Introduction to Data Science

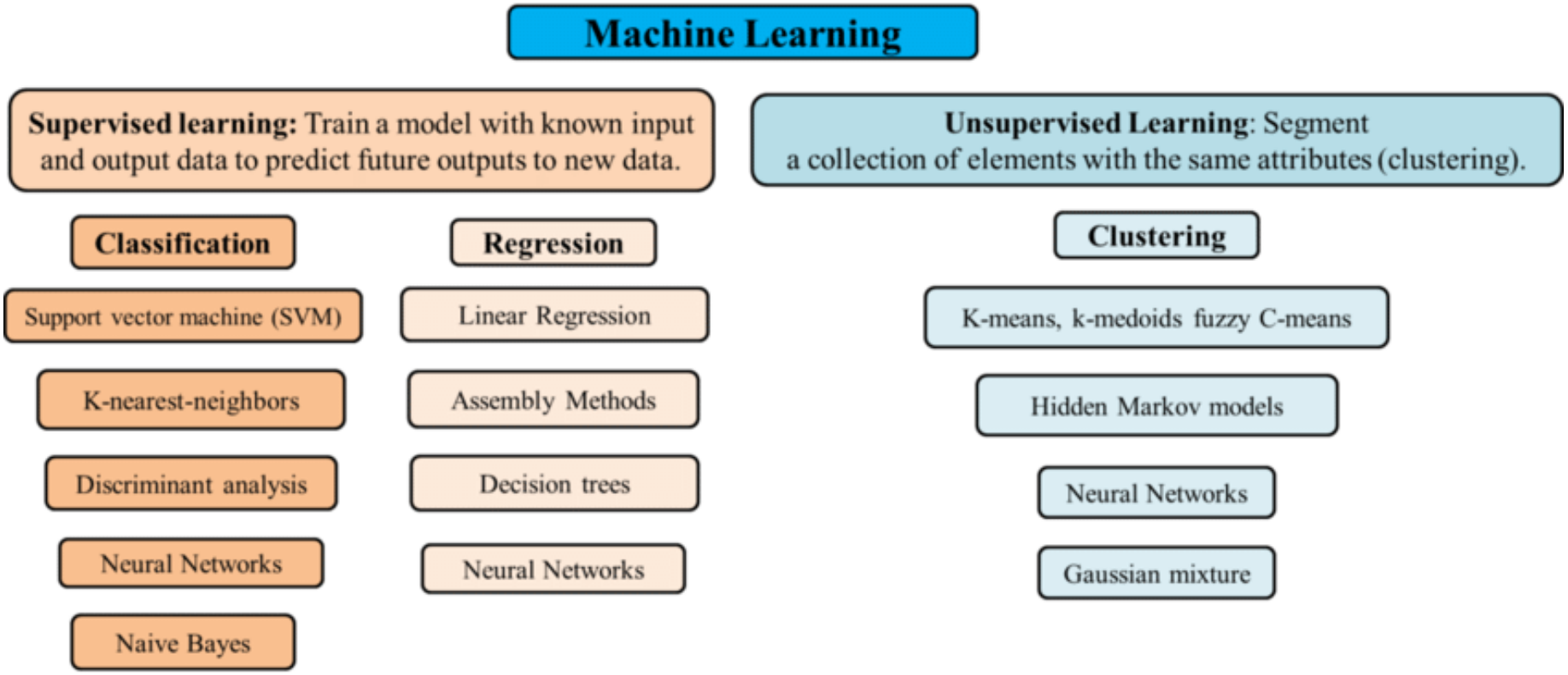
Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 11: March 25 - 29, 2024)

Outline

- Confusion Matrix
- Naïve Bayes

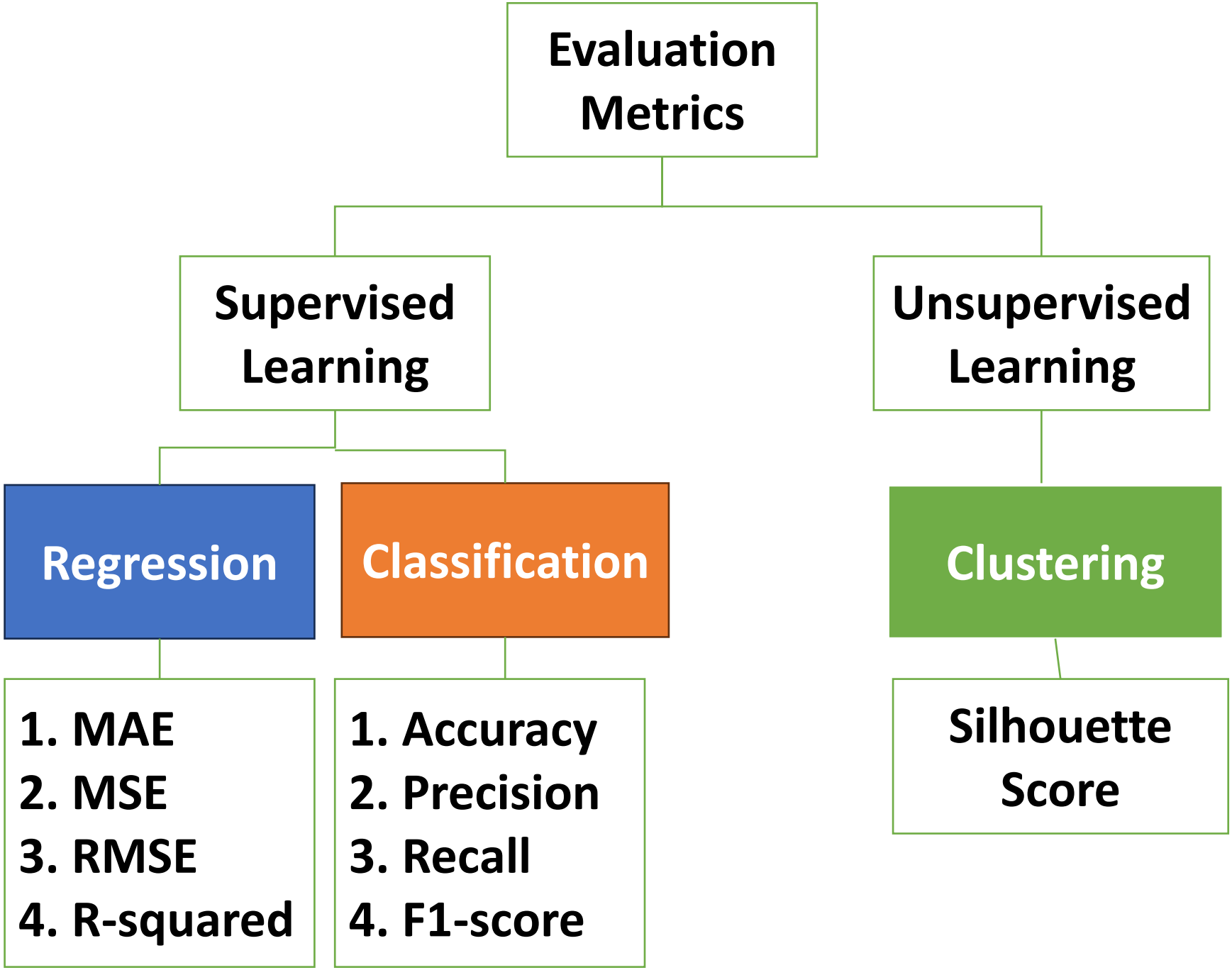
Machine Learning Algorithms



Evaluation Metric

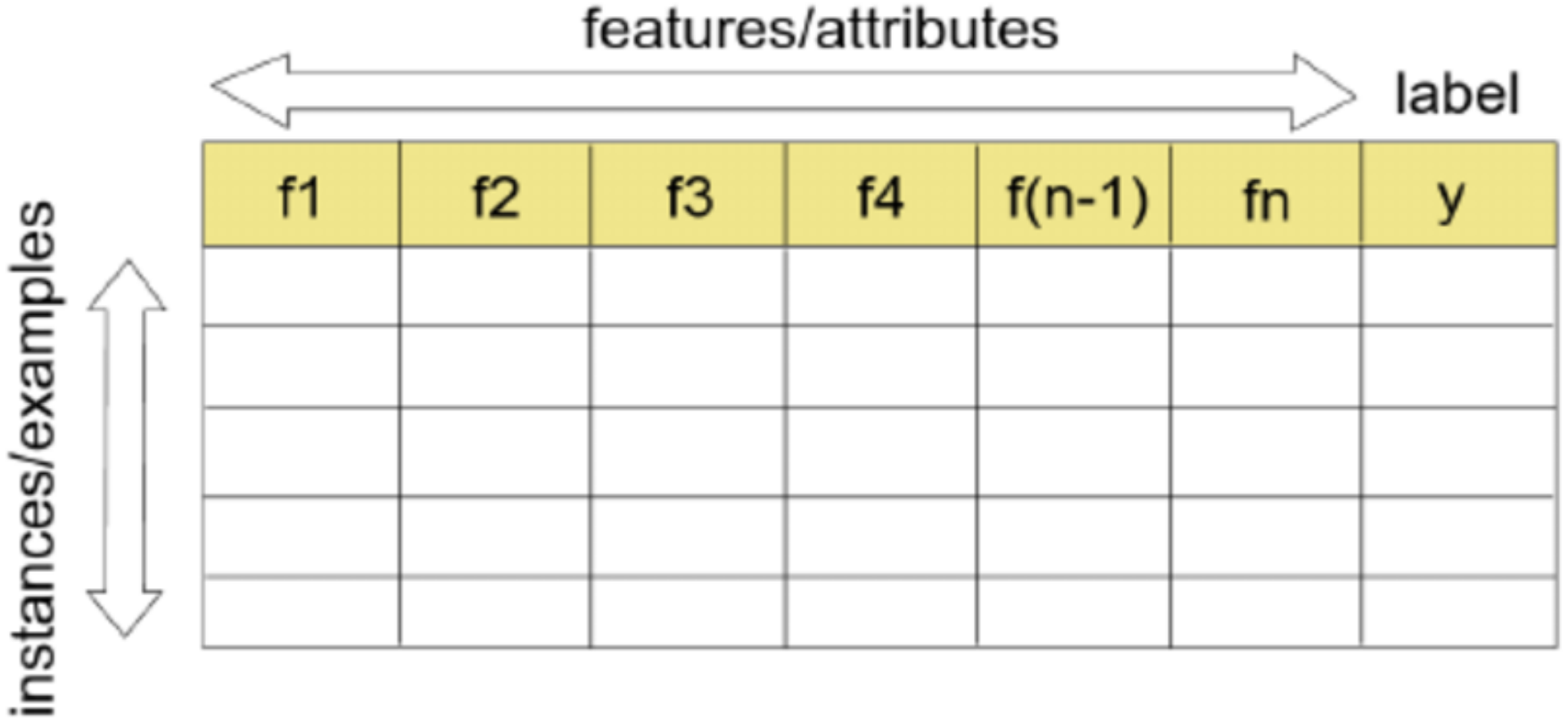
- Evaluation metrics are quantitative measures used to assess the performance and effectiveness of a statistical or machine learning model.

- To evaluate the performance or quality of a Machine Learning model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics.

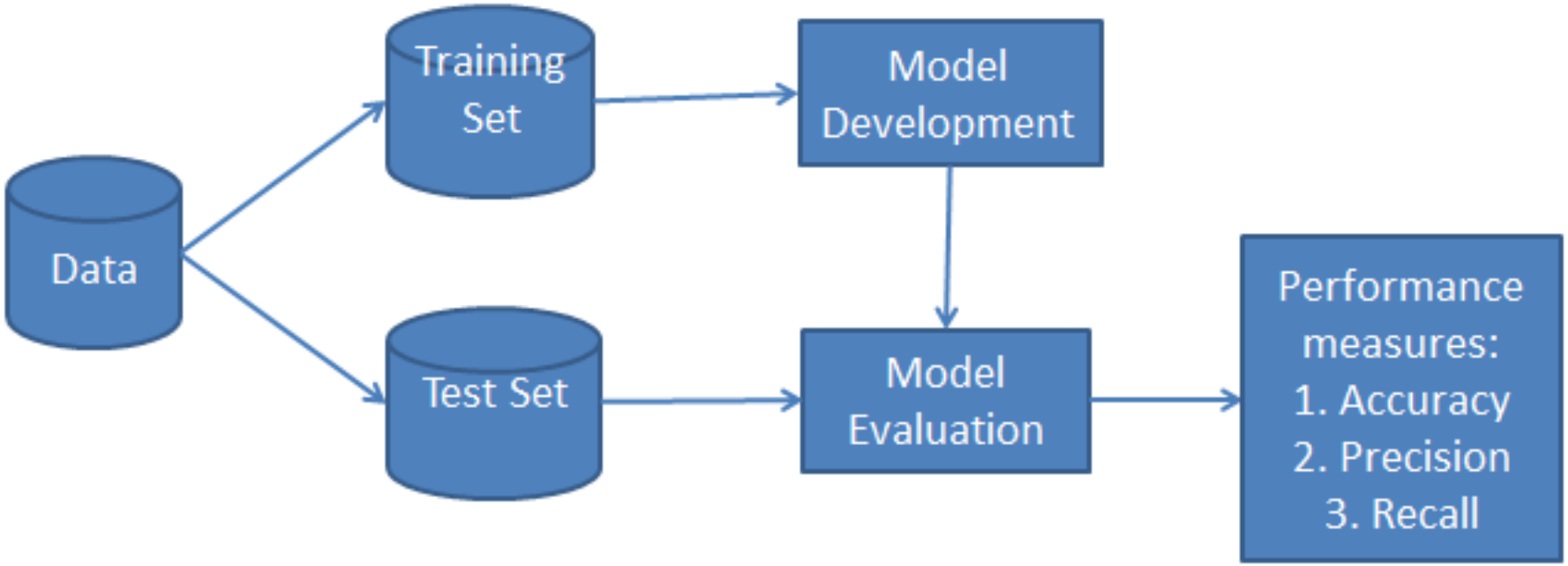

Evaluation Metrics
Supervised Learning
Unsupervised Learning
Regression
Classification
Clustering
1. MAE
2. MSE
3. RMSE
4. R-squared
1. Accuracy
2. Precision
3. Recall
4. F1-score
Silhouette Score

Supervised vs. Unsupervised Machine Learning

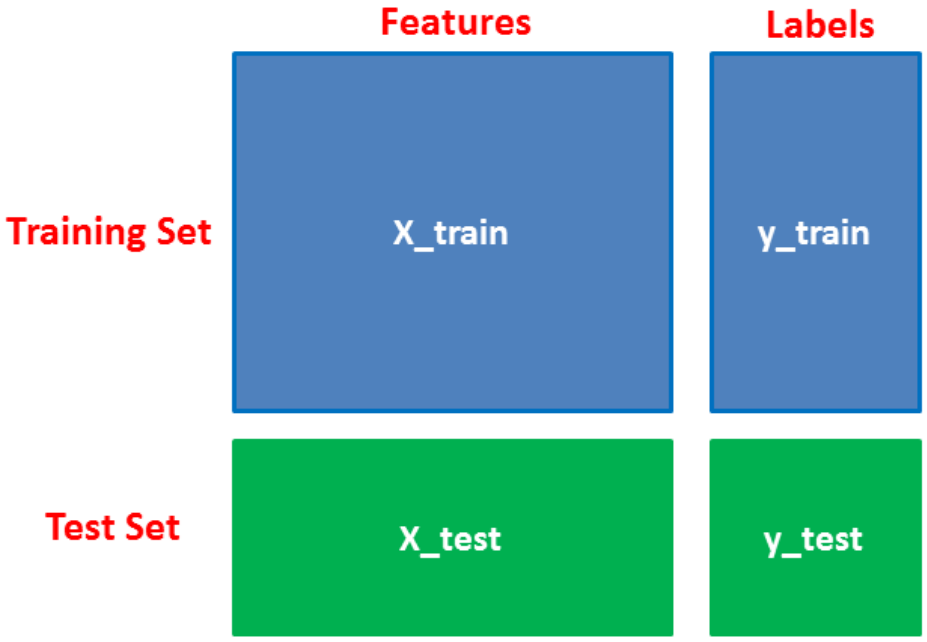
Supervised Machine Learning

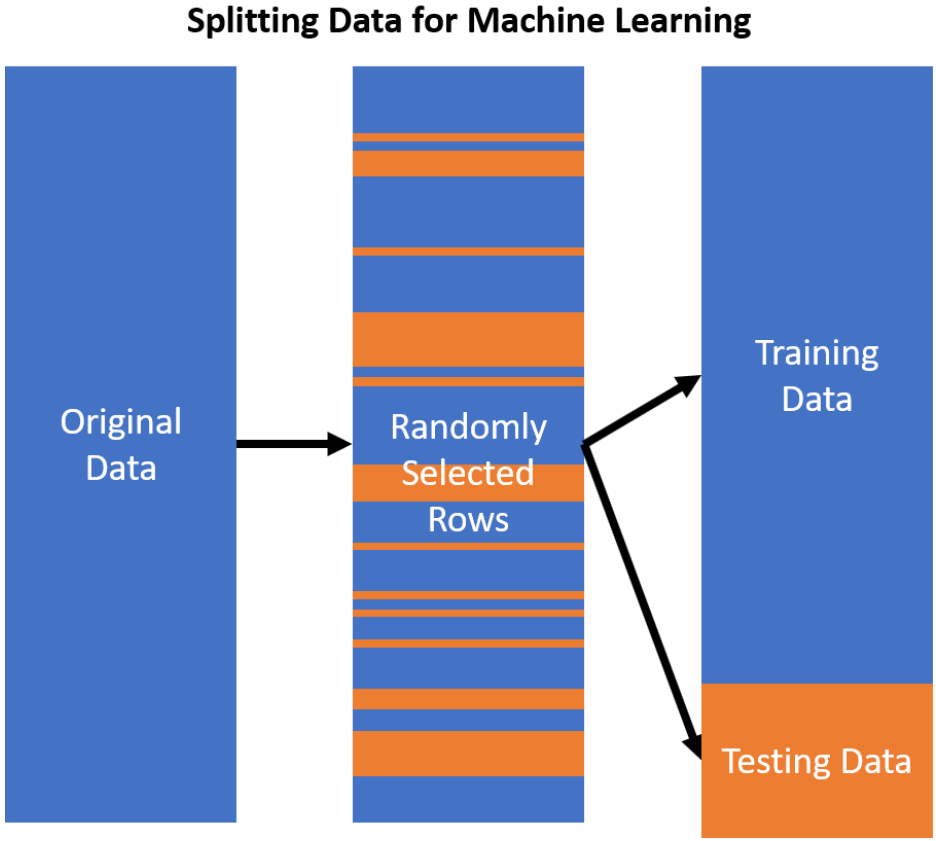


Train-Test-Split



Train-Test-Split





TRAIN_TEST_SPLIT SPLITS DATA INTO TRAINING DATA AND TEST DATA

Original Data

X_1	X_2	X_p	Y

`train_test_split()`

X_train

X_1	X_2	X_p

y_train

Y

X_test

X_1	X_2	X_p

y_test

Y

Classification Model

- Classification is a technique where we categorize data into a given number of classes.

- The main goal of a classification problem is to identify the category/class to which a new data will fall under.

Classification Model

- **Classifier:** An algorithm that maps the input data to a specific category.
- **Classification model:** A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
- **Feature:** A feature is an individual measurable property of a phenomenon being observed.
- **Binary Classification:** Classification task with two possible outcomes. Eg: Gender classification (Male / Female)
- **Multi-class classification:** Classification with more than two classes. In multi class classification each sample is assigned to one and only one target label. Eg: An animal can be cat or dog but not both at the same time
- **Multi-label classification:** Classification task where each sample is mapped to a set of target labels (more than one class). Eg: A news article can be about sports, a person, and location at the same time.

Classification Model

The following are the steps involved in building a classification model:

- **Initialize** the classifier to be used.
- **Train the classifier:** All classifiers in scikit-learn uses a fit(X, y) method to fit the model(training) for the given train data X and train label y.
- **Predict the target:** Given an unlabeled observation X, the predict(X) returns the predicted label y.
- **Evaluate** the classifier model

Evaluating a Classification Model

```
> source('E:/Spring2021/RProgs/SpamFilter.R')
Loading required package: RColorBrewer
Loading required package: NLP
ham --> ham
 spam --> spam
 ham --> ham
 ham --> spam
 ham --> ham
 ham --> ham
 ham --> ham
 ham --> spam
 ham --> ham
 ham --> ham
[1] "Done"
```

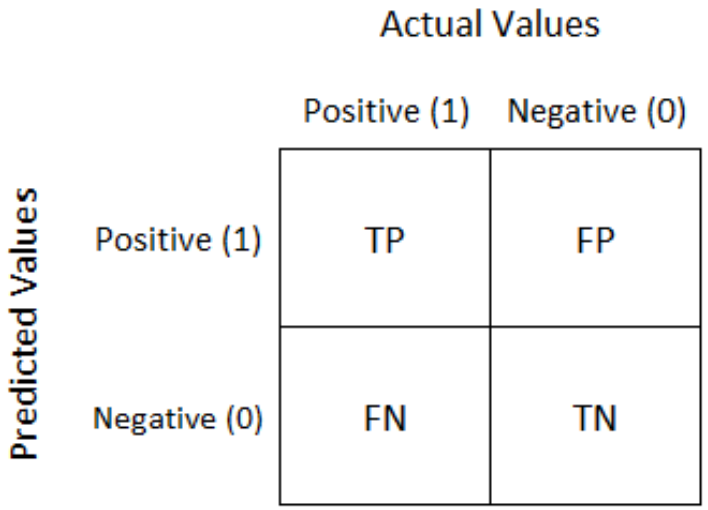
Actual

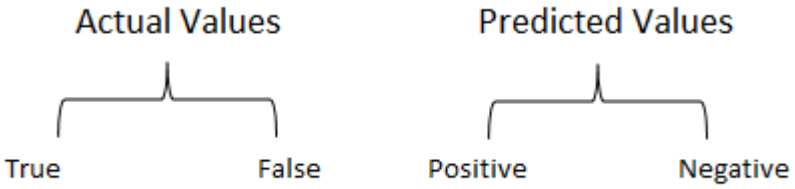
Predicted

Confusion Matrix

- A confusion matrix is a table that is often used to describe the **performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

Confusion Matrix



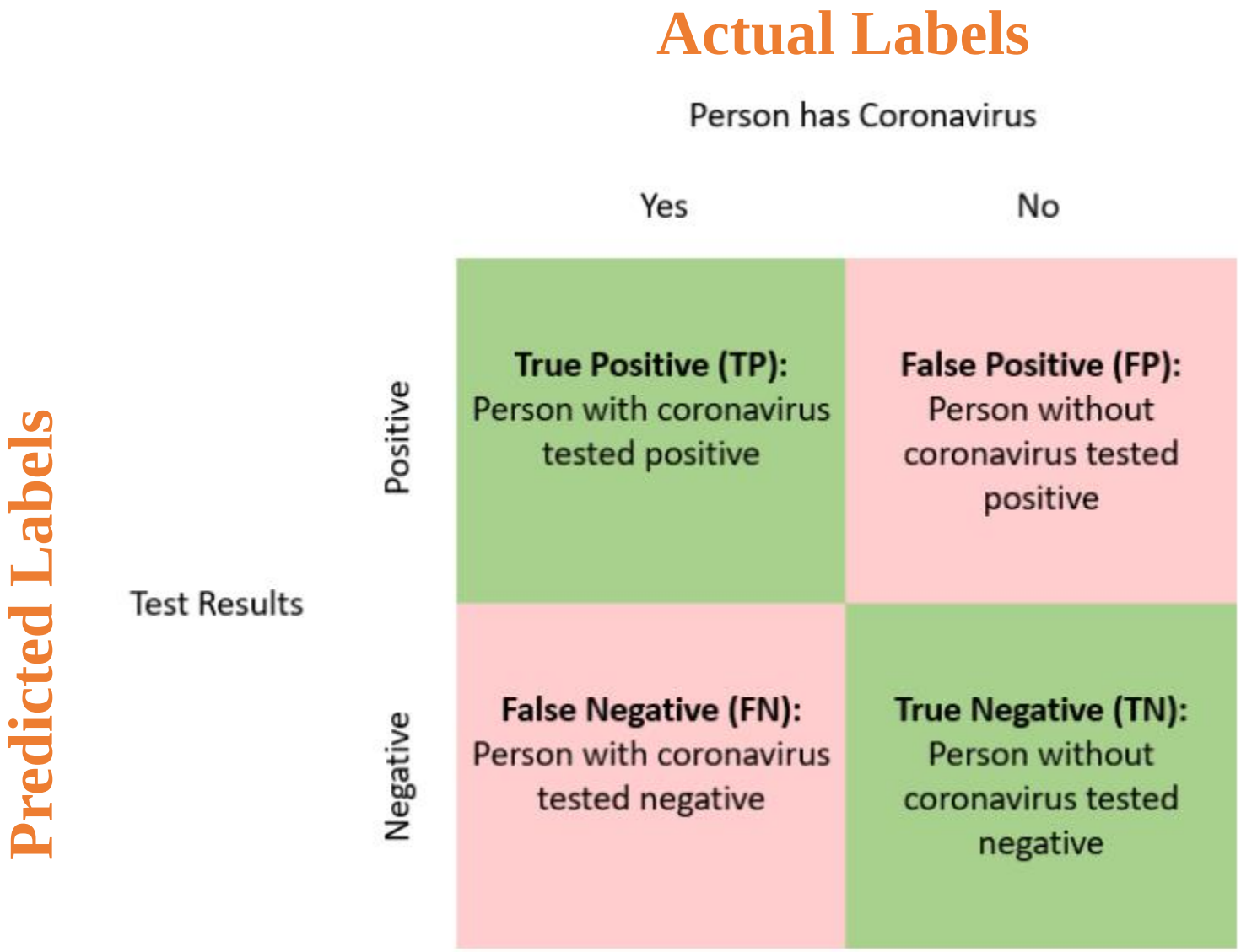


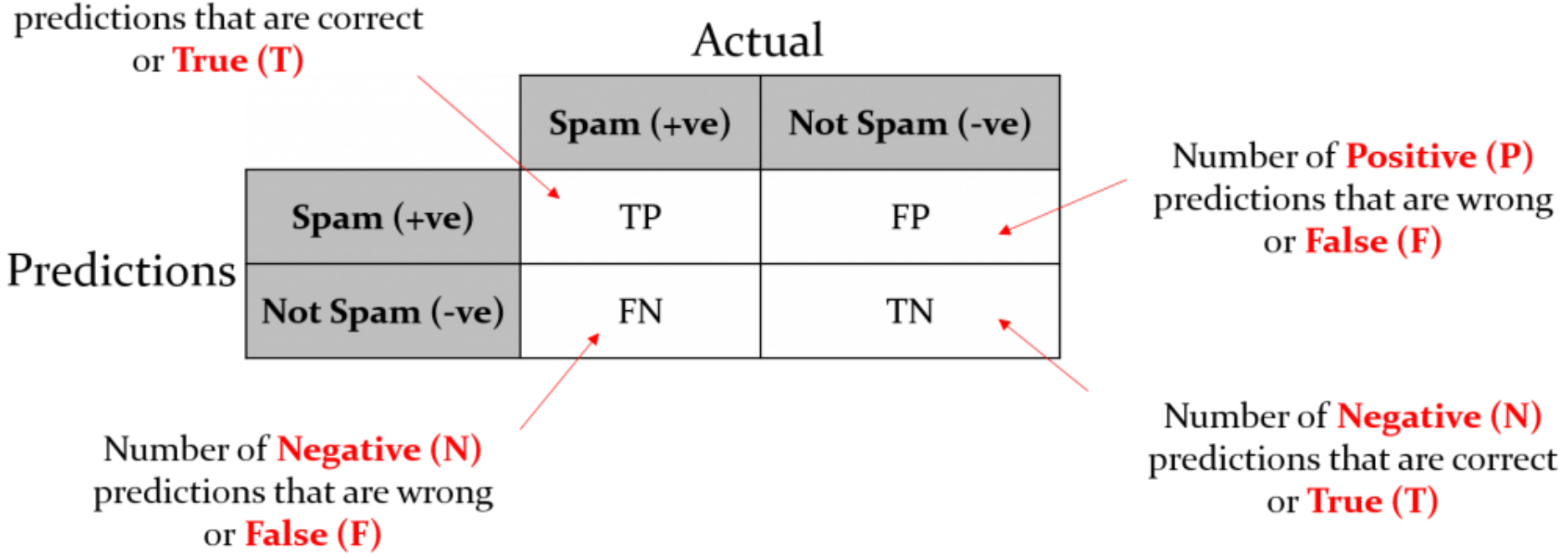
True Positive (TP): You predicted positive and it's true.

True Negative (TN): You predicted negative and it's true.

False Positive (FP): You predicted positive and it's false.

False Negative (FN): You predicted negative and it's false.


Actual Labels
Person has Coronavirus
Yes
No
Predicted Labels
Test Results
Positive
Negative
True Positive (TP): Person with coronavirus tested positive
False Positive (FP): Person without coronavirus tested positive
False Negative (FN): Person with coronavirus tested negative
True Negative (TN): Person without coronavirus tested negative


Number of Positive (P) predictions that are correct or True (T)
Actual
Spam (+ve)
Not Spam (-ve)
Predictions
Spam (+ve)
Not Spam (-ve)
TP
FP
FN
TN
Number of Positive (P) predictions that are wrong or False (F)
Number of Negative (N) predictions that are wrong or False (F)
Number of Negative (N) predictions that are correct or True (T)

Confusion Matrix Terminology

- Classification **Accuracy** is the ratio of correct predictions to total predictions made.

$$\text{Accuracy} = \frac{Correct\ Predictions}{Total\ Predictions}$$

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

Confusion Matrix Terminology

- **Precision** is calculated as the number of correct positive predictions divided by the total number of positive predictions.

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

- **Recall** is calculated as the number of correct positive predictions divided by the total number of positives.

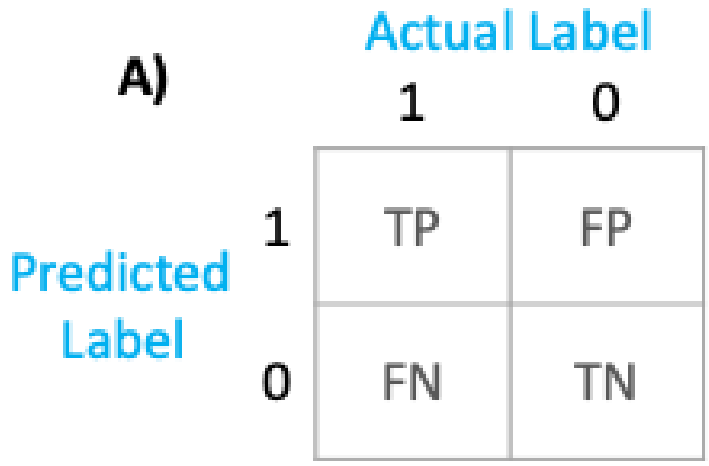
$$\text{Recall} = \frac{TP}{TP + FN}$$

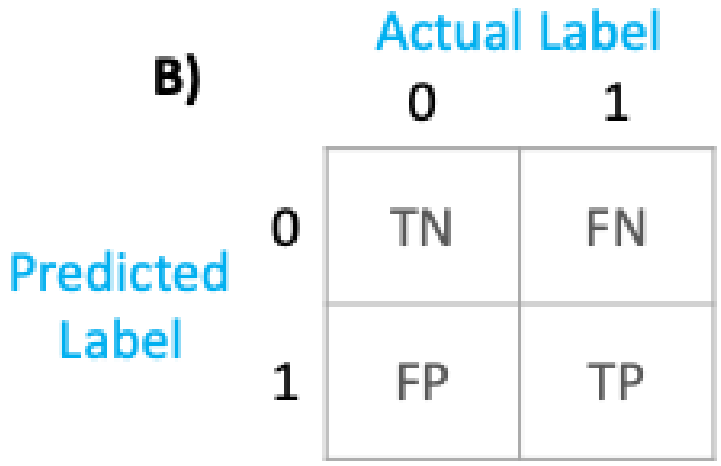
Confusion Matrix Terminology

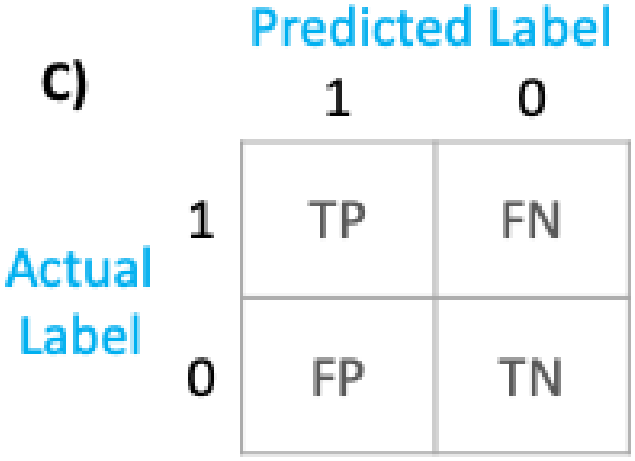
- **F1-score** is the harmonic mean of precision and recall and is a better measure than accuracy.

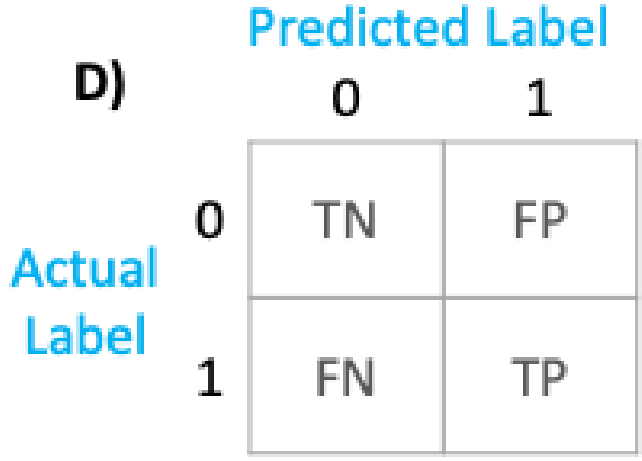
$$\text{F1-score} = \frac{2 * Recall * Precision}{Recall + Precision}$$

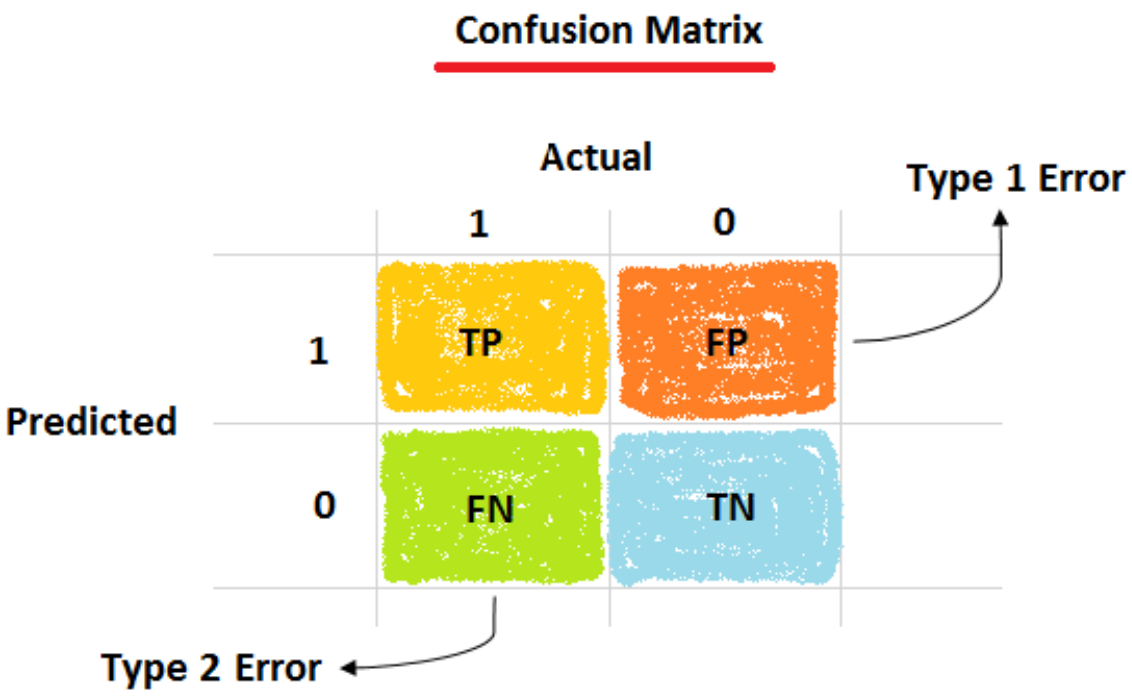
Confusion Matrix Variations











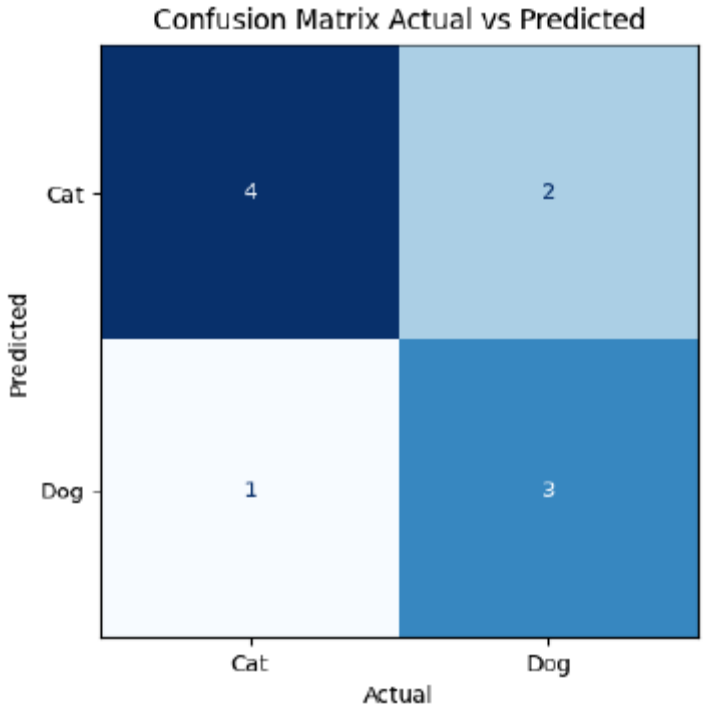
Type I Error (False Positive):— Fire alarm rings when there is no fire.

Type II Error (False Negative):— Fire alarm fails to ring when there is fire.

Confusion Matrix Example

a) The output of a machine learning classifier is given below in the form of actual and predicted data. Draw the Confusion Matrix of this classifier and calculate its accuracy.

Actual	Dog	Dog	Cat	Dog	Cat	Cat	Cat	Dog	Dog	Cat
Predicted	Cat	Dog	Cat	Dog	Dog	Cat	Cat	Dog	Cat	Cat

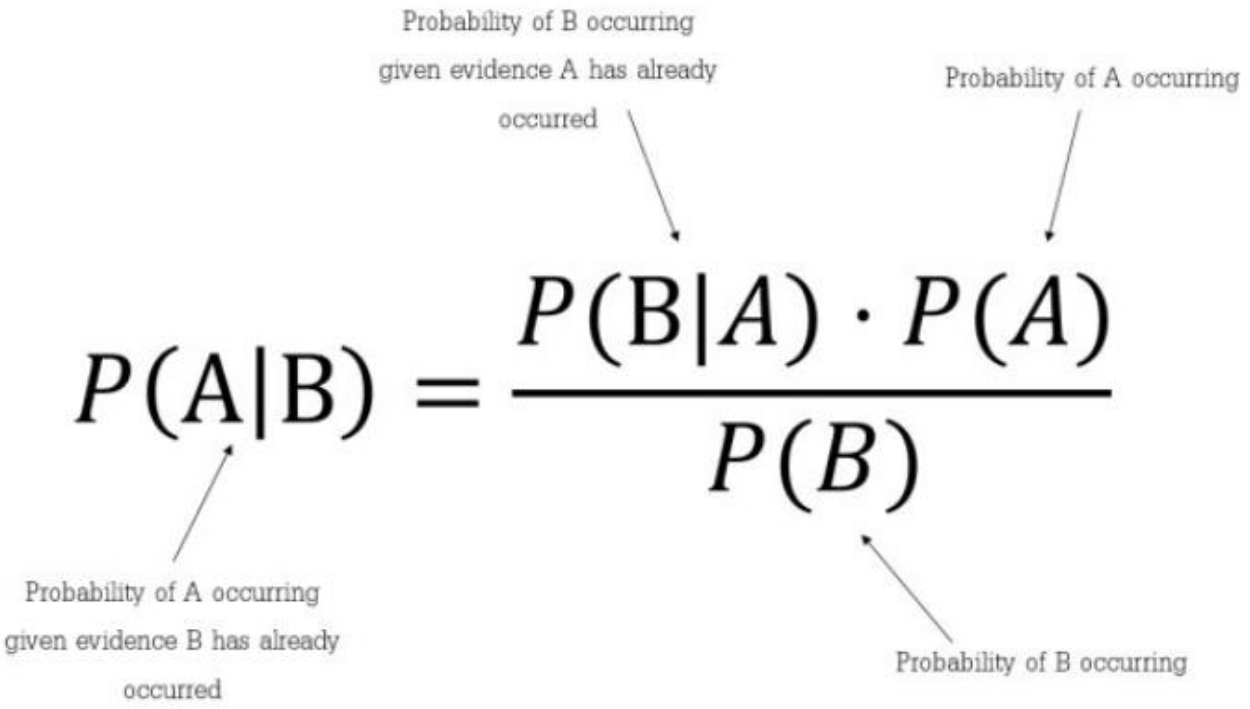


Confusion Matrix Implementation

- Implementation of Confusion Matrix

Bayes Theorem

- Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities.
- Conditional probability is a measure of the probability of an event occurring given that another event has (by assumption or evidence) occurred.



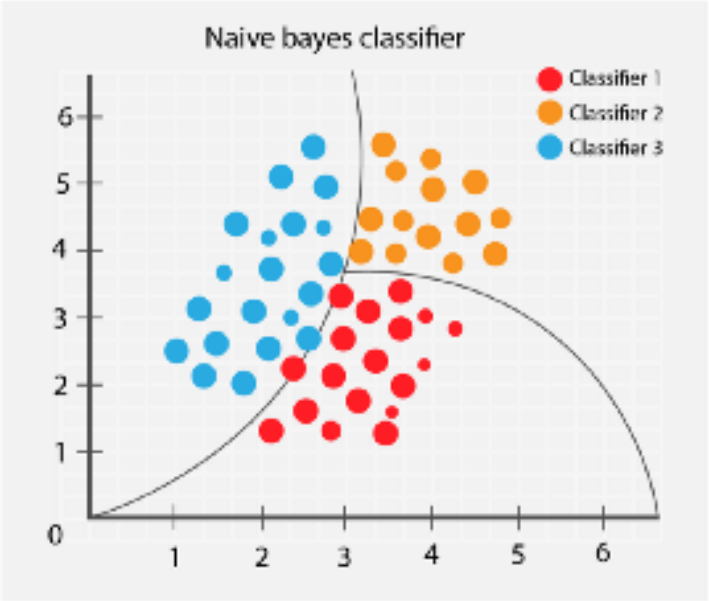
Naïve Bayes

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



Naïve Bayes Classifiers

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.

- It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

- The reason why it is called 'Naïve' because it requires rigid independence assumption between input variables.

Assumptions of Naïve Bayes Classifiers

- The fundamental Naïve Bayes assumption is that each feature makes an:
  - independent
  - equal
- contribution to the outcome.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

- We assume that **no pair of features are dependent**. For example, the color being ‘Red’ has nothing to do with the Type or the Origin of the car. Hence, the features are assumed to be Independent.
- Secondly, each feature is given the same importance. For example, knowing the only Color and Type alone can’t predict the outcome perfectly. So, none of the attributes are irrelevant and assumed to be **contributing Equally** to the outcome.

How Naïve Bayes work?

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

$$X = (x_1, x_2, x_3, ..., x_n)$$

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

How Naïve Baves work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

How Naïve Bayes work?

Counts:

Outlook			Temperature			Humidity			Windy			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
sunny	2	3	hot	2	2	high	3	4	F	6	2	9	5
overc	4	0	mild	4	2	norm	6	1	T	3	3		
rainy	3	2	cool	3	1								

Relative frequencies:

Outlook			Temperature			Humidity			Windy			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
s	2/9	3/5	h	2/9	2/5	h	3/9	4/5	F	6/9	2/5	9/14	5/14
o	4/9	0/5	m	4/9	2/5	n	6/9	1/5	T	3/9	3/9		
r	3/9	2/5	cl	3/9	1/5								

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

$$P(Yes|today) \propto \frac{3}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

How Naïve Bayes work?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

$$P(No|today) \propto \frac{2}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

How Naïve Bayes work?

today = (Sunny, Hot, Normal, False)

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

$$P(Yes|today) \propto \frac{2}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

The Zero-Frequency Problem

- One of the disadvantages of Naïve-Bayes is that if you have no occurrences of a class label and a certain attribute value together then the frequency-based probability estimate will be zero. And this will get a zero when all the probabilities are multiplied.

- An approach to overcome this ‘zero-frequency problem’ in a Bayesian environment is to add one to the count for every attribute value-class combination when an attribute value doesn’t occur with every class value.

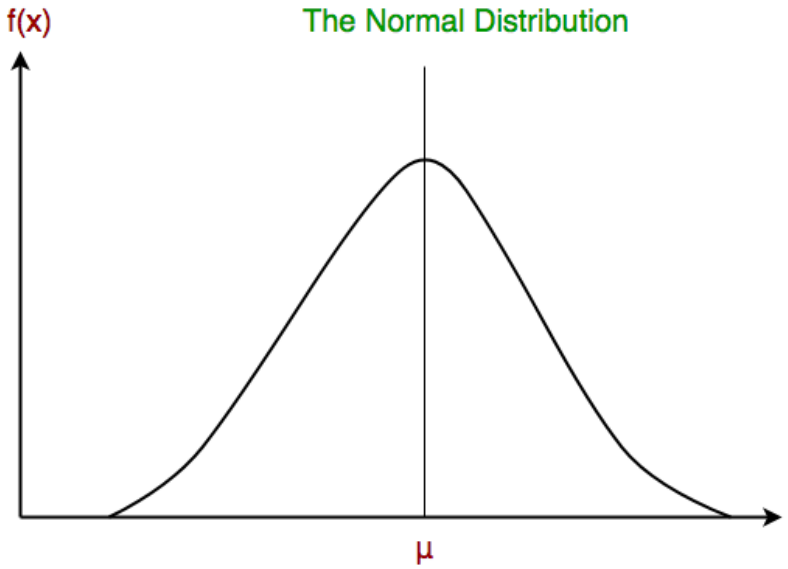
Types of Naïve Bayes Classifiers

- **Multinomial Naive Bayes:** This is mostly used for document classification problem, i.e., whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

- **Bernoulli Naive Bayes:** This is similar to the multinomial naive bayes but the predictors are Boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

Types of Naïve Bayes Classifiers

- **Gaussian Naive Bayes:** Continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.



Naïve Bayes Implemntation

Implement Naïve Bayes Classifier

Naïve Bayes Exercise

Permissions	Type	Size	Class
Read	Executable	Small	Infected
Write	Non-Executable	Large	Clean
Read	Executable	Medium	Infected
Read	Executable	Medium	Infected
Write	Executable	Medium	Clean
Read	Non-Executable	Large	Clean
Write	Executable	Small	Infected

File = (Write, Executable, Large) = ?

Summary

- Confusion Matrix
- Naïve Bayes