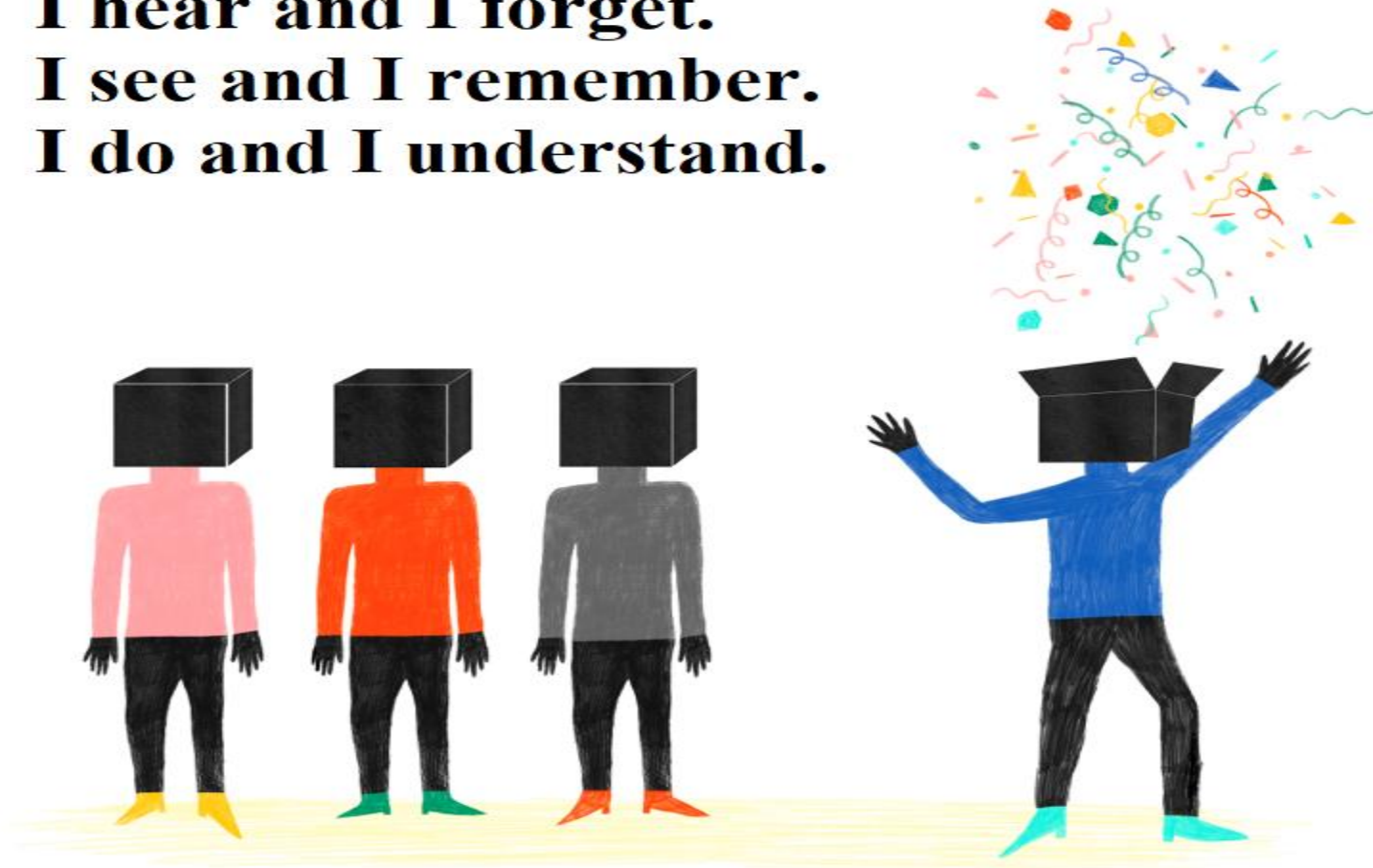


**I hear and I forget.
I see and I remember.
I do and I understand.**



Chandan Verma

Corporate Trainer(Machine Learning,AI,Cloud Computing,IOT)

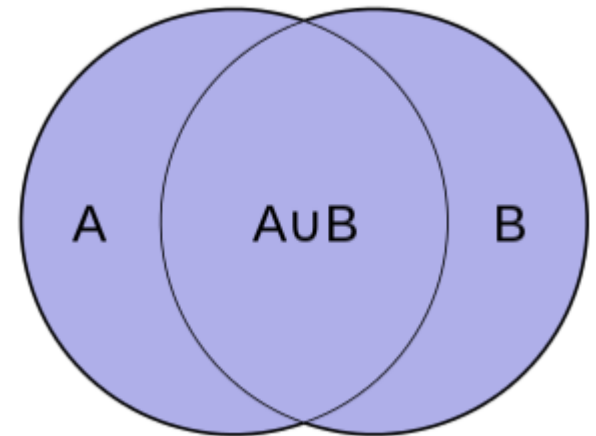
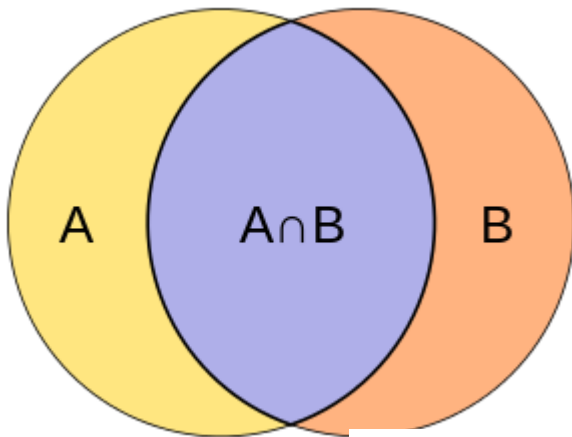
www.facebook.com/verma.chandan.070

www.chandanverma.com



Jaccard similarity

Jaccard similarity or intersection over union is defined as size of intersection divided by size of union of two sets



$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Confusion Matrix

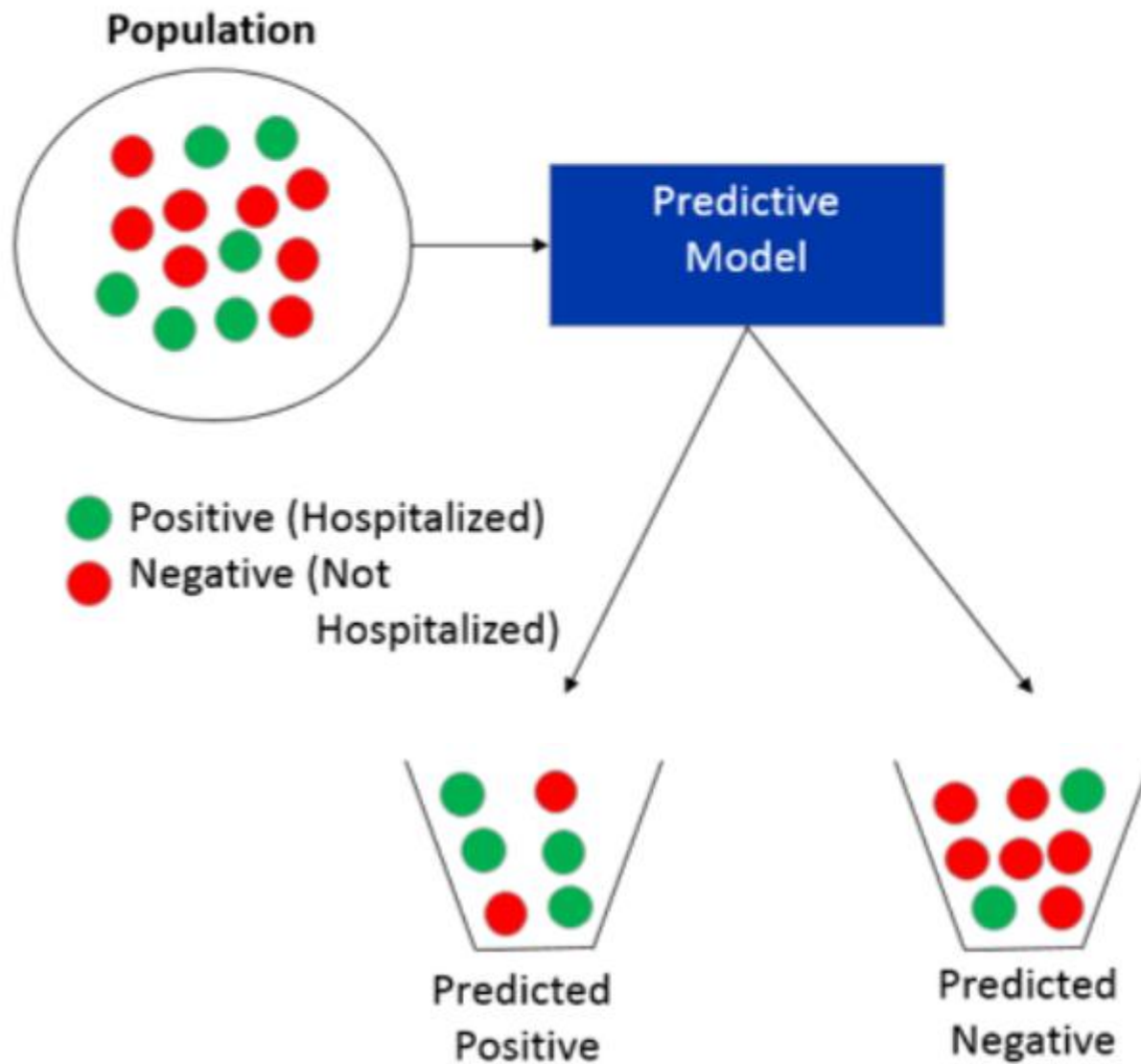
The confusion matrix is simply a square matrix that reports the counts of the **true positive**, **true negative**, **false positive** and false **negative**.

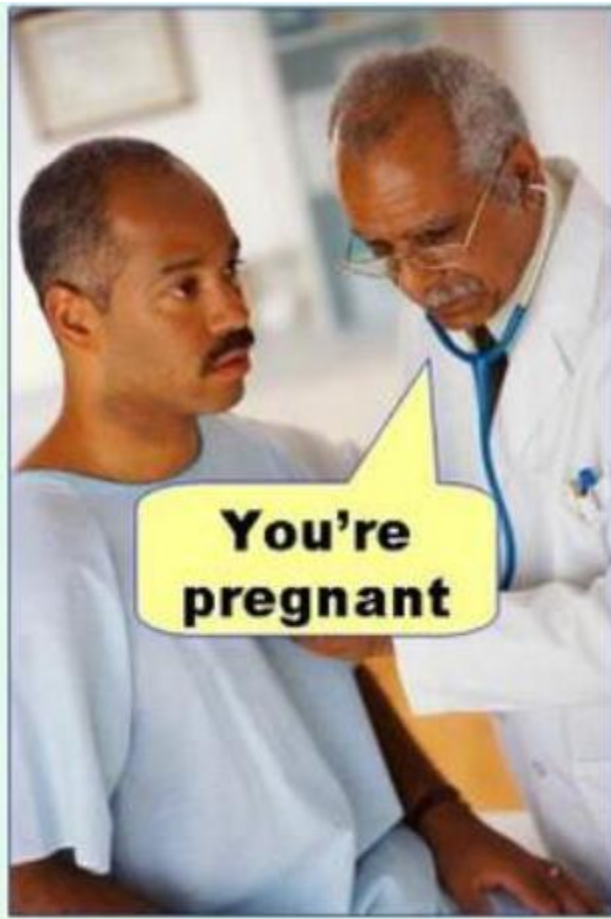
Confusion matrix is a table used to investigate the performance of a classification model where the actual test values are known.

It has two rows and two columns describing the true positives, false positives, false negatives and true negatives.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

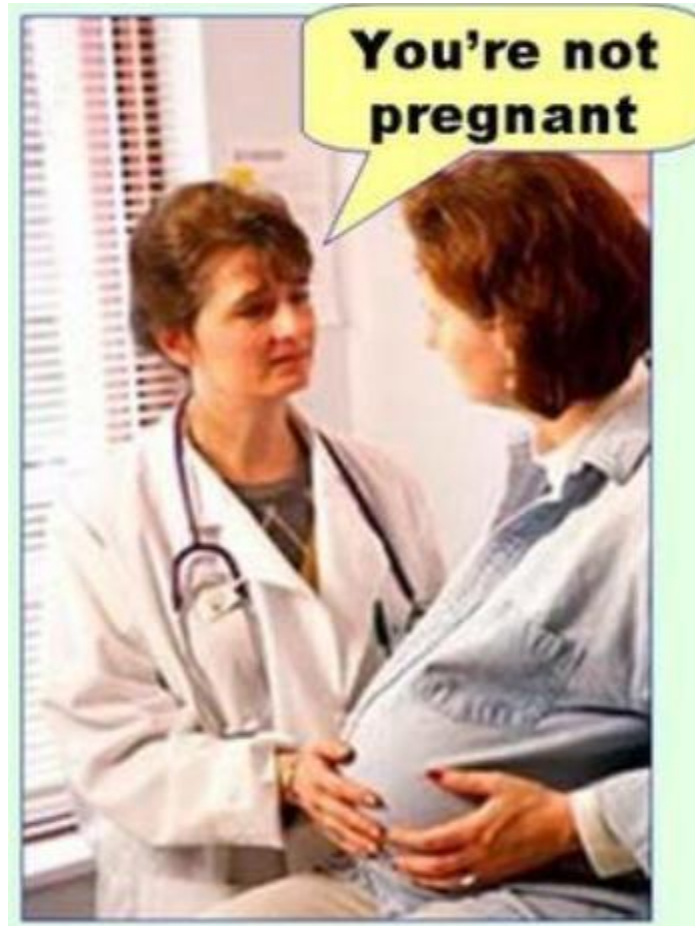
The **Confusion Matrix** is used to explain **Type I** and **Type II** errors from your results. These results are also referred to as false positives and false negatives





False Positives(FP)

A **false positive** is when something is predicted to occur but does not occur. A **false negative** is when something is predicted to not occur, but it does occur



False Negatives(FN)

True Positives(TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

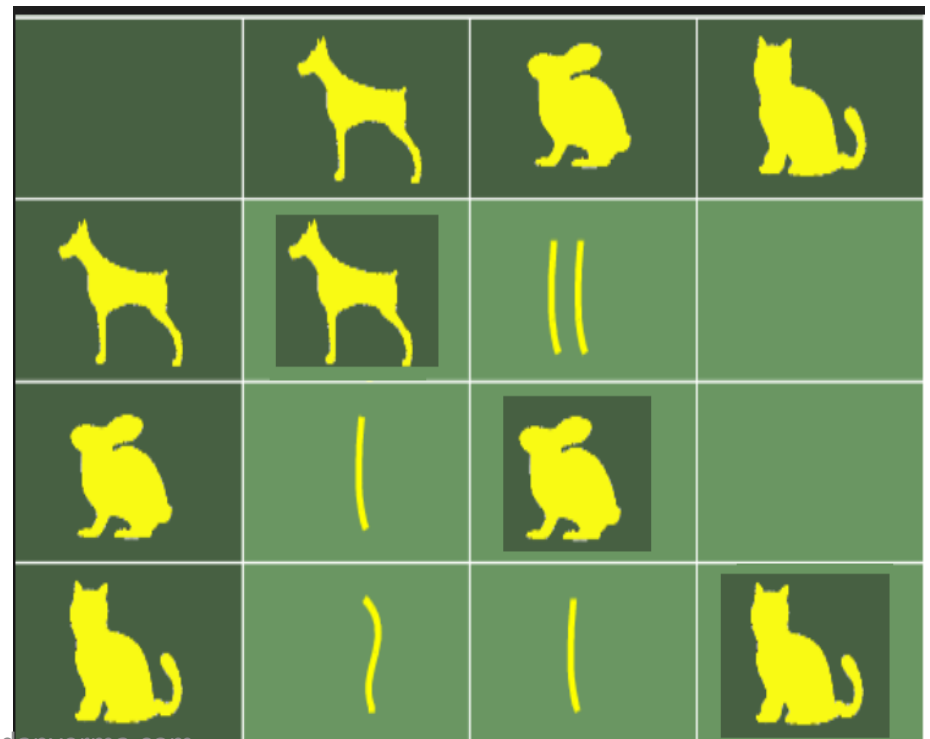
False Positives(FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

False Negatives(FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

True Negatives(TN): We predicted no, and they don't have the disease..



The confusion matrix will be very helpful in this situation to assess the performance of our model.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



confusion matrix

A confusion matrix can give a quick overview of how the prediction model has performed. It is used to see accuracy in Logistic Regression and K-Nearest Neighbor classification

		\hat{y} (Predicted DV)	
		0	1
y (Actual DV)	0	35	5 
	1	10 	50

False Positive (Type I Error)

False Negative (Type II Error)

1. Accuracy Rate = Correct / Total
 $AR = 85/100 = 85\%$

2. Error Rate = Wrong / Total
 $ER = 15/100 = 15\%$

Optimizing a classification model

Both the prediction error and accuracy provide general information about how many samples are misclassified.

Accuracy(ACC): Percentage for correct predictions

$$\text{Accuracy} = \frac{\text{all correct}}{\text{all}}$$

$$\frac{TP + TN}{TP + FN + FP + TN}$$

Prediction error(ERR) or Misclassification Rate: Percentage for incorrect predictions.

$$\text{Misclassification Rate} = \frac{\text{all incorrect}}{\text{all}}$$

$$\frac{FN + FP}{TP + FN + FP + TN} = 1 - \text{Acc}$$

Precision-When it predicts yes, how often is it correct?

		prediction outcome		total
		p	n	
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

$$\text{Precision} = \frac{TP}{TP + FP}$$

Out of all the classes, how much we predicted correctly. It should be high as possible.

Recall

Sensitivity(Recall): Percentage of correct predictions for the actual positives(True Positive Rate).

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall: ability of a classification model to identify all relevant instances

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F1-Score

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time.

single metric that combines recall and precision using the harmonic mean

F1 Score

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

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

sklearn.metrics.confusion_matrix

```
sklearn.metrics.confusion_matrix(y_true,  
y_pred, labels=None, sample_weight=None)
```

y_true :Ground truth (correct) target values.

y_pred :Estimated targets as returned by a classifier.

labels :List of labels to index the matrix.

Receiver operating characteristic curves

ROC (Receiver Operating Characteristic) Curve tells us about how good the model can distinguish between two things

The ROC plot is a model-wide evaluation measure that is based on two basic evaluation measures – **specificity and sensitivity**. Specificity is a performance measure of the whole negative part of a dataset, whereas sensitivity is a performance measure of the whole positive part

Receiver operating characteristic curves

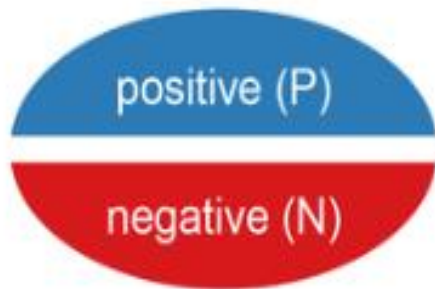
ROC is a plot of signal (True Positive Rate) against noise (False Positive Rate)

The model performance is determined by looking at the area under the ROC curve (or AUC).

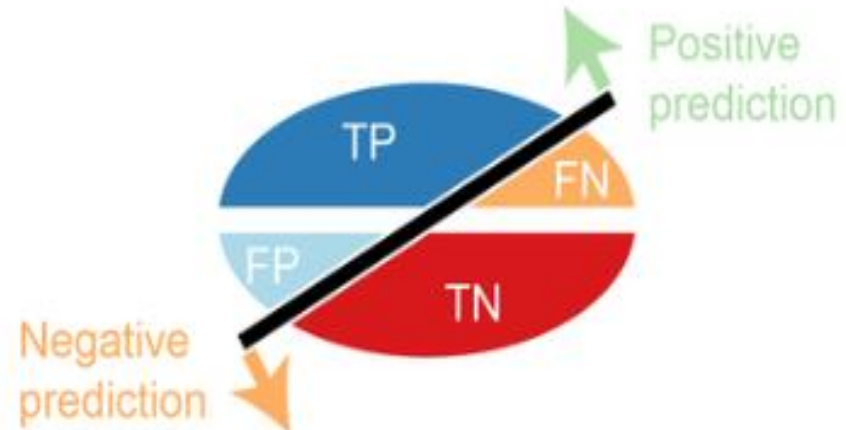
$$TPR = \frac{\text{True Positive}}{\text{Total Positive}}$$

$$FPR = \frac{\text{False Positive}}{\text{Total Negative}}$$

Observed labels

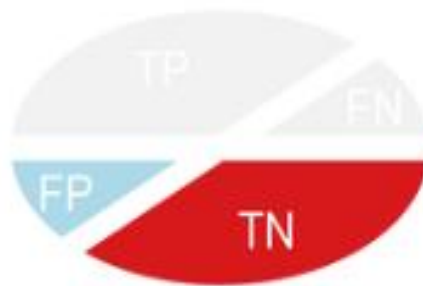


Four outcomes of a classifier



x-axis

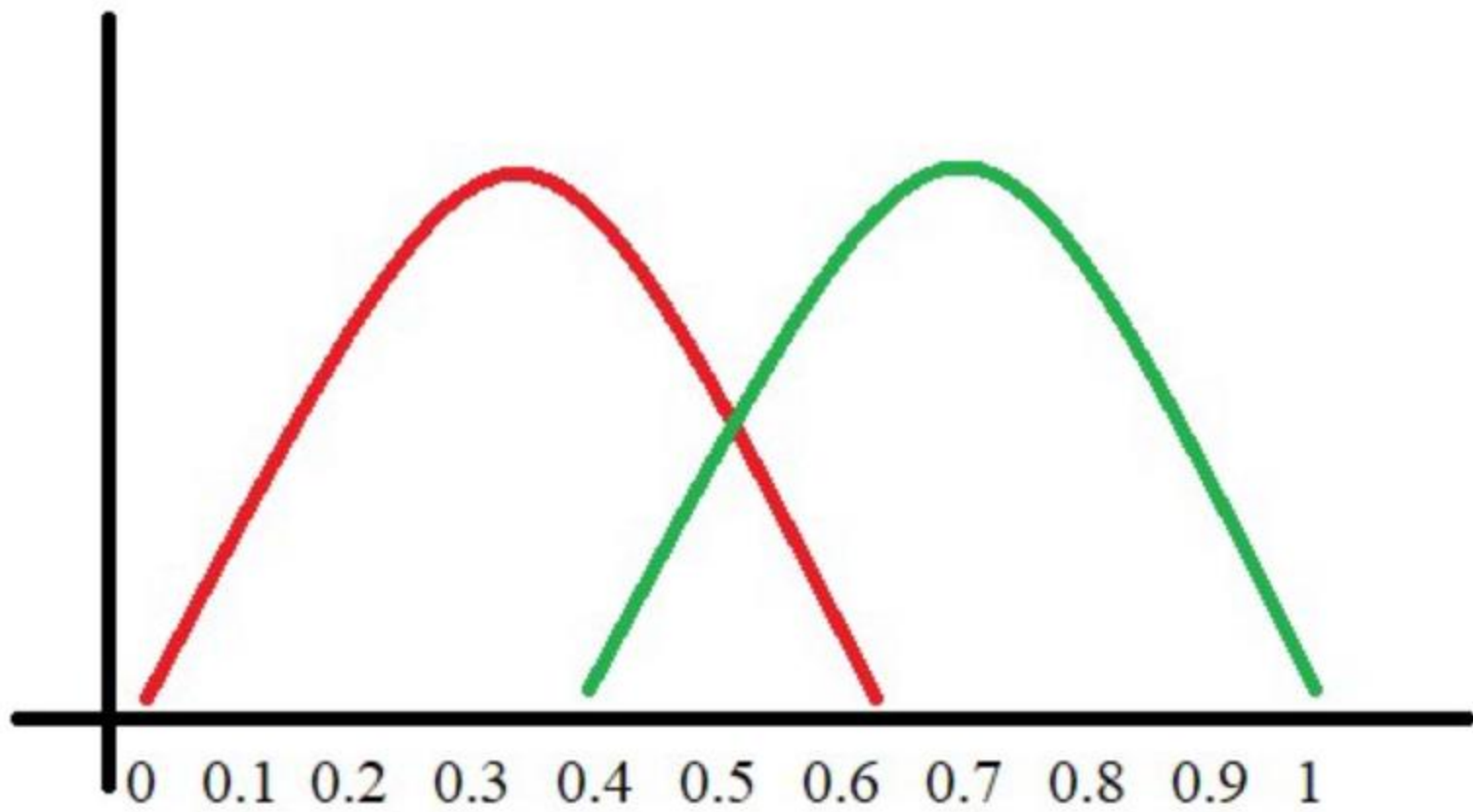
1 - Specificity
False positive rate



y-axis

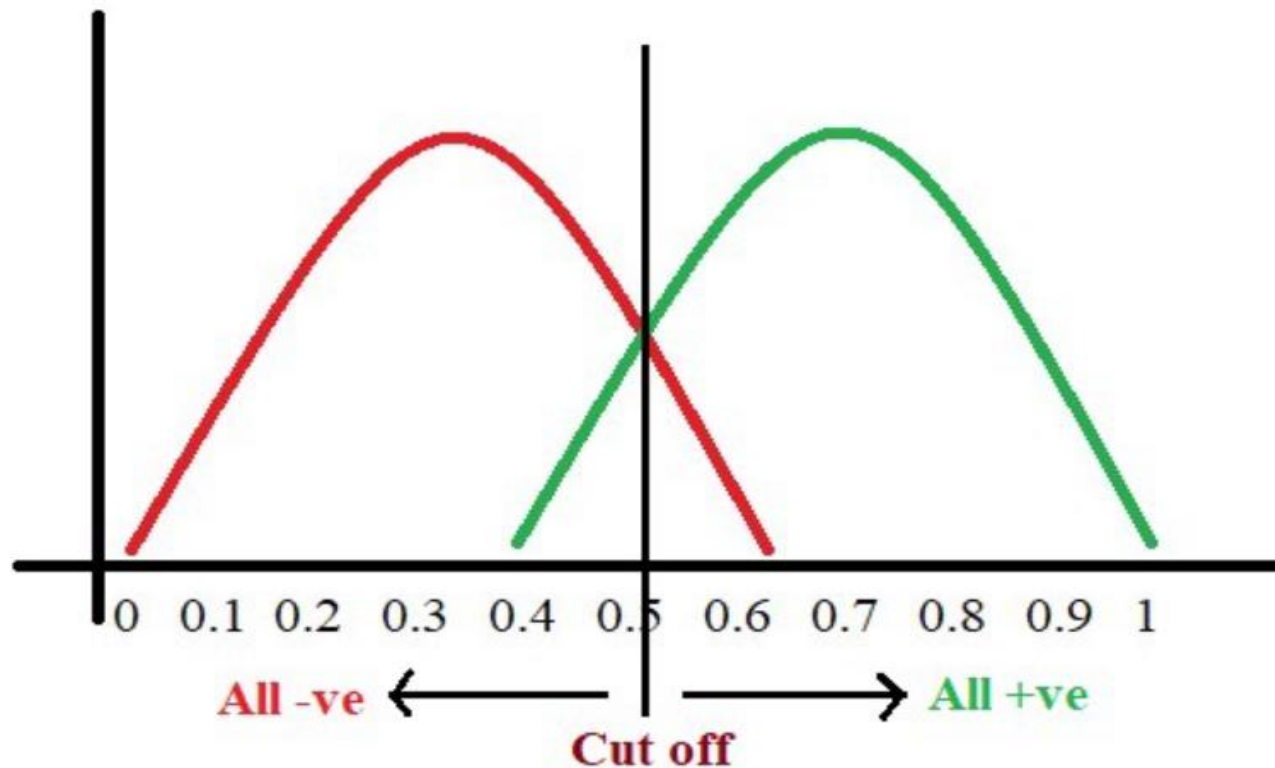
Sensitivity
True positive rate

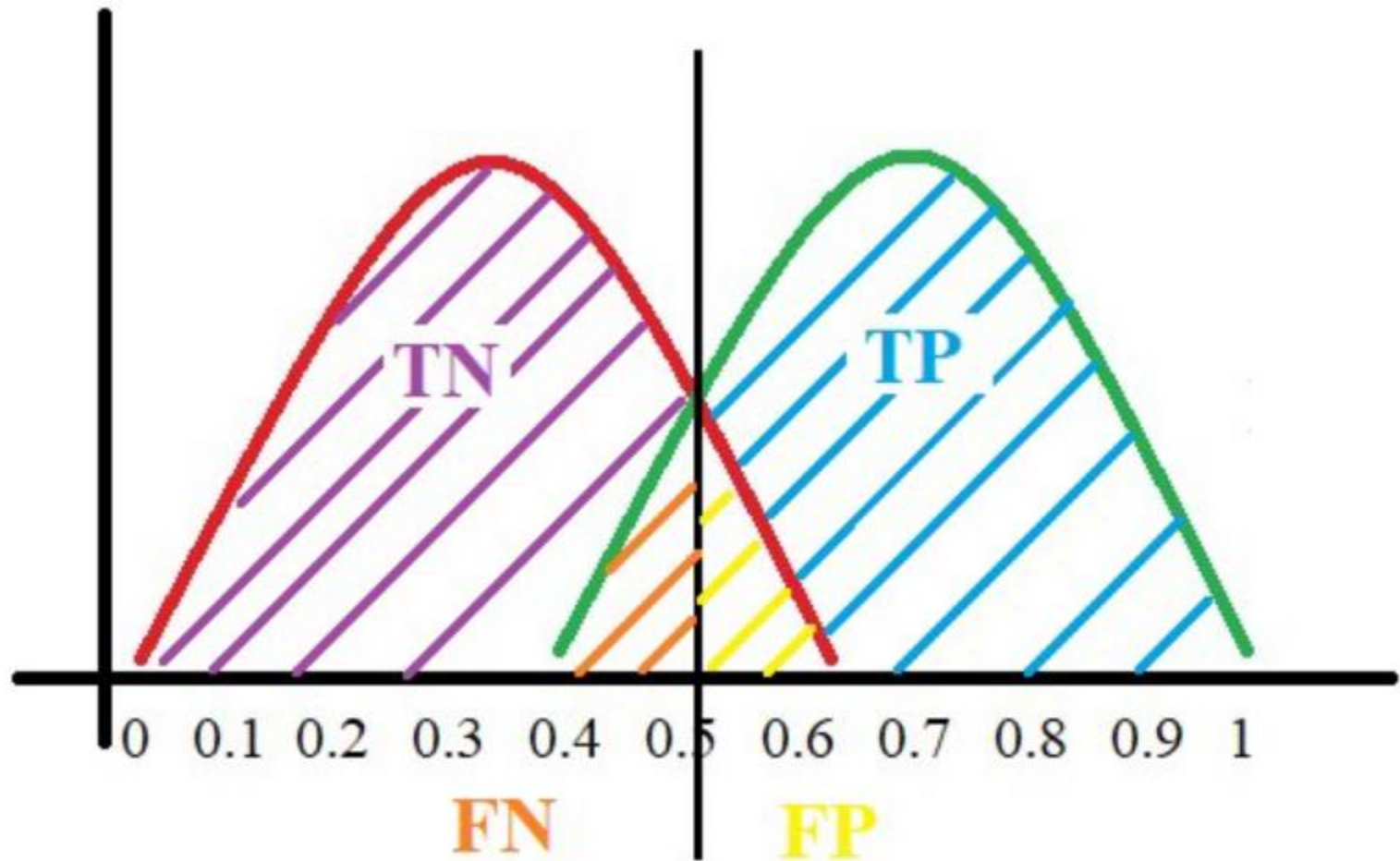




All the positive values above the threshold will be “**True Positives**” and the negative values above the threshold will be “**False Positives**” as they are predicted incorrectly as positives.

All the negative values below the threshold will be “**True Negatives**” and the positive values below the threshold will be “**False Negative**” as they are predicted incorrectly as negatives.





Sensitivity

	positive	negative
positive	<i>TP</i>	<i>FN</i>
negative	<i>FP</i>	<i>TN</i>

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity

	positive	negative
positive	TP	FN
negative	FP	TN

$$\text{Specificity} = \frac{TN}{TN + FP}$$

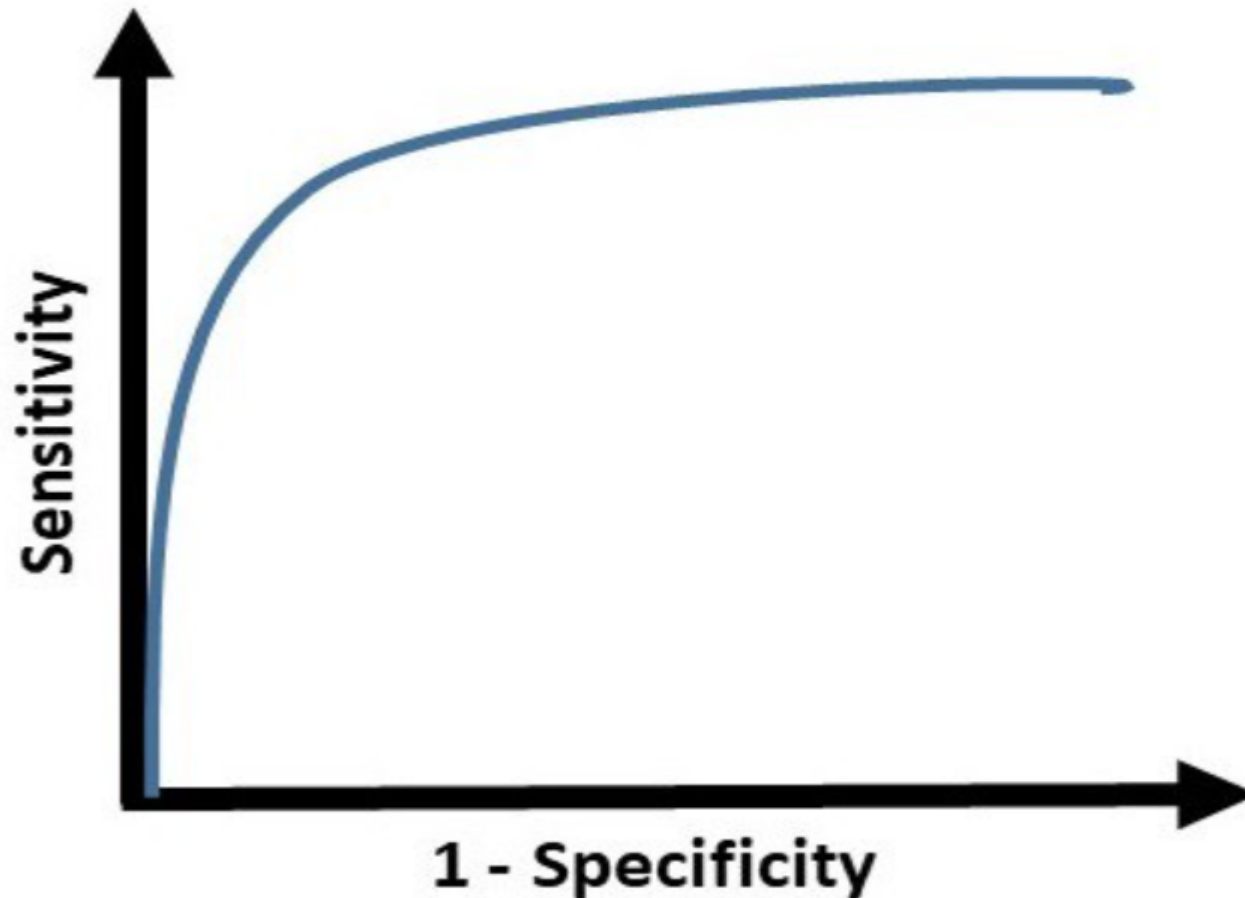
Trade-off between Sensitivity and Specificity

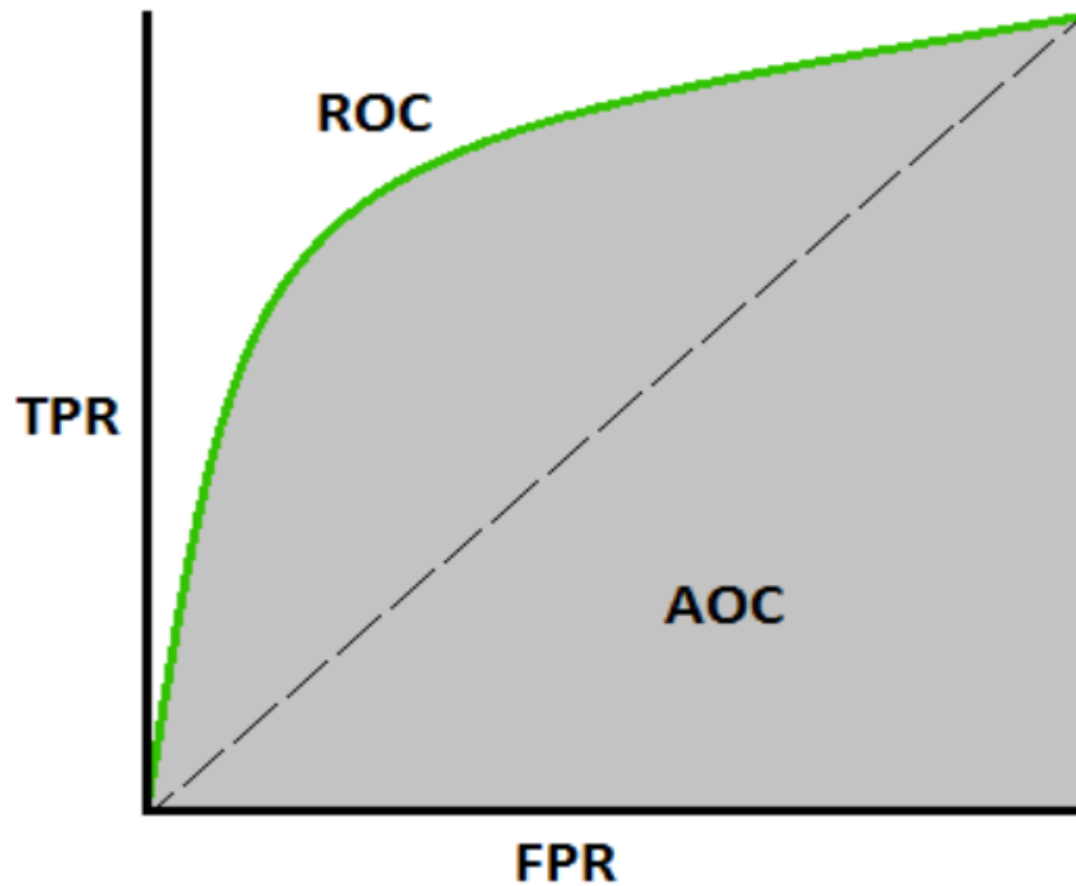
As Sensitivity  Specificity 

As Specificity  Sensitivity 

sensitivity can be called as the “*True Positive Rate*” and $(1 - \text{Specificity})$ can be called as the “*False Positive Rate*”

Receiver operating characteristic curves





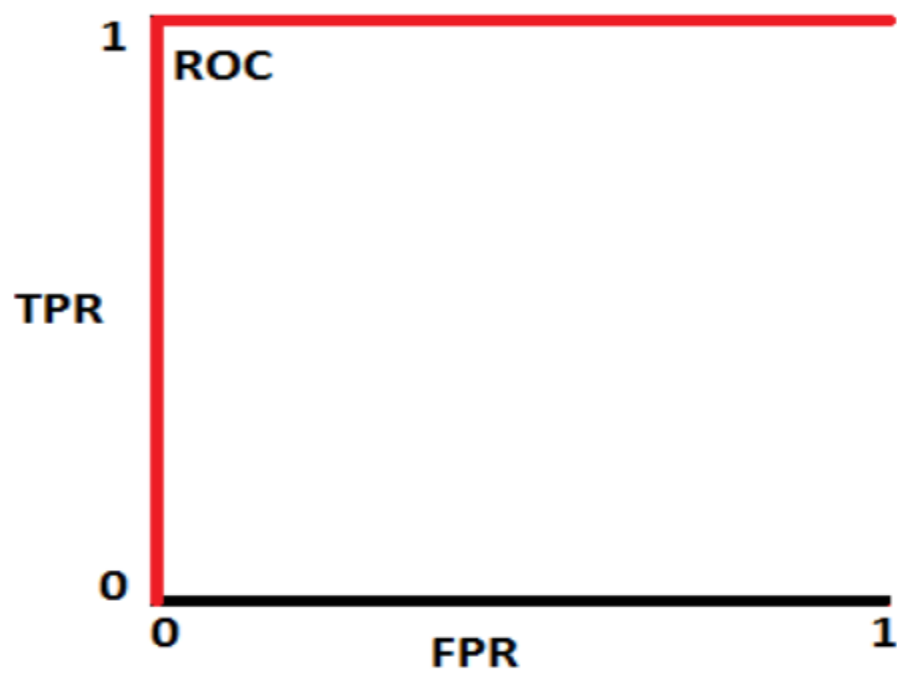
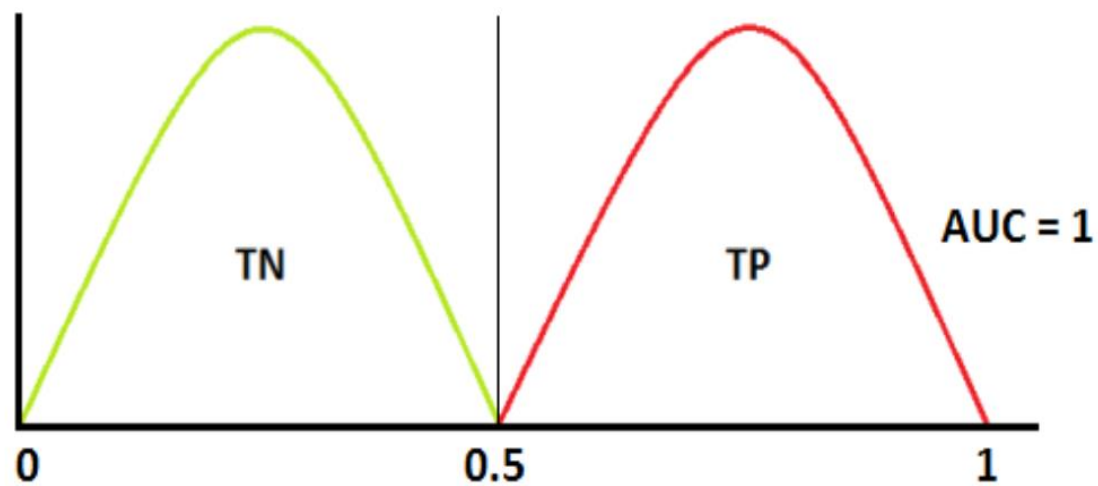
$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

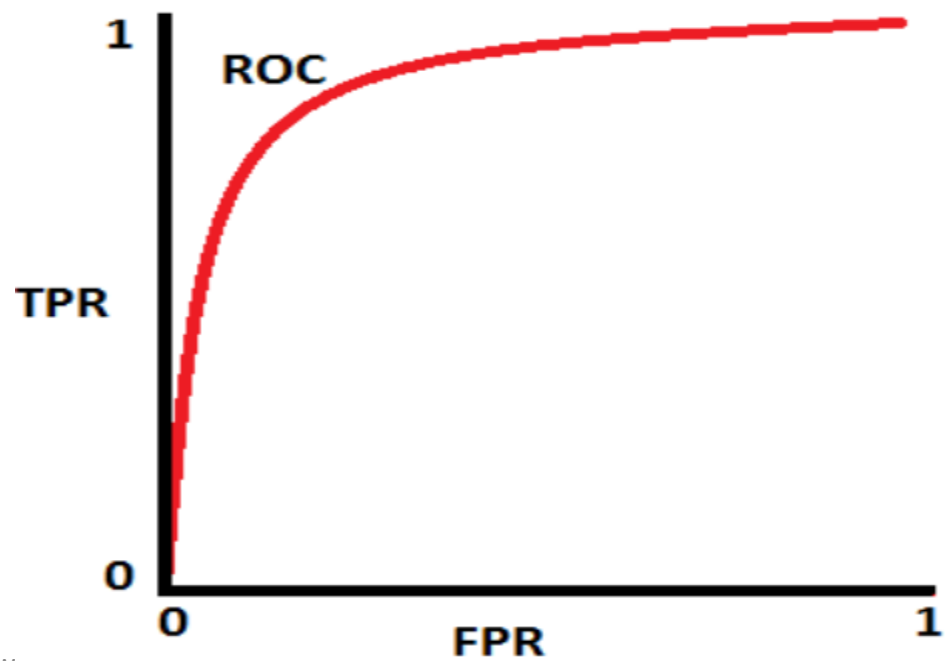
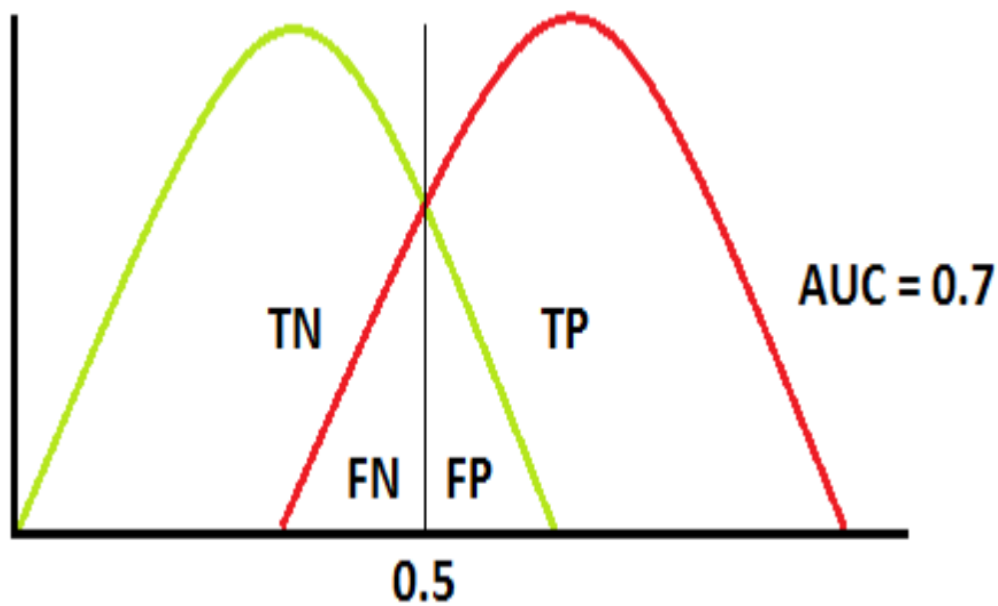
$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

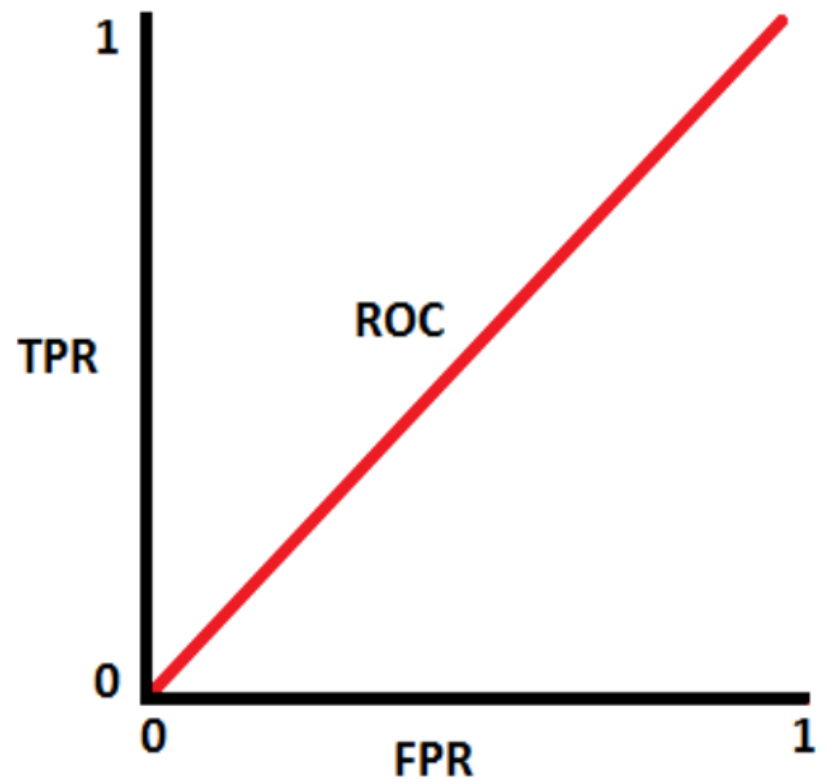
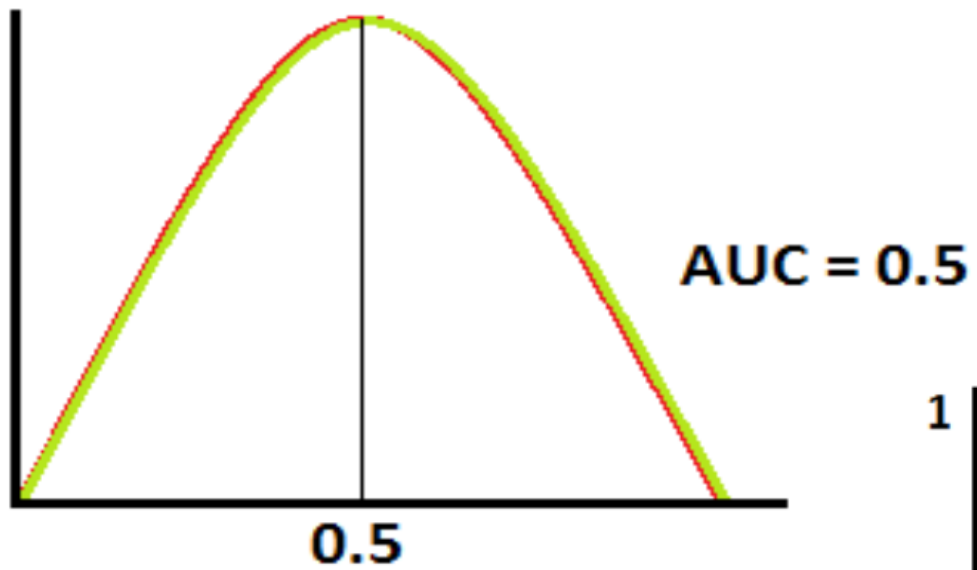
Area Under the Curve

The AUC is the area under the ROC curve. This score gives us a good idea of how well the model performs. It tells how much model is capable of distinguishing between classes.

- An excellent model has AUC near to the 1 which means it has good measure of separability.
- A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s.
- when AUC is 0.5, it means model has no class separation capacity whatsoever







Area Under the Curve

The AUC is the area under the ROC curve. This score gives us a good idea of how well the model performs. It tells how much model is capable of distinguishing between classes.

