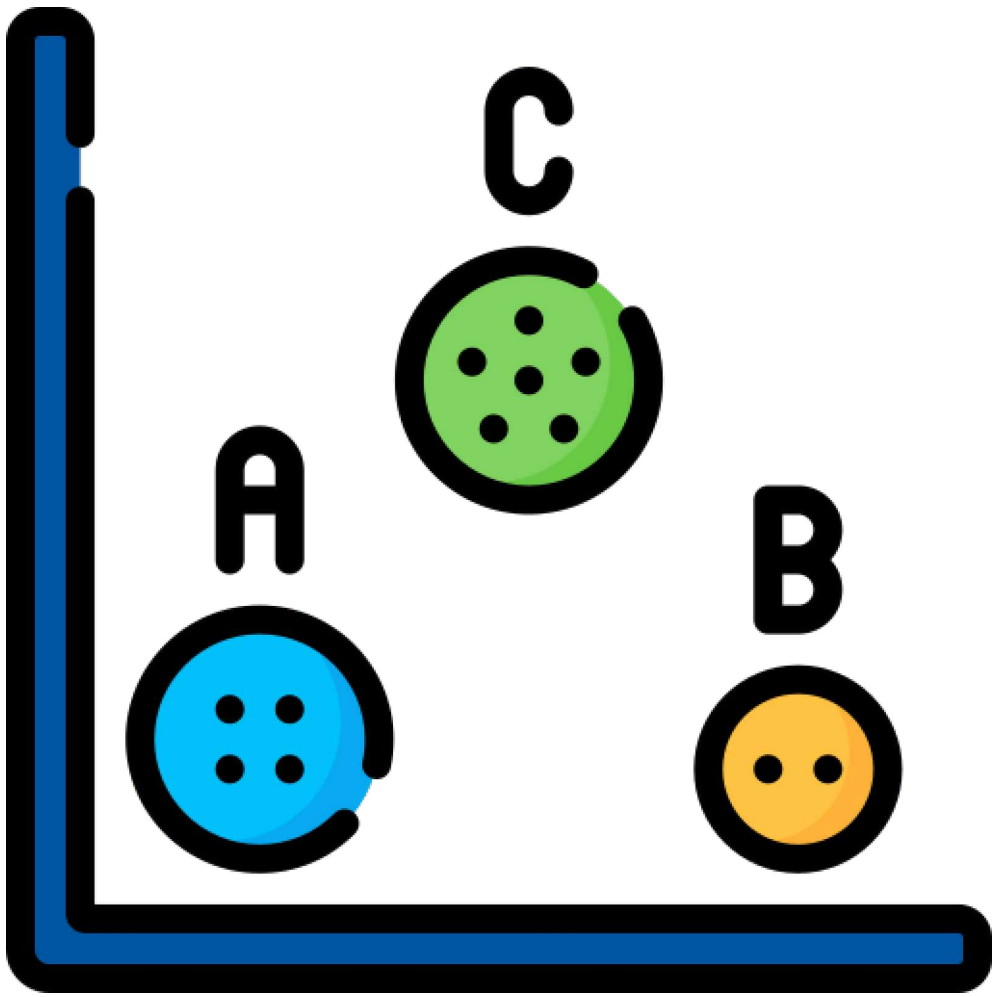
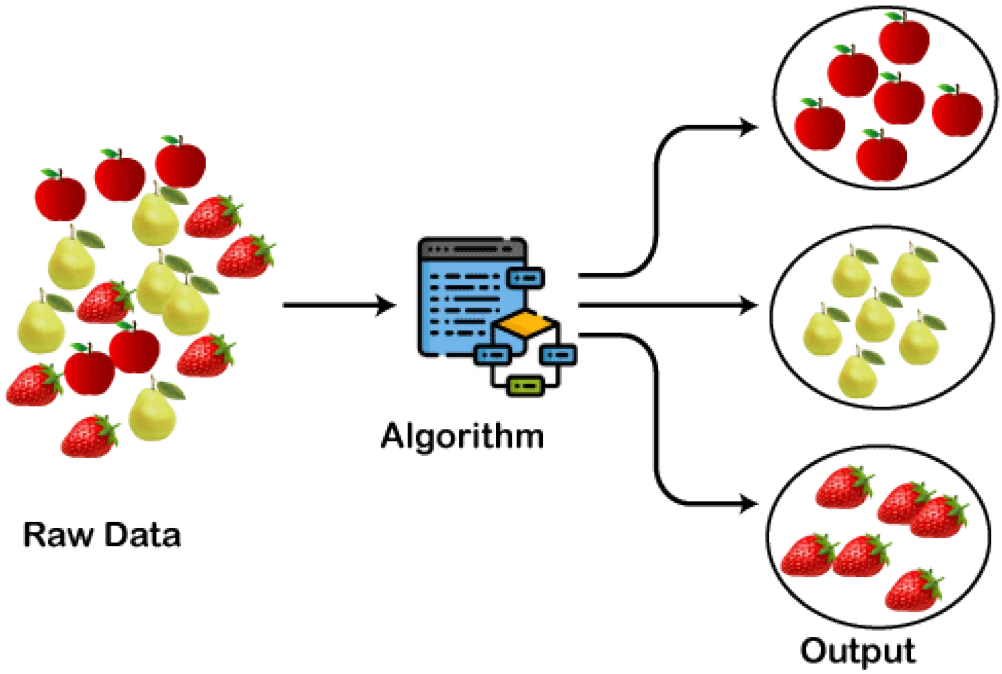
Clustering



What is Clustering?



Clustering is the most popular version of unsupervised learning.

In unsupervised learning, the goal is to identify patterns or structures in the data without any prior knowledge of what to expect.

In Clustering, the goal is to group data points based on their similarity.

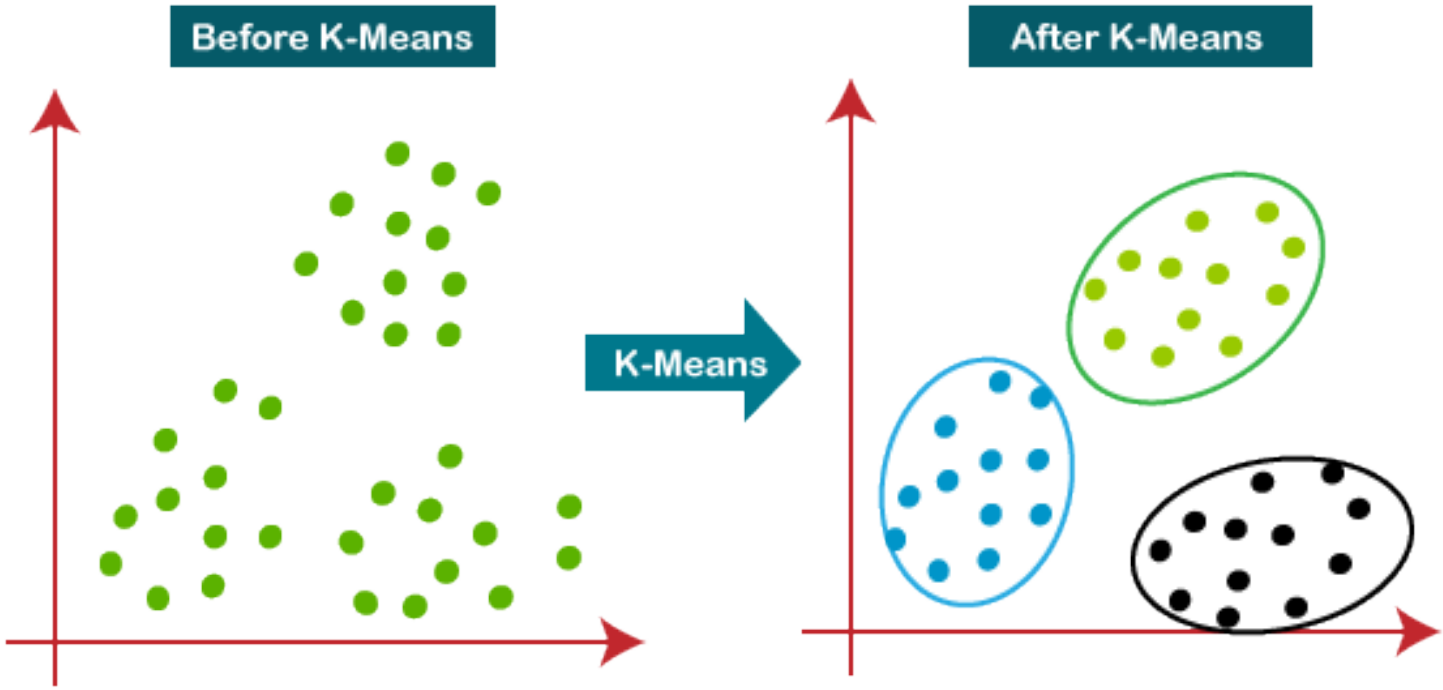
Why it is useful?

Suppose you are the head of a retail store and wish to understand the preferences of your customers. Can you look at the details of each customer and devise a unique business strategy for each one of them?
What you can do is cluster all of your customers into, say 5 groups based on their purchasing habits and use a separate strategy for each group.

Desirable Properties of a Clustering Algorithm

1. Scalability (in terms of both time and space)
2. Ability to deal with different data types
3. Minimal requirements for domain knowledge to determine input parameters
4. Able to deal with noise and outliers
5. Insensitive to the order of input records
6. Incorporation of user-specified constraints
7. Interpretability and usability

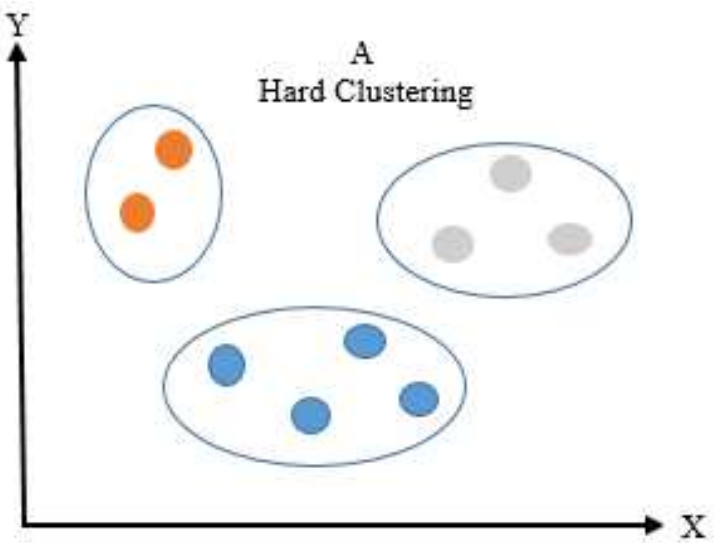
Clustering Algorithms

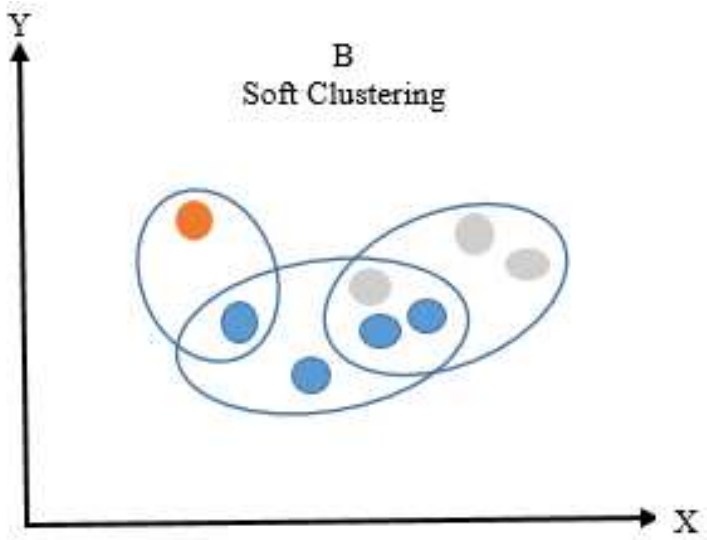


a. Exclusive Clustering

Exclusive clustering is a form of grouping that requires a data point to 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 Algorithms

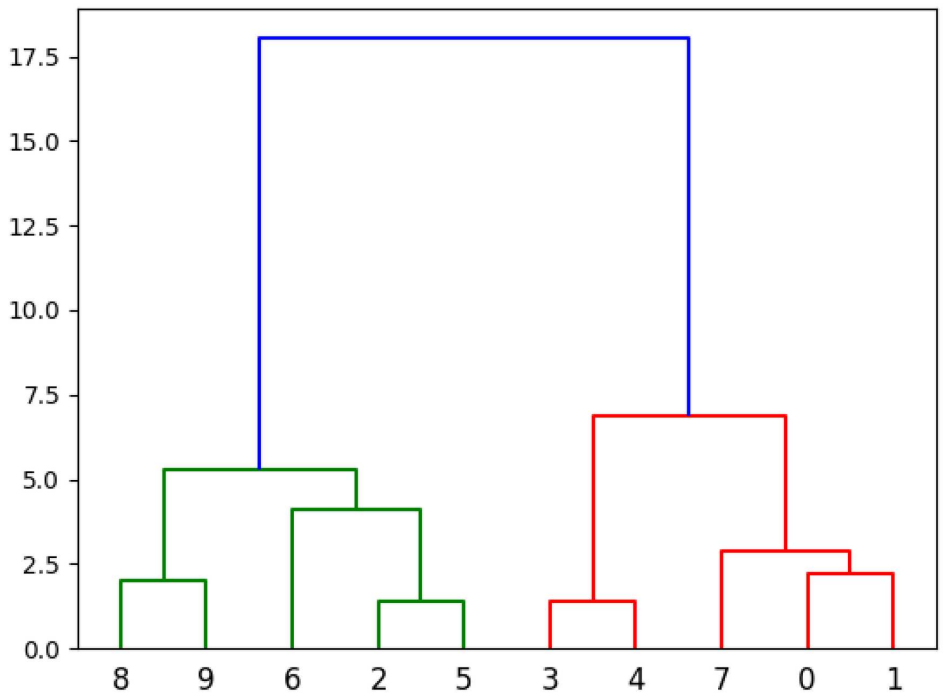




b. Overlapping Clustering

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.

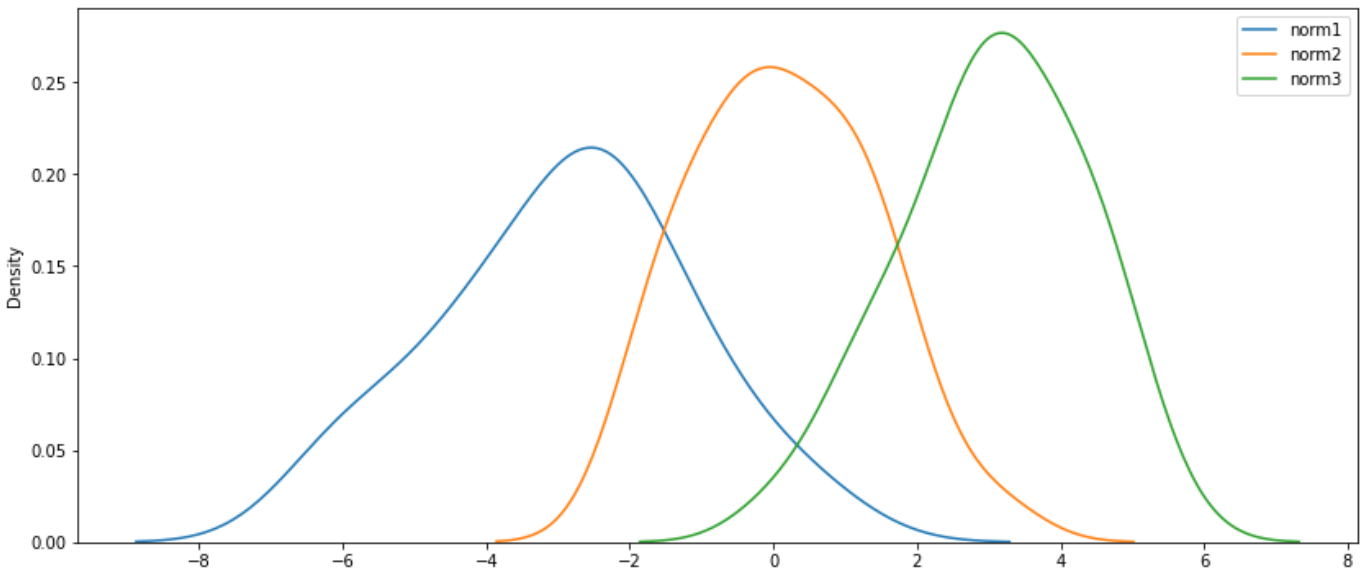
Clustering Algorithms



c. Hierarchical Clustering

Hierarchical clustering can be categorized in two ways; 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.

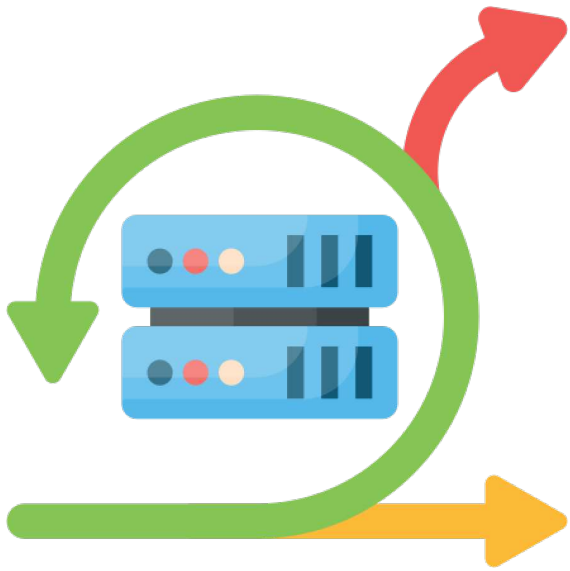
Clustering Algorithms



d. Probabilistic Clustering

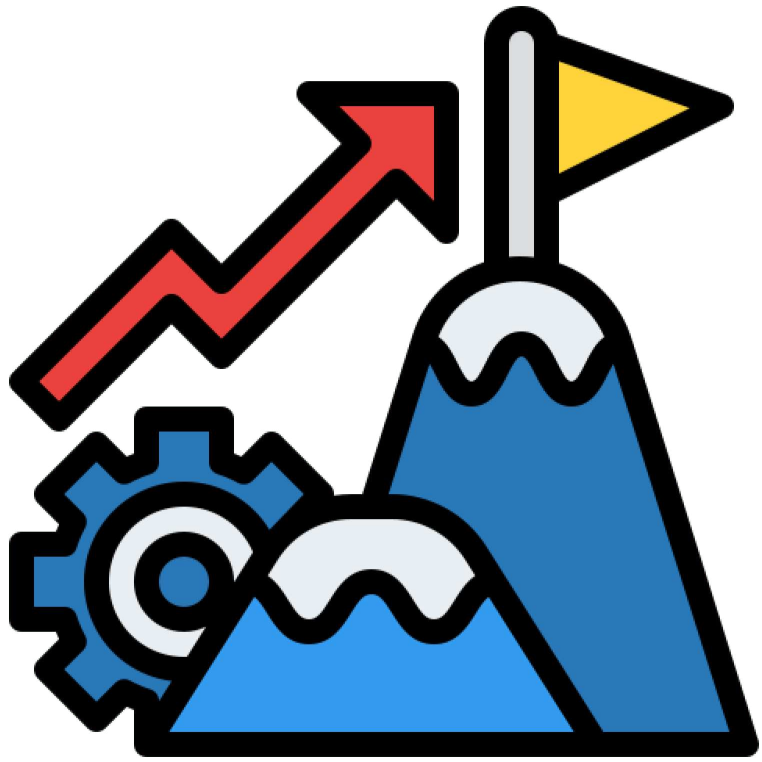
In probabilistic clustering, data points are clustered based on the likelihood that they belong to a particular distribution. The Gaussian Mixture Model (GMM) is one of the most commonly used probabilistic clustering methods.

Applications of Clustering



1. Marketing: To characterize & discover customer segments for marketing purposes.
2. Biology: For classification among different species of plants and animals.
3. Libraries: Clustering different books on the basis of topics and information.
4. City Planning: To make groups of houses and to study their values based on their geographical locations and other factors present.

Challenges



1. Computational complexity due to a high volume of training data.

2. Longer training times

3. Higher risk of inaccurate results

4. Human intervention to validate output variables

5. Lack of transparency into the basis on which data was clustered

Follow **#DataRanch** on LinkedIn for more...

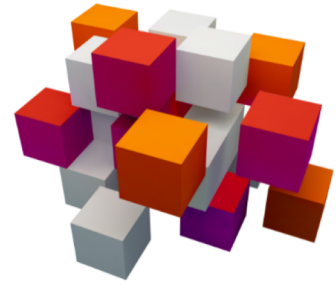
What is Supervised Learning?

DATARANCH.org

What is Unsupervised Learning?

DATARANCH.org

Data Wrangling Steps



DATARANCH.org

Common data fallacies to watch out for...

DATARANCH.org

Data Cleaning Steps

DATARANCH.org

Data Analysis Steps

DATARANCH.org

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

info@dataranch.org

linkedin.com/company/dataranch