# Performance Metrics

IT461: Practical Machine Learning

# Supervised Learning



- Supervised learning requires labeled data.
- Supervised learning model is used to **predict target variables** based on features or input data.
- Based on predictions or output from supervised learning models, the predictive tasks are primarily divided into two categories based on the nature of the output variable: **classification** and **regression**.
- Classification tasks employ algorithms to categorize data into different classes or labels, while regression focuses on predicting continuous or numeric values based on input features.
- In supervised learning, the dataset is divided into training and testing sets. The training set teaches the model, while the testing set assesses its performance on unseen data.

# Model Performance



- The model's performance on the testing set provides ***an estimate of how well it will perform on unseen data***.
  - If the model performs well on the testing set, it indicates that it has learned the underlying patterns and can make accurate predictions.
  - If the model performs poorly on the testing set, it may be overfitting the training data or failing to capture the true relationships.
- **Overfitting** occurs when a model becomes too complex and memorizes the training data, resulting in poor performance on unseen data.
- **Underfitting** happens when a model is too simplistic and fails to capture data complexity.

# Evaluation metrics for regression

- Several metrics can assess the effectiveness of regression models, each providing a different perspective on the model's overall accuracy and fit, such as **R-squared ($R^2$)**, **MSE**, and **MAE**.
- Evaluating regression models requires carefully considering the metrics in the context of the specific application and the implications of prediction errors.
- Considering multiple metrics to get a comprehensive view of the model's performance is often helpful.

# R-squared ($R^2$)



- **R-squared**, also known as *the coefficient of determination*, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables.
- $R^2$ values range from 0 to 1, where a value of 1 indicates that the regression predictions perfectly fit the data.
- R-squared measures how effectively the model explains the dependent variable.
  - $R^2 = 0.9$ would indicate that 90% of the variance of the dependent variable being studied is explained by the variance of the independent variable.
  - The more variance can be explained, the better the model is.
- Although a higher $R^2$ might suggest a better fit, it doesn't always mean a better model, especially if the model is overfitting.

# Adjusted $R^2$



- $R^2$ is influenced by *the number of independent variables used*.
  - The more independent variables are included in the model, the greater the variance resolution $R^2$.
  - $R^2$ either increases or remains the same when new predictors are added to the model.
  - To resolve this, the adjusted $R^2$ is used.

- **Adjusted $R^2$** is a modified version of $R^2$ that has been adjusted for the number of predictors in the model.
- The adjusted R2 *decreases* when adding the extra predictor variables that don't improve the existing model.

# Mean Squared Error (MSE)



- Mean Squared Error (MSE) represents the average squared differences between observed and predicted values.
- MSE emphasizes more significant errors over smaller errors since it squares the residuals. This means that it is *sensitive to outliers*.
- A model with a lower MSE is generally considered better.

# Mean Absolute Error (MAE)



- Mean Absolute Error (MAE) calculates the average absolute differences between observed and predicted values.
- Unlike MSE, MAE treats all errors equally, so it is less sensitive to outliers compared to MSE.
- Similar to MSE, a lower MAE indicates a better model fit to the data.

# Evaluation metrics for classification

- Classification is a primary task in machine learning.
- Once you build a classification model, you need to evaluate its performance accurately.
- Multiple metrics have been developed to measure the effectiveness of classification models, including **accuracy**, **precision**, **recall**, and the **confusion matrix**.
- While building classification models, it's important not to rely on just one metric.
- The choice of the metric should align with the specific goals and requirements of the problem at hand.

# Accuracy



- **Accuracy** represents the ratio of correctly predicted instances to the total instances in the data set.
- It measures the correctness of a model's predictions.
- The accuracy metric is beneficial when the class distribution is balanced. However, in cases where there's a class imbalance, accuracy can be misleading.
  - For instance, if 95% of the samples belong to Class A and only 5% belong to Class B, a naive model predicting everything as Class A will still have an accuracy of 95%.

# Precision and recall



- **Precision** is the proportion of positive instances that were correctly predicted by the model, to the total number of positive predictions (true and false positives).
- Precision is an indicator of the accuracy of the positive predictions made by the model.
- Recall measures the proportion of positive instances correctly detected by the classifier.
- Recall is a good indicator of the ability of the model to identify the positive class.
- Precision and recall are especially important when false positives and false negatives carry different costs.
  - For instance, a false negative (disease present but not detected) can be far more dangerous in medical tests than a false positive (disease absent but indicated as present).
- Models inherently tradeoff between precision and recall; typically, the higher the precision, the lower the recall, and vice versa.

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

# F1 Score



- **F1 score** is a single evaluation metric that aims to account for and optimize both precision and recall.
- It is defined as the harmonic mean between precision and recall.
- A model will have a high F1 score if both precision and recall are high.
- However, a model will have a low F1 score if one factor is low, even if the other is 100 percent.

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

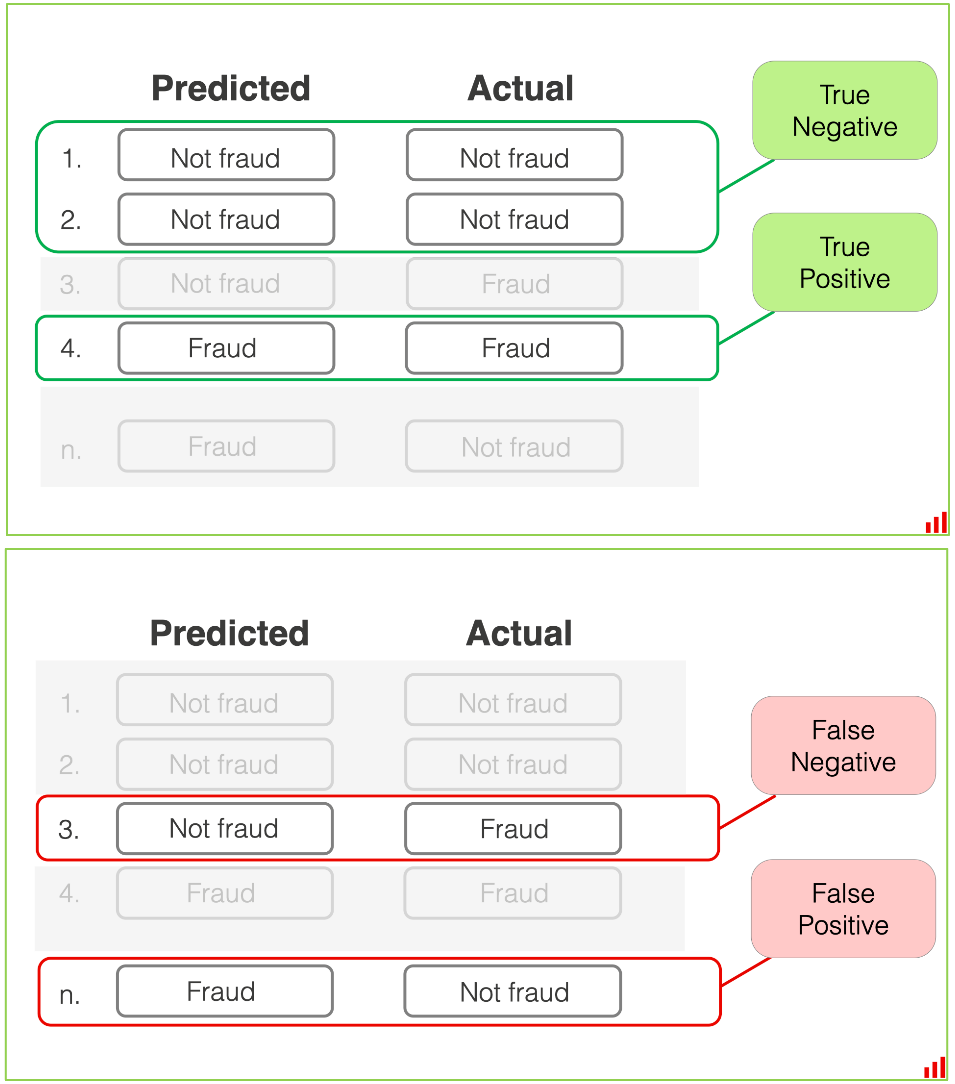
$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

# Confusion Matrix

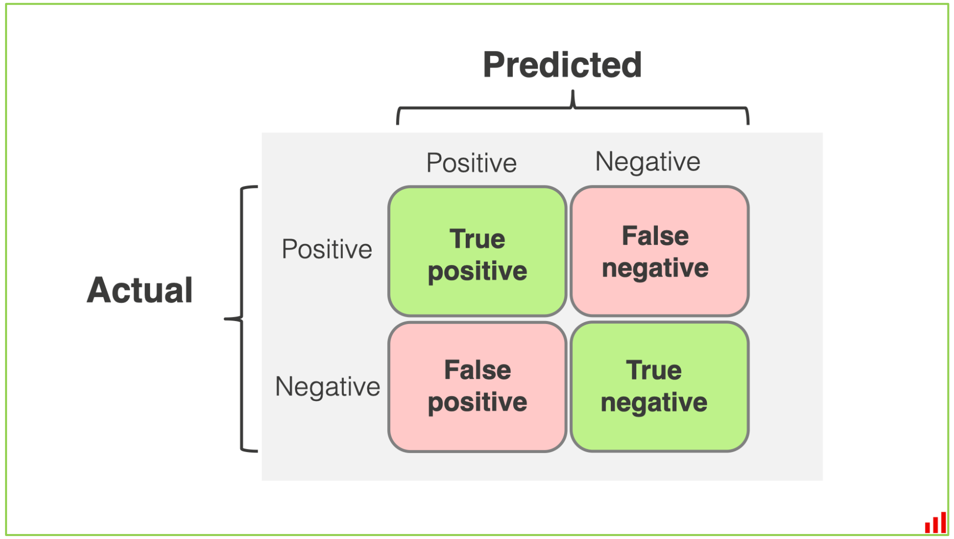


- The confusion matrix is a table that describes the performance of a classification model on a set of data for which you know the true values.
- The matrix typically consists of four values: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).
- True positives and true negatives indicate correct classifications, whereas false positives and false negatives indicate errors.
- Beyond precision and recall, the confusion matrix provides a detailed view of how the predictions distribute across different classes and where the model makes mistakes.

# Confusion Matrix

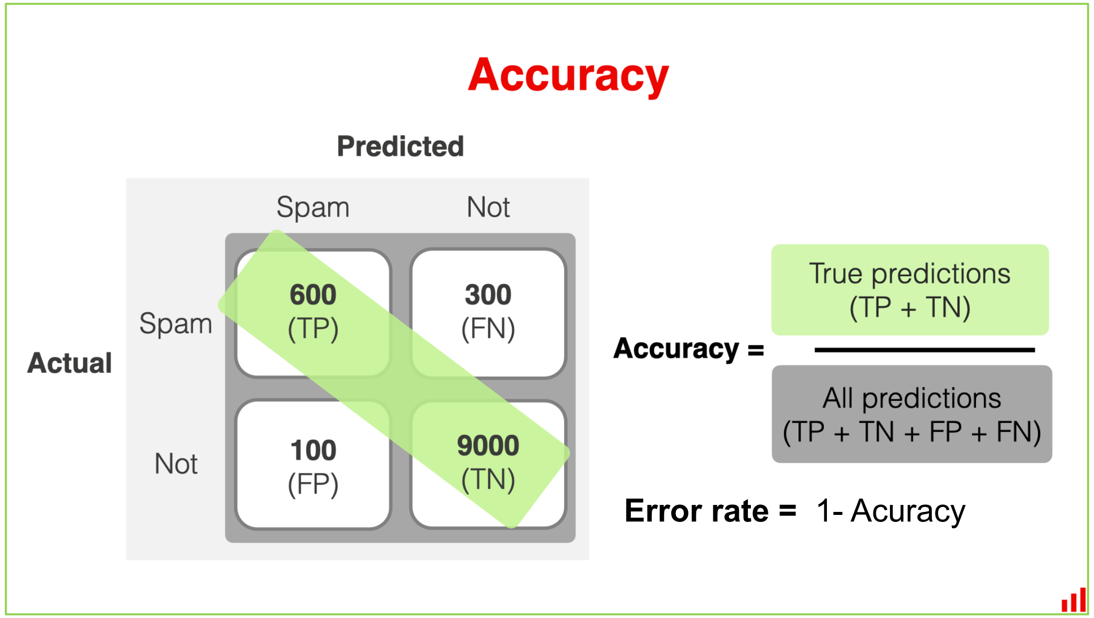


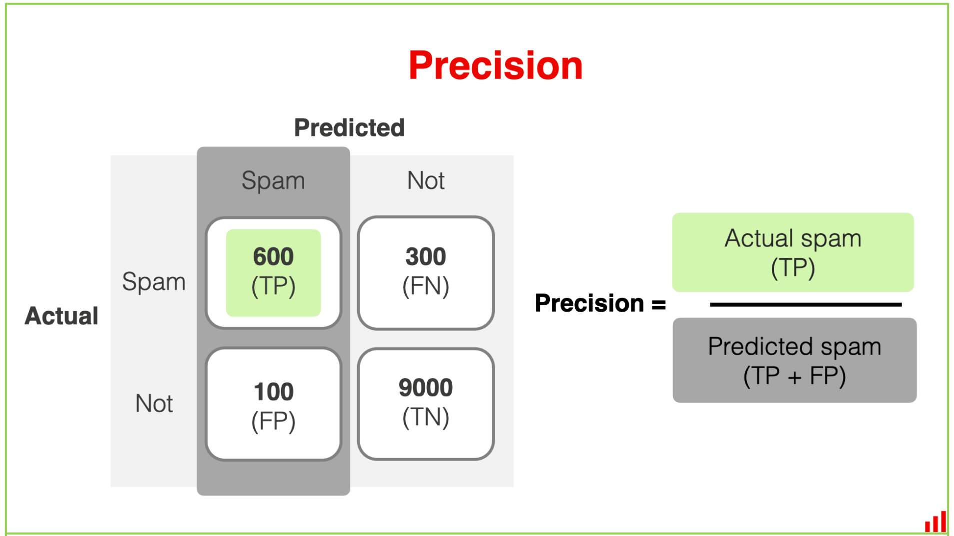

Confusion Matrix

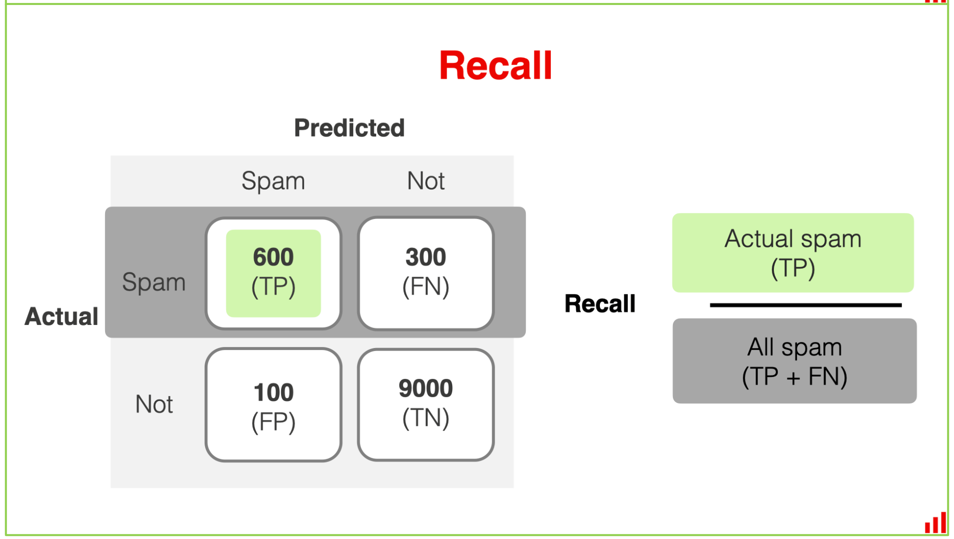


# Example: Precision vs. Recall





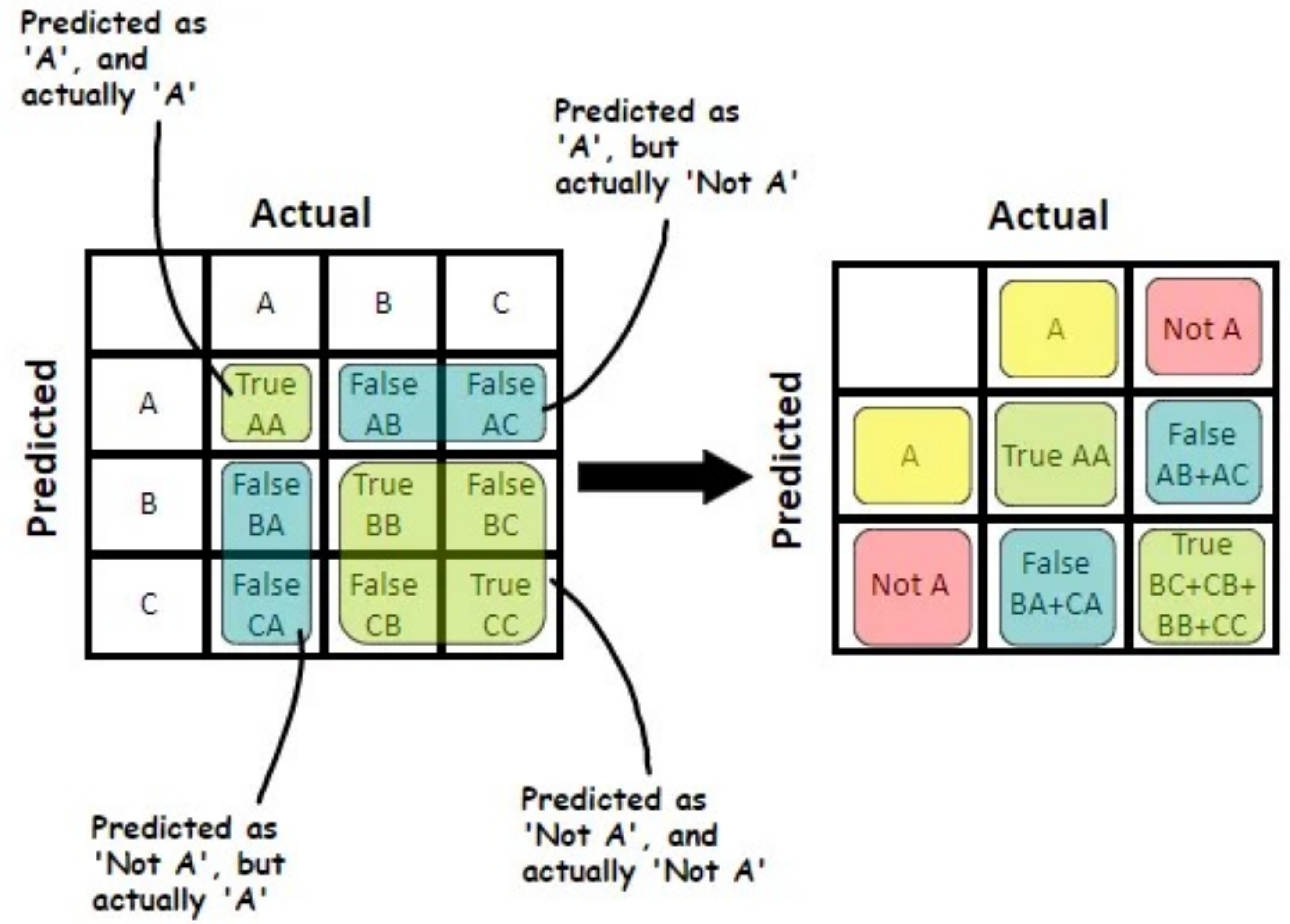




Precision is **600/(600+100)= 0.86**.
*When predicting "spam," the model was correct in 86% of cases.*

Recall is **600/(600+300)= 0.67**.
*The model correctly found 67% of spam emails.*

# Confusion Matrix for Multiclassification



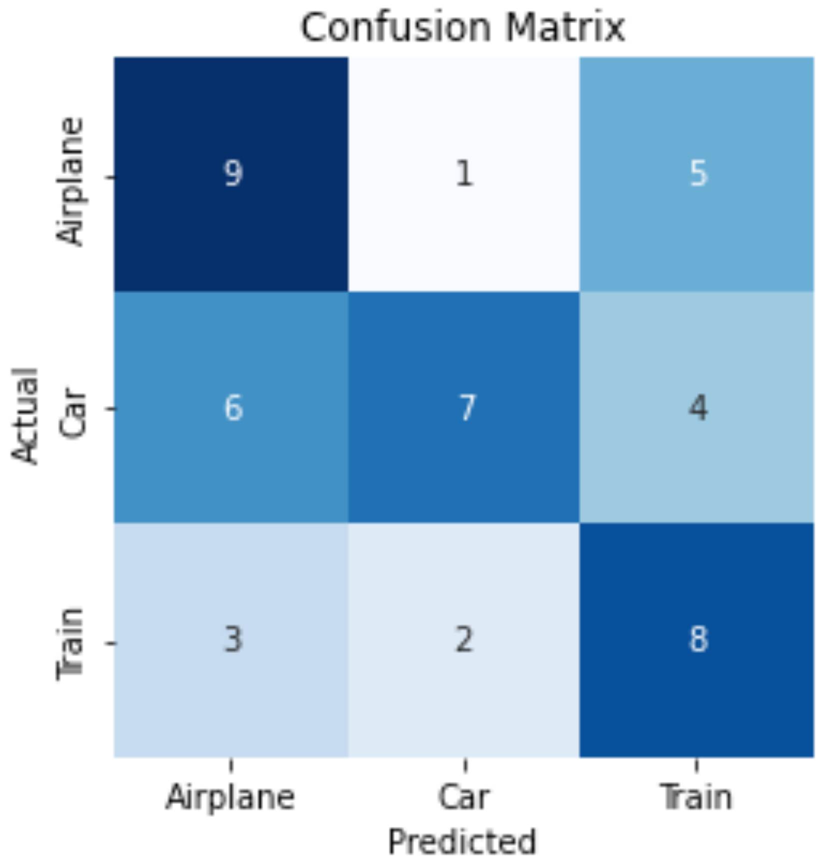
# Example : Confusion Matrix



Calculation of class "Airplane":

- TP = 9
- FN = 1+5 = 6
- FP = 6+3 = 9
- TN = 7+4+2+8 = 21

- Precision = TP/(TP+FP) = 9/(9+9) = 0.5
- Recall = TP/(TP+FN) = 9/(9+6) = 0.6
- F1 = 2*(0.5*0.6)/(0.5+0.6) = 5.55



# Classification Report



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Airplane | 0.50 | 0.60 | 0.55 | 15 |
| Car | 0.70 | 0.41 | 0.52 | 17 |
| Train | 0.47 | 0.62 | 0.53 | 13 |
| accuracy | | | 0.53 | 45 |
| macro avg | 0.56 | 0.54 | 0.53 | 45 |
| weighted avg | 0.57 | 0.53 | 0.53 | 45 |

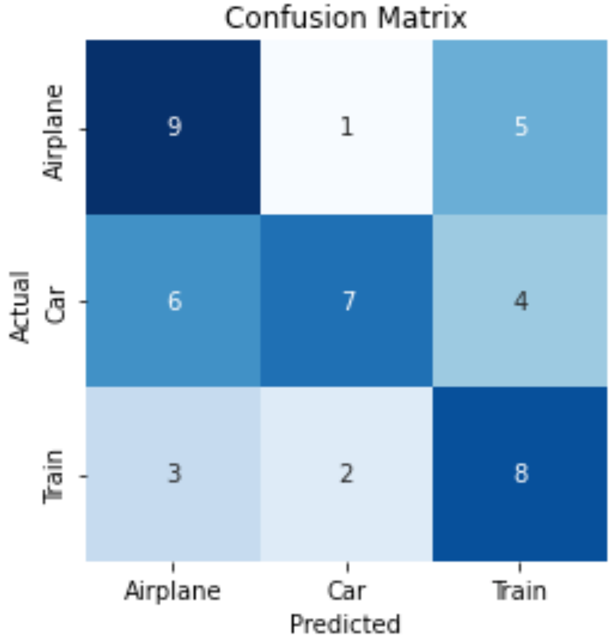
- **Precision**: It is referred to the proportion of correct predictions among all predictions for a particular class.
- **Recall**: It is referred to the proportion of examples of a specific class that have been predicted by the model as belonging to that class.
- **F1 Score**: The Harmonic mean of precision and recall.
- **Accuracy (Micro Precision):** It is calculated by considering the total TP, TN, FN, and TN irrespective of class to calculate Precision.
- **Macro avg**: It is referred to as the unweighted mean of the measure for each class.
- **Weighted avg**: Unlike macro, it is the weighted mean of the measure. **Weights** are the total number of samples per class.

# Example : Classification Report



- Global TP = TP(*airplane*) + TP(*car*) + TP(*train*) = 9+7+8 = 24
- Global FP = FP(*airplane*) + FP(*car*) + FP(*train*) = (6+3) + (1+2) + (5+4) = 21
- **Micro Precision** = Global TP / (Global TP + Global FP )
  = 24 / (24+21) = 0.533
- **Macro Precision** = (P(*airplane*) + P(*car*) + P(*train*)) / #classes
  = (0.50 + 0.70 + 0.47) / 3 = 0.556
- **Weighted Precision** = (**15**×0.50 + **17**×0.70 + **13**×0.47) / **45** = 0.566

(15 airplanes, 17 cars, and 13 trains which aggregated to 45 in total.)

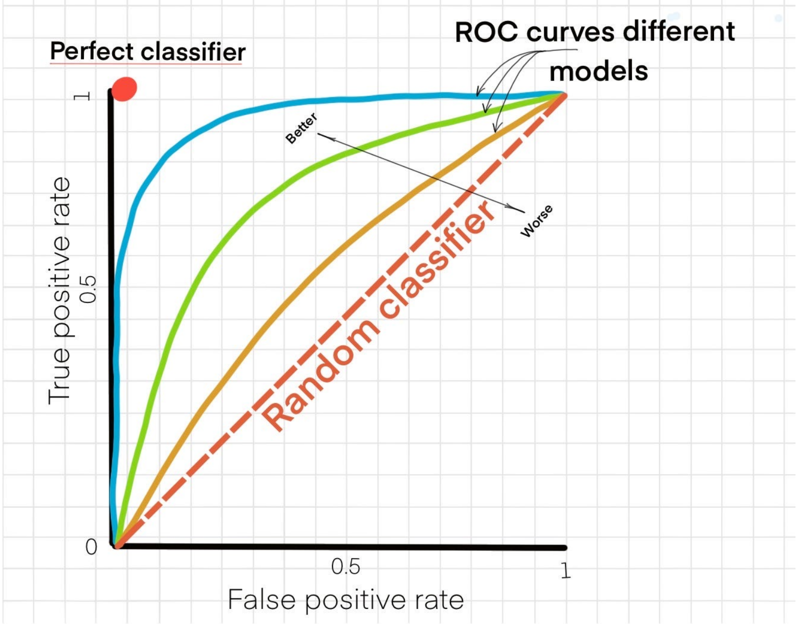



Exercise: Calculate Macro Recall and Weighted Recall.

# ROC



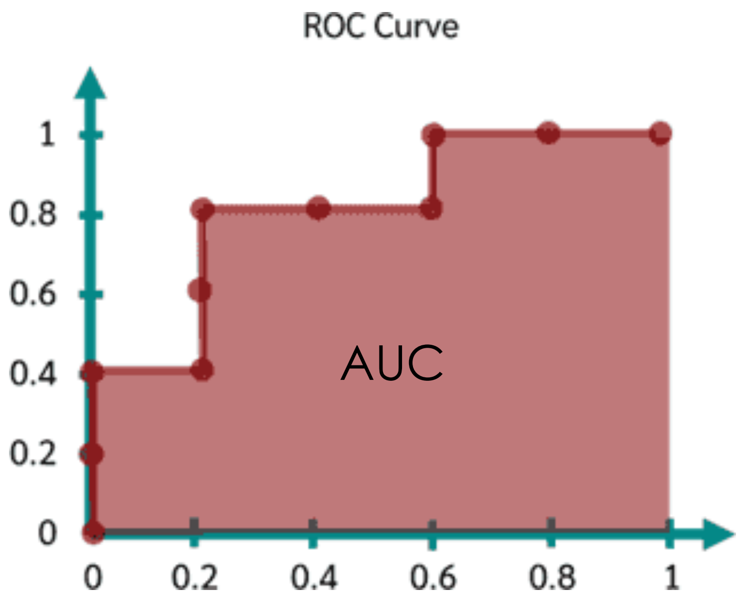
- ROC stands for **Receiver Operating Characteristic**.
- A ROC curve is a graphical representation of the performance of a binary classification model for all classification thresholds.
- The ROC curve plots the True Positive rate (recall) against the False Positive rate.
- Using the ROC curve, we can compare different classification models.



# Area under the Curve (AUC) value



- The **AUC** represents the area under the ROC curve and gives a scalar value, which indicates ***the model's ability to distinguish between the positive and negative classes***.
- AUC represents the probability that a true positive and true negative data points will be classified correctly.
- An AUC value of 0.5 suggests no discrimination (equivalent to random guessing), while a value of 1 indicates perfect discrimination.



# True Positive Rate (TPR) False Positive Rate (FPR)



$$\textbf{\textit{True Positive Rate}} = \frac{\textit{True Positives}}{\underbrace{\textit{True Positives}}_{\text{Correctly classified as "diseased"}} + \underbrace{\textit{False Negatives}}_{\text{Incorrectly classified as "healthy"}}}$$

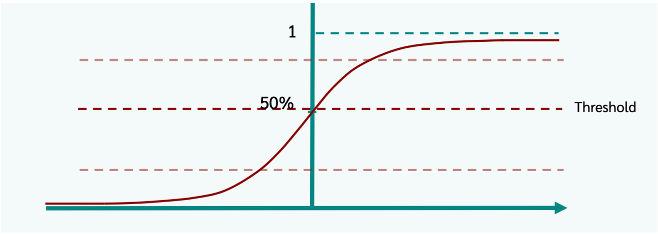
$$\textbf{\textit{False Positive Rate}} = \frac{\textit{False Positives}}{\underbrace{\textit{False Positives}}_{\text{Healthy persons misclassified as "diseased"}} + \underbrace{\textit{True Negatives}}_{\text{Correctly classified as "healthy"}}}$$

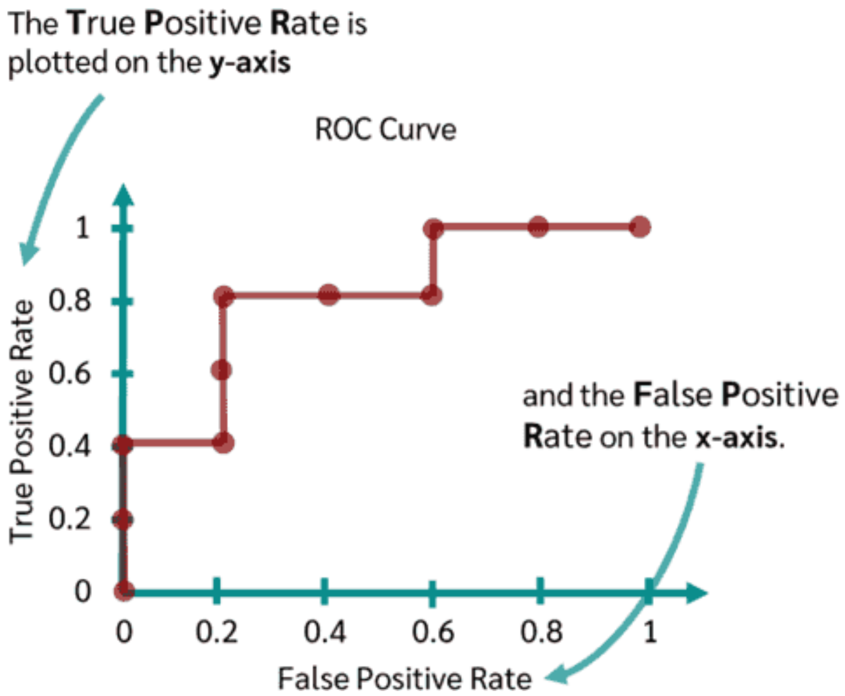
| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

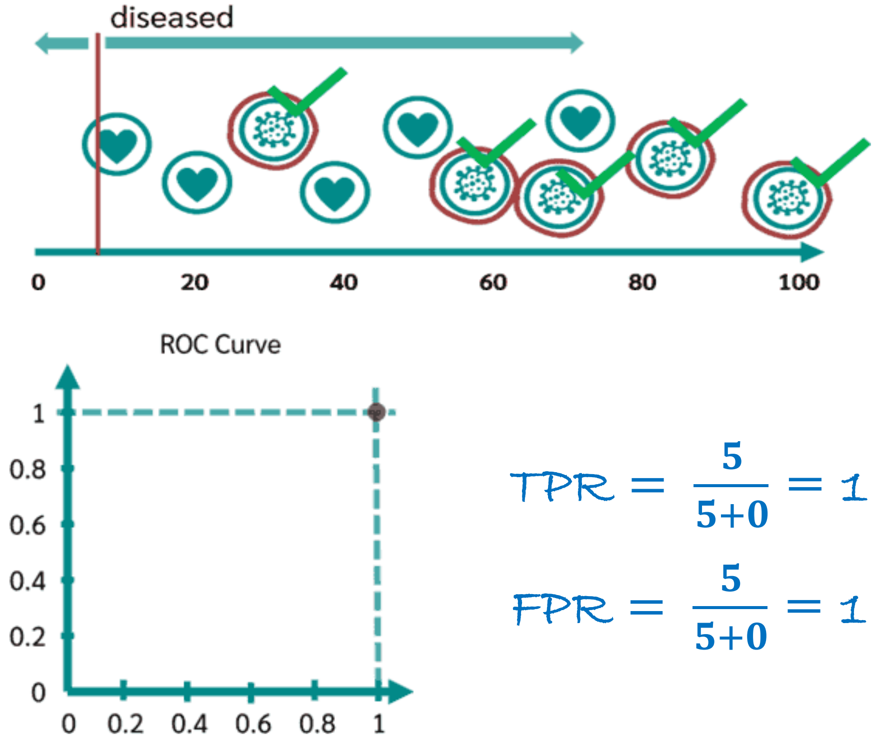
# Plot the ROC Curve



- For each threshold, calculate the TPR and the FPR.
- Then, these two values are plotted on the ROC curve.
- The TPR is plotted on the y-axis and the FPR on the x-axis.



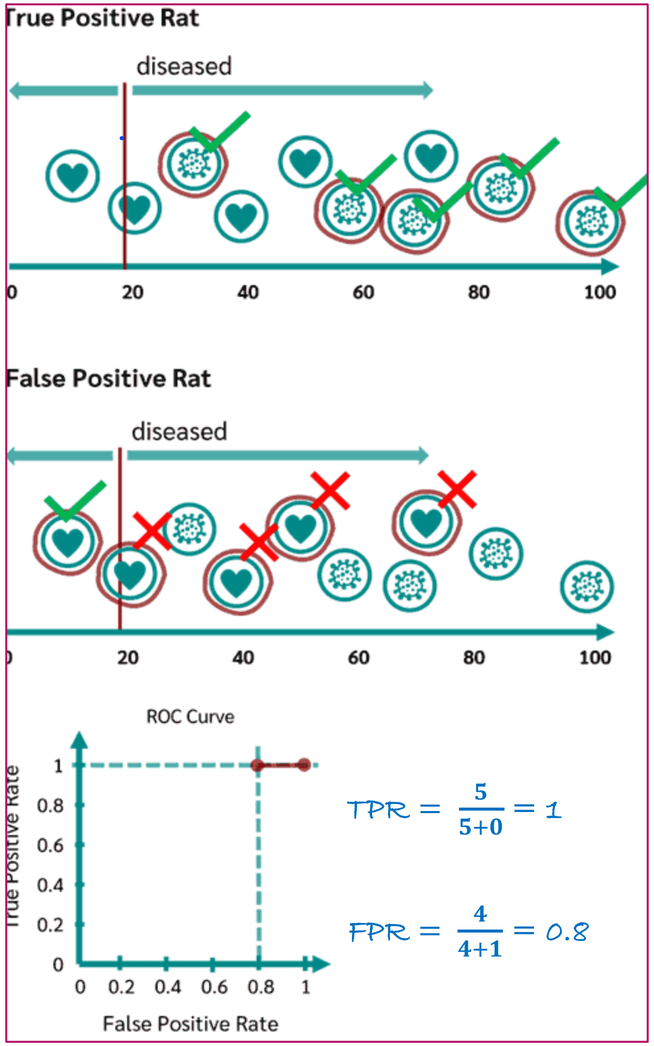




$$TPR = \frac{5}{5+0} = 1$$

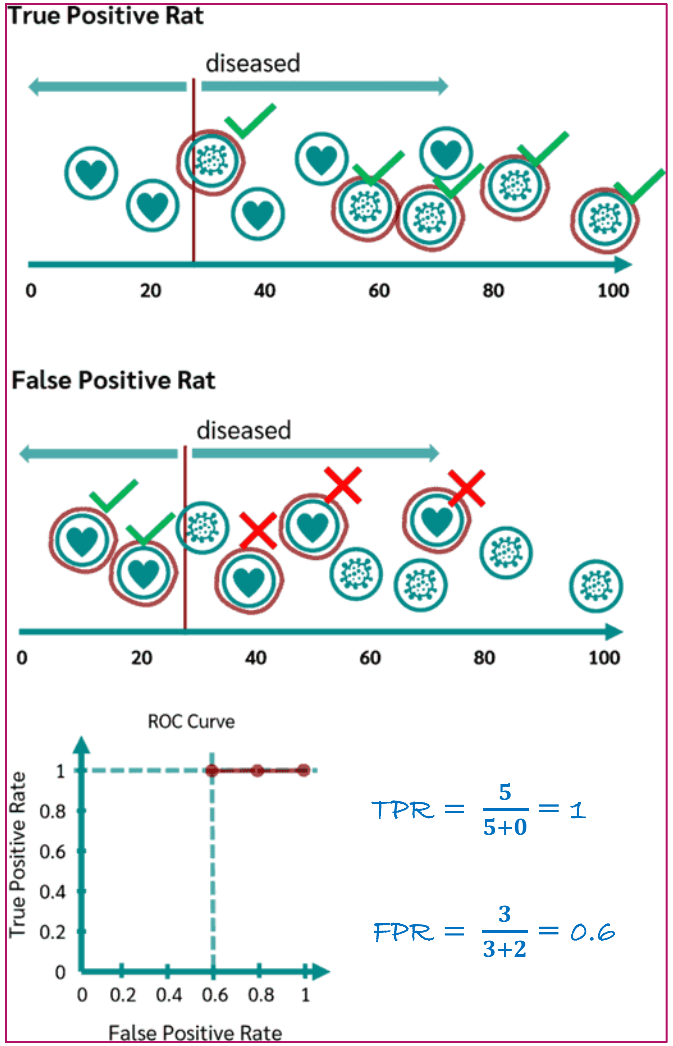
$$FPR = \frac{5}{5+0} = 1$$

# Plot the ROC Curve



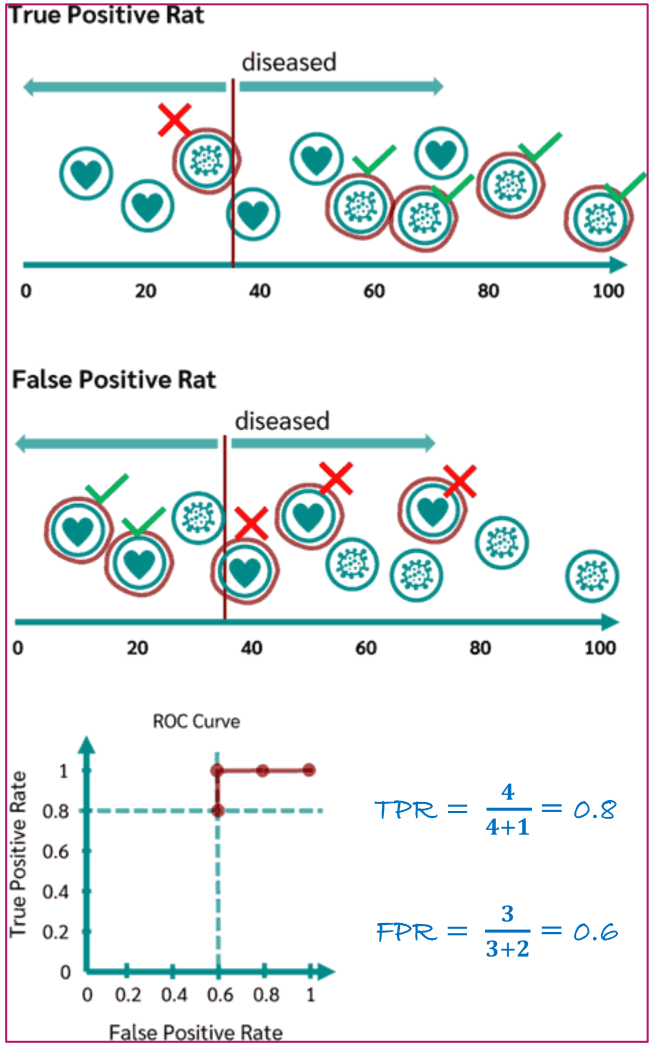

$$TPR = \frac{5}{5+0} = 1$$

$$FPR = \frac{4}{4+1} = 0.8$$




$$TPR = \frac{5}{5+0} = 1$$

$$FPR = \frac{3}{3+2} = 0.6$$



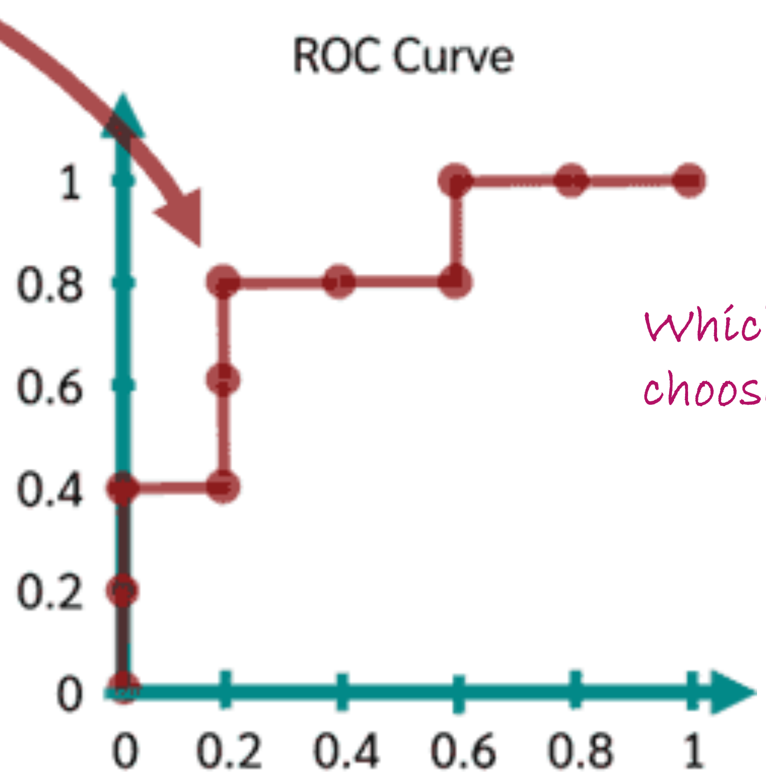

$$TPR = \frac{4}{4+1} = 0.8$$

$$FPR = \frac{3}{3+2} = 0.6$$

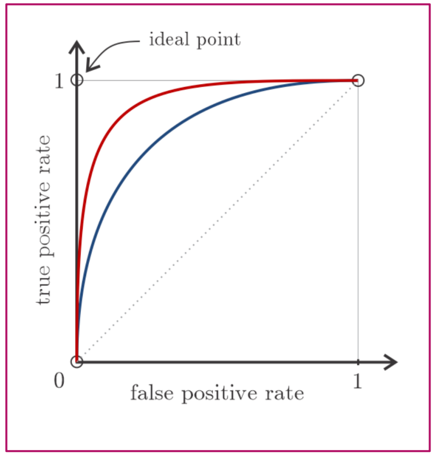
# Plot the ROC Curve



*At this point, 80% of the diseased people were correctly classified as "diseased" and 20% of the healthy people were incorrectly classified as "diseased".*




Which threshold would you choose?"



# Exercise:



- we have a data set with $n = 15$ points and predictions $s(x) \in [0, 1]$
- there are $n_1 = 7$ positive and $n_0 = 8$ negative examples

| Prediction | True class | |
|---|---|---|
| | | $\tau = 1.001$ |
| 0.953 | 1 | |
| 0.920 | 1 | |
| 0.799 | 1 | |
| | | $\tau = 0.794$ |
| 0.788 | 0 | |
| | | $\tau = 0.769$ |
| 0.750 | 1 | |
| 0.679 | 1 | |
| 0.612 | 1 | |
| | | $\tau = 0.598$ |
| 0.583 | 0 | |
| 0.477 | 0 | |
| 0.378 | 0 | |
| | | $\tau = 0.373$ |
| 0.367 | 1 | |
| | | $\tau = 0.306$ |
| 0.248 | 0 | |
| 0.214 | 0 | |
| 0.157 | 0 | |
| 0.112 | 0 | |
| | | $\tau = 0.000$ |

- Plot the ROC curve and Calculate the AUC.