Section 1 : Introduction 0:00
Hello everyone. Welcome to this tutorial on clustering. Clustering is a powerful unsupervised machine learning method used to group similar data points based on their features. Today I will demonstrate two popular clustering techniques K means clustering and DBSCAN clustering. Let's go through our tutorial. This is a clustering overview. Dataset Overview. Let's see the data set overview. We will use the Mall customer dataset to segment customers based on their spending habits and income levels. By the end of this tutorial, you will understand how the method works, how to implement there in Python, and how to compare their performance. This data set includes age, annual income, and spending score. For this clustering, we will use annual income and spending score

Section 2: Dataset Overview 0:59
I already told, you that for this clustering we will use two key features, annual income, and spending score, to ensure the clustering algorithm works effectively. We will normalize these features using standard scaling. Let's go through our code. Here we load and explore the data set. You can see in here exploring the data set.

Section 3: Features Selection 1:32
Then feature selection and scaling, Annual income, and spending scores as our features. Next, we select the features for clustering and scale them to better performance. You can see here we have used the standard scalar. So the scatter plot shows the scale data points for annual income and spending score.

Section 4: K means Clustering 1:57
So k means clustering. What is the K mean clustering, K means is a centroid basic clustering technique. It divides the data into k clusters by minimizing the variance within each cluster to determine the optimal number of clusters, we use the Elbow Method, which plots the within-cluster sum of the square. we are calling WCSS for different values of K. Let's implement it

Section 5: Elbow Method 3:10
This code calculates wcss for k values from 1 to 10 and, plots the Elbow Method. You can see here. We can see here the results of the Elbow Method. The elbow is number five. The elbow means the point of infection on the curve is the best value of k, so k is equal to five.After that we have applied the K means with the optimal k5 from the Elbow Method we already taken.

Section 6: Visualizing Clusters 3:22
Now we are going to visualize the cluster. You can see here KMeans Clustering. Five clusters areshown here. Finally, let's evaluate Kmeans by using matrices like the Silhouette Score and Davies-Bouldin Index. Let's see how it's going. First of all, let's discuss about what is the Silhouette Score. The Silhouette Score is the matrix used to evaluate the quality of a clustering algorithm. It measures how well data points are clustered by comparing how close they are to points within their own cluster versus points in the nearest neighboring cluster. These is the major points. So you can see the formula of the Silhouette Score, a, mean average distance between a point and other points in the same cluster. B, mean average distance between points and points in the nearest cluster. So this is assumed as single data point

The Silhouette Score range, as will define perfect clustering. Points are very close to their own cluster and far from others, and it's if it's zero. This means overlapping clusters. It means points are equally close to their own and the nearest cluster. And if the value is a negative value, it means poor clustering. It means points are closer to another cluster than their own. As an interpretation higher Silhouette Score (it means close to one) indicates well-separated clusters. And low Silhouette Score (close to zero or negative) indicates overlapping of poorly formed clusters. Now we are going to discuss about Davies-Bouldin Index

Section 7: **Davies-Bouldin Index score**.
Davis bowling score evaluates the cluster quality based on the ratio of intra-cluster dispersion to inter-cluster Separation. So lower scores indicate better clustering in the Davies-Bouldin index. So this is the formula of the Davies-Bouldin index. You can see it here.
The range of Davis bounding index: The lower is better. It means close, closer to zero. It's indicate it's better. And lower value indicates compact clusters that are far apart from one another. A high value indicates a poorly separated or overlapping cluster. As an interpretation of the Davies-Bouldin index, Davies-Bouldin index, good clustering in compact, well-separated clusters, and we can take a High Davies-Bouldin Index score: poor clustering with overlapping or disparate clusters.

Section 8: DBSCAN Clustering 7.36

Let's discuss about DBSCAN cluster. What is DBSCAN?
DBSCAN is a density-based clustering technique. K means a center-based clustering technique. Unlike K means DBSCAN doesn't require us to predefine the number of clusters. Instead of, it groups points that are densely packed together and mark sparse points as noise.
DBSCAN clustering, the parameters are eps,
eps means the maximum distance between points. The min_samples, It's means the minimum number of points to form a cluster. Let's implement DBSCAN with examples of parameters. Here we have used eps =0.5 min_samples =5. We apply DBSCAN and visualize the clusters. The noise points are labeled as minus one.

Section 9: Comparison
This is a scatter plot of the DBSCAN clustering. You can see here the final results of the comparison table. According to that results KMeans has higher Silhouette score than DBSCAN. Suggested that K means produce clusters where data points are more similar to their own cluster and less similar to other clusters and the Davies-Bouldin index, Kmeans has lower Davies-Bouldin index than DBSCAN.
Implying that K means forms clusters that are better separated and less similar to each other. Overall , based on these two matrix K means appear to have produce better clustering results for this specific data set.

We can see the comparison of clustering techniques. So as features, I have taken some features to compare both clustering techniques. Handles outlier is poor in K means cluster, and it's excellent in the DBSCAN cluster, Predefined K, yes, we need Predefined K in KMeans clustering, but we don't need to take number of k in DBSCAN .I have already shown in the code,Clusters Shape, spherical works with the simple cluster in the KMean so arbitrary works

with the complex shapes in the DBSCAN. Parameter sensitivity , moderate in the KMean because at that one, need to define K, In DBSCAN it's high, because EPs and the mins-samples must be tuned.

The conclusion, K means works effectively with the well separate spherical clusters, but is sensitive to outliers and requires to case to be predefined. So DBSCAN handle noise well, and it capable to detecting clusters of arbitrary shape, but it's sensitive to parameter selection. So this is the summary of both clusters.

Section 10 –End 11:08
I hope you will understand about the both clustering techniques, K Means and DBSCAN.
Thank you for watching.

Github repository include ppt ,code and dataset.
GITHUB - https://github.com/SanduniMarasingha/Individual-assignment-Machine-learning-tutorial/tree/main