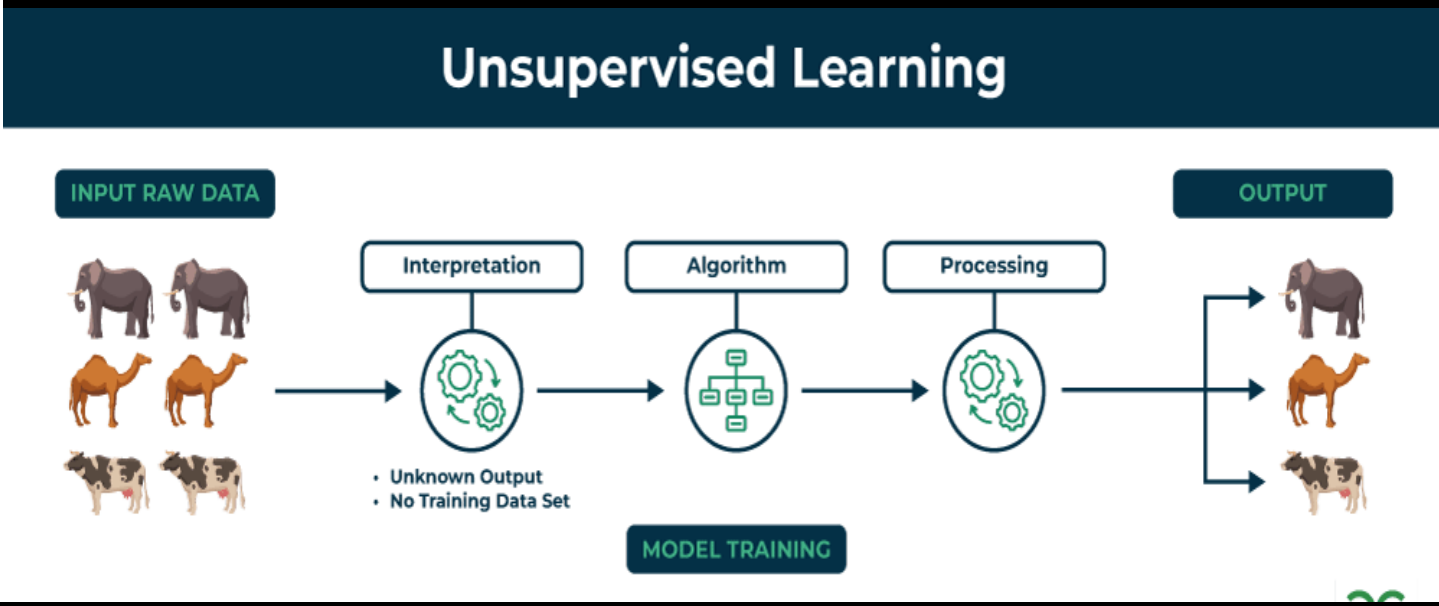
Artificial Intelligence Fullstack
[Course]
Week 11 – Unsupervised Learning – Clustering Techniques (K-Means, Hierarchical, DBSCAN)

[See examples / code in GitHub code repository]

It is not about Theory, it is 20% Theory and 80% Practical – Technical/Development/Programming [Mostly Python based]

ML – What is Unsupervised Learning?

Unsupervised learning is a branch of machine learning that deals with unlabeled data. Unlike supervised learning, where the data is labeled with a specific category or outcome, unsupervised learning algorithms are tasked with finding patterns and relationships within the data without any prior knowledge of the data's meaning.



The image shows set of animals: elephants, camels, and cows that represents raw data that the unsupervised learning algorithm will process.

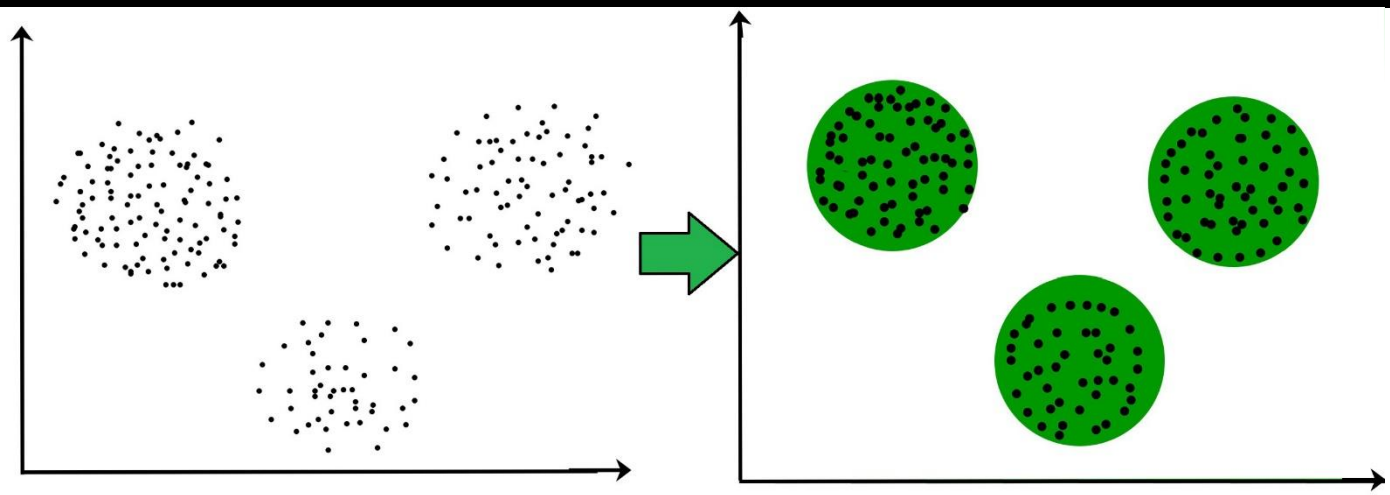
The "Interpretation" stage signifies that the algorithm doesn't have predefined labels or categories for the data. It needs to figure out how to group or organize the data based on inherent patterns.
Algorithm represents the core of unsupervised learning process using techniques like clustering, dimensionality reduction, or anomaly detection to identify patterns and structures in the data.
Processing stage shows the algorithm working on the data.

Reference:

https://www.geeksforgeeks.org/machine-learning/unsupervised-learning/
https://www.ibm.com/think/topics/unsupervised-learning
https://www.datacamp.com/blog/introduction-to-unsupervised-learning
https://www.statisticalaid.com/unsupervised-learning/

ML – Clustering in Machine Learning

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of unsupervised learning, which aims at gaining insights from unlabelled data points.



Types of Clustering

Hard Clustering: In this type of clustering, each data point belongs to a cluster completely or not. For example, Let's say there are 4 data point and we have to cluster them into 2 clusters. So each data point will either belong to cluster 1 or cluster 2.

Data Points	Clusters
A	C1
B	C2
C	C2
D	C1

ML – Soft Clustering

Soft Clustering: In this type of clustering, instead of assigning each data point into a separate cluster, a probability or likelihood of that point being that cluster is evaluated. For example, Let's say there are 4 data point and we have to cluster them into 2 clusters. So we will be evaluating a probability of a data point belonging to both clusters. This probability is calculated for all data points.

Data Points	Probability of C1	Probability of C2
A	0.91	0.09
B	0.3	0.7
C	0.17	0.83
D	1	0

Reference:
https://www.geeksforgeeks.org/machine-learning/clustering-in-machine-learning/

ML – K means Clustering

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called **flat clustering** algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means

Applications of K-Means Clustering

K-Means clustering is a versatile algorithm with various applications in several fields. Here we have highlighted some of the important applications −

Image Segmentation

K-Means clustering can be used to segment an image into different regions based on the color or texture of the pixels. This technique is widely used in computer vision applications, such as object recognition, image retrieval, and medical imaging.

Customer Segmentation

K-Means clustering can be used to segment customers into different groups based on their purchasing behavior or demographic characteristics. This technique is widely used in marketing applications, such as customer retention, loyalty programs, and targeted advertising.

Anomaly Detection

K-Means clustering can be used to detect anomalies in a dataset by identifying data points that do not belong to any cluster. This technique is widely used in fraud detection, network intrusion detection, and predictive maintenance.

Genomic Data Analysis

K-Means clustering can be used to analyze gene expression data to identify different groups of genes that are co-regulated or co-expressed. This technique is widely used in bioinformatics applications, such as drug discovery, disease diagnosis, and personalized medicine.

Reference:

https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/

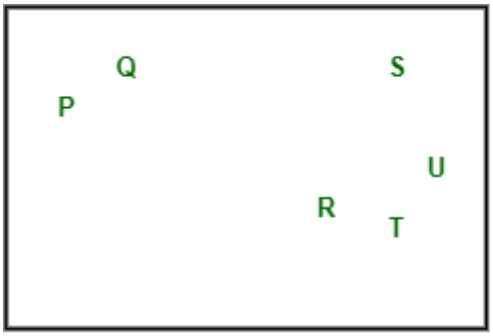
https://www.tutorialspoint.com/machine_learning/machine_learning_k_means_clustering.htm

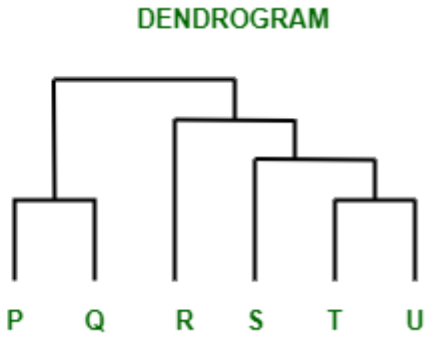
ML – Hierarchical clustering

Hierarchical clustering is used to group similar data points together based on their similarity creating a **hierarchy or tree-like structure**. The key idea is to begin with each data point as its own separate cluster and then progressively merge or split them based on their similarity.

Dendogram

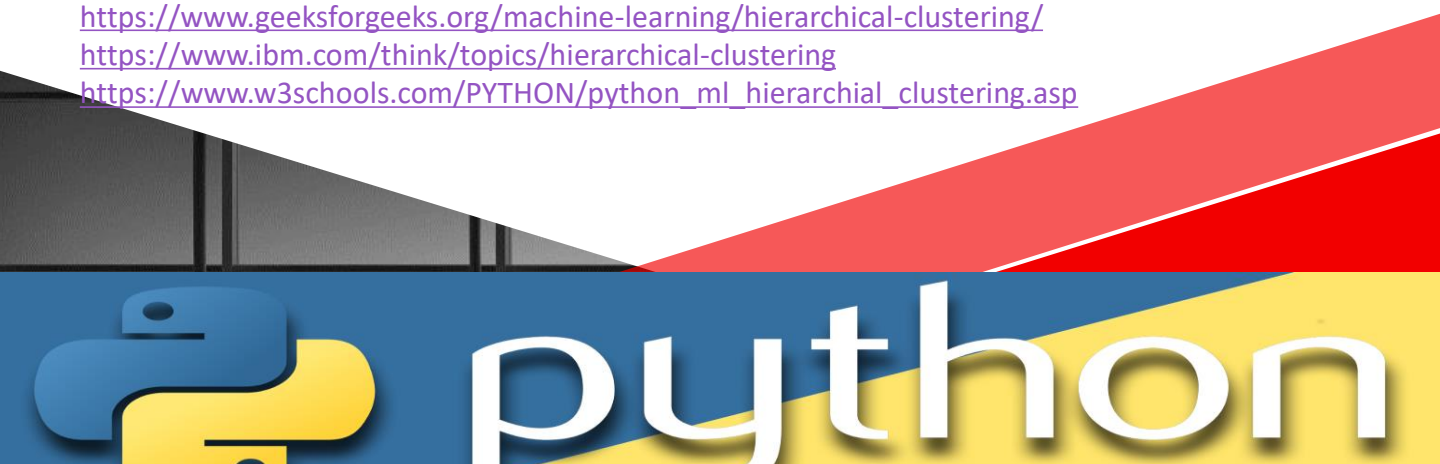
A **dendrogram** is like a family tree for clusters. It shows how individual data points or groups of data merge together. The bottom shows each data point as its own group, and as you move up, similar groups are combined. The lower the merge point, the more similar the groups are. It helps you see how things are grouped step by step. The working of the dendrogram can be explained using the below diagram:





Resources:

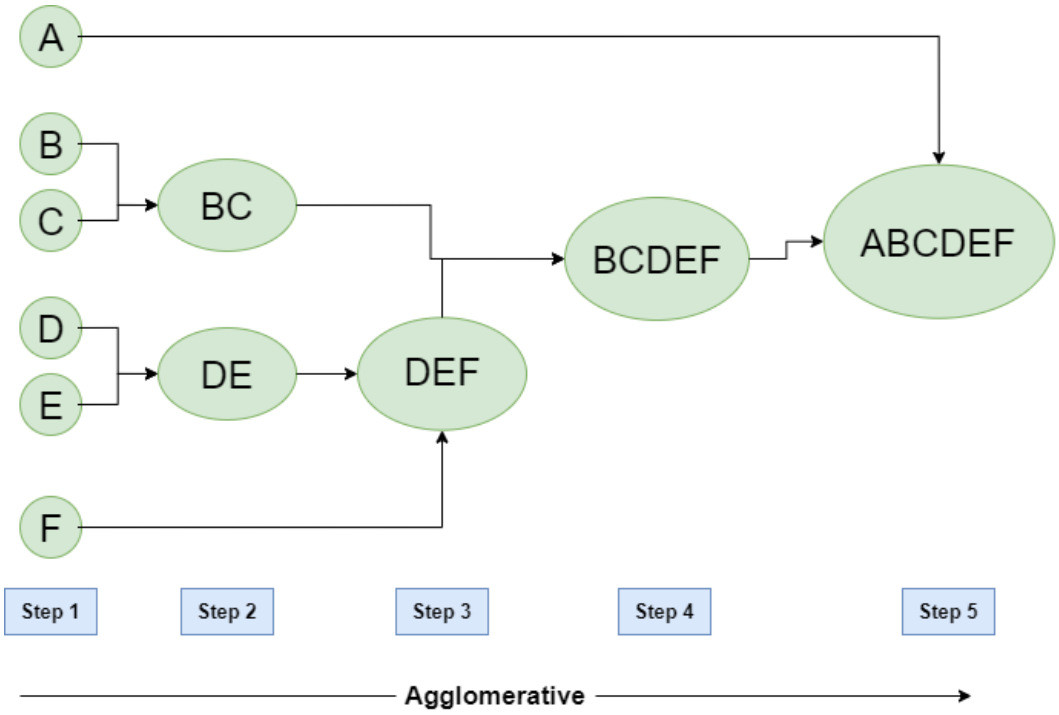
https://www.geeksforgeeks.org/machine-learning/hierarchical-clustering/
https://www.ibm.com/think/topics/hierarchical-clustering
https://www.w3schools.com/PYTHON/python_ml_hierarchial_clustering.asp



ML | Types of Hierarchical Clustering

1. Hierarchical Agglomerative Clustering

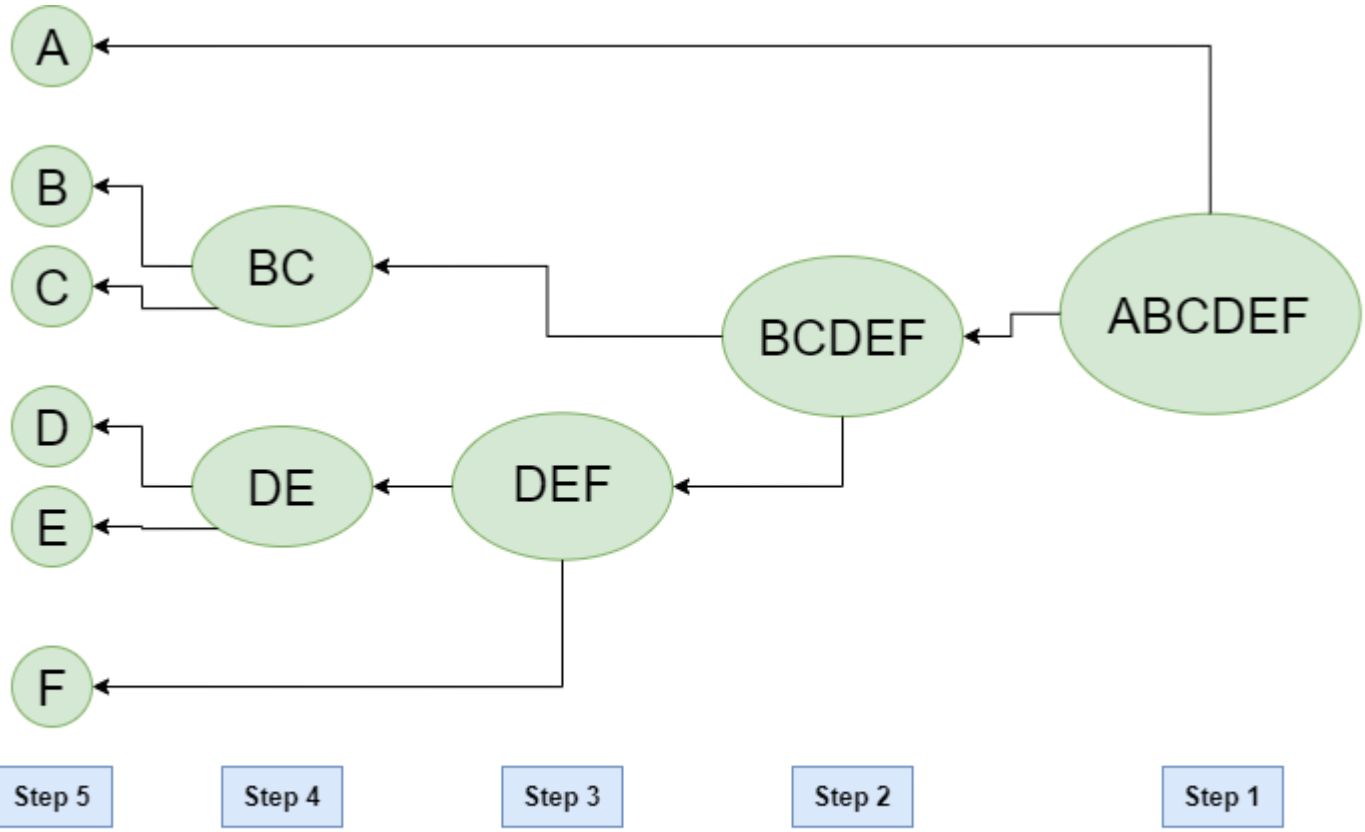
It is also known as the bottom-up approach or hierarchical agglomerative clustering (HAC). Unlike flat clustering hierarchical clustering provides a structured way to group data. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all data.



ML | Types of Hierarchical Clustering

2. Hierarchical Divisive clustering

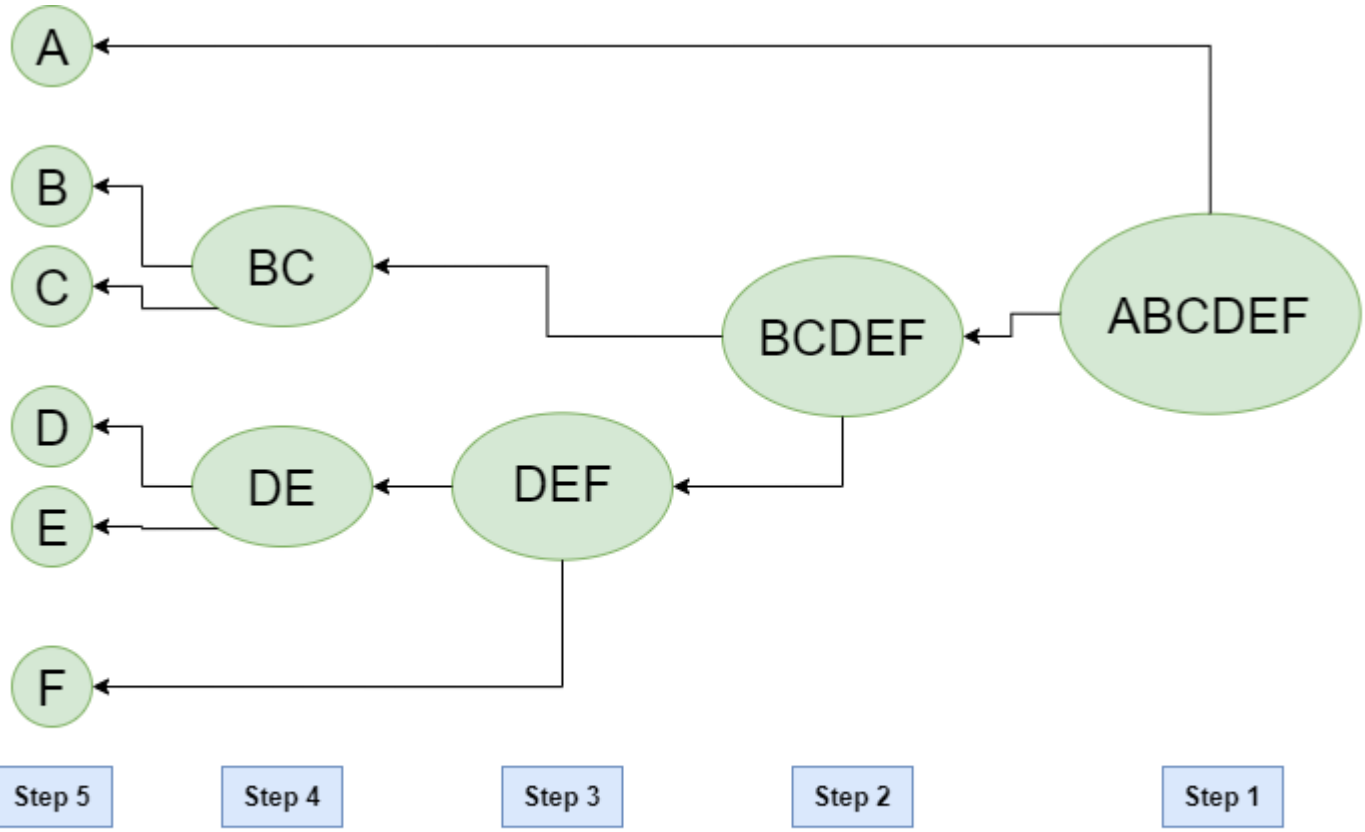
It is also known as a **top-down approach**. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.



ML | Types of Hierarchical Clustering

2. Hierarchical Divisive clustering

It is also known as a **top-down approach**. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.



ML | DBSCAN Clustering

DBSCAN is a density-based clustering algorithm that groups data points that are closely packed together and marks outliers as noise based on their density in the feature space. It identifies clusters as dense regions in the data space separated by areas of lower density. Unlike K-Means or hierarchical clustering which assumes clusters are compact and spherical, DBSCAN perform well in handling real-world data irregularities such as:

- **Arbitrary-Shaped Clusters**: Clusters can take any shape not just circular or convex.
- **Noise and Outliers**: It effectively identifies and handles noise points without assigning them to any cluster.

Reference:

https://www.geeksforgeeks.org/machine-learning/dbscan-clustering-in-ml-density-based-clustering/
https://www.datacamp.com/tutorial/dbscan-clustering-algorithm
https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

ML | When Should We Use DBSCAN Over K-Means Clustering?

DBSCAN and K-Means are both clustering algorithms that group together data that have the same characteristic. However they work on different principles and are suitable for different types of data. We prefer to use DBSCAN when the data is not spherical in shape or the number of classes is not known beforehand.

DBSCAN	K-Means
In DBSCAN we need not specify the number of clusters.	It is very sensitive to the number of clusters so it need to specified
Clusters formed in DBSCAN can be of any arbitrary shape.	Clusters formed are spherical or convex in shape
It can work well with datasets having noise and outliers	It does not work well with outliers data. Outliers can skew the clusters in K-Means to a very large extent.
In DBSCAN two parameters are required for training the Model	In K-Means only one parameter is required is for training the model

Exercises

See code here: https://github.com/ShahzadSarwar10/FullStackAI-B-4-SAT-SUN-10AM-TO-12PM/tree/main/Week11

You should be able to analyze – each code statement, you should be able to see trace information – at each step of debugging. "DEBUGGING IS BEST STRATEGY TO LEARN A LANAGUAGE." So debug code files, line by line, analyze the values of variable – changing at each code statement. BEST STRATEGY TO LEARN DEEP.

Let's put best efforts.
Thanks.
Shahzad – Your AI – ML Instructor

Thank you - for listening and participating

- ❑**Questions / Queries**
- ❑**Suggestions/Recommendation**
- ❑**Ideas…..?**

Shahzad Sarwar
Cognitive Convergence
https://cognitiveconvergence.com
shahzad@cognitiveconvergence.com
voice: +1 4242530744 (USA) +92-3004762901 (Pak)