F-Measures

Prepared By: Dr.Mydhili K Nair, Professor, ISE Dept, RIT
For: Machine Learning Class
Target Audience: Sem 6 Students

Performance Measures

Classification:

- Simple Accuracy
- Precision
- Recall
- F-beta measure
- ROC (and AUC)

Regression:

- Sum of Squares Error
- RMS Error
- Mean Absolute Error

Accuracy as a Performance Measure

- What is 95% accuracy?
    - Classification: 95 / 100 shoes correctly classified
    - Regression:Predict 95/100 house prices correctly

$600,000

$400,000 ✗

$599,999 ✗

Limitations of Simple Accuracy

$$Accuracy = \frac{No.\,Samples\ Predicted\ Correctly}{Total\ No.\,of\ Samples}$$

What is wrong with this ?

9,990 Non-Nike

10 Nike

```
def classifier(shoe):
    return False
```

$$Accuracy = \frac{9{,}990}{10{,}000} = \mathbf{99.9\%}$$

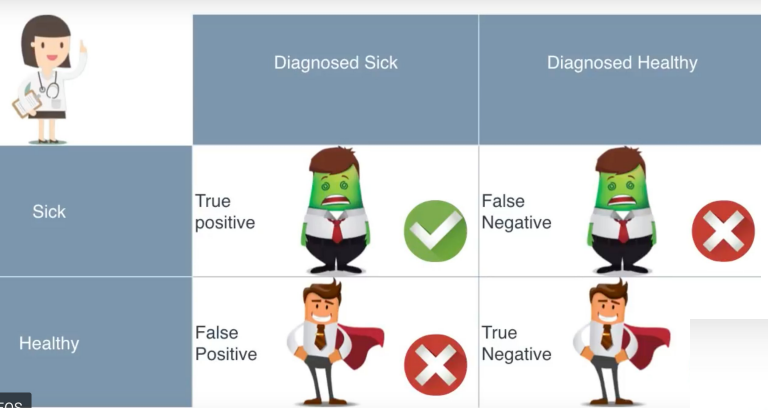
Limitation with Accuracy

Is this tumor cancerous?

very few positive examples

most are negative examples

Class Imbalance Problem



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

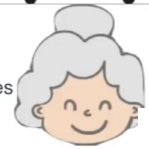
Confusion Matrix

10,000 Patients

(Actual)

(Predicted) 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 

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

1,000
e-mails

Confusion Matrix

(Predicted) Folder

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

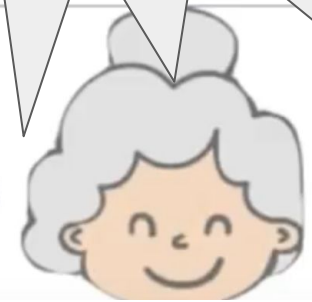
(Actual)

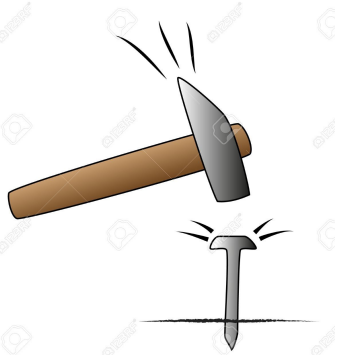
	Diagnosed Sick	Diagnosed Healthy
Sick		False Negative  
Healthy	False Positive  	

Error Rate is very high.
(1-Accuracy Rate) i.e. Off-diagonal values

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

Error Rate is very high.
(1-Accuracy Rate)
i.e. Off-diagonal values





- Simple Accuracy is excellent when we have a Balanced Data Set
- It fails when the Dataset is "Imbalanced".

Precision and Recall as Performance Measure

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

EVALUATION METRICS

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

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

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?

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

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

Recall

E-mail	Folder: Spam Folder	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?

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

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

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?

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

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

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?

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

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

Precision and Recall

Medical Model

Precision: 55.7%

Recall: 83.3%

Spam Detector

Precision: 76.9%

Recall: 37%

Credit Card Fraud

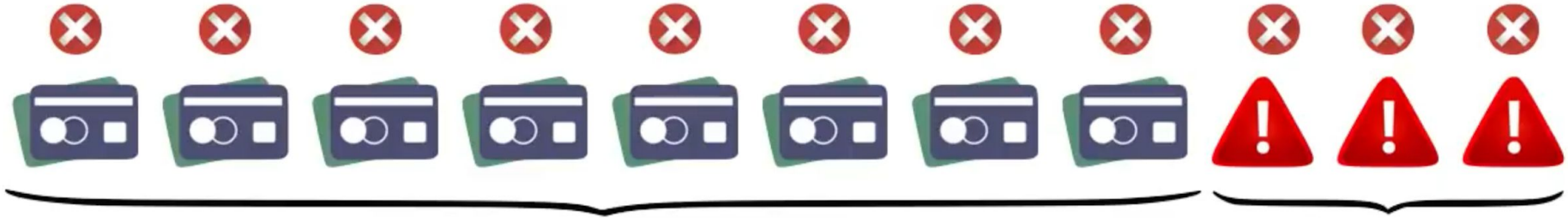
Model: All transactions are good.

Precision = 100%

$$\text{Recall} = \frac{0}{472} = 0\%$$

Average = 50%

Credit Card Fraud



284,335 472

Model: All transactions are fraudulent.

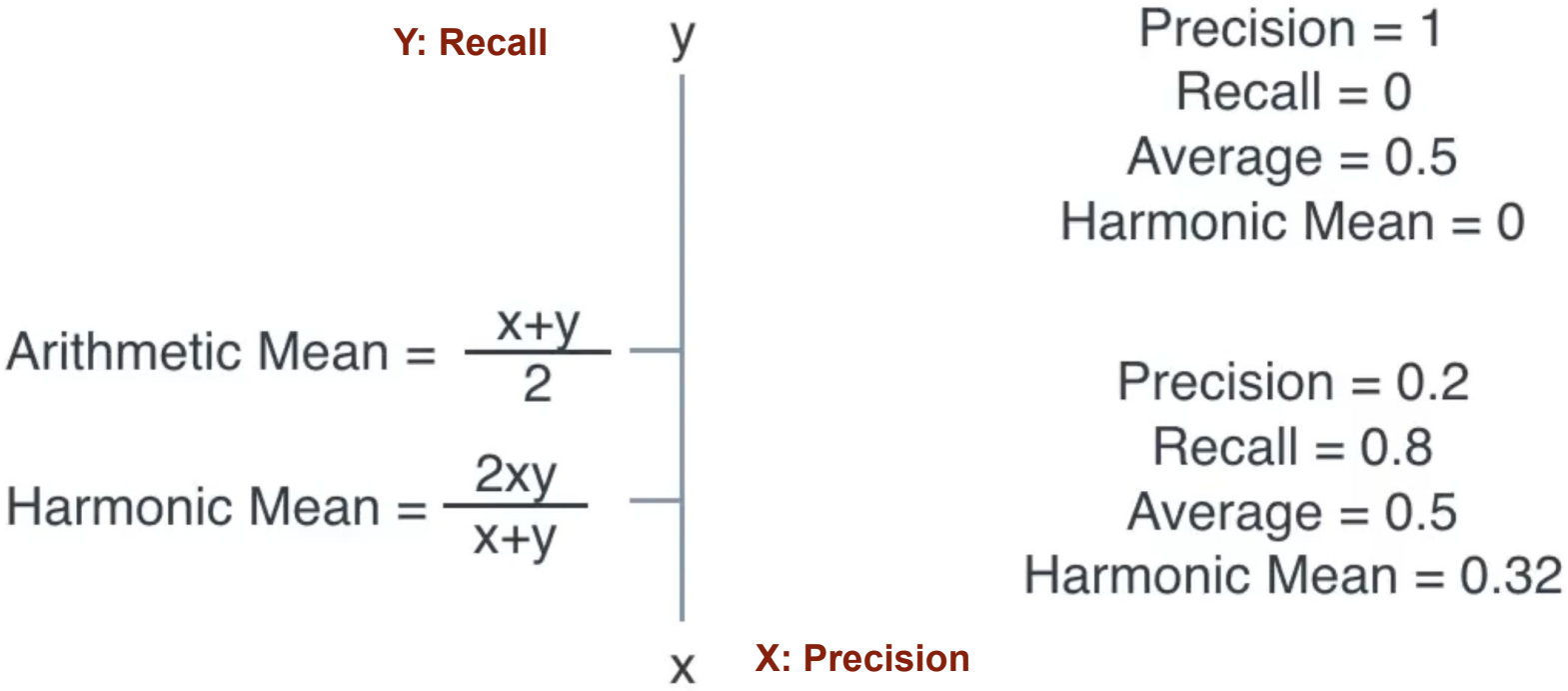
$$\text{Precision} = \frac{472}{284{,}807} = .016\% \qquad \text{Recall} = \frac{472}{472} = 100\%$$

Average = 50.008%

F-Measures as Performance Measure

- Used on imbalanced datasets
- Harmonic Mean of Precision & Recall
- Used because simple mean fails

Harmonic mean

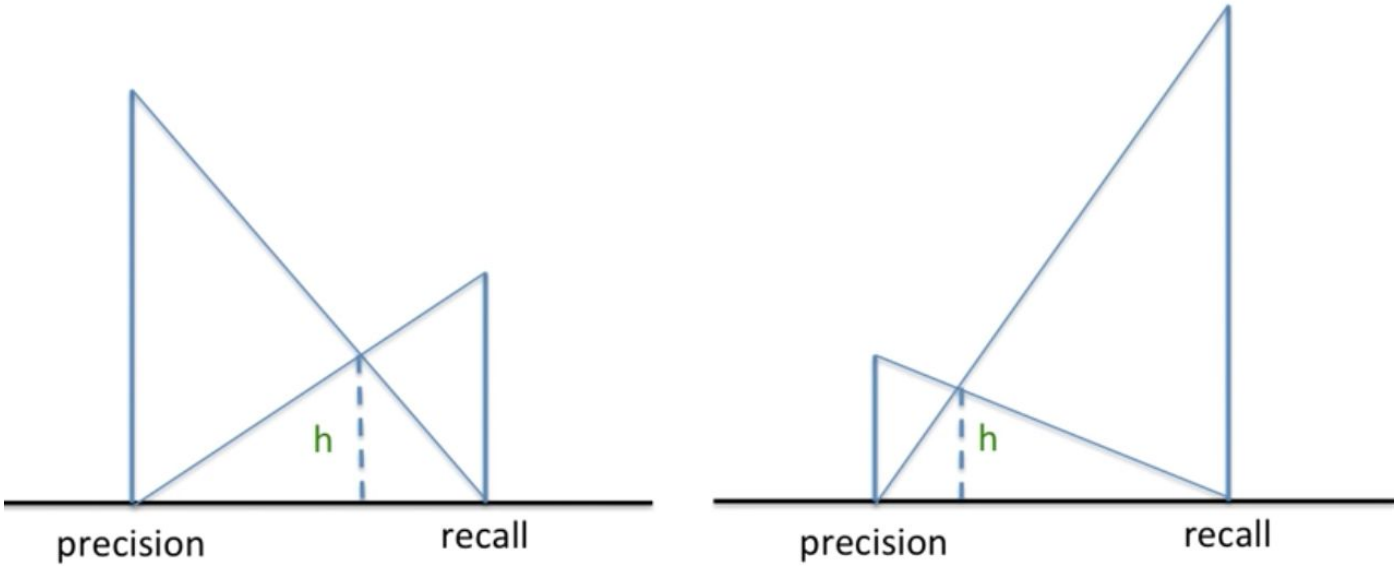


~~Arithmetic Mean(Precision, Recall)~~

F1 Score = Harmonic Mean(Precision, Recall)

Harmonic Mean punishes extreme value more

(Basis of F1 Score)



h is half the harmonic mean

$$\text{F1 Score} = 2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

F1 Score

Medical Model

Precision = 55.7%

Recall = 83.3%

Average = 69.5%

$$\text{F1 Score} = \frac{2 \times 55.7 \times 83.3}{55.7 + 83.3} = 66.76\%$$

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score

Spam Detector Model

Precision = 76.9%

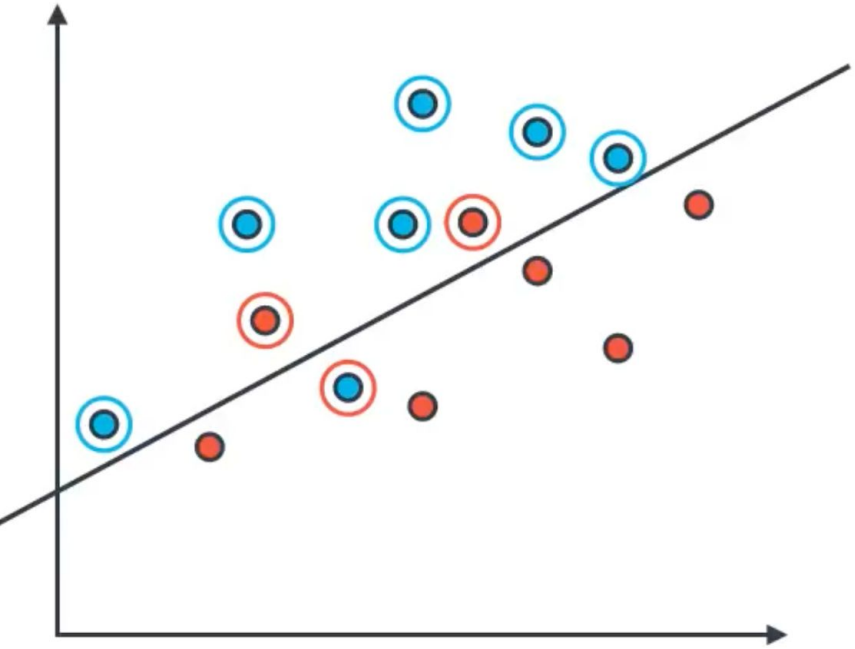
Recall = 37%

Average = 56.95%

$$F1\ Score = \frac{2 \times 76.9 \times 37}{76.9 + 37} = 49.96\%$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 Score



Precision = 75%

Recall = 85.7%

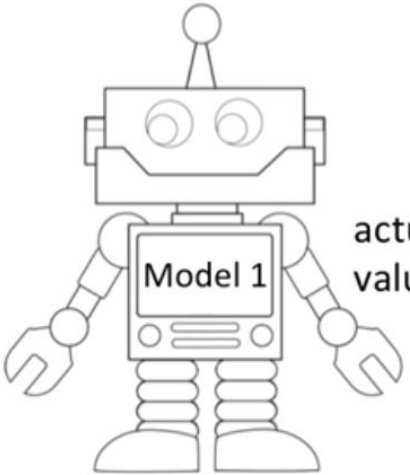
Average = 80.35

$$\text{F1 Score} = \frac{2 \times 75 \times 85.7}{75 + 85.7} = 80\%$$

F1 Score - Best Use Case

Multiclass Classification on imbalanced dataset

Precision of Model 1 (macro average)



actual values

predictions →

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9
	TP: 100 FP: 0	TP: 9 FP: 82	TP: 8 FP: 10	TP: 9 FP: 12

Precision = TP / (TP + FP) P(A) = 1 P(B) = 9/91 P(C) = 8/18 P(D) = 9/21

average precision = P(A) + P(B) + P(C) + P(D) / 4 = 0.492

the number of classes

Recall of Model 1 (macro average)

predictions →

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

TP: 100, FN: 100 R(A) = 100 / 200

TP: 9, FN: 1 R(B) = 9/10

TP: 8, FN: 2 R(C) = 8/10

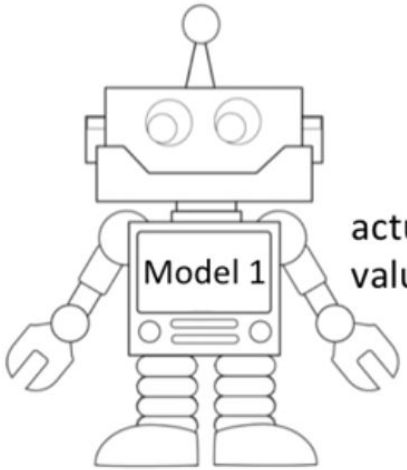
TP: 9, FN: 1 R(D) = 9/10

Recall = TP / (TP + FN)

average recall = R(A) + R(B) + R(C) + R(D) / 4 = 0.775

the number of classes

F1 Score of Model 1



actual values

predictions

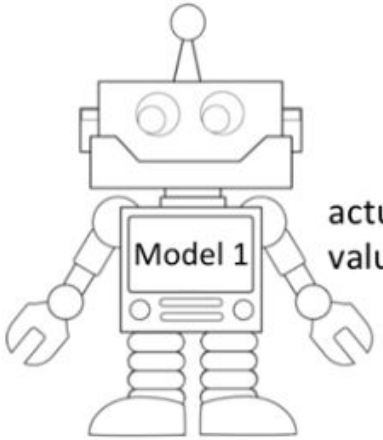
	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

$$\text{F1 Score} = 2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

$$2 \times \frac{0.492 \times 0.775}{0.492 + 0.775}$$

0.601

F1 Score on imbalanced data



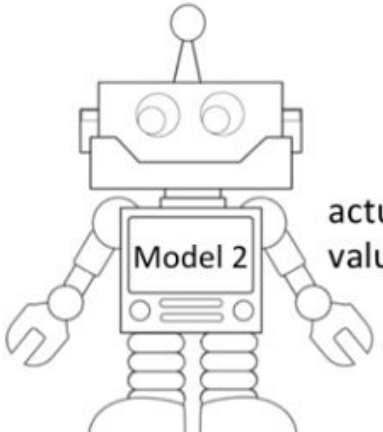
actual values

predictions →

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

F1 Score = **0.601**

accuracy = 0.547



actual values

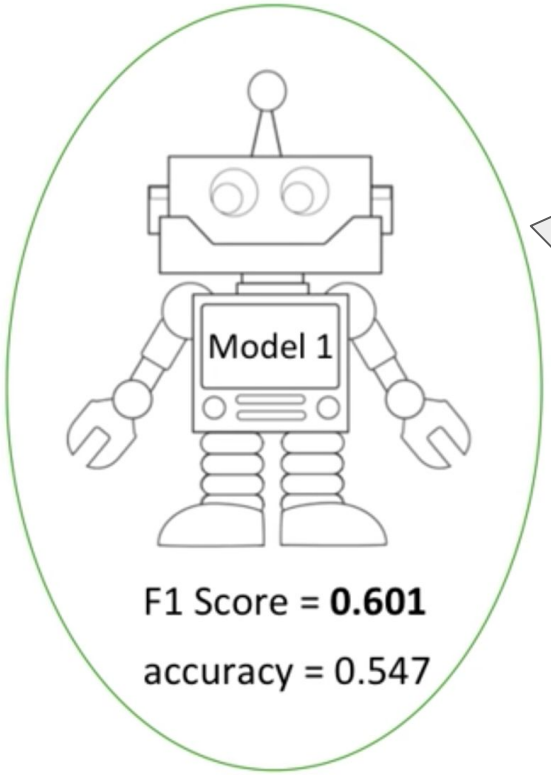
predictions →

	A	B	C	D
A	198	2	0	0
B	7	1	0	2
C	0	8	1	1
D	2	3	4	1

F1 Score = **0.342**

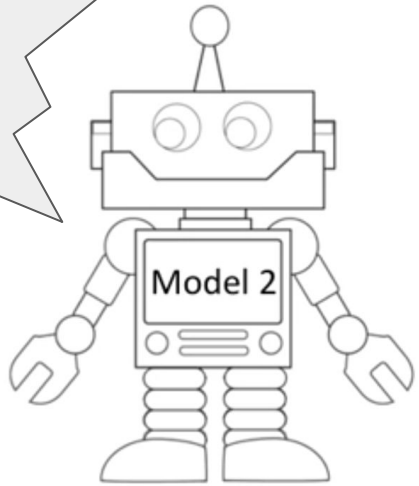
accuracy = 0.87

F1 Score on imbalanced data



Model 1 could be KNN

Which to choose ?



Model 2 could be SVM

Model 1 predicts well on multiple class classification on imbalanced given data, and F1 score is the metric to quantify its performance.

Sources:

1. **Nike Shoes, Performance Measures, Comparing Systems:** https://youtu.be/j-EB6RqqjGI
2. **Grandma Cookie-Cancer Diagnosis:** https://youtu.be/aDW44NPhNw0
3. **Quiz - Prison, Cancer:** https://youtu.be/Clo-t9eeEwg
4. **Robot**: https://youtu.be/HBi-P5j0Kec
5. **Evaluation of a Classification Model:** https://youtu.be/85dtiMz9tSo

Beyond Syllabus
(F_β Score)

F_β Score

Precision **$F_{0.5}$ Score** **F_1 Score** **F_2 Score** **Recall**

Comparing Systems

System 1
- Precision: 70%
- Recall: 60%

?

System 2
- Precision: 80%
- Recall: 50%

$$F_\beta = \frac{1}{\beta \times \frac{1}{Precision} + (1-\beta) \times \frac{1}{Recall}}$$

- Greater β, Greater importance to Precision

Comparing Systems

System 1
- Precision: 70%
- Recall: 60%

?

System 2
- Precision: 80%
- Recall: 50%

$$F_\beta = \frac{1}{\beta \times \frac{1}{Precision} + (1-\beta) \times \frac{1}{Recall}}$$

$\beta = 0.95$

0.6942 < 0.7766

F-Measure

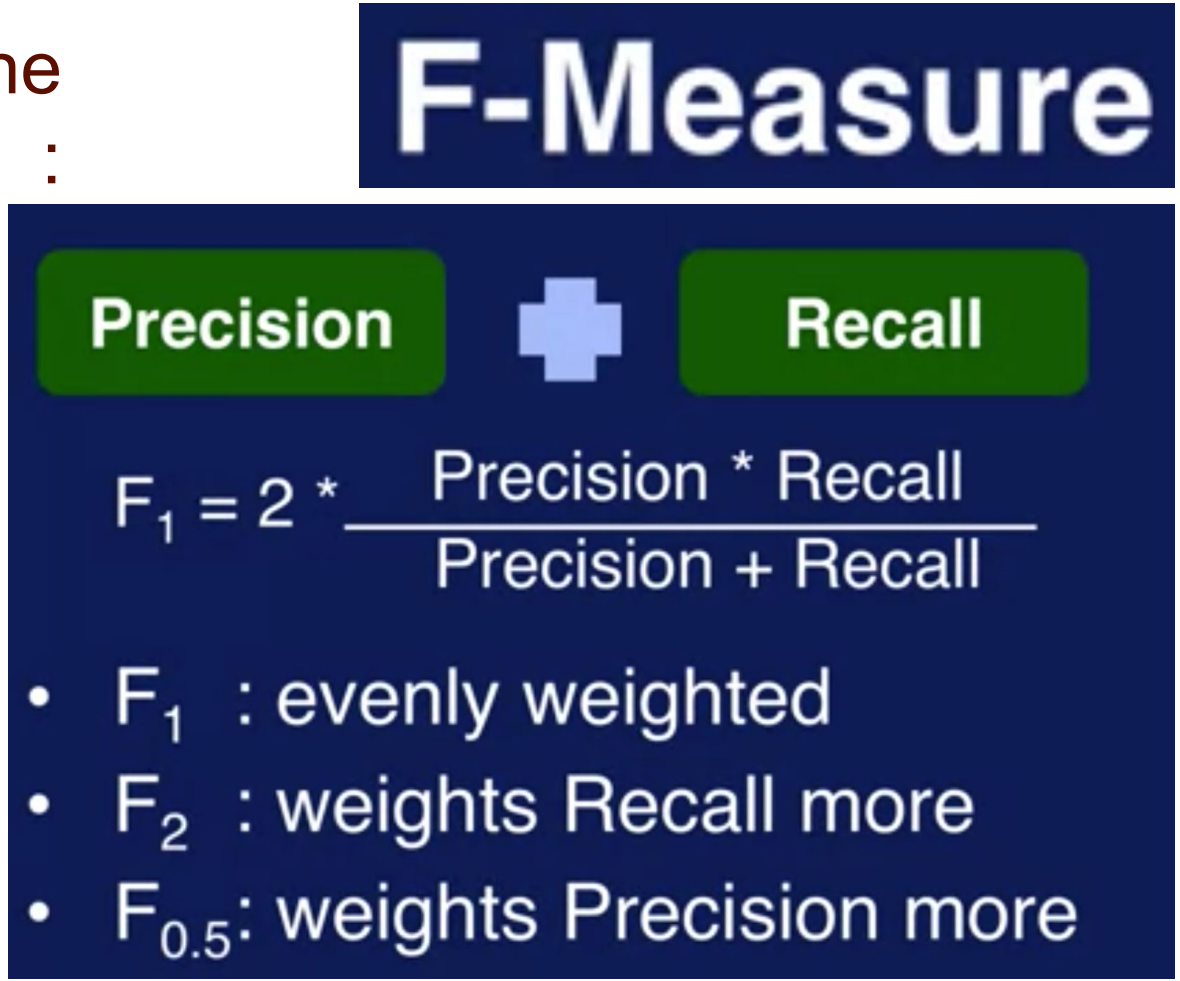
$\beta = 0.5$

$$F_\beta = \frac{1}{0.5 \times \frac{1}{0.7} + (1-0.5) \times \frac{1}{0.6}} = 0.6461$$

$\beta = 0.5$

$$F_\beta = \frac{1}{0.5 \times \frac{1}{0.8} + (1-0.5) \times \frac{1}{0.5}} = 0.6153$$

Measuring Machine Learning Models : F1 Score



QUIZ

In each of the following scenarios which choice of F_1, $F_{0.5}$ or F_2 be the best choice of metric.

#1: *FPR* must be reduced - *Precision* must be high
$F_ß$ where ß must be high. **So F_2**

1. Cancer Detection : If someone is falsely diagnose we may do some extra tests. If someone who actually has cancer is not diagnosed they may die.

2. Convicting to Prison : People are innocent until proven guilty by US Law. We want to avoid false convictions. But we also want criminals to not run free.

#2: *FNR* must be reduced - *Recall* must be high
$F_ß$ where ß must be low. **So $F_{0.5}$**