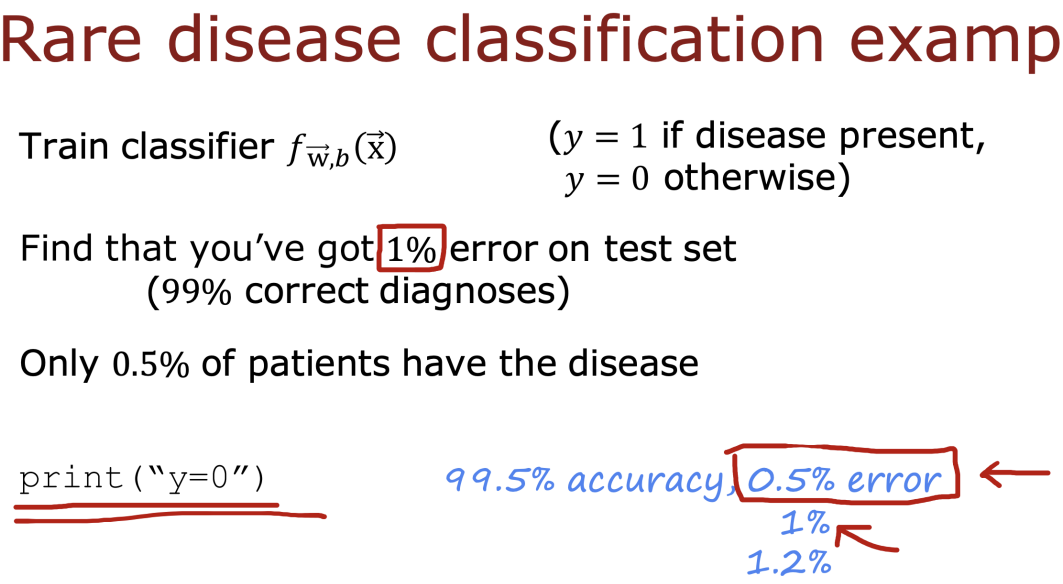
Skewed datasets

In machine learning applications where the ratio of positive to negative examples is very skewed, traditional error metrics like accuracy may not provide a true picture of the model's performance. For instance, in a binary classification problem where the positive class is rare (like diagnosing a rare disease), a model that always predicts the negative class can have a high accuracy but is not useful.

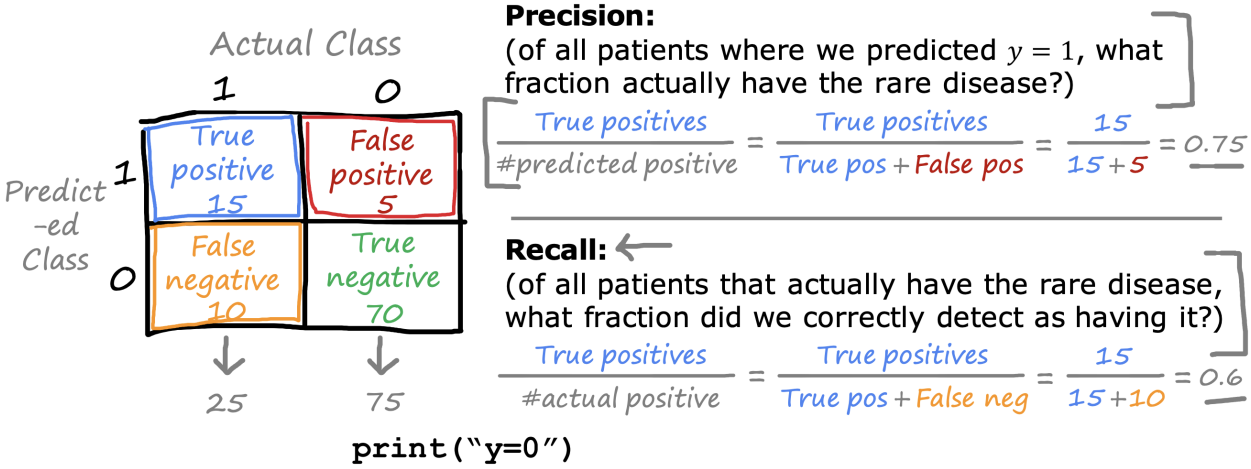


To better evaluate the performance of a model in such scenarios, we use precision and recall.

A **confusion matrix**, a 2×2 table, is used to calculate precision and recall.

Precision/recall

$y = 1$ in presence of rare class we want to detect.




```
print("y=0")
```

These metrics help to identify if a model is just predicting the majority class all the time. A model with either zero precision or zero recall is not useful. By ensuring both precision and recall are reasonably high, we can be more confident that the model is making useful predictions.

> 💡 Precision is the proportion of true positive predictions (correctly identified positive cases) among all positive predictions, while recall is the proportion of true positive predictions among all actual positive cases.

Trading off precision and recall

Logistic regression: $0 < f_{\vec{w},b}(\vec{x}) < 1$

Predict 1 if $f_{\vec{w},b}(\vec{x}) \geq 0.5$ ~~0.7~~ ~~0.9~~ 0.3

Predict 0 if $f_{\vec{w},b}(\vec{x}) < 0.5$ ~~0.7~~ ~~0.9~~ 0.3

$$\text{precision} = \frac{\text{true positives}}{\text{total predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{total actual positive}}$$

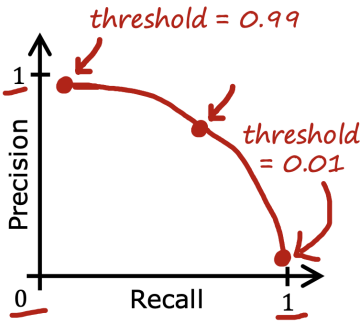
Suppose we want to predict $y = 1$ (rare disease) only if very confident.

→ higher precision, lower recall.

Suppose we want to avoid missing too many cases of rare disease (when in doubt predict $y = 1$)

→ lower precision, higher recall.

More generally predict 1 if: $f_{\vec{w},b}(\vec{x}) \geq$ threshold.



For example, if we want to be very sure before diagnosing a patient with a disease (to avoid unnecessary treatment), we might set a high threshold for diagnosis, leading to high precision but low recall. On the other hand, if we want to avoid missing any cases of the disease, we might set a low threshold for diagnosis, leading to high recall but low precision.

The **F1 score** is a metric that combines precision and recall into a single number, giving more weight to the lower of the two. This can be useful when comparing different algorithms or settings, as it provides a single number to optimize.

F1 score

How to compare precision/recall numbers?

	Precision (P)	Recall (R)	~~Average~~	F_1 score
Algorithm 1	0.5	0.4	~~0.45~~	0.444
Algorithm 2	0.7	0.1	~~0.4~~	0.175
Algorithm 3	0.02	1.0	~~0.501~~	0.0392

print("y=1")

~~Average = $\frac{P+R}{2}$~~

$$F1\ score = \frac{1}{\frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)} = 2\frac{PR}{P+R}$$

However, the best choice of threshold and the relative importance of precision and recall will depend on the specific context and the costs associated with false positives and false negatives.