ACIT4830 – Special Robotics and Control Subject

Evaluating Model Performance

Evi Zouganeli
OsloMet – Oslo Metropolitan University
(evizou@oslomet.no)

Content

(Chapter 10 – evaluating model performance)

- Kappa statistic
- Sensitivity, Specificity
- Precision, Recall, F-measure
- PR-curve
- Hold out method
- K-fold Cross Validation

- Hands-on exercise

overfitting vs. well-performing model

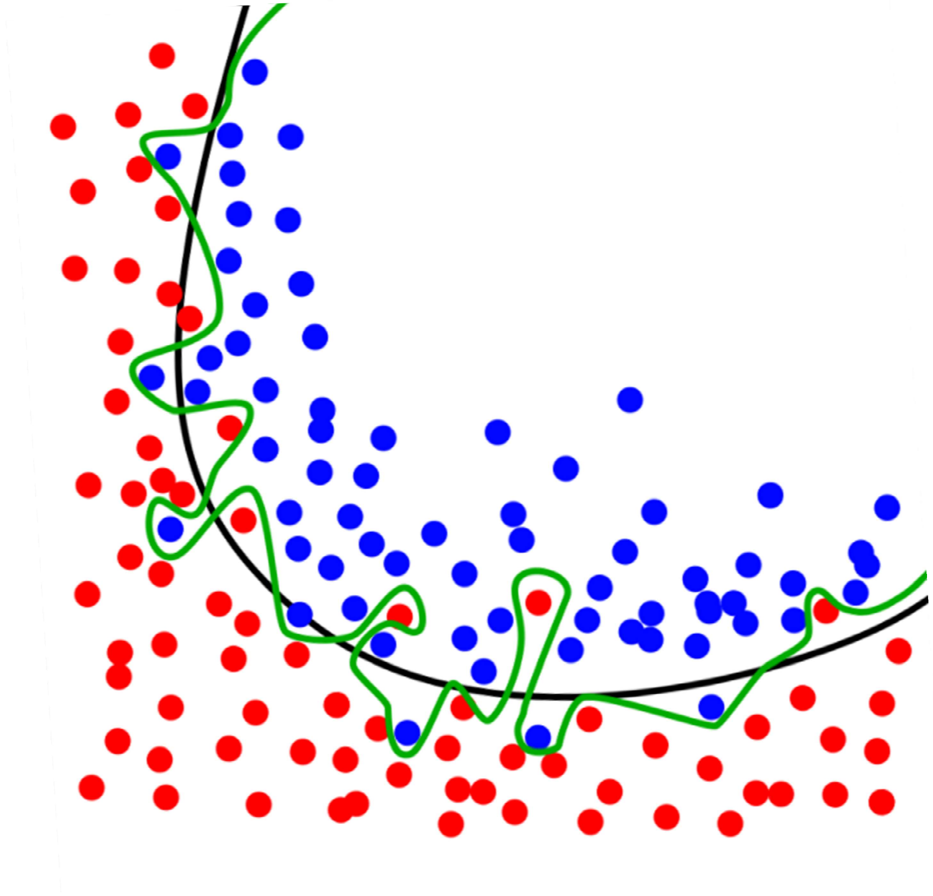
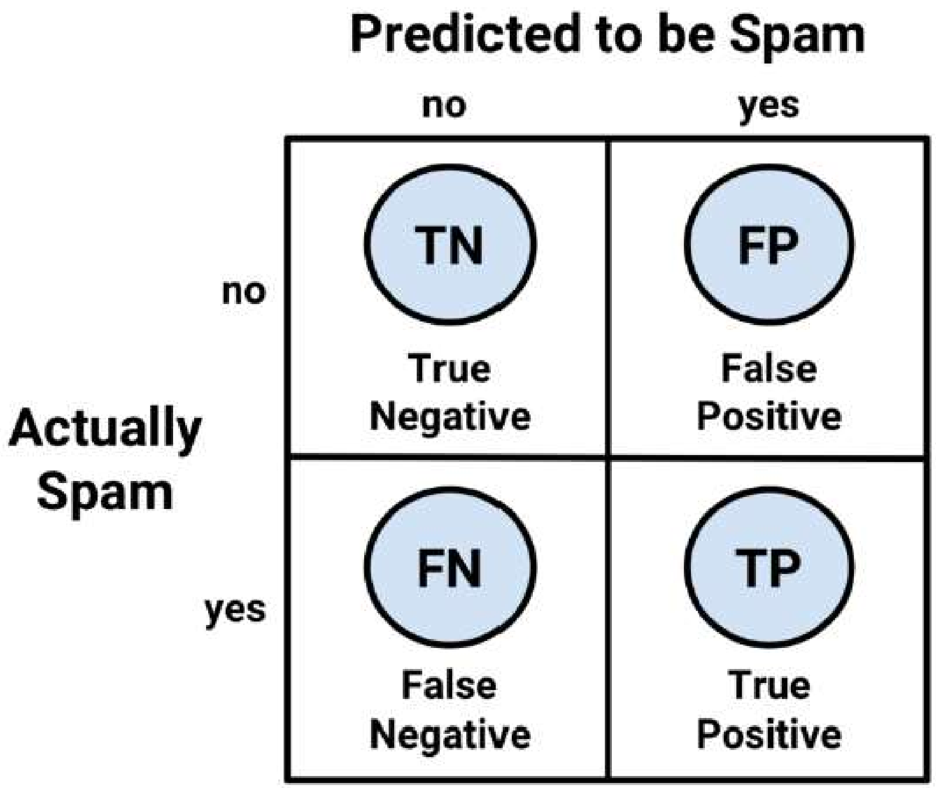
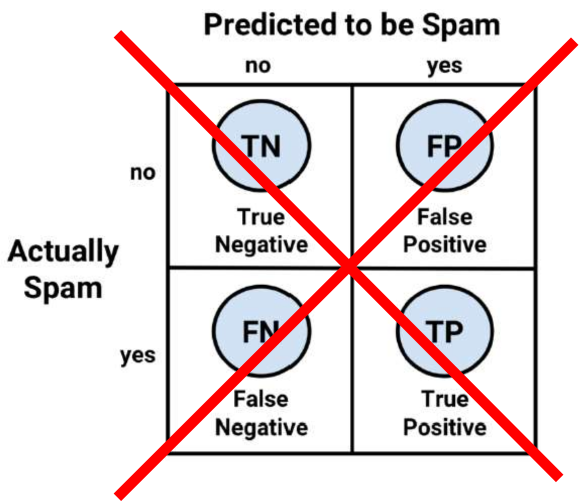


Image Credit: [Overfitting](#) by Chabacano

Confusion Matrix (spam example)





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

$$\text{error rate} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \text{accuracy}$$

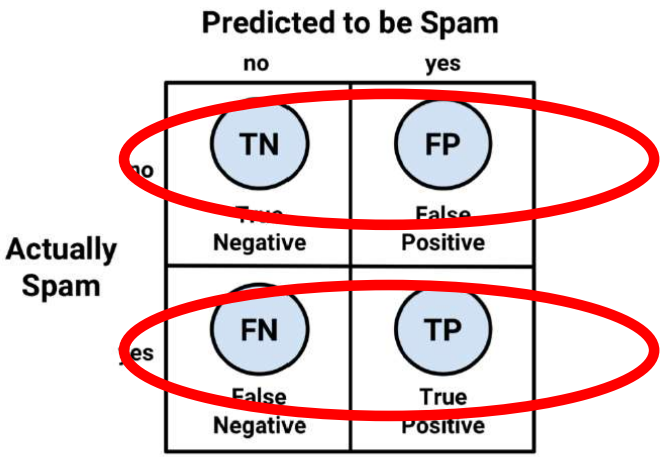
Kappa statistic

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Pr(a): actual agreement between classifier & true values

Pr(e): expected agreement by chance alone

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

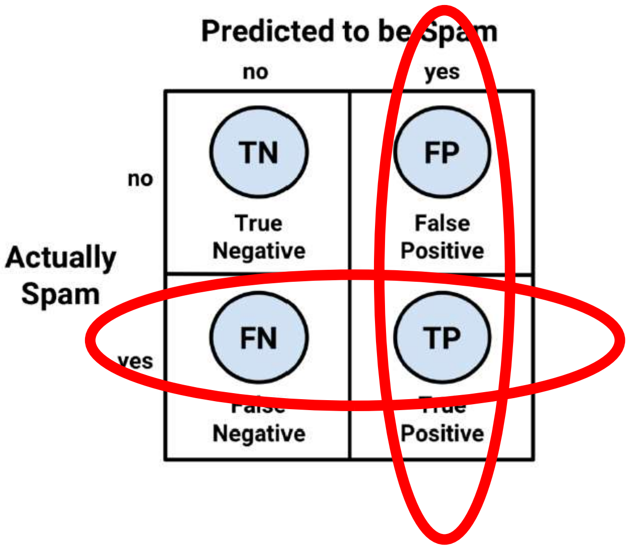


$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

(True Negative rate)

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

(True Positive rate)



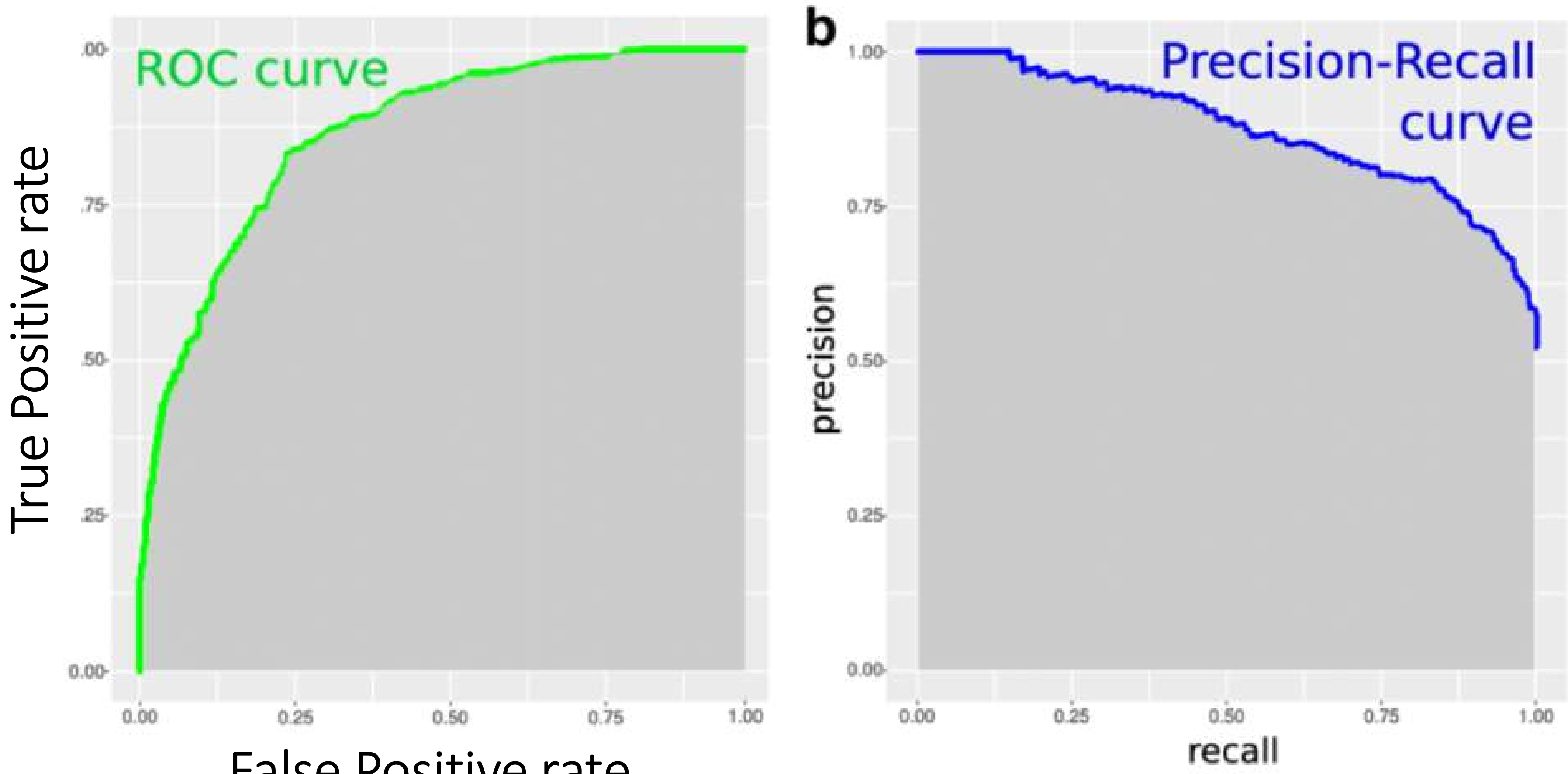
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

(*what proportion of **actual** positives are correctly classified?*)

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

(*what proportion of **predicted** positives are truly positive?*)

$$\text{F - measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$



False Positive rate

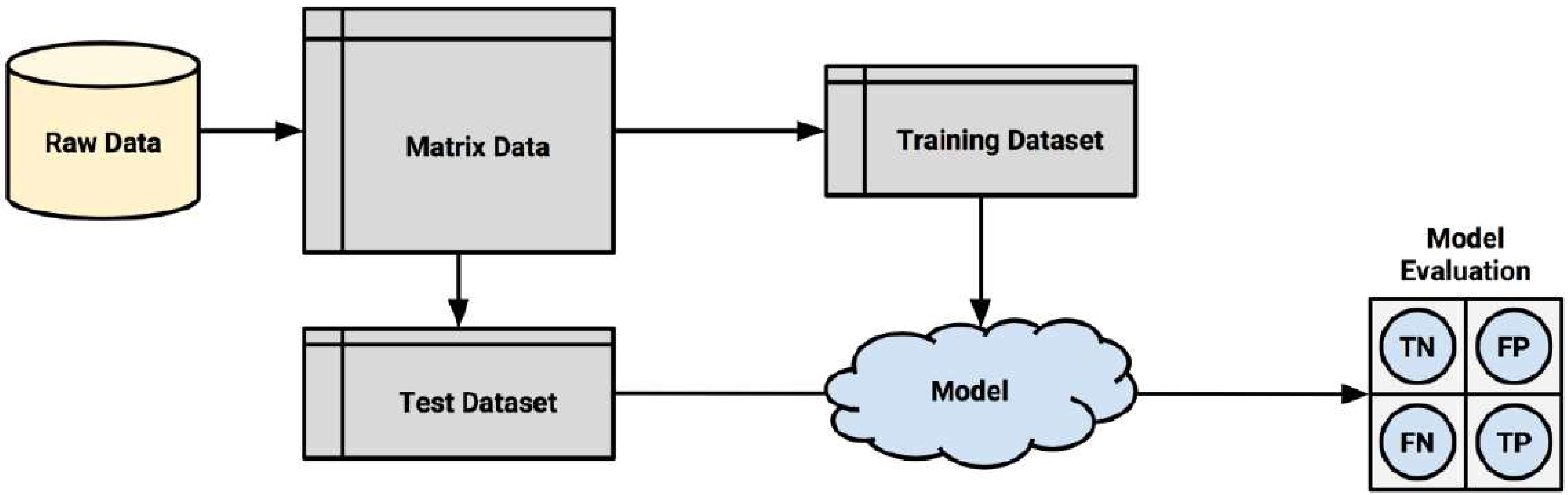
1 – specificity

sensitivity

caret package – performance evaluation

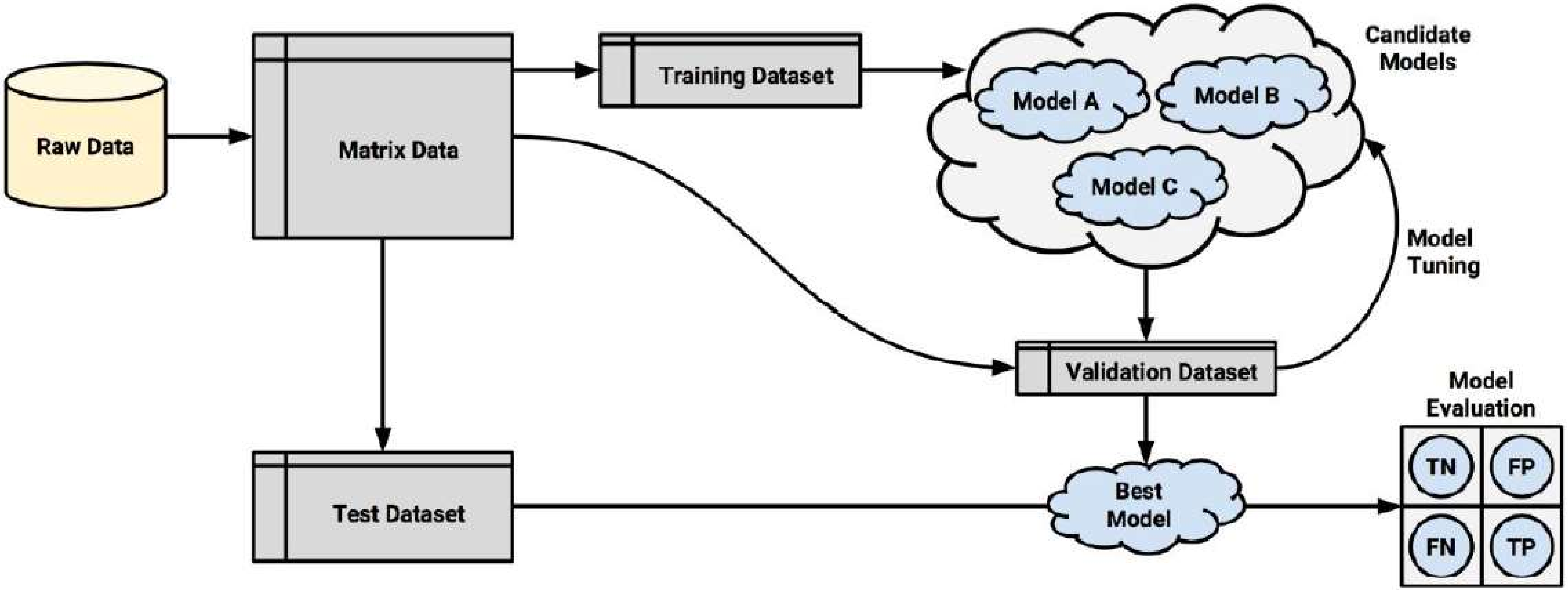
```
Confusion Matrix and Statistics

          Reference
Prediction  ham spam
      ham  1202   29
      spam    5  154

               Accuracy : 0.9755
                 95% CI : (0.966, 0.983)
    No Information Rate : 0.8683
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8867
 Mcnemar's Test P-Value : 7.998e-05

            Sensitivity : 0.8415
            Specificity : 0.9959
         Pos Pred Value : 0.9686
         Neg Pred Value : 0.9764
             Prevalence : 0.1317
         Detection Rate : 0.1108
   Detection Prevalence : 0.1144

       'Positive' Class : spam
```

Hold out method



Typically: 2/3 Training, 1/3 Test (70:30 - 80:20 training: test split)

Hold out with Validation



Typically: 50% Training, 25% Validation, 25% Test

Cross-validation

The repeated holdout is the basis of a technique known as **k-fold cross-validation** (or k-fold CV), which has become the industry standard for estimating model performance. But rather than taking repeated random samples that could potentially use the same record more than once, k-fold CV randomly divides the data into *k* completely separate random partitions called folds.

Standard k= 10