Supervised ML

$$\left\{ \begin{array}{l} X \longleftrightarrow \boxed{Y} \;\Big\}\ \text{Train} \\ \text{input} \qquad\qquad y_{pred} \\ X_{test} \longleftrightarrow Y_{test} \quad 20 \implies \text{Model evaluation} \end{array} \right.$$

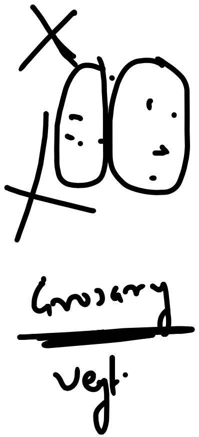
Reg	Classification
	Accuracy

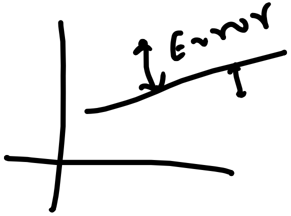
Unsupervised Learning

Yes No
No No

Clustering

Act. y	Pred Yes	No
	10	2
tgt No	3	5

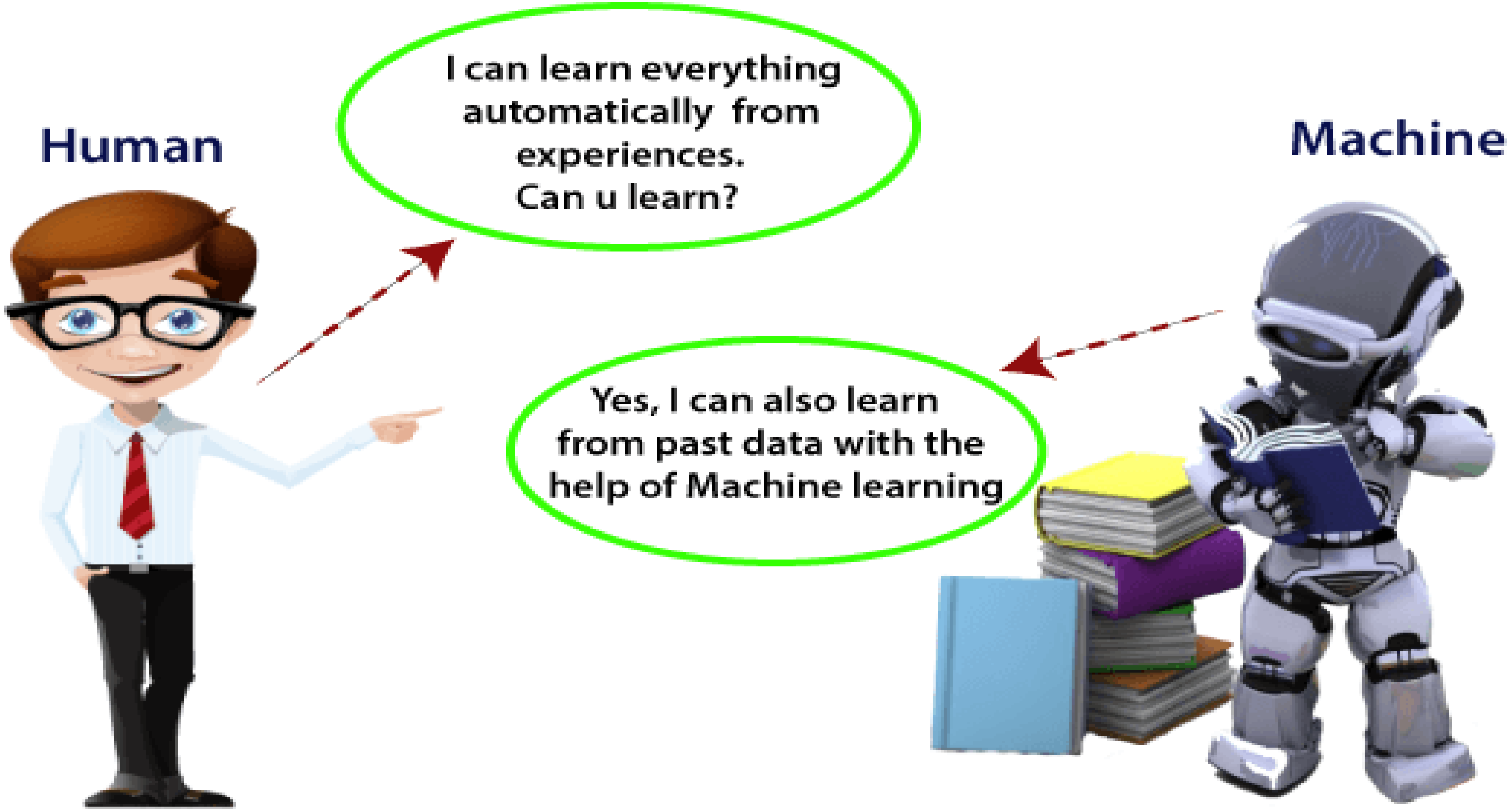




$$(y-\hat{y})^2 = \left(y_{act} - y_{pred}\right)$$

$$= \frac{1}{n}\sum \left(y-\hat{y}\right)^2$$

MSE, RMSE



Human

I can learn everything automatically from experiences.
Can u learn?

Machine

Yes, I can also learn from past data with the help of Machine learning

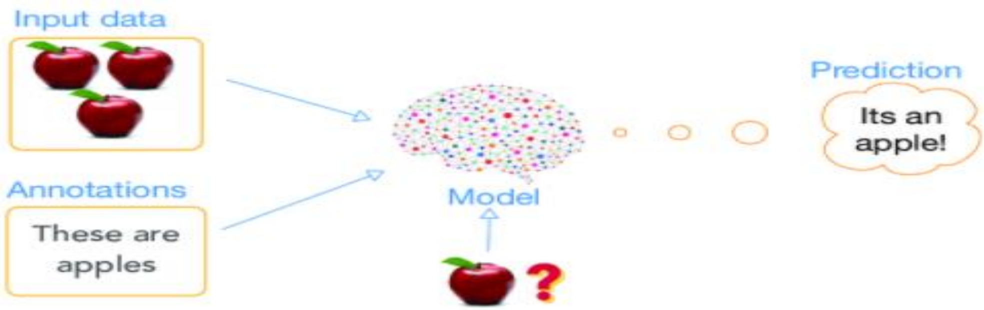
Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

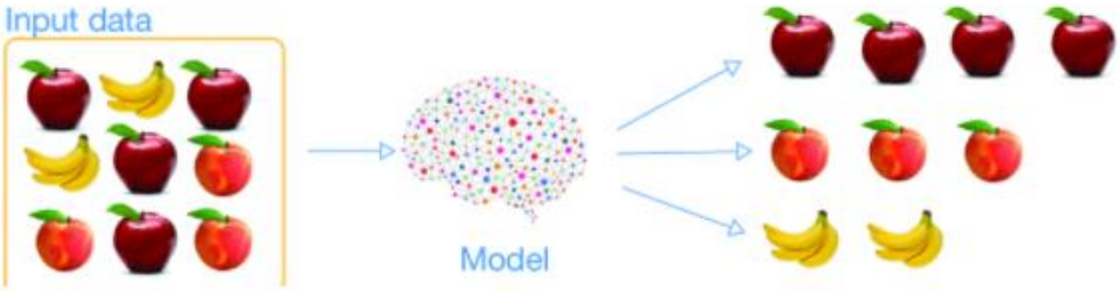
Types of Machine Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

supervised learning



unsupervised learning



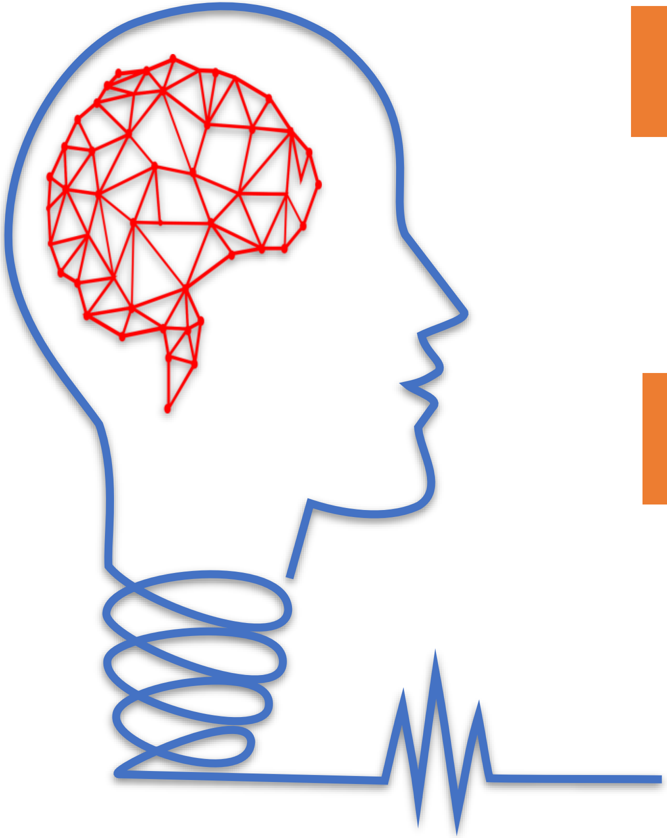
Unsupervised Learning

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Why is it called Unsupervised?

This is called unsupervised learning because unlike supervised learning above there are no correct answers and there is no teacher. Algorithms are left to their own device to discover and present the interesting structure in the data.

Types of Unsupervised Learning



01

Clustering

- Hierarchical Clustering
- K-means Clustering
- Density-Based Clustering

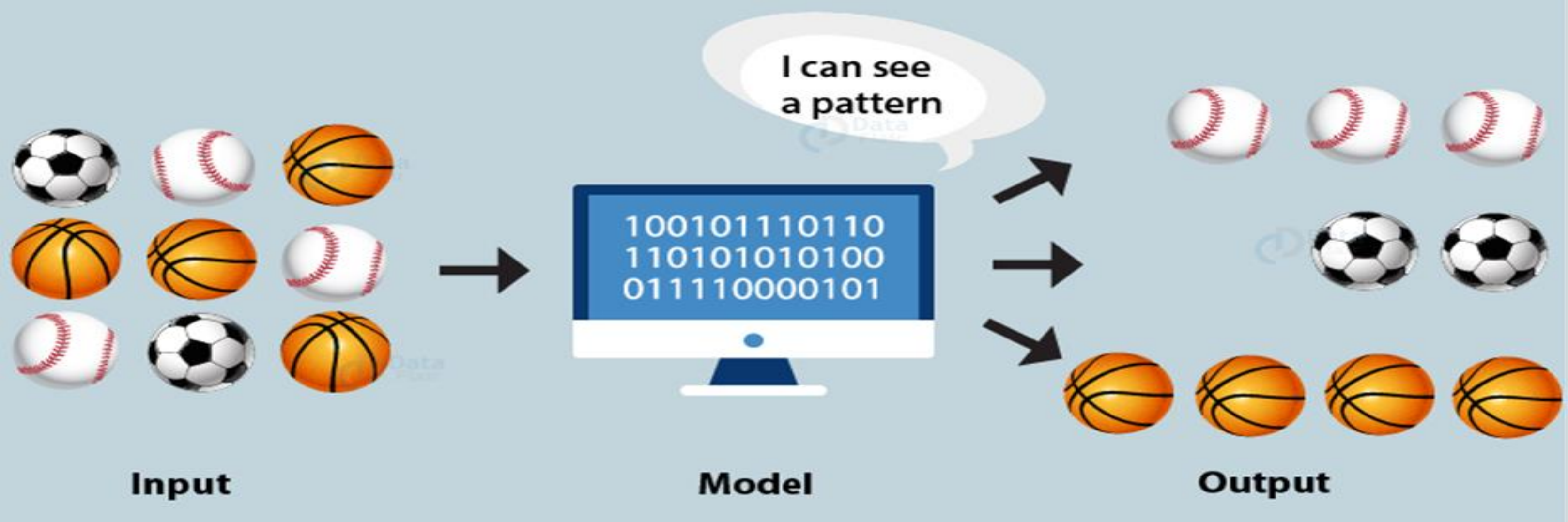
02

Dimensionality Reduction

- Principal Component Analysis

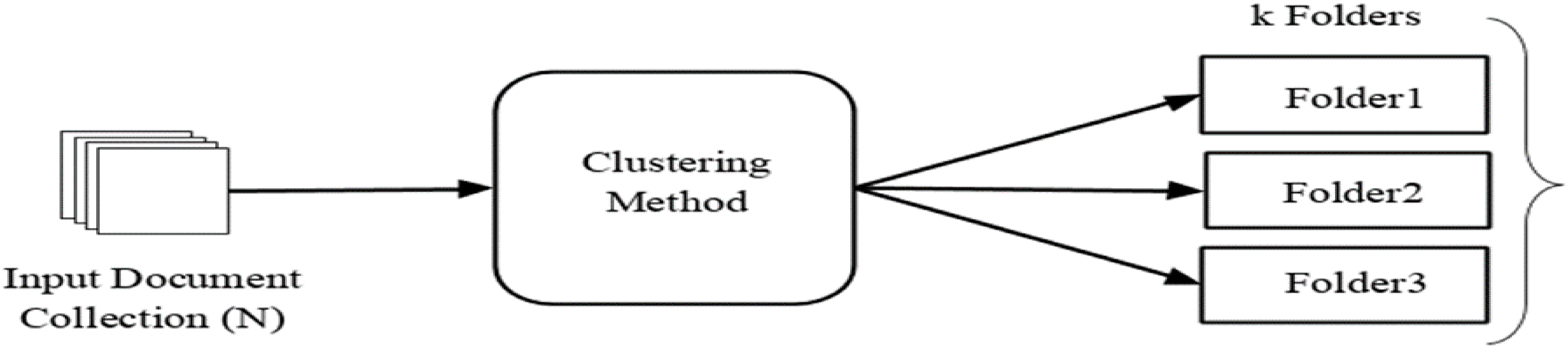
Clustering

Clustering is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Unlike supervised learning, clustering algorithms only interpret the input data and find natural groups or clusters in feature space. It is basically a collection of objects on the basis of similarity and dissimilarity between them



Why Clustering is Unsupervised Learning?

Clustering is an unsupervised machine learning task that automatically divides the data into clusters, or groups of similar items. It does this without having been told how the groups should look ahead of time.



Given a collection of N documents, Clustering segregates them into k groups based on some similarity/dissimilarity measure

Distance Metrics

Distance are used to measure the Similarity

There are many Ways to measures the distance between two instance

Minkowski

$$\left(\sum_{i=1}^{k} \left(|x_i - y_i| \right)^q \right)^{1/q}$$

$$Manhattan = \sum_{i=1}^{n} |x_i - y_i|$$

$$Euclidean = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

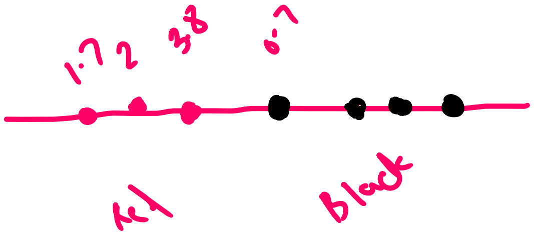
$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Total = 7
Grouping = 2

Age



Fixed

20

Doc Gov
termination { }
0.89

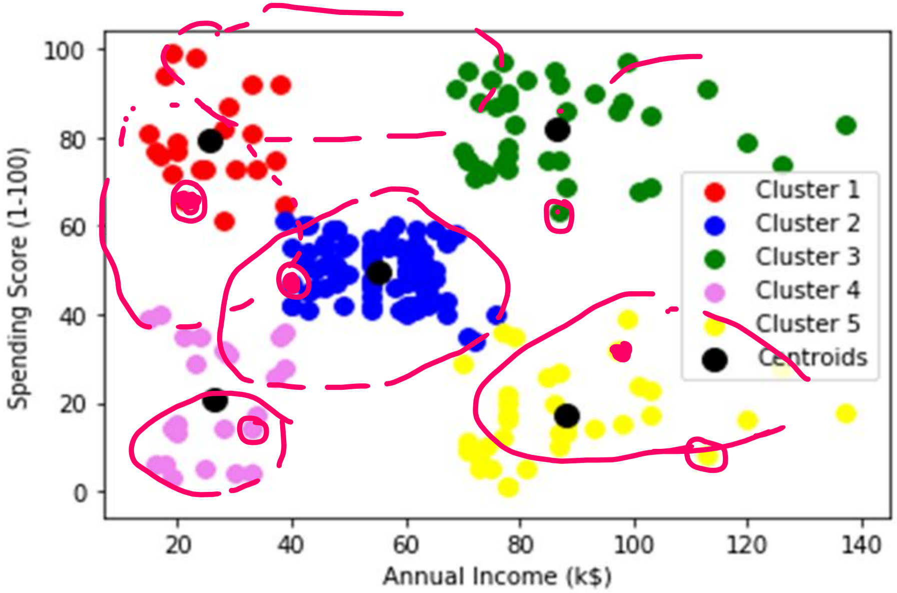
Heath { }

Cancer
-1.3

K-Means Clustering

The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid

k = 5



Steps

1. Choose the number of clusters k
2. Select k random points from the data as centroids
3. Assign all the points to the closest cluster centroid
4. Re-compute the centroids of newly formed clusters (using mean)
5. Repeat 2 to 4

Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

How to Choose the Right Number of Clusters in K-Means Clustering?

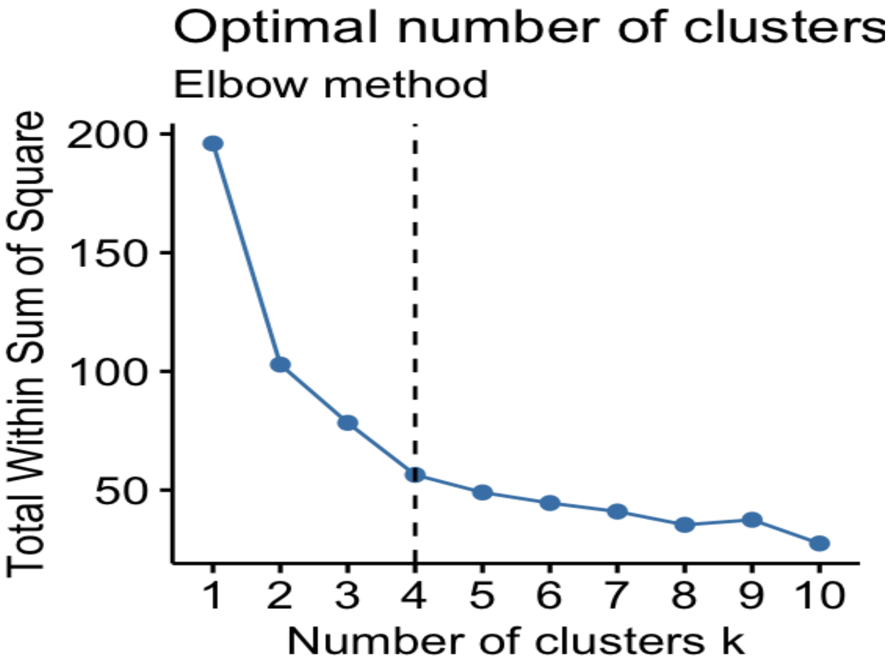
The Elbow Method

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

Within-Cluster-Sum of Squared Errors sounds a bit complex.

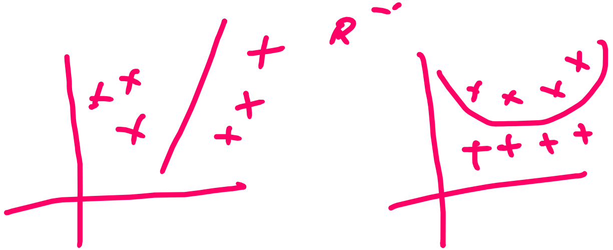
Let’s break it down:

- ➔ The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
- ➔ The WSS score is the sum of these Squared Errors for all the points.
- ➔ Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.


Optimal number of clusters
Elbow method
Total Within Sum of Square
200
150
100
50
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Advantages:

1. Can be applied to any form of data – as long as the data has numerical (continuous) entities.
2. Much faster than other algorithms.
3. Easy to understand and interpret.



Drawbacks:

1. Fails for non-linear data.
2. It requires us to decide on the number of clusters before we start the algorithm – where the user needs to use additional mathematical methods and also heuristic knowledge to verify the correct number of centers.
3. This cannot work for Categorical data.
4. Cannot handle outliers.