

# **MINOR PROJECT REPORT**

**ON**

## **Clustering Based Application**

**Submitted for the partial fulfillment of the requirement for the award of a Degree**

**B.TECH IN**

**COMPUTER SCIENCE & ENGINEERING**

**UNIVERSITY INSTITUTE OF**

**TECHNOLOGY, RGPV**



**Submitted by:**

**Khushi Bhawsar(0101CS201060)**

**Ajay Gurjar(0101CS201008)**

**Anushka Bhasme(0101CS201022)**

**Guided By:**

**Dr. Anjana Deen**

**Professor, DoCSE**

**Prof. Praveen K Kaithal**

**Assistant Professor, DoCSE**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**  
**UNIVERSITY INSTITUTE OF TECHNOLOGY, RGPV**  
**BHOPAL-462036**



**DECLARATION BY THE CANDIDATES**

We hereby declare that the work which is being presented in the Report of Minor Project entitled — **“Clustering Based Application”** is our own work, submitted for the partial fulfillment of the requirement for the award of a bachelor's degree in Computer Science & Engineering. The work which has been carried out at the University Institute of Technology RGPV, Bhopal in the session 2021-2022, and an authentication record of our work carried out under the guidance of **Dr. Anjana Deen (Professor) & Prof. Praveen K Kaithal (Assistant Professor)** DoCSE, University Institute of Technology, RGPV Bhopal.

I further declare that, to the best knowledge, the matter written in this project has not been submitted for the award of any other Degree.

Name of the students:

Date:

**Khushi Bhawsar(0101CS201060)**

Place: Bhopal

**Ajay Gurjar (0101CS201008)**

**Anushka Bhasme (0101CS201022)**

# UNIVERSITY INSTITUTE OF TECHNOLOGY

## RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL



### CERTIFICATE

This is to certify that **Ajay Gurjar, Khushi Bhawsar and Anushka Bhasme** of B.Tech 3rd Year, CSE UIT(RGPV) have completed their **Minor Project** entitled “**Clustering Based Application**” during the academic year **2022- 2023** under my guidance and supervision.

We approve the project for the submission for the partial fulfillment of the requirement for the award of a degree in Computer Science & Engineering.

**Dr. Anjana Deen**

Professor, DoCSE

(Project Guide)

**Head of Department,**

Dr. Uday Chaurasiya

DoCSE, UIT-RGPV

**Prof. Praveen K Kaithal**

Assistant Professor, DoCSE

(Project Guide)

**Prof. Manish Ahirwar**

Associate Professor, DoCSE

(Minor Project Coordinator)

## ACKNOWLEDGEMENT

We would like to offer our heartfelt appreciation to our Project Guides , **Dr. Anjana Deen** and **Prof. Praveen K Kaithal**, for their invaluable advice and assistance. This project would not have been possible without their enthusiasm, hard work, and excellent counsel. Their thorough approach has increased the project's precision and clarity.

We are thankful to **Dr. Uday Chourasia**, Head of the Department of Computer Science and Engineering, University Institute of Technology RGPV, Bhopal, for his unwavering support and inspiration for the project, ethics, and morals. The concepts we've acquired from him have always been a source of exponential inspiration for our path in this life.

We are also thankful to all other members and Staff of the Department who were involved in the project directly or indirectly for their valuable co-application.

We are grateful to all of our co-workers for inspiring us and creating a nice atmosphere to learn and grow.

**Name of the students:**

**Date:**

**Khushi Bhawsar(0101CS201060)**

**Place:Bhopal**

**Ajay Gurjar (0101CS201008)**

**Anushka Bhasme (0101CS201022)**

<b>Title</b>	<b>Page No.</b>
<b>DECLARATION .....</b>	<b>2</b>
<b>CERTIFICATE .....</b>	<b>3</b>
<b>ACKNOWLEDGMENT .....</b>	<b>4</b>
<b>LIST OF FIGURES.....</b>	
<b>ABSTRACT .....</b>	<b>9</b>
 <b>CHAPTER 1 .....</b>	 <b>10 - 16</b>
<b>INTRODUCTION .....</b>	<b>10</b>
1.1 Overview .....	11
1.2 Categories Of Clustering .....	11
1.2.1 Partitioning Based Clustering .....	11
1.2.2 Density Based Clustering .....	11
1.2.3 Distribution Model Based Clustering .....	11
1.3 Objective .....	12
1.4 Motivation Of Work .....	13
1.5 Organization Of Reports .....	16
 <b>CHAPTER 2 .....</b>	 <b>17</b>
<b>LITERATURE SURVEY .....</b>	<b>17</b>
2.1 Survey Report .....	17
2.2 Existing Work .....	26
2.2.1 Amazon Recommendation System .....	26
2.2.2 Cancer Cell Detection and Classification .....	26
2.2.3 Clustering Analysis .....	26
2.3 Critical Paper Reviews .....	27
2.3.1 Paper 1 .....	27
2.3.2 Paper 2 .....	27
2.3.3 Paper 3 .....	27

<b>CHAPTER 3 .....</b>	<b>29</b>
<b>PROBLEM DESCRIPTION .....</b>	<b>29</b>
3.1 Customer Segmentation. ....	29
3.2 Iris Data Set .....	30
 <b>CHAPTER 4 .....</b>	 <b>33</b>
<b>PROPOSED WORK .....</b>	<b>33</b>
4.1 Customer Segmentation .....	33
4.2 Iris Data Set.....	35
 <b>CHAPTER 5 .....</b>	 <b>37</b>
<b>IMPLEMENTATION &amp; RESULT .....</b>	<b>37</b>
5.1 Customer Segmentation .....	37
5.1.1 Customer Segmentation using K means In Clustering.....	37
5.1.2 Choosing the Annual Income Column and Spending Score Column....	37
5.1.3 Choosing the No. of Clusters.....	38
5.1.4 Training the K means Clustering Model.....	39
5.1.5 Visualizing all The Clusters.....	40
5.2 Iris Data Set .....	41
5.2.1 History.....	41
5.2.2 Introduction to Data Set.....	42
5.2.3 Applying K means Clustering.....	43
5.2.4 Original Data vs Our Clustered Data.....	49

<b>CHAPTER 6 .....</b>	<b>51</b>
<b>TOOLS AND TECHNOLOGIES .....</b>	<b>51</b>
6.1 Python.....	51
6.2 Jupyter Notebook.....	51
6.3 NumPy.....	51
6.4 Pandas.....	51
6.5 SKiKit Learn.....	52
6.6 MatPlotLib.....	52
6.7 Seaborn.....	52
.	
<b>CHAPTER 7 .....</b>	<b>53</b>
<b>CONCLUSION AND FUTURE WORK .....</b>	<b>53</b>
7.1 Conclusion .....	53
7.2 Future Work .....	53
7.2.1 Feature Engineering .....	54
7.2.2 Evaluation Algorithm .....	54
7.2.3 Alternative Clustering Algorithm .....	54
7.2.4 Visualization technique .....	54
7.2.5 Dynamic Segmentation .....	54
<b>References .....</b>	<b>55</b>

## LIST OF FIGURES

FIGURE	PAGE NO.
Fig.No.1: Flow of Data in Clustering.....	18
Fig.No.2: Partitioning Clustering.....	19
Fig.No.3: Density Based Clustering.....	18
Fig.No.4: Distribution Based Clustering.....	19
Fig.No.5: Hierarchical Based Clustering .....	18
Fig.No.6: Fuzzy Clustering.....	19
Fig.No.7: Work Flow Chart 1.....	18
Fig.No.8: Work Flow Chart 2.....	



## ABSTRACT

**Clustering** is a fundamental task in machine learning that aims to discover hidden patterns and structures within **unlabeled data**. It plays a crucial role in various domains, including **data mining, pattern recognition, image analysis, and customer segmentation**. The objective of clustering algorithms is to group similar data points together while maximizing the dissimilarity between different groups.

In recent years, numerous clustering algorithms have been developed, each with its own strengths and weaknesses. Traditional approaches such as **k-means, hierarchical clustering, and density-based clustering methods like DBSCAN** have been widely used. These algorithms rely on different similarity metrics, distance measures, and optimization criteria to partition the data into meaningful clusters.

Evaluating the quality of clustering results is another essential aspect in clustering research. **External and internal evaluation measures are commonly used to assess the accuracy, compactness, and separation of clusters**. These measures help researchers and practitioners select appropriate clustering algorithms and tune their parameters for optimal results.

# CHAPTER 1

## Introduction

### 1.1 Overview:

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering. Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No supervision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so-called training set, i.e., a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

Clustering is a technique used in machine learning with various practical applications across different domains. Some of the key clustering-based applications include customer segmentation, anomaly detection, image and document organization, recommender systems, gene expression analysis, social network analysis, and spatial data analysis. Clustering algorithms help in grouping similar data points together based on their intrinsic properties or similarity measures.

## **1.2 Categories of Clustering:**

### **1.2.1 Partitioning Clustering:**

**Partitioning Clustering** is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.

In this type, the dataset is divided into a set of  $k$  groups, where  $K$  is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.

### **1.2.2 Density-based clustering:**

**Density-based clustering** method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

### **1.2.3 Density-based clustering:**

**Distribution model-based clustering** method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

## 1.3 Objective

The objective of clustering-based applications is to discover patterns, structures, or natural groupings in data that can provide insights and facilitate decision-making processes. The specific objectives may vary depending on the application, but here are some common objectives:

### **Data Exploration:**

Clustering allows for the exploration and understanding of complex datasets. By grouping similar data points together, it helps identify inherent structures, relationships, and patterns within the data.

### **Pattern Recognition:**

Clustering can be used to recognize patterns or similarities among data points. It helps in identifying groups or clusters that share common characteristics, behaviors, or attributes, enabling deeper analysis and understanding of the data.

### **Data Segmentation:**

Clustering aids in segmenting data into meaningful subsets or clusters. This segmentation provides a way to organize and categorize data based on similarities, which can be useful for further analysis or targeted actions.

## **1.4 Motivation for work:**

The motivation behind developing a clustering-based application can vary depending on the specific problem or domain. Here are some common motivations that drive the development of clustering-based applications:

### **Data Organization and Exploration:**

Clustering helps to organize large and complex datasets into meaningful groups, allowing users to explore and understand the data more effectively. By visually representing clusters, patterns, and relationships within the data, users can gain insights and make data-driven decisions.

### **Pattern Discovery:**

Clustering facilitates the discovery of hidden patterns or structures in the data that may not be immediately apparent. Uncovering such patterns can lead to new knowledge, insights, and understanding of the underlying processes or phenomena..

### **Anomaly Detection:**

Clustering can help detect anomalies or outliers in the data. By identifying data points that deviate significantly from their clusters, clustering-based applications can assist in anomaly detection, fraud detection, and error identification.

## **Customer Segmentation and Personalization:**

Clustering allows for customer segmentation based on their behaviors, preferences, or demographic attributes. This segmentation enables businesses to tailor their marketing strategies, personalize recommendations, and deliver customized experiences to different customer groups.

## **Efficiency and Resource Optimization:**

Clustering-based applications can aid in optimizing resource allocation and improving efficiency. By grouping similar data points together, businesses can allocate resources more effectively, identify areas of improvement, and streamline processes.

## **Improved Data Representation and Visualization:**

Clustering helps in representing complex data in a more interpretable and visual manner. By visualizing clusters and their relationships, users can grasp the structure and characteristics of the data more intuitively, leading to better understanding and communication.

## **Research and Exploration:**

Developing clustering-based applications can be motivated by the desire to advance research in the field of clustering algorithms, data analysis, and machine learning. These applications can serve as platforms for experimenting with new techniques, evaluating algorithm performance, and exploring innovative clustering approaches.

Overall, the motivation behind developing a clustering-based application lies in the need for data organization, pattern discovery, decision support, anomaly detection, personalization, efficiency improvement, and advancement in research.

## CHAPTER 2

### Literature Survey & Related Work

#### 2.1 Literature Survey

A Clustering or cluster analysis is a machine learning technique, which groups the unlabeled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.[1]

It does it by finding some similar patterns in the unlabeled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabelled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML systems can use this id to simplify the processing of large and complex datasets. The clustering technique is commonly used for statistical data analysis.

**Note:** Clustering is somewhere similar to the classification algorithm, but the difference is the type of dataset that we are using. In classification, we work with the labeled data set, whereas in clustering, we work with the unlabeled dataset.

**Example:** Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. T-shirts are grouped in one section, and trousers are in other sections. Similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily

find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.

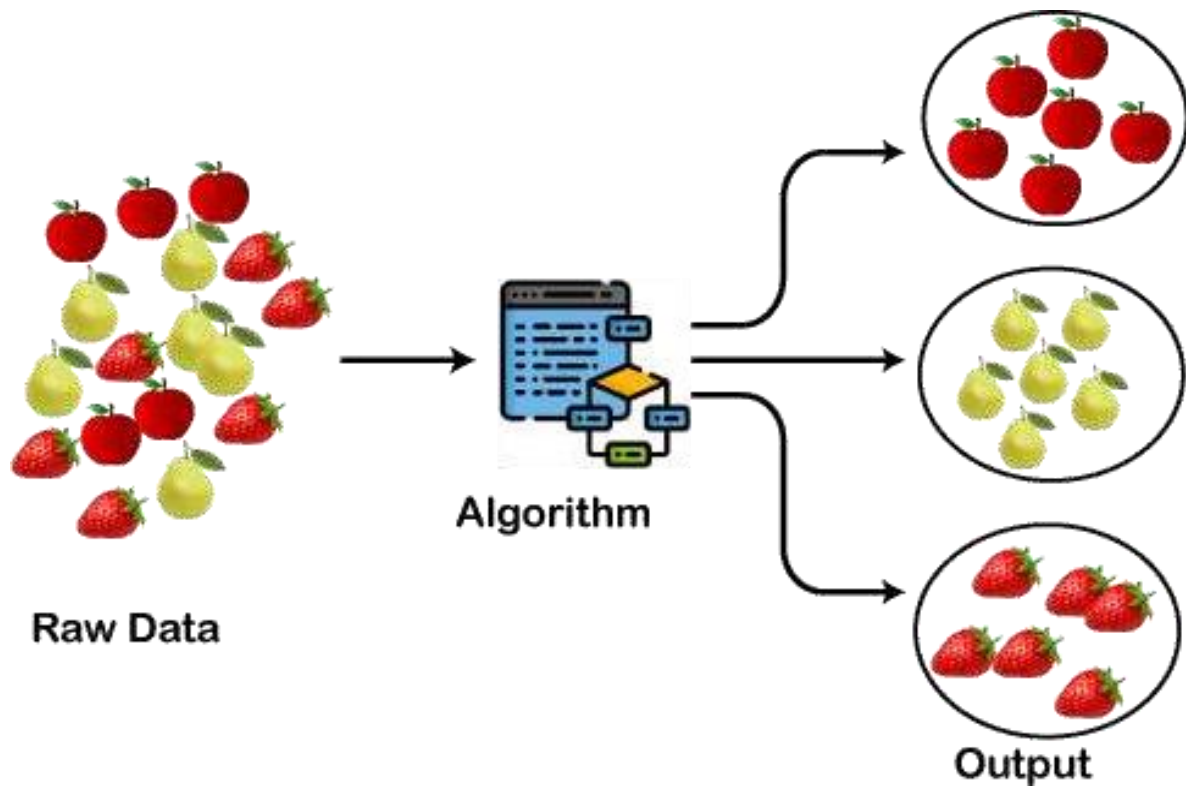
The clustering technique can be widely used in various tasks. Some most common uses of this technique are

- **Market Segmentation**
- **Statistical data analysis**
- **Social network analysis**
- **Image Segmentation**
- **Anomaly detection, etc**

Apart from these general usages, it is used by Amazon in its recommendation system to provide the recommendations as per the past search of products. Netflix also uses this technique to recommend the movies and web-series to its users as per the watch history.



The below diagram explains the working of the clustering algorithm. We can see the Different fruits are divided into several groups with similar properties.



**Figure 1 Flow of Data in clustering**

## **Types Of Clustering[2]**

### Types of Clustering Methods

The clustering methods are broadly divided into Hard clustering (datapoints belongs to only one group) and Soft Clustering (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

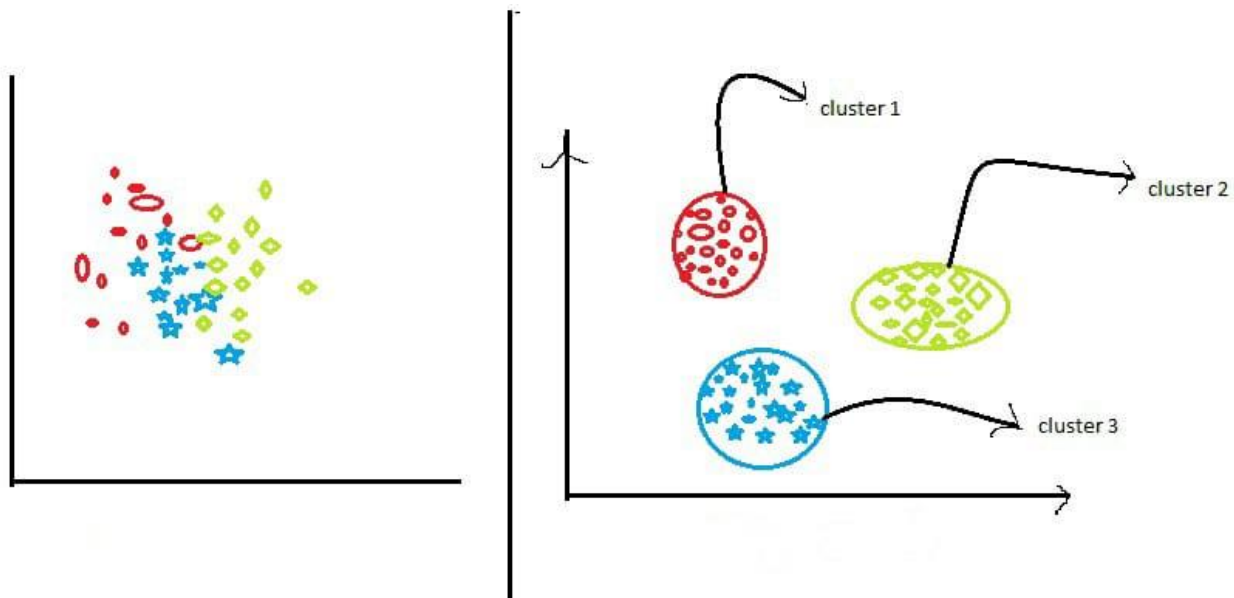
#### 1. Partitioning Clustering

2. Density-Based Clustering
3. Distribution Model-Based Clustering
4. Hierarchical Clustering
5. Fuzzy Clustering

### **Partitioning Clustering:**

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.

In this type, the dataset is divided into a set of  $k$  groups, where  $K$  is used to define the number of pre-defined groups. The cluster centre is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



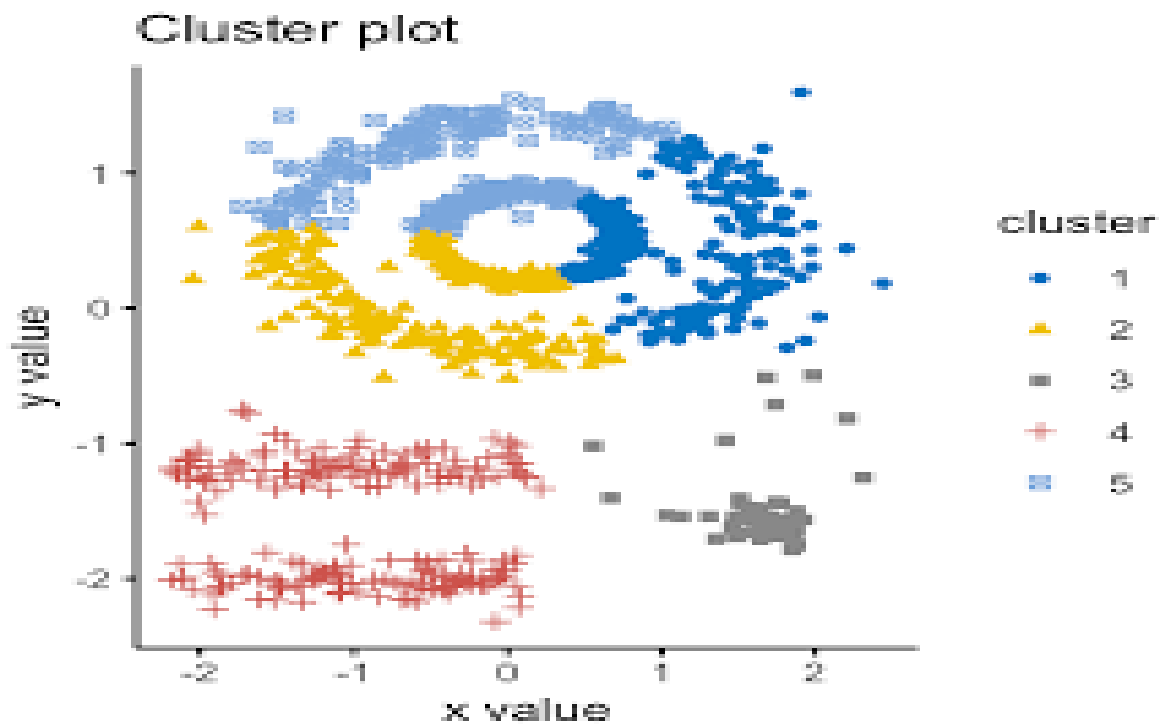
**Figure 2 Partitioning Clustering**

### **Density-Based Clustering:**

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This

algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

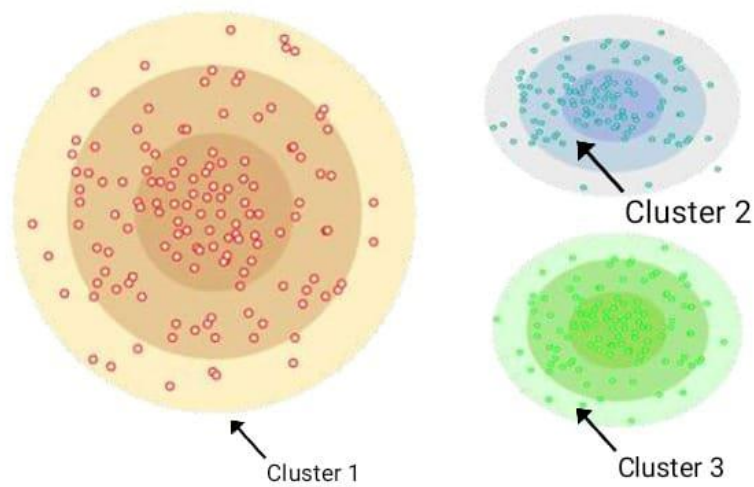


**Figure 3 Density Based Clustering**

### **Distribution Model-Based Clustering:**

In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly Gaussian Distribution.

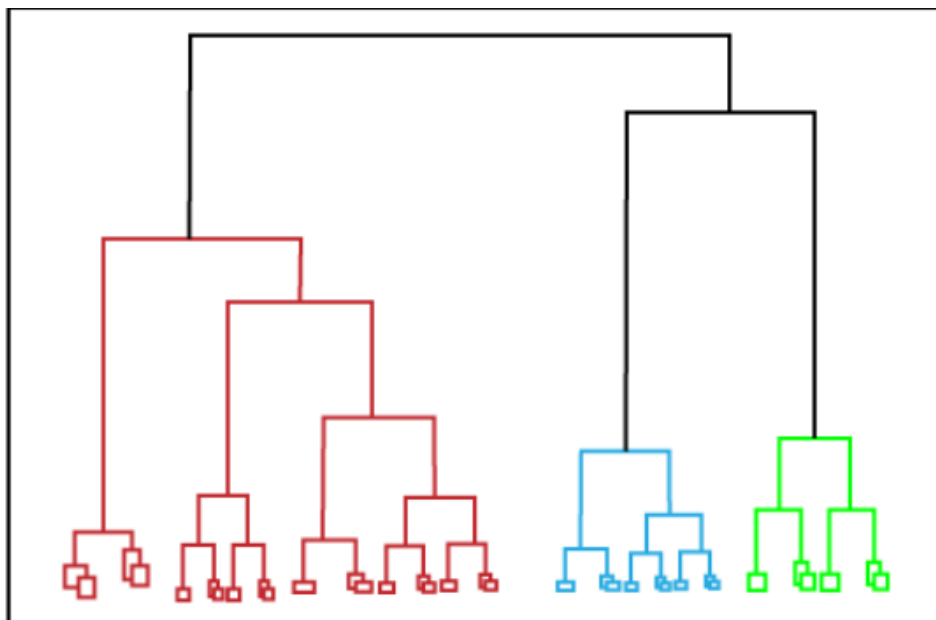
The example of this type is the Expectation-Maximization Clustering algorithm that uses Gaussian Mixture Models (GMM).



**Figure 4 Distribution Based Clustering**

### **Hierarchical Clustering:**

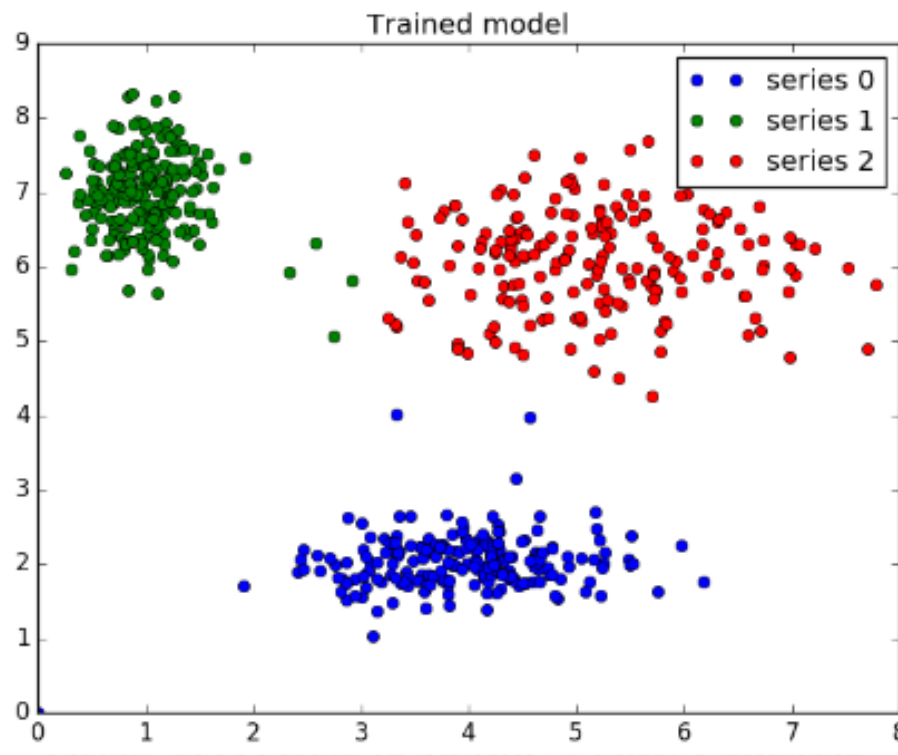
Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a dendrogram. The observations or any number of clusters can be selected by cutting the tree at the correct level.



**Figure 5 Hierarchical Based Clustering**

## Fuzzy clustering:

Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster. Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster. Fuzzy C-means algorithm is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm



**Figure 6 Fuzzy Clustering**

## 3. Clustering Algorithms.[3]

The Clustering algorithms can be divided based on their models that are explained above. There are different types of clustering algorithms published, but only a few are commonly used. The clustering algorithm is based on the kind of data that we are using. Such as, some algorithms need

to guess the number of clusters in the given dataset, whereas some are required to find the minimum distance between the observations of the dataset.

Here we are discussing mainly popular Clustering algorithms that are widely used in machine learning:

### **K-Means algorithm:**

The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of  $O(n)$ .

### **Mean-shift algorithm:**

Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model that works on updating the candidates for centroid to be the center of the points within a given region.

## **4. Clustering Application.[4]**

Below are some commonly known applications of clustering technique in Machine Learning:

### **In Identification of Cancer Cells:**

The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.

### **Customer Segmentation:**

It is used in market research to segment the customers based on their choice and preferences.

## **In Biology:**

It is used in the biology stream to classify different species of plants and animals using the image recognition technique.

## **2.2 Existing Work:**

### **2.2.1 Amazon Recommendation System.**

Amazon uses customer segmentation techniques like clustering to group customers based on their shared characteristics. They collect data on purchase history, browsing behavior, demographics, and more. Relevant features are selected and preprocessed to ensure data quality. Clustering algorithms like K-means, hierarchical clustering, or DBSCAN are applied to identify customer segments. The clusters are evaluated using metrics like silhouette score or WCSS. Amazon interprets the segments to understand customer patterns and preferences. This information is used to personalize experiences and targeted marketing campaigns for different customer segments.

### **2.2.1 Cancer Cell Detection and Classification Using Clustering Techniques.**

This research paper presents a novel approach for cancer cell detection and classification using clustering techniques. Early detection of cancer cells plays a crucial role in improving patient outcomes and survival rates. Traditional methods for cancer cell detection often rely on manual analysis, which is time-consuming and prone to errors. In this study, we propose a clustering-based approach that leverages advanced machine learning algorithms to automate the process of cancer cell detection and classification.

### **2.2.3 Clustering Analysis of the Iris Dataset for Species Identification.**

This research paper presents a comprehensive clustering analysis of the Iris dataset, a popular benchmark dataset in machine learning. The Iris dataset consists of measurements of sepal and

petal lengths and widths for three different Iris flower species. Our study focuses on applying clustering techniques to discover patterns and similarities within the dataset, aiding in species identification and classification.

## **2.3 CRITICAL PAPER REVIEWS**

**Poornachandra Sarang, Centroid-Based Clustering: Clustering Algorithms for Hard Clustering, L. (March 2023). Book of Thinking Data Science, pp. 171-183.[5]**

Many clustering requirements are based on small- to medium-sized datasets. The centroid-based clustering algorithms do an excellent clustering job in such situations. These algorithms find the best value for centroids and group all nearby objects into their respective clusters. These are unsupervised learning algorithms and thus you do not have to worry about creating labeled datasets. The optimization is NP-hard and usually provides only approximate solutions. Though you need to provide an estimate of the number of clusters in prior, these algorithms become an excellent starting point for solving complex clustering problems. In this chapter, I discuss two widely used clustering algorithms in this category—K-means and K-medoids, the latter one is robust to outliers. I also provide you with the techniques for estimating the number of clusters in your dataset before applying the algorithm on a full dataset.

**Duraimoni Neguja and A. Senthil Rajan A Review of Clustering Techniques on Image Segmentation for Reconstruction of Buildings, L. (February 2023), In book: Advanced Communication and Intelligent Systems, First International Conference, ICACIS 2022, Virtual Event, October 20-21, 2022, Revised Selected Papers (pp.401-410).[6]**

The discovery of a new clustering technique on segmented images based on building structures is a challenging process for researchers. In this chapter, clustering techniques on image segmentation in buildings reform is a mingled process of segments of image. This chapter suggests review of various clustering techniques and the improved strategy for the assembling of partitioned image segments of a representation into several areas according to a similarity trial value. In this chapter,



various clustering techniques on distributed particles of image segments is studied as a more complicated procedure that results in computerized model but still common algorithm is not in function. Hence 41 years ago, finding a centralized algorithmic method in clustering separately with help of available data is changed by the lively growth of a broad variety of extremely fussy techniques. The simplest example of the use of a manual interference throughout the charge of clustering outputs on the concern of available procedures. Beyond the kind of input data, the machinist will have to warily choose the finest bespoke process, which major of the moment cannot be done in a routine forum. The prejudiced position of sight of the person is mandatory. Fuzzy C-means clustering, Parallel K-means clustering, Hierarchical density-based clustering and more clustering procedures are compared and studied.

**Jinan Redha, DA Review of Clustering Algorithms, Al-Mustansiriya University, L. (October 2022), DOI:10.5281/zenodo.7243829.[7]**

Clustering is an unsupervised artificial intelligence methodology that has emerged as a good learning tool for evaluating the massive amounts of datasets made available by today's applications. There is an affluence of information available in the field of clustering, and many endeavors have been made to identify and evaluate it for a spectrum of uses; however, the major disadvantages of someone using a classical classification algorithm for big data analysis are their elevated complicity, humongous volume, variety, and generation rate. As a result, typical clustering algorithms for processing such data are rapidly becoming obsolete. In this review, we categorize the review of big data with techniques clustering by identifying the main research concerns. Then, for each subject, we therefore provide an up-to-date review of the research paper.

## **CHAPTER 3**

### **PROBLEM STATEMENT**

#### **3.1 CUSTOMER SEGMENTATION**

Customer Segmentation for Mall Data:

The problem is to segment customers of a mall based on their salary and spending behavior using clustering techniques. The dataset includes information on the salary of each customer and their corresponding spending in the mall. The objective is to identify distinct customer groups with similar salary and spending patterns in order to better understand and cater to their preferences.

The specific goals of this customer segmentation project are:

##### **Group customers based on Salary and Spending:**

Cluster customers into meaningful groups based on their salary and spending behavior. The clusters should capture similarities and differences in both salary and spending patterns.

##### **Discover Customer Segments:**

Identify and interpret different customer segments based on the clustering results. Each segment should represent a distinct group of customers with similar salary and spending characteristics.

##### **Profile Customer Segments:**

Analyze and describe the characteristics of each customer segment, such as average salary range, average spending amount, and any notable trends or patterns observed within each segment.

## **Provide Insights for Business Strategy:**

Extract actionable insights from the customer segmentation results to support marketing and business strategies. This may involve tailoring marketing campaigns, offering personalized promotions, or adjusting product offerings based on the preferences and behaviors of each customer segment.

## **Evaluate and Validate Results:**

Assess the quality and validity of the clustering results using appropriate evaluation metrics. This ensures that the customer segmentation is meaningful, reliable, and aligns with the underlying data distribution.

By addressing this problem statement, the mall management can gain a deeper understanding of their customer base and develop targeted strategies to enhance customer satisfaction, loyalty, and overall business performance.

## **3.2 IRIS DATA SET**

### **Clustering Analysis on Iris Dataset**

The problem is to perform clustering analysis on the Iris dataset, which contains measurements of various attributes of different iris flowers. The objective is to identify distinct groups or clusters within the dataset based on the available features.

The specific goals of this clustering analysis project are:

### **Group Iris Flowers:**

Cluster the iris flowers based on their measurements of sepal length, sepal width, petal length, and petal width. The clusters should capture similarities and differences in these attributes, enabling the identification of distinct groups of iris flowers.

### **Discover Natural Patterns:**

Uncover any underlying patterns or structures in the Iris dataset through clustering. By analyzing the resulting clusters, identify potential relationships or groupings that may exist among the iris flowers.

### **Evaluate Clustering Performance:**

Assess the quality and validity of the clustering results using appropriate evaluation metrics such as silhouette score, Dunn index, or within-cluster sum of squares. This ensures that the clustering analysis is reliable and provides meaningful insights.

### **Visualize Clusters:**

Visualize the identified clusters to gain a better understanding of the distribution and separation of iris flowers based on their attribute measurements. Utilize visual representations such as scatter plots or cluster profiles to facilitate interpretation and analysis.

### **Provide Insights and Interpretations:**

Analyze and interpret the characteristics of each cluster. Identify any distinctive patterns or differences in attribute measurements that define each cluster. This analysis can provide insights into the different species or variations of iris flowers present in the dataset.

### **Potential Applications:**

Discuss potential applications or implications of the clustering analysis on the Iris dataset. This may include species classification, anomaly detection, or further analysis of specific clusters for domain-specific purposes.

By addressing this problem statement, the clustering analysis on the Iris dataset aims to provide insights into the natural grouping or patterns within the iris flowers based on their attribute

measurements. This analysis can contribute to the understanding of the dataset, facilitate species classification and potentially lead to broader application in the field of botany or plant sciences.

## **CHAPTER 4**

### **PROPOSED METHOD**

#### **4.1 CUSTOMER SEGMENTATION**

Customer Segmentation for Mall Data:

To implement customer segmentation of mall data using the WSS (Within- Cluster Sum of Squares) method, elbow point graph, and K-means clustering in Python, you can follow these steps:

##### **Data Preprocessing:**

Start by importing the necessary libraries in Python, load the mall data, and perform any required data preprocessing steps such as handling missing values, scaling, or encoding categorical variables.

##### **Determining Optimal Number of Clusters:**

Use the K-means algorithm to fit multiple models with varying numbers of clusters (k) on the preprocessed data. For each model, calculate the WSS, which represents the sum of squared distances between data points and their respective cluster centroids. Iterate over different values of k and store the WSS values.

##### **Elbow Point Analysis:**

Plot an elbow point graph using the obtained WSS values. The graph will have k on the x-axis and WSS on the y-axis. This graph helps to visualize the point of diminishing returns in terms of reduction in WSS as the number of clusters increases. Look for the "elbow" or the point where the WSS starts to level off significantly.

## **Selecting Optimal Number of Clusters:**

Determine the optimal number of clusters based on the elbow point in the graph. The elbow point represents a balance between the reduction in WSS and the complexity of adding more clusters. This value of  $k$  will be used for the final K-means clustering.

## **K-means Clustering:**

Fit the K-means algorithm with the selected optimal number of clusters on the preprocessed data. Extract the cluster labels assigned to each data point.

## **Cluster Analysis:**

Analyze the obtained clusters by examining the characteristics and behavior of customers within each cluster. Compute descriptive statistics, visualize cluster profiles using box plots or scatter plots, and identify any distinct patterns or differences among the clusters.

## **Interpretation and Application:**

Interpret the results of the customer segmentation and derive actionable insights for business strategies. This may include tailoring marketing campaigns, personalizing customer experiences, or optimizing resource allocation based on the identified customer segments.

Python provides various libraries for implementing these steps, including scikit-learn for K-means clustering, matplotlib for visualizations, and pandas for data preprocessing. By following this approach, you can effectively implement customer

segmentation of mall data using the WSS method, elbow point

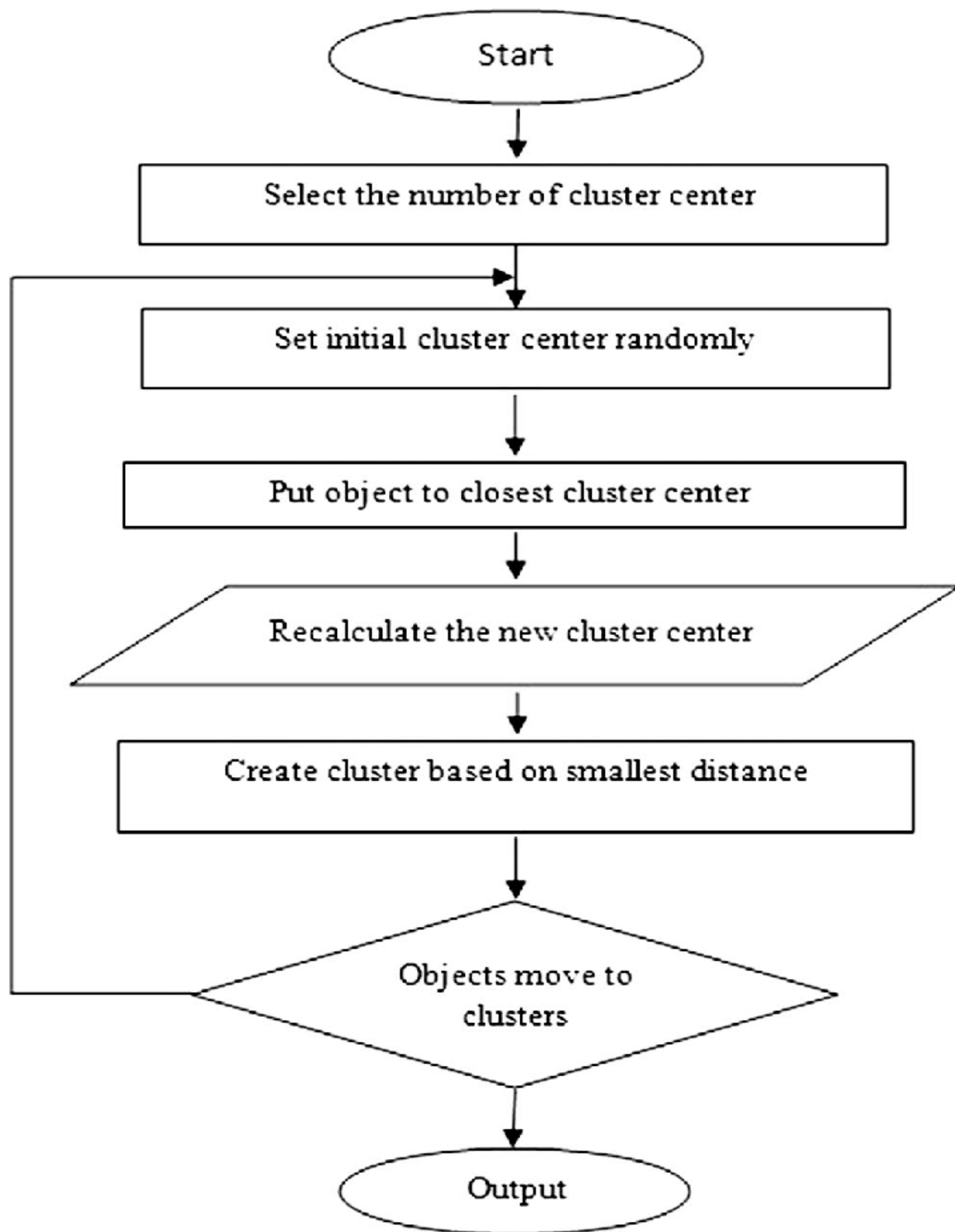


Figure 7 Work Flow Chart 1



## 4.2 IRIS DATASET

Clustering Analysis on Iris Dataset:

To cluster the Iris dataset using the WSS (Within-Cluster Sum of Squares) method, elbow point graph, and K-means clustering in Python, you can follow these steps:

### **Data Preprocessing:**

Start by importing the necessary libraries in Python and load the Iris dataset. Perform any required data preprocessing steps such as scaling or normalization.

### **Determining Optimal Number of Clusters:**

Use the K-means algorithm to fit multiple models with varying numbers of clusters ( $k$ ) on the preprocessed data. For each model, calculate the WSS, which represents the sum of squared distances between data points and their respective cluster centroids. Iterate over different values of  $k$  and store the WSS values.

### **Elbow Point Analysis:**

Plot an elbow point graph using the obtained WSS values. The graph will have  $k$  on the x-axis and WSS on the y-axis. This graph helps visualize the point of diminishing returns in terms of reduction in WSS as the number of clusters increases. Look for the "elbow" or the point where the WSS starts to level off significantly.

### **Selecting Optimal Number of Clusters:**

Determine the optimal number of clusters based on the elbow point in the graph. The elbow point represents a balance between the reduction in WSS and the complexity of adding more clusters. This value of  $k$  will be for the used final K-means clustering.

### **K-means Clustering:**

Fit the K-means algorithm with the selected optimal number of clusters on the preprocessed Iris dataset. Extract the cluster labels assigned to each data point.

### **Visualize the Clusters:**

Plot the clusters using scatter plots, where each data point is colored according to its assigned cluster label. This helps visualize the separation and grouping of the iris flowers based on the chosen features.

### **Interpretation and Analysis:**

Analyze the characteristics of each cluster, such as the average values of the features within each cluster. Compare the clusters to understand the differences and similarities between the iris flowers belonging to different clusters. You can also evaluate the quality of the clustering using appropriate metrics like silhouette score or inter-cluster distance.

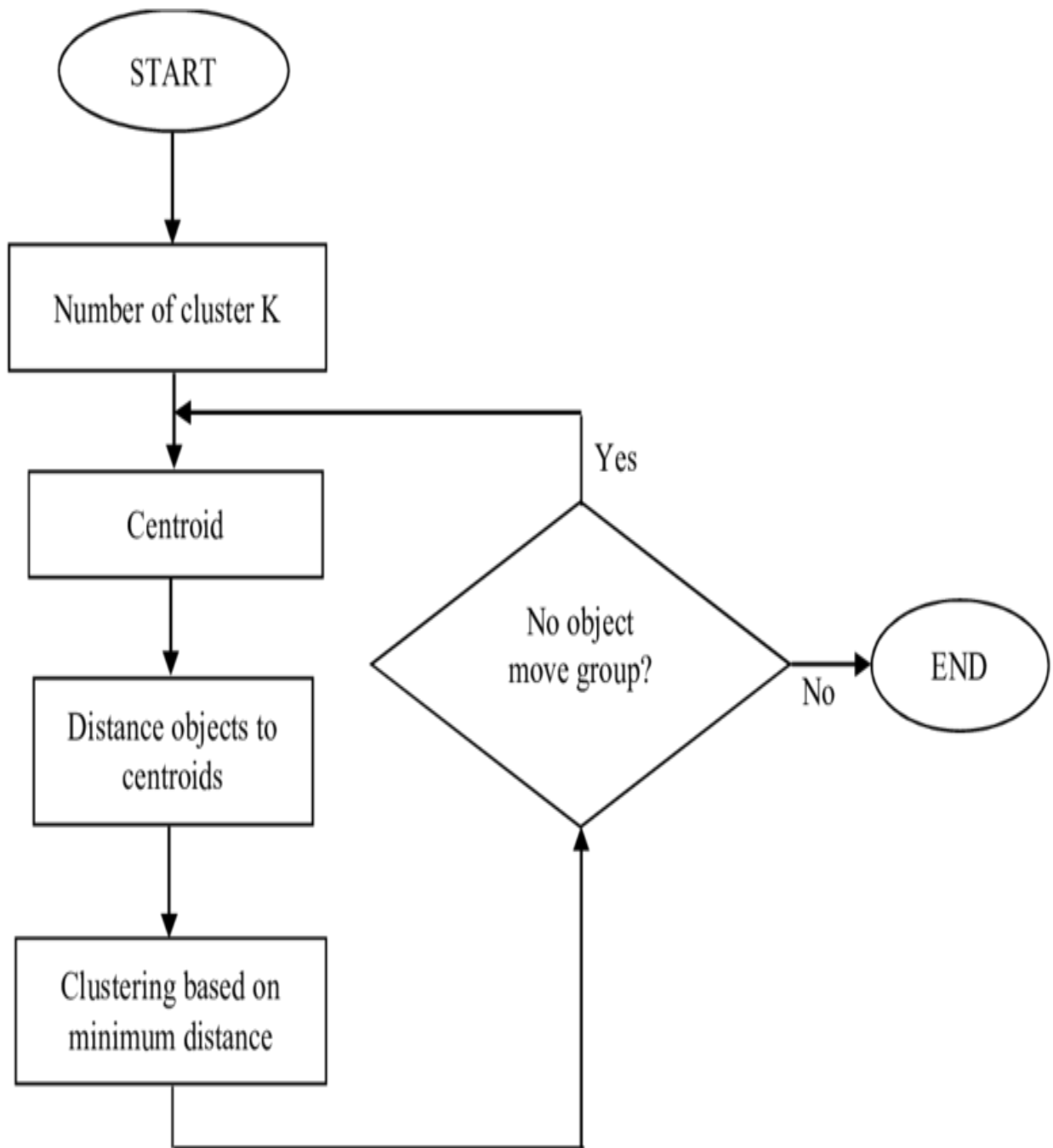
### **Further Analysis and Applications:**

Utilize the obtained clusters for various purposes, such as species classification, anomaly detection, or pattern recognition. For example, you can train a classifier using the cluster labels as target values to predict the species of new iris flowers.

Python provides libraries such as scikit-learn, matplotlib, and pandas that can be used for implementing these steps. By following this method, you can effectively cluster the Iris dataset

using the WSS method, elbow clustering in Python.

point graph, and K-means.

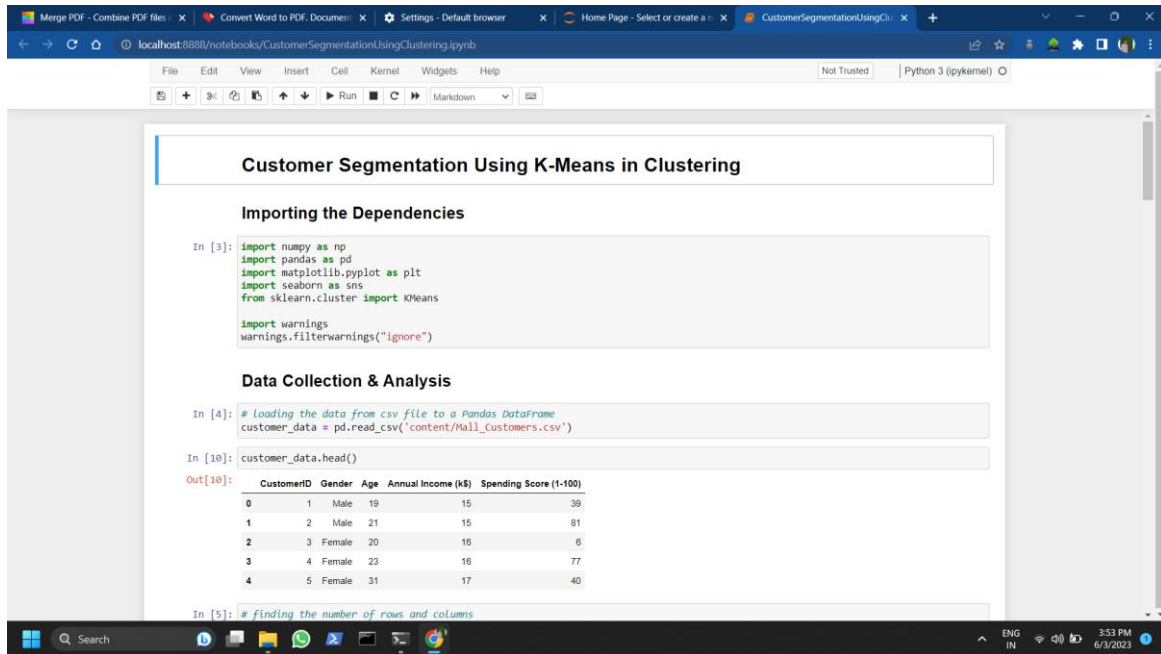


**Figure 8 Work Flow Chart 2**

# CHAPTER 5

## IMPLEMENTATION AND RESULT

### 5.1 CUSTOMER SEGMENTATION



```
File Edit View Insert Cell Kernel Widgets Help
Not Trusted Python 3 (ipykernel)
```

### Customer Segmentation Using K-Means in Clustering

#### Importing the Dependencies

```
In [3]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
import warnings
warnings.filterwarnings("ignore")
```

#### Data Collection & Analysis

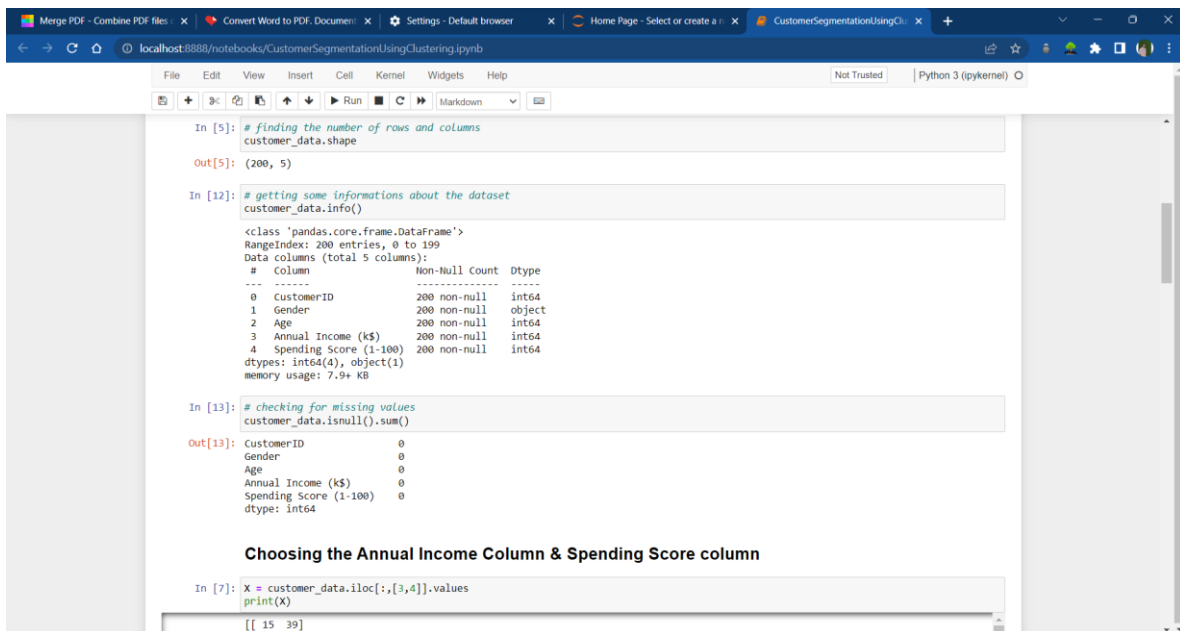
```
In [4]: # loading the data from csv file to a Pandas DataFrame
customer_data = pd.read_csv('content/Mall_Customers.csv')
```

```
In [10]: customer_data.head()
```

```
Out[10]:
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

```
In [5]: # finding the number of rows and columns
```



```
File Edit View Insert Cell Kernel Widgets Help
Not Trusted Python 3 (ipykernel)
```

```
In [5]: # finding the number of rows and columns
customer_data.shape
```

```
Out[5]: (200, 5)
```

```
In [12]: # getting some informations about the dataset
customer_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                Non-Null Count  Dtype
---  ---                ---
 0   CustomerID            200 non-null   int64
 1   Gender                200 non-null   object
 2   Age                   200 non-null   int64
 3   Annual Income (k$)    200 non-null   int64
 4   Spending Score (1-100) 200 non-null   int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

```
In [13]: # checking for missing values
customer_data.isnull().sum()
```

```
Out[13]: CustomerID    0
Gender              0
Age                 0
Annual Income (k$)  0
Spending Score (1-100) 0
dtype: int64
```

#### Choosing the Annual Income Column & Spending Score column

```
In [7]: X = customer_data.iloc[:, [3,4]].values
print(X)
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 ...]
```

Choosing the Annual Income Column & Spending Score column

```
In [7]: X = customer_data.iloc[:,[3,4]].values
print(X)
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 [ 17  76]
 [ 18   6]
 [ 18  94]
 [ 19   3]
 [ 19  72]
 [ 19  14]
 [ 19  99]
 [ 20  15]
 [ 20  77]
 [ 20  13]
 [ 20  79]
 [ 21  35]
 [ 21  66]
 [ 23  29]]
```

Choosing the number of clusters

WCSS -> Within Clusters Sum of Squares

```
In [8]: # finding wcss value for different number of clusters
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)
```

Choosing the number of clusters

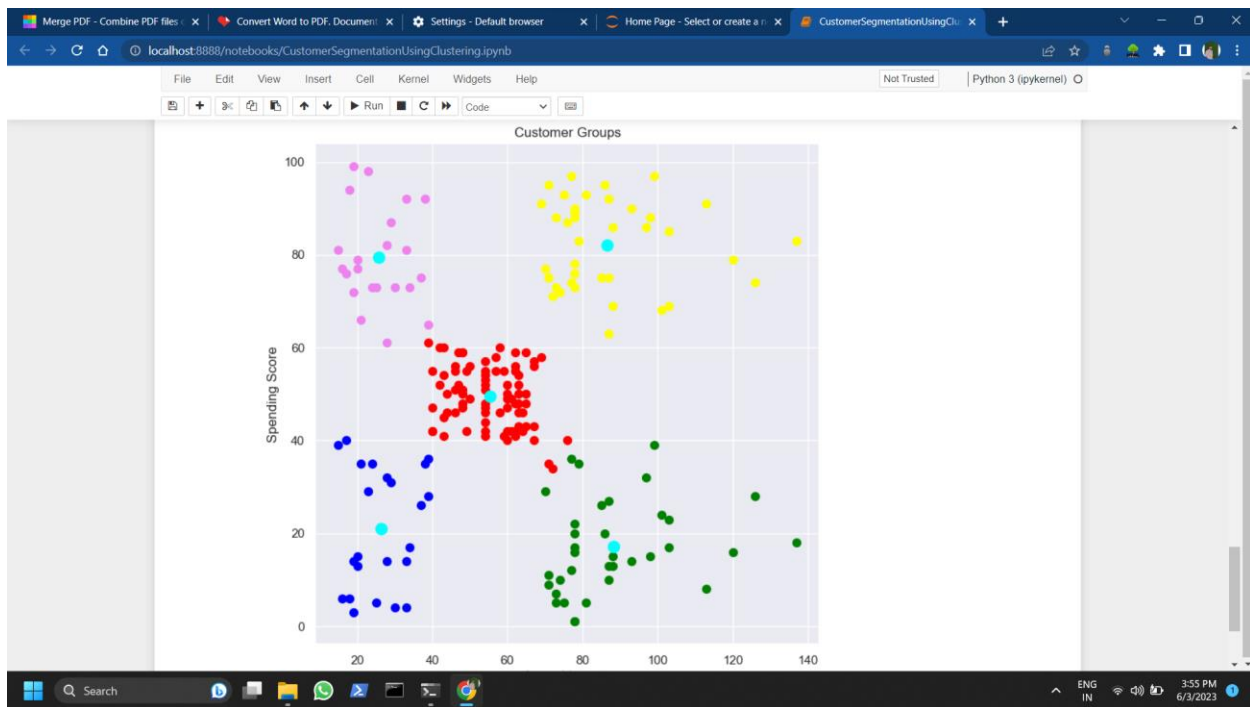
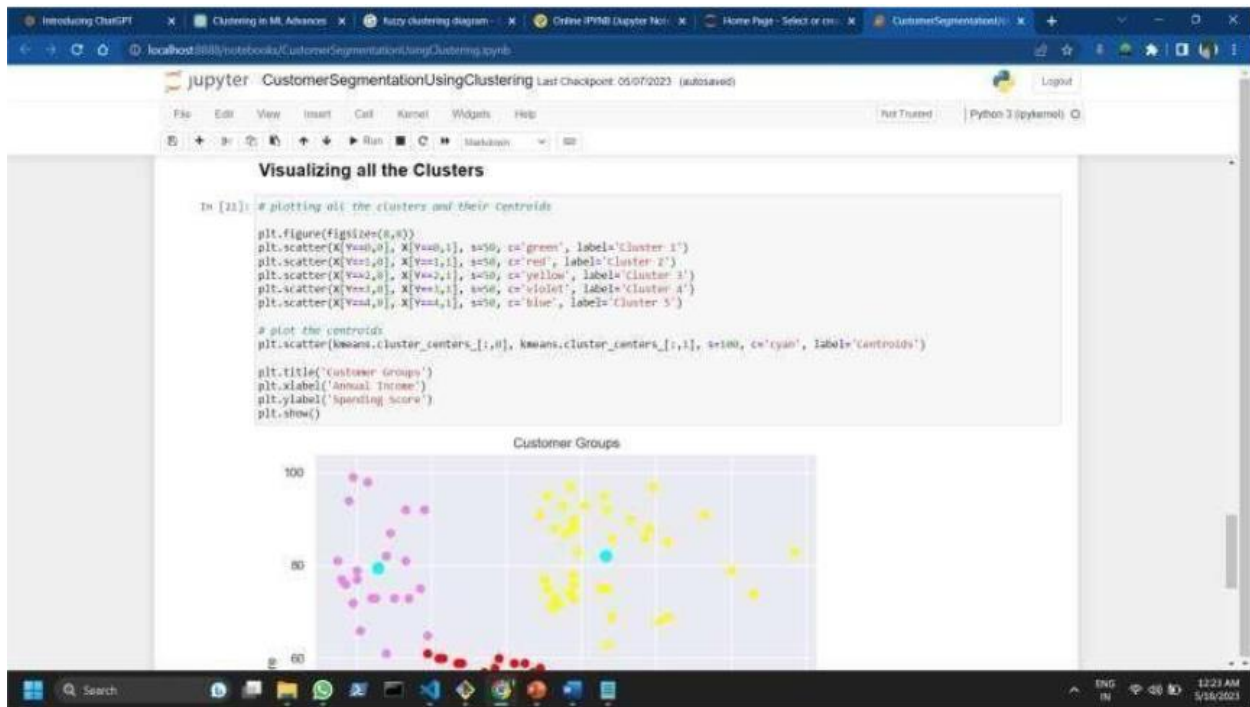
WCSS -> Within Clusters Sum of Squares

```
In [8]: # finding wcss value for different number of clusters
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

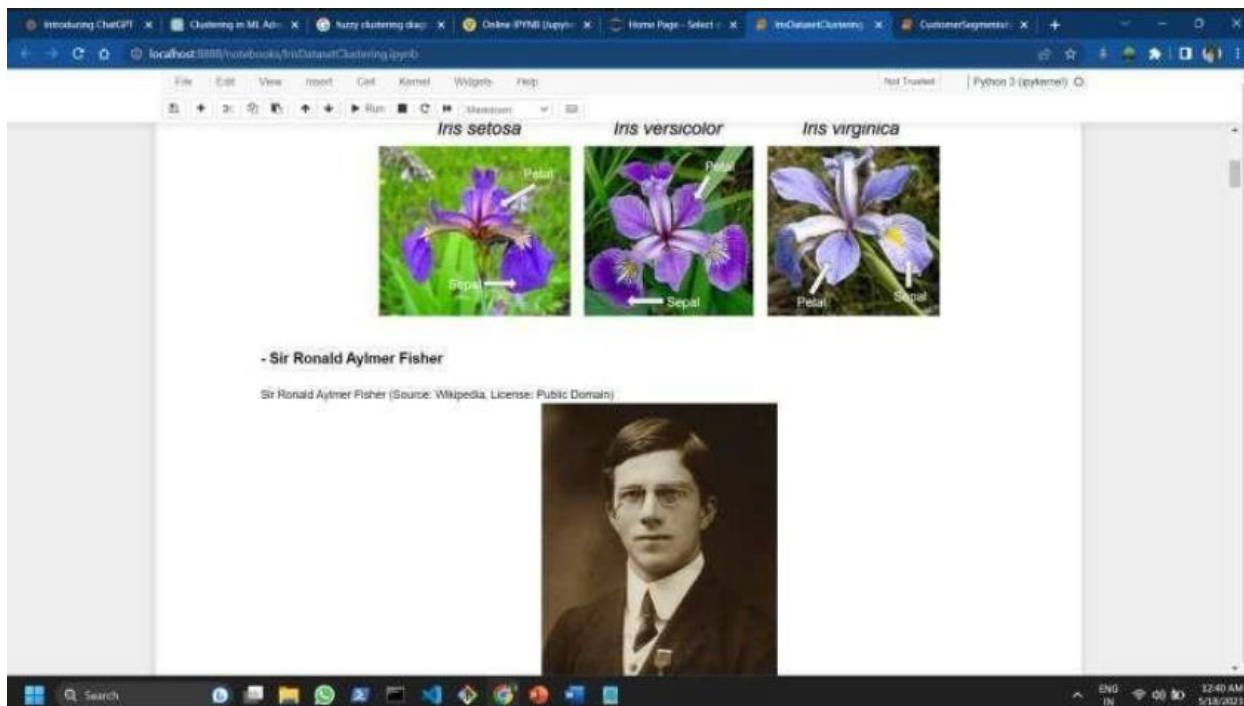
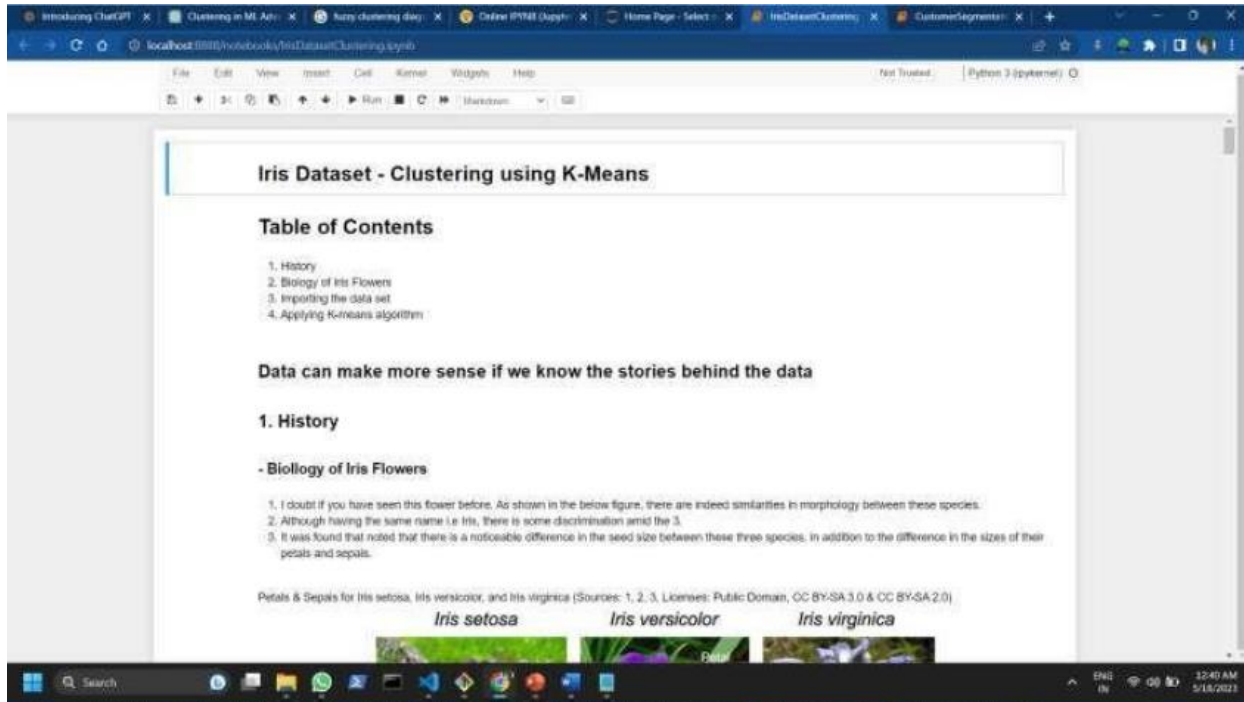
In [16]: # plot an elbow graph
sns.set()
plt.plot(range(1,11), wcss)
plt.title('The Elbow Point Graph')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```

The Elbow Point Graph





## 5.2 Iris Data Set





1. His contribution to statistics is way beyond the Fisher's exact test. For example, he developed the maximum likelihood estimation and the analysis of variance (more commonly known as its acronym ANOVA) test. For these important contributions, he has been highly regarded in the history of modern statistics, as noted on his Wikipedia page.

- Link: <https://online.library.wiley.com/doi/10.1111/j.1469-1808.1936.tb02137.x>
- The t- and z-test methods developed in the 20th century were used for statistical analysis until 1915, when Ronald Fisher created the analysis of variance method.
- ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.

2. Fisher developed and evaluated a linear function to differentiate Iris species based on the morphology of their flowers. It was the first time that the sepal and petal measures of the three Iris species as mentioned above appeared publicly.

3. Fisher developed and evaluated a linear function to differentiate Iris species based on the morphology of their flowers. It was the first time that the sepal and petal measures of the three Iris species as mentioned above appeared publicly. A snapshot of the original data table is provided below. Please note that these measures were recorded in centimeters.

**- Canada's Gaspé Peninsula**



**- Morphological Measures of Iris Flowers (Part of the Iris Dataset, Source & License)**

Table I

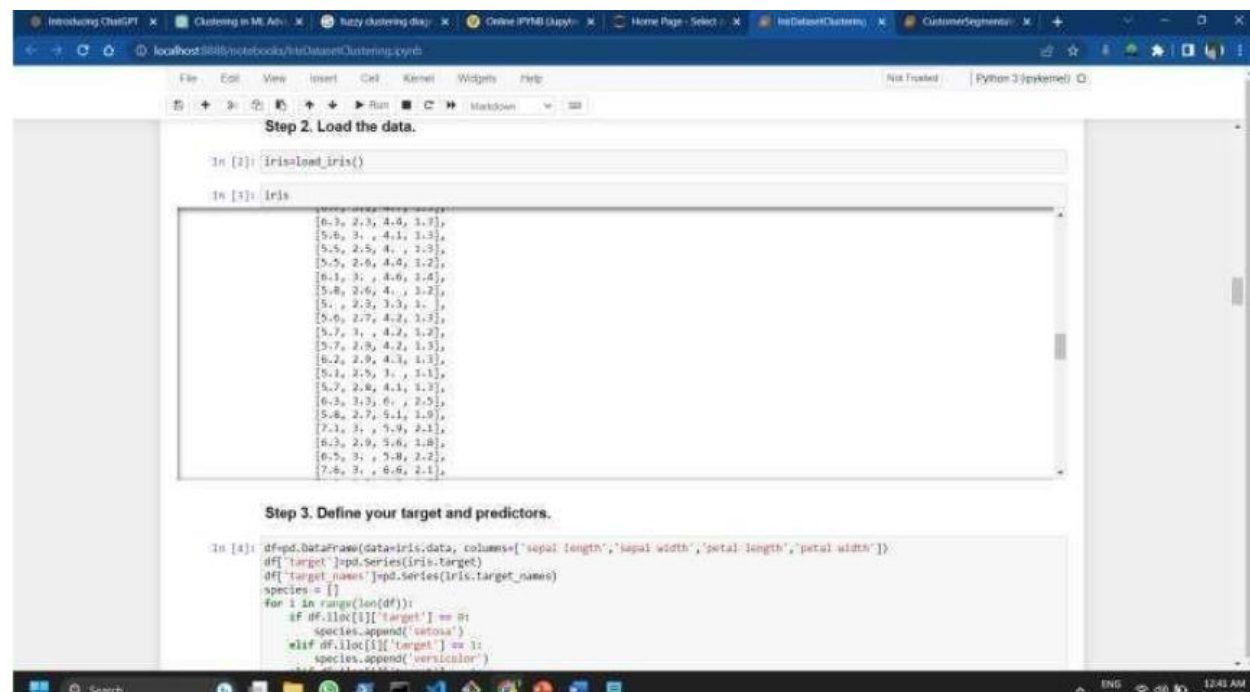
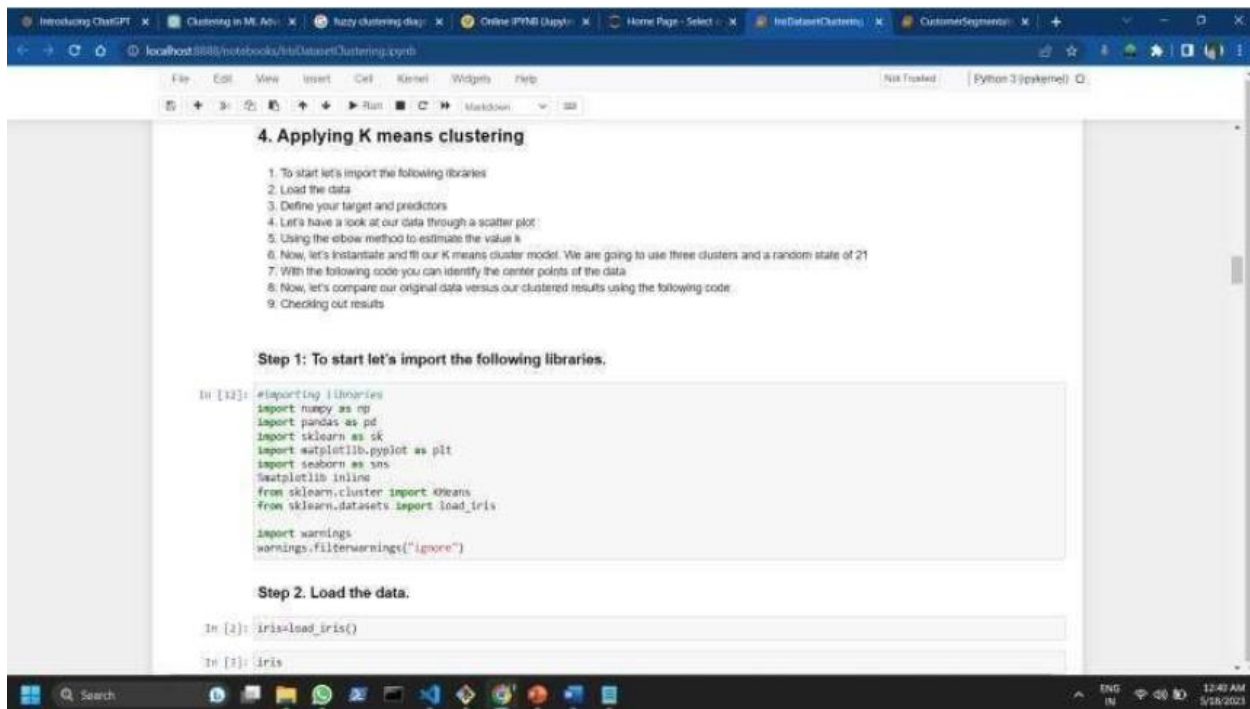
<i>Iris setosa</i>				<i>Iris versicolor</i>				<i>Iris virginica</i>			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.6	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	3.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	3.6	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	3.8	4.3	1.3	7.6	3.0	6.6	2.1
4.8	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	3.9	4.6	1.3	6.7	2.5	5.8	1.8

**2. Introduction to data set**

- The Iris dataset contains the data for 50 flowers from each of the 3 species - Setosa, Versicolor and Virginica.
- The data gives the measurements in centimeters of the following variables for each of the flowers:
  - Sepal length and width
  - Petal length and width
- The task it poses of discriminating between three species of Iris from measurements of their petals and sepals is simple but challenging.

Note: Do your bit for biological correctness in citing the plants concerned carefully. *Iris setosa*, *Iris versicolor* and *Iris virginica* are three species (not varieties, as in some statistical accounts); their binominals should be presented in *italic*, as here, and *Iris* as genus name and the other names indicating particular species should begin with upper and lower case respectively.

**- Goal of the study:**



Introducing ChatGPT x Outlier in ML Adv x fuzzy clustering day x Online IPYNB Deploy x Home Page - Select x IrisDatasetCustom x CustomerSegmentation x

localhost:8888/notebooks/IrisDatasetCustom.ipynb

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Step 3. Define your target and predictors.

```
In [4]: df=pd.DataFrame(data=iris.data, columns=['sepal length','sepal width','petal length','petal width'])
df['target']=pd.Series(iris.target)
df['target_names']=pd.Series(iris.target_names)
species = []
for i in range(len(df)):
    if df.iloc[i]['target'] == 0:
        species.append('setosa')
    elif df.iloc[i]['target'] == 1:
        species.append('versicolor')
    elif df.iloc[i]['target'] == 2:
        species.append('virginica')
df['species'] = species
```

```
In [5]: df
```

```
Out[5]:
```

	sepal length	sepal width	petal length	petal width	target	target_names	Species
0	5.1	3.5	1.4	0.2	0	setosa	setosa
1	4.9	3.0	1.4	0.2	0	versicolor	setosa
2	4.7	3.2	1.3	0.2	0	versicolor	setosa
3	4.6	3.1	1.6	0.2	0	NaN	setosa
4	5.0	3.6	1.4	0.2	0	NaN	setosa
...	...	...	...	...	...	...	...
145	5.7	3.0	5.2	2.3	2	NaN	virginica
146	5.5	2.5	5.0	1.9	2	NaN	virginica
147	5.5	3.0	5.2	2.0	2	NaN	virginica
148	5.2	3.4	5.4	2.3	2	NaN	virginica
149	5.9	3.0	5.1	1.8	2	NaN	virginica

150 rows x 7 columns

12:41 AM 5/18/2023

Introducing ChatGPT x Outlier in ML Adv x fuzzy clustering day x Online IPYNB Deploy x Home Page - Select x IrisDatasetCustom x CustomerSegmentation x

localhost:8888/notebooks/IrisDatasetCustom.ipynb

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

```
In [8]: x=iris.data
x
```

```
Out[8]:
```

```
[[5.4, 1.7, 1.5, 0.2],
 [4.8, 1.4, 1.6, 0.2],
 [4.8, 1.4, 1.4, 0.2],
 [4.9, 3.1, 1.5, 0.1],
 [5.8, 4.1, 1.2, 0.2],
 [5.7, 4.4, 1.5, 0.4],
 [5.4, 3.9, 1.3, 0.4],
 [5.1, 3.5, 1.4, 0.3],
 [5.7, 3.8, 1.7, 0.3],
 [5.1, 1.8, 1.5, 0.3],
 [5.4, 1.4, 1.7, 0.2],
 [5.1, 1.7, 1.5, 0.4],
 [4.6, 1.6, 1.1, 0.2],
 [5.1, 3.3, 1.7, 0.5],
 [4.8, 3.4, 1.9, 0.2],
 [5.1, 3.1, 1.6, 0.2],
 [5.1, 1.4, 1.6, 0.4],
 [5.2, 3.5, 1.5, 0.2],
 [5.2, 3.4, 1.4, 0.2]]
```

Step 4: Let's have a look at our data through a scatter plot.

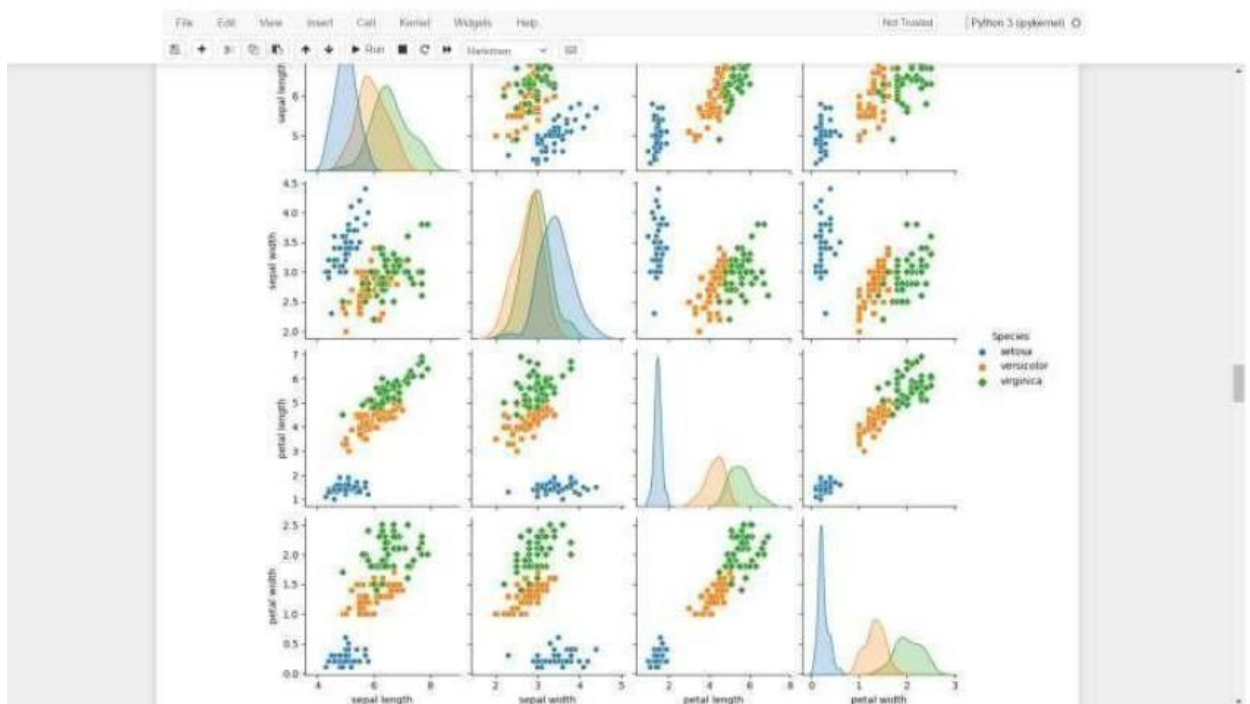
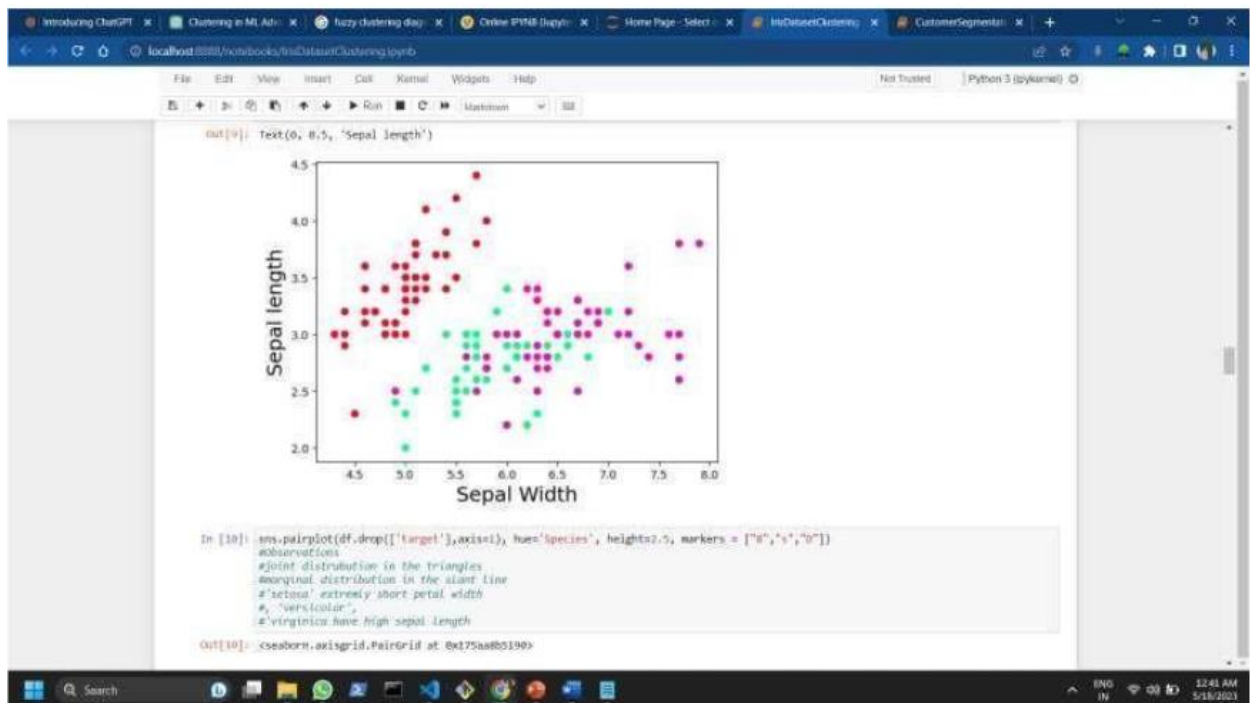
```
In [9]: #scatter plot of Sepal length & width of actual data set
import matplotlib.pyplot as plt
# axes=None, alpha=None, linewidth=None, verts=None, edgecolors=None, *, plotnonfinite=False, data=None, **kwargs)
# Ref: https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.scatter.html
# plt.scatter(x=x[,0], y=x[:,1], c=y, cmap='gist_rainbow')
plt.scatter(x=df['sepal length'], y=df['sepal width'], c=df['target'], cmap='gist_rainbow') #try using cmap='rainbow'

plt.xlabel('sepal width', fontsize=18)
plt.ylabel('sepal length', fontsize=18)
```

```
Out[9]: Text(0, 0.5, 'sepal length')
```

45

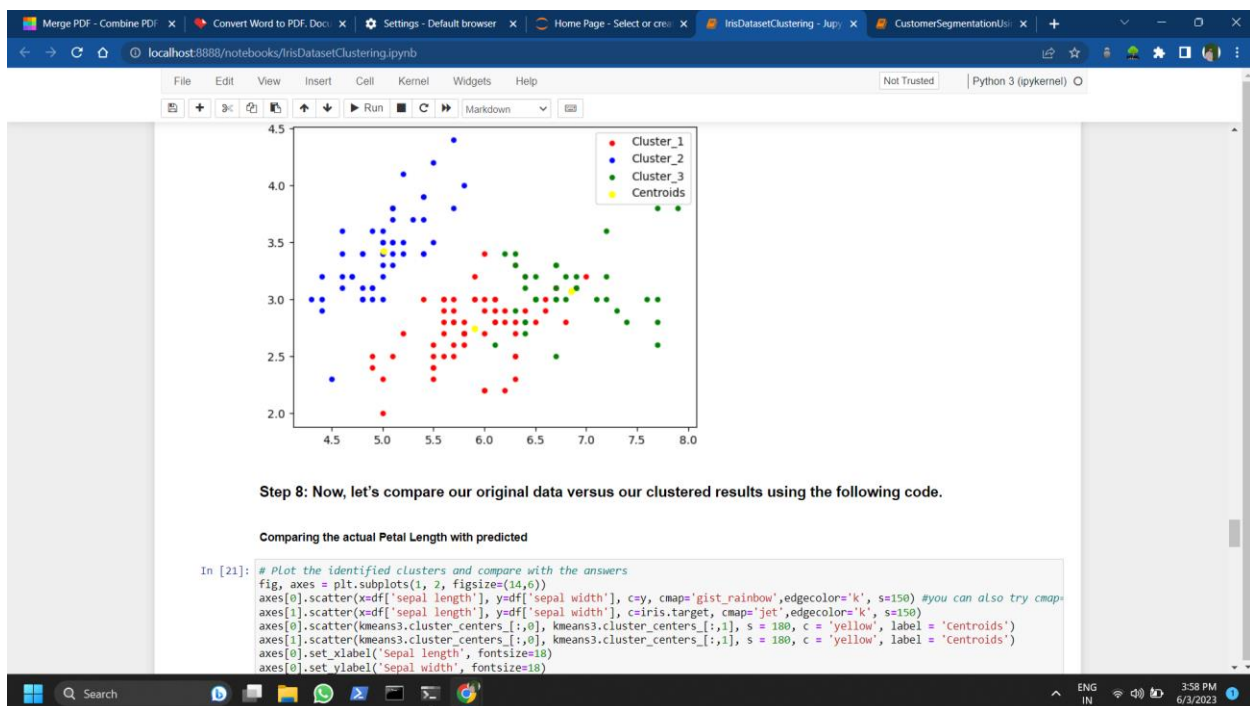
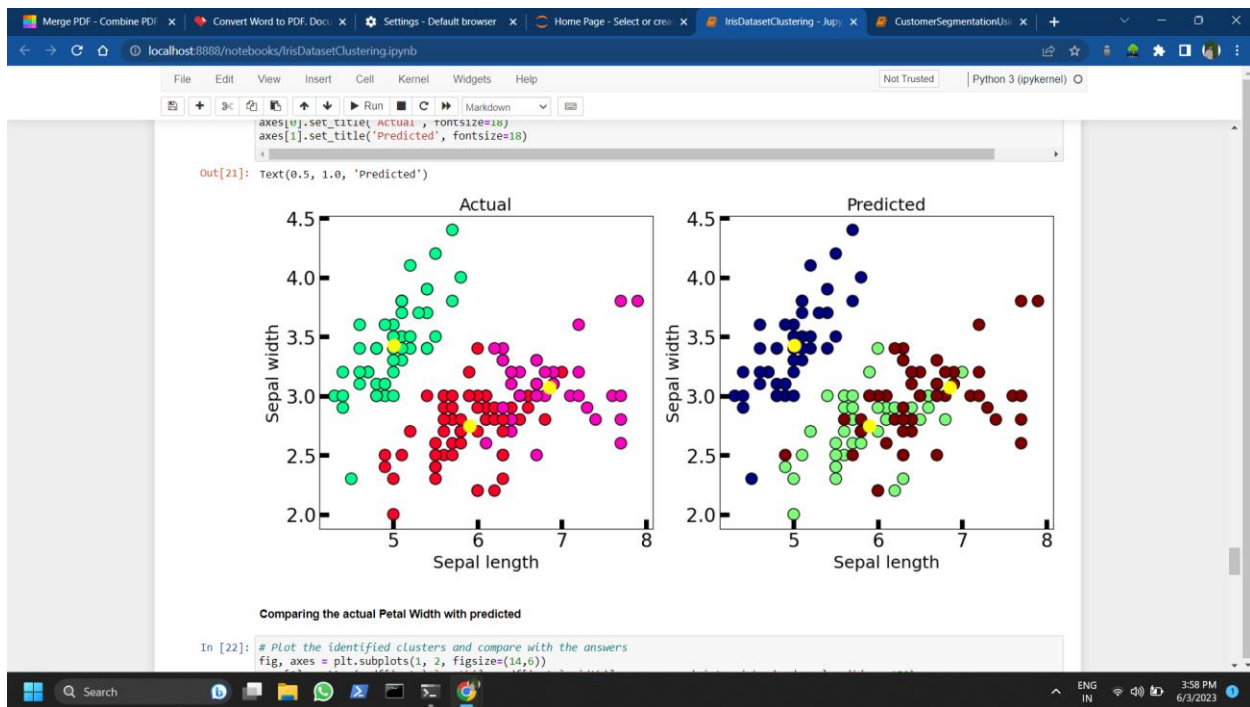
12:41 AM 5/18/2023

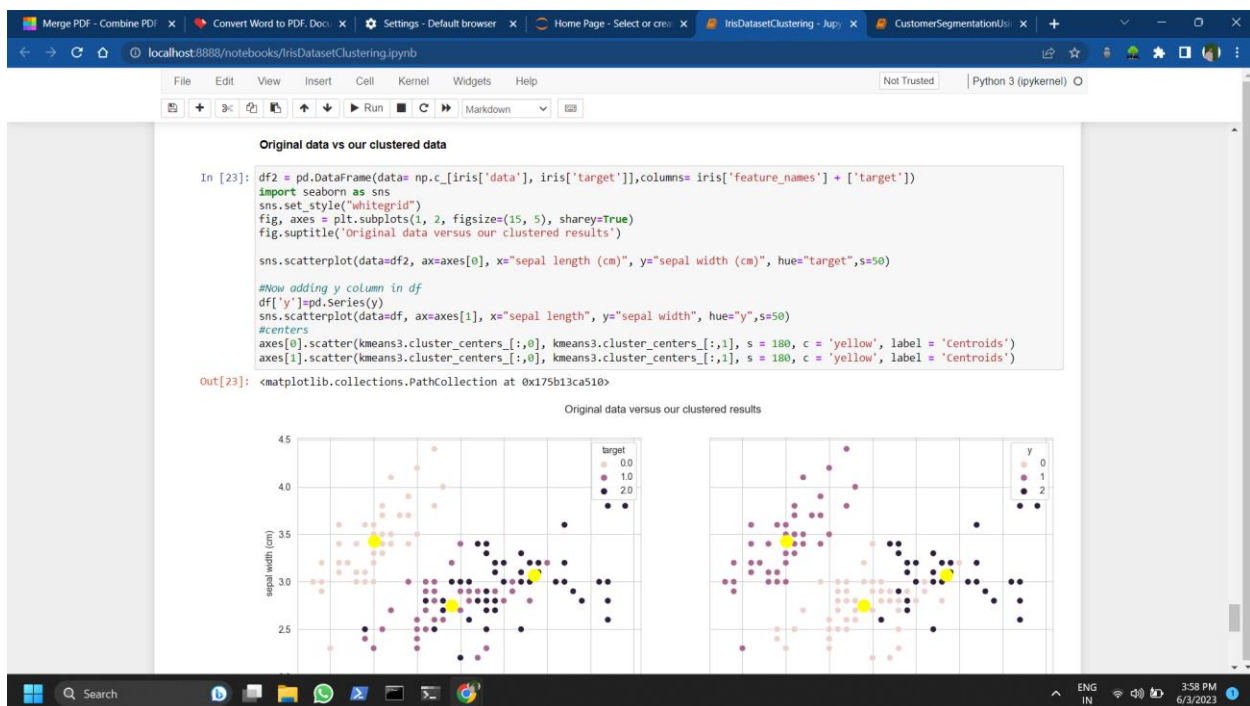
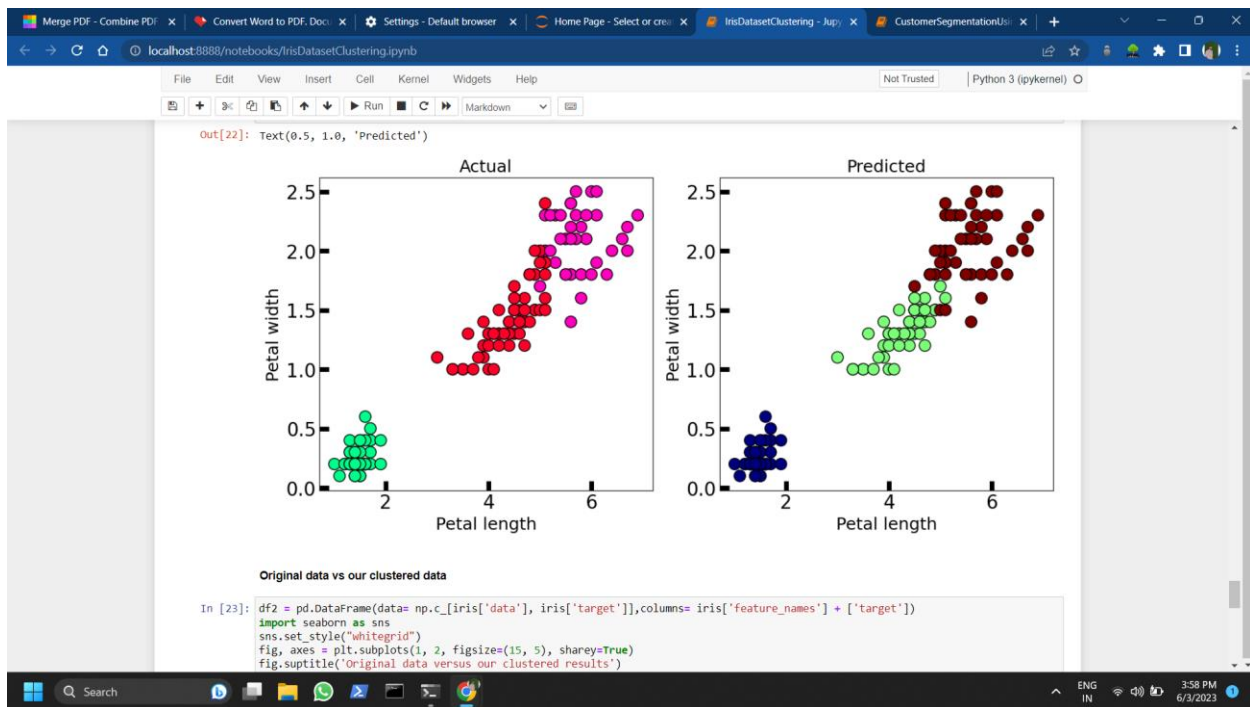




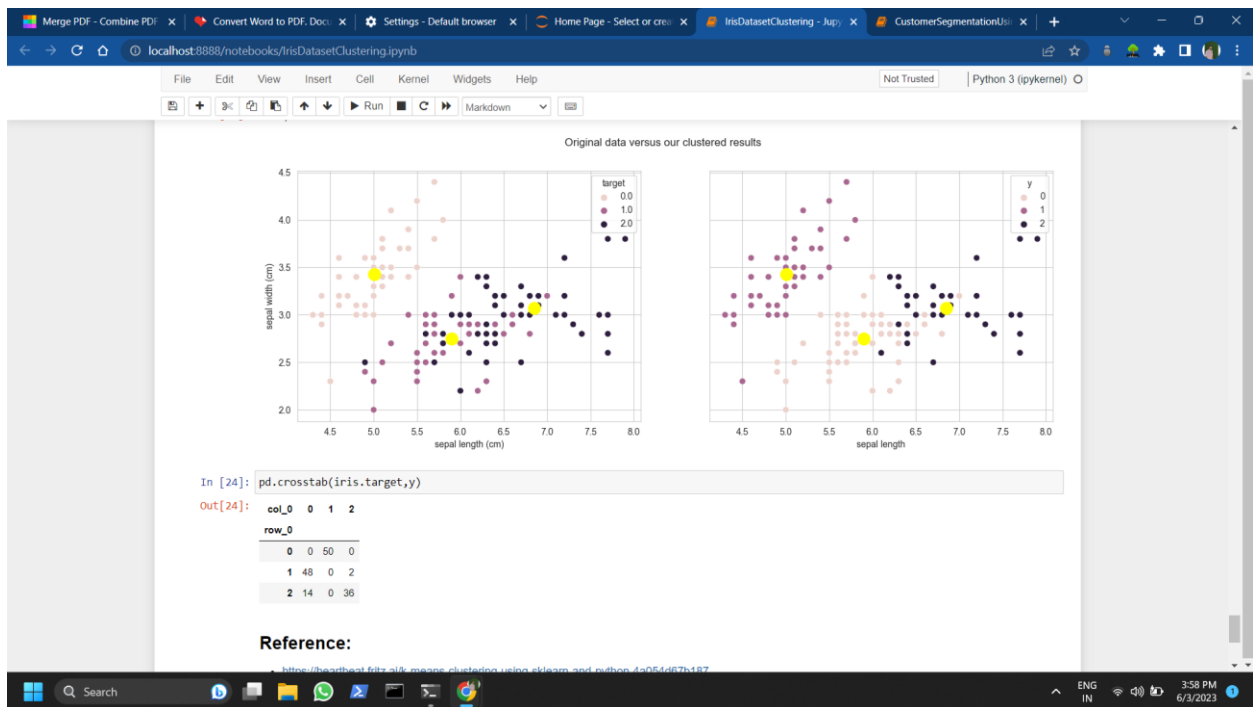












## CHAPTER 6

### TOOLS AND TECHNOLOGIES

To perform clustering using the WSS method, elbow point graph, and K- means clustering, you can utilize the following tools and technologies, including:

#### **Python:**

Python is a popular programming language for data analysis and machine learning tasks. It provides a wide range of libraries and frameworks that are essential for implementing clustering algorithms and data analysis techniques.

#### **Jupyter Notebook:**

Jupyter Notebook is an open-source web application that allows you to create and share documents that contain code, visualizations, and narrative text. It provides an interactive environment where you can run and modify code cells, making it convenient for data exploration, analysis, and visualization.

#### **NumPy:**

NumPy is a fundamental library for numerical computing in Python. It provides efficient data structures and mathematical functions to handle large multidimensional arrays and perform various numerical operations required in data analysis and clustering.

#### **Pandas:**

pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrames, which allow you to efficiently handle and manipulate structured data. pandas is often

used for data preprocessing, cleaning, and transforming the Iris dataset before performing clustering.

### **Scikit-learn:**

scikit-learn is a widely used machine learning library in Python. It provides a comprehensive set of tools and algorithms for various tasks, including clustering. scikit-learn offers a K-means implementation, evaluation metrics, and utility functions that make it easier to perform clustering on the Iris dataset.

### **Matplotlib:**

matplotlib is a plotting library for creating static, animated, and interactive visualizations in Python. It offers various plot types and customization options that help in visualizing the Iris dataset, cluster assignments, and the elbow point graph.

### **Seaborn:**

seaborn is a statistical data visualization library that works on top of matplotlib. It provides high-level interfaces for creating aesthetically pleasing and informative visualizations. seaborn can be used to enhance the visual representation of clusters and perform additional data exploration.

These tools and technologies, including Jupyter Notebook, Python, NumPy, pandas, scikit-learn, matplotlib, and seaborn, provide a powerful and comprehensive environment for implementing clustering algorithms, analyzing data, and visualizing the results on the Iris dataset.

## **CHAPTER 7**

### **CONCLUSION AND FUTURE-WORK**

#### **7.1 Conclusion**

In this project, we successfully implemented clustering techniques. By applying the WSS method, elbow point graph, and K-means clustering algorithm, we were able to group data based on their behavior. The clustering analysis provided valuable insights into customer segments, allowing for targeted marketing strategies, personalized promotions, and improved customer experiences.

The implementation of customer segmentation on the mall data demonstrated the effectiveness of clustering algorithms in identifying distinct customer groups. By leveraging Python and its libraries such as scikit-learn, matplotlib, and pandas, we performed data preprocessing, determined the optimal number of clusters, applied K-means clustering, and analyzed the resulting segments

The implementation of clustering on the Iris dataset showcased the capability of clustering algorithms in identifying distinct clusters within the data. Utilizing Python and libraries such as scikit-learn, matplotlib, and pandas, we performed data preprocessing, determined the optimal number of clusters, applied K-means clustering, and analyzed the resulting clusters.

.

#### **7.2 FUTURE WORK**

##### **7.2.1 Feature Engineering:**

Explore additional features, such as demographic data, customer preferences, or transaction history, to improve the accuracy and granularity of customer segmentation. Incorporating more relevant variables can enhance the understanding of customer behavior and preferences.

##### **7.2.2 Evaluation Metrics:**

Consider using alternative evaluation metrics, such as silhouette score, Dunn index, or entropy, to assess the quality and stability of the clustering results. Evaluating multiple metrics can provide a more comprehensive evaluation of the clustering performance.

### **7.2.3 Alternative Clustering Algorithms:**

Experiment with other clustering algorithms, such as hierarchical clustering, DBSCAN, or Gaussian mixture models, to compare their performance and explore different perspectives of customer segmentation. Each algorithm has its own strengths and limitations, and exploring alternatives can provide additional insights.

### **7.2.4 Visualization Techniques:**

Explore advanced visualization techniques, including interactive plots, 3D visualizations, or dimensionality reduction techniques, to visualize the customer segments in a more intuitive and informative way. Visualizations can aid in understanding the relationships and differences between customer groups.

### **7.2.5 Dynamic Segmentation:**

Develop a system or pipeline that can perform real-time or dynamic customer segmentation. This can involve updating the segmentation as new customer data becomes available, allowing for adaptive marketing

## References

1. T. Handhayani and I. Wasito, "Fully unsupervised clustering in nonlinearly separable data using intelligent Kernel K-Means," 2014 International Conference on Advanced Computer Science and Information System, Jakarta, Indonesia, 2014, pp. 450-453, doi: 10.1109/ICACSYS.2014.7065891. Accessed:23 Dec 2022 [Online].
2. Mann, Amandeep Kaur, and Navneet Kaur. "Survey paper on clustering techniques." International journal of science, engineering and technology research 2.4 (2013): 803-806. Accessed:23 Dec 2022 [Online].
3. K. Bindra and A. Mishra, "A detailed study of clustering algorithms," 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2017, pp. 371-376, doi: 10.1109/ICRITO.2017.8342454. Accessed:23 Dec 2022 [Online].
4. Harrigan, K.R. (1985), An application of clustering for strategic group analysis. Strat. Mgmt. J., 6: 55-73. <https://doi.org/10.1002/smj.4250060105>, Accessed:23 Dec 2022 [Online].
5. Poornachandra Sarang, Centroid-Based Clustering: Clustering Algorithms for Hard Clustering, L. (March 2023). Book of Thinking Data Science, pp. 171-183, Accessed:23 Dec 2022 [Online].

6. Duraimoni Neguja and A. Senthil Rajan, A Review of Clustering Techniques on Image Segmentation for Reconstruction of Buildings, L. (February 2023), In book: Advanced Communication and Intelligent Systems, First International Conference, ICACIS 2022, Virtual Event, October 20-21, 2022, Revised Selected Papers (pp.401-410). Accessed:23 Dec 2022[Online].
7. Jinan Redha, DA Review of Clustering Algorithms, Al-Mustansiriya University, L. (October 2022), DOI:10.5281/zenodo.7243829. Accessed:23 Dec 2022 [Online].