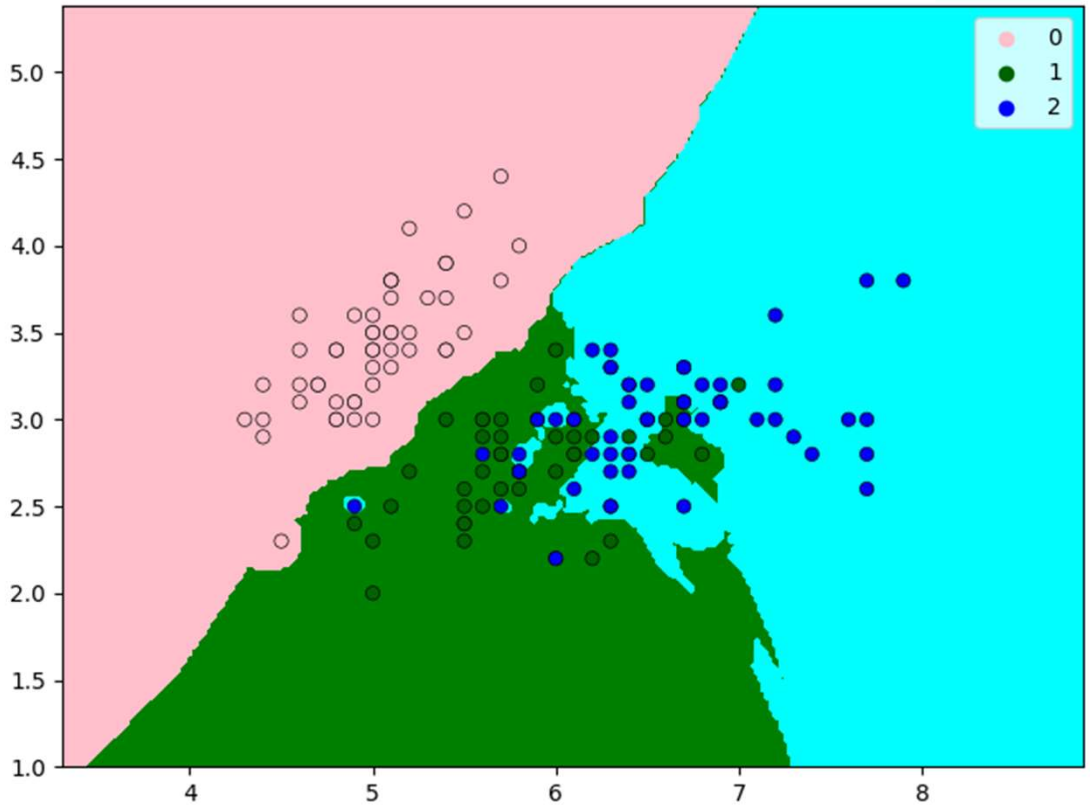
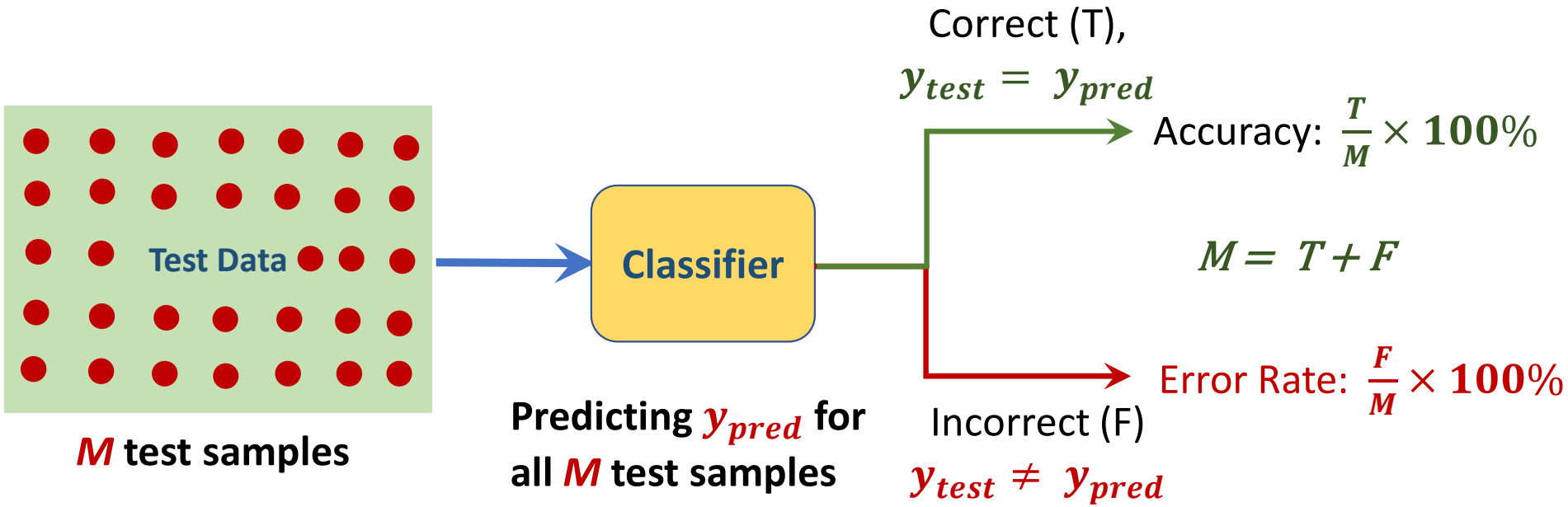
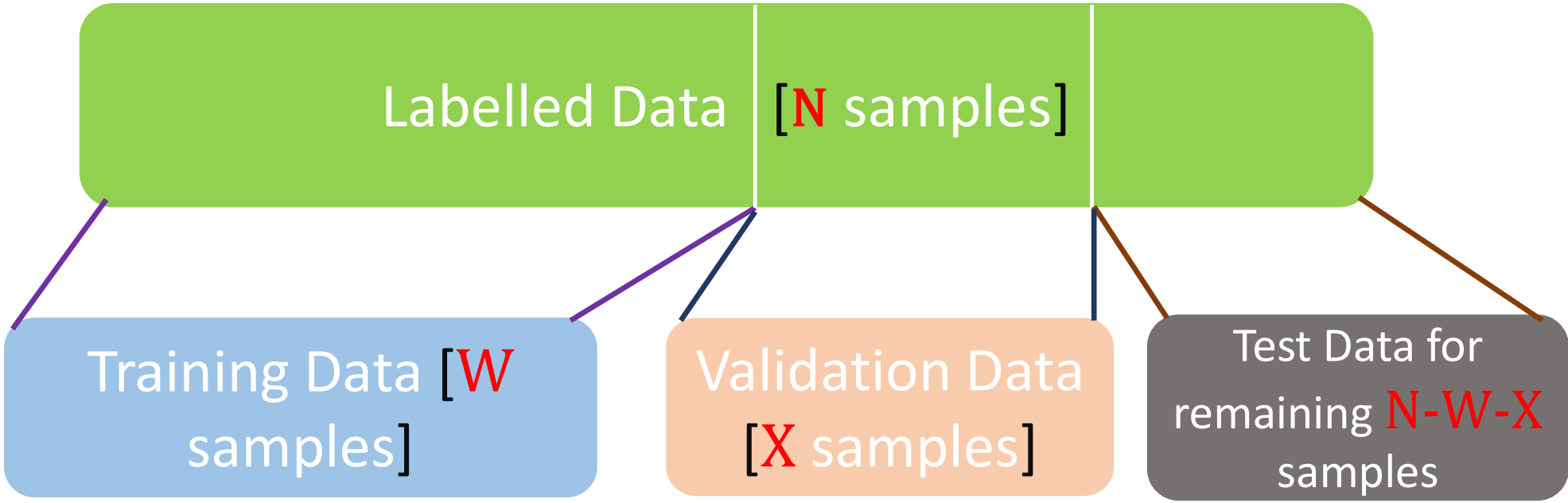
# Decision Boundaries - Revisit



# Evaluating a Classifier - Accuracy



# ML Based on Training - Validation - Testing Data




- What if our selection of validation/test set is biased
  - What happens to our model during training?
  - What happens to our estimate of accuracy?

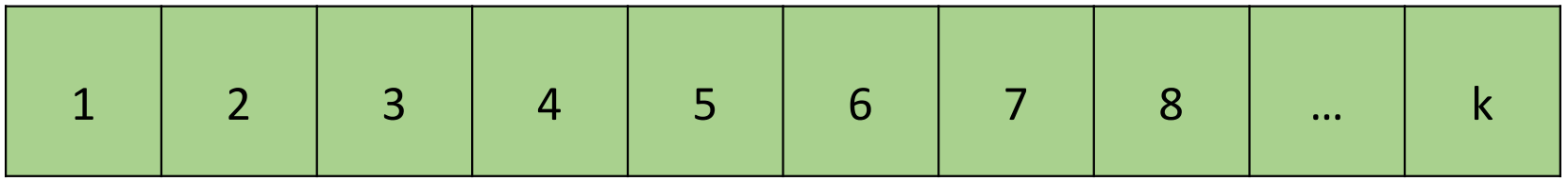
# K-fold Cross Validation

# Cross Validation (CV)

- It is a process in which the original dataset is divided is divided into two parts only- the ‘training dataset’ and the ‘testing dataset’

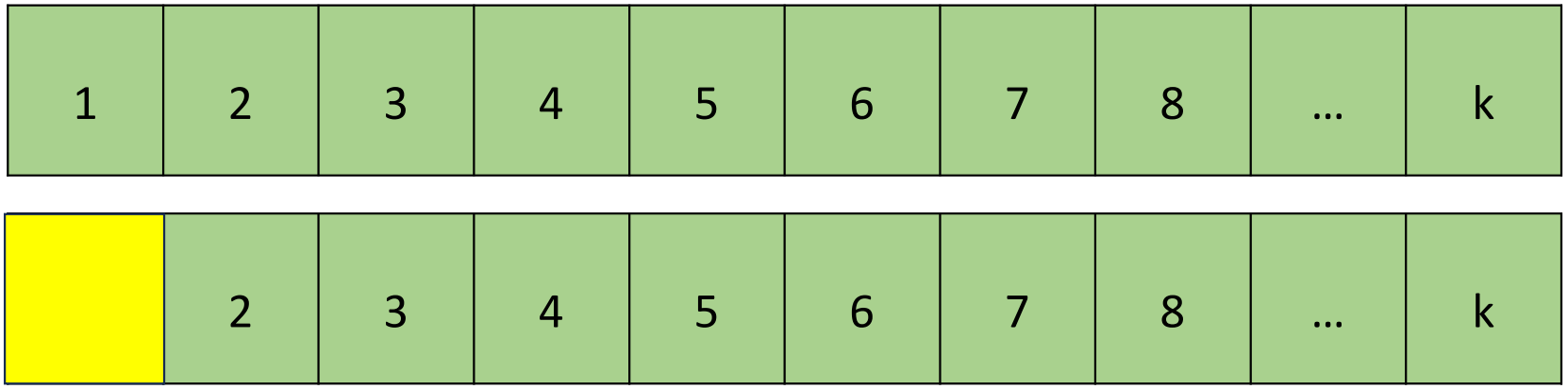
1. Shuffle the dataset randomly

2. Split the dataset into k groups [also known as k-fold]

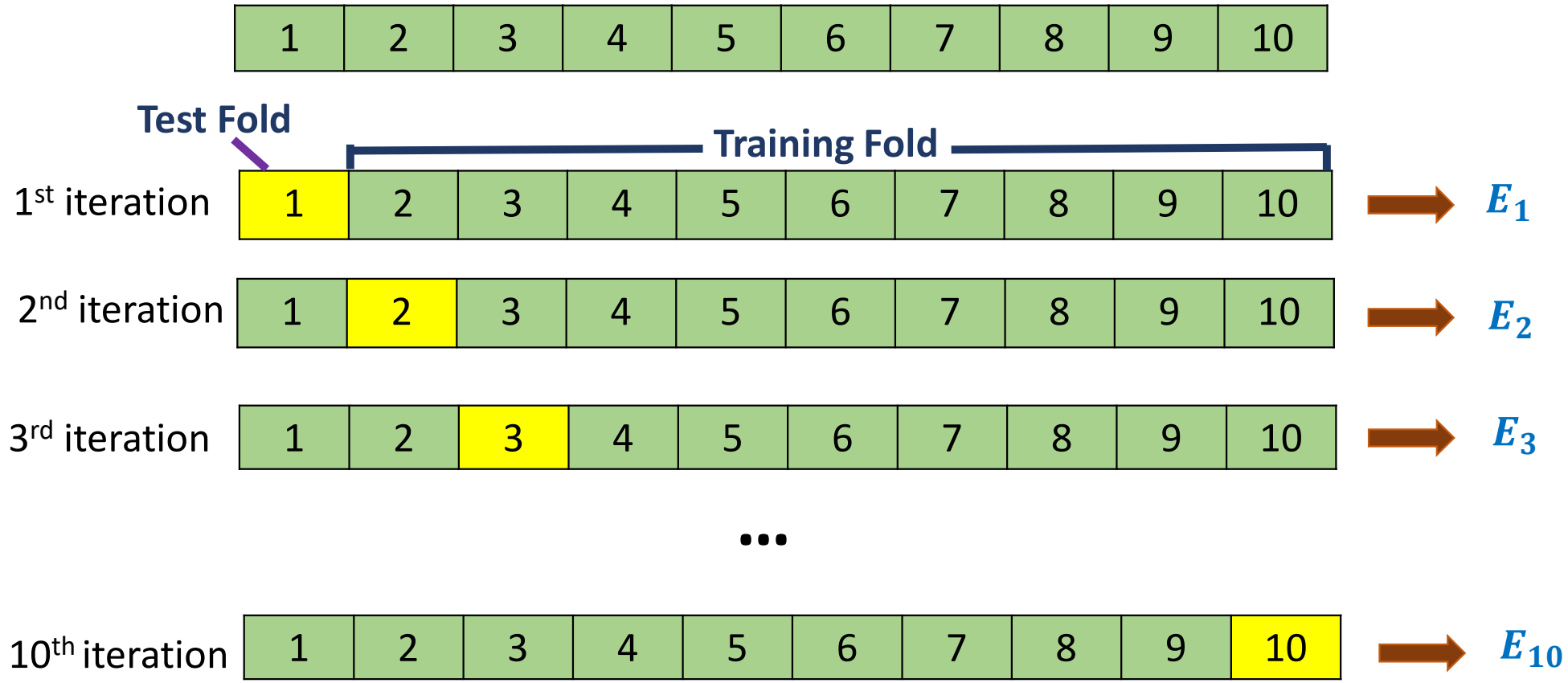


# Cross Validation - CV

3. For each unique group:
   - Take the group as a hold out or test data set
   - Take the remaining groups as a training data set
   - Fit a model on the training set and evaluate it on the test set
   - Retain the evaluation score and discard the model

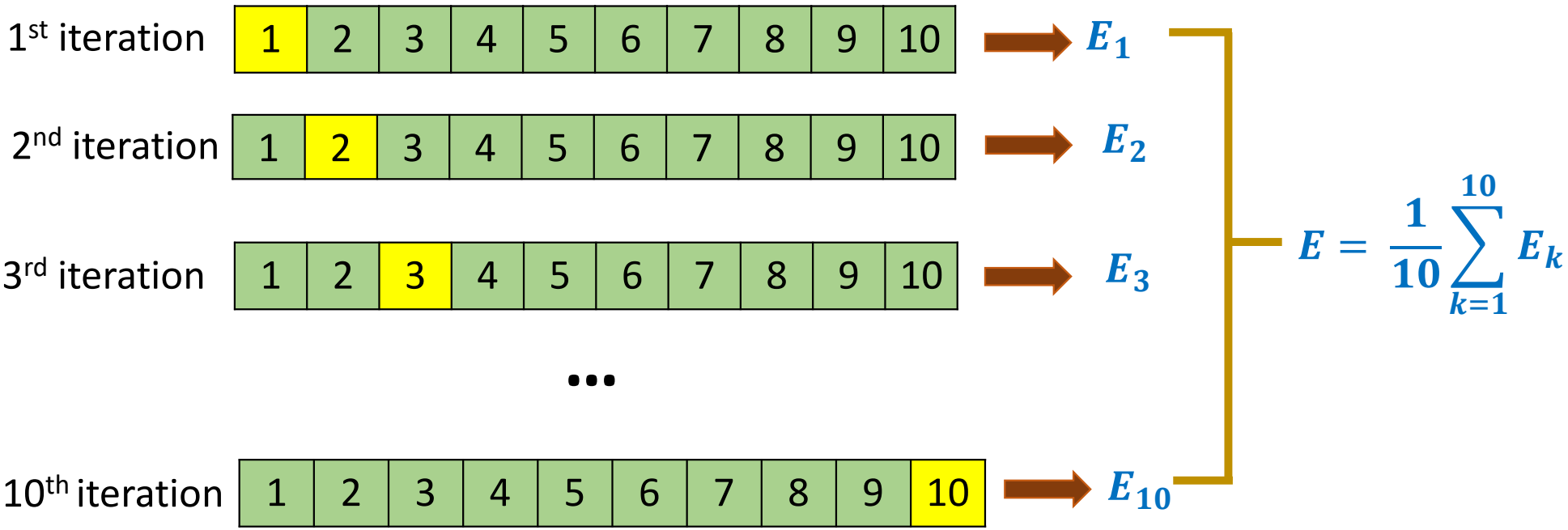


# Cross Validation - CV



# Cross Validation - CV

4. Summarize the skill of the model using the sample of model evaluation scores



# Cross-Validation Notes

- Significantly reduces the variance of the performance estimate
  - Note: Each estimate may not be reliable, but the mean is. Guarantees that the model's score does not depend on how the train/test set are picked
- Provides a confidence interval for final estimate
- Avoids overfitting by exposing the model to different subsets of data
- Does not produce a single trained model
  - May train the model using the whole labelled data
- Takes more time to train

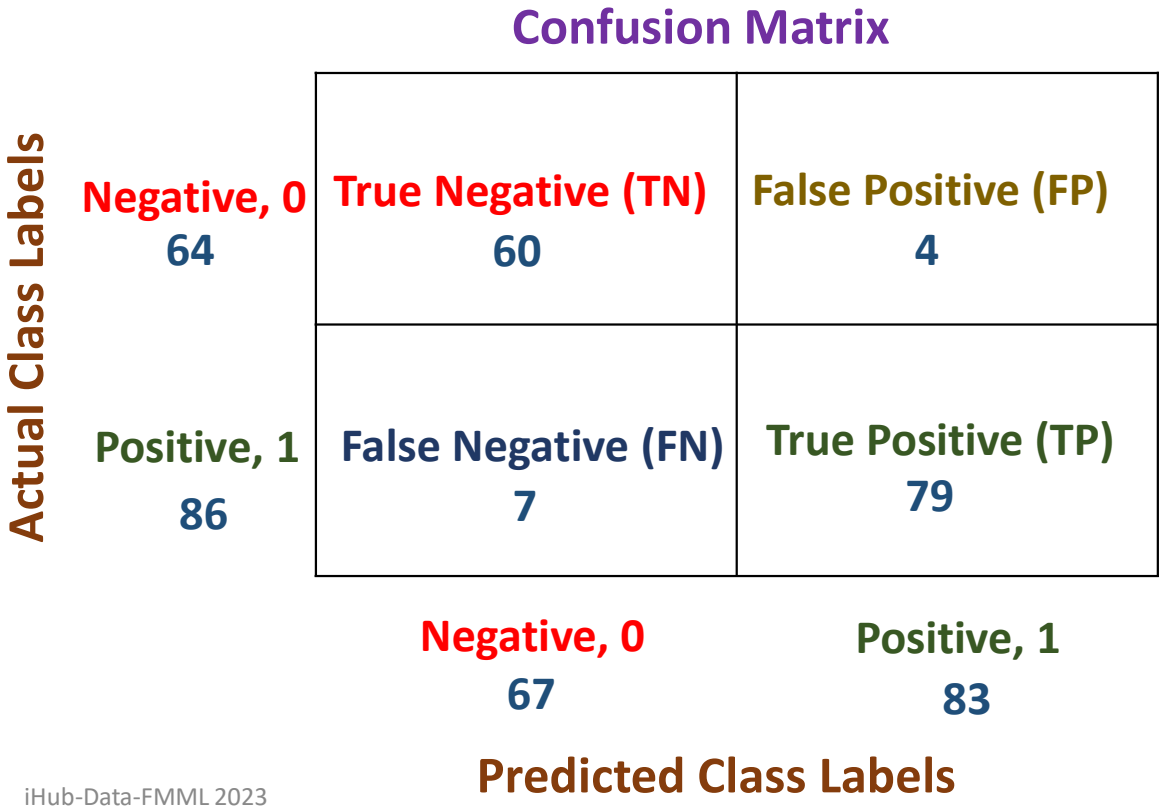
# Metrics to Measure Classification Performance

"All models are wrong, but some are useful", George Box

# Confusion Matrix

- A common way of presenting True Positive (**TP**), True Negative (**TN**), False Positive (**FP**), and False Negative (**FN**) predictions.

- Accuracy $= \frac{TP+TN}{TP+FP+TN+FN} = \frac{60+79}{60+4+7+79} = 0.93$



# Confusion Matrix

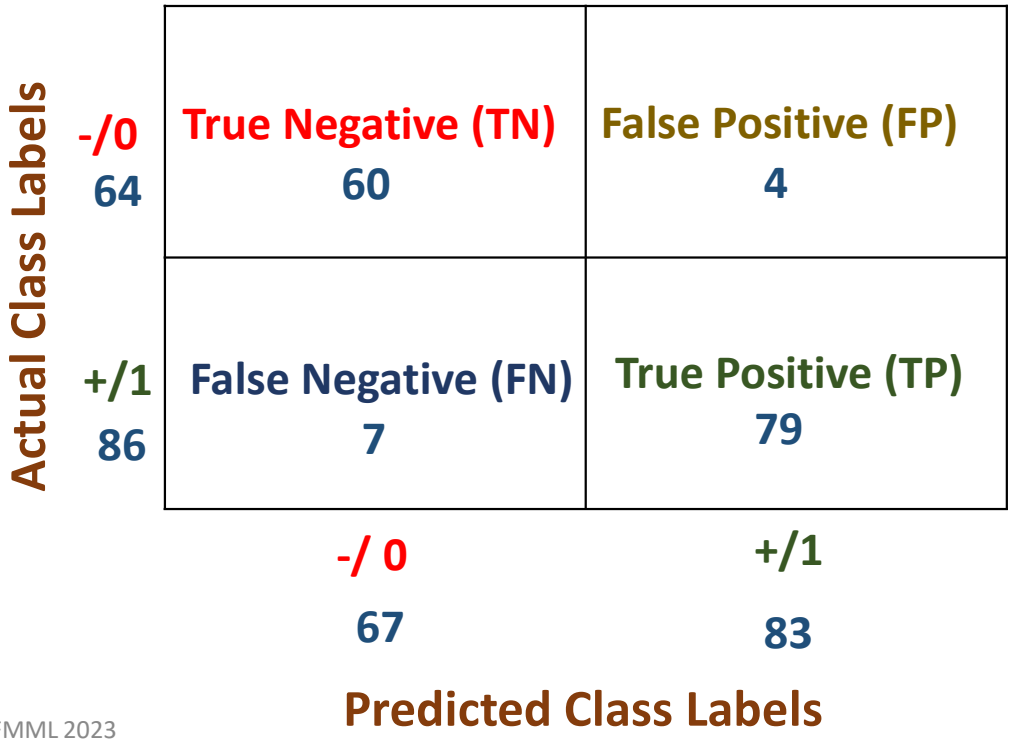
- Misclassification $= \frac{FP+FN}{TP+FP+TN+FN} = \frac{4+7}{60+4+7+79} = 0.09$

- True Positive Rate (TPR)/Sensitivity

$$= \frac{TP}{TP+FN} = \frac{79}{79+7} = 0.92$$

- False Positive Rate (FPR)/False Alarm

$$= \frac{FP}{TN+FP} = \frac{4}{60+4} = 0.06$$



# Confusion Matrix

- Misclassification $= \frac{FP+FN}{TP+FP+TN+FN} = \frac{4+7}{60+4+7+79} = 0.09$

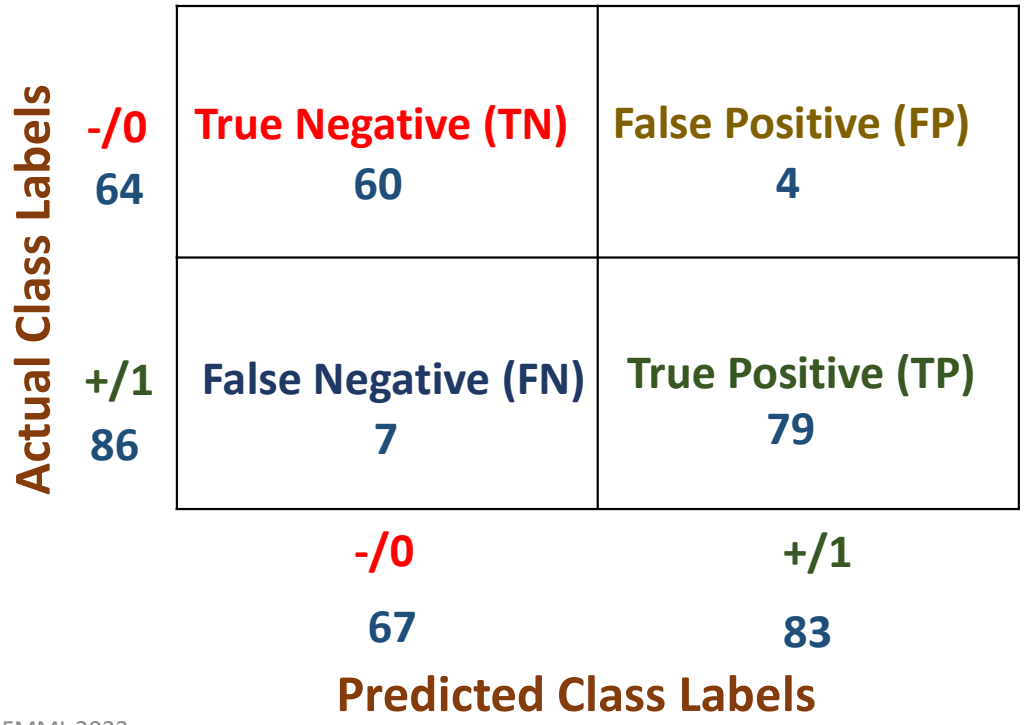
- True Negative Rate (TNR)/Specificity

$$= \frac{TN}{TN+FP} = \frac{60}{60+4} = 0.94$$

- False Negative Rate (FNR)

$$= \frac{FN}{TP+FN} = \frac{7}{79+7} = 0.08$$

- Check: FPR = 1-TNR = 1- Specificity



# Precision, Recall

- Precision = $\frac{TP}{TP+FP} = \frac{79}{79+4} = 0.95$

✓ Ratio of correctly predicted positive observations to the total predicted positive observations

- Recall = $\frac{TP}{TP+FN} = \frac{79}{79+7} = 0.92$

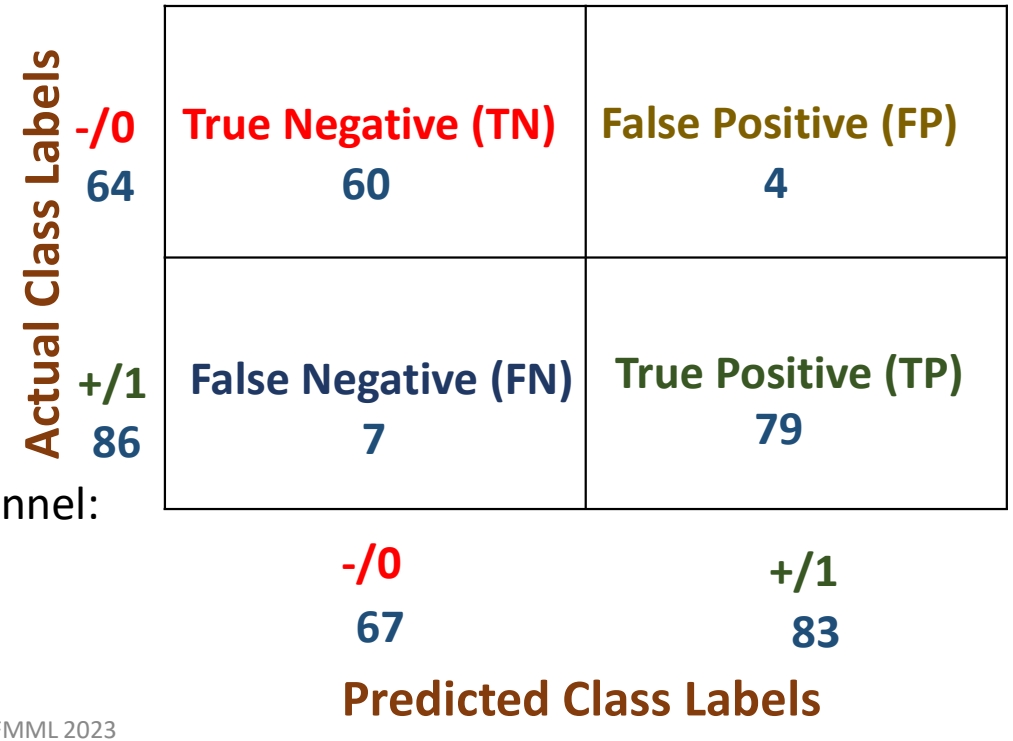
✓ Ratio of correctly predicted positive observations to the total positive observations in actual class

| Actual Class Labels | | |
|---|---|---|
| -/0 64 | True Negative (TN) 60 | False Positive (FP) 4 |
| +/1 86 | False Negative (FN) 7 | True Positive (TP) 79 |
| | -/0 67 | +/1 83 |

Predicted Class Labels

# Which rates are important?

- Screening for a terminal disease.
  - Do not want to miss anyone, Low False Negative, High Recall
- Automatic bombing on detecting a target from a drone
  - Should not hurt civilians: Zero False Alarms/Zero False Positive
- Giving access to a secure installation
  - Should not give access to unauthorized personnel: Low False Positives, High Precision



# Extension to Multi-class Classifier – Confusion Matrix

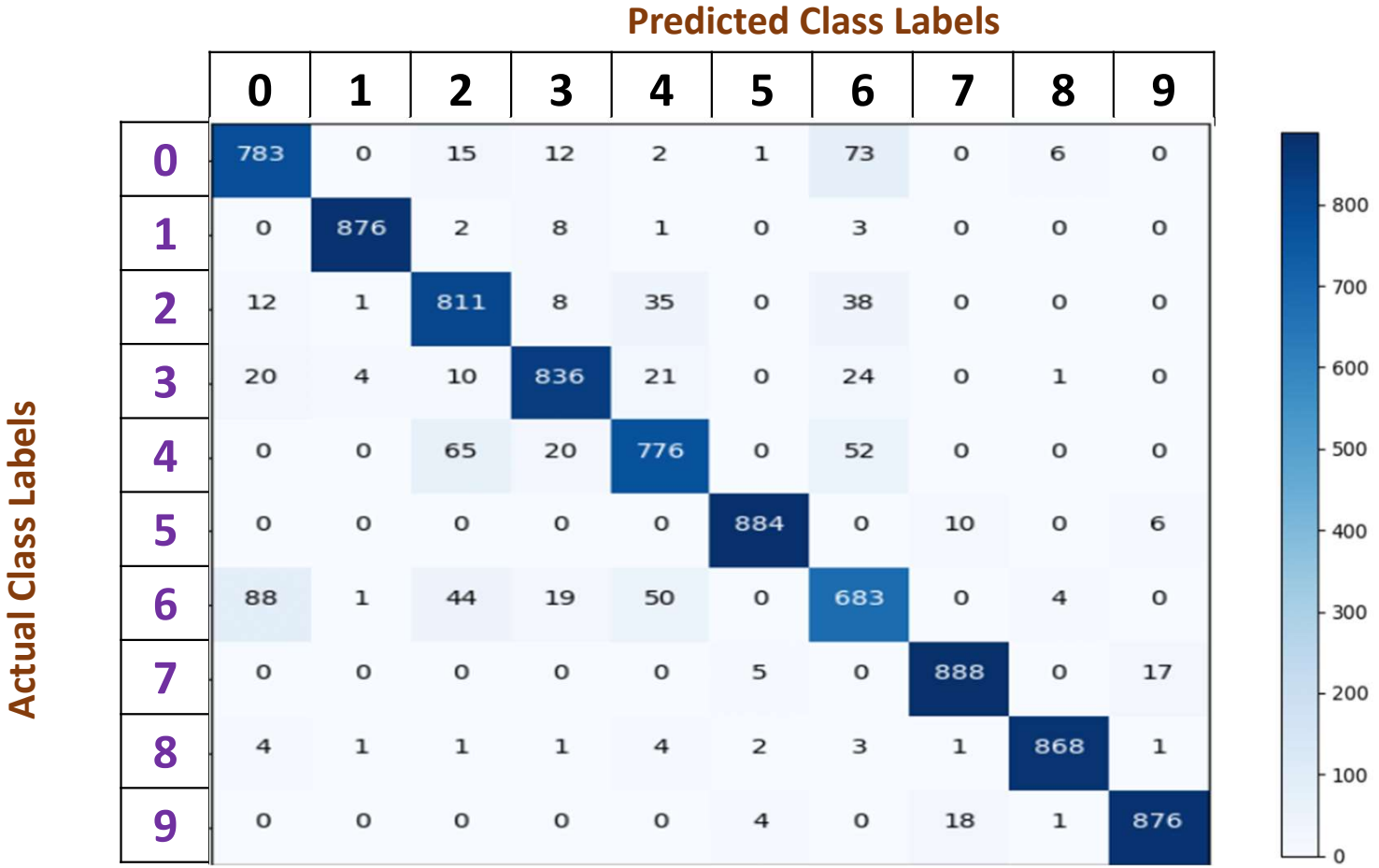



Image Source: https://www.researchgate.net/profile/Lina-Yao-4/publication/325921786/figure/fig3/AS:640163516526592@1529638284399/Confusion-Matrix-on-the-prediction-Fashion-MNIST-dataset.png

# Precision vs Recall

## The Tradeoff

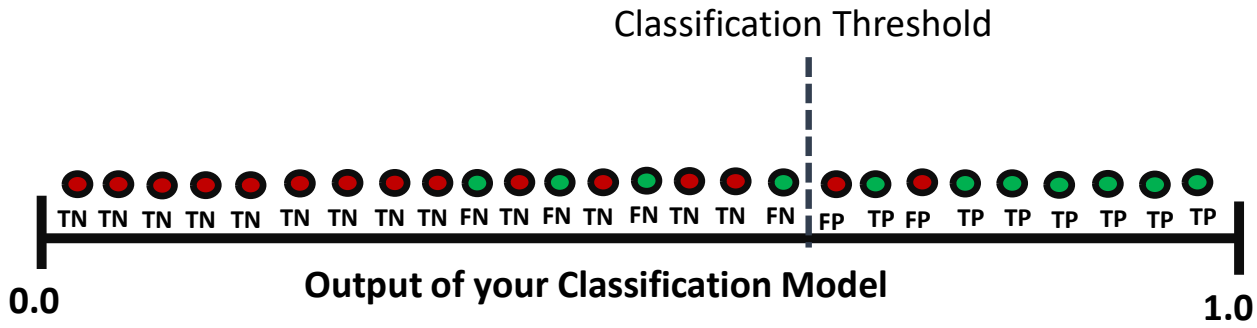
# Classification Threshold

- A **classification threshold** (also called the **decision threshold**) is a value that determines how the model assigns data points to one of the 2 classes
- A value above the threshold indicates “Class 1"; a value below indicates “Class 0."
- It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.

# Classification Threshold - Dependence

| Actual Class | Prediction Probability | Threshold > 0.5 | Threshold > 0.6 | Threshold > 0.7 | Threshold > 0.8 |
|---|---|---|---|---|---|
| 0 | 0.98 | 1 | 1 | 1 | 1 |
| 1 | 0.67 | 1 | 1 | 0 | 0 |
| 1 | 0.58 | 1 | 0 | 0 | 0 |
| 0 | 0.78 | 1 | 1 | 1 | 0 |
| 1 | 0.85 | 1 | 1 | 1 | 1 |
| 0 | 0.86 | 1 | 1 | 1 | 1 |
| 0 | 0.79 | 1 | 1 | 1 | 0 |
| 0 | 0.89 | 1 | 1 | 1 | 1 |
| 1 | 0.82 | 1 | 1 | 1 | 1 |
| 0 | 0.86 | 1 | 1 | 1 | 1 |

# Classifying Diseases




Actual Values
- Has Disease
- No Disease

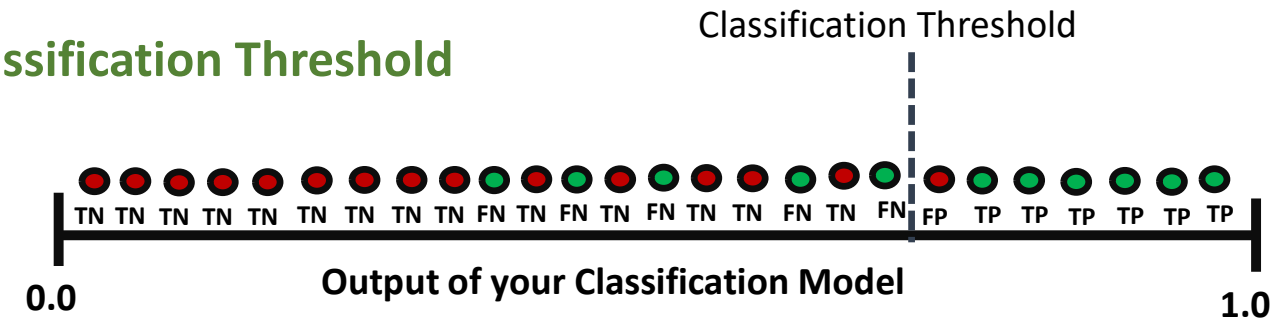
| | |
|---|---|
| True Positive (TP) : | False Positives (FP) : |
| False Negatives (FN): | True Negatives (TN): |

$$Precision = \frac{TP}{TP+FP} = \frac{7}{7+2} = 0.77$$

$$Recall = \frac{TP}{TP+FN} = \frac{7}{7+4} = 0.63$$

# Classifying Diseases

**Increasing Classification Threshold**




Actual Values
- Has Disease
- No Disease

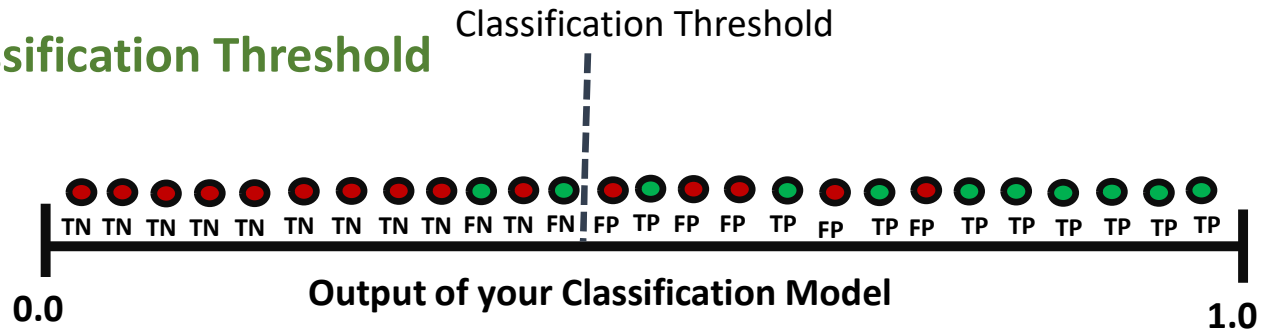
| | |
|---|---|
| True Positive (TP) : 6 | False Positives (FP) : 1 |
| False Negatives (FN): 5 | True Negatives (TN): 14 |

$$Precision = \frac{TP}{TP+FP} = \frac{6}{6+1} = 0.85 \ (0.77)$$

$$Recall = \frac{TP}{TP+FN} = \frac{6}{6+5} = 0.54 \ (0.63)$$

# Classifying Diseases

**Decreasing Classification Threshold**




Actual Values
- Has Disease
- No Disease

| | |
|---|---|
| True Positive (TP) : 9 | False Positives (FP) : 5 |
| False Negatives (FN): 2 | True Negatives (TN): 10 |

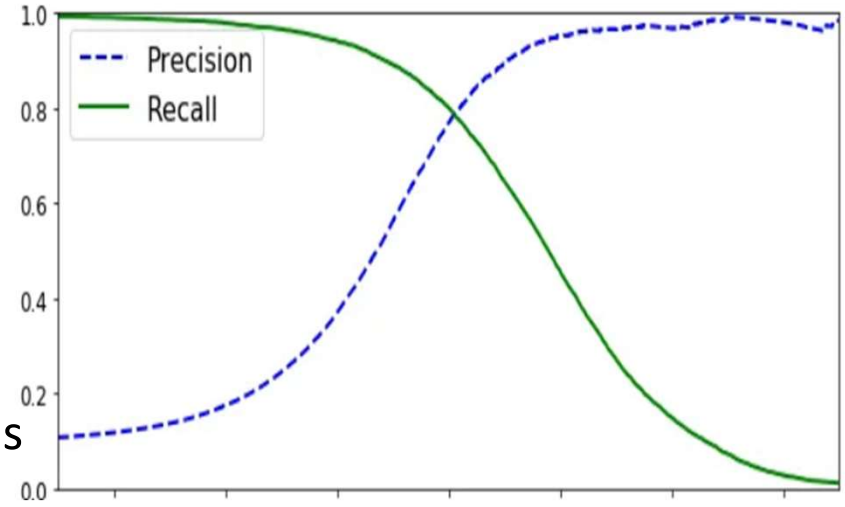
$$Precision = \frac{TP}{TP+FP} = \frac{9}{9+5} = 0.64 \ (0.77) \ (0.85)$$

$$Recall = \frac{TP}{TP+FN} = \frac{9}{9+2} = 0.81 \ (0.63) \ (0.54)$$

# Precision Recall Trade-off

- Precision measures how accurate the model's positive predictions are
- Recall is the measure of how well the model can identify all the positive instances in the data

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$




- The trade-off between precision and recall occurs, as the classification threshold is changed because improving one usually comes at the expense of the other

# F-1 Measure: A Single Metric

- One classifier has high Precision but lower Recall; Another behaves exactly the opposite
- Definition of Recall and Precision have same numerator, different denominators. A sensible way to combine them is harmonic mean.
- F-1 Measure (Information Retrieval): Combines precision and recall in a single score

$$F_1 = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} = 2\,\frac{Precision \times Recall}{Precision+Recal}$$

- Punishes extreme values more

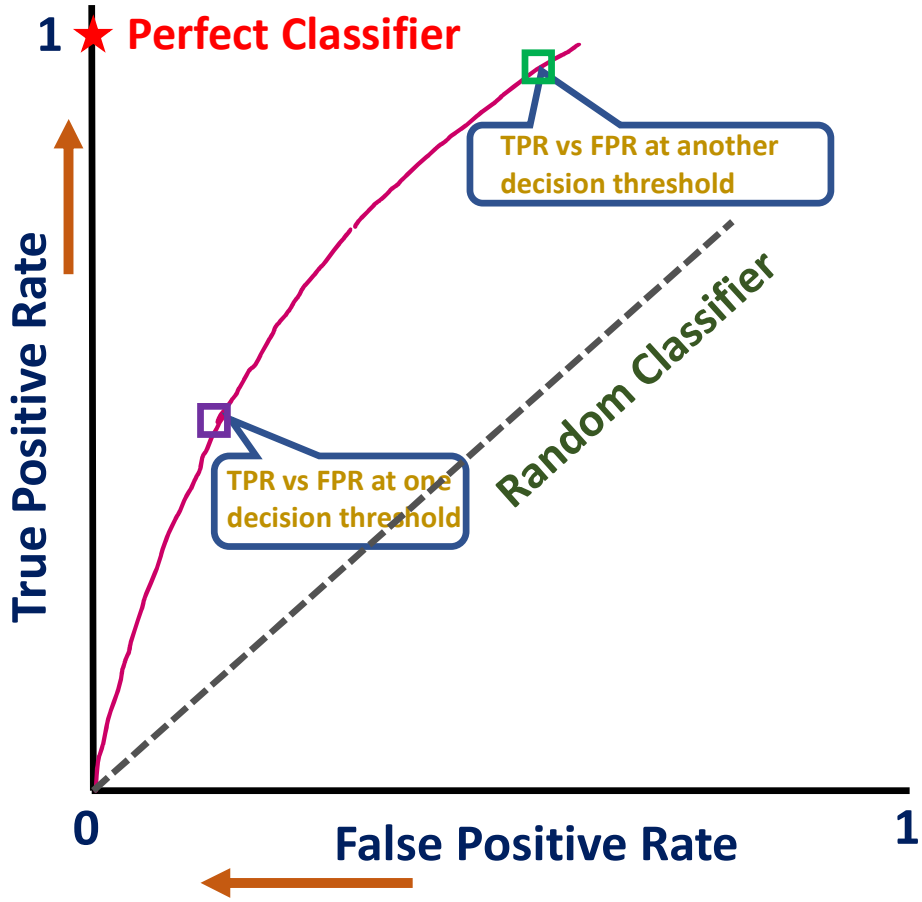
# ROC Curve

- An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
  - True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

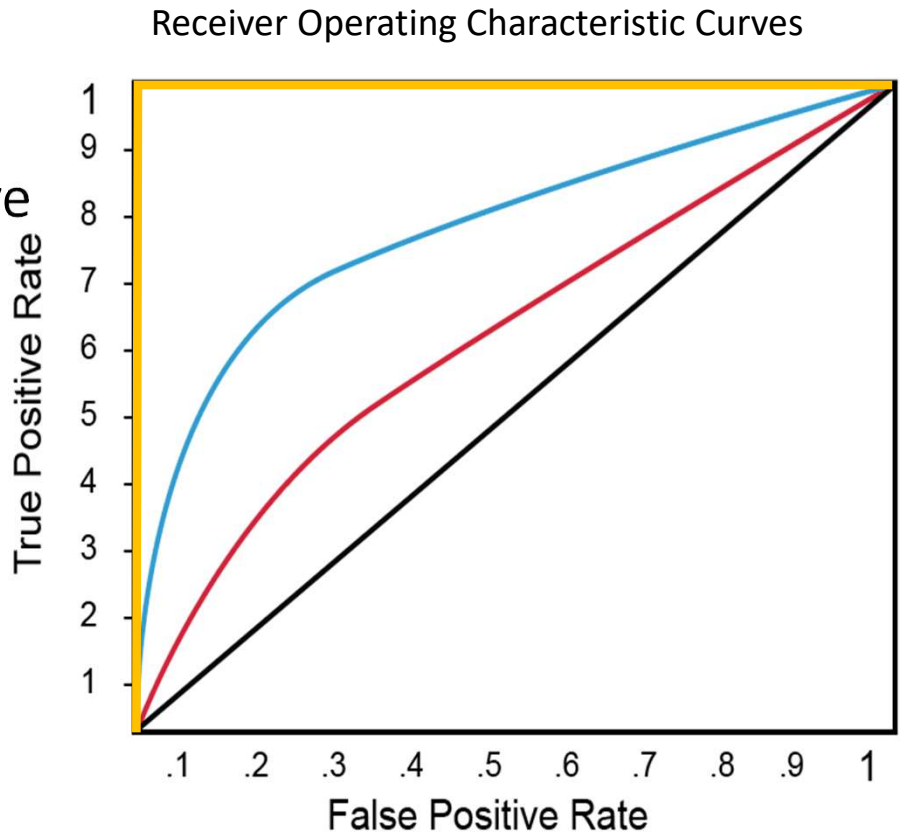
  - False Positive Rate

$$FPR = \frac{FP}{TN + FP}$$



# ROC Curve

- As threshold varies, we move along a curve
- Different representations / distance metrics / algorithms produce different curves
- Blue > Red > Black
- Ideal

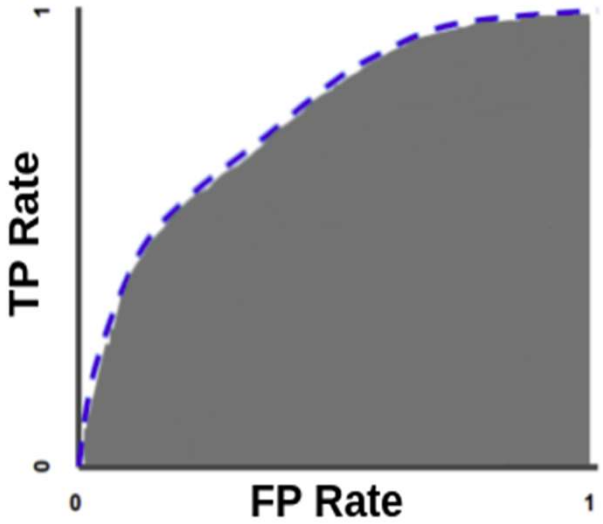


# Area under the Curve

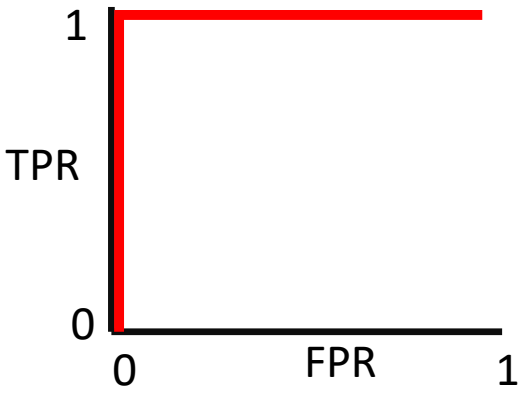
- Area Under Curve measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

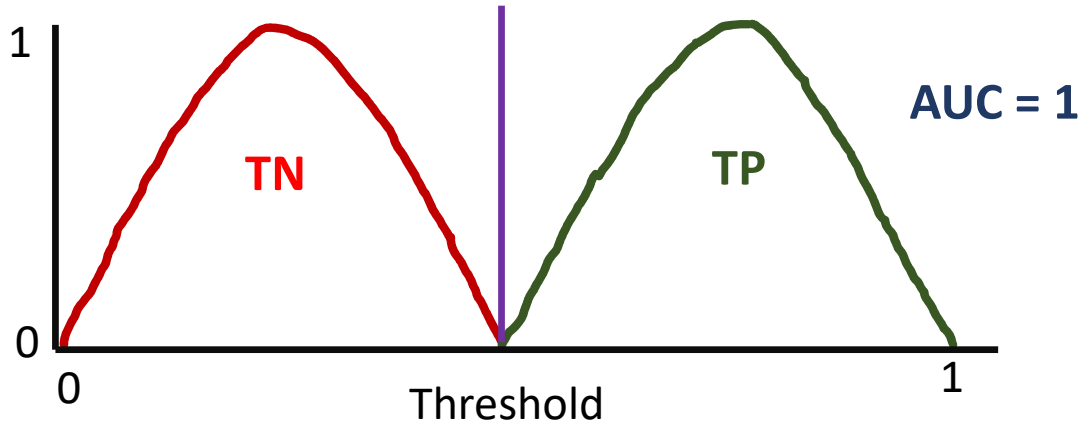
- An excellent model has AUC near to the 1 which means it has a good measure of separability.

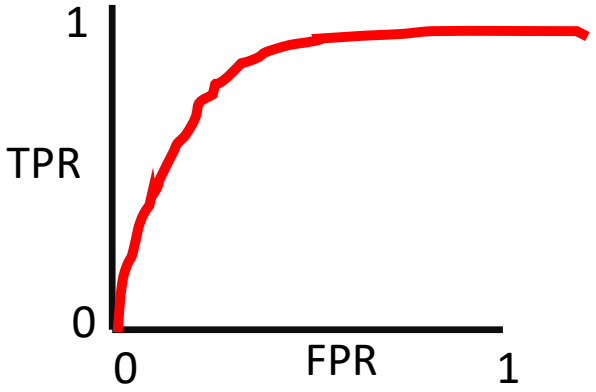
- A poor model has an AUC near 0 which means it has the worst measure of separability.

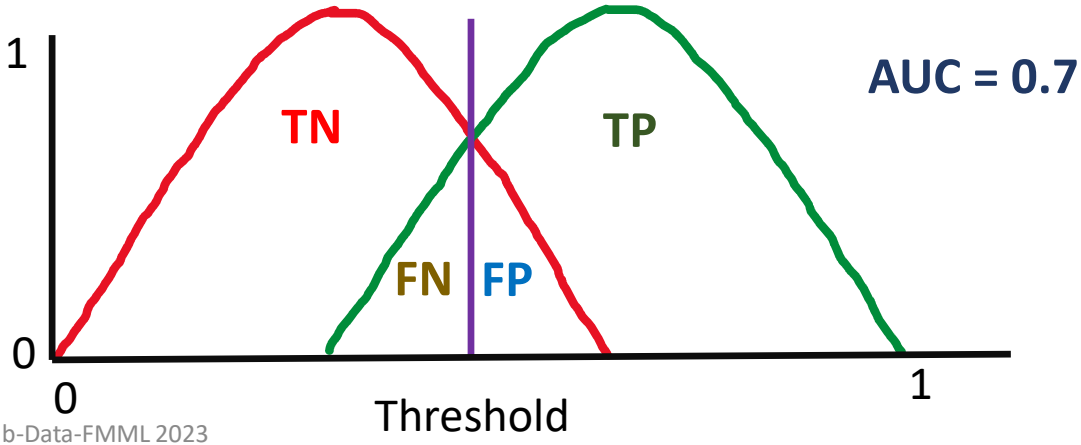


# AUC – Some Illustrations

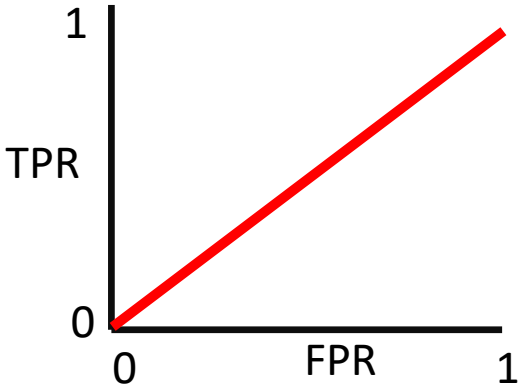


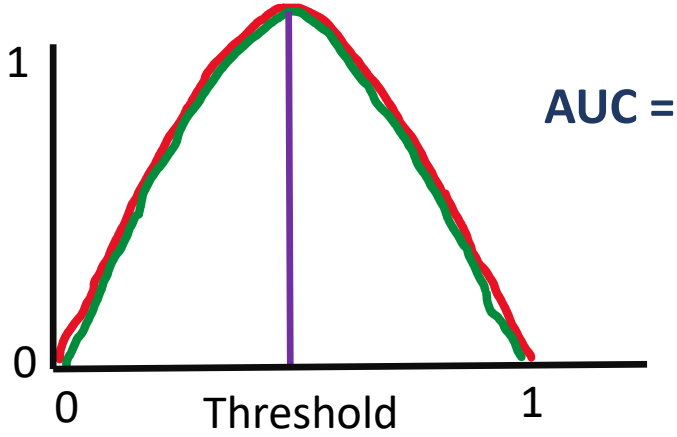



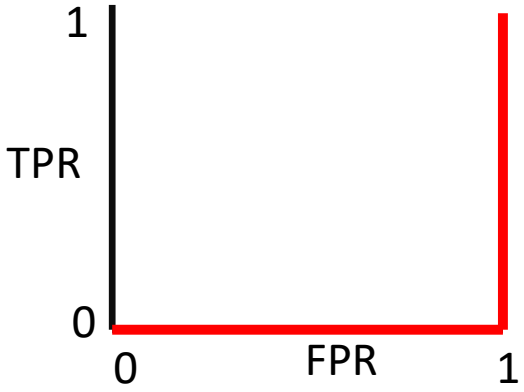



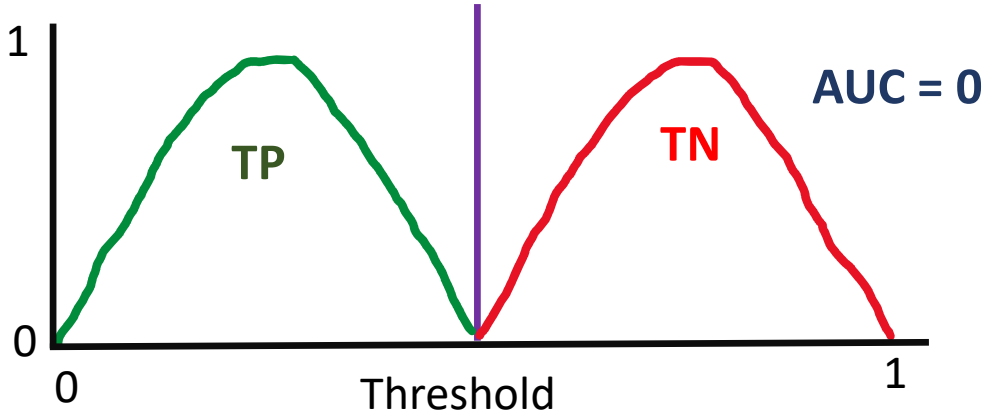



# AUC – Some Illustrations



# Notes on Performance Metrics

- Use the right metric based on the type of problem
- Use a chart that best demonstrates/compares the results
- Use cross-validation whenever possible
  - ✓Report the standard deviation of accuracies
  - ✓Use it to decide if a difference is significant
- Use Single Metrics when appropriate
  - ✓F-Measure
  - ✓Area Under Curve