**Name: Sahil Vishwas Shendkar**

**Roll No: TYITB118**

**Assignment No. 3**

**AIM:** Assignment on K-Means on Salary Dataset.

**OBJECTIVE: To apply K-Means Clustering on the Iris Dataset to group similar data points and uncover natural patterns or groupings without using predefined labels.**

**PREREQUISITE:** Python programming

**THEORY:**

**K-Means Clustering is a popular unsupervised machine learning algorithm used to identify and group similar data points into clusters based on their features. Unlike supervised learning, where the algorithm is trained on labeled data, clustering works without any predefined class labels. The goal is to explore the data and detect natural groupings or patterns among the instances.**

**In K-Means, the number of clusters (K) is defined beforehand. The algorithm partitions the dataset into K clusters in such a way that the distance between the data points and the center of their assigned cluster is minimized. Each cluster is represented by its centroid, which is the average of all the data points in the cluster.**

**K-Means is widely used in various real-world scenarios, including market segmentation, image compression, document classification, fraud detection, and customer behavior analysis. Its simplicity, speed, and scalability make it a reliable choice for exploratory data analysis and feature-based segmentation.**

**DATABASE:**

**For this assignment, we use the Iris Dataset, one of the most widely used datasets for understanding clustering, classification, and basic data science concepts. Originally introduced by British biologist and statistician Ronald A. Fisher, the Iris dataset includes measurements of 150 iris flowers, classified into three species:**

* **Iris Setosa**
* **Iris Versicolor**
* **Iris Virginica**

**Each record in the dataset includes four features (input variables):**

* **Sepal Length (cm)**
* **Sepal Width (cm)**
* **Petal Length (cm)**
* **Petal Width (cm)**

**Although the dataset contains species labels, these labels are not used during the clustering process, because clustering is an unsupervised method. Instead, K-Means uses the feature values to find natural groupings in the data. After clustering, we can optionally compare the predicted clusters to the known species labels to assess the quality of clustering.**

**The Iris dataset is typically stored in CSV format or accessed directly from libraries like scikit-learn in Python, and is an ideal dataset for visualizations due to its small size and structure.**

**Working of K-Means Clustering**

**K-Means follows an iterative approach to find the best grouping of the data. Here's a step-by-step explanation:**

1. **Specify the number of clusters (K) you want the algorithm to find. In the case of the Iris dataset, K is often set to 3, corresponding to the three species.**
2. **Randomly choose K initial centroids in the feature space.**
3. **Assign each data point to the nearest centroid, forming K clusters. The "closeness" is typically measured using a distance metric such as Euclidean distance.**
4. **Recalculate the centroids by finding the mean of the data points in each cluster.**
5. **Repeat the assignment and update steps until the cluster assignments no longer change or the maximum number of iterations is reached.**

**This process ensures that clusters are as compact and as distinct as possible, minimizing intra-cluster distance and maximizing inter-cluster separation.**

**Application to the Iris Dataset**

**In this assignment, we apply the K-Means algorithm to the Iris Dataset with the following goals:**

* **Group the 150 iris flower records into three clusters (since the dataset includes three known species).**
* **Identify how well the algorithm clusters the data based purely on sepal and petal dimensions.**
* **Evaluate the clustering results by comparing the predicted clusters with the known species labels.**

**This application provides valuable insight into whether natural groupings exist in the data that correspond to biological categories. Even though clustering does not use the species labels during training, a high correspondence between the clusters and the species confirms the strength of the underlying feature relationships.**

**Assessing Clustering Performance**

**Since clustering does not use labeled output, traditional accuracy scores are not directly applicable. However, we can assess the effectiveness of K-Means clustering using the following methods:**

**1. Inertia**

**Inertia measures the compactness of the clusters. It calculates the total distance of each data point from its assigned cluster centroid. Lower inertia values indicate better clustering.**

**2. Silhouette Score**

**This score indicates how similar each point is to its own cluster compared to other clusters. A higher silhouette score implies well-separated, meaningful clusters.**

**3. Visual Analysis**

**By reducing the dataset dimensions using Principal Component Analysis (PCA) or t-SNE, we can plot the clusters in 2D space and visually inspect the separation and cohesion of clusters.**

**4. Comparing with Actual Labels**

**Although unsupervised, we can still match the clusters formed with the actual species in the dataset to see how well the clustering aligns with biological classification.**

**Importance of K-Means Clustering**

**K-Means is significant for several reasons:**

* **It provides quick and interpretable grouping results for unlabelled data.**
* **It is scalable and can handle large datasets efficiently.**
* **It helps in discovering patterns, segmenting markets, and identifying outliers.**
* **It is often used as a preprocessing step in complex data pipelines for customer segmentation, feature engineering, or recommendation systems.**

**In the Iris dataset context, it helps to:**

* **Detect natural divisions among flower species based on measurements.**
* **Validate feature relevance (e.g., petal length and width are more informative than sepal width).**
* **Provide a benchmark for testing more advanced clustering or classification methods.**

**CONCLUSION:**

**K-Means Clustering is a fundamental and efficient algorithm for identifying structure in unlabeled datasets. By applying K-Means to the Iris Dataset, we can uncover meaningful groupings based solely on flower measurements. The resulting clusters offer insight into the similarities and differences among flower species, highlighting the value of unsupervised learning techniques. Even without labels, the algorithm successfully reveals natural clusters that, in many cases, align closely with biological species. This reinforces the idea that K-Means is not just a useful tool for grouping, but also for gaining deeper understanding and insight from raw data.**