# SEGMENTATION OF SHOPPING MALL CUSTOMERS USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

**D.V.Jaya Surya ( Reg.No - 39110250 )**
**M.V.Tharun kumar ( Reg.No – 39110578 )**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF COMPUTING**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**
**(DEEMED TO BE UNIVERSITY)**
**Accredited with Grade “A” by NAAC | 12B Status by UGC | Approved by AICTE**
**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**
**CHENNAI - 600119**

**APRIL - 2023**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

## DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

## BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Dasa Venkata Jaya Surya (Reg.No-39110250)** who carried out the project Phase-2 entitled "**Segmentation of shopping mall customers using machine learning"** under my supervision from January 2023 to April 2023.

**Internal Guide**
Dr.M.D.Anto Praveena M.E.,Ph.D.,

**Head of the Department**
Dr. L. Lakshmanan M.E., Ph.D.,

**Submitted for Viva-voce Examination held on 20.04.2023**

**Internal Examiner** **External Examiner**

# DECLARATION

I **Dasa Venkata Jaya Surya**(Reg.No-39110250),hereby declare that the ProjectPhase-2 Report entitled "**Segmentation of shopping mall customers using machine learning**" done by us under the guidance of**Dr.M.D.Anto Praveena M.E.,Ph.D.,**is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering.**

Venkata jaya Surya

**DATE:20.04.2023**

**PLACE: Chennai** **SIGNATURE OF THECANDIDATE**

# ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D.**, **Dean**, School of Computing **Dr.L.Lakshmanan M.E., Ph.D.,** Heads ofthe Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressivereviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide,**Dr.M.D.Anto Praveena M.E.,Ph.D.,**for her valuable guidance, suggestions, and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in manywaysfor the completion of the project.

# ABSTRACT

In this modern era, everything and everyone is innovative, where everyone competes with being better than others. The emergence of many entrepreneurs, competitors, and business interested people has created a lot of insecurities and tension among competing businesses to find new customers and hold the old customers. Because of this one should need and maintain exceptional customer service and it becomes very appropriate irrespective of the business scale. And also, it is equally important to understand the needs of customers specifically to provide greater customer support and to advertise them with the most appropriate products. In the pool of these online products customers are confused about what to buy and what not to and also the company or the business people are confused about which section of customers to be targeted for selling their particular type of products. This confusion will probably be possible by the process called CUSTOMER SEGMENTATION. The process of segmenting the customers with similar interests and similar shopping behavior into the same segment and with different interests and different shopping patterns into different segments is called customer segmentation. Customer segmentation and pattern extraction are the major aspects of a business decision support system. Each segment has the same set of customers who most probably has the same kind of interests and shopping patterns. In this paper, we planned to do this customer segmentation using three different clustering algorithms namely K means clustering algorithm, Mini batch means, and hierarchical clustering algorithms and also going to compare all these clustering algorithms based on their efficiency and root mean squared errors.

***Keywords:*** Customer segmentation, Clustering, K-means clustering, Mini Batch Kmeans clustering, hierarchical clustering.

# TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF ABBREVIATIONS

| ABBREVIATIONS | EXPANSION |
|---|---|
| AI | Artificial Intelligence |
| EDA | Exploratory Data Analysis |
| GUI | Graphical User Interface |
| ML | Machine Learning |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| RFM | Recency ,Frequency, and Monetary |
| RMSE | Root Mean Square Error |
| SNA | Social Network Analysis |

# CHAPTER 1

# INTRODUCTION

## 1.1 OVERVIEW

Data is very precious in today's ever-competitive world. Every day organizations and people are encountered with a large amount of data. A most efficient way to handle this data is to classify or categorize the data into Clusters, set of groups, or partitions. "Usually, the classification methods are either supervised or unsupervised, depending on whether they have labeled datasets or not". Unsupervised classification is the exploratory data analysis where there won't be any training data set and having to extract hidden patterns in the data set with no labeled responses is achieved whereas classification of supervised learning model is machine learning task of deducing a function from training data set. The main focus is to enhance the propinquity or closeness in data points belonging to the same group and increase the variance among various groups and all this is achieved through some measure of similarity. Exploratory- by data analysis is all about dealing with a wide range of applications such as " engineering, text mining, pattern recognition, bioinformatics, spatial data analysis, - mechanical engineering, voice mining, textual document collection, artificial intelligence, image segmentation, ". This diversity explains the importance of clustering in scientific research but this diversity can lead to contradictions due to different purposes and nomenclature.

Maintaining and Managing relationships of a customer have always played a very key role to provide business intelligence to companies to build, develop and manage very important long-term relationships with customers. The importance of treating customers as a main asset to the organization is increasing in the present- day era. By using clustering techniques like k-means, mini-batch k-means, hierarchical clustering customers with the same habits are clustered as one cluster. Segmentation of customers helps the team of marketing to recognize different customer segments that think differently and follow different purchasing techniques and strategies. Customer segmentation helps figure out the customers who vary in terms of purchasing habits, expectations, desires, preferences, and attributes. The important purpose of doing customer segmentation is to group customers, who have the same interests so that the marketing or business team can converge in an effective marketing plan. The techniques of clustering consider data tuples as objects. They partition the data objects into clusters or groups. Customer Segmentation is the process where one has to divide the customers into various groups called customer segments so each customer segment comprises customers who have similar interests and patterns. The segmentation process is mostly based on the similarity or the identical nature in different ways that are relevant to marketing features like age, gender, interests, and miscellaneous spending habits.

Customer segmentation has importance as it includes, the ability to modify the programs of the market so that it is suitable to each of the segments, support in a business decision, identification of products associated with each customer segment, and managing the demand and supply of that product, and predicting customer defection, identifying and targeting the potential customer base, providing directions in finding the solutions. Clustering is an iterative process of knowledge discovery from unorganized and huge amounts of data that is raw. Clustering is one of the kinds of exploration of data mining that is used in several applications, those are classification, machine learning, and recognition of patterns

## 1.2 MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine learning could be a subfield of computer science (AI). The goal ofmachine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient processapproaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge to modify decision-making processes supported knowledge inputs.

### *1.2.1 History and relationships to other fields*

The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer, and pioneer in the field of computer gaming and artificial intelligence. Also, the synonym self-teaching computers were used in this period. A representative book of machine learning research during the 1960s was Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973.In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.

Modern-day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning to train it to classify the cancerous moles. Whereas, a machine learning algorithm for stock trading may inform the trader of future potential predictions.

## 1.3 MACHINE LEARNING APPROACHES

In machine learning, tasks square measure istypically classified into broad classes. These classes square measure supported however learning is received or however, feedback on the education is given to the system developed. Two of the foremost wide adopted machine learning strategies are square measure supervised learning that trains algorithms supported example input and output information that's tagged by humans, and unattended learning that provides the algorithmic program with no tagged information to permit it to search out structure at intervals its computer file.

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

***Supervised learning:*** The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

***Unsupervised learning:*** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

***Reinforcement learning:*** A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, theprogram is provided feedback that's analogous to rewards, which it tries to maximize.

### *1.3.1 SupervisedLearning*

In supervised learning, the pc is given example inputs that square measure labeledwith their desired outputs.The aim of this technique is for the algorithmic program to be ready to “learn” by comparing its actual output with the “taught” outputs to search out errors, and modify the model consequently. Supervised learning thus uses patterns to predict label values on extra unlabeled information. For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical information to predict statistically probably future events. It's going to use historical stock exchange info to anticipate approaching fluctuations or be used to filter spam emails. In supervised learning, labeled photos of dogs are often used as input files to classify unlabeled photos of dogs.

Types of supervised learning algorithms include active learning, classification, and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification,andspeakerverification.

### *1.3.2 Unsupervised Learning*

In unsupervised learning, information is unlabeled, and the learning rule is left to seek out commonalities among its input file. The goal of unattended learning is also as easy as discovering hidden patterns at intervals in a dataset, however, it should even have a goal of feature learning, that permits the procedure machine to mechanically discover the representations that square measure required to classify data.In this information fed into the Associate in Nursing unattended learning rule, it should be determined that ladiesof a definite age vary UN agency obtain unscented soaps square measure probably to be pregnant, and so a promotingcampaignassociatedwithphysiological conditionandbabywillbemerchandised.

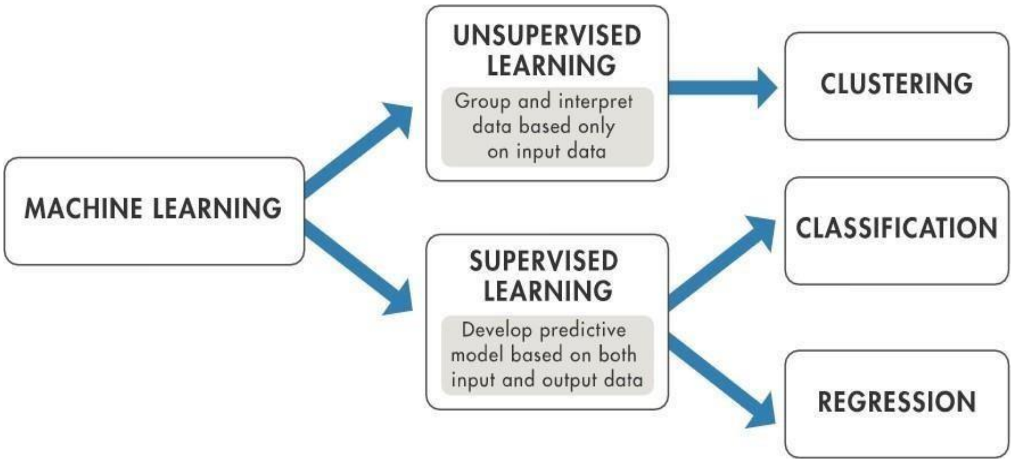
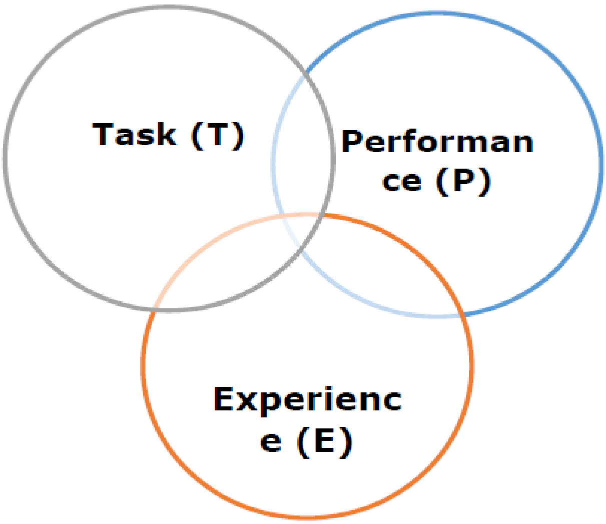

MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

**Fig.1.1Machine LearningClassification**

**In Fig 1.1**Unsupervised learning is usually used for transactional information. You will have an oversized dataset of consumers and their purchases, however, as a person, you'll probably not be able to add up what similar attributes will be drawn from client profiles and their styles of purchases.



### 1.2 Machine Learning Task

**In Fig 1.2** The task of clustering is subjective which means there are many ways of achieving the goal of clustering. Each methodology has its own set of rules to segregate data points into different clusters. There is n number of clustering algorithms in which these are few mostly used algorithms such as K means clustering algorithm, Hierarchical clustering algorithms, and Mini-batch K means clustering algorithm, etc.

### 1.4 CLUSTERING

Clustering is the task of dividing the data points into definite groups such that the data points in the same group have similar characteristics or similar behavior. In short, segregating the data points into different clusters based on their similar traits.

Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labeled output variables. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets.

It depends on the type of algorithm we use which decides how the clusters will be created. The inferences that need to be drawn from the data sets also depend upon the user as there is no criterion for good clustering.

### *1.4.1 Types of Clustering*

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.

### *1.4.2 Density-Based Clustering*

In this method, the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered clusters.

The data points in the sparse region (the region where the data points are very few) are considered as noise or outliers. The clusters created in these methods can be of arbitrary shape.

### *1.4.3 Hierarchical Clustering*

Hierarchical Clustering groups (Agglomerative or also called Bottom-Up Approach) or divides (Divisive or also called Top-Down Approach) the clusters based on the distance metrics. In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one.

Divisive is the opposite of Agglomerative, it starts with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram.

### *1.4.4 Centroid-based*

Centroid-based clustering is the one you probably hear about the most. It's a little sensitive to the initial parameters you give it, but it's fast and efficient.

These types of algorithms separate data points based on multiple centroids in the data. Each data point is assigned to a cluster based on its squared distance from the centroid. This is the most commonly used type of clustering.

K-Means clustering is one of the most widely used algorithms. It partitions the data points into k clusters based upon the distance metric used for the clustering. The value of 'k' is to be defined by the user. The distance is calculated between the data points and the centroids of the clusters.

The data point which is closest to the centroid of the cluster gets assigned to that cluster. After an iteration, it computes the centroids of those clusters again and the process continues until a pre-defined number of iterations are completed or when the centroids of the clusters do not change after an iteration.

It is a very computationally expensive algorithm as it computes the distance of every data point with the centroids of all the clusters at each iteration. This makes it difficult for implementing the same for huge data sets.

### *1.4.5 Applications of Clustering*

Clustering is used in our daily lives such as in data mining, in academics, in web cluster engines, in bioinformatics, in image processing, and many more. There are a few common applications where clustering is used as a tool are Recommendation engines, Market segmentation, Customer segmentation, Social Network Analysis(SNA), Search result Clustering, Identification of cancer cells, biological data analysis, and medical imaging analysis.

[4] **Scalability** − Some clustering algorithms work well in small data sets including less than 200 data objects; however, a huge database can include millions of objects. Clustering on a sample of a given huge data set can lead to biased results. There are highly scalable clustering algorithms are required.

- **Ability to deal with different types of attributes** − Some algorithms are designed to cluster interval-based (numerical) records. However, applications can require clustering several types of data, including binary, categorical (nominal), and ordinal data, or a combination of these data types.

- **Discovery of clusters with arbitrary shape** − Some clustering algorithms determine clusters depending on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to discover spherical clusters with the same size and density. However, a cluster can be of any shape. It is essential to develop algorithms that can identify clusters of arbitrary shapes.
- **Minimal requirements for domain knowledge to determine input parameters** − Some clustering algorithms needed users to input specific parameters in cluster analysis (including the number of desired clusters). The clustering results are quite sensitive to input parameters. Parameters are hard to decide, specifically for data sets including high-dimensional objects. This not only burdens users but also creates the quality of clustering tough to control.
- **Ability to deal with noisy data** − Some real-world databases include outliers or missing, unknown, or erroneous records. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
- **Insensitivity to the order of input records** − Some clustering algorithms are responsive to the order of input data, e.g., the similar set of data, when presented with multiple orderings to such an algorithm, and it can generate dramatically different clusters. It is essential to develop algorithms that are unresponsive to the order of input.
- **High dimensionality** − A database or a data warehouse can include several dimensions or attributes. Some clustering algorithms are best at managing low- dimensional data, containing only two to three dimensions. Human eyes are best at determining the quality of clustering for up to three dimensions. It is disputing to cluster data objects in high-dimensional space, especially considering that data in high-dimensional space can be very inadequate and highly misrepresented.

**Constraint-based clustering** − Real-world applications can be required to performclustering under several types of constraints. Consider that your job is to select the areas for a given number of new automatic cash stations (ATMs)in a city.

# CHAPTER 2

# LITERATURE SURVEY

## 2.1 RELATED WORK

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex task. This is because customers may be different according to their demands, desires, preferences and so on. Instead of "one-size-fits-all" approach, customer segmentation clusters the customers into groups sharing the same properties or behavioural characteristics. According to, customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, demographical conditions, data geographical conditions and economic conditions as well as behavioural patterns. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, plan the marketing budget, determining new market opportunities, making better brand strategy, identifying customers retention.

Algorithm Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space. According to K-means algorithm in one of the most popular centroid based algorithms. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, C1,……Ck that is, $Ci \subset D$ and $Ci \cap Cj = \emptyset$ for $(1 \leq i, j \leq k)$. A centroid-based partitioning technique uses the centroid of a cluster, Ci, to represent that cluster. Conceptually, the centroid of a cluster is its centre point. The difference between an object p ∈ Ci and ci, the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y.

**Algorithm**: The k-means algorithm for partitioning, where each cluster's centre is represented by the mean value of the objects in the cluster. Input: k: the number of clusters, D: a data set containing n objects. Output: A set of k clusters. Method: (1) arbitrarily choose k objects from D as the initial cluster centres; (2) repeat (3) (re)assigns each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

In conclusion,the literature review highlights the significance of customer segmentation in attracting the customers towards the products which in turn aids the increase in the business scale in the market. Segmenting the customer group.

# CHAPTER 3

# METHODOLOGY

## 3.1 EXISTING SYSTEM

The existing model for the customer segmentation depicts that it is based on the K-means clustering algorithm which comes under centroid-based clustering.The suitable K value for the given dataset is selected appropriately which represents the predefined clusters. Raw and unlabeled data is taken as input which is further divided into clusters until the best clusters are found.Centroid based algorithm used in this model is efficient but sensitive to initial conditions and outliers

## 3.2 PROPOSEDSYSTEM

In the proposed system, the customer segmentation model includes not only centroid-based but also hierarchical clustering. • The three clustering algorithms K means, Minibatch K means and the hierarchical algorithm has been selected from the literature survey. • By deploying the three different algorithms, the clusters are formed and analyzed respectively. • The most effective and efficient algorithm is determined by comparing and evaluating the precision rate among the three algorithms

## 3.3 OBJECTIVE OFPROJECT

Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarities among customers in each group. The main objective of segmenting customers is to decide how to relate to customers in each segment to maximize the value of each customer to the business

The emergence of many competitors and entrepreneurs has caused a lot of tension among competing businesses to find new buyers and keep the old ones. As a result of the predecessor, the need for exceptional customer service becomes appropriate regardless of the size of the business.Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service.

## 3.4 SOFTWARE AND HARDWAREREQUIREMENTS

### *3.4.1 SoftwareRequirements:*

- Python
- Anaconda
- Jupyter Notebook

### *3.4.2 HardwareRequirements:*

- Processor: Intel Corei5
- RAM:8GB
- OS:Windows

### *3.4.3 Libraries:*

- **Tkinter**- Tkinter is a library of python used often by everyone. It is a library that is used to create GUI-based applications easily. It contains so many widgets like radio buttons, textfiles, and so on. We have used this for creating an account registration screen, log in or register screen, prediction interface which is a GUI basedapplication

- **Sklearn**- Scikit Learn also known as sklearn is anopen-source library for python programming used for implementing machine learning algorithms. It features various classification, clustering, regression machine learning algorithms. In this, it is used for importing machine learning models, getting accuracy, get a confusion matrix.
- **Pandas-** Library of python which can be used easily. It gives speed results and is also easily understandable. It is a library that can be used without any cost. We have used it for data analysis and to read thedataset.
- **Matplotlib-** Library of python used for visualizing the data using graphs, scatterplots, and so on. Here, we have used it for datavisualization.
- **Numpy-** Library of python used for arrays computation. It has so many functions. We have used this module to change the 2-dimensional array into a contiguous flattened array by using the ravelfunction.

## 3.5 PROGRAMMING LANGUAGES

### *3.5.1 Python*

Python is the best programing language fitted to Machine Learning. In step with studies and surveys, Python is the fifth most significant language yet because the preferred language for machine learning and information science. It's owing to the subsequent strengths that Python has –

- **Easy to be told and perceive-** The syntax of Python is simpler; thence it's comparatively straightforward, even for beginners conjointly, to be told and perceive thelanguage.
- **Multi-purpose language −** Python could be a multi-purpose programing language as a result of it supports structured programming, object-oriented programming yet as practicalprogramming.
- **Support of open supply community −** As being open supply programing language, Python is supported by a giant developer community. Because of

this, the bugs square measure is simply mounted by the Python community. This characteristic makes Python strong andadaptative.

### *3.5.2 Domain*

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches. In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coachonknowledgeinputsanduseappliedmathanalysistooutputvaluesthat fall inside a particular vary. Thanks to this, machine learning facilitates computers in building models from sample knowledge tomodify decision-making processes supported knowledgeinputs.

## 3.6 SYSTEMARCHITECTURE

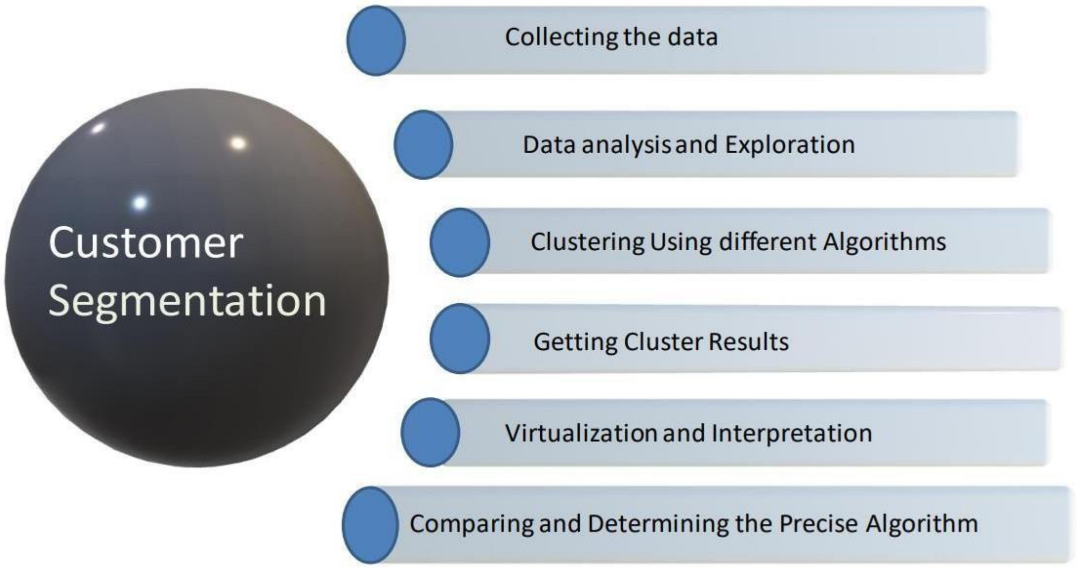

Customer Segmentation
Collecting the data
Data analysis and Exploration
Clustering Using different Algorithms
Getting Cluster Results
Virtualization and Interpretation
Comparing and Determining the Precise Algorithm

**Fig 3.1 System Architecture**

**InFig3.1-**Thedata are usually divided into two types: *Structured* and *Unstructured*. The simplest example of structured data would be a .xls or .csv file where every column stands for an attribute of the data. Unstructured data could be represented by a set of text files, photos, or video files. Often, business dictates how to organize the collection and storage of data.

### *A. Collect data*

This is a data preparation phase. The feature usually helps to refine all data items at a standard rate to improve the performance of clustering algorithms.[12] Each data point varies from grade 2 to +2. Integration techniques that include min-max, decimal, and Z-point are the standard z-signing strategy used to make things uneven before the dataset. While you'll be occupied with analyzing the dataset, you should also start the process of collecting your data in the right shape and format. It could be the same format as in the reference dataset (if that fits your purpose), or if the difference is quite substantial – some other format.

The data are usually divided into two types: *Structured* and *Unstructured*. The simplest example of structured data would be a .xls or .csv file where every column stands for an attribute of the data. Unstructured data could be represented by a set of text files, photos, or video files. Often, business dictates how to organize the collection and storage of data.

### *B. Data Analysis and Exploration*

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, to better understand the nature of the data.

Data exploration, also known as exploratory data analysis (EDA), is a process where users look at and understand their data with statistical and visualization methods. This step helps identify patterns and problems in the dataset, as well as decide which model or algorithm to use in subsequent steps. Although sometimes researchers tend to spend more time on model architecture design and parameter tuning, the importance of data exploration should not be ignored.

### *C. Clustering using different Algorithms*

Considering the knowledge gained from the literature survey, the three most used and efficient algorithms are taken into account for clustering the customers. K means clustering algorithm; Mini batch k means clustering algorithm and Hierarchical clustering algorithm. The three algorithms are all set to be deployed on the dataset respectively.

### *D. Cluster Results*

By deploying the three selected algorithms on the dataset the customer data has been clustered and clusters are formed. Further analyzing the clusters formed by different algorithms the results of the cluster are obtained for three different algorithms which are deployed respectively.Because clustering is unsupervised, no “truth” is available to verify results. The absence of truth complicates assessing quality.

### *E. Comparison and Determination of Precise Algorithm*

Checking the quality of clustering is not a rigorous process because clustering lacks “truth”.Implementing a clustering model with no target to aim, it is not possible to calculate the accuracy score. Henceforth, the aim is to create clusters with distinct or unique characteristics. The two most common metrics to measure the distinctness of clusters are SilhouetteCoefficient and Davies- BouldinIndex.Comparing the metric scores produced by the three algorithms, the most precise algorithm is determined.

## 3.7 ALGORITHMS USED

### *3.7.1 K means Clustering*

K means clustering method is one of the unsupervised partition-based clustering techniques that decomposes the unlabeled dataset into different clusters. The algorithms work by determining the appropriate K value -no of clusters, wherein turn to find the 'K' centroids. Therefore, forms the clusters by assigning each data point to its closest k-center.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

### *3.7.2 Hierarchical Clustering*

Agglomerative hierarchical clustering deviates from partition-based clustering as it builds a binary merge tree with leaves containing data elements and the root that contains the full data set. The graphical representation of that tree that implants the nodes on the plane is called a dendrogram.

The hierarchical clustering technique has two approaches:

1. ***Agglomerative*:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. ***Divisive:*** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

The closest distance between the two clusters is crucial for hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called Linkage methods.

***Single Linkage:*** It is the Shortest Distance between the closest points of the clusters.

***Complete Linkage*:** It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.

***Average Linkage*:** It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

***Centroid Linkage*:** It is the linkage method in which the distance between the centroid of the clusters is calculated.

### *3.7.3 Minibatch K means clustering*

There is no doubt that k-means is one of the most popular clustering algorithms because of its performance and low cost of time but with an increase in the size of the datasets being taken into consideration for analysis the computation time of k-means increases. To overcome this, a different approach is introduced called the Minibatch k-means algorithm whose main idea is to divide the whole dataset into small- fixed-size batches of data and use a new random mini batch from the dataset and update the clusters where this iteration is repeated till theconvergence.

Mini Batch K-means algorithm‘s main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. Each mini-batch updates the clusters using a convex combination of the values of the prototypes and the data, applying a learning rate that decreases with the number of iterations. This learning rate is the inverse of the number of data assigned to a cluster duringthe process. As the number of iterations increases, the effect of new data is reduced, so convergence can be detected when no changes in the clusters occur in several consecutive iterations.

### *3.7.4 Elbow Method*

Determining the optimal no of clusters for the given dataset is the most fundamental step for any unsupervised algorithm. The Elbow method helps us to determine the best value of k. the k value is selected where the point starts to flatten out forming an elbow in the graph plotted using the sum of squared distance between the data points and their respective assigned cluster centroids. Therefore, the optimal number of clusters is determined.

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

WCSS= $\sum_{P_i \text{ in Cluster1}} \text{distance}(P_i\, C_1)^2 + \sum_{P_i \text{ in Cluster2}} \text{distance}(P_i\, C_2)^2 + \sum_{P_i \text{ in CLuster3}} \text{distance}(P_i\, C_3)^2$

### 3.8 MODULES

The project contains three parts:

- **Dataset Collection**- We had collected datasets from Kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 200 rows and 5columns.
- **Train and test the model**- We had used three clustering algorithms named K-means clustering algorithm, Hierarchical clustering, and mini-batch K-means algorithm to train the dataset. After training, we had tested the model and found their clusters, silhouette score, and DaviesBoulding score.
- **Deploy the models**- Deployed the model to get the clusters formed.The cluster shows the different segmentation of customers based on many
- attributes. By this, we will get the silhouette score and DaviesBoulding scores of the model as the output.

**Following are the steps to do this project (use Jupyter Notebook):**

- Collect thedataset.
- Import the necessarylibraries.
- Visualize thedataset.
- Train the dataset using K-means clustering algorithm, Hierarchical clustering, and mini-batch K-means algorithms.
- Test the model and find the clusters and silhouette score and DaviesBoulding score.
- Deploy the model
- Based on the scorespredict which algorithm is best for customer segmentation and go ahead with that clustering algorithm.

# CHAPTER 4

## RESULTS AND DISCUSSION

### 4.1. PERFORMANCE ANALYSIS

Unlike supervised algorithms such as a linear regression model, there is a target to predict where the accuracy can be measured by using metrics such as RMSE, MAPE, MAE, etc., Implementing a clustering model with no target to aim, it is not possible to calculate the accuracy score. Henceforth, the aim is to create clusters with distinct or unique characteristics. The two most common metrics to measure the distinctness of clusters are:

**SilhouetteCoefficient:**

The silhouette score is a measure of the average similarity of the objects within a cluster and their distance to the other objects in the other clusters.

For each data point I, we first define:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Secondly, we define:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

Finally, we define the silhouette score of a data point I as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

This score ranges between -1 and 1, where the clusters are more well-defined and distinct with higherscores.

**Davies-BouldinIndex:**

The Davies-Bouldin (DB) criterion is based on a ratio between “within- cluster” **and** “between-cluster” distances.

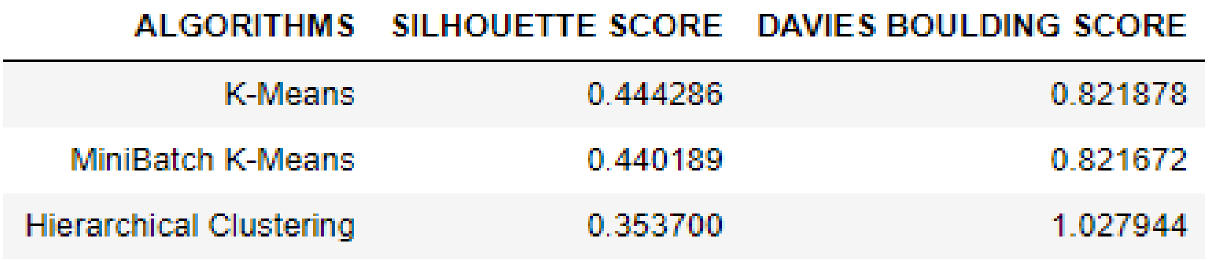
$$\mathrm{DB}(\mathcal{C}) = \frac{1}{k}\sum_{i=1}^{k} \max_{j\leq k, j\neq i} D_{ij}, \;\; k = |\mathcal{C}|,$$

Dij is the "within-to-between cluster distance ratio" for the ith and jth clusters.

$$D_{ij} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{ij}},$$

where d¯ i is the average distance between every data point in cluster I and its centroid, similar for d¯ j. dij is the Euclidean distance between the centroids of the two clusters.

On contrary to the Silhouette score, this score measures the similarity among the clusters which defines that the lower the score the better clusters are formed.



| ALGORITHMS | SILHOUETTE SCORE | DAVIES BOULDING SCORE |
|---:|---:|---:|
| K-Means | 0.444286 | 0.821878 |
| MiniBatch K-Means | 0.440189 | 0.821672 |
| Hierarchical Clustering | 0.353700 | 1.027944 |

**Table 4.1 Performance Score Comparison**

# CHAPTER 5

## CONCLUSION :

The significance of customer segmentation in attracting the customers towards the products which in turn aids the increase in the business scale in the market. Segmenting the customer group into the different groups according to the similarities they possess, on one hand, helps the marketers to provide customized ads, products, and offers. where on other hand it supports the customers by avoiding them from the confusion of the products to buy. It couldn't be said that the K means is the most effective clustering algorithm every time. It depends on the various factors such as the size of the data, attributes of the data,etc.,

Comparing the clusters obtained by deploying the three different clustering algorithms on the customers' data using the metrics that measure the distinctness and uniqueness of the clusters. It is observed that the K means algorithm produces the best clusters by obtaining the highest Silhouette score and the least Davies Bouldin score followed by hierarchical clustering and minibatch k means clustering.

## FUTURE ENHANCEMENTS :

This Project can further be enhanced by including different clustering algorithms that may depict more proficiency and by considering the large datasets which in turn increases the efficiency. In near future we implement the proposed data model using visual studio IDE and provide the performance and merits and limitation of our proposed work.

# REFERENCES

[1] Customer Segmentation Using Machine Learning, Prof. Ilavedhan A. Aman Banduni, School of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.

[2] Kamalpreet Bindra and Anuranjan Mishra, A Detailed Study of Clustering Algorithms, CSE Department, Noida International University, 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), In September 2017, AIIT, Amity University Uttar Pradesh, Noida, India.

[3] KAI PENG(Member, IEEE), VICTOR C. M. LEUNG, (Fellow, IEEE), AND QINGJIA HUANG, "Clustering Approach based on Mini batch Kmeans ", In 2018, College of Engineering, Huaqiao University, Quanzhou 362021, China.

[4] Fionn Murtagh and Pedro Contreras, "Methods of Hierarchical Clustering", In 2018, Science Foundation Ireland, Wilton Place, Dublin, Ireland Department of Computer Science, Royal Holloway, University of London.

[5]D.P. Yash Kushwaha, Deepak Prajapati, “CustomerSegmentation using K-Means Algorithm,” 8th Semester.

[6] Manju Kaushik, Bhawana Mathur, “Comparative Study of K-Means and Hierarchical Clustering Techniques”, In June 2014, JECRC University, Jaipur.

[7] Ali Feizollah, Nor Badrul Anuar, Rosli Salleh, and Fairuz Amalina, "Comparative Study of K-means and Mini Batch K-means Clustering Algorithms", Computer System and Technology Department, The University

of Malaya, Kuala Lumpur, Malaysia, International Journal of Software and Hardware Research in Engineering.

[8] Asith Ishantha, "Mall Customer Segmentation using Clustering Algorithm", Future University Hakodate, Conference Paper, March 2021.

[9] Onur Dogan, Dokuz eylul University, Ejder Aycin, Kocaeli University, Zeki Atil Bulut, Dokuz Eylul University, "Customer Segmentation by using RFM model and clustering methods: A case Study in Retail Industry", In July 2018, International Journal of Contemporar Economics and Administrative Sciences.

[10] Juni Nurma Sari,Ridi Ferdiana,Lukito Nugroho,Paulus Insap Santosa,"Review on Custom- er Segmentation Technique",Department of Electrical

Engineering and Information Tech- nology, University of Gadjah Mada, Jogjakarta, Indonesia, Department of Informatics Technology, Polytechnic Caltex Riau, Pekanbaru, Indonesia.

[11] I. S. Dhillon and D. M. Modha, "Concept decompositions for huge sparse textual content records using clustering," Machine Learning, vol. 42, trouble 1, pp. 143-175, 2001.

[12] Jayant Tikmani, Sudhanshu Tiwari and Sujata Khedkar, "Mall customer segmentation based on cluster analysis An Approach to Customer Classification using k-means", IJIRCCE, 2015.

[13] Puwanenthiren Premkanth, ―Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research

Editeur: Global Journals Inc. (USA). 2012. Stampa ISSN: 09755853. Volume 12 Edizione1.

[14] nedetti, F., Beneventano, D., Bergamaschi, S., Simonini, G. Aspects of Information Systems 80, 136–147 (2019)

# APPENDICES

## A. SOURCE CODE

```
import NumPy as np
import pandas as PD
import matplotlib.pyplotasplt
import seaborn as sns

df=pd.read_csv("Mall_Customers.csv")
df.head()
df.shape
df.describe()
df.dtypes
df.isnull().sum()
df.drop(["CustomerID"],axis=1, inplace=True)
df.head()

plt.figure(1,figsize=(15,6))
n=0
for x in ['Age','No. of Purchases','Spending Score (1-100)']:
   n += 1
plt.subplot(1,3,n)
plt.subplots_adjust(hspace = 0.5,wspace = 0.5)
sns.distplot(df[x],bins = 20)
plt.title('Distplot of {}'.format(x))
plt.show()
plt.figure(figsize=(15,5))
sns.countplot(y='Gender',data=df)
plt.show()
plt.figure(1,figsize=(15,7))
n=0
for cols in ['Age','No. of Purchases','Spending Score (1-100)']:
   n+=1
plt.subplot(1 , 3 , n)
sns.set(style="whitegrid")
plt.subplots_adjust(hspace = 0.5, wspace = 0.5)
sns.violinplot(x = cols, y = 'Gender', data = df)
plt.ylabel('Gender' if n == 1 else '')
plt.title('Violin Plot')
plt.show()
age_18_25 = df.Age[(df.Age>= 18) & (df.Age<= 25)]
age_26_35 = df.Age[(df.Age>= 26) & (df.Age<= 35)]
age_36_45 = df.Age[(df.Age>= 36) & (df.Age<= 45)]
age_46_55 = df.Age[(df.Age>= 46) & (df.Age<= 55)]
age_55above = df.Age[df.Age>= 56]

agex = ["18-25","26-35","36-45","46-55","55+"]
agey =
[len(age_18_25.values),len(age_26_35.values),len(age_36_45.values),len(age_46_55.val
ues),len(age_55above.values)]

plt.figure(figsize=(15,6))
sns.barplot(x = agex, y = agey, palette ="mako")
plt.title("Number of Customer and Ages")
plt.xlabel("Age")
plt.ylabel("Number of Customer")
plt.show()
sns.relplot(x="No. of Purchases", y="Spending Score (1-100)", data = df)
ss_1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) &
(df["Spending Score (1-100)"] <= 20)]
ss_21_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) &
(df["Spending Score (1-100)"] <= 40)]
ss_41_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) &
(df["Spending Score (1-100)"] <= 60)]
ss_61_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) &
(df["Spending Score (1-100)"] <= 80)]
ss_81_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) &
(df["Spending Score (1-100)"] <= 100)]

ssx = ["1-20", "21-40", "41-60", "61-80", "81-100"]
ssy = [len(ss_1_20.values), len(ss_21_40.values), len(ss_41_60.values),
len(ss_61_80.values), len(ss_81_100.values)]

plt.figure(figsize=(15,6))
sns.barplot(x=ssx, y=ssy, palette="rocket")
plt.title("Spending Scores")
plt.xlabel("Score")
plt.ylabel("Number of Customer Having the Score")
plt.show()


ai0_30 = df["No. of Purchases"][(df["No. of Purchases"] >= 0) & (df["No. of Purchases"] <=
30)]
ai31_60 = df["No. of Purchases"][(df["No. of Purchases"] >= 31) & (df["No. of Purchases"]
<= 60)]
ai61_90 = df["No. of Purchases"][(df["No. of Purchases"] >= 61) & (df["No. of Purchases"]
<= 90)]
ai91_120 = df["No. of Purchases"][(df["No. of Purchases"] >= 91) & (df["No. of Purchases"]
<= 120)]
ai121_150 = df["No. of Purchases"][(df["No. of Purchases"] >= 121) & (df["No. of
Purchases"] <= 150)]

aix = ["$ 0 - 30,000","$ 30,001 - 60,000", "$ 60,000 - 90,000", "$ 90,001 - 120,000",
"120,001 - 150,000"]
aiy = [len(ai0_30.values), len(ai31_60.values), len(ai61_90.values), len(ai91_120.values),
len(ai121_150.values)]
```

```
plt.figure(figsize=(15,6))
sns.barplot(x=aix, y=aiy, palette="Spectral")
plt.title("No. of Purchases")
plt.xlabel("Income")
plt.ylabel("Number of Customer")
plt.show()

X1=df.loc[:, ["Age","Spending Score (1-100)"]].values

from sklearn.cluster import KMeans
wcss = []
fork in range(1,11):
kmeans = KMeans(n_clusters=k, init="k-means++")
kmeans.fit(X1)
wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
kmeans = KMeans(n_clusters=4)

label = kmeans.fit_predict(X1)

print(label)
print(kmeans.cluster_centers_)
plt.scatter(X1[:,0],X1[:,1], c=kmeans.labels_, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1], color='black')
plt.title('Clusters of Customers')
plt.xlabel('Age')
plt.ylabel('Spending Score(1-100)')
plt.show()
X2=df.loc[:, ["No. of Purchases","Spending Score (1-100)"]].values

from sklearn.cluster import KMeans
```

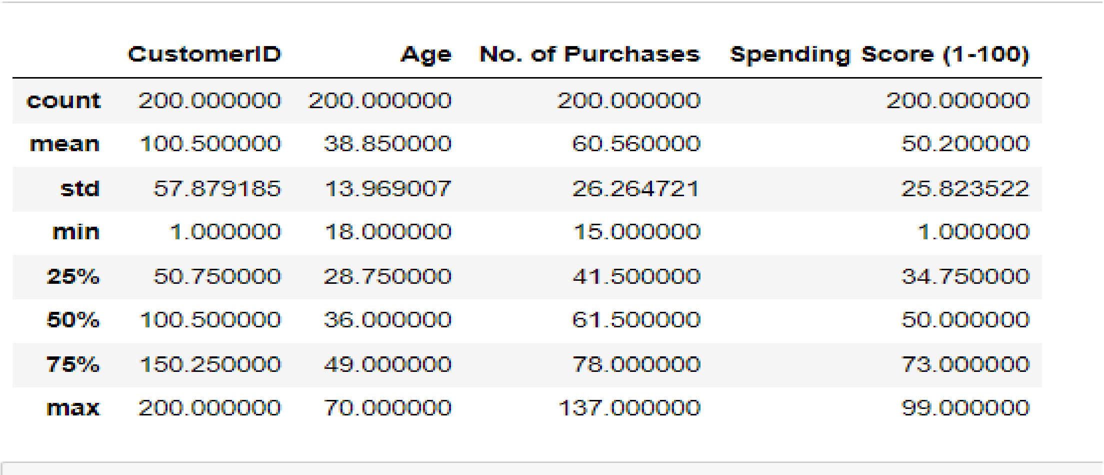
```
wcss = []
for k in range(1,11):
kmeans = KMeans(n_clusters=k,init="k-means++")
kmeans.fit(X2)
wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker = "8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
kmeans = KMeans(n_clusters=5)

label = kmeans.fit_predict(X2)

print(label)
print(kmeans.cluster_centers_)
plt.scatter(X2[:,0], X1[:,1], c=kmeans.labels_, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0] ,kmeans.cluster_centers_[:,1], color='black')
plt.title('Clusters of Customers')
plt.xlabel('No. of Purchases')
plt.ylabel('Spending Score(1-100)')
plt.show()
X3=df.iloc[:,1:]

wcss = []
for k in range(1,11):
kmeans = KMeans(n_clusters=k, init="k-means++")
kmeans.fit(X3)
wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
kmeans = KMeans(n_clusters = 5)
```

```
label = kmeans.fit_predict(X3)

print(label)
print(kmeans.cluster_centers_)
clusters = kmeans.fit_predict(X3)
df["label"] = clusters

from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.Age[df.label == 0], df["No. of Purchases"][df.label == 0], df["Spending Score (1-100)"][df.label == 0], c='blue',s=60)
ax.scatter(df.Age[df.label == 1], df["No. of Purchases"][df.label == 1], df["Spending Score (1-100)"][df.label == 1], c='red',s=60)
ax.scatter(df.Age[df.label == 2], df["No. of Purchases"][df.label == 2], df["Spending Score (1-100)"][df.label == 2], c='green',s=60)
ax.scatter(df.Age[df.label == 3], df["No. of Purchases"][df.label == 3], df["Spending Score (1-100)"][df.label == 3], c='orange',s=60)
ax.scatter(df.Age[df.label == 4], df["No. of Purchases"][df.label == 4], df["Spending Score (1-100)"][df.label == 4], c='purple',s=60)
ax.view_init(30, 185)

plt.xlabel("Age")
plt.ylabel("No. of Purchases")
ax.set_zlabel('Spendiing Score (1-100)')

plt.show()
from sklearn.cluster import MiniBatchKMeans
from sklearn import metrics
minikm = MiniBatchKMeans(n_clusters=5,init='random',batch_size=100000)
minikm_labels = minikm.fit_predict(X3)
print(minikm_labels)
mini_cluster = minikm.fit_predict(X3)
df["minikm_labels"] = mini_cluster
```

```
from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.Age[df.label == 0], df["No. of Purchases"][df.label == 0], df["Spending Score
(1-100)"][df.label == 0], c='blue',s=60)
ax.scatter(df.Age[df.label == 1], df["No. of Purchases"][df.label == 1], df["Spending Score
(1-100)"][df.label == 1], c='red',s=60)
ax.scatter(df.Age[df.label == 2], df["No. of Purchases"][df.label == 2], df["Spending Score
(1-100)"][df.label == 2], c='green',s=60)
ax.scatter(df.Age[df.label == 3], df["No. of Purchases"][df.label == 3], df["Spending Score
(1-100)"][df.label == 3], c='orange',s=60)
ax.scatter(df.Age[df.label == 4], df["No. of Purchases"][df.label == 4], df["Spending Score
(1-100)"][df.label == 4], c='purple',s=60)
ax.view_init(30, 185)

plt.xlabel("Age")
plt.ylabel("No. of Purchases")
ax.set_zlabel('Spendiing Score (1-100)')

plt.show()
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage
agglo_clustering
=AgglomerativeClustering(n_clusters=5,affinity='euclidean',linkage='ward')
agglo_clustering_labels = agglo_clustering.fit_predict(X3)
agglo_clusters = agglo_clustering.fit_predict(X3)
df["agglo_clustering_labels"] = agglo_clusters

from mpl_toolkits.mplot3d import Axes3D
Z = linkage(X3,method = 'ward')
dendro = dendrogram(Z)
plt.title('Dendrogram')
plt.ylabel('Euclidean disance')
plt.show
algorithms = ["K-Means","Hierarchical Clustering","MiniBatch K-Means"]
#Silhoutte Score
```

```
ss =
[metrics.silhouette_score(X3,label),metrics.silhouette_score(X3,minikm_labels),metrics.sil
houette_score(X3,agglo_clustering_labels)]
#Davies Bouldin Score
db =
[metrics.davies_bouldin_score(X3,label),metrics.davies_bouldin_score(X3,minikm_labels)
,metrics.davies_bouldin_score(X3,agglo_clustering_labels)]
comparision = {"ALGORITHMS":algorithms,"SILHOUETTESCORE":ss,"DAVIES
BOULDING SCORE":db}
compdf = pd.DataFrame(comparision)
display(compdf.sort_values(by=["SILHOUETTE SCORE"],ascending = False))
```

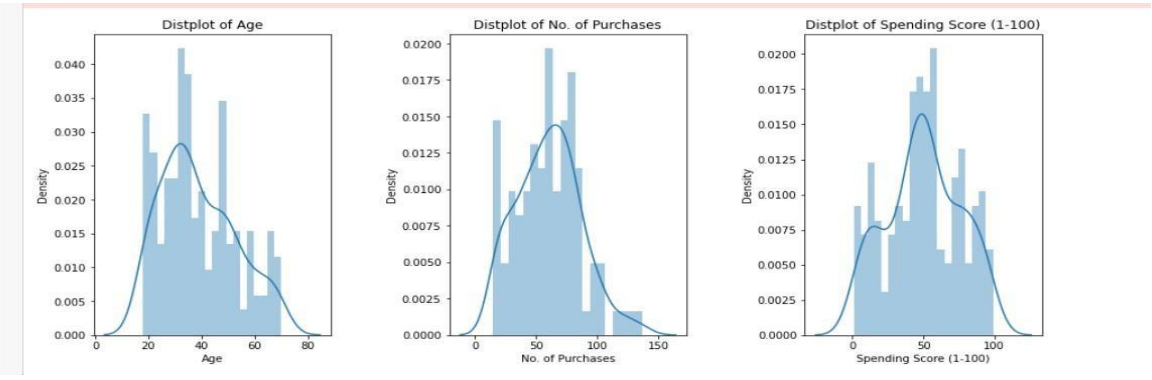
## B. SCREENSHOTS



| | CustomerID | Age | No. of Purchases | Spending Score (1-100) |
|---|---|---|---|---|
| **count** | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| **mean** | 100.500000 | 38.850000 | 60.560000 | 50.200000 |
| **std** | 57.879185 | 13.969007 | 26.264721 | 25.823522 |
| **min** | 1.000000 | 18.000000 | 15.000000 | 1.000000 |
| **25%** | 50.750000 | 28.750000 | 41.500000 | 34.750000 |
| **50%** | 100.500000 | 36.000000 | 61.500000 | 50.000000 |
| **75%** | 150.250000 | 49.000000 | 78.000000 | 73.000000 |
| **max** | 200.000000 | 70.000000 | 137.000000 | 99.000000 |

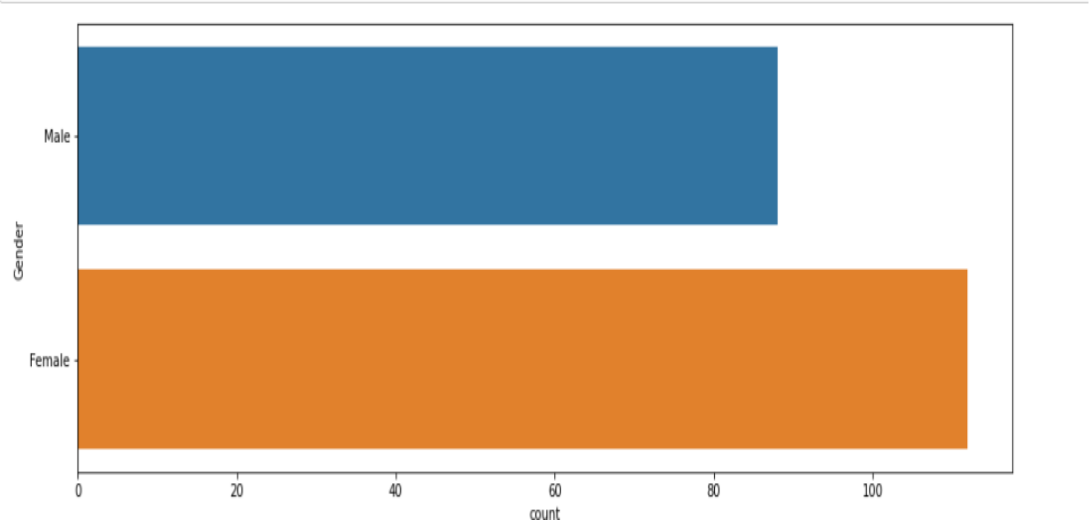
B.1. Dataset

In Fig B.1- Age of the customers ranges from 18-70. This shows that the mall attracts has shops and things which suite all age group people.
Average age of customers is 39.
Average income of customers is 60 K$.
Average spending score of customers is 50.



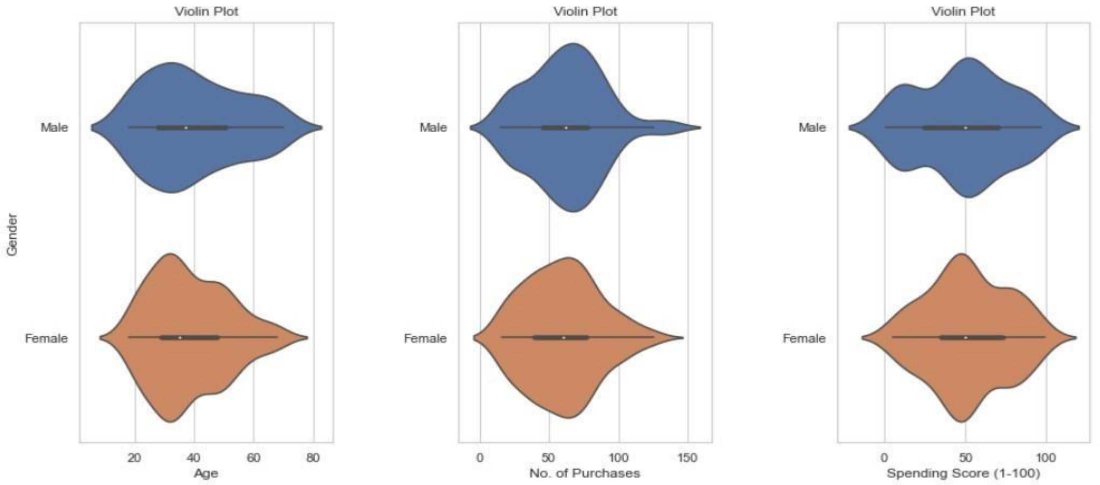
B.2. Distplot of Features

In Fig B.2- The first satge of Data exploration and also known as data analysis,is the process of looking at and visualizing data in order to quickly again insights,locate specific reas or patterns or both for the purpose of doing more extensive study.



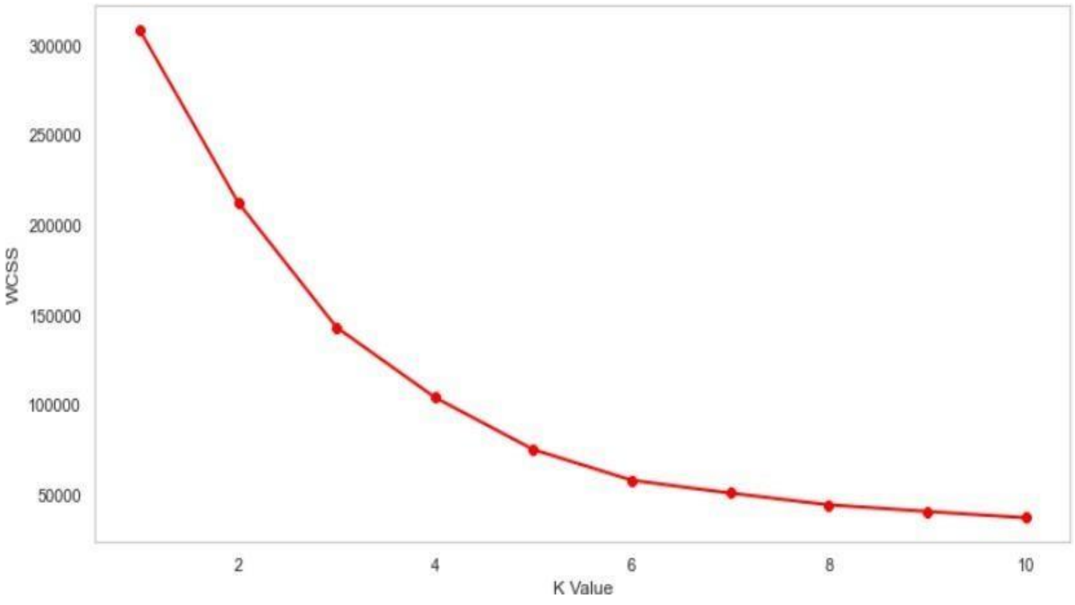
B.3. Plotting of Gender

In Fig B.3- The data set including records of consumer information has been taken into account by kaggle. More is preferable in this instance science it helps us discover more patterns and trends with in the information.



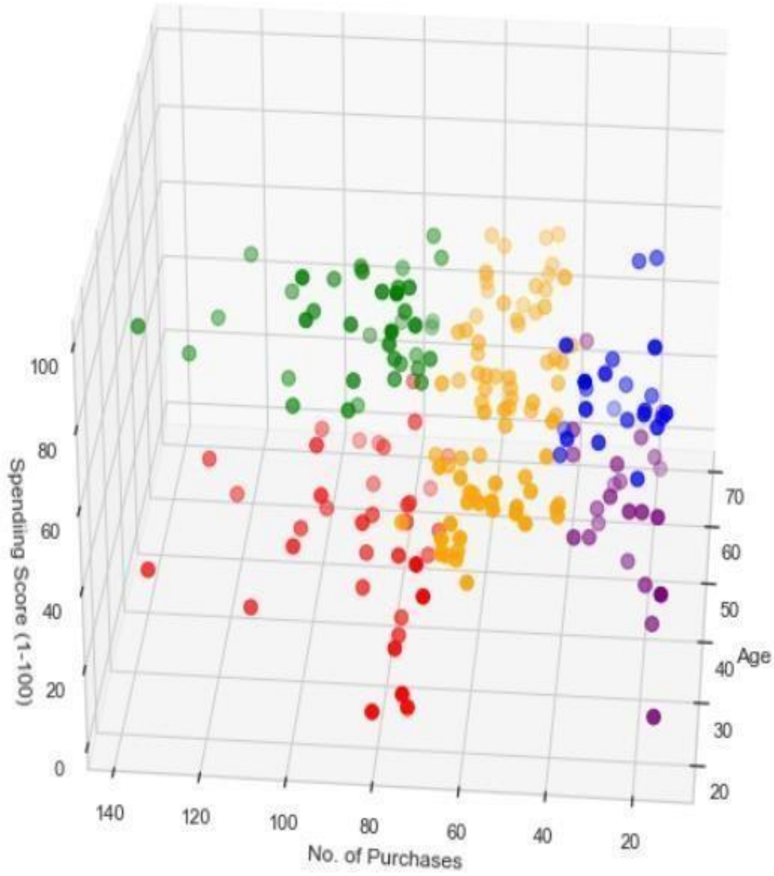
B.4. Violin plot for features

In Fig B.4- This is the most crucial stage that will enables us to discover tha data and patterns, We may have a better understanding of the clients interests, preferences. And purchasing habits through this . This will help identify the characteristics that are most closely associated to clients and the company as a whole.



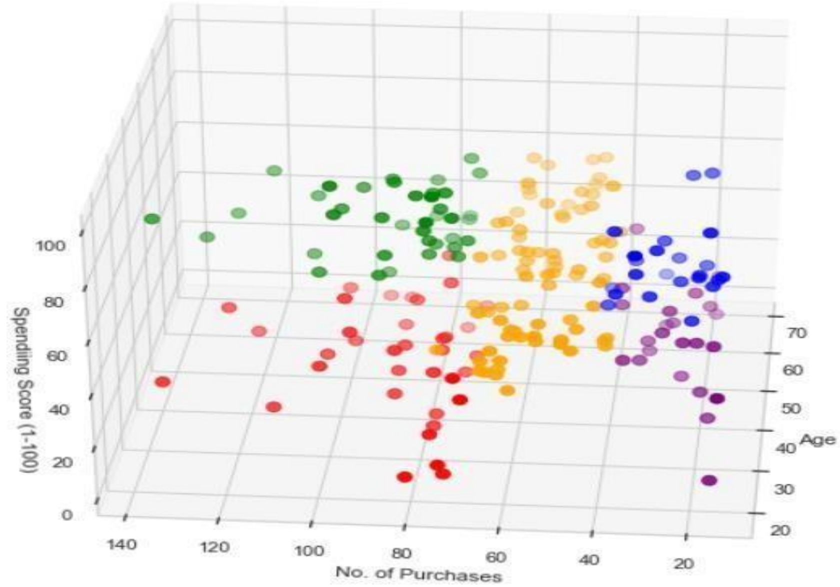
B.5. Elbow method for determining No. of Clusters

In Fig B.5- The number of clusters in this study should be K=5,or five,according to the graph above.



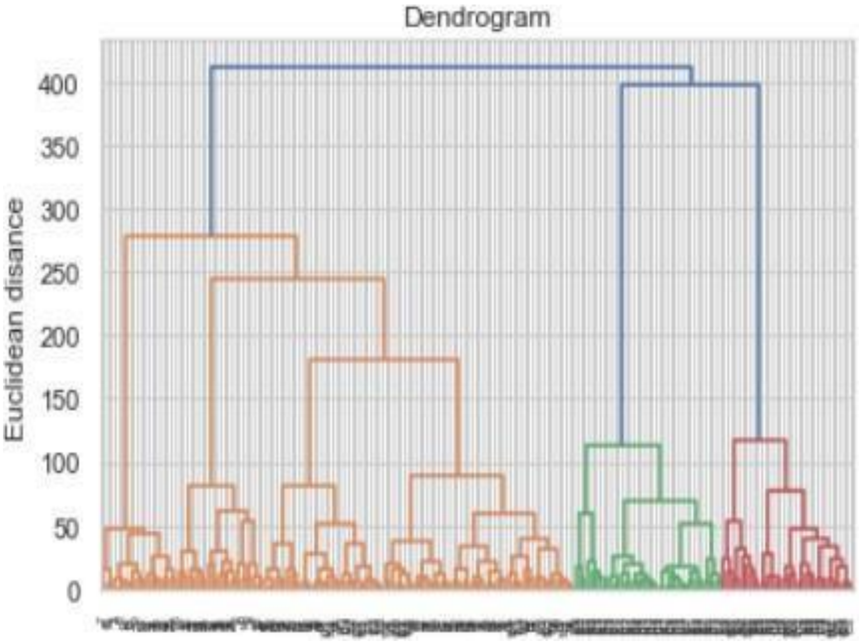
B.6. 3D plot for K means clustering

In Fig B.6 - we will be using Iris dataset as we did in the previous one. We already know that Iris dataset contains 3 different types of flowers and 4 features for each flower. But we will be using only 3 features for this tutorial since we can't visualize a 4 dimensional space. Therefore, it is a smart idea to choose 3 random cluster centers.



B.7. 3D plot for Mini batch K means Clustering

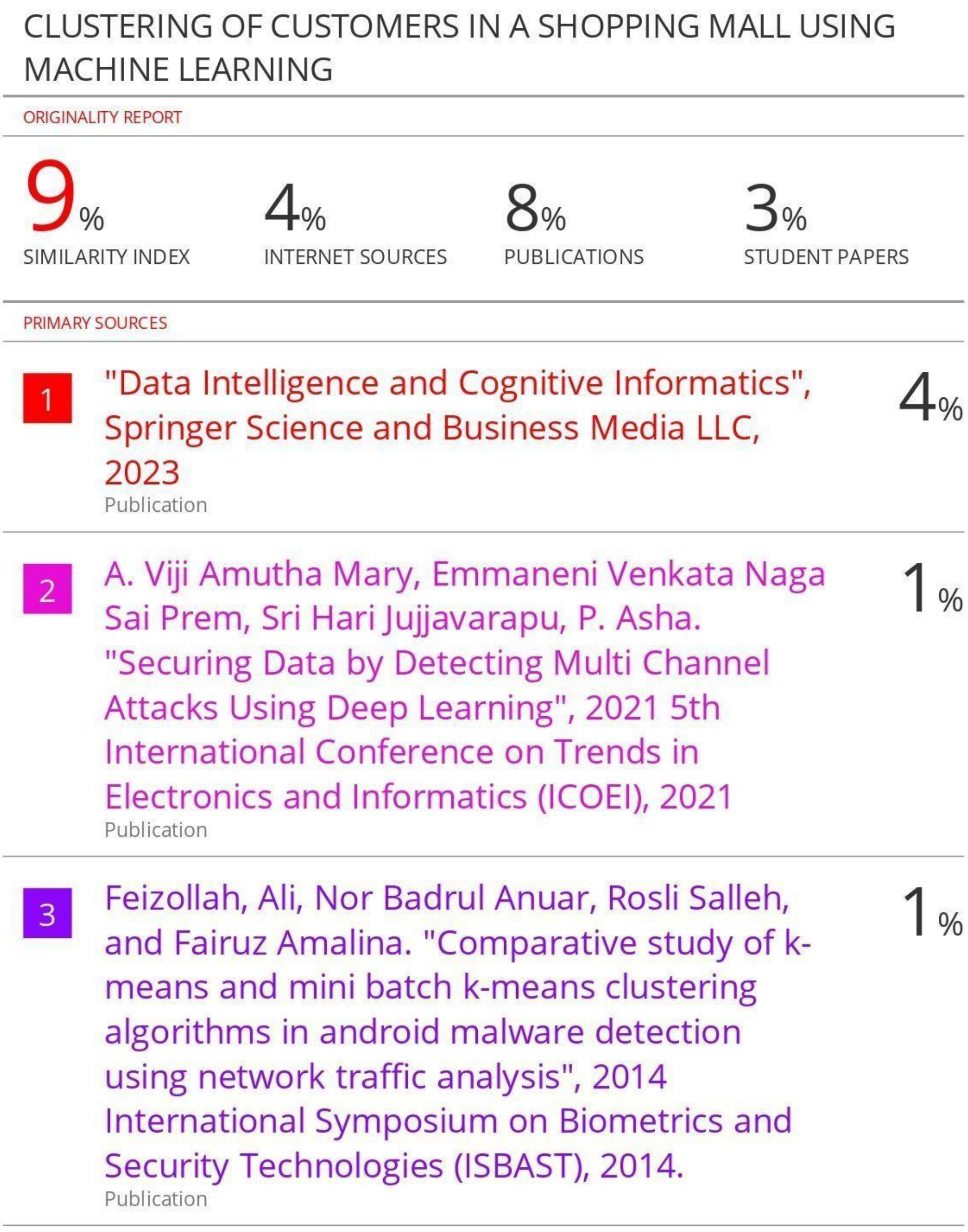
In Fig B.7- Mini Batch K-means algorithm's main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence.



B.8. Dendrogram for Hierarchical Clustering

In Fig B.8- A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering.

## C.PLAGIARISM REPORT



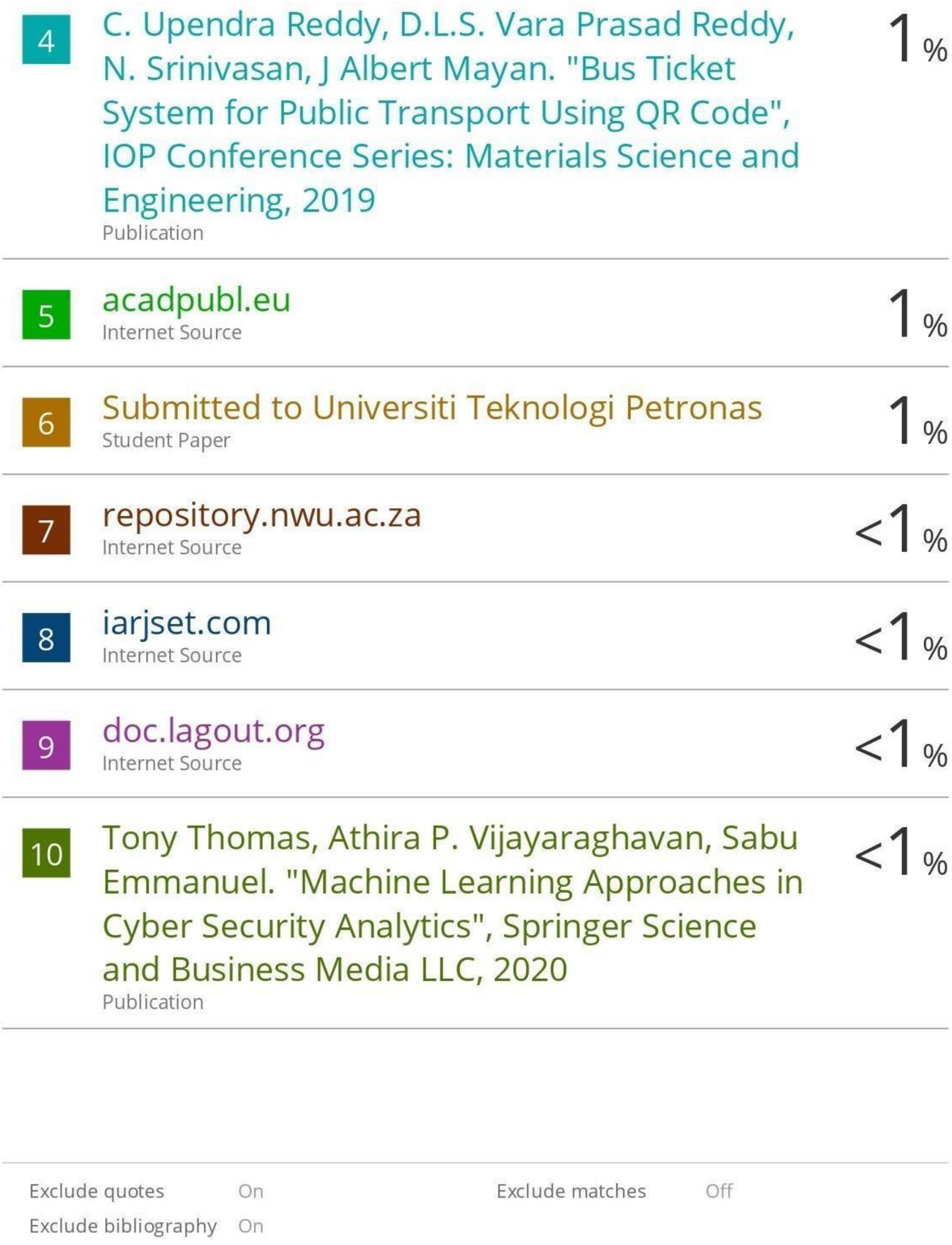
CLUSTERING OF CUSTOMERS IN A SHOPPING MALL USING MACHINE LEARNING

ORIGINALITY REPORT

| 9% | 4% | 8% | 3% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | "Data Intelligence and Cognitive Informatics", Springer Science and Business Media LLC, 2023<br>Publication | 4% |
| 2 | A. Viji Amutha Mary, Emmaneni Venkata Naga Sai Prem, Sri Hari Jujjavarapu, P. Asha. "Securing Data by Detecting Multi Channel Attacks Using Deep Learning", 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021<br>Publication | 1% |
| 3 | Feizollah, Ali, Nor Badrul Anuar, Rosli Salleh, and Fairuz Amalina. "Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis", 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 2014.<br>Publication | 1% |



| | | |
|---|---|---|
| 4 | C. Upendra Reddy, D.L.S. Vara Prasad Reddy, N. Srinivasan, J Albert Mayan. "Bus Ticket System for Public Transport Using QR Code", IOP Conference Series: Materials Science and Engineering, 2019<br>Publication | 1% |
| 5 | acadpubl.eu<br>Internet Source | 1% |
| 6 | Submitted to Universiti Teknologi Petronas<br>Student Paper | 1% |
| 7 | repository.nwu.ac.za<br>Internet Source | <1% |
| 8 | iarjset.com<br>Internet Source | <1% |
| 9 | doc.lagout.org<br>Internet Source | <1% |
| 10 | Tony Thomas, Athira P. Vijayaraghavan, Sabu Emmanuel. "Machine Learning Approaches in Cyber Security Analytics", Springer Science and Business Media LLC, 2020<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |