Metrics for Evaluating Classifier Performance

Prepared by

Dr.Siddique Ibrahim

Basic Terminology

- **Positivie Tuples -**Metrics for Evaluating Classifier Performance
- **Nagative Tuples -**all other tuples

Accuracy

- Suppose we use our classifier on a test set of labeled tuples.
- **P** is the number of positive tuples and **N** is the number of negative tuples.
- For each tuple, we compare the classifier's class label prediction with the tuple's known class label.

Four additional terms we need to know that are the "building blocks"

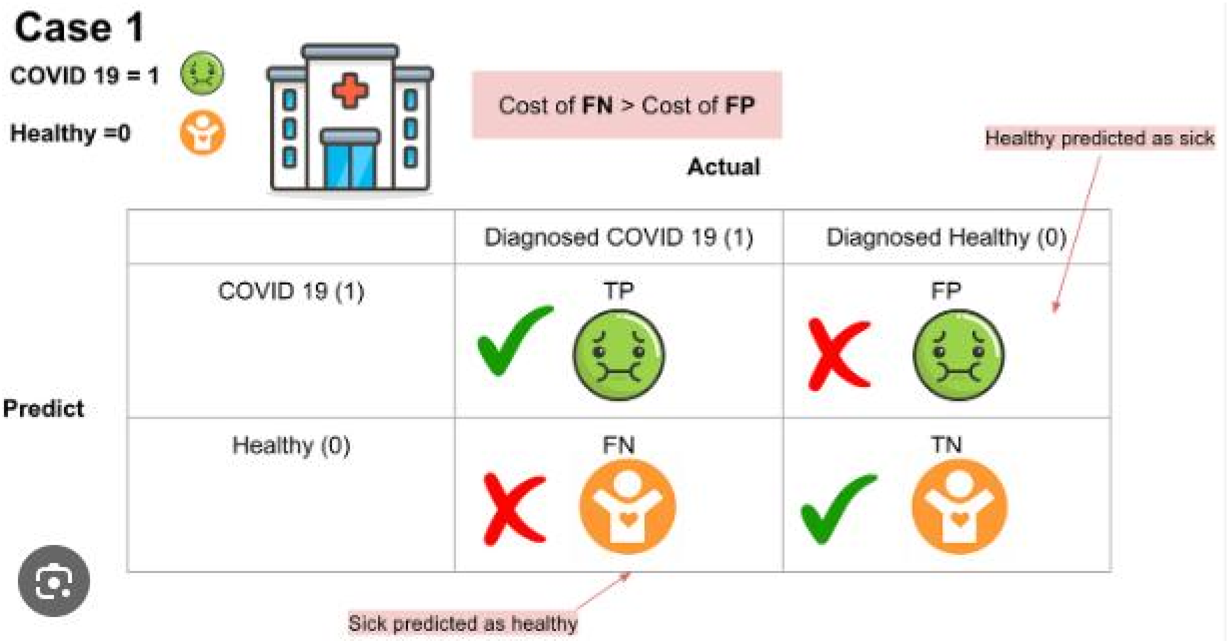
- **True Positives (TP):** These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- **True Negatives(TN):** These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- **False Positives (FP):** These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer = yes). Let FP be the number of false positives.
- **False Negatives (FN):** These are the positive tuples that were mislabeled as neg_x0002_ative (e.g., tuples of class buys computer = yes for which the classifier predicted buys computer = no). Let FN be the number of false negatives.

- Suppose we use our classifier on a test set of labeled tuples.
- **P** is the number of positive tuples and **N** is the number of negative tuples.
- For each tuple, we compare the classifier's class label prediction with the tuple's known class label.

Confusion matrix

- The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes.
- **TP and TN** tell us when the classifier is getting things right,
- while **FP and FN** tell us when the classifier is getting things wrong

	Predicted Class		
		Yes	No
Actual Class	Yes	**TP**	**FN**
	No	**FP**	**TN**



Case 1

COVID 19 = 1

Healthy =0

Cost of **FN** > Cost of **FP**

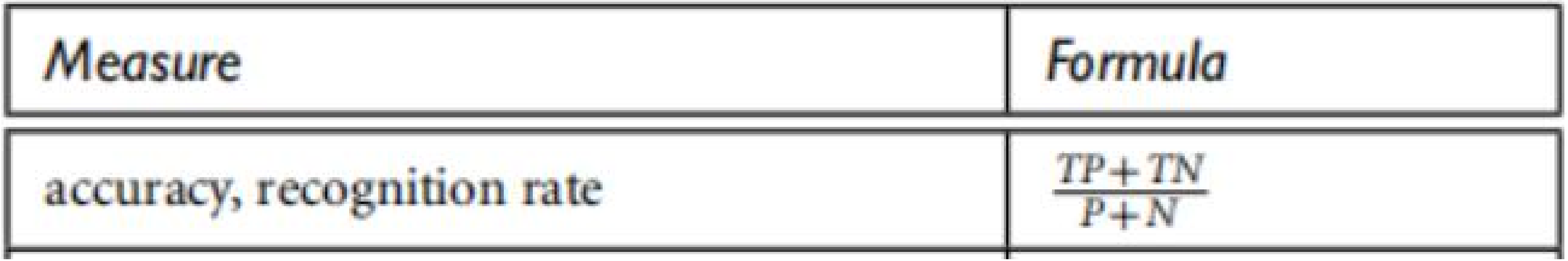
Actual

Predict

	Diagnosed COVID 19 (1)	Diagnosed Healthy (0)
COVID 19 (1)	TP ✓	FP ✗
Healthy (0)	FN ✗	TN ✓

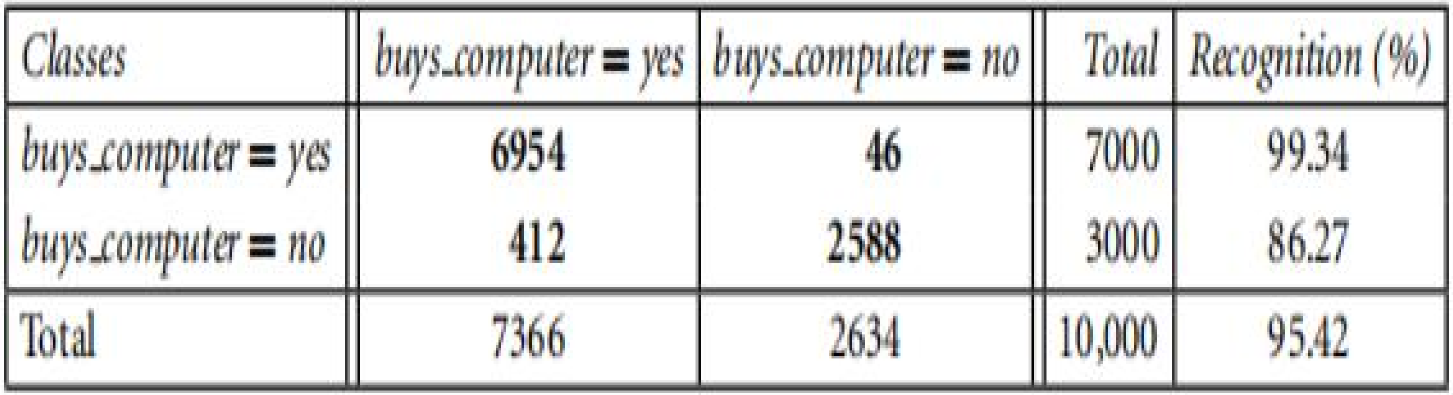
Healthy predicted as sick

Sick predicted as healthy



Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$

Evaluation Measure-Accuracy



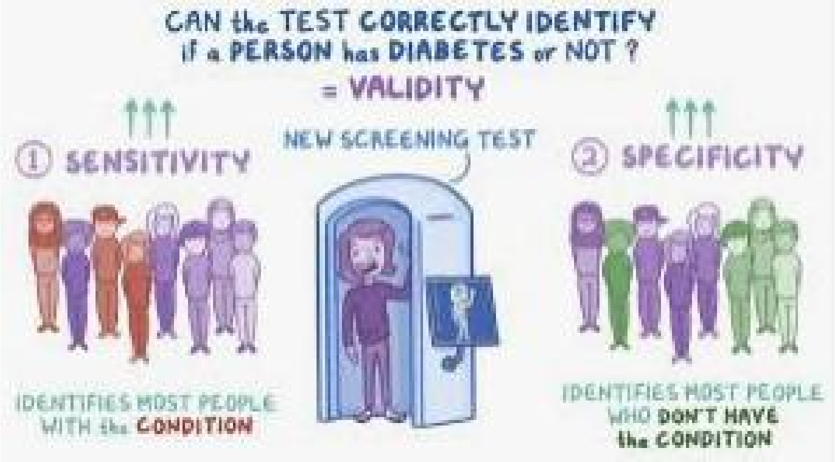
Classes	*buys_computer = yes*	*buys_computer = no*	*Total*	*Recognition (%)*
buys_computer = yes	6954	46	7000	99.34
buys_computer = no	412	2588	3000	86.27
Total	7366	2634	10,000	95.42

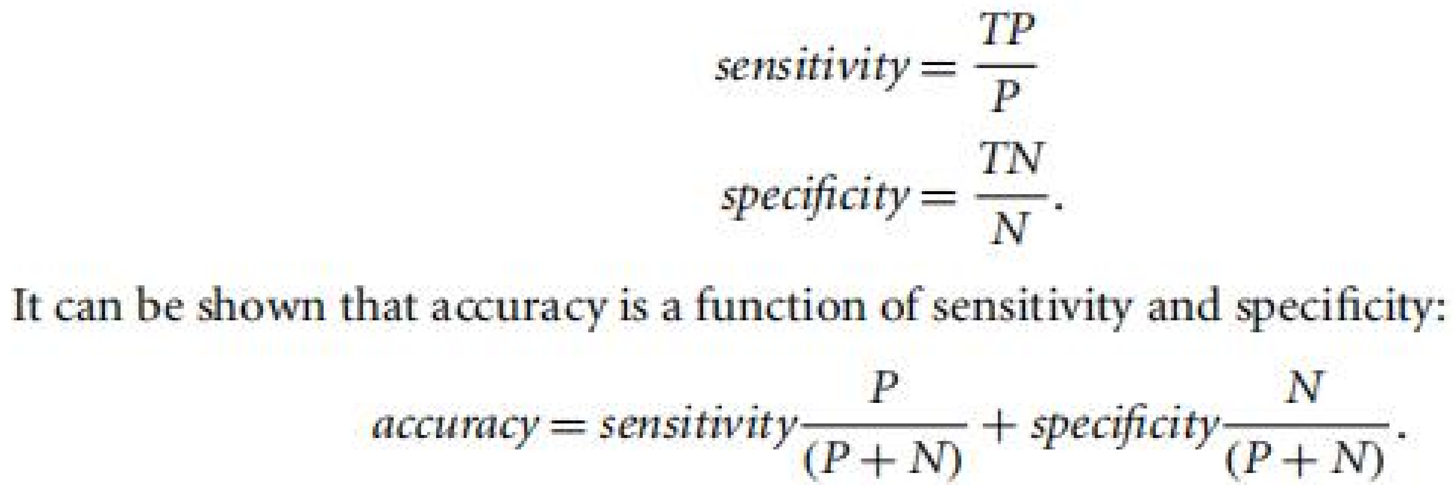
In the pattern recognition literature, this is also referred to as the overall recognition rate of the classifier, that is, it reflects how well the classifier recognizes tuples of the various classes.

Accuracy is most effective when the class distribution is relatively balanced.

sensitivity and specificity

- Sensitivity is also referred to as the true positive (recognition) rate (i.e., the proportion of positive tuples that are correctly identified), while specificity is the true negative rate (i.e., the proportion of negative tuples that are correctly identified). These measures are defined as


CAN the TEST CORRECTLY IDENTIFY if a PERSON has DIABETES or NOT ?
= VALIDITY
① SENSITIVITY
NEW SCREENING TEST
② SPECIFICITY
IDENTIFIES MOST PEOPLE WITH the CONDITION
IDENTIFIES MOST PEOPLE WHO DON'T HAVE the CONDITION



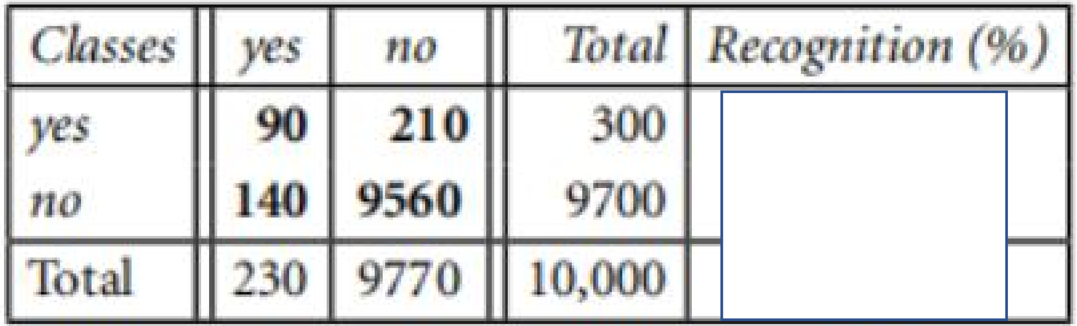
$$sensitivity = \frac{TP}{P}$$

$$specificity = \frac{TN}{N}.$$

It can be shown that accuracy is a function of sensitivity and specificity:

$$accuracy = sensitivity \frac{P}{(P+N)} + specificity \frac{N}{(P+N)}.$$

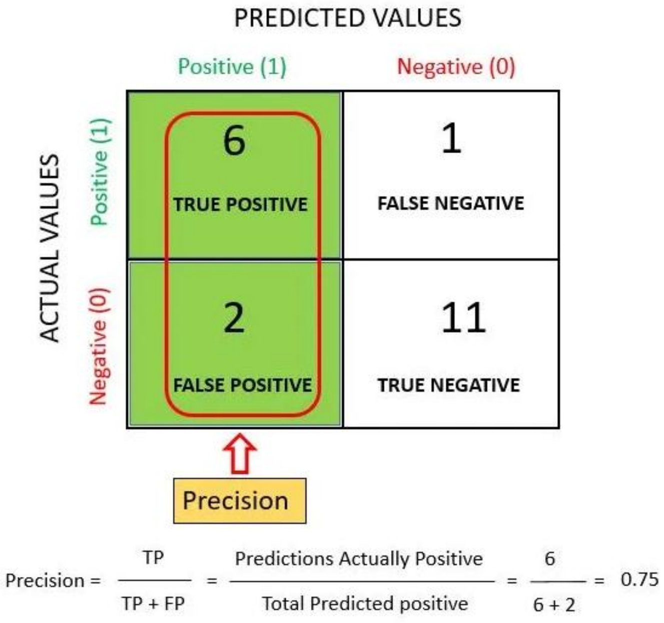
Example



Classes	yes	no	Total	Recognition (%)
yes	90	210	300	
no	140	9560	9700	
Total	230	9770	10,000	

the classifier is 90 /300 = 30.00%. The specificity is 9560 /9700 = 98.56%.
The classifier’s over_x0002_all accuracy is 9650 /10,000 = 96.50%.
Thus, we note that although the classifier has a high accuracy, it’s ability to correctly label the positive (rare) class is poor given its low sen_x0002_sitivity. It has high specificity, meaning that it can accurately recognize negative tuples

Precision



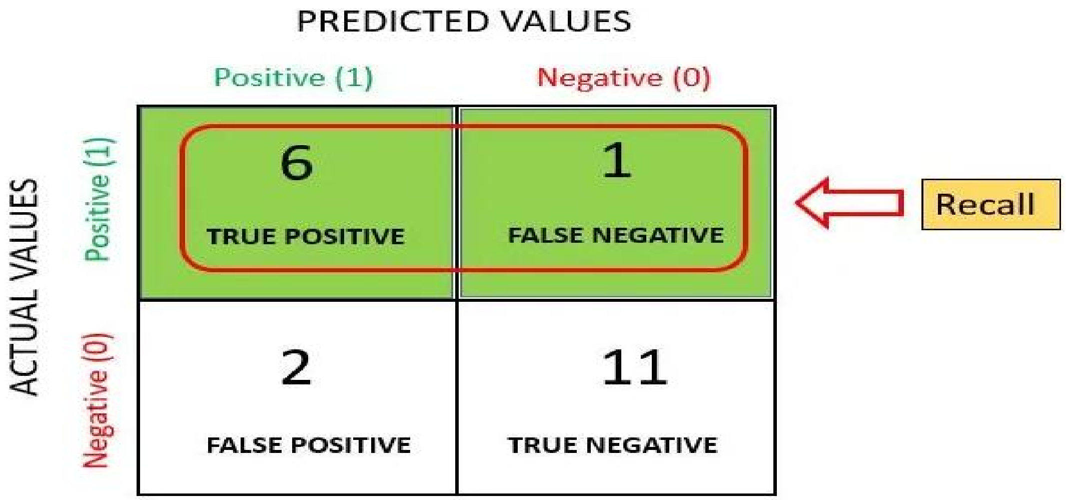
ACTUAL VALUES	PREDICTED VALUES: Positive (1)	PREDICTED VALUES: Negative (0)
Positive (1)	6 TRUE POSITIVE	1 FALSE NEGATIVE
Negative (0)	2 FALSE POSITIVE	11 TRUE NEGATIVE

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} = \frac{\text{Predictions Actually Positive}}{\text{Total Predicted positive}} = \frac{6}{6+2} = 0.75$$

Precision can be thought of as a measure of exactness (i.e., what percentage of tuples labeled as positive are actually such

Recall

*“**Recall** is a useful metric in cases where **False Negative** trumps **False Positive**”*



		PREDICTED VALUES	
		Positive (1)	Negative (0)
ACTUAL VALUES	Positive (1)	6 TRUE POSITIVE	1 FALSE NEGATIVE
	Negative (0)	2 FALSE POSITIVE	11 TRUE NEGATIVE

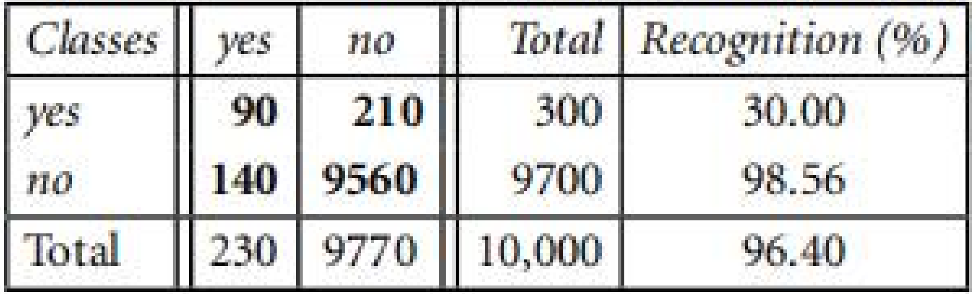
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Predictions Actually Positive}}{\text{Total Actual positive}} = \frac{6}{6+1} = 0.85$$

Precision and recall. The precision of the classifier in the Figure for the yes class is

90/ 230 = 39.13%.

The recall is 90/300 = 30.00%, which is the same calculation for sensitivity

in Example



Classes	*yes*	*no*	*Total*	*Recognition (%)*
yes	**90**	**210**	300	30.00
no	**140**	**9560**	9700	98.56
Total	230	9770	10,000	96.40

High recall but low precision implies that most ground-truth objects have been detected, but most detections are incorrect (many false positives).
High precision but low recall implies that most the predicted boxes are correct, but most ground-truth objects have not been detected (many false negatives).
High precision and high recall implies an ideal detector that has detected all ground-truth objects correctly.
Low precision and low recall implies a poor detector that does not detect most ground-truth objects (many false negatives), and most detections are incorrect (many false positives).