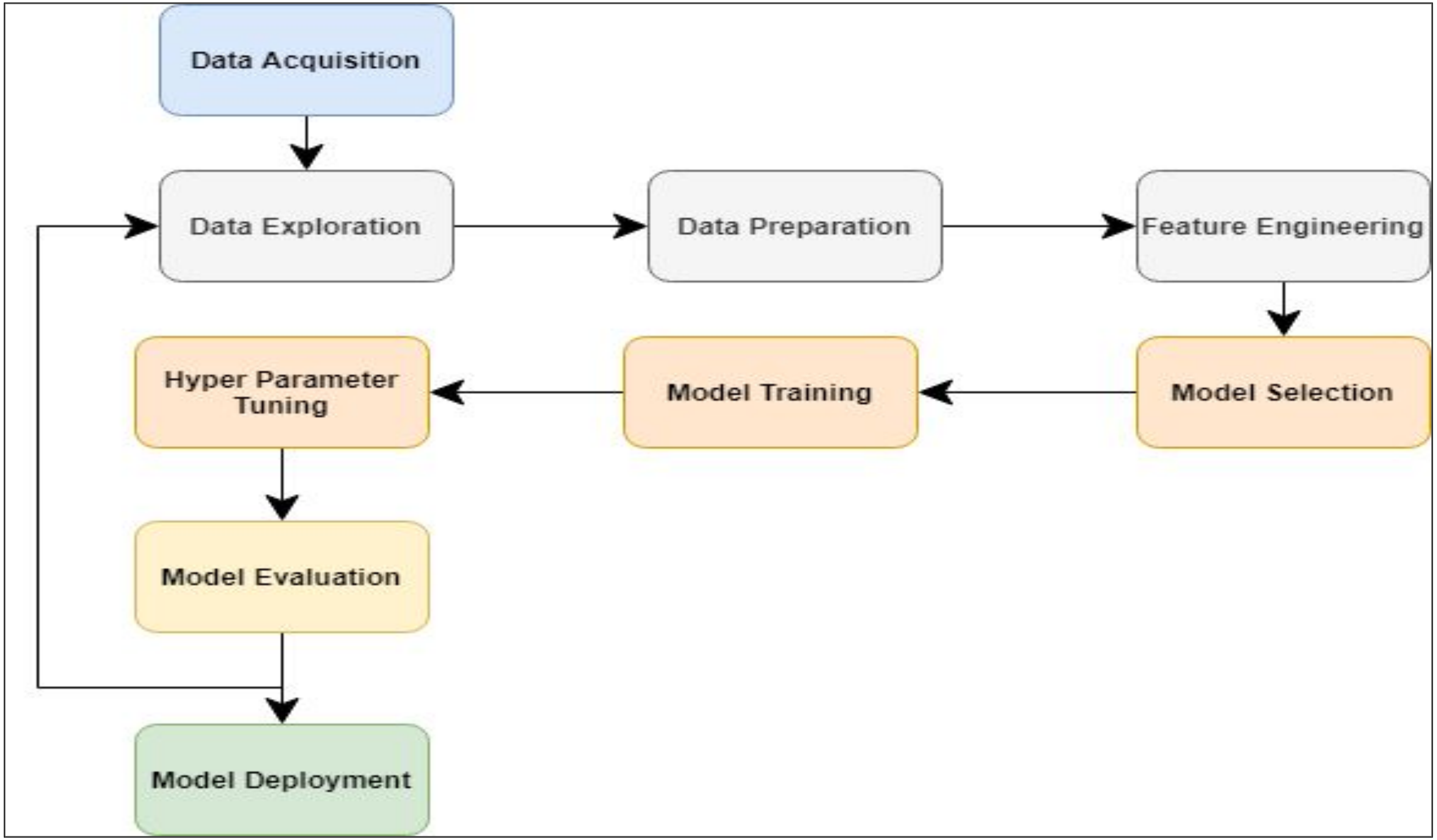
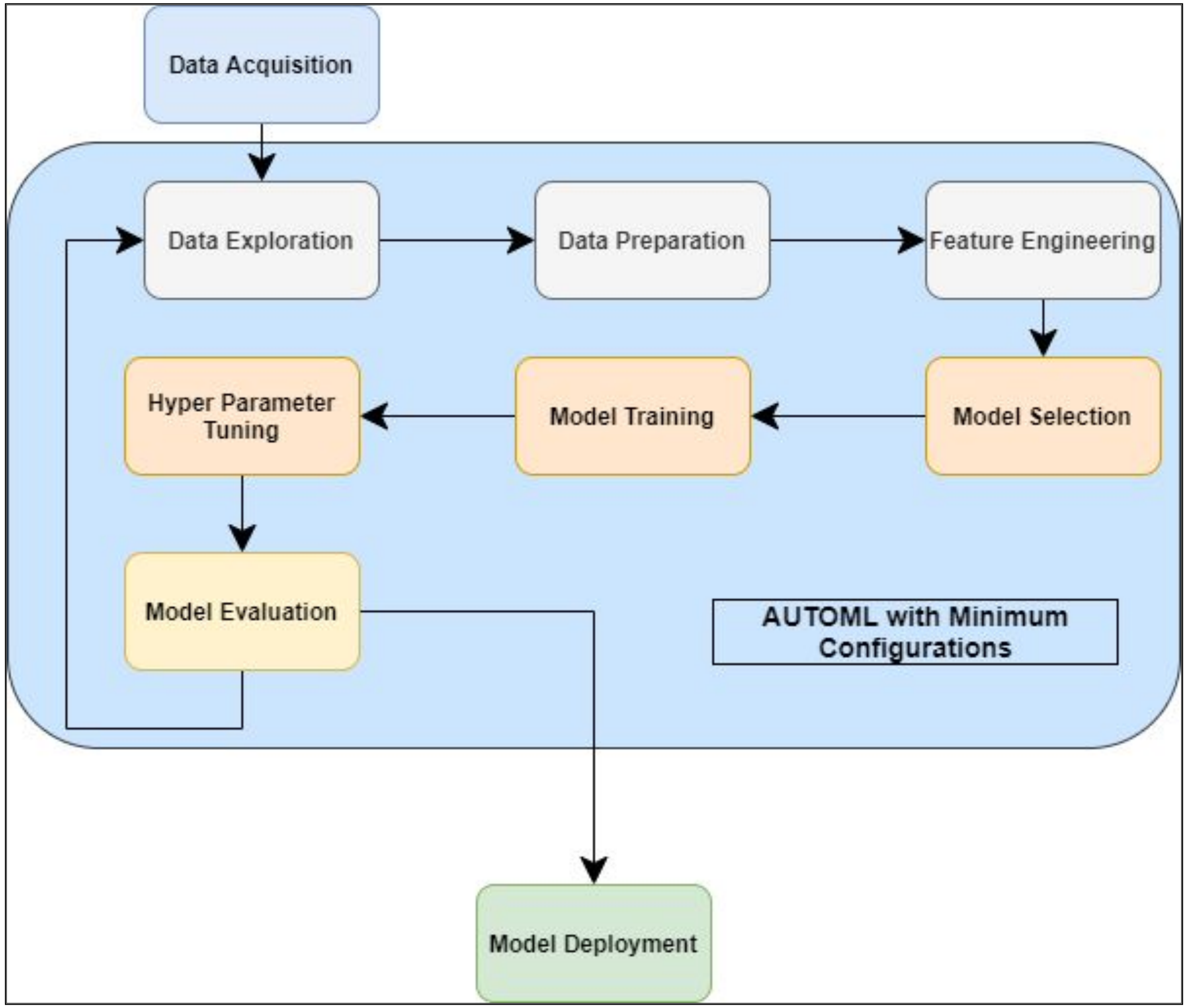
Supervised Learning

Data Mining Process



AutoML

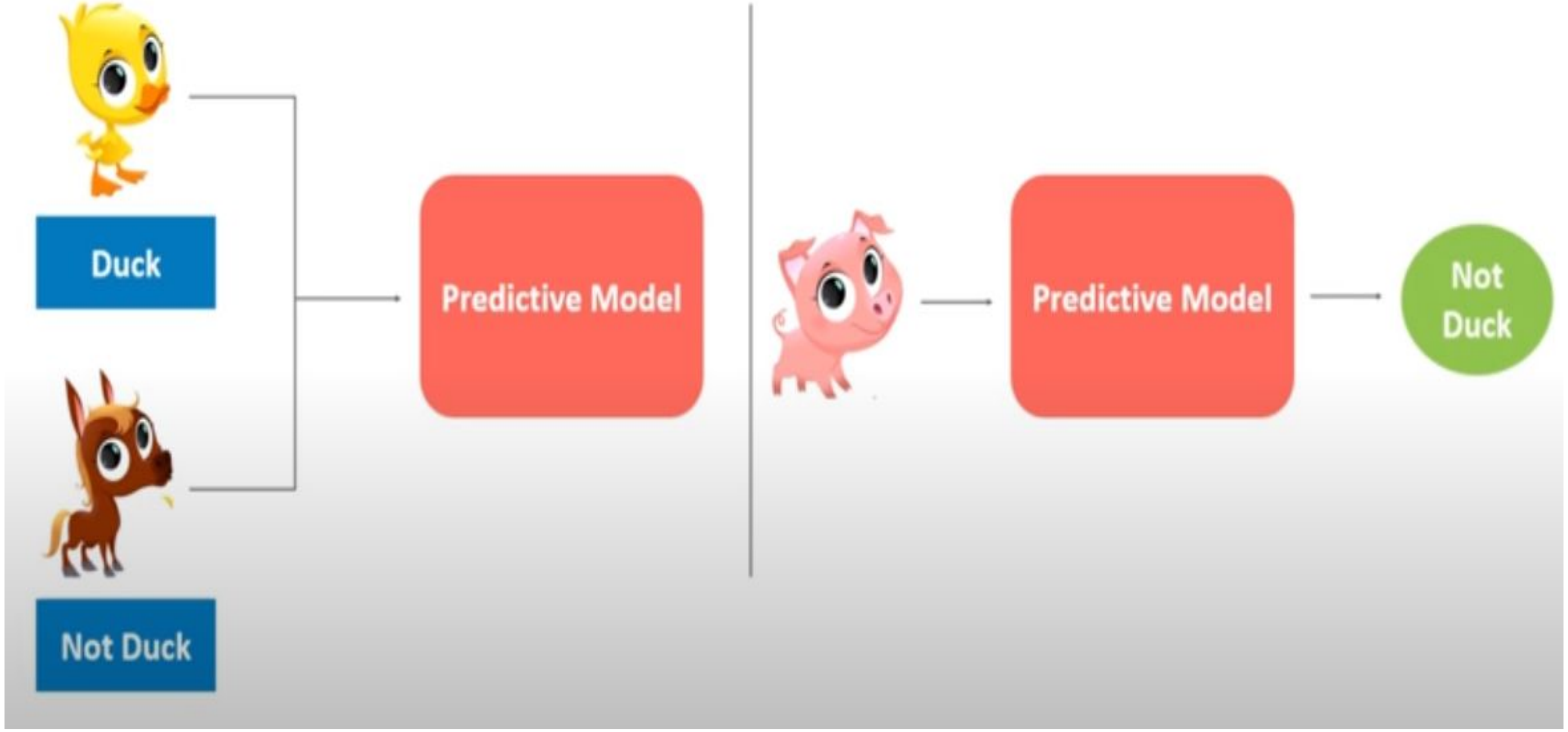


Data Acquisition
Data Exploration
Data Preparation
Feature Engineering
Hyper Parameter Tuning
Model Training
Model Selection
Model Evaluation
AUTOML with Minimum Configurations
Model Deployment

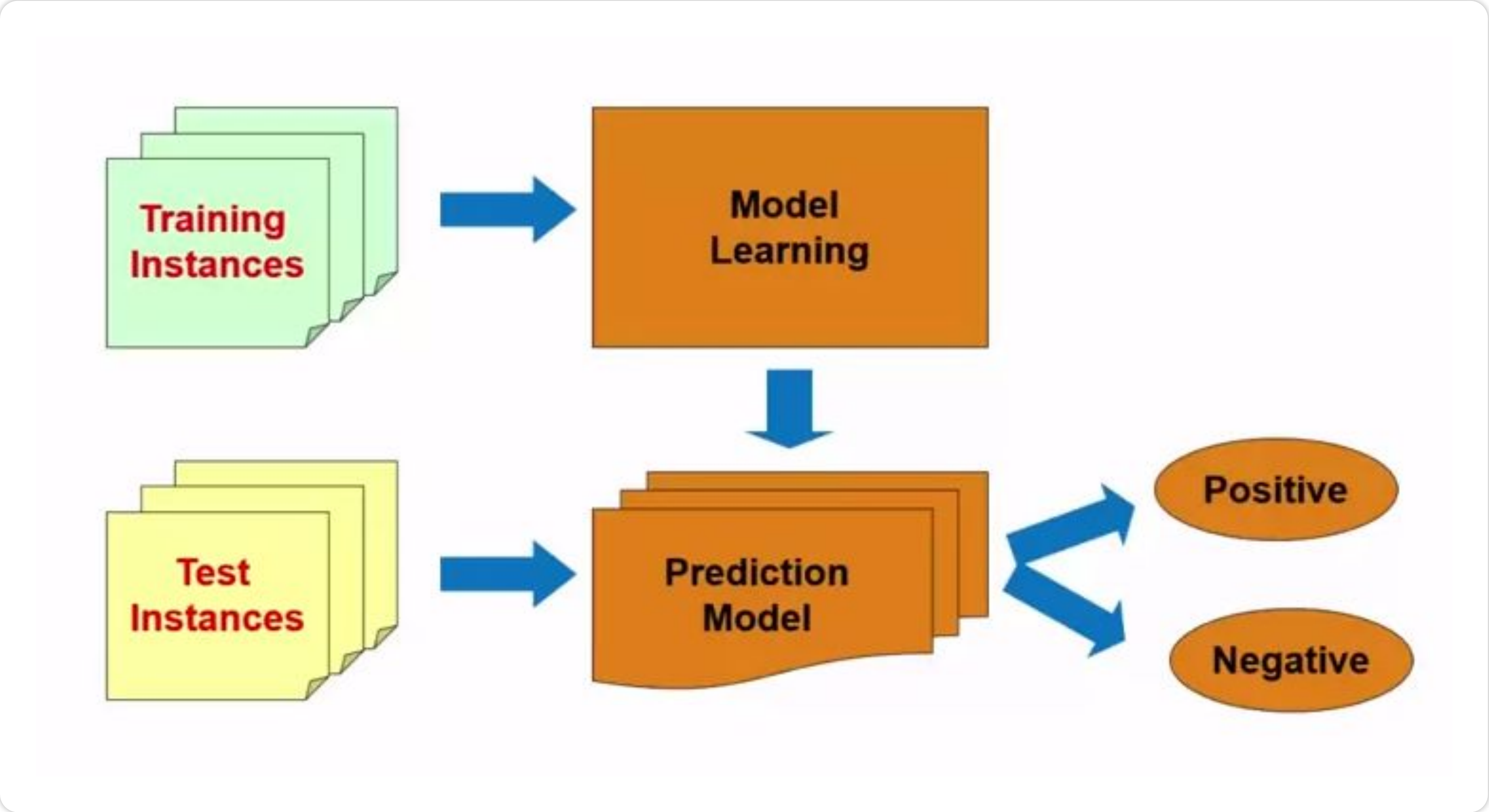
Supervised Learning

- The computer is provided with example inputs that are labeled with their desired outputs.
- Supervised learning uses patterns to predict label values on additional unlabeled data.
- An algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water.
- By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

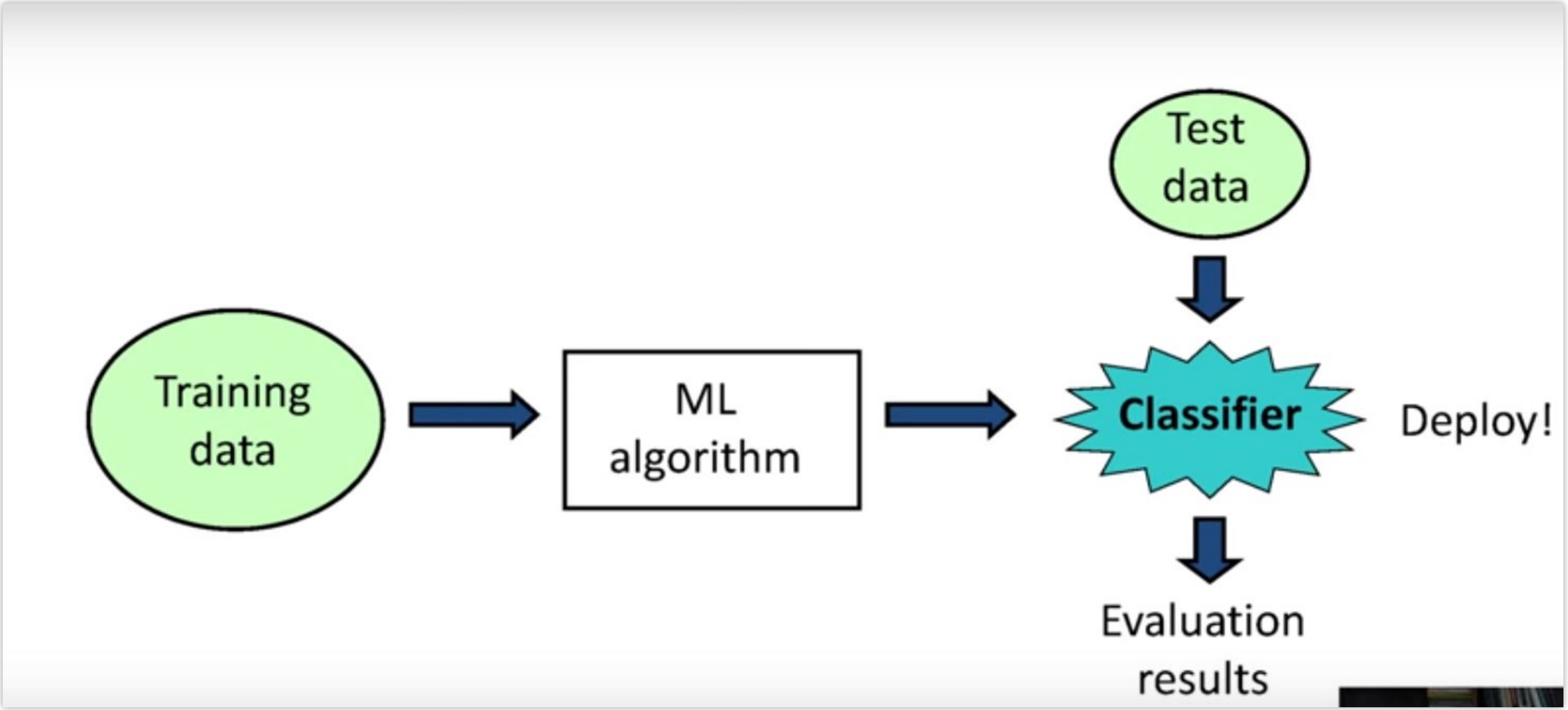
Supervised Learning



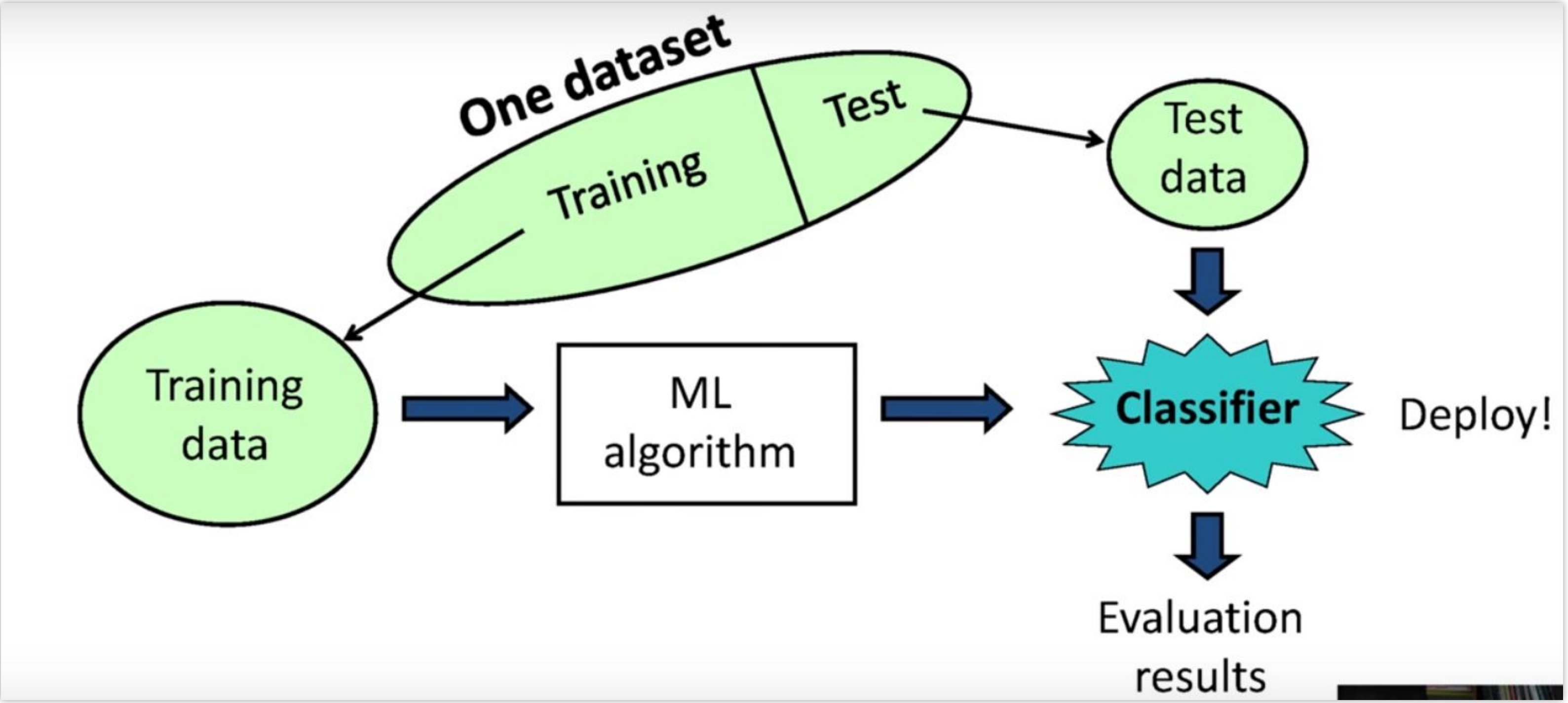
What is Classification



Training & Testing

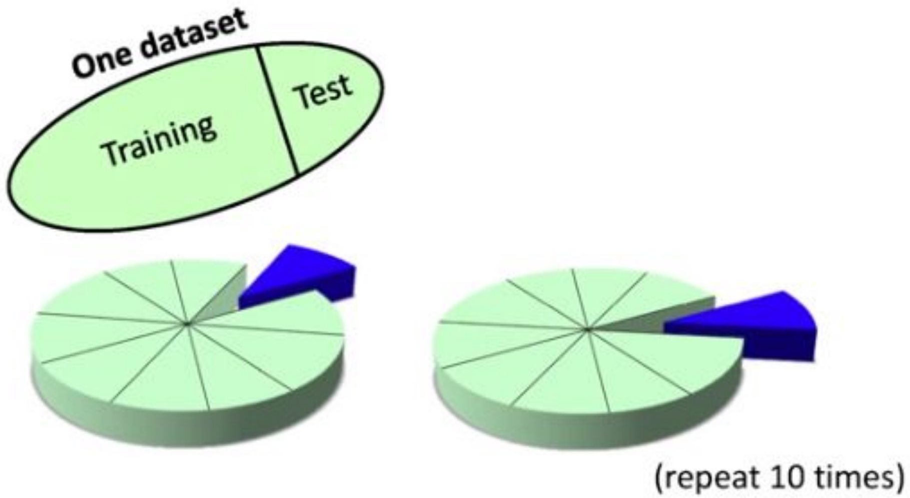


Percentage Split



Cross Validation

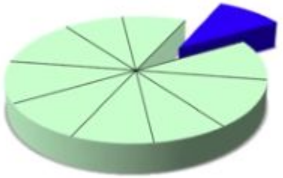
- Repeated holdout
  (in Lesson 2.3, hold out 10% for testing, repeat 10 times)



Cross Validation

10-fold cross-validation

- Divide dataset into 10 parts (folds)
- Hold out each part in turn
- Average the results
- Each data point used once for testing, 9 times for training



Stratified cross-validation

- Ensure that each fold has the right proportion of each class value

In Simple Terms?

Age 30

Cars 2

Buy a Bike
Probability 65%

Age 60

Cars 1

Buy a Bike
Probability 20%

Iris-setosa	Iris-versicolor	Iris-virginica

<table>
<tr><td colspan="2"></td><td colspan="2">True condition</td><td colspan="3"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td rowspan="2" colspan="2"></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Accuracy of NB

- Error Rate
  - (T - C) / T
    - T – Number of Objects
    - C – correctly classified objects.

- Confusion matrix

Predicated Class	True Class		
	A	**B**	**C**
A	**8**	1	1
B	2	**9**	2
C	0	0	**7**

Classification Performance Measures

- Recall = a/(a+c) where a + c > 0 (o.w. undefined).
  - Did we find all of those that belonged in the class?

- Precision = a/(a+b) where a+b>0 (o.w. undefined).
  - Of the times we predicted it was “in class”, how often are we correct?

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

F1 Score / F1 Measure

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Matthews Correlation Coefficient

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC doesn't depend on which class is the positive one, which has the advantage over the F1 score to avoid incorrectly defining the positive class.

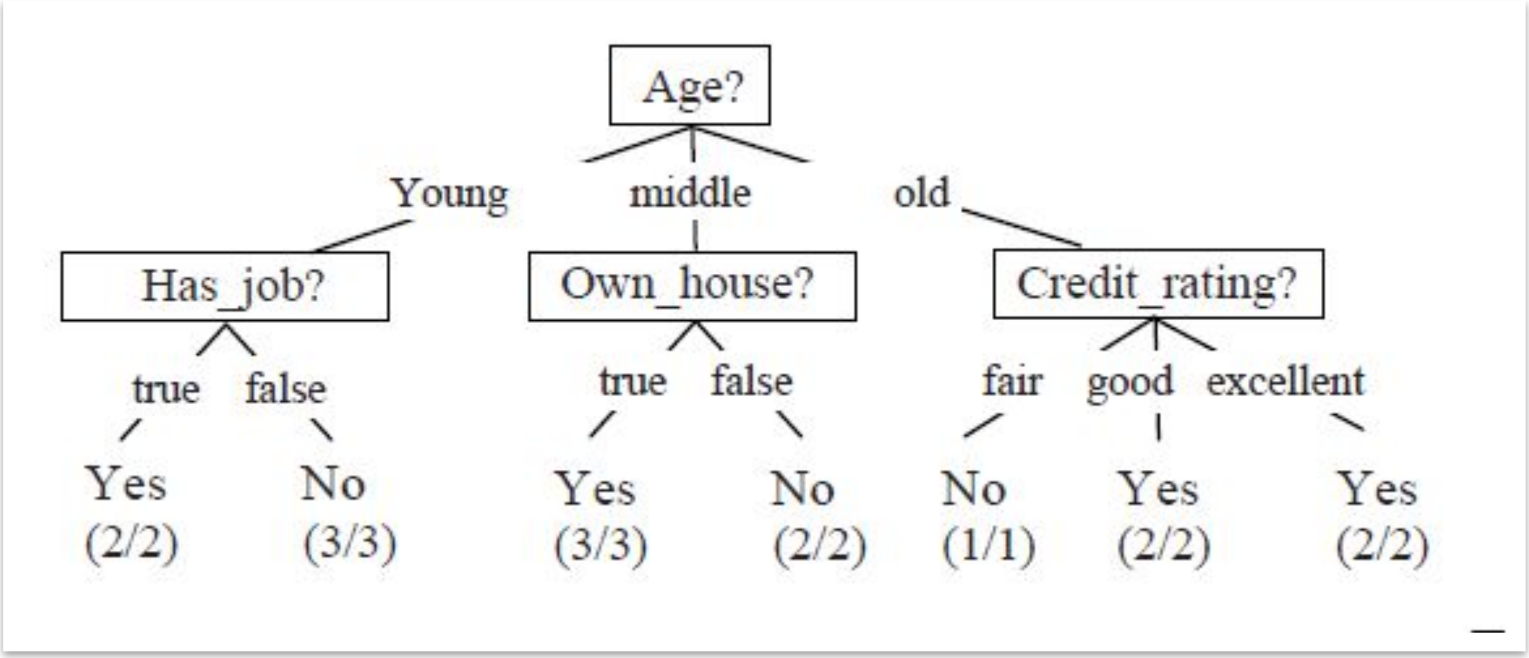
- False Positive (FP)
  - FP cases are those that did not belong to a class but allocated to it.
- False Negative (FN)
  - FN are cases that belong to a class but were not allocated to it.

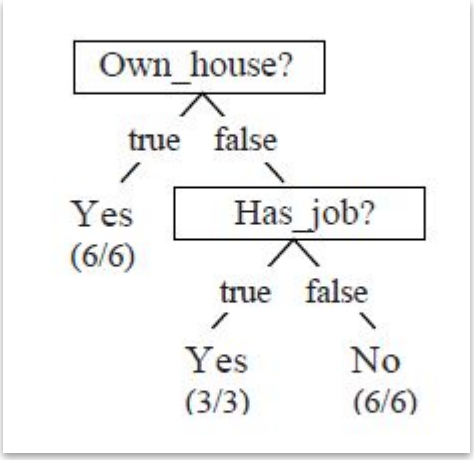
	FP	FN
Class 1	2	2
Class 2	4	1
Class 3	0	3

- Sensitivity = TP / (TP + FN)
- Specificity = TN/(TN + FP)

Decision Trees

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**


Age?
Young
middle
old
Has_job?
Own_house?
Credit_rating?
true
false
true
false
fair
good
excellent
Yes
(2/2)
No
(3/3)
Yes
(3/3)
No
(2/2)
No
(1/1)
Yes
(2/2)
Yes
(2/2)


Own_house?
true
false
Yes
(6/6)
Has_job?
true
false
Yes
(3/3)
No
(6/6)