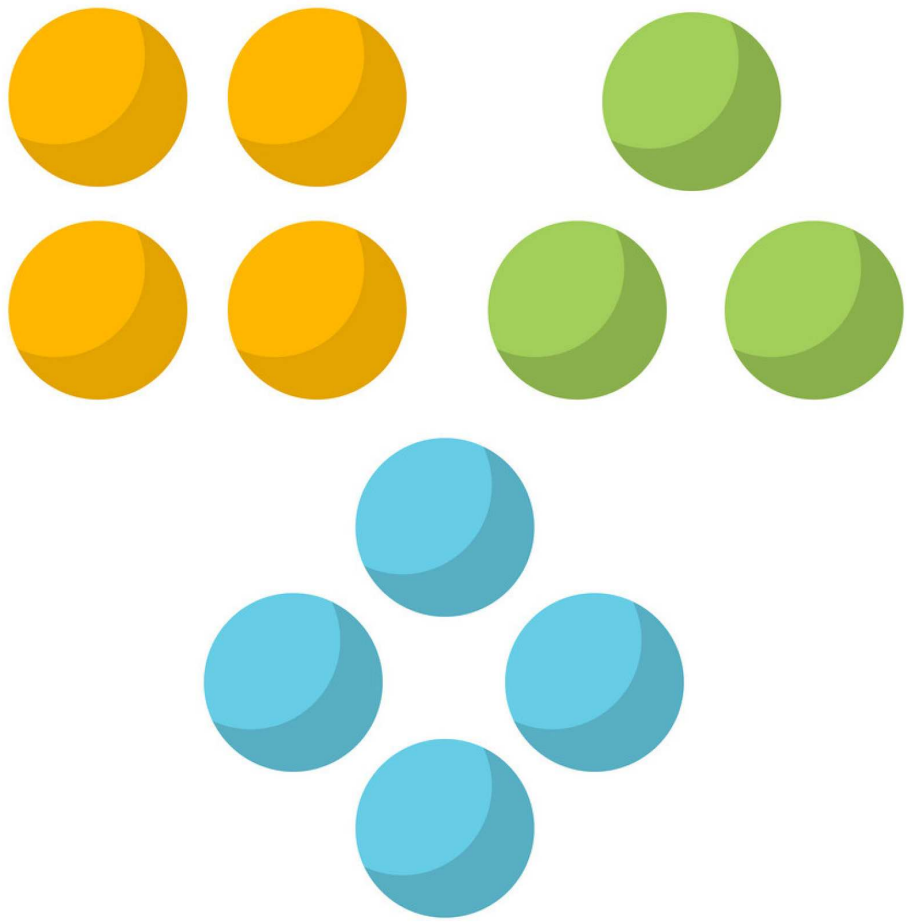
What is Unsupervised Learning?



DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

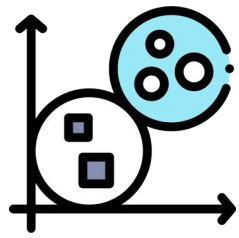
Unsupervised Learning

In unsupervised learning, an algorithm is trained on a dataset without any labeled output or target variable. The goal of unsupervised learning is to identify patterns or structures in the data without any prior knowledge of what to expect.

Unsupervised learning algorithms are ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

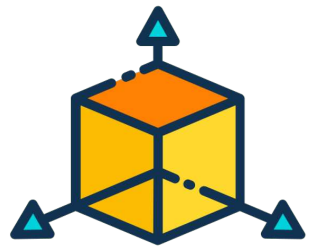
Categories of Unsupervised Learning

1. Clustering



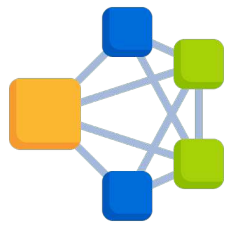
Clustering is a process of grouping similar data points together based on their similarity or differences.

2. Dimensionality Reduction



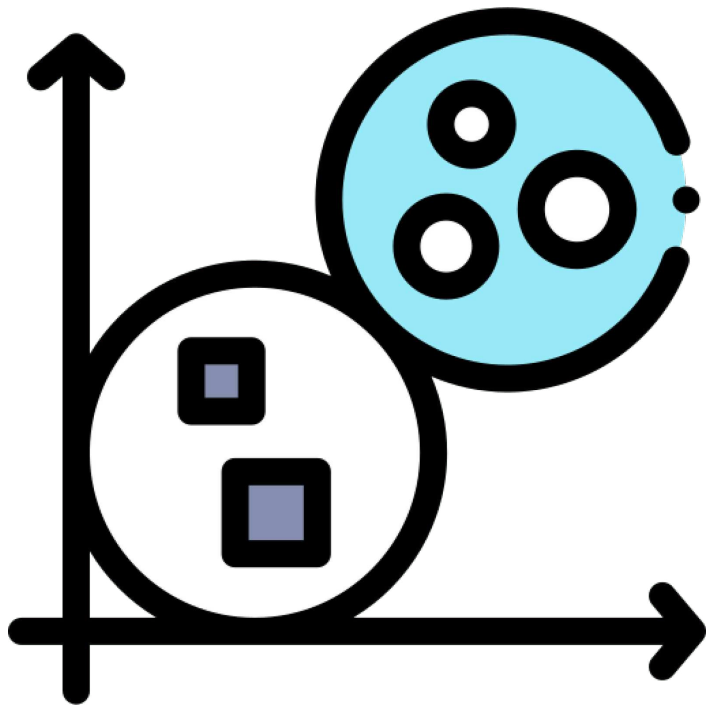
In dimensionality reduction, the number of data inputs are reduced to a manageable size while also preserving the integrity of the dataset as much as possible.

3. Association Rules



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

1. Clustering



Clustering is a process of grouping similar data points together based on their similarity or differences. The four main types of clustering algorithms are:

a. Exclusive Clustering
b. Overlapping Clustering
c. Hierarchical Clustering
d. Probabilistic Clustering

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.

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.

Divisive clustering can be defined as the opposite of agglomerative clustering; instead it takes a “top-down” approach. In this case, a single data cluster is divided based on the differences between data points.

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 the one of the most commonly used probabilistic clustering methods.

2. Dimensionality Reduction

Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset are too high. It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset. The three main types of dimensionality reduction:

a. Principal component analysis
b. Singular value decomposition
c. Autoencoders

Dimensionality Reduction Algorithms

a. Principal Component Analysis

Principal component analysis (PCA) 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.

Dimensionality Reduction Algorithms

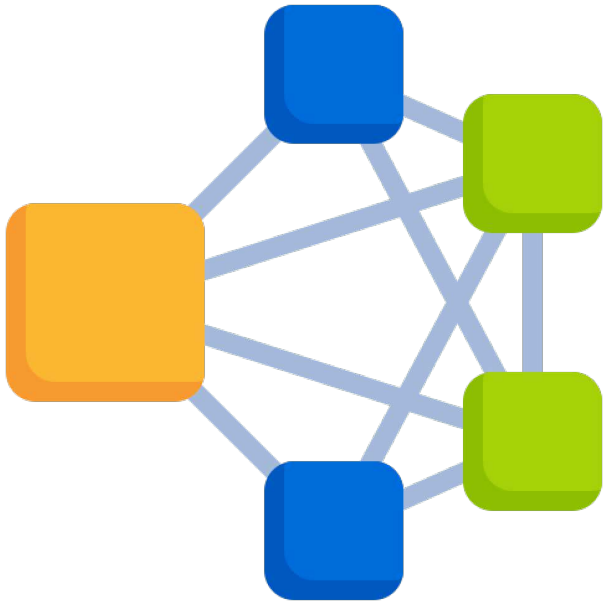
b. Singular Value Decomposition

Singular value decomposition (SVD) is another dimensionality reduction approach which factorizes a matrix, A, into three, low-rank matrices. SVD is denoted by the formula, A = USVT, where U and V are orthogonal matrices. S is a diagonal matrix, and S values are considered singular values of matrix A.

c. Autoencoders

Autoencoders leverage neural networks to compress data and then recreate a new representation of the original data's input. The stage from the input layer to the hidden layer is referred to as "encoding" while the stage from the hidden layer to the output layer is known as "decoding."

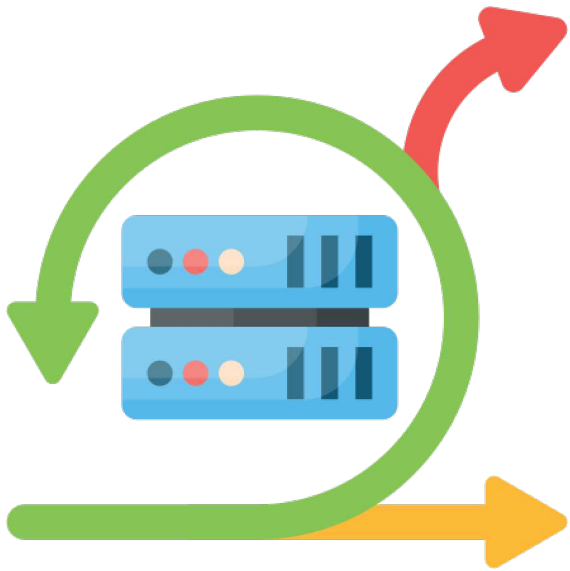
3. Association Rules



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.

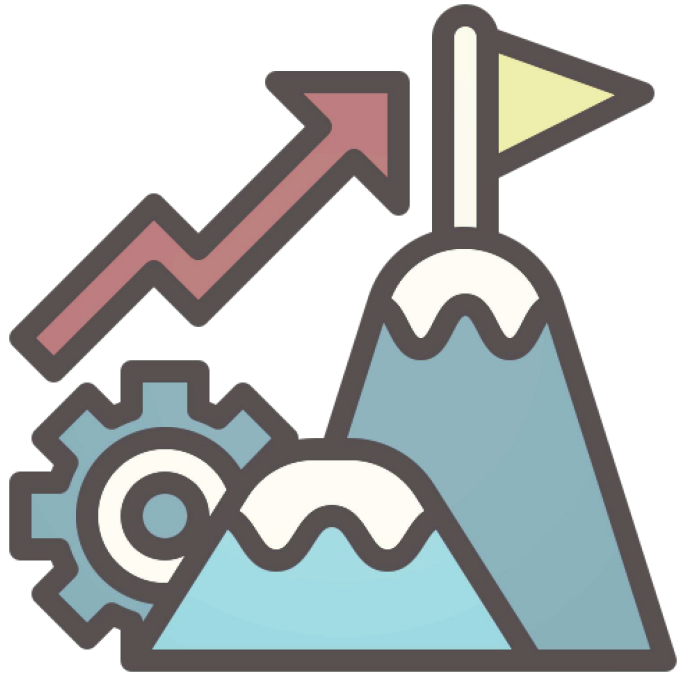
While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

Usage of Unsupervised Learning



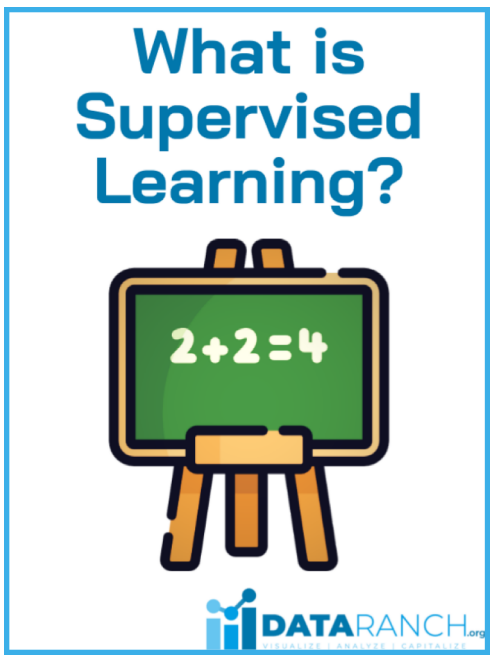
a. Computer vision
b. Medical imaging
c. Anomaly detection
d. Customer personas
e. Recommendation engines
f. News sections

Challenges of Supervised Learning



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** for more...



DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

info@dataranch.org

linkedin.com/company/dataranch