Powered by:

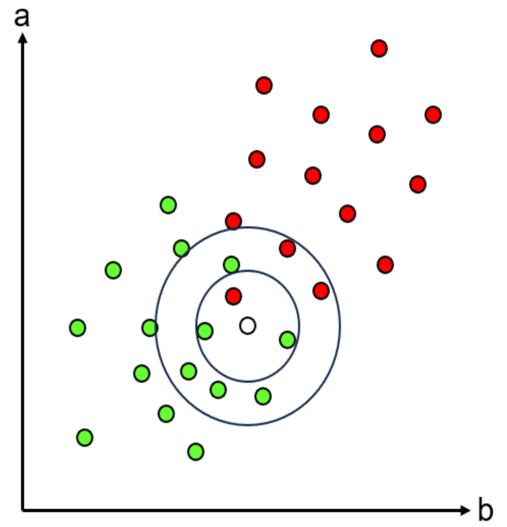
Media Partner:

TECHAD BOOTCAMP

BATCH 8

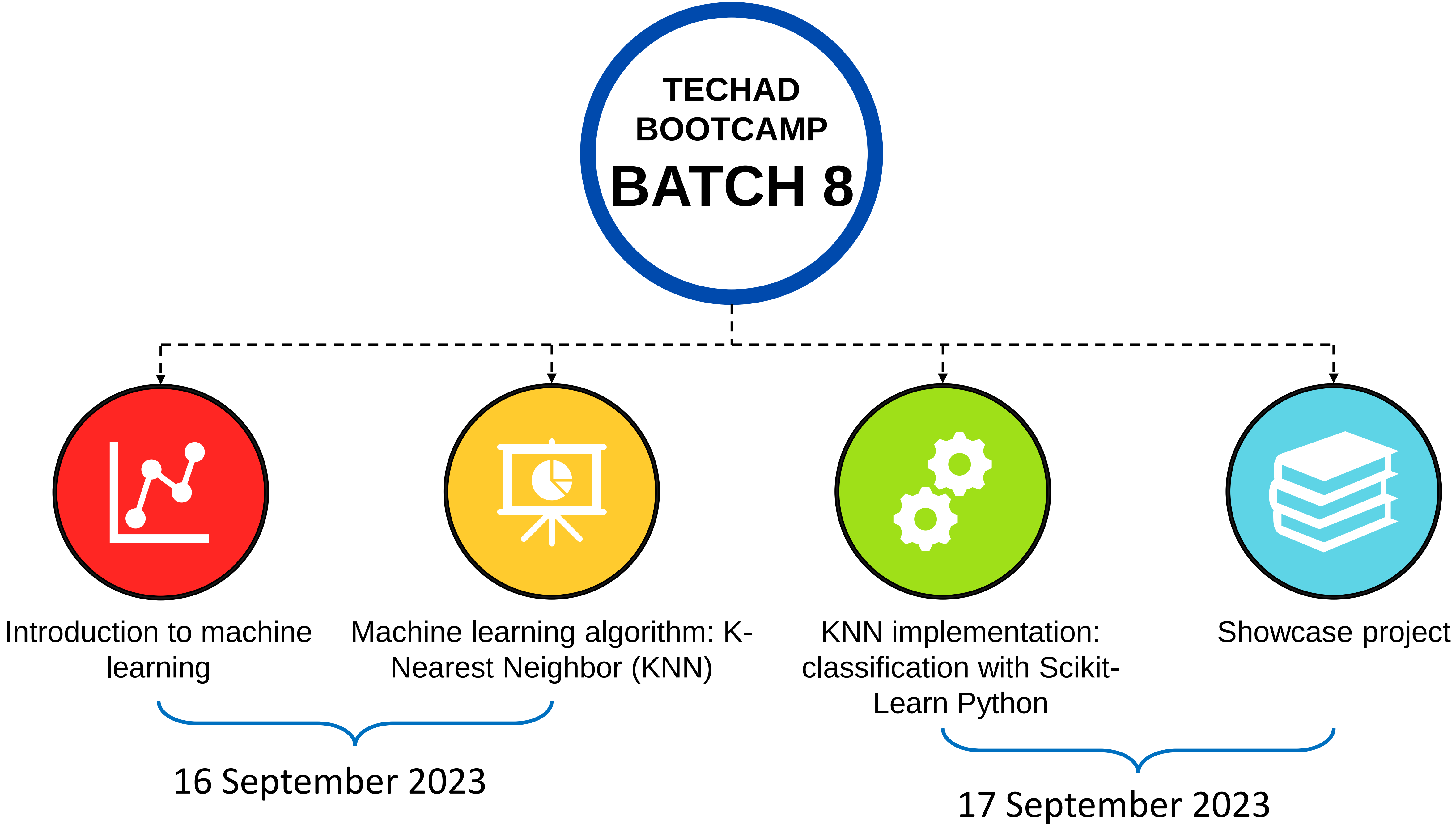
TOPIC:

Machine Learning: An Introduction to KNN and Its Implementation



FARID YULI MARTIN ADIYATMA, S.T.

16 September 2023


TECHAD
BOOTCAMP
BATCH 8
Introduction to machine learning
Machine learning algorithm: K-Nearest Neighbor (KNN)
KNN implementation: classification with Scikit-Learn Python
Showcase project
16 September 2023
17 September 2023

Powered by:

Media Partner:

Introduction to Machine Learning

Module 1

16 September 2023

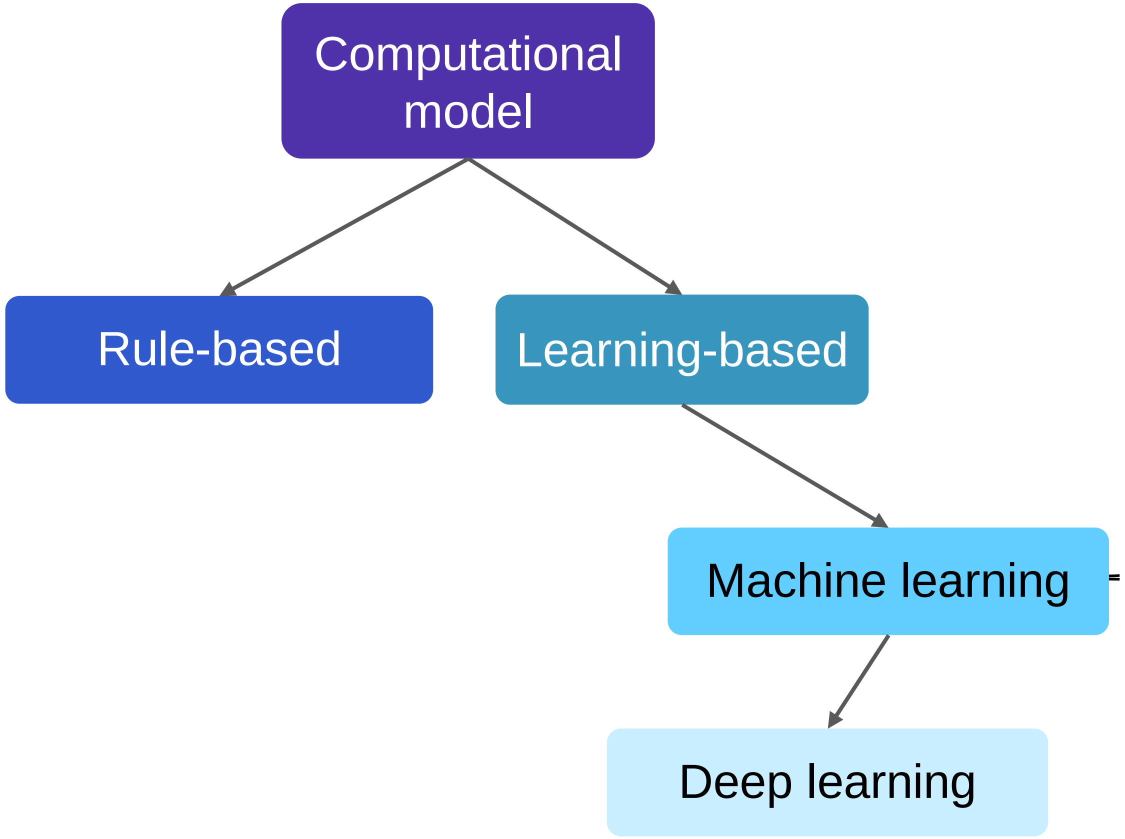
TECHAD BOOTCAMP

BATCH 8

TOPIC:

Machine Learning:
An Introduction to KNN
and Its Implementation

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior.

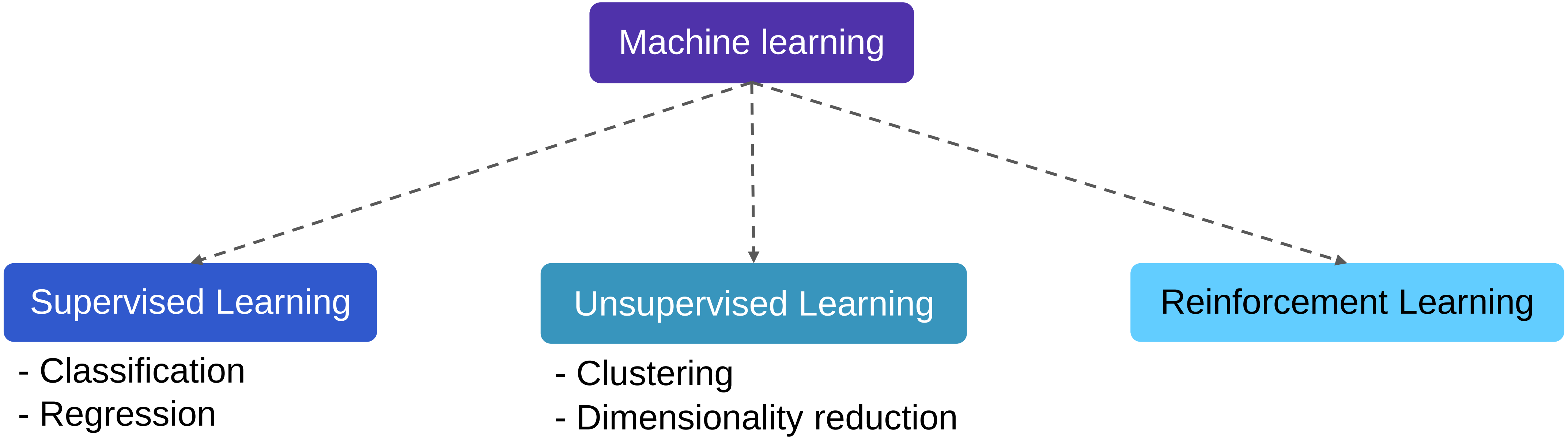


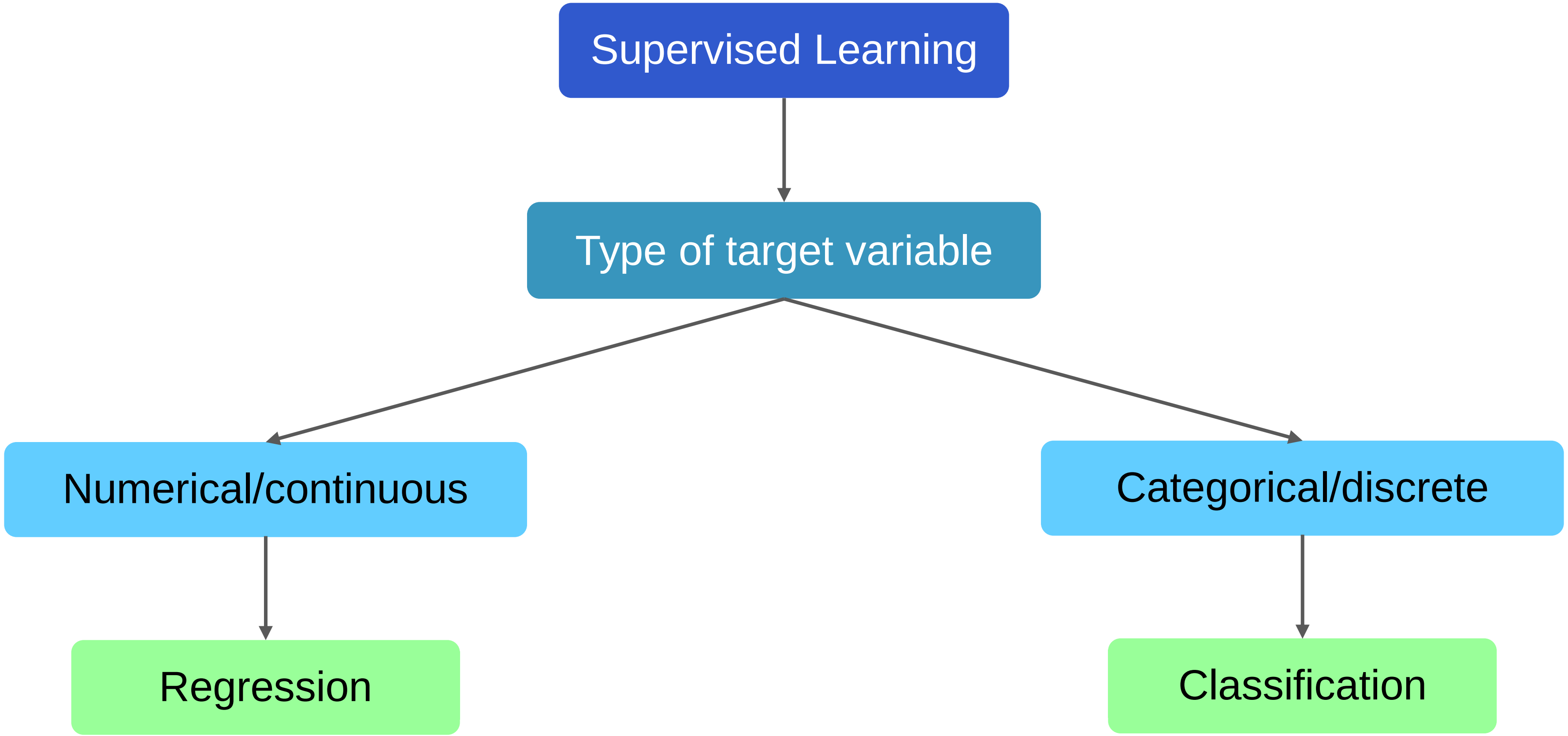
- **Supervised learning**
- Unsupervised learning
- Reinforcement learning
- Semi-supervised learning

AI categories based on the domain/input data

- Natural Language Processing
- Computer vision/image processing
- **Numerical or categorical data processing**

Source:
https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained


Machine learning
Supervised Learning
- Classification
- Regression
Unsupervised Learning
- Clustering
- Dimensionality reduction
Reinforcement Learning



Source:
https://www.datacamp.com/blog/classification-machine-learning

Regression

crim	tax	nox	Medv
632	296	538	24
2.731	242	469	21.6
2.729	242	469	34.7
3.237	222	458	33.4
6.905	222	458	36.2
2.985	222	458	28.7
8.829	311	524	22.9
14.455	311	524	27.1
21.124	311	524	16.5

Target

Features

Classification

No.	Sepal Length	Sepal Width	Species
1	5.3	3.7	Setosa
2	5.1	3.8	Setosa
3	7.2	3	Virginica
4	5.4	3.4	Setosa
5	5.1	3.3	Setosa
6	5.4	3.9	Setosa
7	7.4	2.8	Virginica
8	6.1	2.8	Versicolor
9	7.3	2.9	Virginica
10	6	2.7	Versicolor
11	5.8	2.8	Virginica
12	6.3	2.3	Versicolor
13	5.1	2.5	Versicolor
14	6.3	2.5	Versicolor
15	5.5	2.4	Versicolor

Target

Features

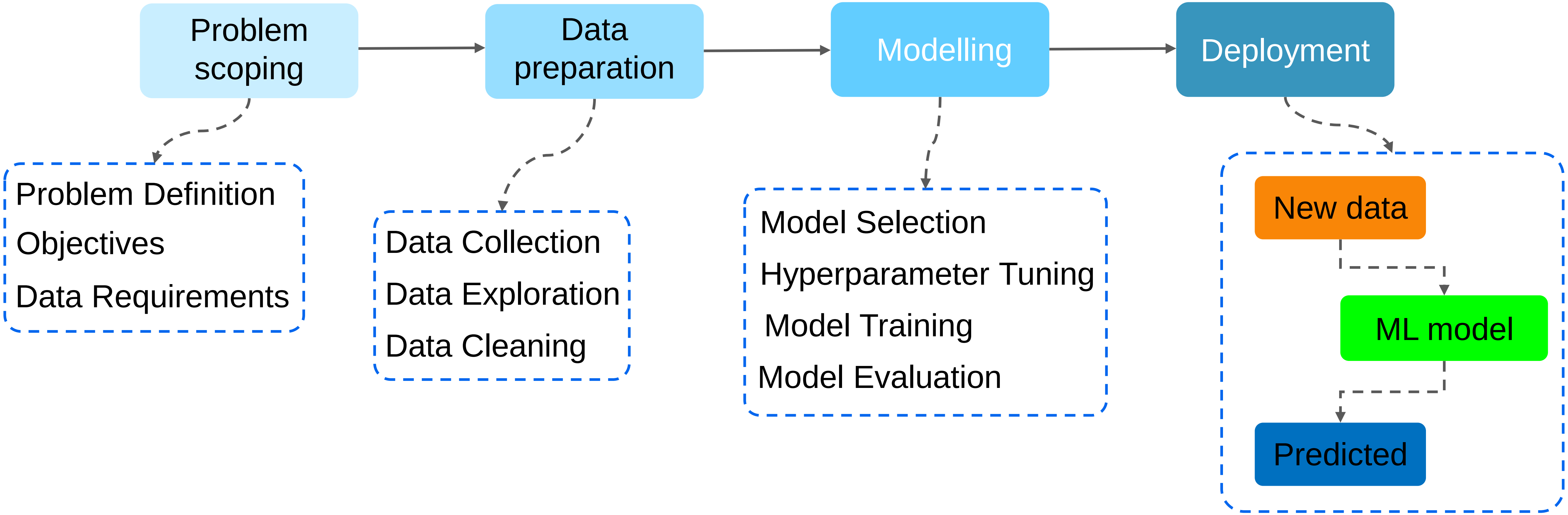
Regression

- Linear Regression
- Polynomial Regression
- Lasso Regression
- Ridge Regression
- Logistic Regression
- KNN
- SVR
- Decision Tree
- Random Forest
- Neural Network

Classification

- KNN
- SVM
- Naïve Bayes
- Decision Tree
- Ensemble learning
- Neural network

AI Project Cycle



Powered by:

Media Partner:

Machine learning algorithm: K-Nearest Neighbor (KNN)

Module 2

16 September 2023

TECHAD BOOTCAMP

BATCH 8

TOPIC:

Machine Learning: An Introduction to KNN and Its Implementation

Overview of KNN

KNN is a pattern recognition algorithm that can be used to:
- Classification
- Regression

KNN uses a similarity measure, based on distance, to classify the target.

51

Distance function

- Euclidean distance
- Manhattan distance
- Minkowski distance

The most important hyperparameter in KNN is the number of neighbors (K)

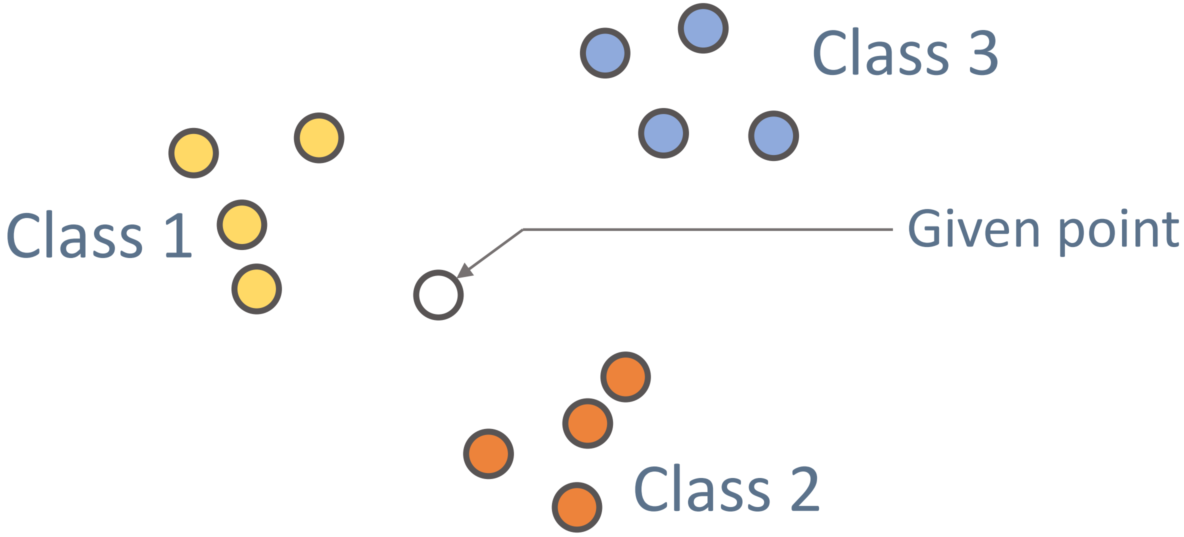
There is no structured method to find the best K

KNN is sensitive to outliers

Source:
https://towardsdatascience.com/k-nearest-neighbors-knn-algorithm-23832490e3f4#:~:text=This%20is%20considered%20as%20overfitting,on%20the%20test%20dataset%20also.

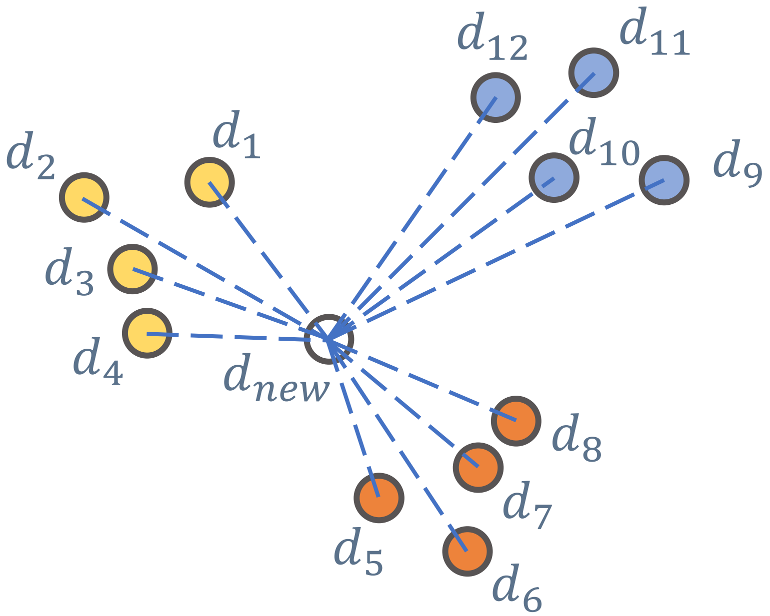
Working principle of KNN

Step 1
Initialization



Working principle of KNN

Step 2
Calculate distance



Distance function in KNN

Euclidean distance

$$d\ (a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Manhattan distance

$$d\ (a, b) = \sum_{i=1}^{n} |a_i - b_i|$$

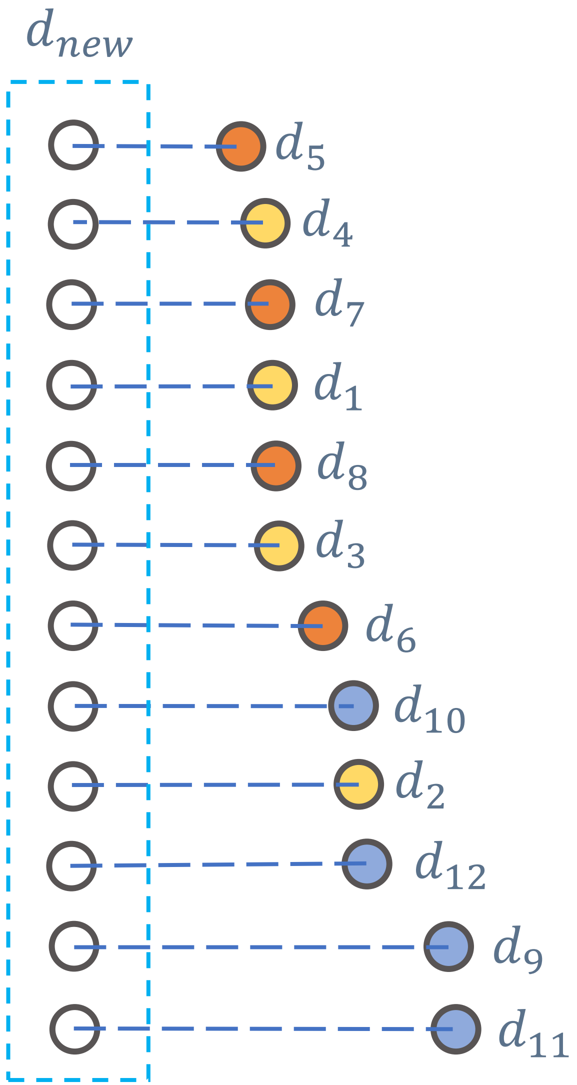
Minkowski distance

$$d\ (a, b) = (\sum_{i=1}^{n} |a_i - b_i|^p)^{1/p}$$

Working principle of KNN

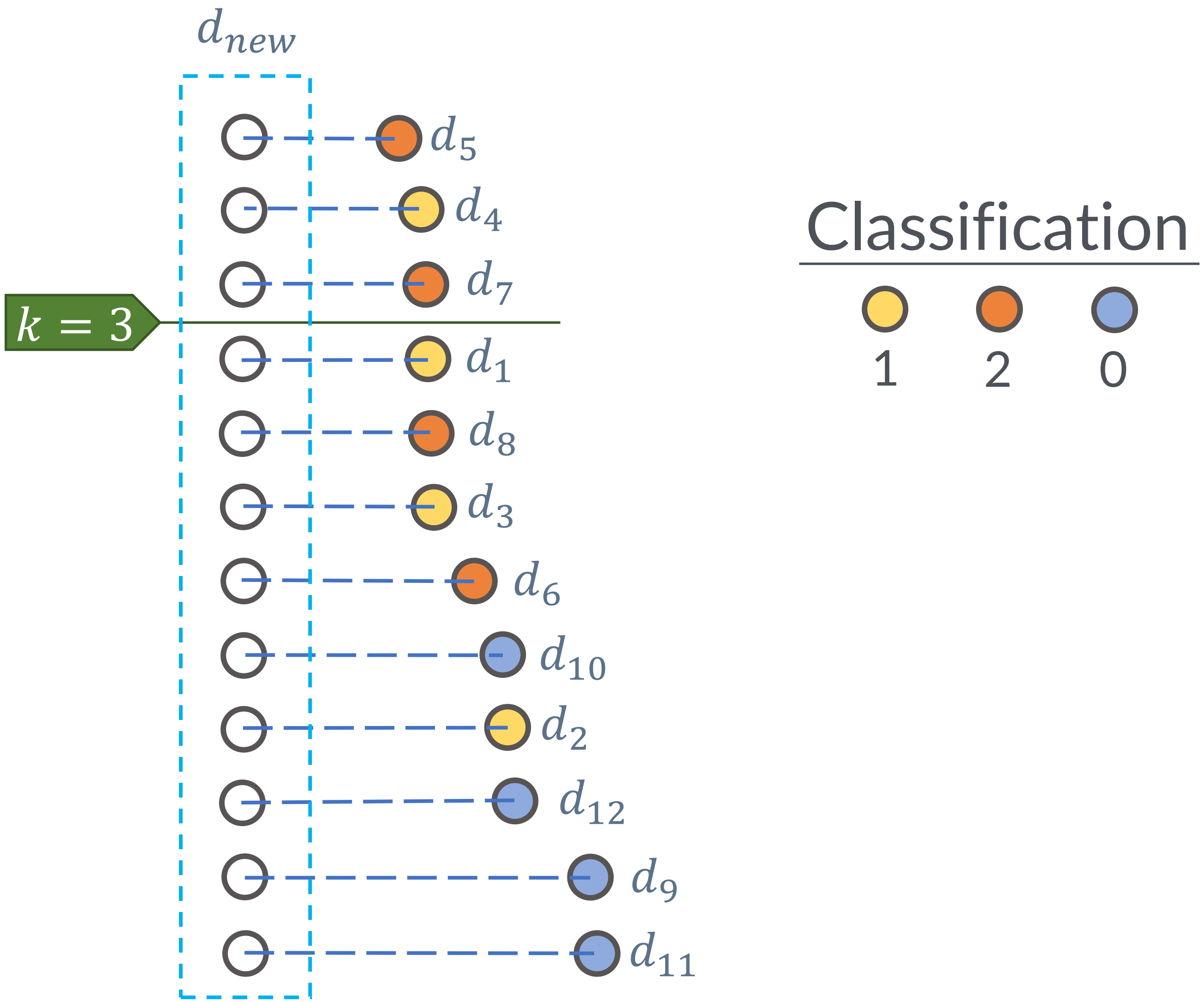
Step 3

Nearest Neighbors

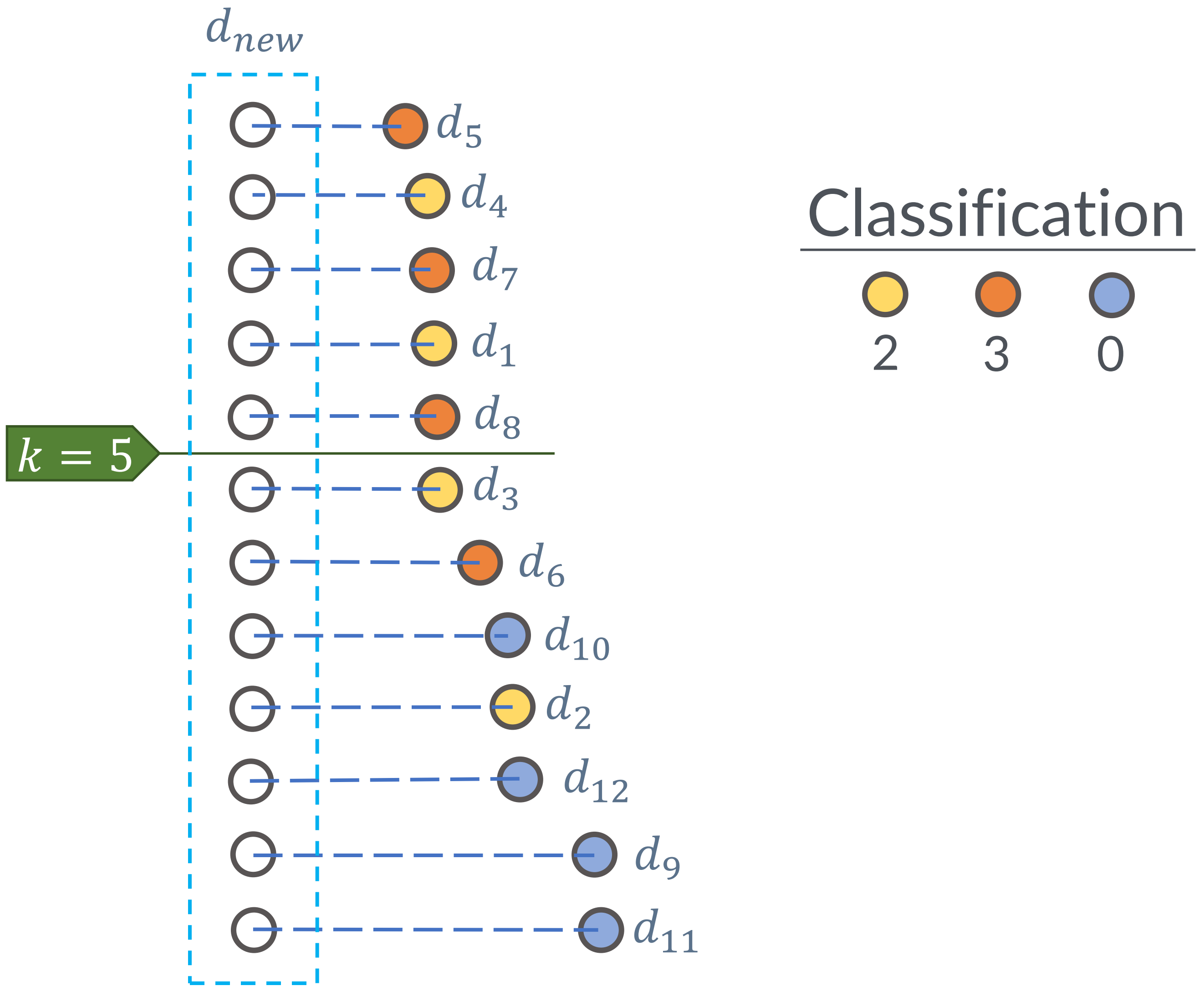


Working principle of KNN

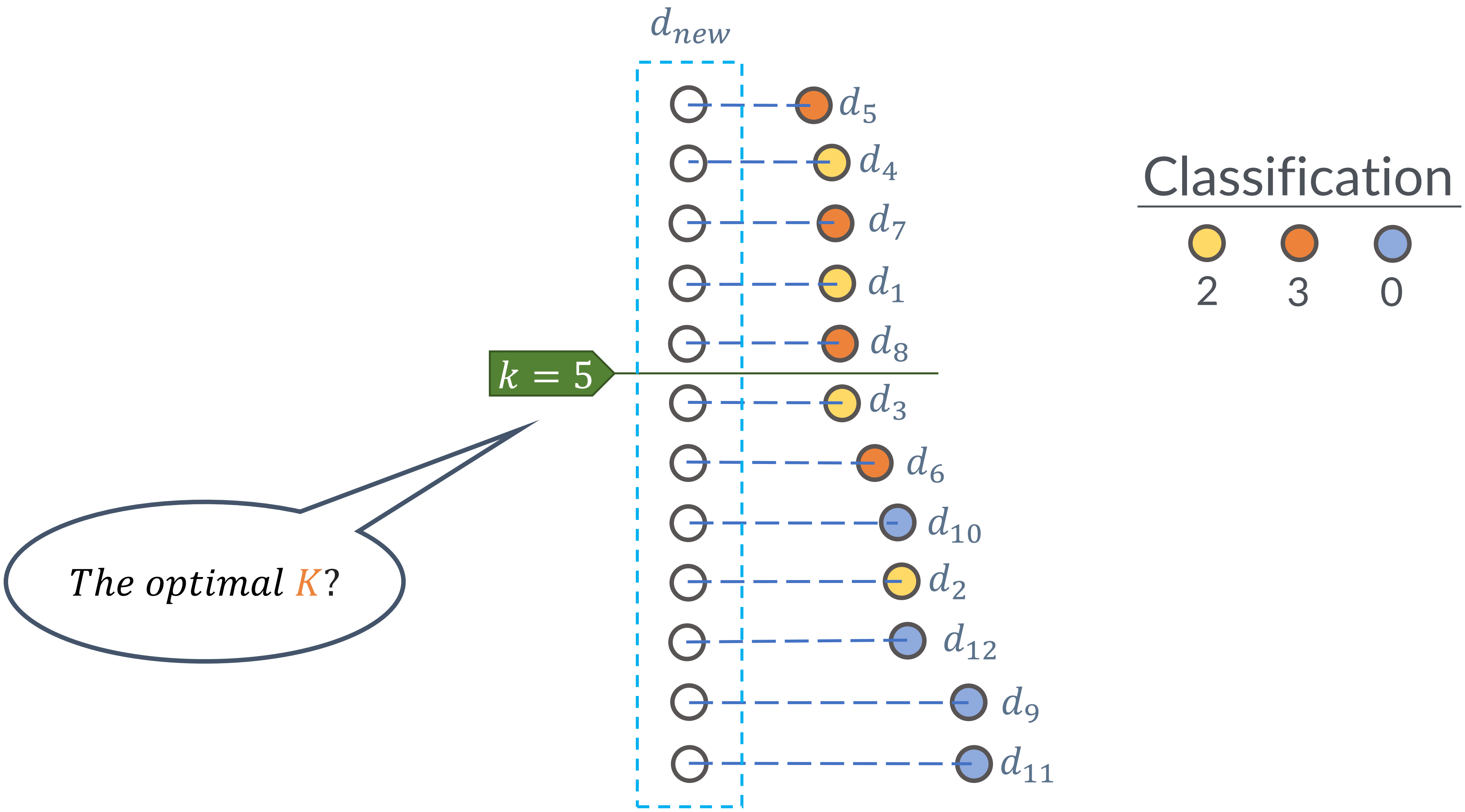
Step 4
Vote



Working principle of KNN



Working principle of KNN



Simple Example of KNN

Dataset

No.	Sepal Length	Sepal Width	Species
1	5.3	3.7	Setosa
2	5.1	3.8	Setosa
3	7.2	3	Virginica
4	5.4	3.4	Setosa
5	5.1	3.3	Setosa
6	5.4	3.9	Setosa
7	7.4	2.8	Virginica
8	6.1	2.8	Versicolor
9	7.3	2.9	Virginica
10	6	2.7	Versicolor
11	5.8	2.8	Virginica
12	6.3	2.3	Versicolor
13	5.1	2.5	Versicolor
14	6.3	2.5	Versicolor
15	5.5	2.4	Versicolor

New data

No.	Sepal Length	Sepal Width	Species
1	5.2	3.1	?

KNN classification using various distance functions

$$Euclidean\ distance = \sqrt{(5.2 - 5.3)^2 + (3.1 - 3.7)^2}$$
$$= 0.608$$

$$Manhattan\ distance = |5.2 - 5.3| + |3.1 - 3.7|$$
$$= 0.7$$

$$Minkowski\ distance = (|5.2 - 5.3|^3 + |3.1 - 3.7|^3)^{\frac{1}{3}}$$
$$= 0.802$$

Source:
https://medium.com/machine-learning-researcher/k-nearest-neighbors-in-machine-learning-e794014abd2a

Simple Example of KNN

Distance between each datapoint and new data

No.	Sepal Length	Sepal Width	Species	Distance
1	5.3	3.7	Setosa	0.608
2	5.1	3.8	Setosa	0.707
3	7.2	3	Virginica	2.002
4	5.4	3.4	Setosa	0.36
5	5.1	3.3	Setosa	0.22
6	5.4	3.9	Setosa	0.82
7	7.4	2.8	Virginica	2.22
8	6.1	2.8	Versicolor	0.94
9	7.3	2.9	Virginica	2.1
10	6	2.7	Versicolor	0.89
11	5.8	2.8	Virginica	0.67
12	6.3	2.3	Versicolor	1.36
13	5.1	2.5	Versicolor	0.6
14	6.3	2.5	Versicolor	1.25
15	5.5	2.4	Versicolor	0.75

New data

No.	Sepal Length	Sepal Width	Species
1	5.2	3.1	?

Using Euclidean Distance

Source:
https://medium.com/machine-learning-researcher/k-nearest-neighbors-in-machine-learning-e794014abd2a

Simple Example of KNN

Sorted dataset

No.	Sepal Length	Sepal Width	Species	Distance
5	5.1	3.3	Setosa	0.22
4	5.4	3.4	Setosa	0.36
13	5.1	2.5	Versicolor	0.6
1	5.3	3.7	Setosa	0.608
11	5.8	2.8	Virginica	0.67
2	5.1	3.8	Setosa	0.707
15	5.5	2.4	Versicolor	0.75
6	5.4	3.9	Setosa	0.82
10	6	2.7	Versicolor	0.89
8	6.1	2.8	Versicolor	0.94
14	6.3	2.5	Versicolor	1.25
12	6.3	2.3	Versicolor	1.36
3	7.2	3	Virginica	2.002
9	7.3	2.9	Virginica	2.1
7	7.4	2.8	Virginica	2.22

$k = 5$

For k = 5

No.	Sepal Length	Sepal Width	Species
1	5.2	3.1	Setosa

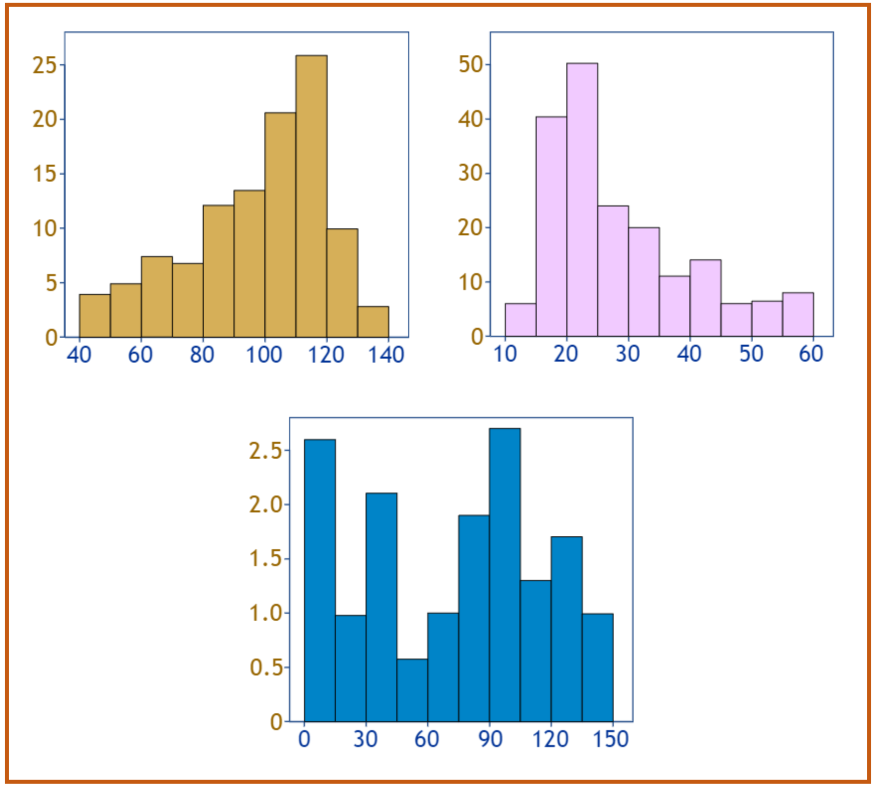
Voting

Setosa	Versicolor	Virginica
3	1	1

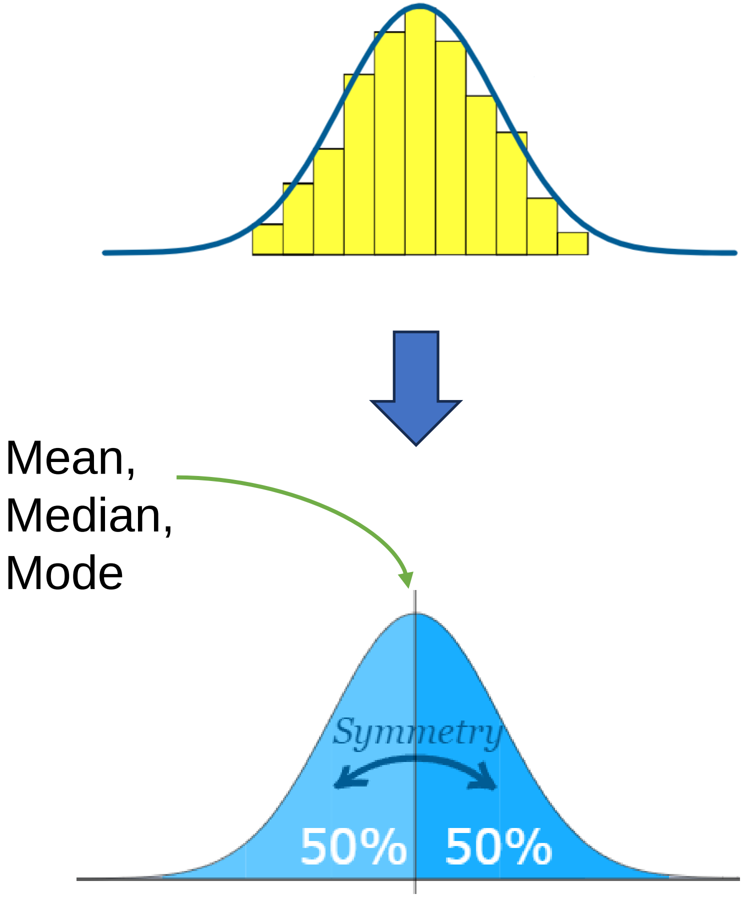
Source:
https://medium.com/machine-learning-researcher/k-nearest-neighbors-in-machine-learning-e794014abd2a

Normal Distribution (Gaussian)

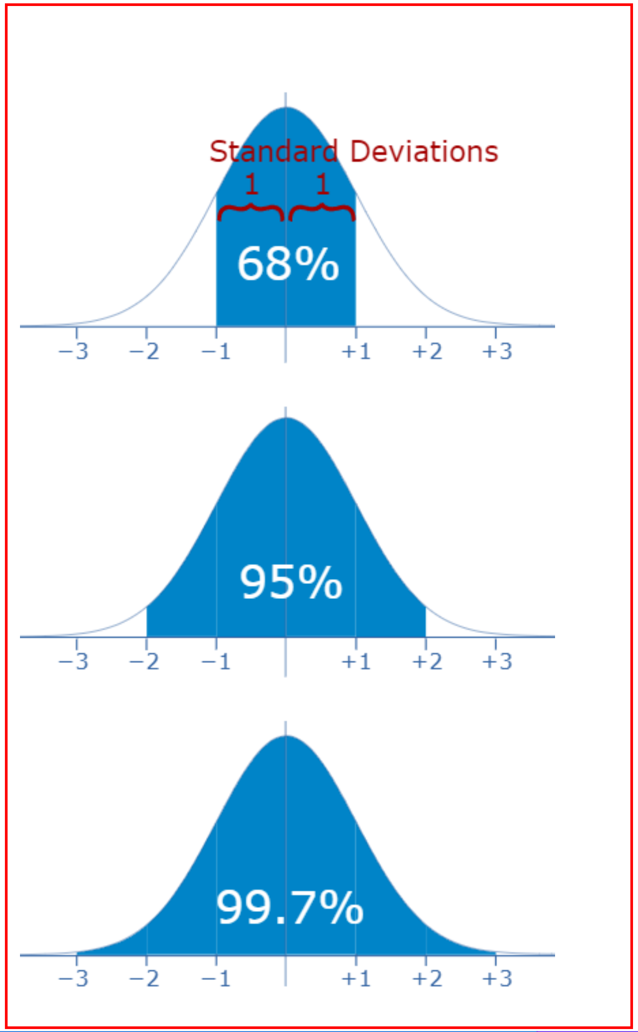
Data can be "distributed" (spread out) in different ways.



Normal distribution illustrated as bell curve



Three standard deviations



Source:
https://www.mathsisfun.com/data/standard-normal-distribution.html

Data Cleaning

1. Missing data

Missing data is a common issue in data analysis and machine learning, and how it is handled can significantly impact the quality and reliability of analytical results and predictive models.

Mitigation technique, e.g., Deletion, mean, median, mode, forward and backward fill, interpolation and regression

Missing data

Sepal Length	Sepal Width	Species
6.1	2.8	Versicolor
6	2.7	Versicolor
6.3		Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

Deletion

Sepal Length	Sepal Width	Species
6.1	2.8	Versicolor
6	2.7	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

Mean

Sepal Length	Sepal Width	Species
6.1	2.8	Versicolor
6	2.7	Versicolor
6.3	2.58	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

Forward fill

Sepal Length	Sepal Width	Species
6.1	2.8	Versicolor
6	2.7	Versicolor
6.3	2.5	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

Data Cleaning

2. Outliers

Outliers are data points that significantly differ from the majority of the data and can distort statistical analyses and machine learning models.

Effective outlier mitigation is crucial to ensure the reliability and accuracy of data-driven insights and models.

Detection technique, e.g., Z-score, interquartile range (IQR)

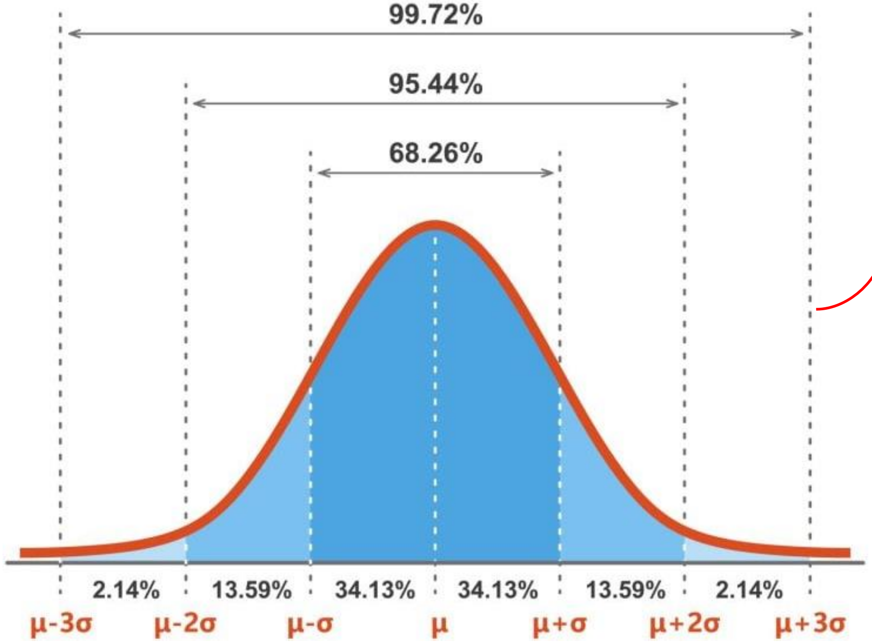
Z-score

$$\boxed{Z = \frac{x - \mu}{\sigma}}$$

x = data

μ = mean

σ = standard deviation



Rule

If Z > 3 or Z < -3 , then Z is outlier

Imbalanced Data

Imbalanced data in machine learning refers to a dataset where the distribution of the target class is not equal.

This means that one class (the majority class) has a significantly higher number of observations than the other class (the minority class).

There are a number of techniques that can be used to handle imbalanced data in machine learning:

1. **Oversampling**: This involves creating synthetic examples of the minority class.
2. **Undersampling**: This involves removing examples of the majority class.

Majority class

```
Normal      757
Disease     205
```

Minority class

Feature Scaling

Feature scaling in machine learning is the process of transforming the features in a dataset so that their values share a similar scale.

Normalization rescales the values of a feature to a specific range, typically [0, 1] or [-1, 1].

Standardization does not bound values to a specific range like normalization. Instead, it scales data to have a mean of 0 and a standard deviation of 1.

Example:

Dataset

Employee	Age	Salary
1	44	7300000
2	27	4700000
3	30	5300000
4	38	6200000
5	40	5700000
6	35	5300000

New data

Age	Salary
48	7800000

Employee 1

$$Euclidean\ distance = \sqrt{(7800000 - 7300000)^2 + (48 - 44)^2}$$

$$= 500000$$

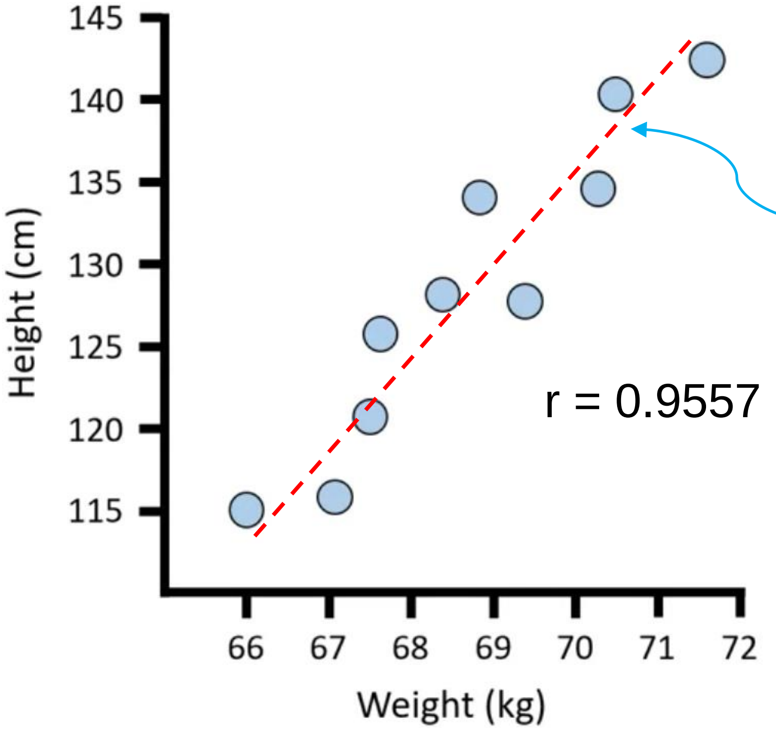
Feature Selection

The higher the number of features, the more computational time is needed.

Some features do not exhibit a strong correlation with the target.

The concept of Pearson correlation

Participant	Weight (kg)	Height (cm)
1	66.0	115.0
2	67.2	116.3
3	67.6	120.8
4	67.8	125.7
5	68.5	127.5
6	69.4	126.9
7	69.0	134.2
8	70.3	134.9
9	70.7	140.6
10	71.8	144.1



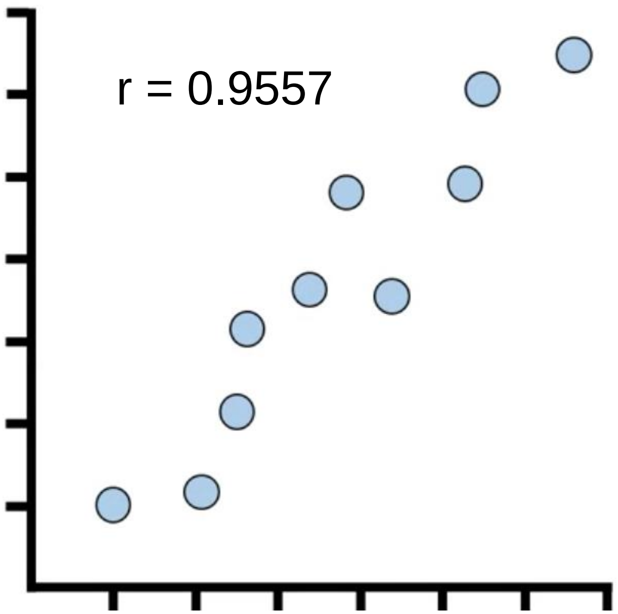
A Pearson correlation measures the strength and direction of linear correlation

Source:
https://www.youtube.com/watch?v=e4ApDqG6MGE

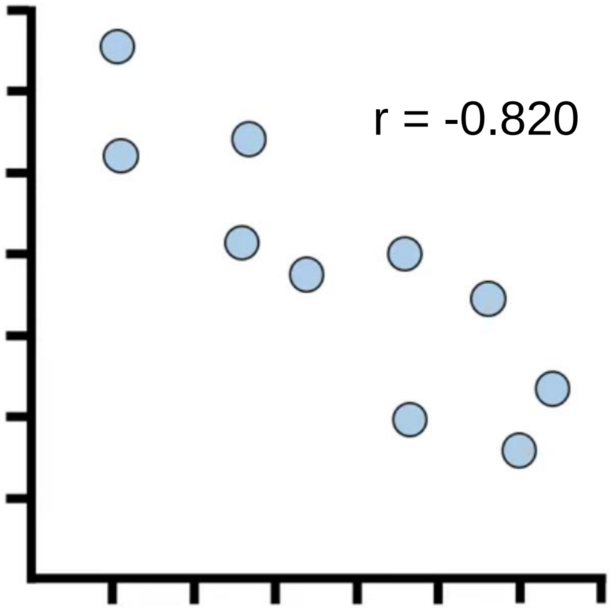
Correlation

Direction of correlation

Positive correlation, r > 0



Negative correlation, r < 0

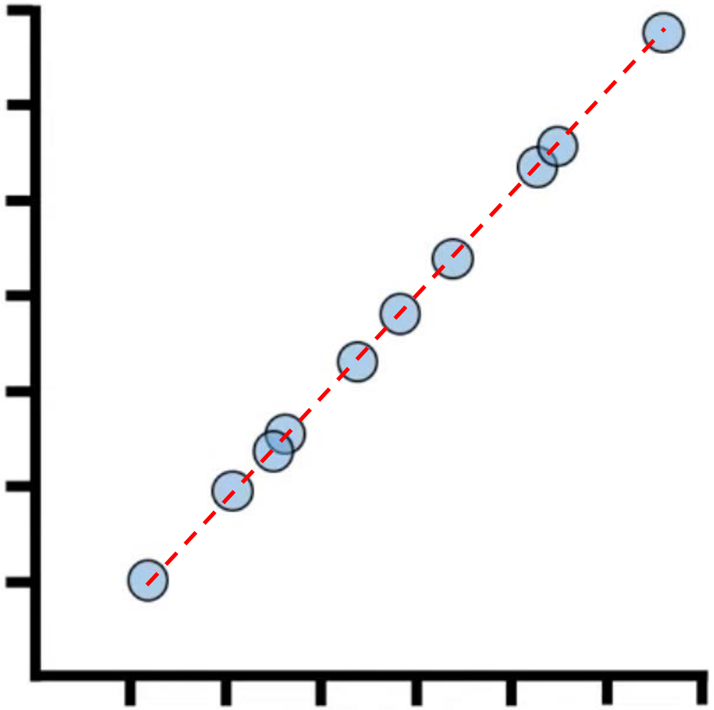


Source:
https://www.youtube.com/watch?v=e4ApDqG6MGE

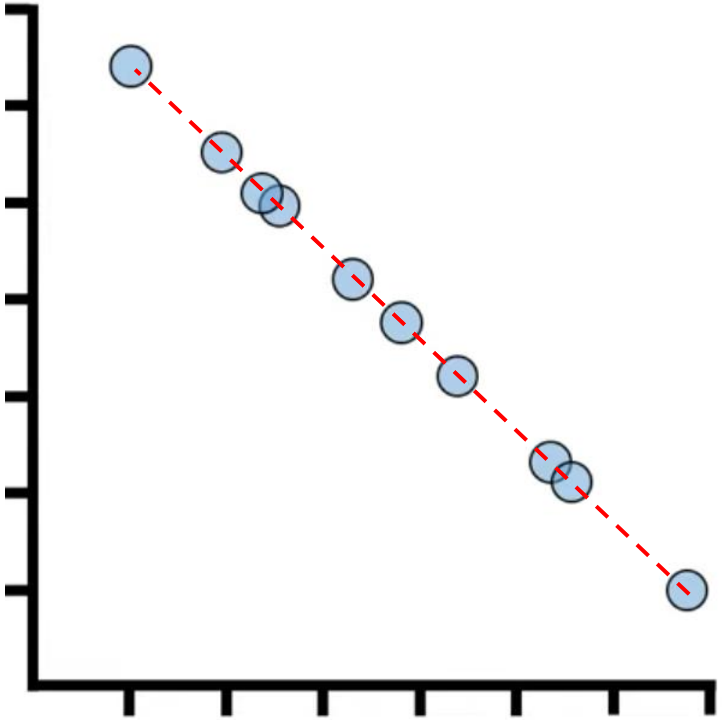
Correlation

Strength of correlation

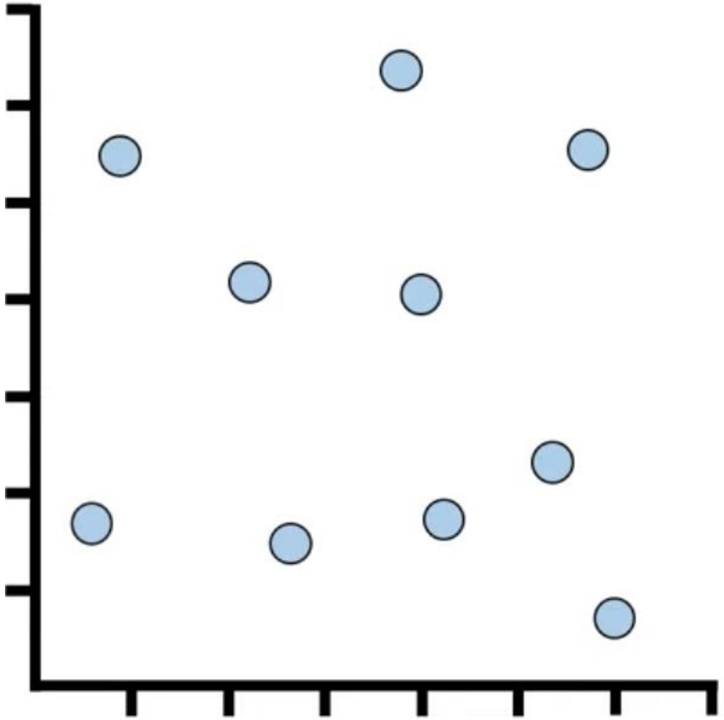
Perfect positive correlation, r =1



Perfect negative correlation, r = -1



No correlation, r = 0



Source:
https://www.youtube.com/watch?v=e4ApDqG6MGE

Evaluation Metrics

Confusion Matrix	Actually Positive	Actually Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

A good model has high TP and TN and low FP and FN.

Accuracy

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

Precision

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

Recall

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

F1 Score

$$Recall = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Terima kasih

ขอบคุณมาก