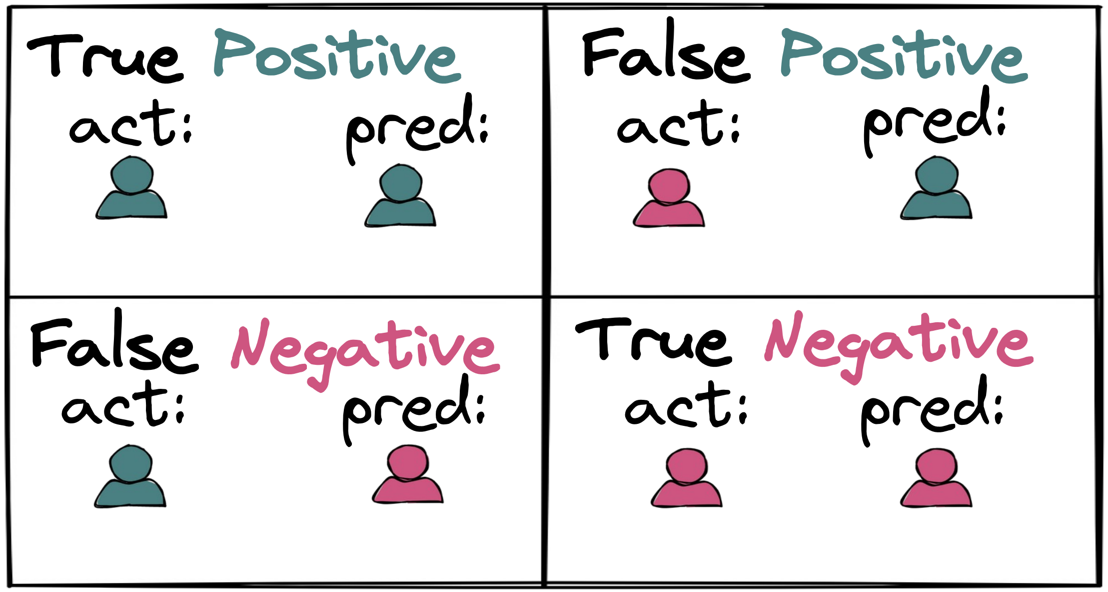
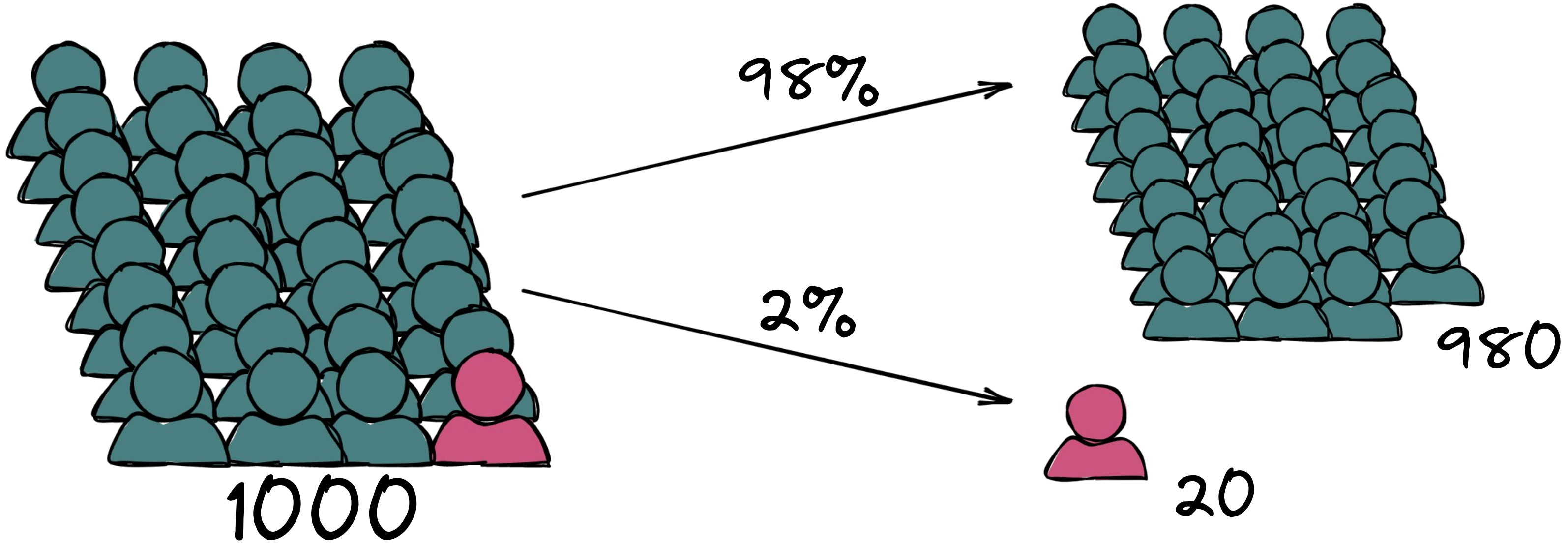
# Classification metrics




Week 14

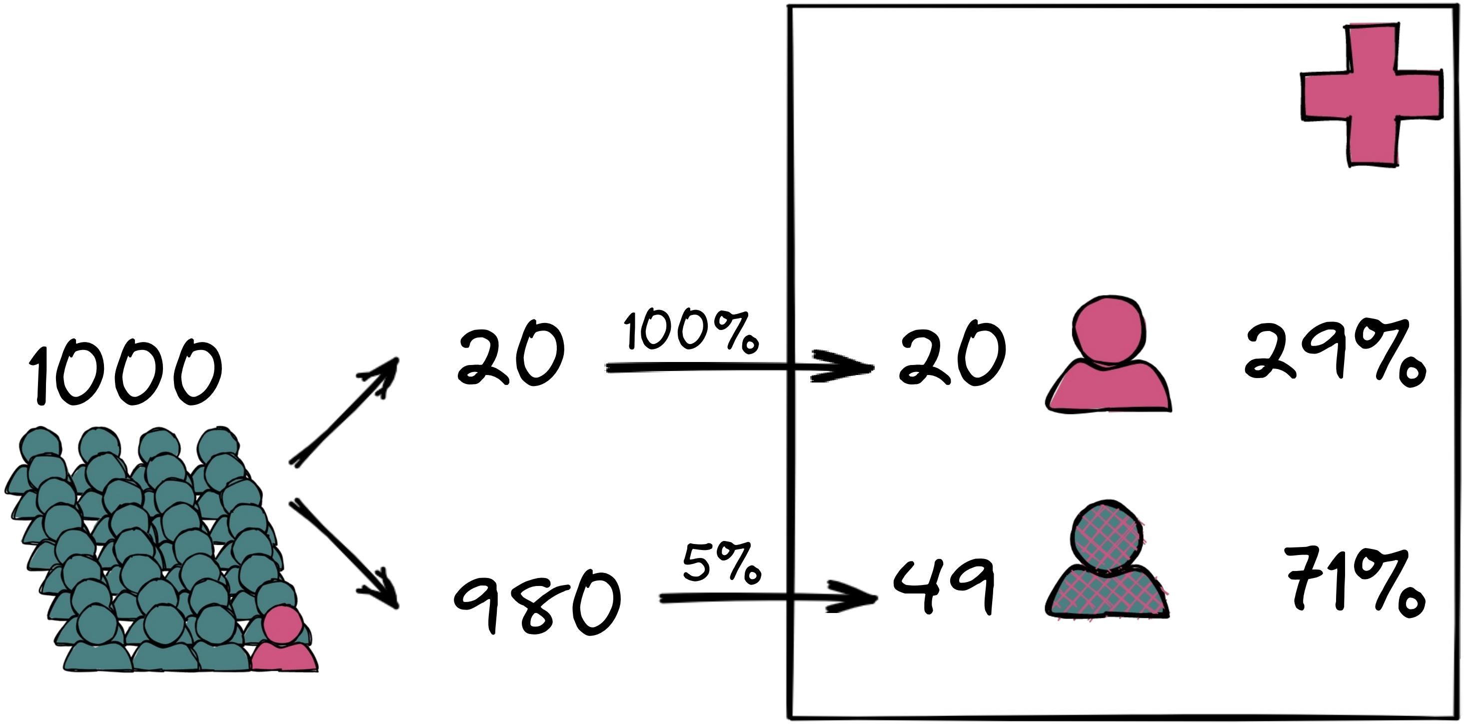
Middlesex University Dubai: Fall 2021;
CST4050; Instructor: Ivan Reznikov

# Classification metrics: Case1




Disease Test: 100% correct if ill
95% correct if not ill

# Classification metrics: Case1




71% of people in the hospital aren't ill.
The error is much higher than 5%

# Confusion matrix

| | actual class: True | actual class: False |
|---|---|---|
| predicted: True | True Positive (TP)<br>actual:<br>predicted: | Type I error<br>False Positive (FP)<br>actual:<br>predicted: |
| predicted: False | Type II error<br>False Negative (FN)<br>actual:<br>predicted: | True Negative (TN)<br>actual:<br>predicted: |

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

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN}$$

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

$$\text{NPV}^{*} = \frac{TP}{TP+FN}$$

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

*NPV - negative predicted value

# Confusion matrix: metrics

Precision relates to the amount of false positive results (type I errors): high precision means small amount of type I errors.

Recall relates to the amount of false negative results (type II errors): high recall means small amount of type II errors

One may use the following observation to remember:
<u>**p**</u>recision – <u>**p**</u>ositive, r<u>**e**</u>call – n<u>**e**</u>gative

F1-score is a metric, taking both precision and recall into account. One of the main reasons, why we multiple precision and recall instead of taking, say, their average, is to deal with extremely low values.

# Confusion matrix: Case1

| | actual class | |
|---|---|---|
| predicted class | True Positive (TP)<br>act: pred:<br>931 | False Positive (FP)<br>act: pred:<br>0 |
| | False Negative (FN)<br>act: pred:<br>49 | True Negative (TN)<br>act: pred:<br>20 |

$$\text{Precision} = \frac{931}{931+0} = 1$$

$$\text{Sensitivity (Recall)} = \frac{931}{931+49} = 0.95$$

$$\text{Specificity} = \frac{20}{20+0} = 1$$

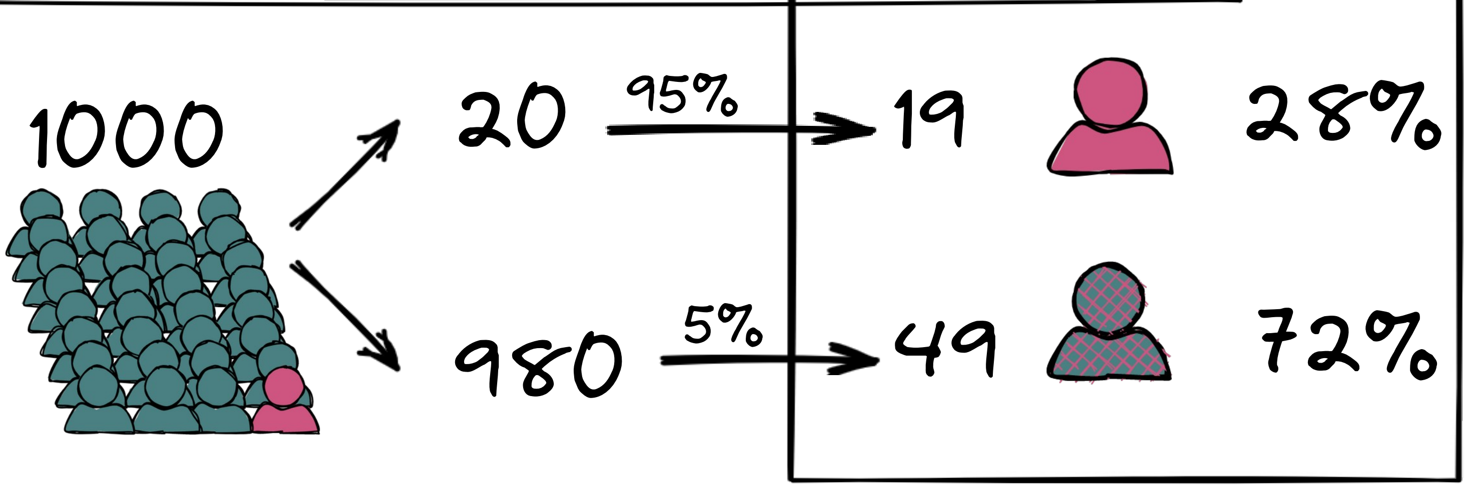
$$\text{NPV} = \frac{20}{20+49} = 0.29$$

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

$$= 0.97$$

# Classification metrics: Case2

Disease Test2: 95% correct if ill
95% correct if not ill




72% of people in the hospital aren't ill

# Confusion matrix: Case2

Disease Test2: 95% correct if ill
95% correct if not ill

The mentioned TN, TP, FN, FP are easily calculated using <u>Bayes' Theorem</u>:

**TP:**

P (ill if test says ill) = P (B) * P (A | B) = 0.98*0.95 = 0.931 = 93.1%

**FN:**

P (ill if test says not ill) = 0.98*0.05 = 0.049 = 4.9%

**TN:**

P (not ill if test says not ill) = 0.02*0.95 = 0.019 = 1.9%

**FP:**

P (not ill if test says ill) = 0.02*0.05 = 0.001 = 0.1%

# Confusion matrix: Case2

actual class

predicted class

| | |
|---|---|
| True Positive (TP)<br>act: pred:<br>931 | False Positive (FP)<br>act: pred:<br>1 |
| False Negative (FN)<br>act: pred:<br>49 | True Negative (TN)<br>act: pred:<br>19 |

$$\text{Precision} = \frac{931}{931+1} = 0.99$$

$$\text{Sensitivity (Recall)} = \frac{931}{931+49} = 0.95$$

$$\text{Specificity} = \frac{19}{19+1} = 0.95$$

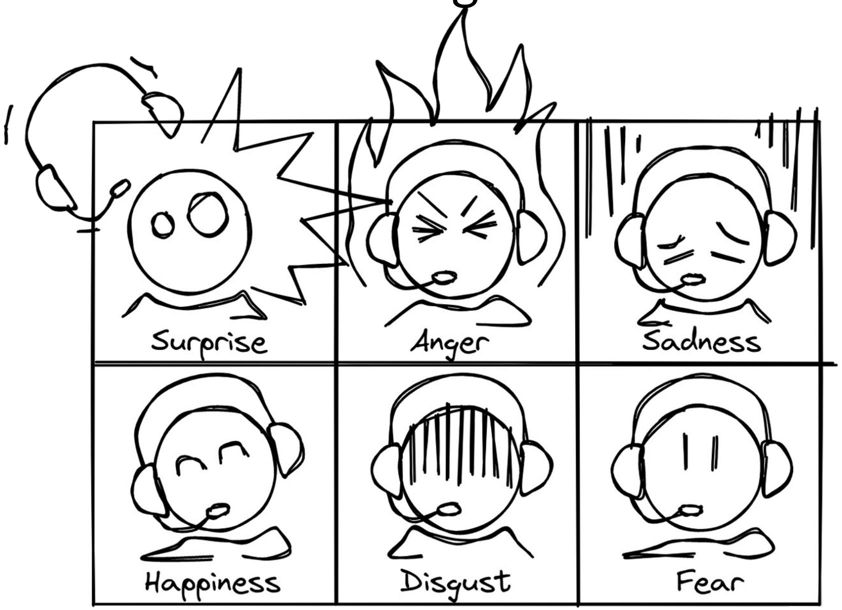
$$\text{NPV} = \frac{19}{19+49} = 0.28$$

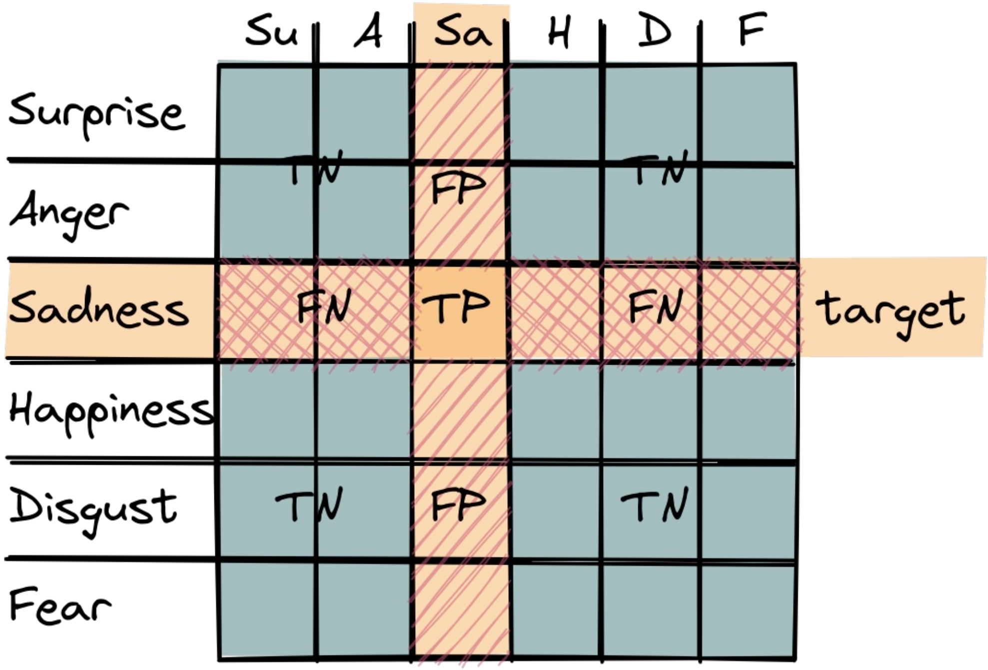
$$F1 = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$= 0.97$

# Confusion matrix: Case3

Imagine you're building a system to detect emotions via audio.
Among six classes of emotions, you need to detect sadness. The confusion matrix won't change much:




TN – True Negative, TP – True Positive
FN – False Negative, FP – False Positive

# Confusion matrix: errors

$H_0$ – patient doesn't have COVID-19.

**Type I error** (first kind error; *<u>false positive</u>*) – your prediction/observation is *<u>positive</u>*, and it's *<u>false</u>*. Type I error involves rejecting a true null hypothesis – the patient has no COVID-19, but a PCR test falsely returned "positive." Another example is a fire alarm going on, but there is no fire ("false alarm").

**Type II error** (second kind error; *<u>false negative</u>*) – your prediction/observation is *<u>negative</u>*, and it's *<u>false</u>*. Type II error deals with failing to reject a false null hypothesis – the patient has COVID-19, but the PCR test fails to detect it. Another example: fire alarm, silent, whereas fire, is all around the building.

Usually, <u>type II errors are far worse than type I errors</u>, as the consequences of type II errors – not detecting the disease or fire is more harmful than setting up a false alarm.

# Confusion matrix: metrics

Of course, besides the discussed metrics we can use basic accuracy:

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

or ROC-AUC: The Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) probability curve. The ROC curve displays represent all confusion matrixes for a given model. The closer the AUC to 1 – the better the model.

We'll cover ROC-AUC deeper in one of the following lectures.