

DATA SCIENCE

CONFUSION MATRIX

Confusion Matrix: table to describe the performance of a classifier

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Example: Test for presence of disease

NO = negative test = False = 0

YES = positive test = True = 1

- *How many classes are there?*
- *How many patients?*
- *How many times is disease predicted?*
- *How many patients actually have the disease?*

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

Basic Terminology:

- *True Positives (TP)*
- *True Negatives (TN)*
- *False Positives (FP)*
- *False Negatives (FN)*

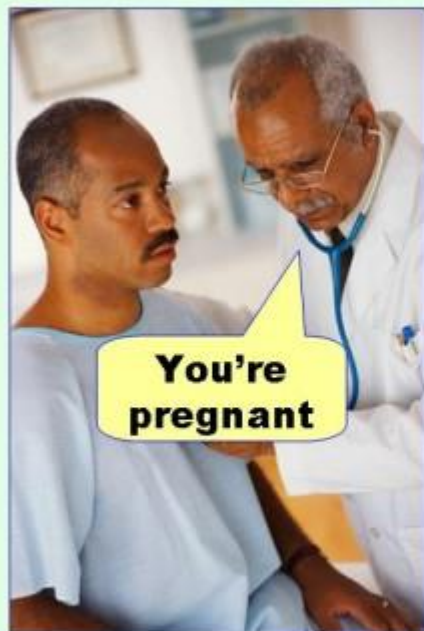
Accuracy:

- *Overall, how often is it **correct**?*
- $(TP + TN) / total = 150 / 165 = 0.91$

Misclassification Rate (Error Rate):

- *Overall, how often is it **wrong**?*
- $(FP + FN) / total = 15 / 165 = 0.09$

Type I error
(false positive)



Type II error
(false negative)



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

False Positive Rate:

- When actual value is **negative**, how often is prediction **wrong**?
- $FP / \text{actual no} = 10/60 = 0.17$

Sensitivity:

- When actual value is **positive**, how often is prediction **correct**?
- $TP / \text{actual yes} = 100/105 = 0.95$
- “True Positive Rate” or “Recall”

Specificity:

- When actual value is **negative**, how often is prediction **correct**?
- $TN / \text{actual no} = 50/60 = 0.83$