It256_lab8 k means clustering

Screenshots and short notes
By: vivek vittal biragoni 211AI041

K-means clustering is an unsupervised machine learning algorithm used to partition data into groups or clusters based on their similarity. Here are some key points about k-means clustering:

1. Objective: The goal of k-means clustering is to minimize the within-cluster sum of squares, which measures the distance between each data point and the centroid of its assigned cluster.

2. Number of Clusters: The algorithm requires specifying the number of clusters (k) in advance. It assumes that the data can be divided into k distinct groups.

3. Algorithm Steps:
   - Initialization: Randomly select k data points as initial cluster centroids.
   - Assignment: Assign each data point to the nearest centroid, forming k clusters.
   - Update: Calculate the new centroids as the mean of the data points within each cluster.
   - Repeat the assignment and update steps until convergence or a maximum number of iterations.

4. Euclidean Distance: K-means clustering typically uses the Euclidean distance metric to measure the similarity between data points.

5. Centroids: The centroids represent the center of each cluster and are updated iteratively during the algorithm's execution.

6. Convergence: The algorithm converges when the assignments of data points to clusters no longer change significantly or when the maximum number of iterations is reached.

7. Cluster Assignment: Each data point is assigned to the cluster with the nearest centroid based on the Euclidean distance.

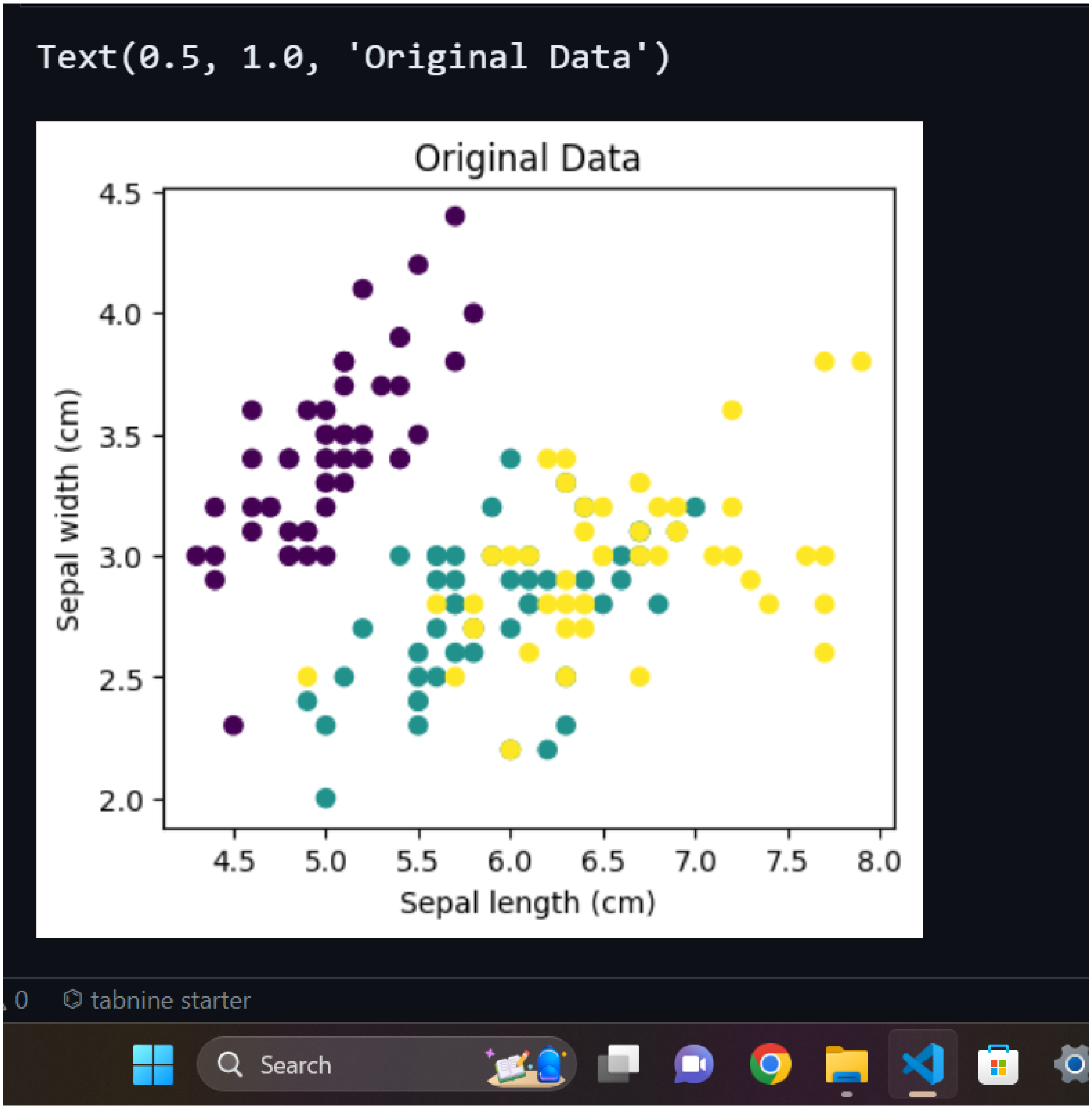
8. Accuracy Evaluation: In the absence of ground truth labels, evaluating the accuracy of k-means clustering can be challenging. One common approach is to compare the cluster

assignments with the known labels if available. However, this may not always be possible in unsupervised settings.

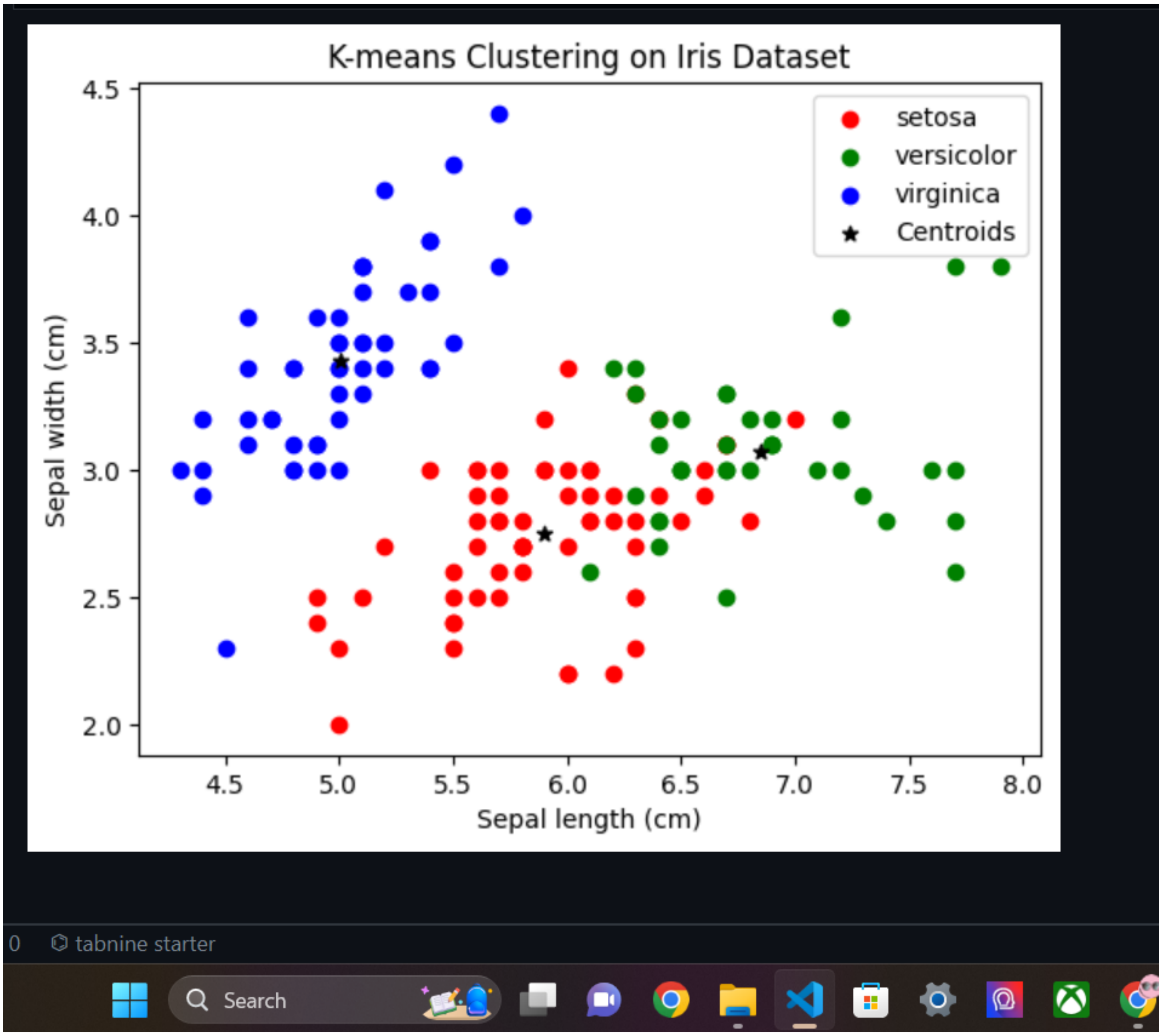
9. Visualization: Visualizing the results of k-means clustering can provide insights into the structure and grouping of the data. Plotting the data points colored by their assigned clusters and the cluster centroids helps understand the separation and compactness of the clusters.

10. Limitations:
- Sensitivity to Initial Centroids: The algorithm's performance can be influenced by the initial choice of centroids, as it may converge to different local optima.
- Determining the Number of Clusters: Selecting the appropriate number of clusters (k) can be subjective and may require domain knowledge or trial and error.
- Sensitive to Outliers: Outliers can significantly impact the centroid calculations and cluster assignments.

Overall, k-means clustering is a widely used algorithm for unsupervised clustering tasks. It provides a simple and interpretable approach to group similar data points together based on their features.



```
Text(0.5, 1.0, 'Original Data')
```



K-means Clustering on Iris Dataset
Sepal width (cm)
Sepal length (cm)
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
● setosa
● versicolor
● virginica
★ Centroids
0
tabnine starter
Search