Measuring Machine Learning Models-Accuracy_Precision_Recall_Confusion Matrix

Prepared By: Dr.Mydhili K Nair, Professor, ISE Dept, RIT

For: Machine Learning Class

Target Audience: Sem 6 Students

Why should we measure Machine Learning Models?

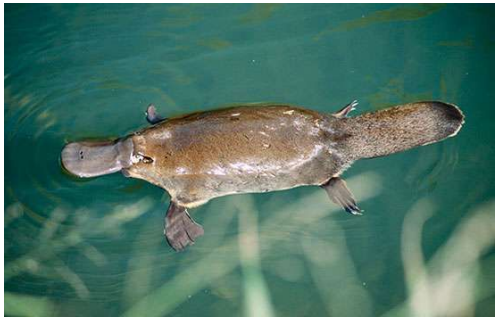
Predicted Values

Actual Values	1	0
1		
0		





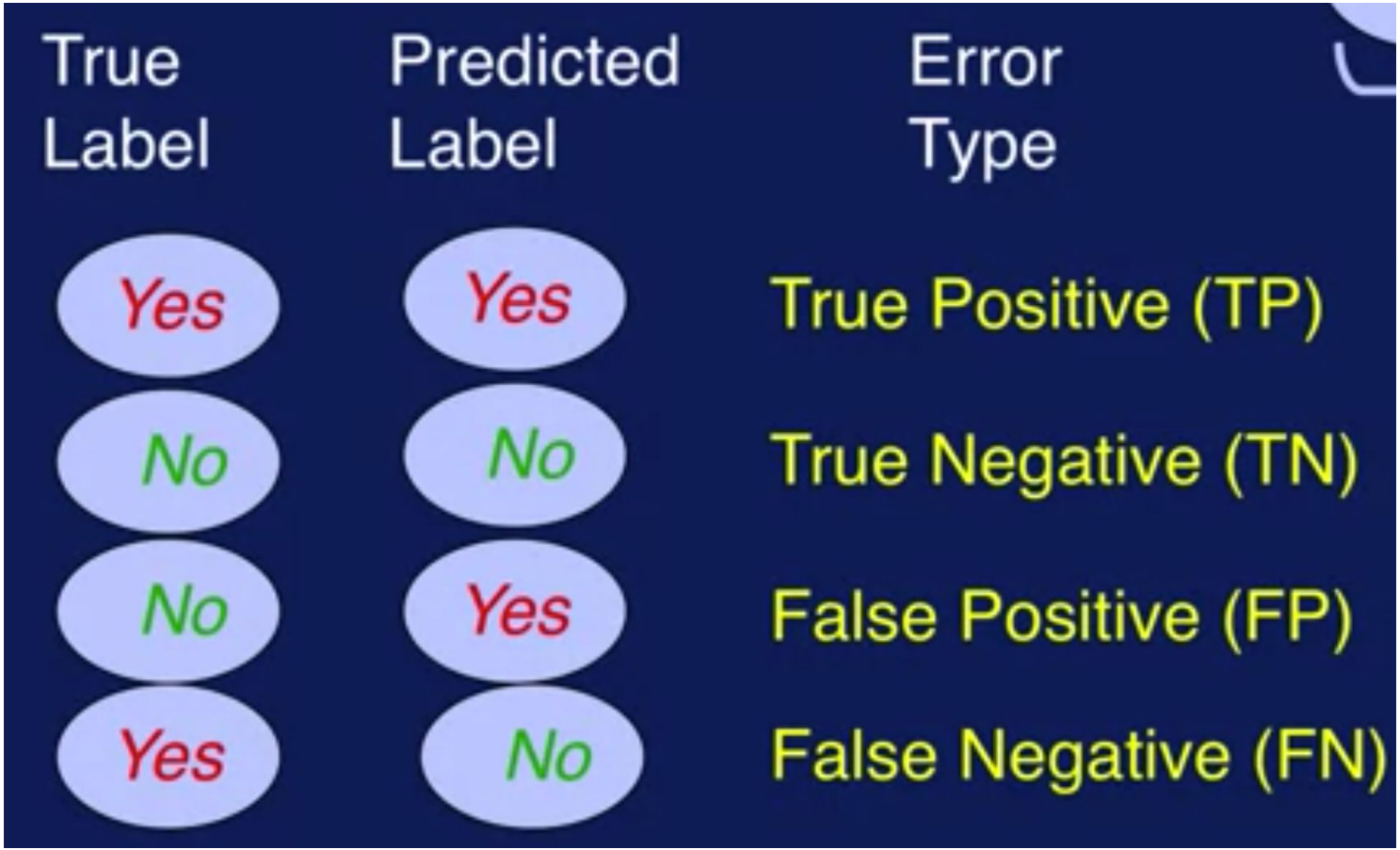





Classic Classification Problem - “Are these animals mammals?”

Prepared By: Dr.Mydhili K Nair, Prof, ISE Dept, Ramaiah Institute of Technology, Bengaluru

Types of Classification Errors



TN

FN

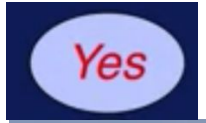
FN

TP

Characteristics

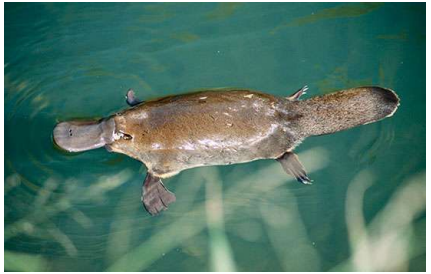
- Mammary Glands
- Fur or hair
- Three middle ear bones

TN



FP

TP



FP

Measuring Machine Learning Models

❑ Classification Metrics

- ✓ Confusion Matrix
- ✓ Accuracy
- ✓ Precision
- ✓ Recall
- ✓ F1 Score
- ✓ F-Beta Score
- ✓ Receiver Operating Characteristic Curve – ROC Curve

❑ Regression Metrics

- ✓ Mean Absolute Error
- ✓ Mean Squared Error
- ✓ R2 Score

$$\text{Accuracy Rate} = \frac{\#\ \text{correct predictions}}{\#\ \text{total predictions}}$$

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

$$\text{Error Rate} = \frac{\#\text{ incorrect predictions}}{\#\text{ total predictions}}$$

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

$$= 1 - \text{Accurate Rate}$$

True Label	Predicted Label
Yes	No
No	No
No	No
Yes	Yes
Yes	Yes
No	No
Yes	No
Yes	Yes
No	No
No	Yes

$$\text{Accuracy Rate} = \frac{\#\text{ correct predictions}}{\#\text{ total predictions}}$$

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

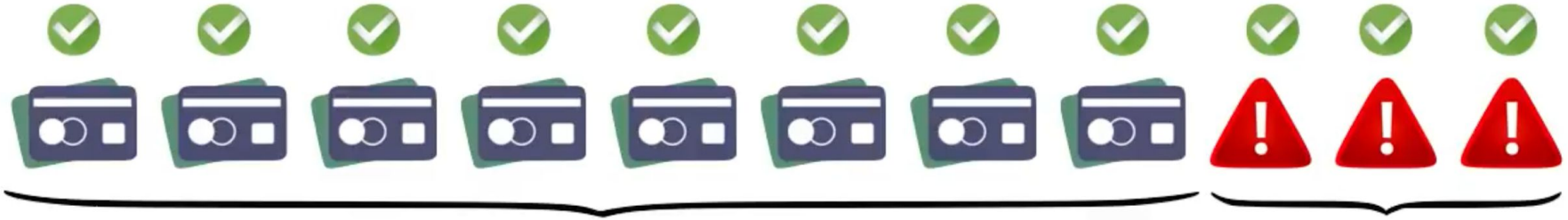
$$= (3 + 4) / 10 = 7 / 10 = 0.7$$

$$\text{Error Rate} = \frac{\#\text{ incorrect predictions}}{\#\text{ total predictions}}$$

$$= 1 - \text{Accuracy Rate}$$

$$= 1 - 0.7 = 0.3$$

Credit Card Fraud

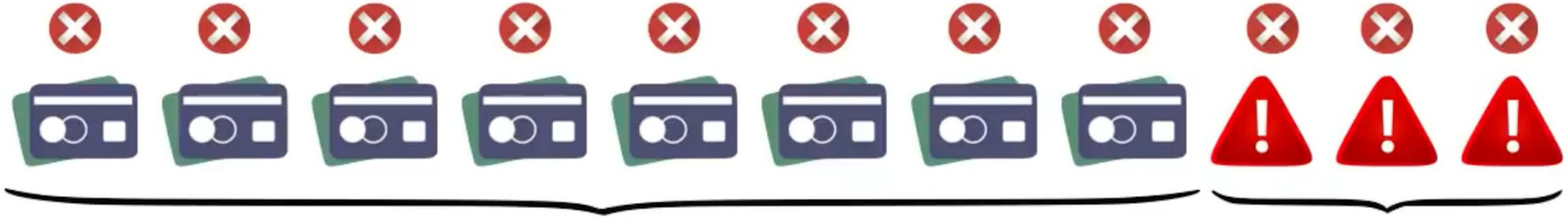


Model: All transactions are good.

$$\text{Correct} = \frac{284{,}335}{284{,}807} = 99.83\%$$

Problem: I'm not catching any of the bad ones!

Credit Card Fraud



Model: All transactions are fraudulent.

Great! Now I'm catching *all* the bad transactions!

Problem: I'm accidentally catching all the good ones!

Limitation with Accuracy

Is this tumor cancerous?

very few positive examples

most are negative examples

Class Imbalance Problem

Limitation with Accuracy

Is this tumor cancerous?

- Say 3% of samples are cancer
- If model always predicts non-cancer
  - Accuracy = 97%

97% of samples are non-cancerous

- But no cancer cases detected!

Class Imbalance Problem

Precision and Recall : Evaluation Metrics required to find out how well our model classifies positive versus negative classes

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

← All samples with Predicted = Yes

Denominator: Total # of samples predicted as being positive

Class Imbalance Problem

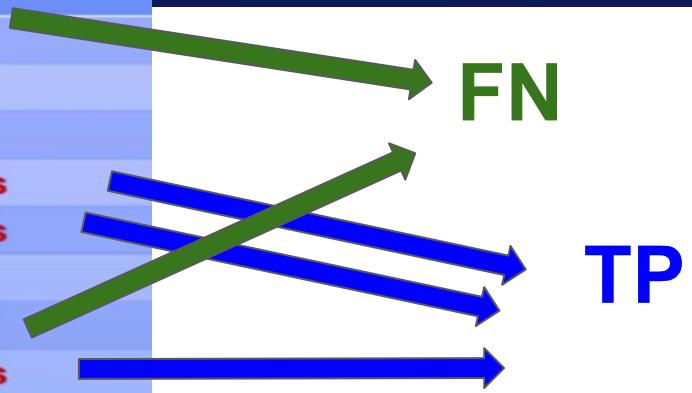
Prepared By: Dr.Mydhili K Nair, Prof, ISE Dept, Ramaiah Institute of Technology, Bengaluru

Precision and Recall : Evaluation Metrics required to find out how well our model classifies positive versus negative classes

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

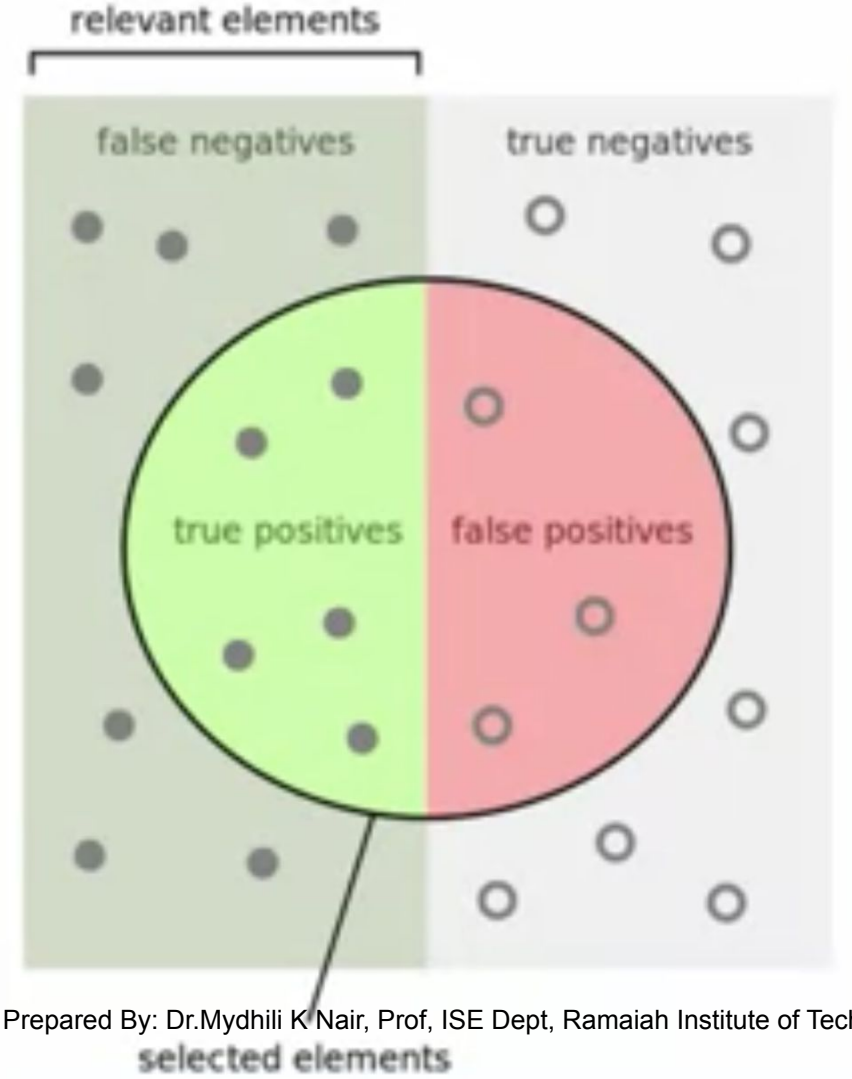
← All samples with True = Yes

True Label	Predicted Label
Yes	No
No	No
No	No
Yes	Yes
Yes	Yes
No	No
Yes	No
Yes	Yes
No	No
No	Yes



Denominator: Total # of samples belonging to the "True Class" or with "True/Actual" Label as "True"

Class Imbalance Problem

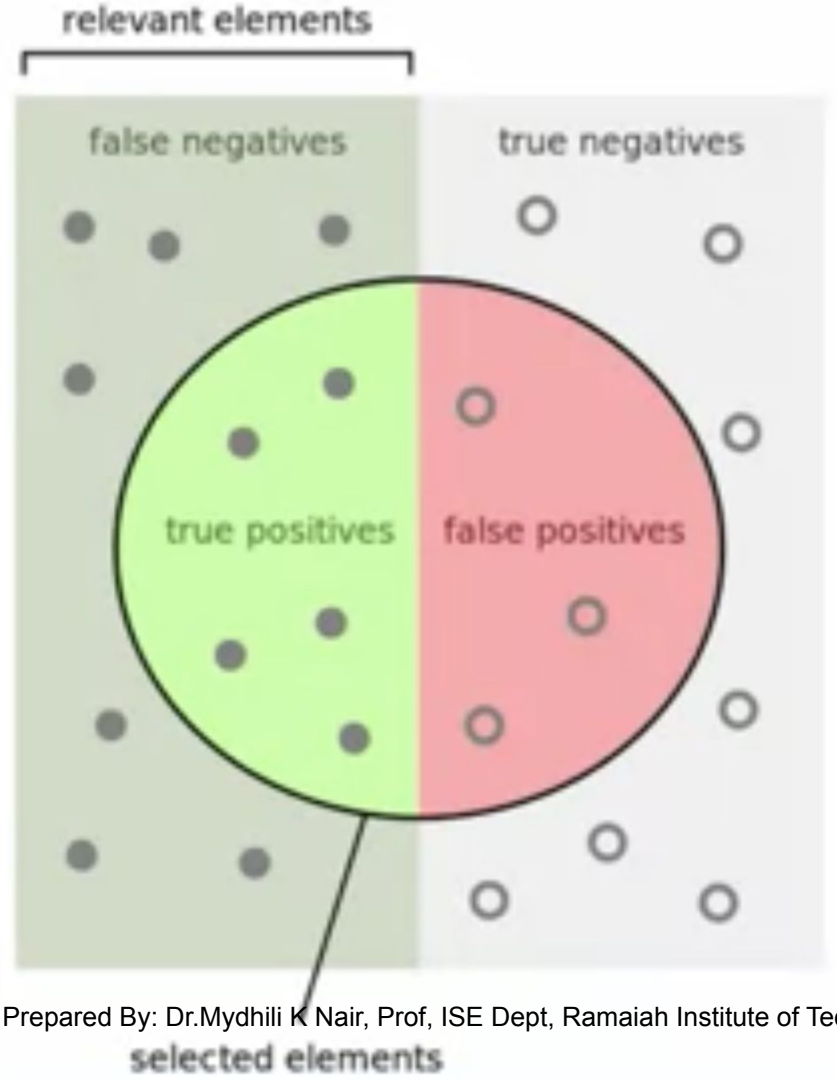


How many relevant items are selected?

Recall =

Samples correctly predicted as Positive

Samples actually Positive



How many selected items are relevant?

$$\text{Precision} = \frac{\text{Samples correctly predicted as Positive}}{\text{Samples predicted as Positive}}$$

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

All samples with Predicted = Yes

Measure of exactness

Predicted as positive which is **actually** in the "positive" class

Calculates the number of positive classes that **the model correctly identified**

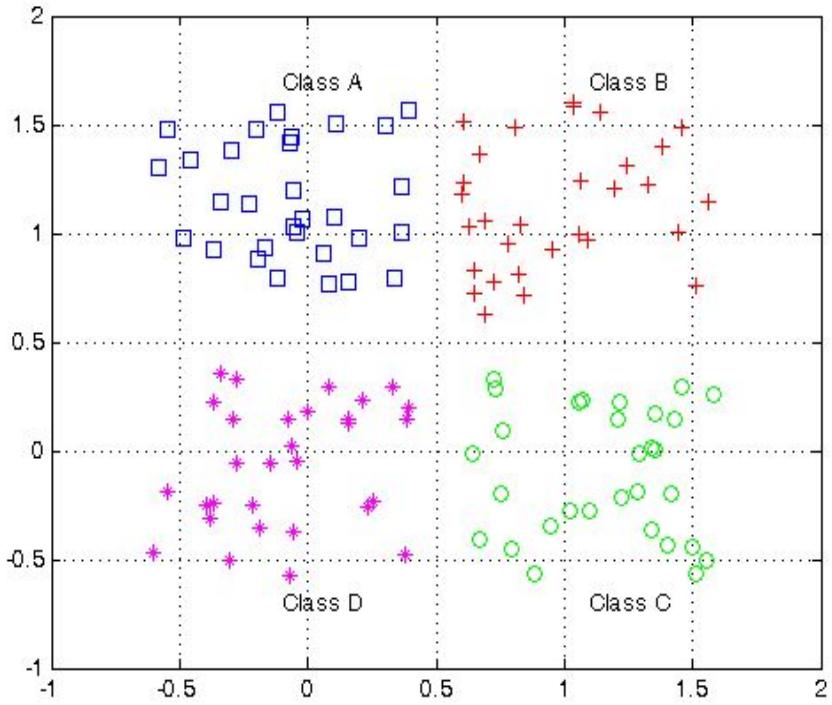
Measure of completeness

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

All samples with True = Yes

Precision + Recall

- Use together
- Goal: Maximize both



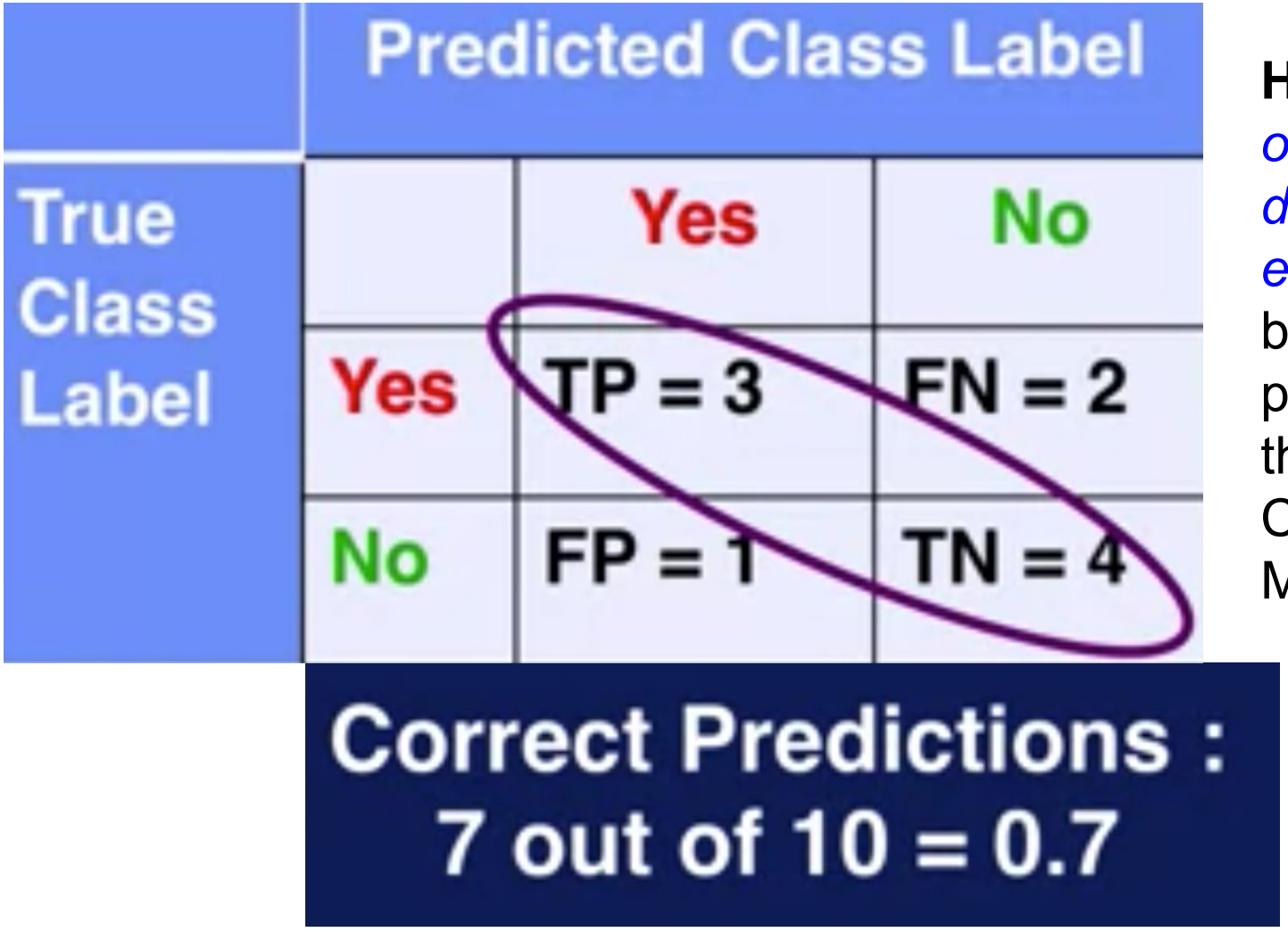
For a Fixed Value of Precision, determine recall and vice-versa.

Perfect **Precision** Score of 1 for Class A samples ⇒ Every sample belonging to Class A were indeed belonging to Class A.
Missing: What about samples of Class A that were predicted / classified incorrectly?
Precision quantifies the number of positive class predictions that actually belong to the positive class.

Perfect **Recall** Score of 1 ⇒ # of samples of Class A correctly labelled as Class A.
Missing: How many samples were incorrectly labelled as Class A?
Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

True Label	Predicted Label
Yes	No
No	No
No	No
Yes	Yes
Yes	Yes
No	No
Yes	No
Yes	Yes
No	No
No	Yes

		Predicted Class Label	
		Yes	No
True Class Label	Yes	TP = 3	FN = 2
	No	FP = 1	TN = 4



Higher *the sum of values of diagonal elements* the better the performance of the Classification Model.

	Predicted Class Label		
True Class Label		Yes	No
	Yes	TP = 3	FN = 2
	No	FP = 1	TN = 4

Lower *the sum of values of off-diagonal elements* the better the performance of the Classification Model.

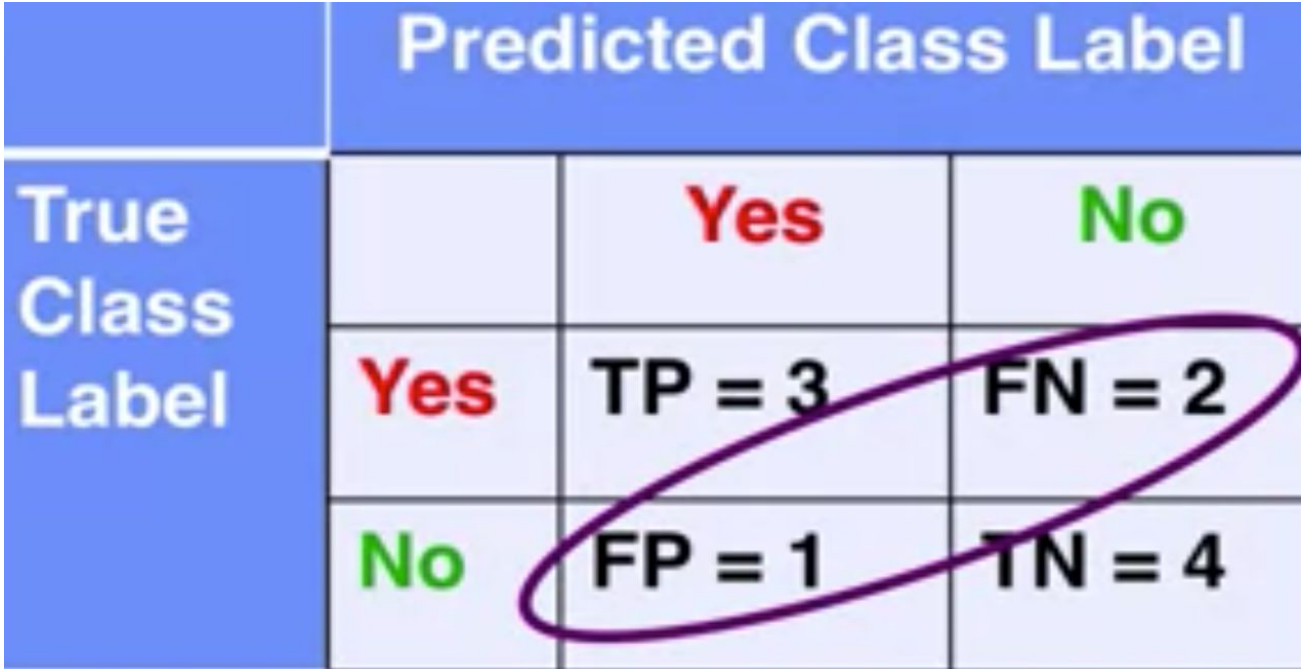
Three mis-classifications!!!

Incorrect Predictions : 3 out of 10 = 0.3

		Predicted Class Label	
		Yes	No
True Class Label	Yes	TP = 3	FN = 2
	No	FP = 1	TN = 4

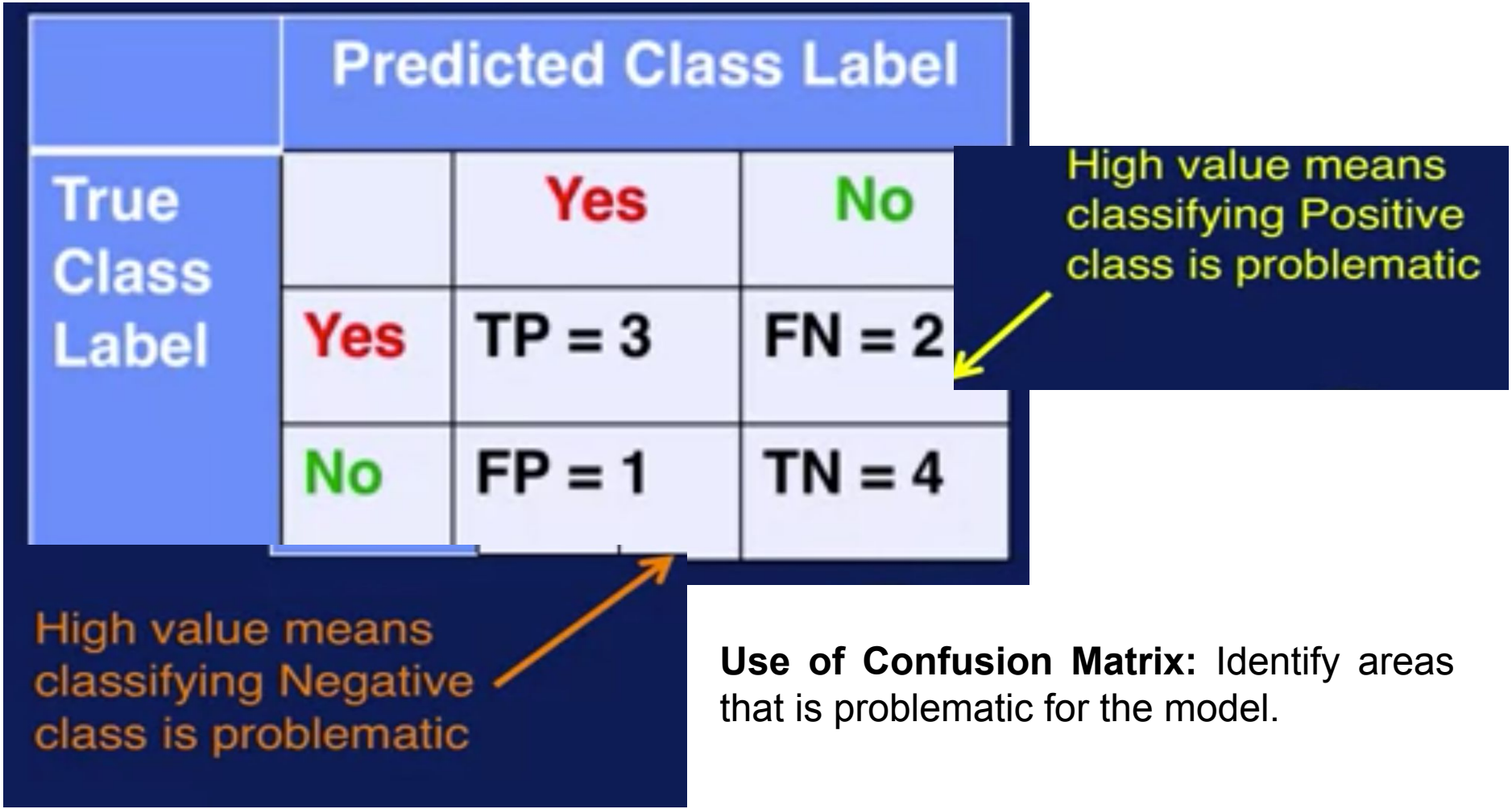
$$\text{Accuracy Rate} = \frac{\#\text{ correct predictions}}{\#\text{ total predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

The sum of the Diagonal Elements is a measure of the **Accuracy Rate**.



The sum of the Off-Diagonal Elements is a measure of the **Error Rate.**

$$\text{Error Rate} = \frac{\#\text{ incorrect predictions}}{\#\text{ total predictions}} = \frac{FN + FP}{TP + TN + FP + FN} = 1 - \text{Accurate Rate}$$



Use of Confusion Matrix: Identify areas that is problematic for the model.

CONFUSION MATRIX

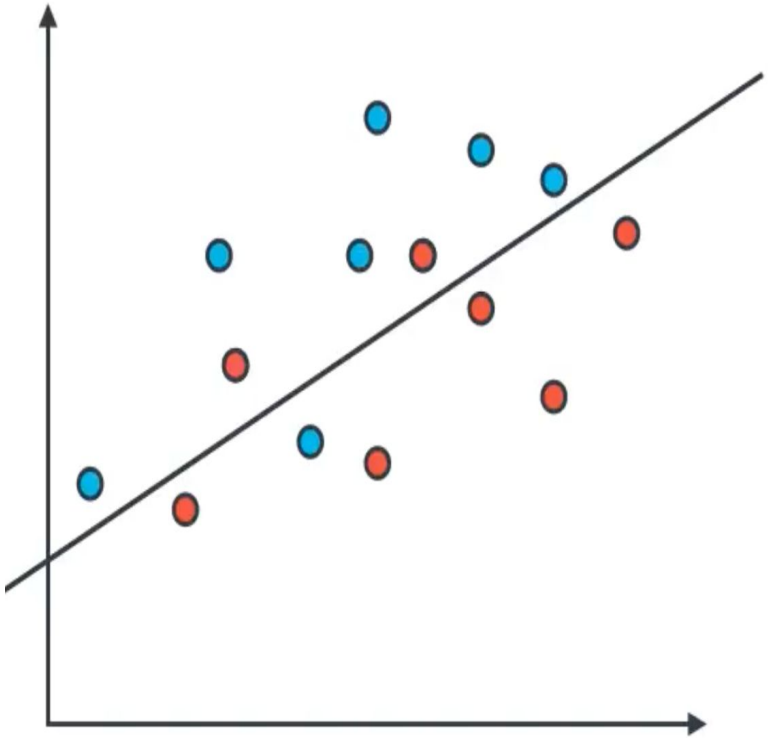
❖ **Confusion Matrix**

- ➢ a table that describes the performance of a model
- ❑ True Positive:
  - Positive point classified as positive.
- ❑ True Negative:
  - Negative point classified as negative.
- ❑ False Positive:
  - Negative point incorrectly classified as positive.
- ❑ False Negative:
  - Positive point incorrectly classified as negative.

	Guessed positive	Guessed Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

- Table used to classify the different types of errors made by a Classifier
- Values used to find the performance of a classifier (accuracy & error rate)
- Indicates what types of errors the model is making

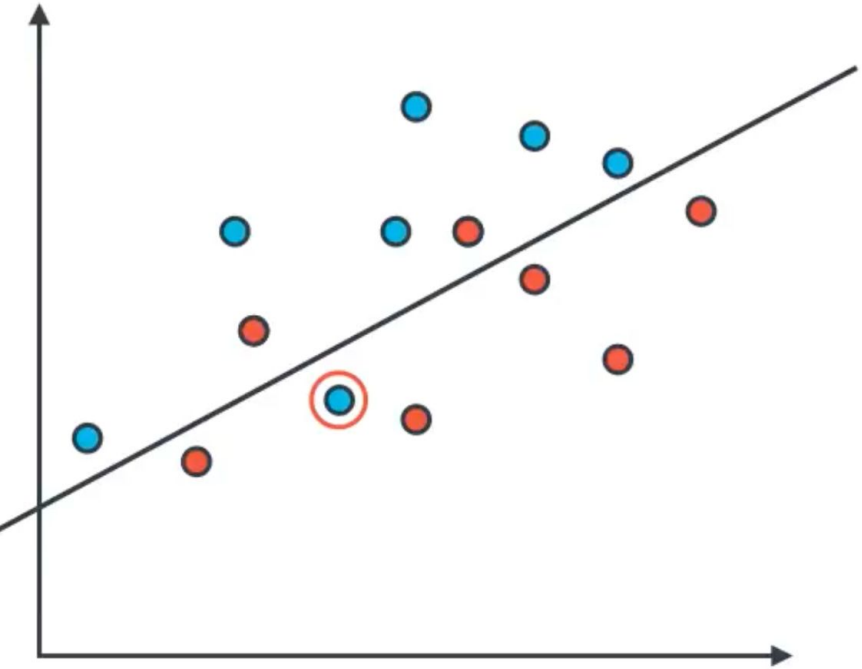
Confusion Matrix



Prediction

Data	Guessed Positive	Guessed Negative
Positive		
Negative		

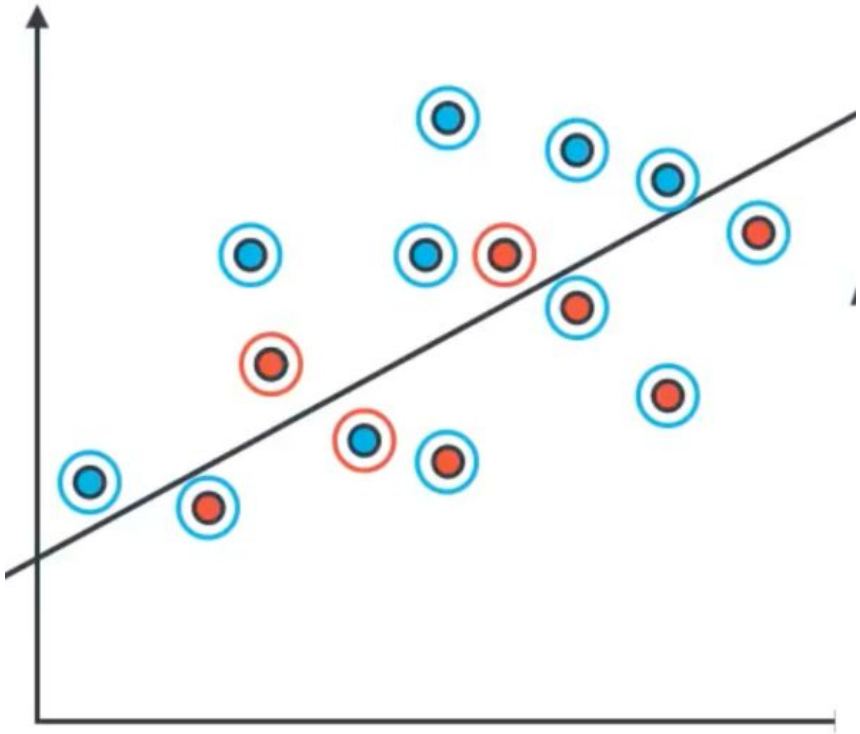
Confusion Matrix



Prediction

Data	Guessed Positive	Guessed Negative
Positive	6 True positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

Precision: Out of all the data, how many points did we classify correctly?

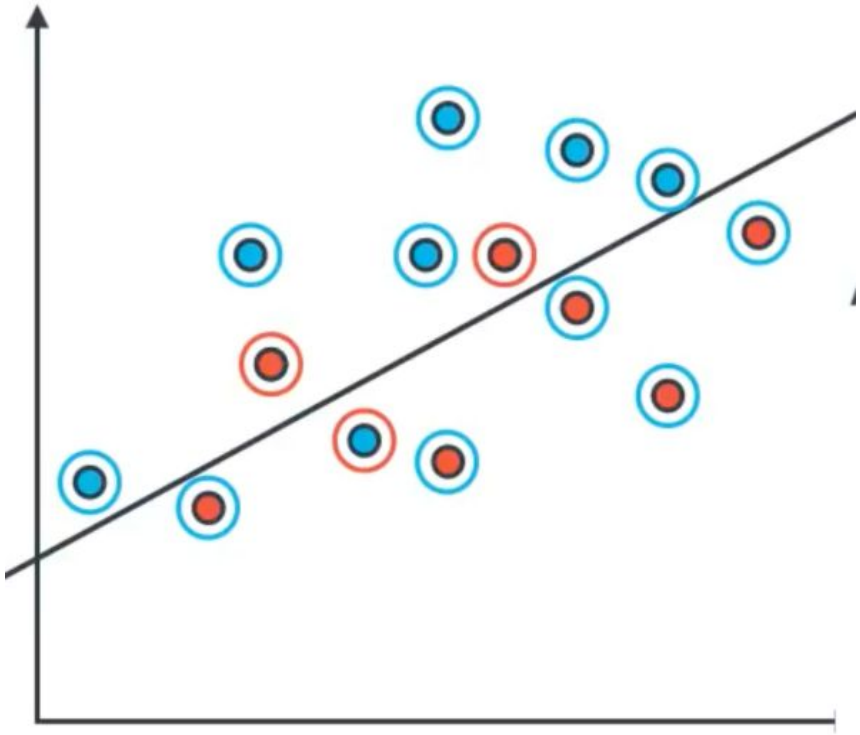


$$\text{Accuracy} = \frac{\text{Correctly Classified points}}{\text{All points}}$$

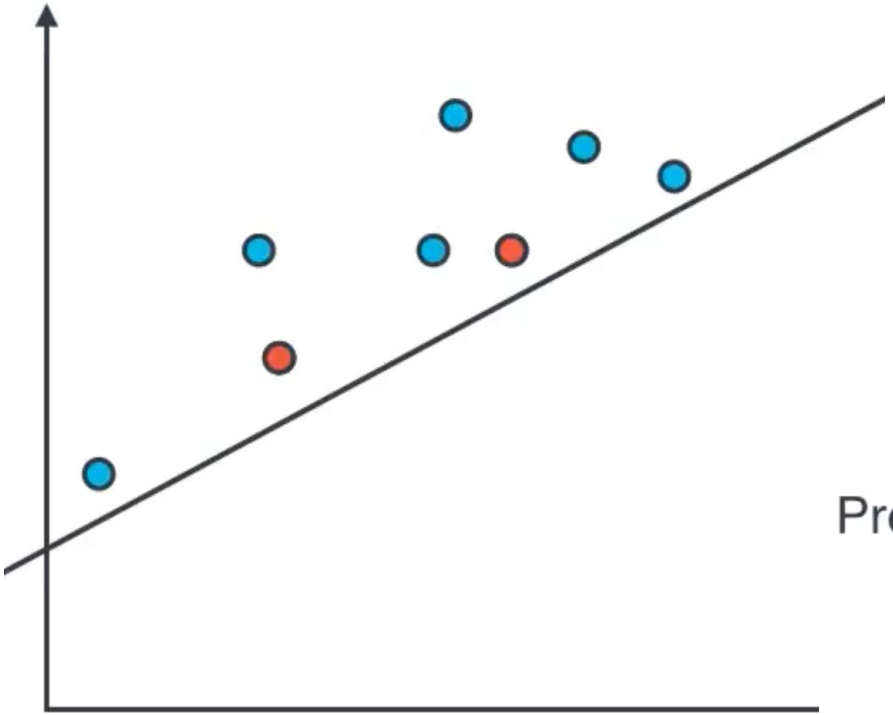
$$= \frac{11}{11 + 3}$$

$$= \frac{11}{14}$$

$$= 78.57\%$$



Precision: Out of the points we've predicted to be positive, how many are correct?



Precision: Out of the points we've predicted to be positive, how many are correct?

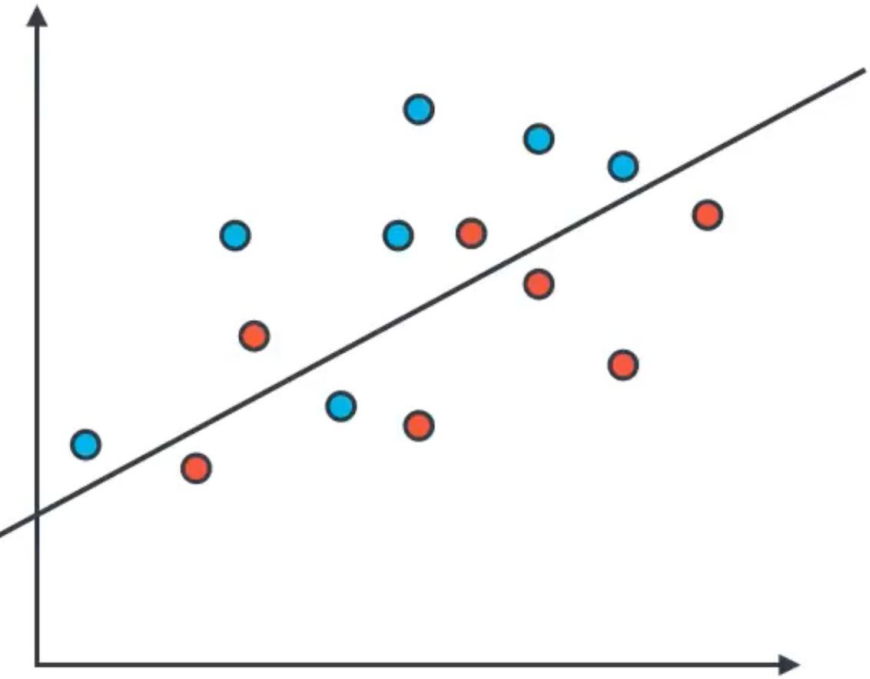
$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False Positives}}$$

$$= \frac{6}{6 + 2}$$

$$= \frac{6}{8}$$

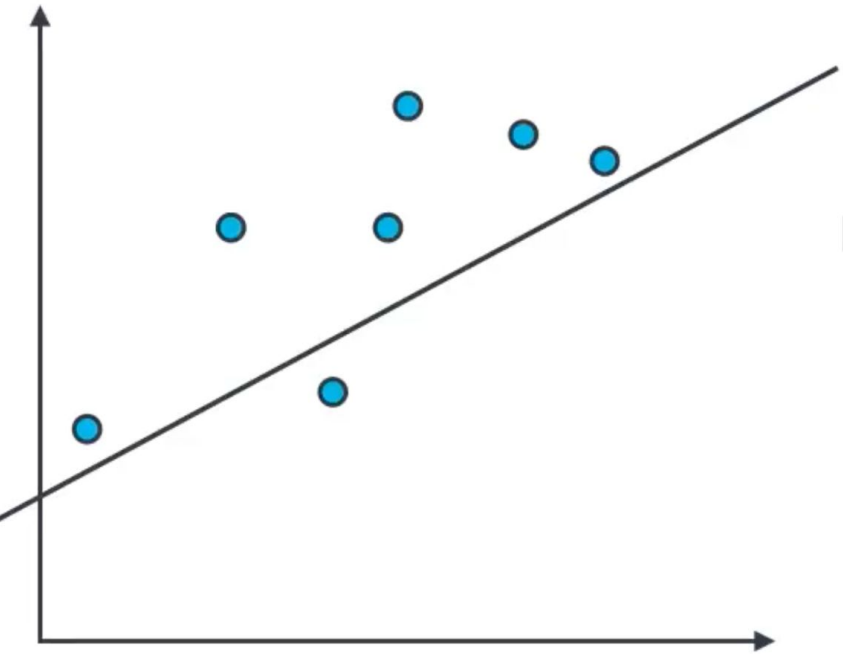
$$= 75\%$$

Recall



Recall: Out of the points labelled positive, how many did we correctly predict?

Recall



Recall: Out of the points labelled positive, how many did we correctly predict?

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}}$$

$$= \frac{6}{6 + 1}$$

$$= \frac{6}{7}$$

$$= 85.7\%$$

Case Study #1:

Medical Model

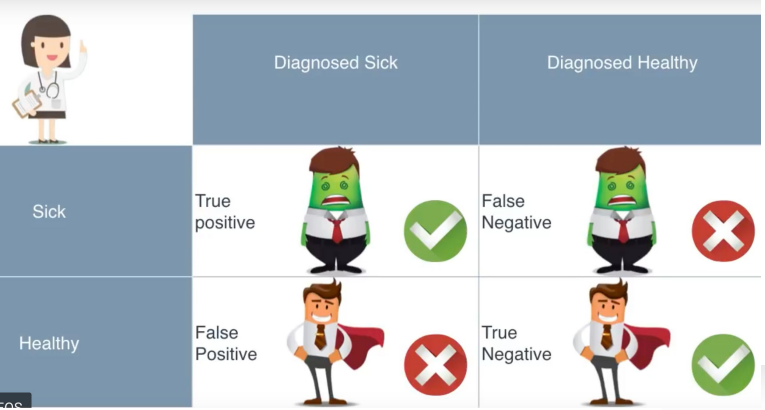
Healthy

Sick

Case Study #2:

Spam Classifier Model

Not spam



$$\text{Accuracy} = \frac{1{,}000 + 8{,}000}{10{,}000} = 90\%$$

Confusion Matrix

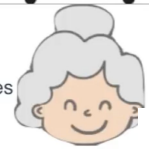
Case Study #1: Medical Model

E.g. of Accuracy is not a useful measure

10,000 Patients

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000 True positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

	Sent to Spam Folder	Sent to Inbox
Spam	True Positives 	False Negatives 
Not Spam	False Positives 	True Negatives 

$$\text{Accuracy} = \frac{100 + 700}{1000} = 80\%$$

Case Study #2: Spam Mail Filter Model

E.g. of Accuracy is not a useful measure

1,000 e-mails

Confusion Matrix

E-mail \ Folder	Spam Folder	Inbox
Spam	100 True positives	170 False Negatives
Not spam	30 False Positives	700 True Negatives

Case Study #1: Medical Model

High RECALL Model

	Diagnosed Sick	Diagnosed Healthy
Sick		**FN Not "OK"** False Negative  
Healthy	False Positive  	

Case Study #2: Spam Mail Filter Model

High PRECISION Model

	Sent to Spam Folder	Sent to Inbox
Spam		False Negatives  
Not Spam	**FP Not "OK"** False Positives  	

FN Not “OK”

False Negative

High RECALL Model

FP Not “OK”

False Positives

High PRECISION Model

EVALUATION METRICS

Medical Model

False positives ok
False negatives **NOT** ok

Find all the sick people
Ok if not all are sick

High Recall Model

Spam Detector

False positives **NOT** ok
False negatives ok

You don't necessarily need to find all spam
But they better all be spam

High Precision Model

Precision

Folder

E-mail

	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Precision: Out of the all the e-mails, sent to the spam inbox, how many were actually spam?

Precision

E-mail \ Folder	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Precision: Out of the all the e-mails, sent to the spam inbox, how many were actually spam?

$$\text{Precision} = \frac{100}{100 + 30} = 76.9\%$$

Precision

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Is Healthy	600	9000

Precision: Out of the patients we diagnosed with an illness, how many did we classify correctly?

Recall

E-mail \ Folder	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Recall: Out of the all the spam e-mails, how many were correctly sent to the spam folder?

Recall

E-mail \ Folder	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Recall: Out of the all the spam e-mails, how many were correctly sent to the spam folder?

$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

Precision

Patients	Diagnosis: Diagnosed sick	Diagnosed Healthy
Sick	1000	200 ✗
Healthy	800	8000

Precision: Out of the patients we diagnosed with an illness, how many did we classify correctly?

$$\text{Precision} = \frac{1{,}000}{1{,}000 + 800} = 55.7\%$$

Recall

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1000	200
Is Healthy	800	8000

Recall: Out of the sick patients, how many did we correctly diagnose as sick?

Recall

Patients \ Diagnosis	Diagnosed Sick	Diagnosed Healthy
Sick	1000	200 ✗
Is Healthy	800	8000

Recall: Out of the sick patients, how many did we correctly diagnose as sick?

$$\text{Recall} = \frac{1{,}000}{1{,}000 + 200} = 83.3\%$$

Precision and Recall

Medical Model

Precision: 55.7%

Recall: 83.3%

Spam Detector

Precision: 76.9%

Recall: 37%

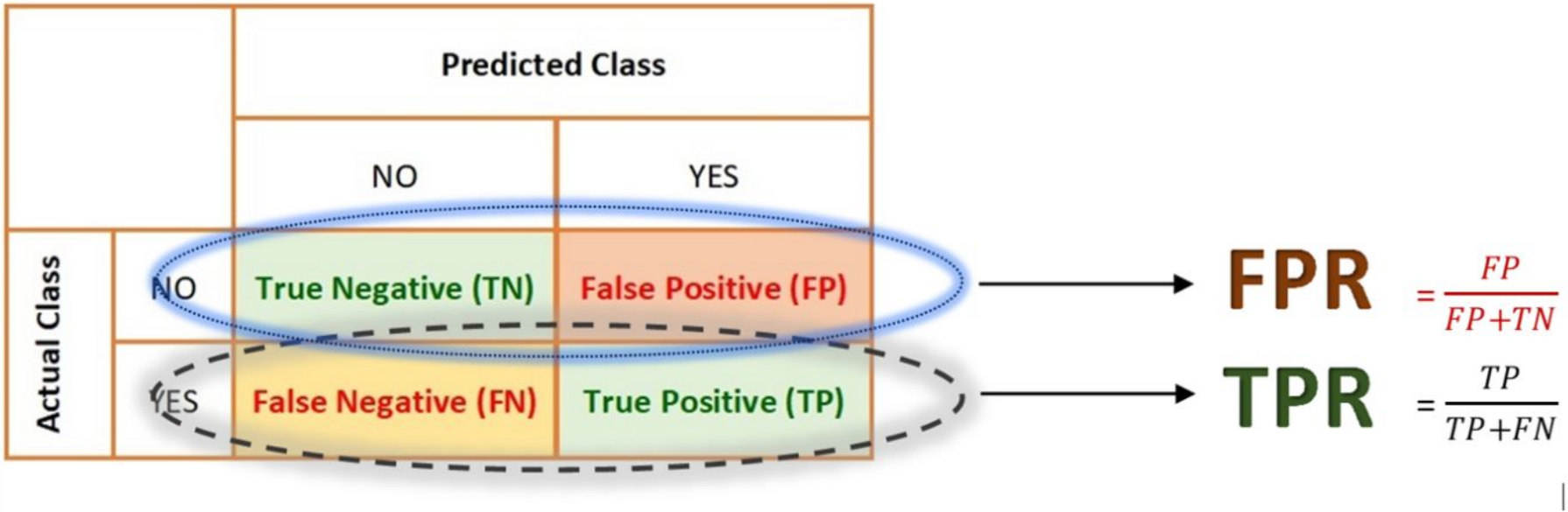
Don't Confuse with this king of Confusion Matrix!

		Predicted Class	
		NO	YES
Actual Class	NO	True Negative (TN)	False Positive (FP)
	YES	False Negative (FN)	True Positive (TP)

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

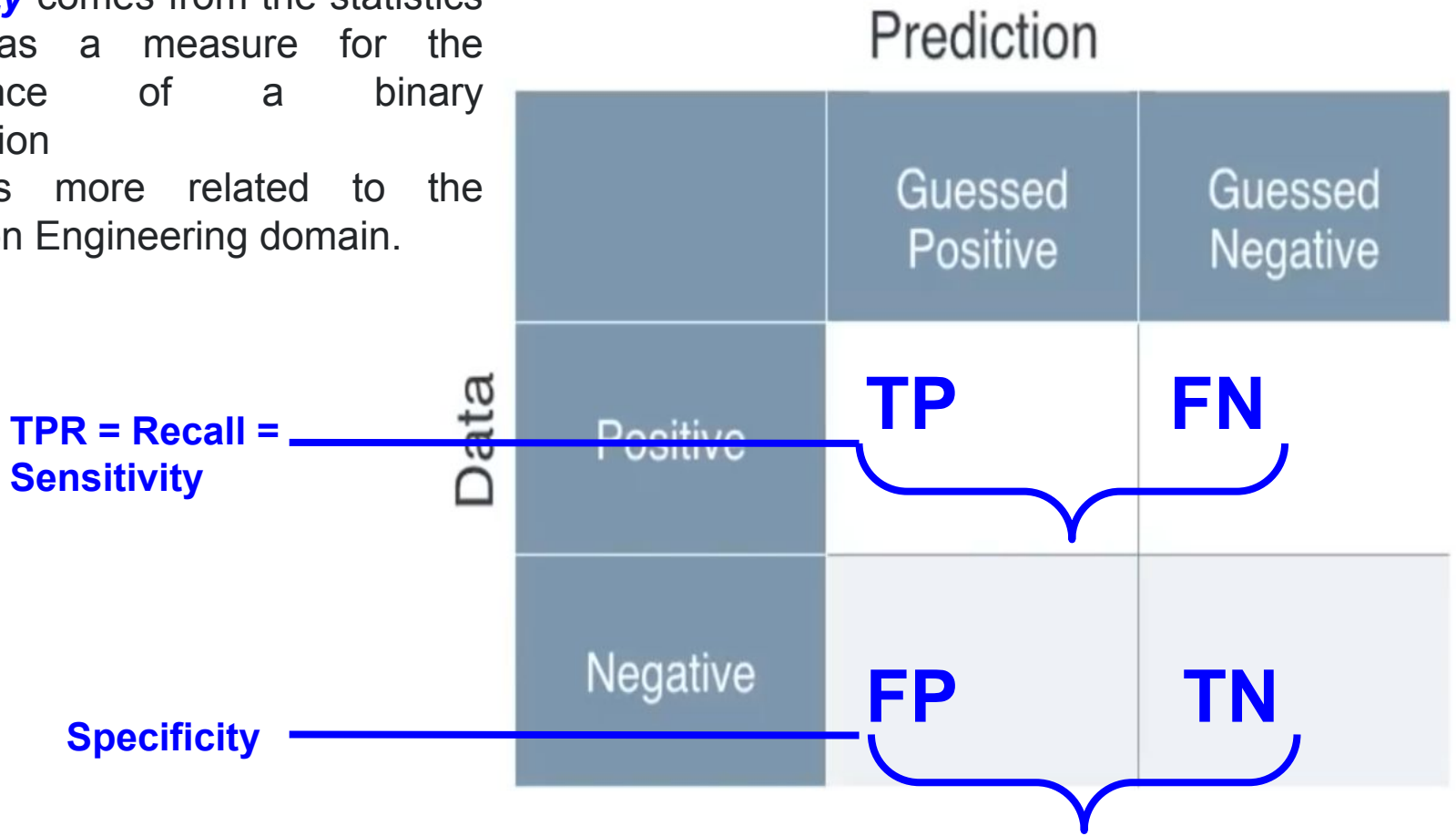
$$FPR = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

Don't Confuse with this king of Confusion Matrix!



Sensitivity comes from the statistics domain as a measure for the performance of a binary classification

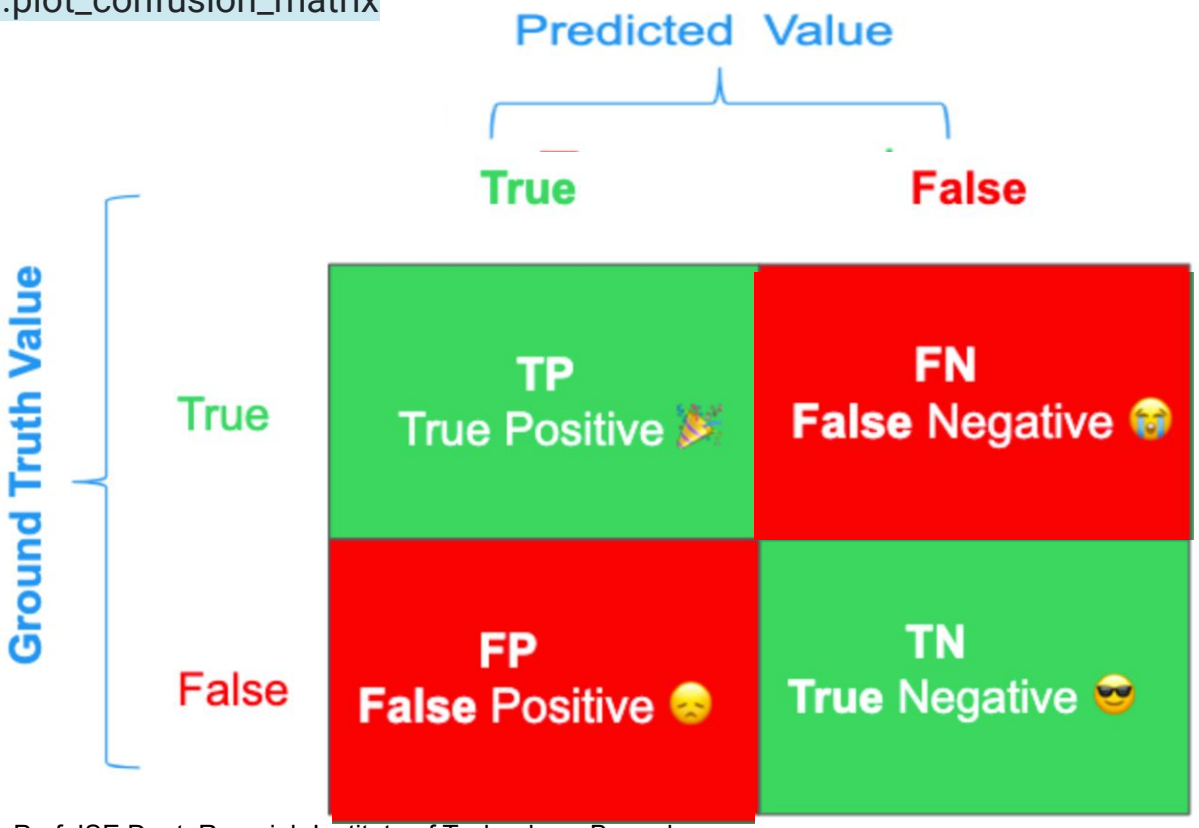
Recall is more related to the Information Engineering domain.



Confusion Matrix

NOTE: Correct Confusion Matrix to study.

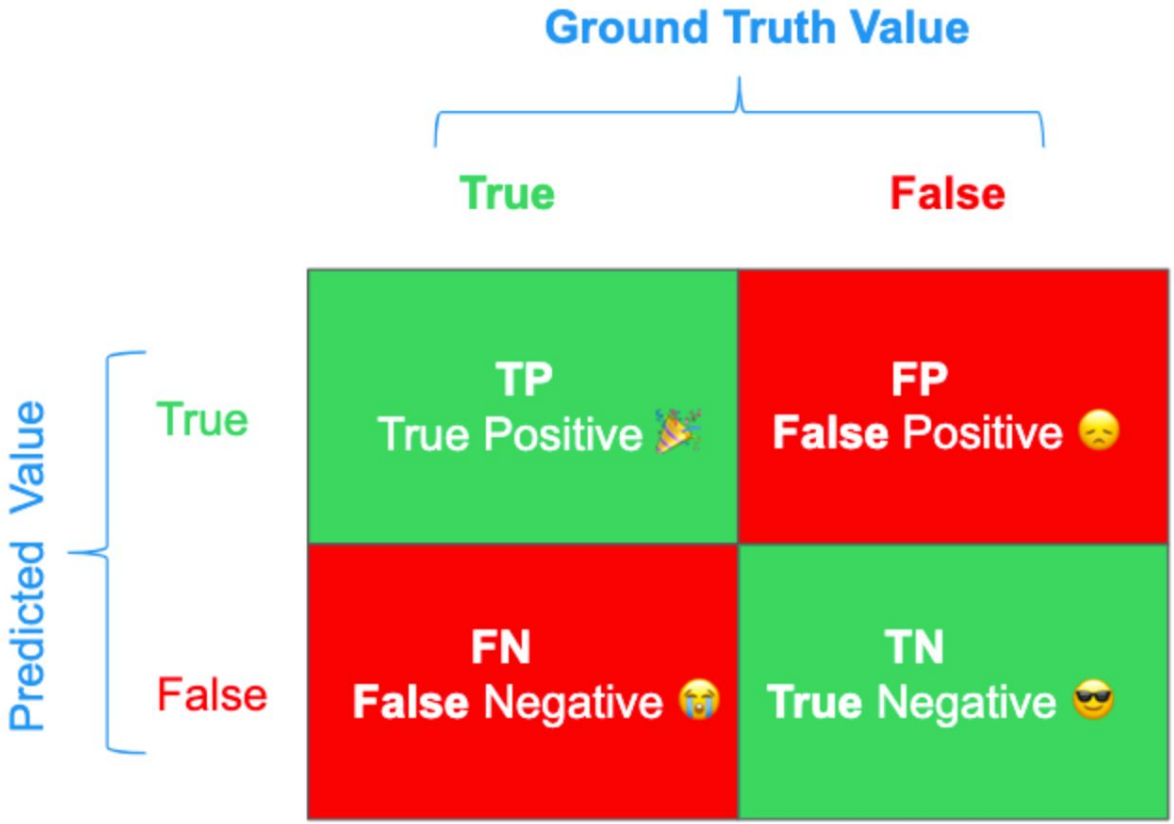
`sklearn.metrics`.plot_confusion_matrix



Confusion Matrix

Difference #1 :

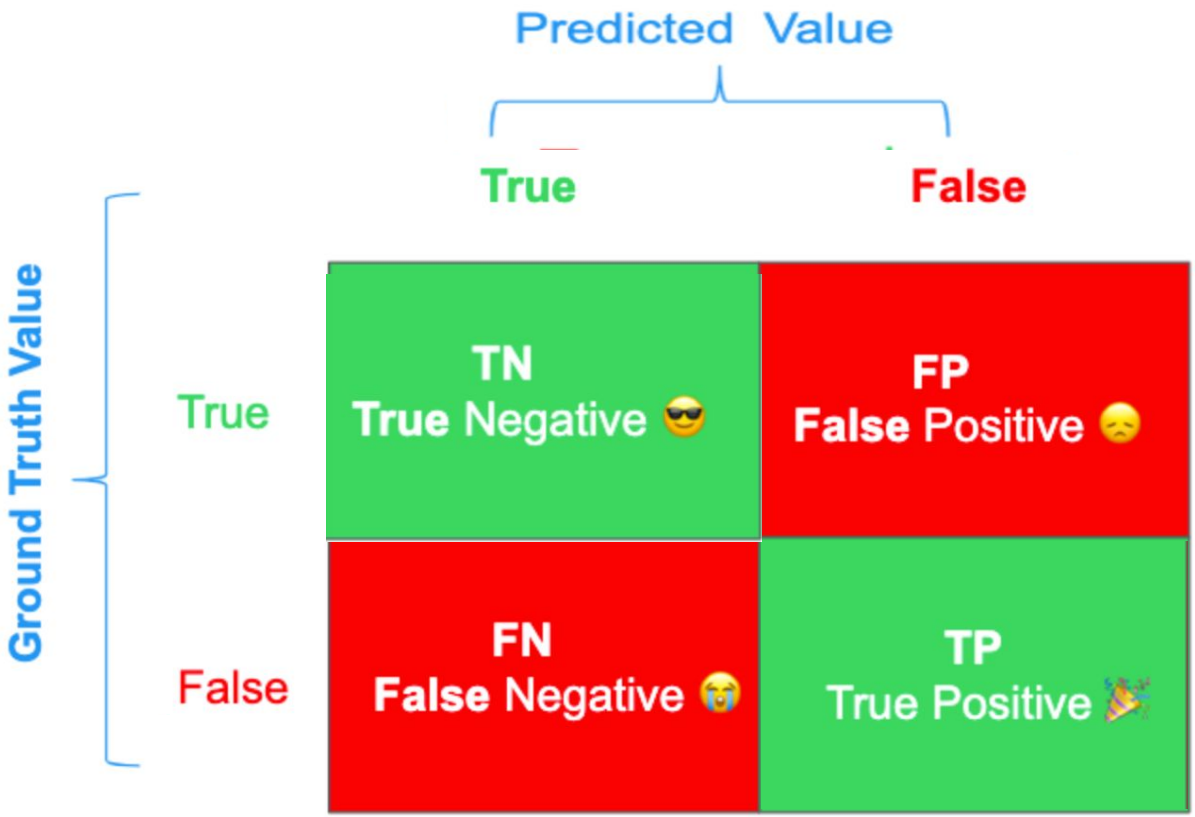
NOTE: Be Careful of the order in which the Confusion Matrix is written. Here the "True Label" and "Predicted Label" positions are reversed. Therefore "FP" and "FN" positions are changed.



Confusion Matrix

NOTE: Be Careful of the order in which the Confusion Matrix is written. Here the “TN”, “TP” , “FN”, “FP” positions are reversed.

Difference #2 :



High Precision or High Recall?

1. People are innocent until proven guilty as per Indian Law. We want to avoid false convictions, even at the cost of criminals running free.
2. In Planet Utopia, the law demands that the number of criminals running free is brought as minimum as possible, even at the cost of wrongly convicting innocent people.
3. In an Examination Results ML Model, we need to correctly identify students who have failed. The model should never send a failed student home by classifying them as pass.

High Precision or High Recall?

(Refer Tutorial #2)

High Precision or High Recall?

5. At a Call Center there are thousands of free customers registering in their website every week. The call center team wants to call them all, but it is impossible, so they ask you, their data analyst, to select those with good chances to be a buyer. The idea is to not call a guy that is not going to buy, but to ensure that potential buyers who showed interest, are identified, so they don't go without buying.

- **Precision**: Call all the customers who enquired. Dont mind “False Positives”
- **Recall** : Call those customers who showed interest at the cost of not calling all of them.
  Remember - Recall involves “Opportunity Cost”. Precision involves “Direct Cost”

Precision	Recall
The ability of a classification model to identify only the relevant data points.	The ability of a model to find all the relevant cases within a dataset.
Precision quantifies the number of positive class predictions that actually belong to the positive class.	Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.
Precision = TP / (TP + FP)	Recall = TP / (TP + FN)
E.g. Spam Mail Filter	E.g. Medical Model

DEMOS

A Pirate's Guide to Accuracy, Precision, Recall, and Other Scores

Refer:
Pirate_Handcoded_Sklearn.ipynb

China's task is to diagnose 100 patients in Wuhan with Corona Virus present in 50% of its general population. Chinese data analysts assume a black box model, where they put in information about patients and receive a score between 0 and 1. They can alter the threshold for labeling a patient as positive (has the disease) to maximize the classifier performance. They will evaluate thresholds from 0.0 to 1.0 in increments of 0.1, at each step calculating the precision, recall, F1, and location on the ROC curve. Following are the classification outcomes at each threshold

Threshold	TP	FP	TN	FN
0.0	50	50	0	0
0.1	48	47	3	2
0.2	47	40	9	4
0.3	45	31	16	8
0.4	44	23	22	11
0.5	42	16	29	13
0.6	36	12	34	18
0.7	30	11	38	21
0.8	20	4	43	33
0.9	12	3	45	40
1.0	0	0	50	50

Outcome of model at each threshold

Refer: Corona Virus.ipynb
For Python Code Solution

Corona Virus: Pen and Paper Solution

Threshold =0.5	Actual Positives	Actual Negatives
Predicted Positives	42 (TP)	16 (FP)
Predicted Negatives	13 (FN)	29 (TN)

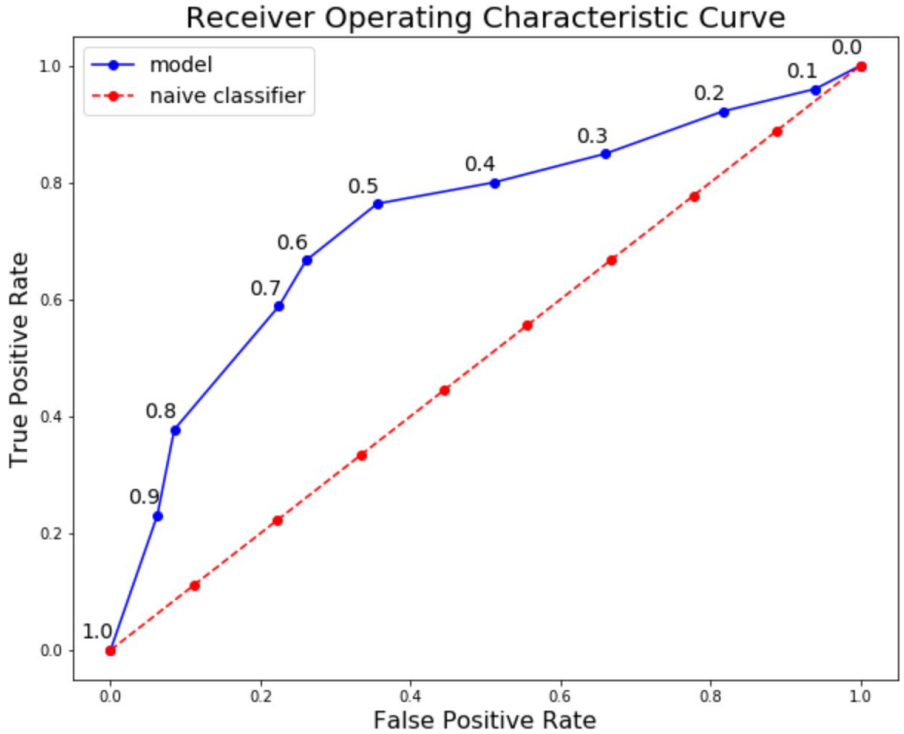
Confusion Matrix for Threshold of 0.5

$$true\ positive\ rate = \frac{TP}{TP + FN} = \frac{42}{42 + 13} = 0.76 \qquad false\ positive\ rate = \frac{FP}{FP + TN} = \frac{16}{16 + 29} = 0.36$$

$$recall = \frac{TP}{TP+FN} = \frac{42}{42+13} = 0.76 \qquad precision = \frac{TP}{TP+FP} = \frac{42}{42+16} = 0.724 \qquad F1\ Score = 2 * \frac{precision*recall}{precision+recall} = 0.74$$

Corona Virus: Pen and Paper Solution

threshold	recall	precision	f1	tpr	fpr
0.0	1	0.5	0.666667	1	1
0.1	0.96	0.505263	0.662069	0.96	0.94
0.2	0.921569	0.54023	0.681159	0.921569	0.816327
0.3	0.849057	0.592105	0.697674	0.849057	0.659574
0.4	0.8	0.656716	0.721311	0.8	0.511111
0.5	0.763636	0.724138	0.743363	0.763636	0.355556
0.6	0.666667	0.75	0.705882	0.666667	0.26087
0.7	0.588235	0.731707	0.652174	0.588235	0.22449
0.8	0.377358	0.833333	0.519481	0.377358	0.0851064
0.9	0.230769	0.8	0.358209	0.230769	0.0625
1.0	0	0	0	0	0



Inference: Based on the **highest value of F1 Score**, the overall best model occurs at a threshold of **0.5.** It is also the point which has the maximum ROC - AUC.

Prepared By: Dr.Mydhili K Nair, Prof, ISE Dept, Ramaiah Institute of Technology, Bengaluru

Sources

1. Pirate Gold Example: https://blog.floydhub.com/a-pirates-guide-to-accuracy-precision-recall-and-other-scores/
2. Corona Virus Solution Code: https://github.com/WillKoehrsen/Data-Analysis/blob/master/recall_precision/recall_precision_example.ipynb
3. Precision, recall, sensitivity and specificity: https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/
4. Confusion Matrix Explained Well: https://towardsdatascience.com/confusion-matrix-and-class-statistics-68b79f4f510b