DEMOGRAPHIC DIVISION OF A MART BY APPLYING CLUSTERING TECHNIQUES

Mrs.J.Mounika[1], I. Monali[2], G. Lakshmi Pooja[3], G. Meghana[4], P. Mary Pushpa[5]

[1]Assistent Professor Department of Information Technology, KKR & KSR Institute Of Technology And Sciences (A), Guntur, India

[2,3,4,5] Undergraduate Students , Department of Information Technology , KKR & KSR Institute Of Technology And Sciences (A), Guntur, India

---***---

Abstract – *Everyone in this contemporary era strives to be more creative and competitive than others. Therefore, to survive in this cutthroat world, we must be superior to others. When we consider business scenarios in the modern era, they depend on innovations that can captivate customers with their offerings. Because we are not employing good techniques, our business will be monotonous and we will incur losses. Therefore, we must employ advanced techniques that can produce results quickly and precisely if we want to have a successful and market-leading business in this era. Targeting customer wish lists and concentrating on customer sales are two strategies many businesses use. As numerous techniques and algorithms were available at the time, machine learning enters the picture. Furthermore, future decision-making and the discovery of buried patterns in the data both use this.*

The first section to target is specified in this concept, and the second segment is equalized using segmentation. Customer segmentation is the division of a customer base into groups and subgroups according to similar needs and behaviors, and then further into individual segments. To segment, the audience for this paper, three different clustering algorithms—K-Means, Agglomerative, and Hierarchical Clustering—were implemented. The results of the clusters produced by these algorithms were then compared. A Python program has been created and trained by applying standard scalar to a dataset with about 2000 training samples obtained from a nearby mart. However, there were a lot of null values in the dataset that we collected. Therefore, we used the data-cleaning process to remove those null values. The dataset was then used for our testing after we removed all the *invalid data and ensured that there were no missing values. Data preprocessing, data cleaning, data transformation, and data evolution are typically the next few steps. It displays the number of men and women arriving at the market and the type of work they are all performing. Each group will then be divided into different clusters for each segment. Finally, we'll have a variety of clusters. Our outcome will be that.*

Key Words**: Customers, Marts, Segmentation, Machine learning, Python, K-Means, Agglomerative algorithm, Clustering Techniques, Data Cleaning, Data Transformation, Preprocessing, Missing Values, Behaviors, Standard Scalar, Dataset, Decision Making.**

1. INTRODUCTION

The need for more marketing strategies has increased for established companies as new businesses continue to open up because the market is becoming increasingly cutthroat. In the world of marketing today, a straightforward rule has emerged: "Change or Die." It clearly states that they must alter their current strategies if they want to stay competitive in the market as better methods and tactics emerge. In the marketing world, this rule was observed. As the customer base grows day by day, it has become difficult for businesses to meet every customer's needs and requirements. At this point, data mining is crucial for revealing the hidden patterns kept in the company's database.

Customer segmentation is one of the applications of data mining that assists in clustering customers who exhibit similar patterns into smaller groups, making it simpler for the company to manage its large customer base. This segmentation can have a direct or indirect impact on the marketing strategy because it opens up some new avenues for research, including which segment the product will be best suited for, tailoring marketing strategies to each segment, offering discounts to certain segments, and understanding the customer-object relationship, which the company had not previously understood. Customer segmentation gives businesses the ability to see what their customers are buying, which motivates them to better serve them and increase customer satisfaction. It also enables companies to identify their target customers and improve their marketing strategies to increase sales from them.

Clustering is an effective method for implementing customer segmentation. Unsupervised learning includes clustering, which is the ability to find clusters in unlabeled datasets. Clustering algorithms include k-means, hierarchical clustering, Agglomerative clustering, and others. Three different clustering algorithms were tested on a dataset with two features and 2000 records in this paper.

2. LITERATURE REVIEW

Aman Banduni[1], The process of customer segmentation using machine learning is explained. They use machine learning to categorize the customers in this by performing a few steps. To perform tasks for any type of segmentation, we first need the dataset. Customer classification, Big Data, data repository, clustering data, and k-means are the steps that came after. Over time, the commercial world has become more competitive as companies like these now need to entice new customers while also meeting the needs and demands of their existing ones. Determining and meeting each customer's needs and requirements is a very difficult task. This is a result of the fact that various clients have various needs, desires, demographics, sizes, preferences, and other traits. Currently, providing equal service to all customers is not a wise business decision. Because of this, businesses engage in customer segmentation. Millions of internally connected sensors send data about their customers, suppliers, and business processes to the outside world. world of technology, including automobiles and smartphones, as well as information gathered from production, sensing, and communications. the ability to improve forecasts, save resources, increase output, and improve several areas, such as traffic management, weather forecasting, disaster prevention, finance, fraud control, business transactions, national security, education, and healthcare. Every data collection effort is made to obtain trustworthy data that will aid in the analysis and production of truthful but inaccurate answers to the questions posed. Through the process of clustering based on commonalities, data is organized into datasets. To analyze datasets based on the given condition, a variety of methods can be used. There are numerous methods for performing k-means and grouping data into groups. The data in this paper comes from the UCI machine Learning repository. In this paper, they cluster the customers by indicating in different colors such as red, orange, and blue which product will be completely sold in a short time

Zhenyu Wang, Yi Zuo atal[2], This paper primarily discusses customer segmentation using a broad learning system. Since the introduction of the first POS (Point of Sale) system in supermarkets in the 1970s, POS data has been regarded as a critical strategic resource. Retailers hope to improve their understanding of consumer purchasing patterns and increase customer loyalty by analyzing POS data. Numerous researchers have discovered through the analysis of POS data that consumers are more committed to a specific product after making multiple purchases of it. As a result, we typically divide consumer loyalty into three categories based on previous purchases: high, middle, and low (POS data). The frequency with which a customer purchases and consumes a specific brand of goods is referred to as the "high level." Pos data cannot explain the purchasing decision-making process of consumers. To address this issue wireless, non-contact RFID technology has emerged as a viable option. In 2011, et al.[1] attached a tiny RFID tag to shopping carts to track customers' in-store activities based on where the carts were placed. Several studies have found that the amount of time customers spend on each component has a positive impact on their consumption behavior. Researchers using RFID data to reflect various consumer behaviors frequently divide the length of time that customers spend in a given interval into three segments (high level, middle level, and low level). To understand the consumers' purchase decision-making process, we divide them into various homogeneous segments using POS and RFID data. Several classification methods are used. SVM models and Neural Networks are the most commonly used segmentation techniques. Both POS and RFID data have a positive impact on customer segmentation; however, only one type of data can fully comprehend a customer's purchasing activity; for example, POS data can only fully comprehend a client's purchase outcomes, whereas RFID data can fully comprehend a customer's shopping behavior. They configured the Neural Network to be a three-layer BP neural network. The output layer has two nodes, while the hidden layer in the middle has ten nodes, and the output layer has two nodes. The features of SVM are limited. In this paper, they used the BLS to identify customer segmentation and compare it to other classification models.

Su-li HAO[3] This paper primarily discusses the segmentation of commercial bank customers using unascertained clustering. Client relationship management and value management are now the primary driving forces behind commercial bank development. Numerous reputable commercial banks worldwide, such as Citibank, Bank of America, HSBC, JP Morgan, and others, practice and benefit from customer relationship management. Commercial banks will profit if they value their customers and work to maintain their high standards. The commercial bank's marketing decisions are based on an assessment of the client's lifetime value. Some of them use customer current value as the measurement standard to evaluate customer value in the current evaluation method, while others use customer potential value as the measuring standard. Others consider the measuring standard to be the simple sum of the current and potential values. The new method must determine the value of the customer currency, non-customer currency, current value, and potential value accurately. To separate the indications of commercial banks' customer lifetime appraisal, quantitative and qualitative indicators can be used. While qualitative indicators can be obtained through professional scoring, quantitative indicators can be obtained directly through calculation. Using unascertained clustering, the study divides commercial bank customers into four categories: quality customers, backbone customers, mass customers, and low-class customers. Unascertained clustering corrects the shortcomings of C-mean value clustering and provides a quantitative description of the sample's properties. It is also more logical to use uncertain clustering to divide commercial bank clients. Furthermore, commercial bank decision-makers should base their decisions on science rather than untrustworthy groupings.

Zhang Xiao-bin, Gao Feng, Xi'an School of Computer Science [4] - In this study, the fuzzy C-means clustering algorithm was improved and proven to be useful for segmenting customers and determining high-value customer group attributes. The goal of this data mining process is to keep extracting and discovering patterns in large data sets using machine learning methods. Customer churn is the most important factor in resolving any problem in any company. The data mining pattern will solve the major issues in the customer segmentation process and accurately predict the customer's value. Customer segmentation is the process of managing and analyzing client information, as well as processing business data stored in companies into more data access knowledge. The Mercer Kernels function is used in this function to take input values and separate the customer output value data. And the Kernels method is used to cluster datasets during the segmentation process. And in this, there is data access by evaluating the usefulness and reliability of the data mining attribute findings. This process approach can save the enterprise money and manpower while improving the situation of data exploration and data deficiency.

XIONG Weiwen, Chen Liang and atal [5], The most common segmentation is customer relationship management. Mall owners primarily focus on the wants and needs of their patrons. Using the correlation model, financial model, frequency model, and recency model, customer behavior has been identified. The algorithm is the algorithm, and the AHP is used to weigh the customer. whether or not their products are sold in the environment of the consumers. At stores or malls, those products are noted as being challenging for those customers. These are determined by RFM values that have been compiled to categorize customers. If value = R.F.M, then Grey correlation analysis is essentially used to identify the closest path for the line-by-line continuation.In this way, we have the majority of three datasets and provide tables, graphs, and diagrams of correlation techniques. A fantastic end logistics division of logistics engineering is a category of platinum that represents various segments of the logistics market. The methods for customer segmentation use empirical analysis to show the effectiveness of the suggested method and its performance. Customers of a certain type pay more attention to the level of service provided in logistics, but they are less valuable. To draw in these customers, they should use effective advertising, cutting-edge marketing strategies, and customized logistics solutions to increase customer loyalty, firmly hold onto customers like these, and generate high profits. The RFM (Recency, Frequency, and Monetary) and grey correlation dimensions are used in this paper to structure the customer segmentation model. There are provided the formulas for calculating the various dimensions. Based on it, this paper chooses an illustration for a case study, and the findings demonstrate the viability of the method.

3. IMPLEMENTATION

Algorithm:

Step 1: Start

Step 2: Collect the Dataset

Step 3: Check whether the data has null values or not, if yes goto step 4 otherwise goto step 6

Step 4: Then Clean the data by using preprocessing techniques

Step 5: If no missing data will be found, if again missing data is observed repeat step 4 again

Step 6: Then apply K-Means and Agglomerative techniques

Step 7: Then the data will be evaluated and divided into Clusters

Step 8: Later Exploratory Data Analysis will be done and then it will be our final Result

Step 9: Stop

4. PROPOSED SYSTEM

The proposed study's main goal is to create a machine learning model to separate customers. The study's specific goals are as follows:

- Recognize the current situation as it relates to the clientele of the business.
- Prior work on customer segmentation should be consolidated.
- Investigate data to discover the relationship between the customer and the attributes that benefit the business.
- Use unsupervised machine learning to perform clustering analysis and evaluate built models.
- For the best results, use tableau for data visualization and interpretation.
- The suggested system is built using unsupervised machine learning techniques like
  - Agglomerative Clustering
  - Elbow Method

to calculate the number of clusters in a dataset. designed for interpretation and validation of consistency within cluster analysis.

Agglomerative Clustering:

It is a bottom-up approach, in which the algorithm starts by taking all data points as a single cluster and then merges them. The foundation of agglomerative clustering is the creation of a hierarchy represented by dendrograms. The algorithm uses the dendrogram as memory to store information about how clusters are formed. The clustering process begins by creating N clusters for N data points, which are then combined along the closest data points in each step so that there is one fewer cluster in the current step than in the previous one.

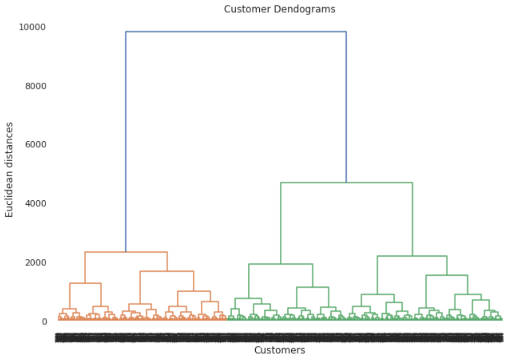


Fig-1: Agglomerative cluster

K-Means Clustering:

It is the most basic clustering algorithm based on the partitioning principle. The algorithm is sensitive to the initialization of the positions of the centroids; the number of K (centroids) is calculated by the elbow method, and after the calculation of K centroids in terms of Euclidean distance, data points are assigned to the cluster's closest centroid. Barycenters are assigned after the cluster is formed. calculated by the means of the cluster, and this process is repeated until there is no change in centroid position.

Elbow Method:

For the K-means clustering algorithm, the elbow method is used to determine the best value for K. The SSE of each data point is calculated using its nearest centroid and a range of K values. The elbow is the point at which we should stop further subdividing the data because it is the value of K at which the SSE is most likely to decline as K increases.

Dendrogram:

The hierarchical representation of an object, the dendrogram, is used to determine the output of hierarchical clustering. The height of each clade (horizontal line) is used to interpret the dendrogram; the lower the height, the more associated data points there are, and the higher the height, the fewer associated data points there are.Fig. 1 demonstrates the dendrogram created using our dataset. The formation of the clusters and their eventual convergence into a single cluster are shown in the figure.

In Fig. 1, we look for the longest vertical line that is not being cut by any of the clades and extends virtually over the entire width of the graph. The second-to-last clade in green has its right leg bearing the longest vertical line that is not being cut by any clade. A dendrogram is used to find the optimal number of clusters to apply for agglomerative clustering. Now, by imagining a horizontal line that cuts through the longest vertical line, we obtain the horizontal line that cuts through a total of five vertical lines, giving us the ideal number of clusters for our dataset

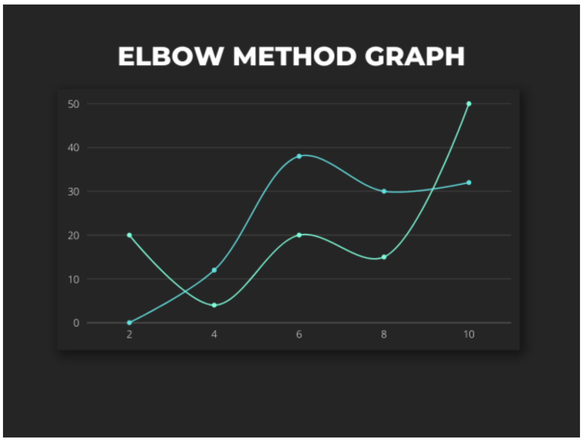


Fig-2: Elbow graph

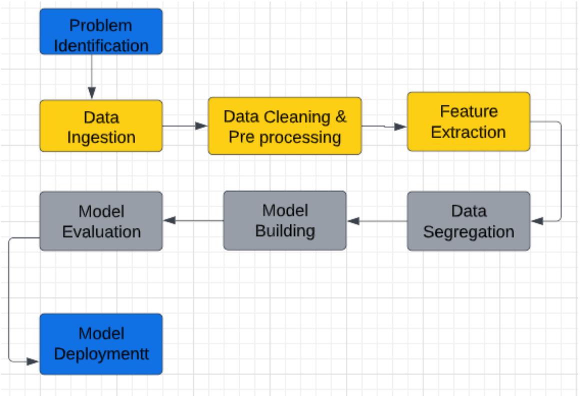
Bandwidth:

The radius of the circle (kernel) describing how many data points should be in the cluster can be thought of as the bandwidth. It is the sole requirement for the input to the Mean shift algorithm with K Nearest Neighbors as a calculator. The initialization of bandwidth has a significant impact on the convergence of the mean shift algorithm; a small value can slow convergence while a large value can accelerate it.

Mean shift clustering:

This clustering algorithm is a non-parametric iterative algorithm that works by treating all data points in the feature space as an empirical probability density function. The algorithm clusters each data point by allowing data points to converge to a region of local maxima, which is accomplished by fixing a window around each data point, determining the mean, shifting the window to the mean, and repeating the steps until all data points converge, forming the clusters.

SYSTEM ARCHITECTURE:



WORKFLOW:

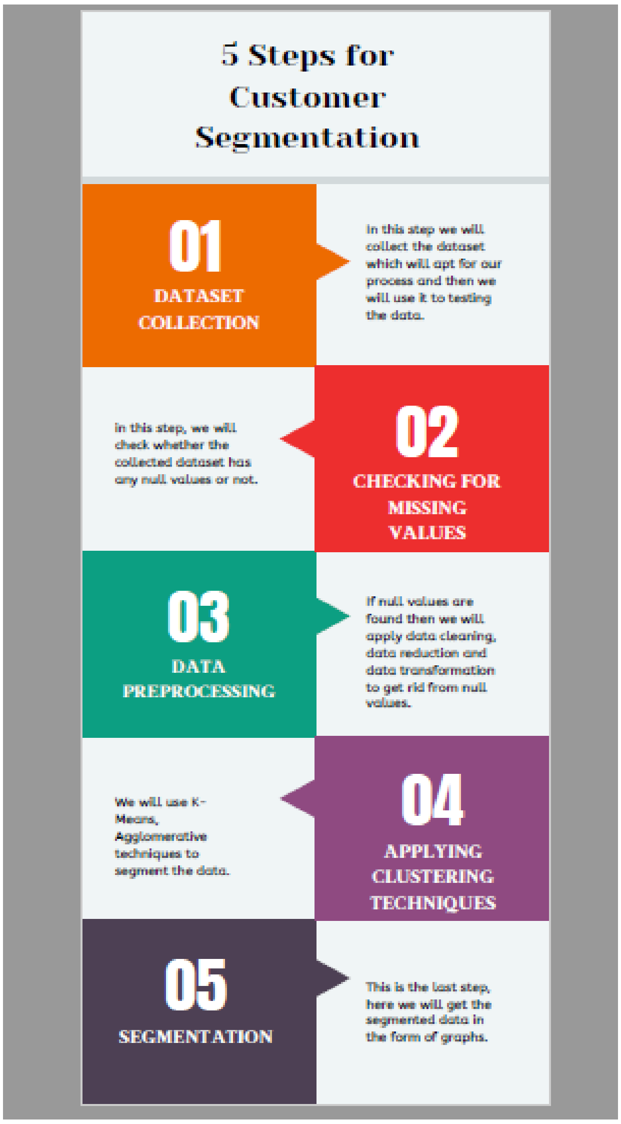
The dataset was obtained from a neighborhood retail store and includes two features: the typical number of visits to the store and the typical amount of shopping done on an annual basis. The dataset was obtained from a neighborhood retail store and includes two features: the typical number of visits to the store and the typical shopping done on an annual basis. The dataset was obtained from a neighborhood retail store and includes two features: the typical number of visits to the store and the typical amount of shopping done on an annual basis.

Data Cleaning:

Data cleaning is the process of removing inaccurate, incomplete, and misleading data from datasets and replacing the missing values. There are a few strategies in data cleaning.

Data Integration:

Assembling various sources into a single dataset One of the key processes in data management is data integration. During data integration, there are a few issues to take into account.



5 Steps for Customer Segmentation

01 DATASET COLLECTION

In this step we will collect the dataset which will apt for our process and then we will use it to testing the data.

02 CHECKING FOR MISSING VALUES

In this step, we will check whether the collected dataset has any null values or not.

03 DATA PREPROCESSING

If null values are found then we will apply data cleaning, data reduction and data transformation to get rid from null values.

04 APPLYING CLUSTERING TECHNIQUES

We will use K-Means, Agglomerative techniques to segment the data.

05 SEGMENTATION

This is the last step, here we will get the segmented data in the form of graphs.

Fig-3: Flow of Work

Data Reduction:

This method facilitates data volume reduction, which facilitates analysis while yielding essentially the same outcome. Additionally, this decrease helps free up storage space. Dimensionality reduction, numerosity reduction, and data compression are some of the techniques for data reduction.

5. RESULTS:

The process of testing involves running a program with the goal of identifying errors. Our software must be error-free in order to function properly. If testing is completed successfully, all software flaws will be fixed. The process of ensuring and validating that software or an application is bug-free satisfies technical specifications as determined by its design and development, and effectively and efficiently meets user requirements while handling all exceptional and boundary cases, is known as testing. The test objectives are listed below.

A program is tested by being run with the goal of identifying any errors. A strong test case is one that has a good chance of spotting an error that hasn't been identified yet. A test that finds an error that has not yet been identified is successful.

The output screens are as follows:

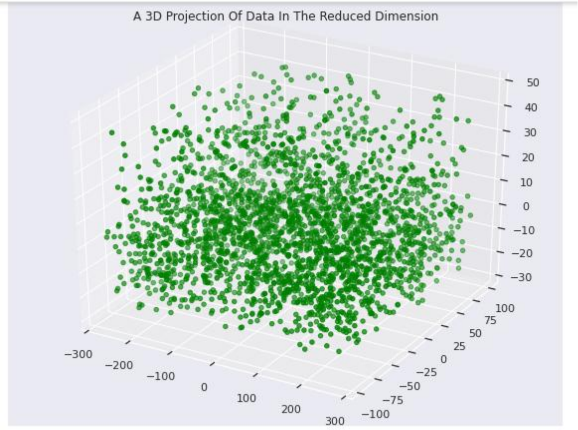


Fig-4: Clustering

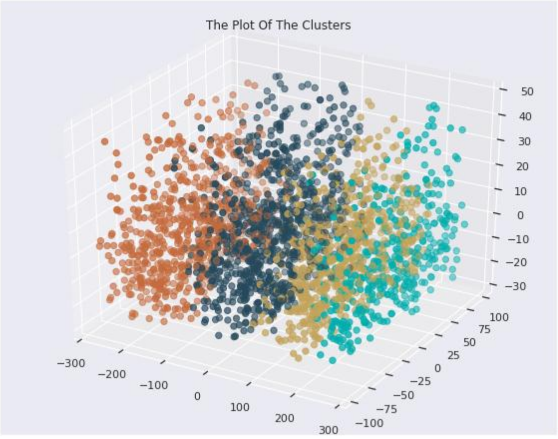


Fig-5: Plot of Clusters

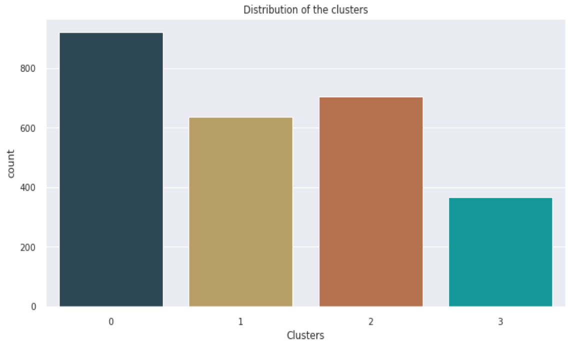


Fig-6: Clusters

6. CONCLUSIONS:

In order to segment customers, the study tried to build unsupervised machine learning models like K Means and hierarchical clustering. The next steps would be to take a closer look at large dataset features, evaluate them, and create an effective model.

First, we began with the pre-processing of the data. Following that, we used clustering algorithms. We chose MiniBatchK-Means as the second model after contrasting these clustering models. The data was then split into six clusters because it is simple to predict customer behavior using data from four clusters. But each of the clusters has its own unique traits.

7. REFERENCES:

[1]AMAN BANDUNI, Prof ILAVENDHAN A, "Customer Segmentation Using Machine Learning" 2019 International Conference of Security, Pattern Analysis.

[2]Z. Wang, Y. Zuo, T. Li, C. L. Philip Chen and K. Yada, "Analysis of Customer Segmentation Based on Broad Learning System," 2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2019, pp. 75-80, doi: 10.1109/SPAC49953.2019.237870.

[3]H. Su-li, "The customer segmentation of commercial banks based on unascertained clustering," 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM), 2010, pp. 297-300, doi: 10.1109/ICLSIM.2010.5461416.

[4]X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and Business Intelligence, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

[5]X. Weiwen, C. Liang, Z. Zhiyong and Q. Zhuqiang, "RFM Value and Grey Relation Based Customer Segmentation Model in the Logistics Market Segmentation," 2008 International Conference on Computer Science and Software Engineering, 2008, pp. 1298-1301, doi: 10.1109/CSSE.2008.79.