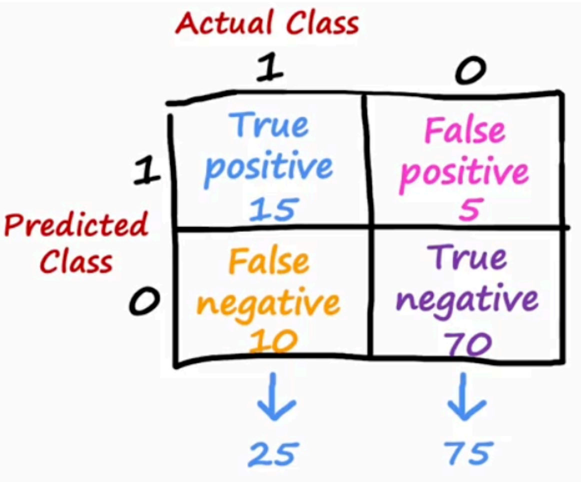
# Skewed datasets

## Error metrics for skewed datasets

- Traditional accuracy metrics can be misleading when dealing with skewed datasets, as they may not reflect the true performance of the model.
- An example illustrates that a simple algorithm predicting all negatives can achieve high accuracy, but is not useful for diagnosis.Here's a paraphrased version of your sentence:

> A dataset is considered **skewed** when there is a significant imbalance between the number of positive and negative examples, meaning one class heavily outnumbers the other.

### Precision/recall




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

**Precision:**(of all patients where we predicted $y = 1$, what fraction actually have the rare disease?)

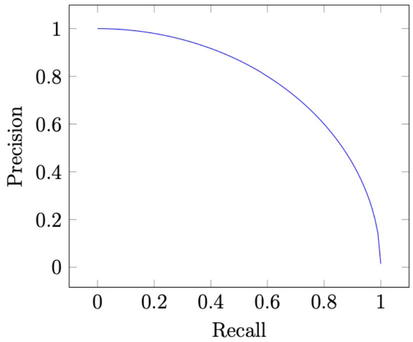
$\frac{TruePositives}{PredictedPositive} = \frac{TruePositives}{TruePos+FalsePos} = \frac{15}{15+5} = 0.75$

**Recall:** (of all patients that actually have the rare disease, what fraction did we correctly detect as having it?)

$\frac{TruePositives}{ActualPositive} = \frac{TruePositives}{TruePos+FalseNeg} = \frac{15}{15+10} = 0.6$

## Trading off precision and recall

The trade-off between Precision and Recall follows this curve:




When selecting a model, it's important to consider both precision $P$ and recall $R$. To simplify this process, we introduce the $F_1 - score$, a metric that combines $P$ and $R$, defined as:

$$F_1 = \frac{1}{\frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)} = \frac{2PR}{P + R}.$$

With this measure, we can select the model that achieves the highest $F_1 - score$.