InTrAinZ Internship

K-Means Clustering

Internship : Artificial Intelligence

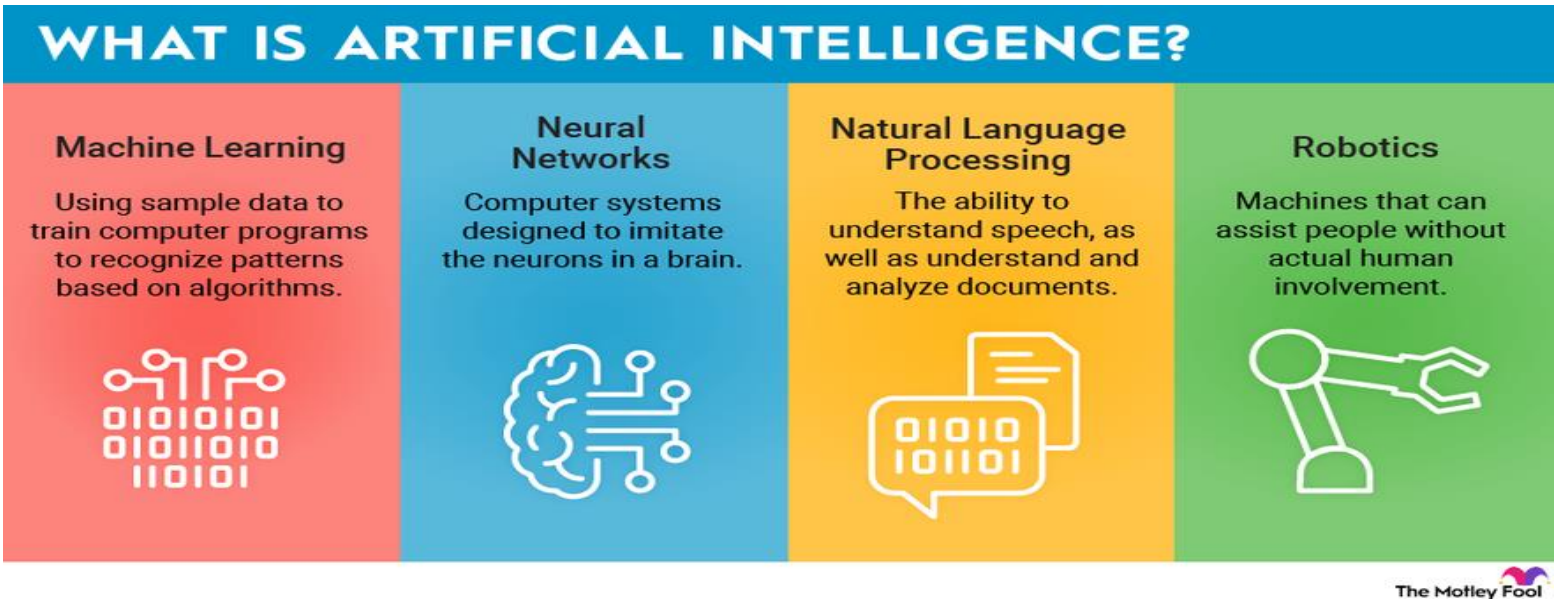
Name : Y Pradeep Reddy College : AITS Tirupati Course : AIML

Topics / Agenda :-

- Artificial Intelligence and Machine Learning
- Types of Machine Learning
- Introduction to Clustering
- Understanding K-Means Clustering
- Simple k-Means clustering program
- Applications of K-Means Clustering
- Strengths and weakness of k-Means

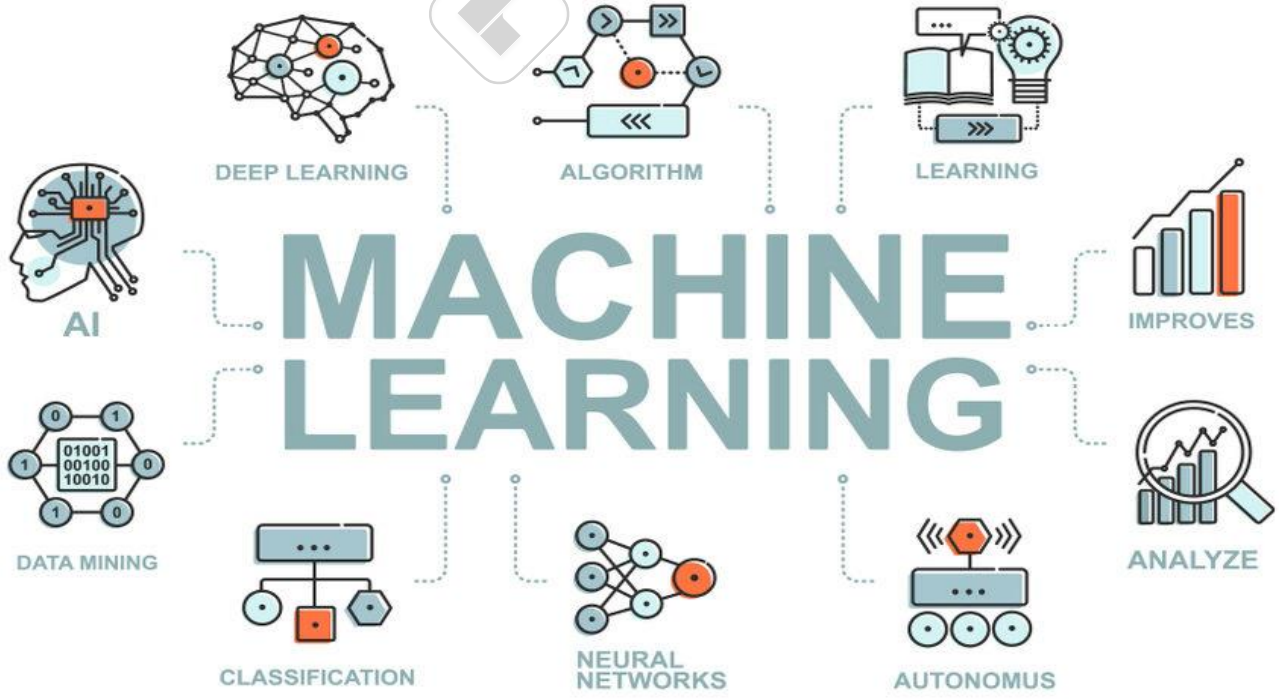
Artificial Intelligence:

Artificial intelligence is the science of making machines that can think like humans. It can do things that are considered "smart." AI technology can process large amounts of data in ways, unlike humans. The goal for AI is to be able to do things such as recognize patterns, make decisions, and judge like humans.



Machine Learning:

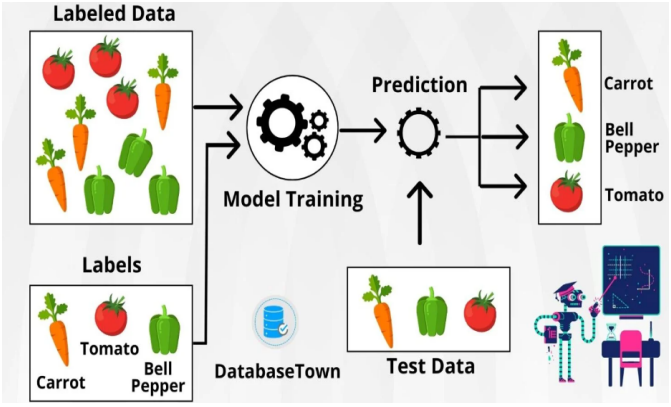
What is machine learning? Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. IBM has a rich history with machine learning.

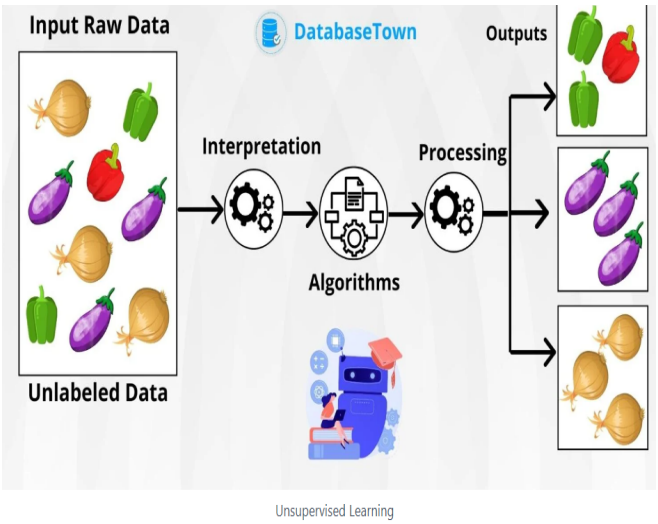


Types of machine learning:

Supervised Machine learning:

supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.



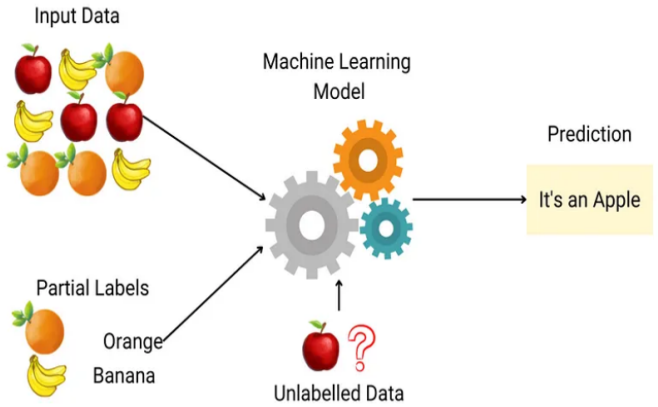


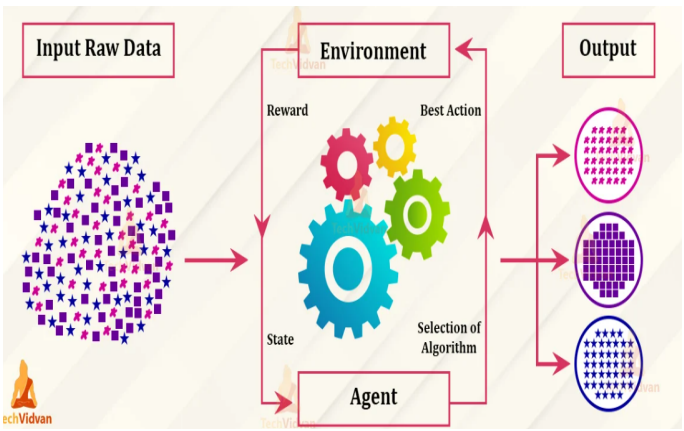
Unsupervised Machine learning:

Unsupervised learning in artificial intelligence is a type of machine learning that learns from data without human supervision. Unlike supervised learning, unsupervised machine learning models are given unlabeled data and allowed to discover patterns and insights without any explicit guidance or instruction.

Semi-supervised Machine learning:

Semi-supervised learning is a broad category of machine learning that uses labeled data to ground predictions, and unlabeled data to learn the shape of the larger data distribution. Practitioners can achieve strong results with fractions of the labeled data, and as a result, can save valuable time and money.





Reinforcement Learning

In reinforcement learning, developers devise a method of rewarding desired behaviors and punishing negative behaviors . This method assigns positive values to the desired actions to encourage the agent to use them, while negative values are assigned to undesired behaviors to discourage them.

Introduction :

Overview of Data Analysis:

Data analysis is a crucial aspect of deriving meaningful insights from raw data. Clustering is a specific technique within data analysis that focuses on grouping similar data points together based on certain characteristics or features.

Purpose of Clustering in Data Analysis:

Clustering serves multiple purposes, including pattern recognition, data compression, and efficient data organization. By grouping similar entities, it simplifies the interpretation of complex datasets, facilitating more effective decision-making processes.

Types of CLustering :

Hierarchical Clustering:

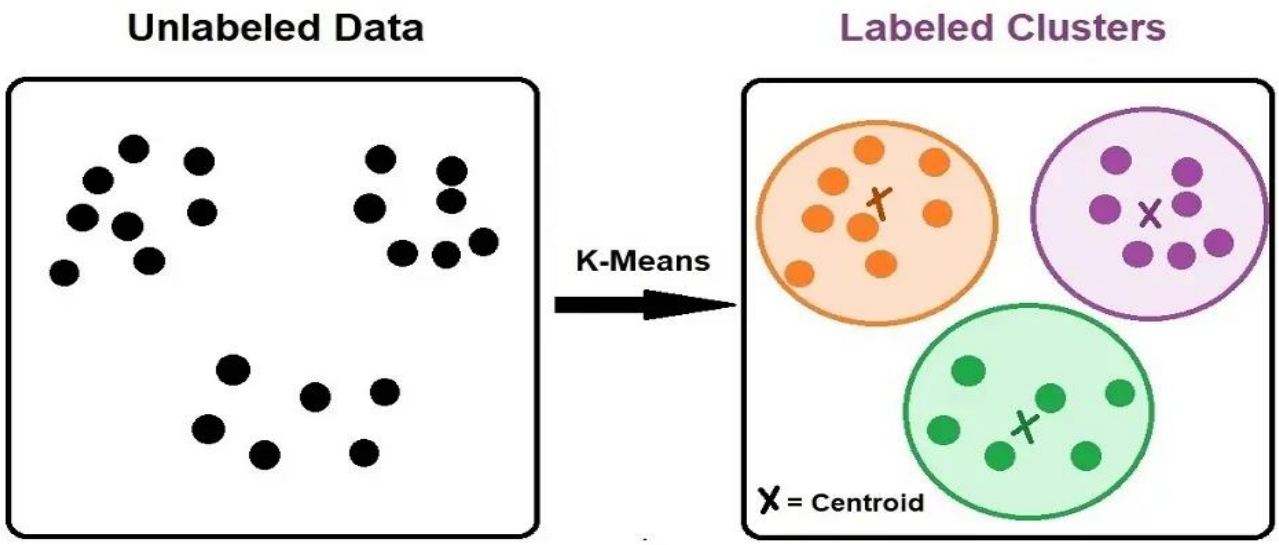
In hierarchical clustering, data points are grouped in a tree-like structure, creating a hierarchy of clusters. This method provides a nuanced understanding of relationships between data points at different levels.

Partitioning Clustering:

Partitioning clustering involves dividing the dataset into distinct , non-overlapping clusters. K-Means, a popular partitioning algorithm, assigns data points to clusters based on their proximity to centroid values.

Density-Based Clustering:

Density-based clustering identifies clusters based on the density of data points in a given region. Algorithms like DBSCAN are particularly useful in discovering clusters of varying shapes and sizes.



Understanding K-Means Clustering :

K-Means is a popular partitioning clustering algorithm that organizes data points into K distinct groups. It is widely used for its simplicity and efficiency in handling large datasets. K-Means has applications in various fields, including machine learning, image analysis , and data mining.

How K-Means Works :

Initialization of Centroids:

K-Means begins by randomly initializing K centroids , which are the representative points for each cluster. These centroids act as the centers of the clusters that will be formed.

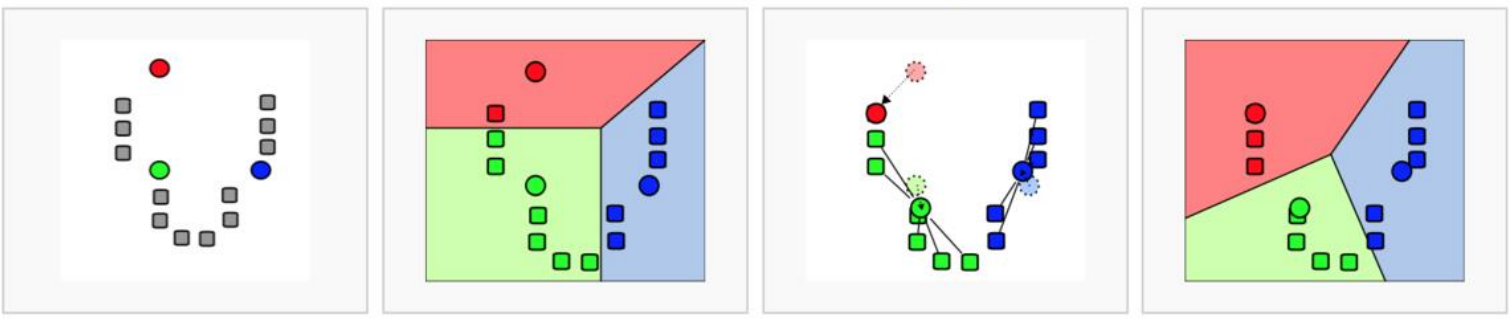
Assignment of Data Points to Clusters:

Data points are assigned to the cluster whose centroid is closest to them. This step is repeated until all data points are assigned, and clusters are formed.

Update of Centroids:

After data points are assigned to clusters, the centroids are recalculated as the mean of all data points within each cluster. This process is iteratively performed until convergence, optimizing the placement of centroids.

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

Simple K-Means Clustering Program:

Python code for clustering the data:
Libraries Used :

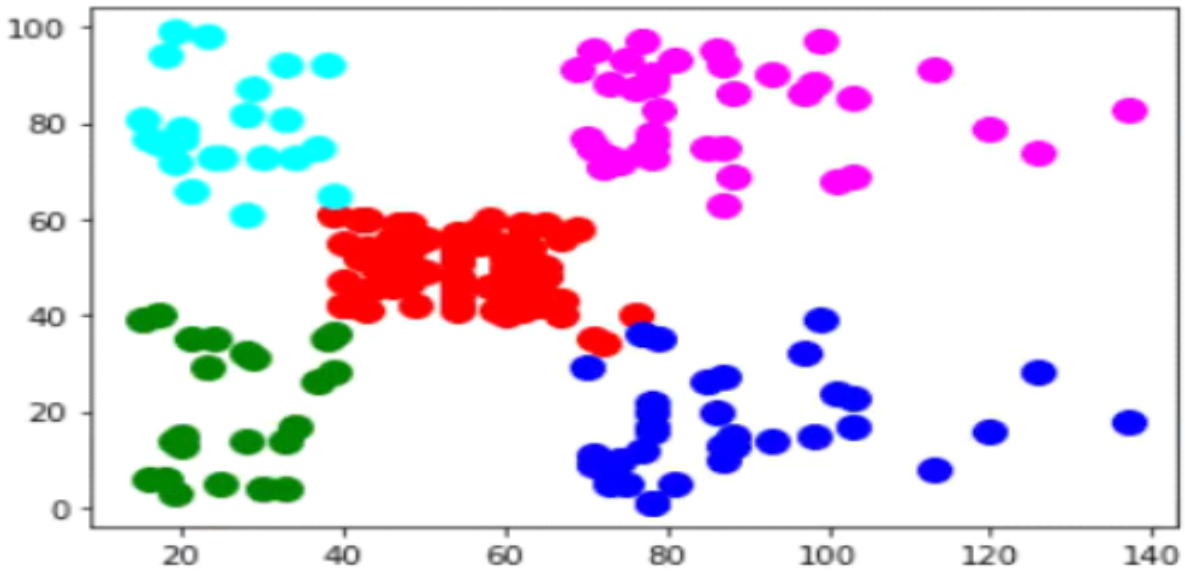
- numpy
- pandas
- sklearn

The required data should be save with ".csv" extension.

Program:

```
import numpy as np
import pandas as pd
#loading the data into "dataset"
dataset = pd.read_csv("filename.csv")
x=dataset.iloc[[: , [3,4]].values
#Elbow method
from sklearn.cluster import KMeans
wcss = []
for in in range(1, 11):
 kmeans= KMeans(n_clusters = i , random_state = 42)
 kmeans.fit(x)
 wcss.append(kmeans.inertia_)
import matplotlib.pyplot as plt
#grouping and plotting the data
plt.plot(range(2, 31), wcss)
kmeans=KMeans(n_clusters = 5, init = "k-means++", random_
state = 42)
y_means = kmeans.fit_predict(x)
plt.scatter(x[y_means == 0, 0], x[y_means ==0, 1], s=100,
c= 'red', label='cluster 1')
plt.scatter(x[y_means == 1, 0], x[y_means ==1, 1],
s=100, c= 'blue', label='cluster 2')
plt.scatter(x[y_means == 2, 0], x[y_means ==2, 1], s=100,
c= 'green', label='cluster 3')
plt.scatter(x[y_means == 3, 0], x[y_means ==3, 1], s=100,
c= 'cyan', label='cluster 4')
plt.scatter(x[y_means == 4, 0], x[y_means ==4, 1], s=100
, c= 'magenta', label='cluster 5')
```

Output:



Applications of K-Means Clustering:

Customer Segmentation:

K-means clustering is an algorithm that can be used to segment customers based on their similarities. The algorithm aims to find k clusters in the data, where each cluster represents a group of customers that are similar to each other

Anomaly Detection:

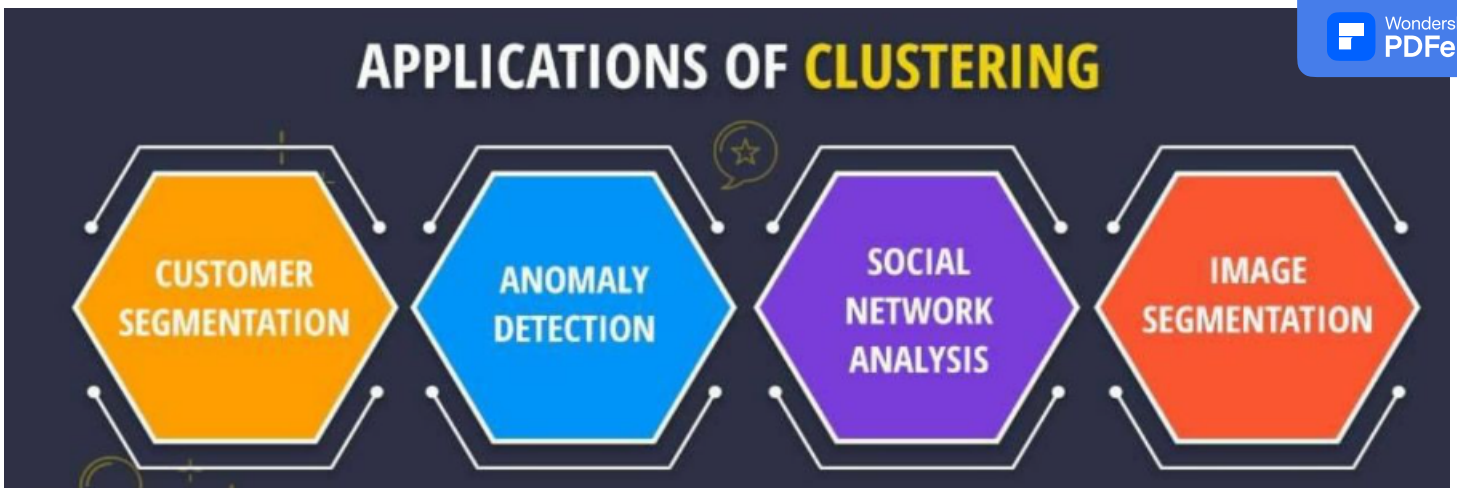
After fitting the K-Means model, the cluster labels for each data point are predicted and counted. A scatter plot can then be created to visualize the clusters along with the centroid. Next, the code calculates the anomaly scores by computing the Euclidean distance from each point to its nearest cluster center.

Image Segmentation:

K -means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background. But before applying K -means algorithm, first partial stretching enhancement is applied to the image to improve the quality of the image.

Social Network Analysis:

The idea behind K-means clustering is to divide a dataset into a specified number of clusters (k), where all the points within the same cluster are similar to one another, and those in different clusters are different.



Strengths and weakness of k-Means:

Strengths of k-means

• Strengths:

– Simple: easy to understand and to implement

– Efficient: Time complexity: O(tkn),

where n is the number of data points, k is the number of clusters, and t is the number of iterations.

– Since both k and t are small. k-means is considered a linear algorithm.

• K-means is the most popular clustering algorithm.

• Note that: it terminates at a local optimum if SSE is used. The global

optimum is hard to find due to complexity.

Weakness of k-means:

• The algorithm is only applicable if the mean is defined.

– For categorical data, k-mode - the centroid is represented by most frequent values.

• The user needs to specify k.

• The algorithm is sensitive to outliers

– Outliers are data points that are very far away from other data points. – Outliers

could be errors in the data recording or some special data points with very different values.

Conclusion:

In conclusion, K-means clustering is a powerful unsupervised machine learning algorithm for grouping unlabeled datasets. Its objective is to divide data into clusters, making similar data points part of the same group. The algorithm initializes cluster centroids and iteratively assigns data points to the nearest centroid, updating centroids based on the mean of points in each cluster.

Thanking you

Y Pradeep reddy