**Department of Computer Science & Engineering**

**Final Year B. Tech. (CSE) – I: 2022-23**

**5CS462: PE5 - Data Mining Lab**

**Assignment No. 6**

**Name:** Shreyash Malu

**PRN:** 2019BTECS00057

**Batch:** B7

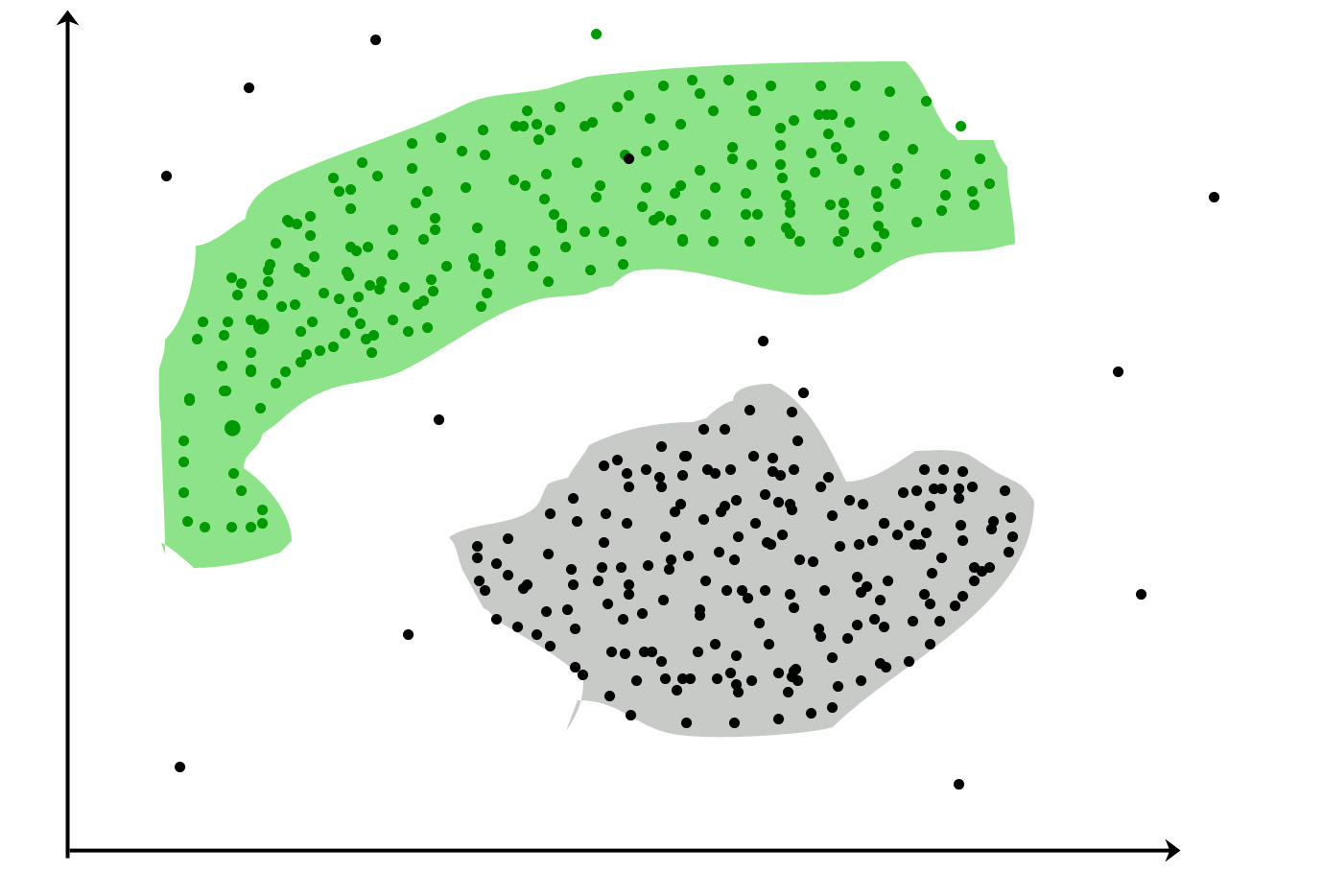
**Aim**:

To study and implement clustering algorithms

**Theory** :

**Clustering:**

**Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. It is not necessary for clusters to be spherical. Such as



**Clustering Methods :**

* **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*, *OPTICS (Ordering Points to Identify Clustering Structure)*, etc.
* **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category

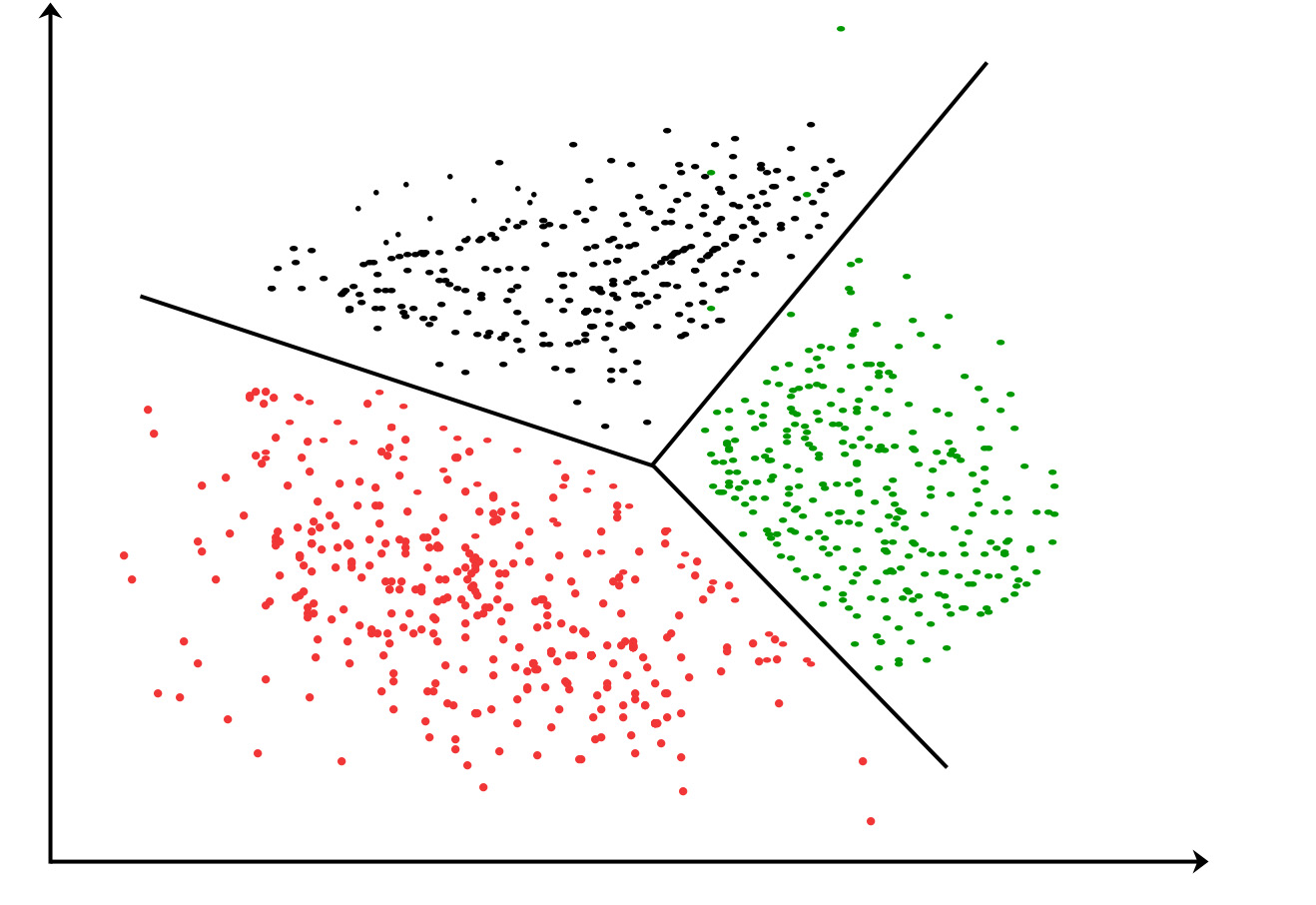
**Agglomerative** (bottom-up*approach*)

**Divisive** (top-down*approach*)

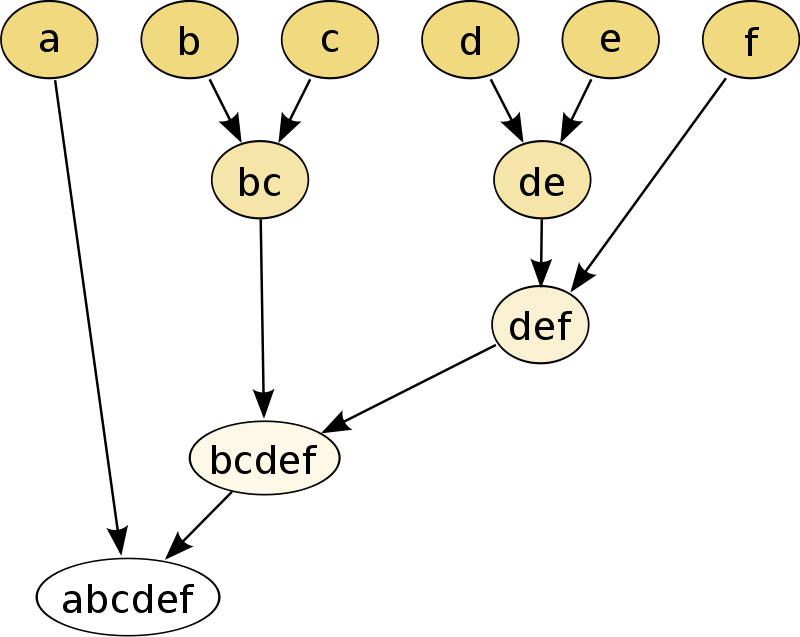
* **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)*, etc.

**Clustering Algorithms :**

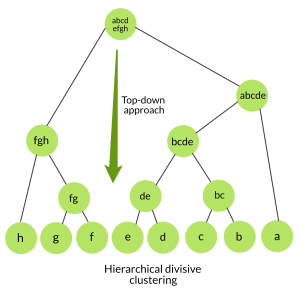
[K-means clustering algorithm](https://www.geeksforgeeks.org/k-means-clustering-introduction/) – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



**Agglomerative**: This is a "[bottom-up](https://en.wikipedia.org/wiki/Top-down_and_bottom-up_design)" approach: Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.



**Divisive**: This is a "[top-down](https://en.wikipedia.org/wiki/Top-down_and_bottom-up_design)" approach: All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

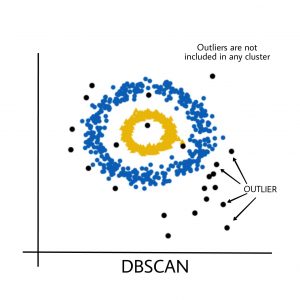


**DBSCAN**:

**DBSCAN algorithm requires two parameters:**

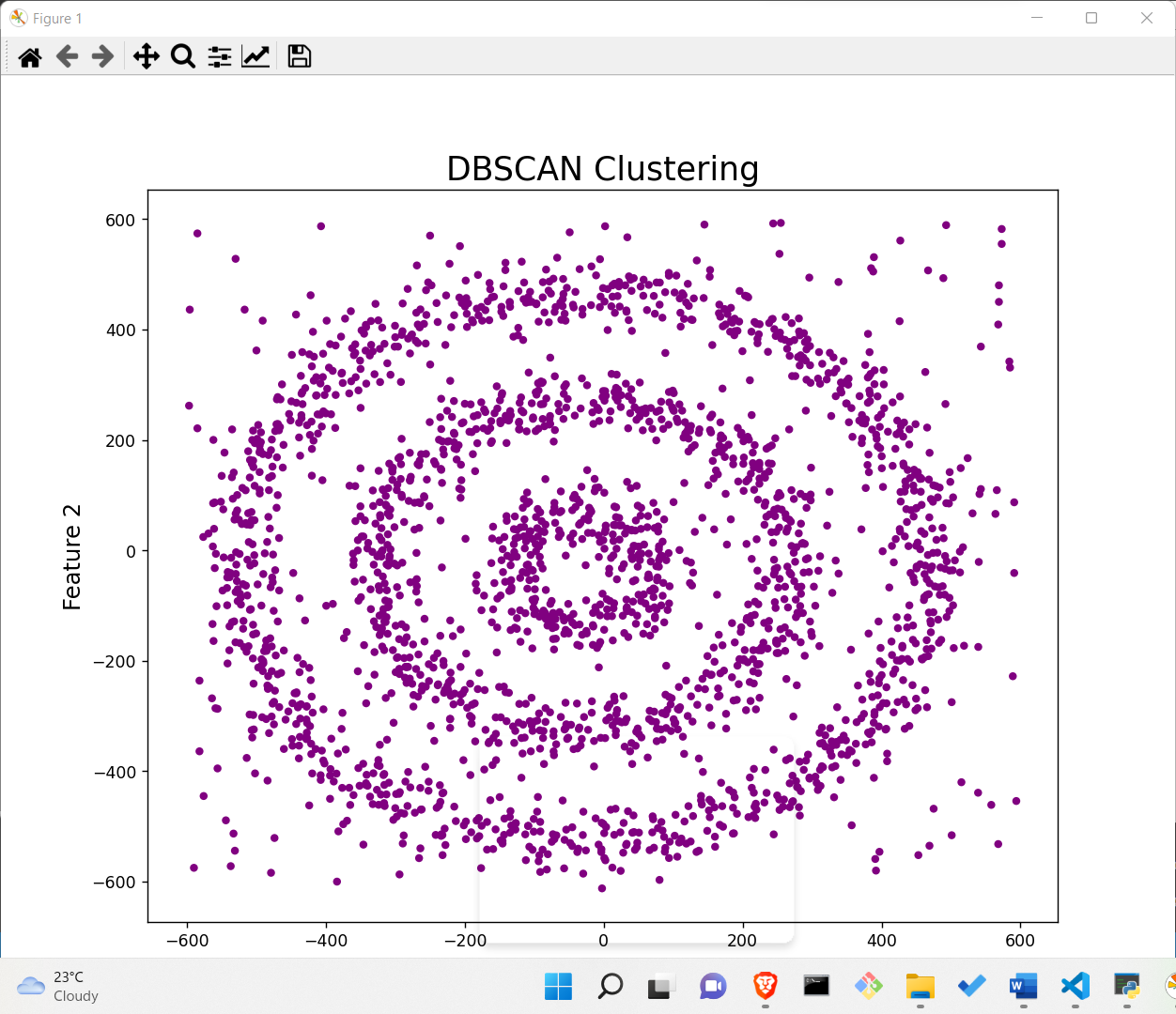
1. **eps** : It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the ***k-distance graph***.
2. **MinPts**: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3.

***In this algorithm, we have 3 types of data points.******Core Point****: A point is a core point if it has more than MinPts points within eps.****Border Point****: A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.****Noise or outlier****: A point which is not a core point or border point.*

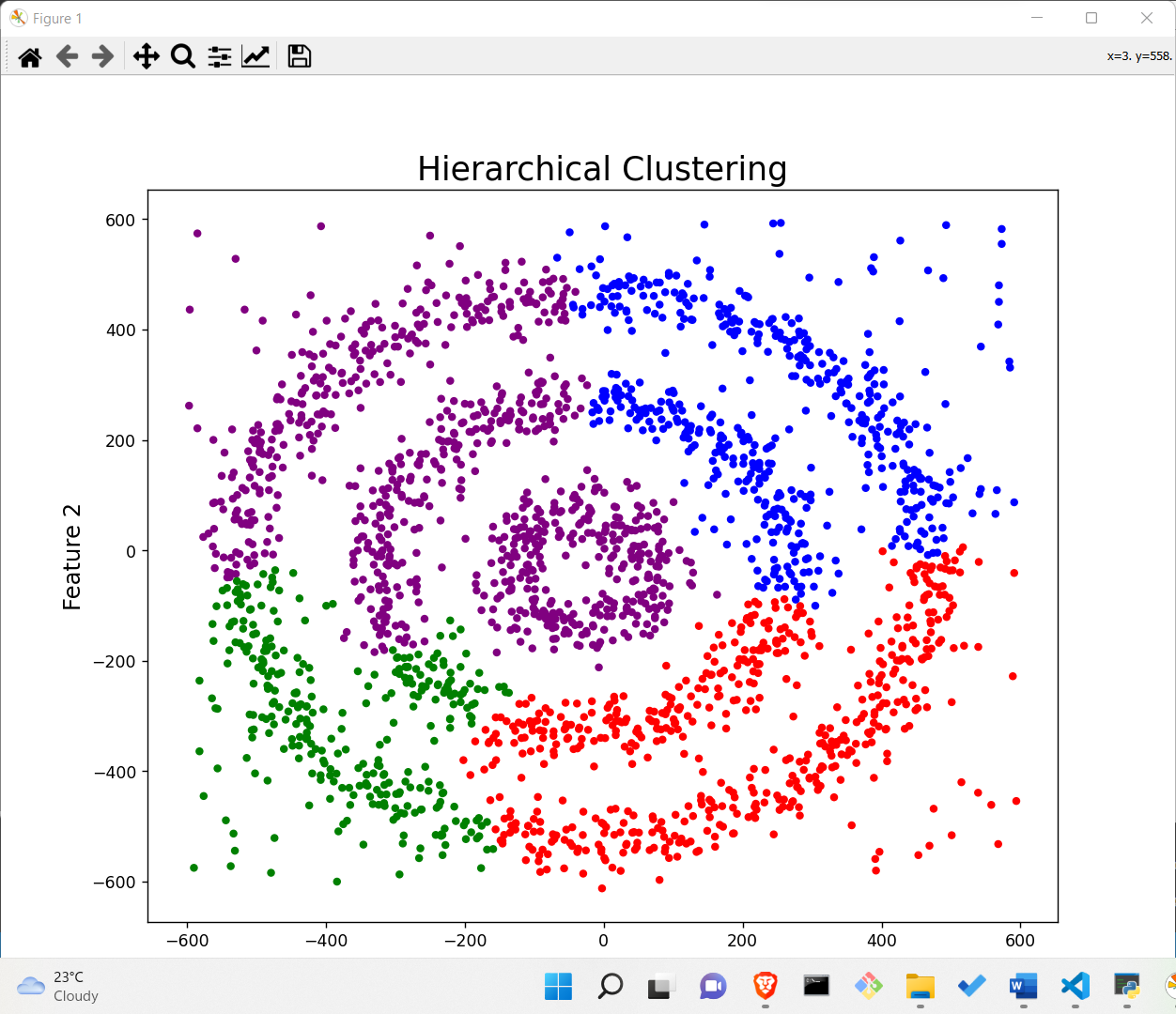


**Output: Enclosed in uploaded pdfs**

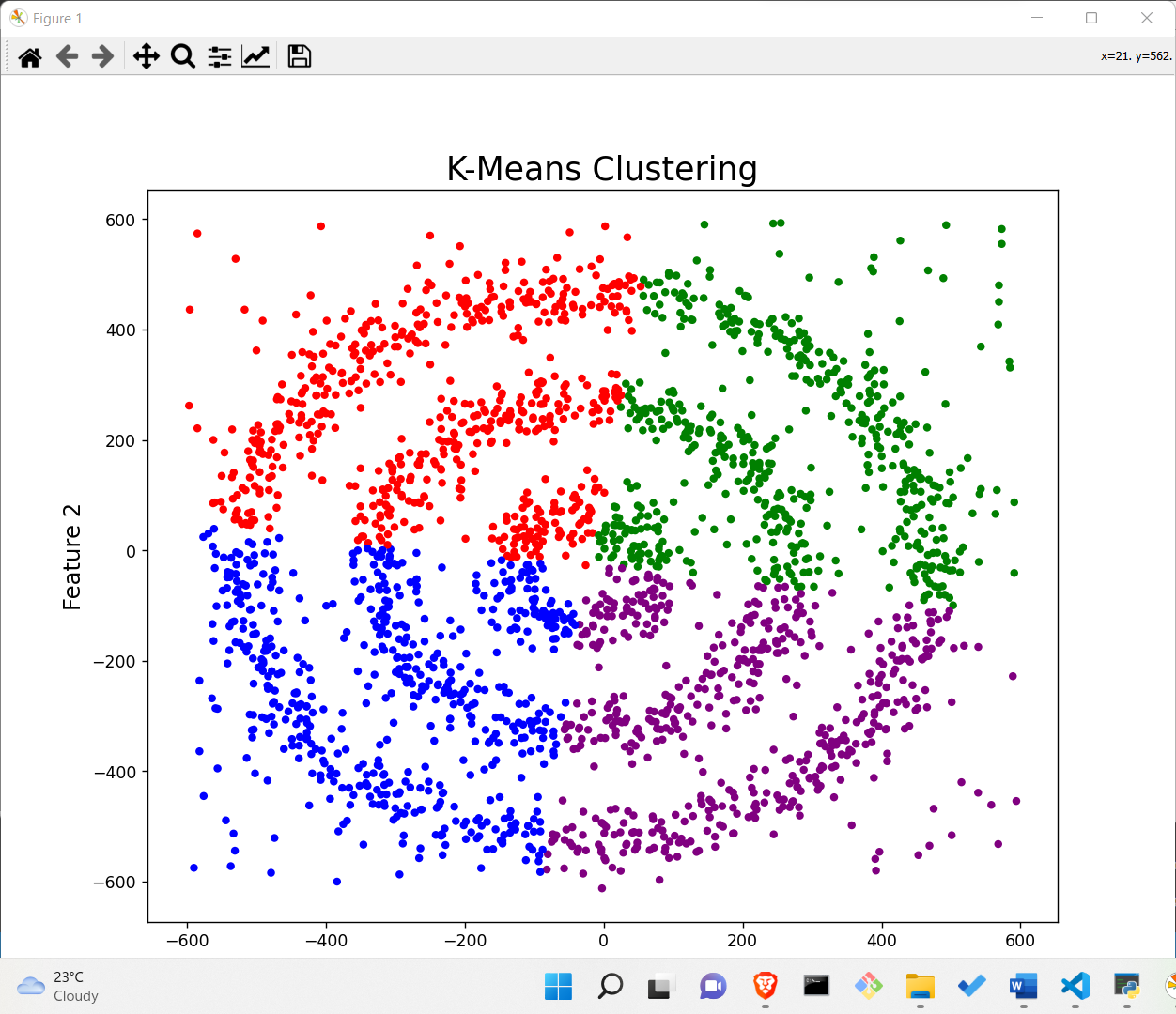
**DBScan**

****

**Hierarchical Clustering**

****

**K-Means**

****

**Conclusion:**

**Implemented different clustering algorithms**

**References:**

[**https://www.geeksforgeeks.org/clustering-in-machine-learning/**](https://www.geeksforgeeks.org/clustering-in-machine-learning/)

[**https://www.geeksforgeeks.org/different-types-clustering-algorithm/**](https://www.geeksforgeeks.org/different-types-clustering-algorithm/)

[**https://www.geeksforgeeks.org/analysis-of-test-data-using-k-means-clustering-in-python/**](https://www.geeksforgeeks.org/analysis-of-test-data-using-k-means-clustering-in-python/)