\left(\frac{p}{1-p}\right)$

$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ where 'p' is the probability of a positive outcome

(odds)

Evaluating the Logistic Regression Model

Confusion Matrix

Confusion matrix is a performance measurement for machine learning classification problem where output can be two or more classes.

Binary Class

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP: TRUE POSITIVE	FP: FALSE POSITIVE
Negative (0)	FN: FALSE NEGATIVE	TN: TRUE NEGATIVE

FP → Type 1 Error

FN → Type 2 Error

Accuracy; Precision, Recall, F1-score ??

Positive class: Pregnant
Negative class: Not pregnant.

True Positive:

Interpretation: You predicted positive and it's true.

Ex. You predicted that a woman is pregnant and she actually is. → TRUE

True Negative:

Interpretation: You predicted negative and it's true.

You predicted that a man is not pregnant and he actually is not.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

You predicted that a man is pregnant but he actually is not.

False Negative: (Type 2 Error)

Interpretation: You predicted negative and it's false.

Ques :

cancer → +ve
No cancer → -ve

		Actual (True) Values	
		Cancer	No Cancer
Positive → Predicted Values	Cancer	45	18
Negative	No Cancer	12	25

Confusion matrix for the cancer example. Image by Author.

True Positive : 45
True Negative : 25
False Positive : 18
False Negative : 12

wrong predictions

Accuracy

Base metric used for the model evaluation is often Accuracy, describing the number of correct often Accuracy, describing the number of correct predictions over all predictions.

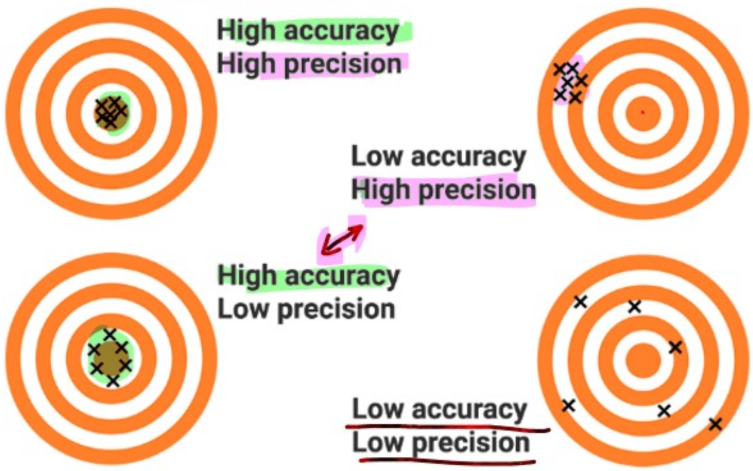
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{45+25}{45+18+25+12}$$

$$= \frac{70}{100} \times 100 = \underline{\underline{70\%}}.$$

Precision

Accuracy vs Precision.

Ideal situation



(X)

Accuracy

Accurately hitting the target (bull's eye) implies you are close to the center of the target.

Precision

Precisely hitting a target means all the hits are closely spaced even if they are very far from the center of the target.

Precision is a measure of how many of the positive predictions made are correct. (True Positives)

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{\text{No. of correctly predicted positive classes}}{\text{No. of total positive classes}}$$

$$= \frac{45}{45+18} = \frac{45}{63} \times 100 = 71.4\%$$

Recall / Sensitivity

Recall / Sensitivity

Recall/Sensitivity is a measure of how many of the positive cases, the classifier correctly predicted, over all the positive cases in the data.

Sen ~ivity = T / I = $\frac{\text{No. of correctly predicted positive instances}}{\text{No. of total positive instances in dataset}}$

$= \frac{45}{45+12} = \frac{45}{57} \times 100 = 78.9\%$

Specificity

Specificity is a measure of how many negative predictions made are correct (True negatives)

I ~ty = "
T I F

$= \frac{\text{No. of correctly predicted negative instances}}{\text{No. of total negative instances in the dataset}}$

$= \frac{25}{25+18} \times 100$

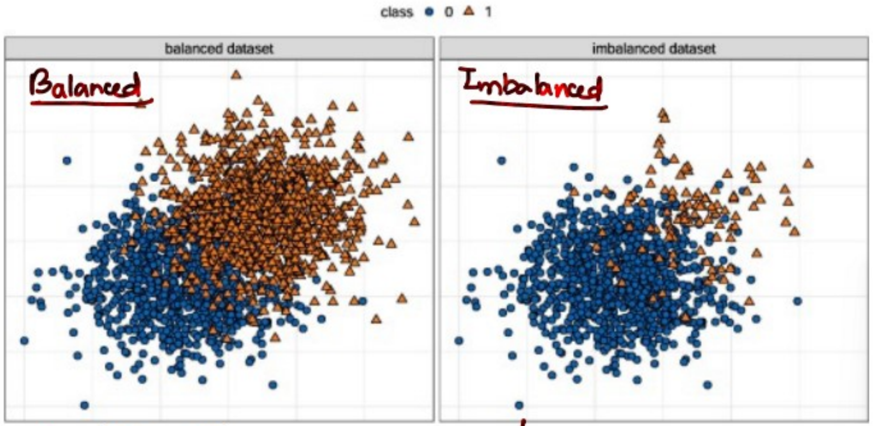
$= \frac{25}{43} \times 100 = 58.1\%$

F1 - score:

It is a measure combining both precision and recall. It is generally described as the harmonic mean of precision and recall.

F1 = * P ci on - Recall
P ci on Recall

Balanced vs Imbalanced datasets



↓ 50% ⟷ 50%
60% ⟷ 40%

Accuracy

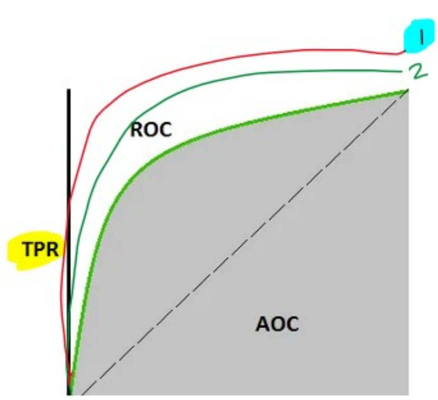
↓ **(F1-score)**

Note #. Accuracy is a good metric when the classes are balanced. However it may not be a suitable metric when there is a significant class imbalances because a model could achieve high accuracy by simply predicting the majority class.

Note: Precision, recall and F1-score provide better insights into the model's performance on positive instances which are often the minority class in the imbalanced datasets

AUC- ROC curve

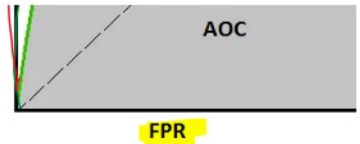
(Area under curve - Receiver operating characteristic) curve



→ Higher accuracy.

True Positive Rate

High AUC ⇒ Model is much better with accuracy → 1 → 1
0 → 0



FPR: False Positive Rate

$$TPR = \text{Recall}/\text{sensitivity} = \left(\frac{TP}{TP+FN}\right)$$

$$FPR = 1 - \text{Specificity} = 1 - \left(\frac{TN}{TN+FP}\right)$$

$$= \frac{\cancel{TN}+FP-\cancel{TN}}{TN+FP} = \left(\frac{FP}{TN+FP}\right)$$