Unsupervised learning

Unsupervised learning

- Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabelled datasets.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

| Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6
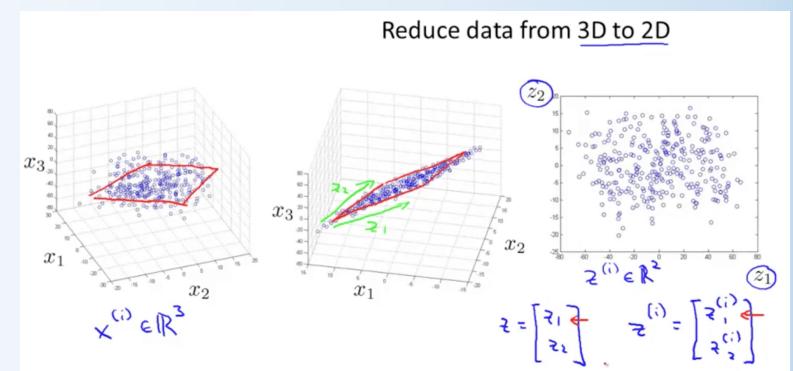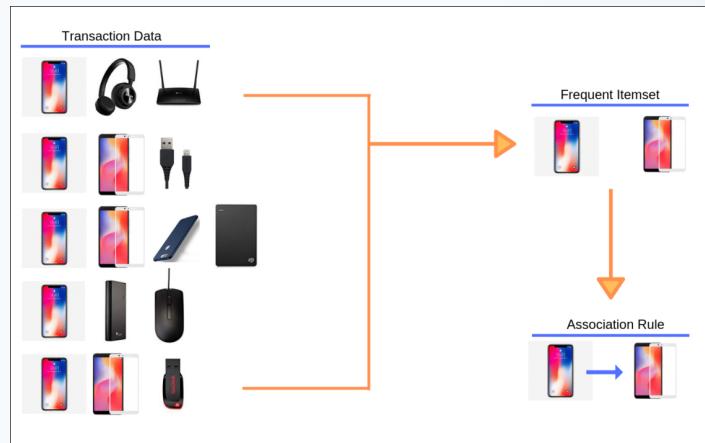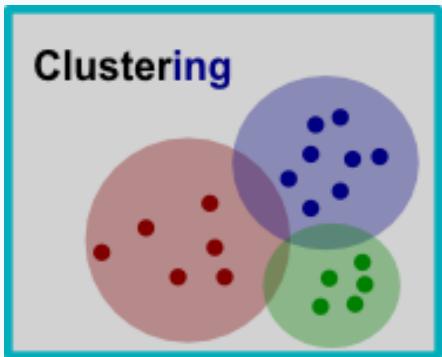
unlabeled

- These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

Common unsupervised learning approaches

Unsupervised learning models are utilized for three main tasks—

- Clustering

- Association

- Dimensionality reduction

Clustering

Clustering

- Clustering is a data mining technique which groups unlabelled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information.

- Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

Exclusive and Overlapping Clustering

- Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster.

- This can also be referred to as "hard" clustering. The K-means clustering algorithm is an example of exclusive clustering.

Clustering

K-means clustering

- It is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid.

- The data points closest to a given centroid will be clustered under the same category.

- A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity.

- K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.
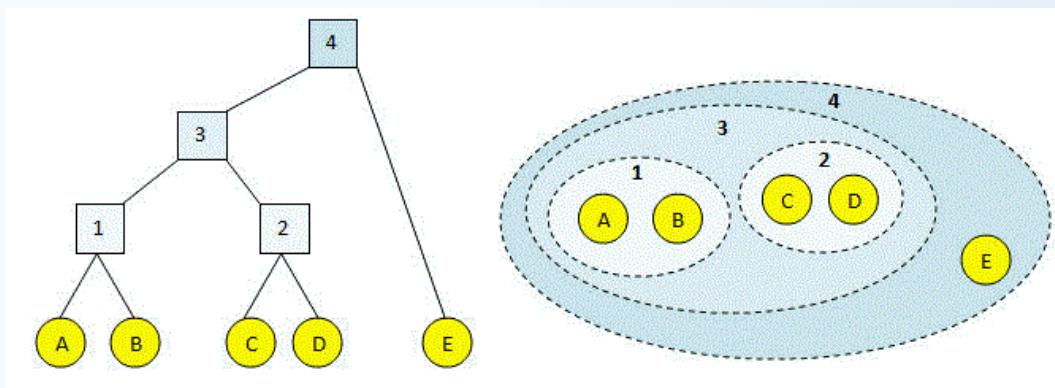
Overlapping clusters differs from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership.
"Soft" or fuzzy k-means clustering is an example of overlapping clustering.

Clustering

Hierarchical clustering

- Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways; they can be agglomerative or divisive.

- Agglomerative clustering is considered a "bottoms-up approach." Its data points are isolated as separate groupings initially, and then they are merged together iteratively on the basis of similarity until one cluster has been achieved.

Association Rules

Association Rules

- An association rule is a rule-based method for finding relationships between variables in a given dataset.

- These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.

- Understanding consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines.

- Examples of this can be seen in Amazon's "Customers Who Bought This Item Also Bought" or Spotify's "Discover Weekly" playlist. While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

Dimensionality Reduction

Dimensionality reduction

- While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. overfitting) and it can also make it difficult to visualize datasets.

- Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high.

- It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible.

- It is commonly used in the preprocessing data stage.

Dimensionality Reduction

Principal component analysis

- Principal component analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and to compress datasets through feature extraction.

- This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset.

- While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component.

- This process repeats based on the number of dimensions, where a next principal component is the direction orthogonal to the prior components with the most variance.

K Means Clustering

K-Means Algorithm

K-Means Clustering Algorithm

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

- In K-Means algorithm we group the unlabelled dataset into different clusters.

- Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

- It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

K-Means Algorithm
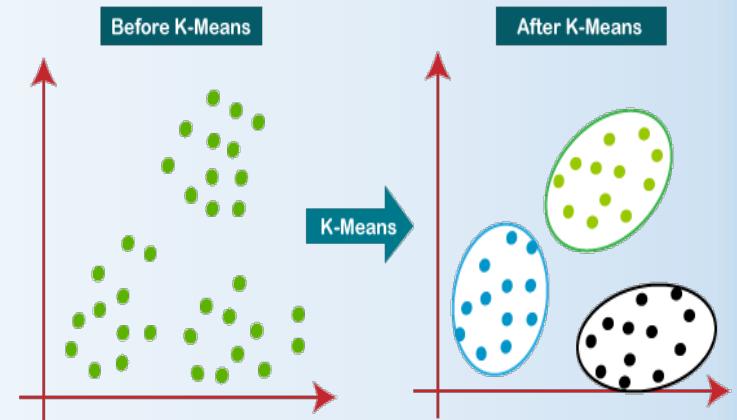
K-Means Clustering Algorithm

- The algorithm takes the unlabelled dataset as input, divides the dataset into

  k-number of clusters, and repeats the process until it does not find the best

  clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative

  process.

- Assigns each data point to its closest k-center. Those data points which are

  near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away

from other clusters.

Before K-Means

After K-Means

K-Means

Working of K-Means Algorithm

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.
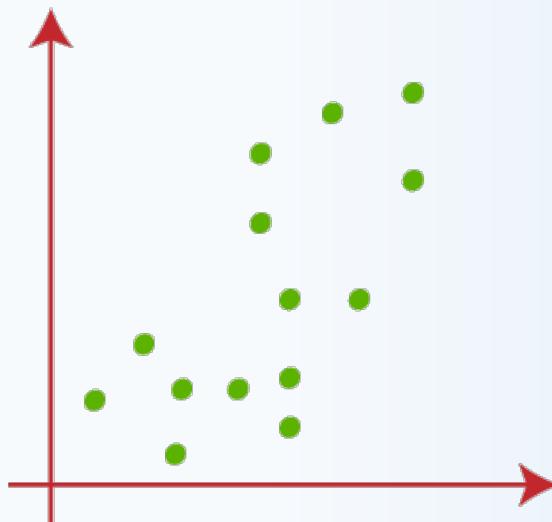
Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to finish.

Step-7: The model is ready.

Working of K-Means Algorithm with Example

Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:

Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try

to group these datasets into two different clusters.
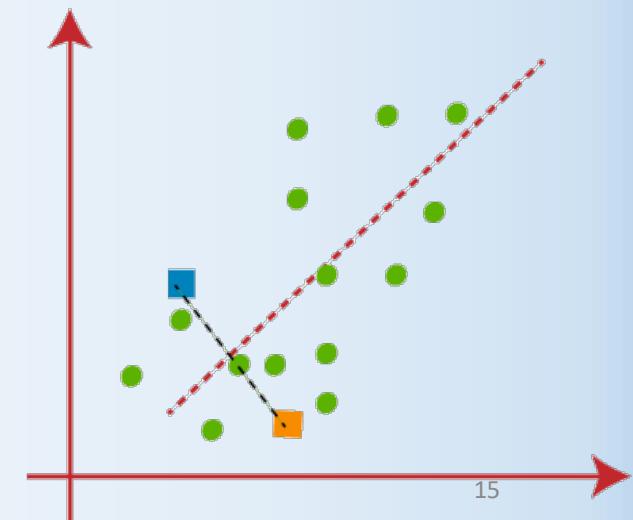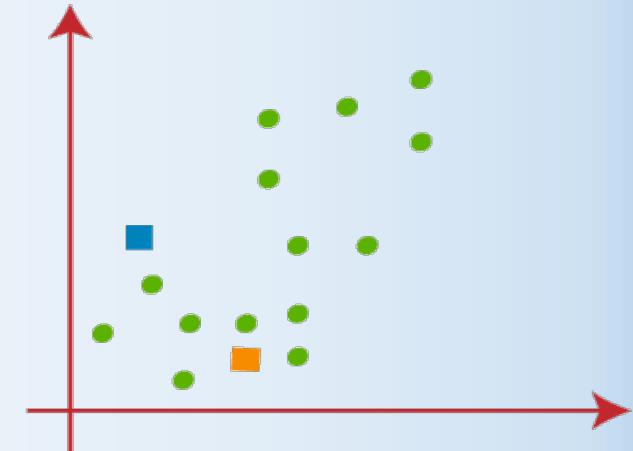
Working of K-Means Algorithm with Example

We need to choose some random k points or centroid to form the cluster.

These points can be either the points from the dataset or any other point. So, here we are selecting the two points as k points, which are not the part of our dataset.
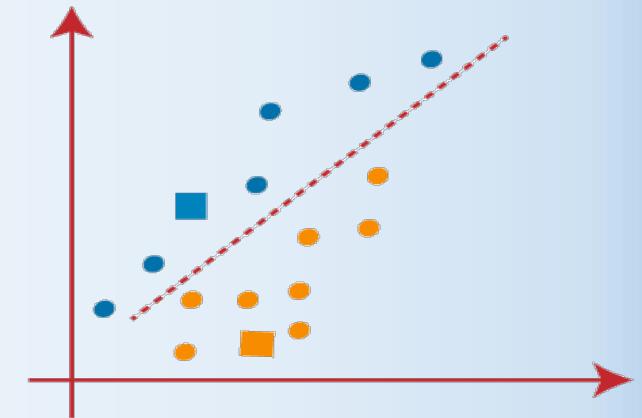
Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points.

So, we will draw a median between both the centroids.
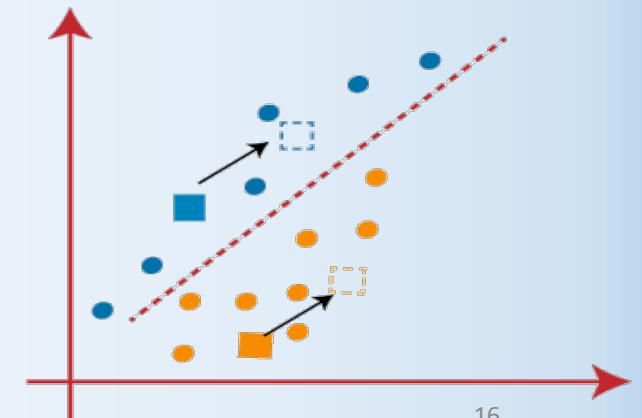
Working of K-Means Algorithm with Example

From the above steps, it is clear that points left side of the line is near to the

K1 or blue centroid, and points to the right of the line are close to the yellow

centroid. Let's color them as blue and yellow for clear visualization.
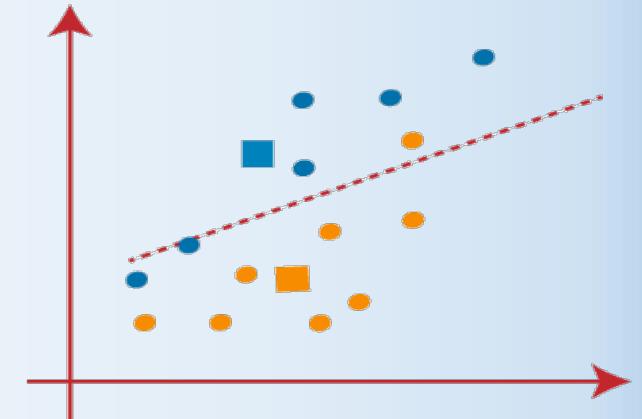
As we need to find the closest cluster, so we will repeat the process by

choosing a new centroid. To choose the new centroids, we will compute the

center of gravity of these centroids, and will find new centroids as below:
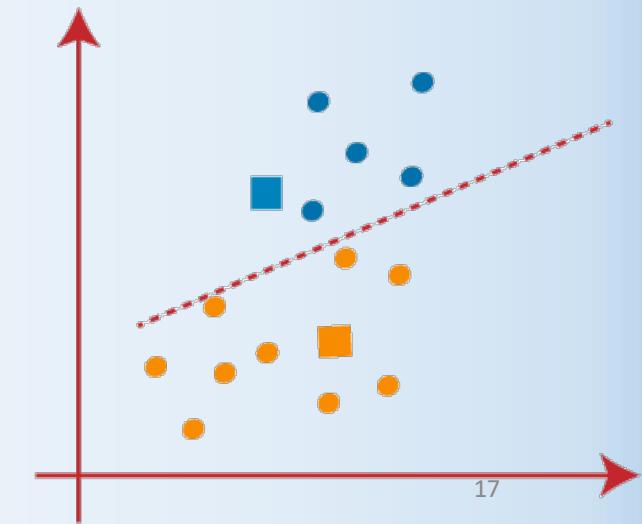
Working of K-Means Algorithm with Example

Next, we will reassign each datapoint to the new centroid. For this, we will

repeat the same process of finding a median line. The median will be like

adjacent image:

From the previous step, we can see, one yellow point is on the left side of the

line, and two blue points are right to the line. So, these points will be assigned

to new centroids.
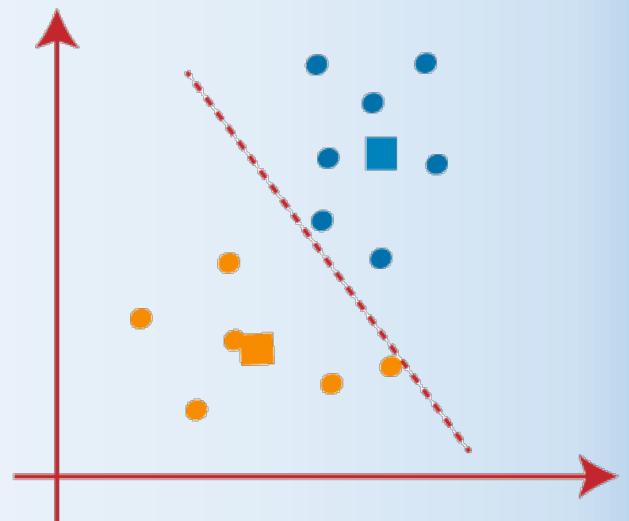
As reassignment has taken place, so we will again go to the step-4, which is

finding new centroids or K-points.

We will repeat the process by finding the center of gravity of centroids, so the

new centroids will be as shown in the adjacent image:
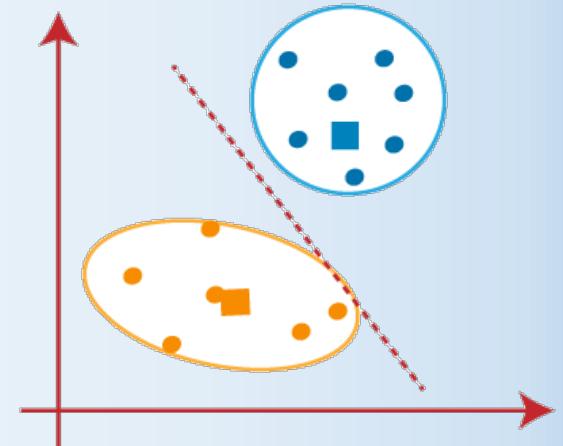
As we got the new centroids so again will draw the median line and reassign

the data points. So, the image will be:
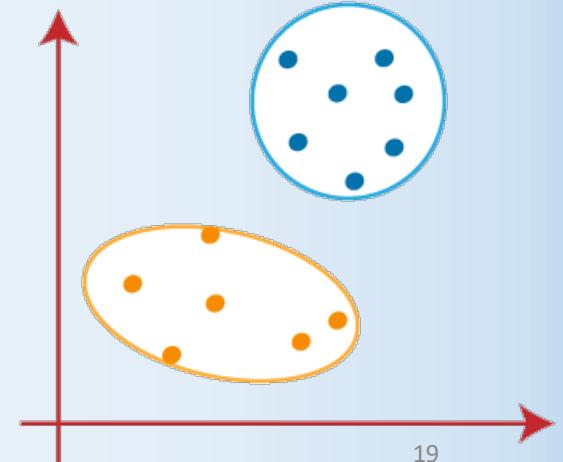
Working of K-Means Algorithm with Example

We can see in the above step; there are no dissimilar data points on either

side of the line, which means our model is formed. Consider the adjacent

image:

As our model is ready, so we can now remove the assumed centroids, and the

two final clusters will be as shown in the adjacent image:

How to choose the value of K

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters.

Elbow Method

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value.

WCSS stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

Steps to choose the value of K
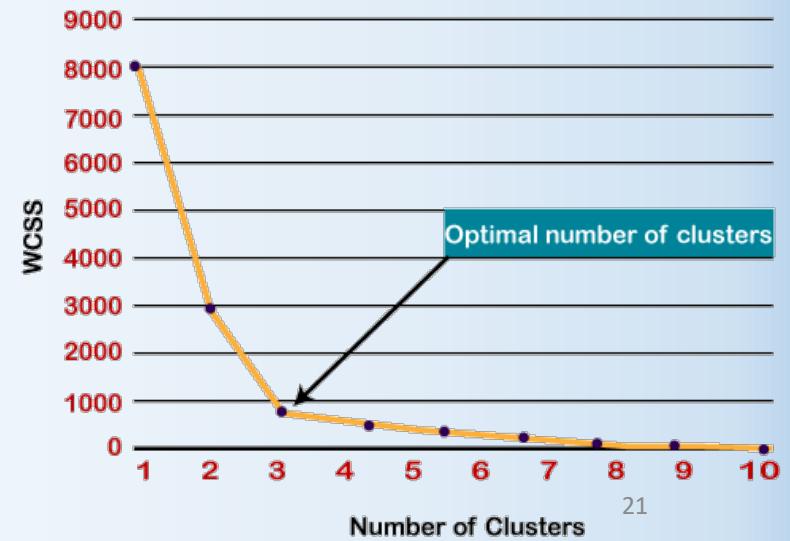
To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).

- For each value of K, calculates the WCSS value.

- Plots a curve between calculated WCSS values and the number of clusters K.

- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow,

hence it is known as the elbow method.

The graph for the elbow method looks like the adjacent image:

Important Hyperparameters

n_clusters : int, default=8

The number of clusters to form as well as the number of centroids to generate.

Init {'k-means++', 'random'}

'k-means++' : selects initial cluster centroids using sampling based on an empirical probability distribution of the points'

contribution to the overall inertia.

'random': choose n_clusters observations (rows) at random from data for the initial centroids.
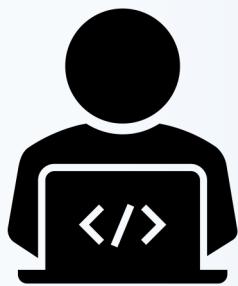
n_init : int, default=10

Number of time the k-means algorithm will be run with different centroid seeds.

Application of Clustering

Clustering is widely used in many industries. Below are some commonly known applications of clustering technique in Machine Learning:

- **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.

- **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.

- **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.

- **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

Keep Learning…………… Keep Coding…………… Keep going……………