Customer Segmentation using K-Means Clustering

Hemashree Kilari[*1], Sailesh Edara[2], Guna Ratna Sai Yarra[3], Dileep Varma Gadhiraju[4]
[1,2,3,4]Student,Department of Computer Science Engineering,
GITAM University, Visakhapatnam,
Andhra Pradesh, India.

Abstract: **Our environment often generates vast volumes of data, and the importance of analyzing that data cannot be overstated. In this modern era of innovation, when everyone is competing to be better than everyone else, the corporate strategy must be tailored to the current conditions. Because so many potential customers are unclear what to buy and what not to buy, today's business relies on new ideas. Businesses are unable to evaluate their target population on their own. This is where machine learning comes in; various algorithms are used to uncover hidden patterns in data in order to make better decisions. Clustering is a machine learning technique that entails comparing data points from several groups. Market research, medical data, search optimization, pattern recognition, image processing, and other applications are among them. Customer segmentation, which falls under market research, is the topic of our project. Customer segmentation is defined as the classification of consumers into groups based on their shared characteristics. In today's environment, it's critical for businesses to divide their clients into groups based on their age, gender, geography, and other characteristics. This allows businesses to focus on certain clients who are most likely to purchase their goods. Machine learning will provide them a competitive advantage over their competitors if they can use it successfully to better their operations. This project's main goal is to utilize the K-means algorithm to divide customers into groups based on their attributes. Finally, using the mean value as the major indication, the data from the various clusters will tell us which group the new client will belong to.**

Keywords: K-Means algorithm, Customer segmentation, Mall Customers, Silhouette method, Elbow method, Python, Machine Learning.

I. INTRODUCTION

Existing businesses must embrace marketing tactics to stay competitive as new enterprises open their doors every day. In today's society, the key marketing rule is "change or perish." Businesses are finding it more difficult to meet the demands of each and every one of their consumers as the number of customers grows [1]. In this case, data mining can assist in identifying hidden tendencies in a company's database.

Client segmentation is a data mining approach that splits a customer base into multiple groups based on characteristics such as gender, age, hobbies, and other buying habits[2]. Customers are split into groups based on shared qualities, in other words. Segmentation can influence marketing strategy directly or indirectly because it opens up many new paths to discover, such as which segment the product will be good for, customizing marketing plans for each segment, providing discounts for a specific segment, and deciphering the customer and object relationship, which was previously unknown to the company[3].

A customer segmentation strategy allows firms to target particular groups of consumers, resulting in more efficient marketing resource allocation and greater potential for cross and up-selling. It's easier for firms to create unique offers to entice customers to spend more when they deliver customized communications to a group of customers as part of a marketing mix tailored to their requirements. Consumer segmentation may help with customer loyalty and retention by improving customer service. Because of their individualized character, marketing materials that employ customer segmentation are more valued and appreciated by the consumer who gets them than impersonal brand communications that ignore purchase history or any type of customer relationship [4].

Customer segmentation has been demonstrated to benefit from clustering. Clustering is a sort of unsupervised learning that allows us to locate clusters in unlabeled datasets. Clustering techniques include K-means, hierarchical clustering, DBSCAN clustering, and others [5]. The major purpose of this work is to apply a data mining strategy to find consumer groups using the K-means clustering algorithm to partition data. The silhouette method yields the most clusters.

II. CUSTOMER SEGMENTATION

Companies have had to grow their profitability and company through time as a result of fierce competition in the business field to meet customer expectations and attract new clients depending on their desires. It's tough and time-consuming to identify and respond to each customer's needs. This is owing to the fact that, among other things, clients have a diverse set of aims, interests, and preferences. Customer segmentation, as opposed to a "one-size-fits-all" strategy, divides customers into groups based on comparable characteristics or habits. Customer segmentation is a marketing strategy that divides a market into distinct, homogeneous groups. The data used in the customer segmentation strategy, which divides customers into categories, is based on a number of factors, including regional circumstances, economic patterns, and demographic trends, and behavioral patterns. A client segmentation technique can help a company's marketing resources be better utilized [6].

III. MACHINE LEARNING

We've seen machine learning in action in a variety of businesses, like Facebook, where it helps us identify ourselves and our friends, and YouTube, where it helps us discover new content., where it recommends movies based on our preferences. Machine learning is divided into two types: unsupervised learning and supervised learning. A data analyst often employs supervised learning to address problems like classification and regression, implying that the data in this case is targetable and that we want to anticipate in the future, such as assessing a student's worth or the amount of monthly costs [7].

Unsupervised learning, on the other hand, may or may not have a label or goal in mind. Because it is based on a mathematical model, clustering, for example, does not have a changeable goal. For instance, we might want to group students depending on their learning interests. or product purchases. Strong competition exists in the marketing business, particularly malls, in order to boost consumer numbers and so produce big profits. [8]

Many retailers and other marketplaces are already using machine learning to achieve this goal. Malls and shopping centres use the information they collect from customers to construct machine learning models that target the right individuals. This not only boosts revenue and visitor numbers, but it also increases business efficiency.

IV. CLUSTERING

Clustering is a technique for finding comparable groups of data in a large dataset. Members of each group are more alike than members of other groupings. Cluster-based segmentation has been popular in data analysis since the 1970s, particularly in marketing. Clustering is not a systematic data analysis approach, and although it provides a lot of flexibility, it is very dependent on the data or sample utilized, as pointed out by. The "tandem technique," which consists of two processes, one of which the first is factor analysis, while the second is cluster analysis, is a combination of factor and cluster analysis., is a statistical strategy employed by this cluster analysis investigation. In a number of pieces, this method has been blasted. All of this is due to a flaw in his method: early factor analysis may wipe away existing cluster formations. Because binary variables are employed, hierarchical cluster analysis may be used instead of tandem cluster analysis. Many scientists questioned the method's validity at the time, and nonhierarchical methodologies have dominated research since the 1980s [9].

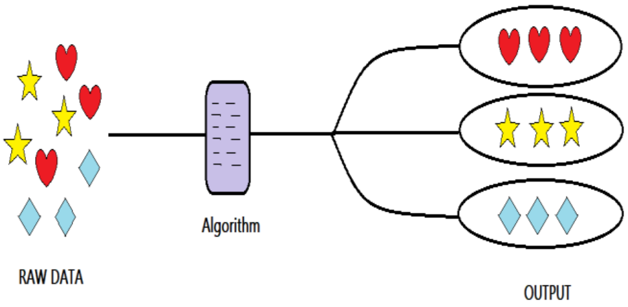


Figure.1. Clustering

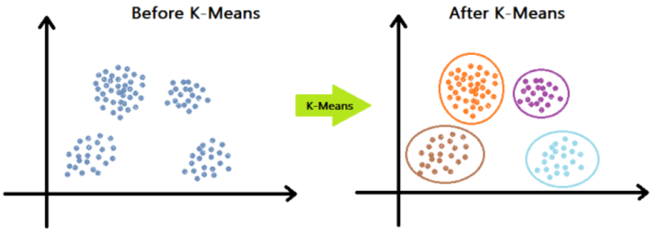
IV.1.CUSTOMER SEGMENTATION AND MACHINE LEARNING:

Using machine learning algorithms to find new segments is one method of client segmentation. Machine learning consumer segmentation enables advanced algorithms to reveal insights and groups that marketers may struggle to obtain on their own. Marketers who create a feedback loop between their segmentation model and campaign results will see their customer groups improve over time. In these circumstances, the machine learning model will be able to not only fine-tune its segment definitions, but also determine whether one segment outperforms the others, thus maximizing marketing performance [10].

IV.2. CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING:

The most well-known unsupervised partitioning clustering approach is K-Means Clustering. This clustering approach, commonly known as the centroid-based technique, divides data into non-hierarchical categories. [11], [12]

The dataset is separated into a collection of k groups in this sort of partitioning, where K is the number of pre-defined groups or clusters. When compared to another cluster centroid, the cluster center is built so that the distance between data points in one cluster is as low as possible



V. ALGORITHM

Step 1: Select the number K to determine the number of clusters.

Step 2: At random, select K locations or centroids. (It's possible that it's not the same as the incoming dataset.)

Step 3: Form the preset K clusters by assigning each data point to the centroid that is closest to it.

Step 4: Calculate the variance and move the centroid of each cluster.

Step 5: Reverse the previous three steps, reassigning each datapoint to the cluster's new closest centroid.

Step-6: Go to step-4 if there is a reassignment; otherwise, go to FINISH.

Step 7: The model is now complete.

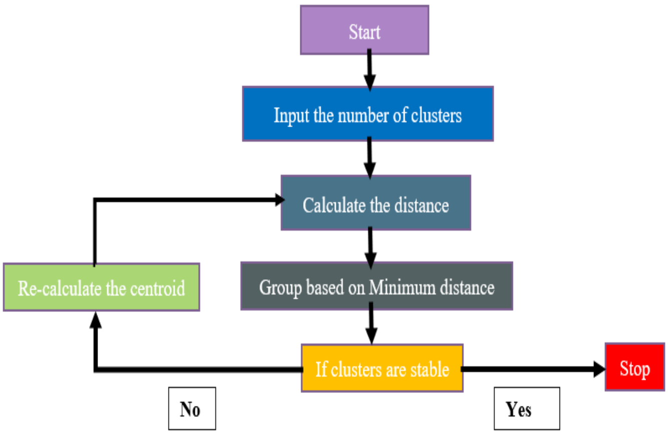

Start
Input the number of clusters
Calculate the distance
Re-calculate the centroid
Group based on Minimum distance
If clusters are stable
Stop
No
Yes

Figure.2. Flow of K-Means algorithm

VI. RELATED WORKS

Using data from an online retail outlet, this research offered an implementation of the k-Means clustering technique for consumer segmentation. Customers were divided into mutually exclusive groups in this model, three clusters in this case [13].

The authors chose internal clustering validation over external clustering verification since the dataset was unbalanced. External clustering verification relies on external data such as labels. Internal cluster validation can be used to choose the best clustering algorithm for the dataset and, conversely, to cluster the data within the cluster appropriately[14].

Machine learning approaches are excellent for evaluating customer data and uncovering patterns and insights. Artificially intelligent models are effective decision-making tools. They have the ability to precisely define client categories, which is far more difficult to achieve manually or with traditional analytical methods. Machine learning algorithms come in a number of flavors, each of which is best suited to a certain context. The model in this study is built on the k-means clustering technique, which is a popular machine learning algorithm for client segmentation challenges[15].

VII. OUTLINE OF EXISTING MODEL

1.The existing base paper uses “Elbow method” to find out the minimum optimal clusters for the K-means clustering.
2.But elbow method does not work effectively in a few cases, for instance take the below scatter plot.

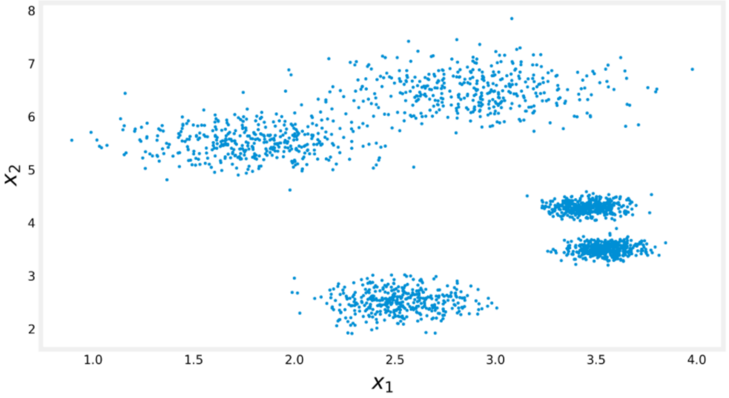


Figure.3. Scatter plot for X1 vs X2

3.Humans may be able to tell that the data originates from five different clusters, but we struggle to perceive high-dimensional data.
4.The Elbow Method, as seen on the left, would most likely lead us to k = 4.
5.The Elbow Method causes us to conclude that two of the clusters are one since they are so close together. This is due to the fact that establishing a centroid in the center of both clusters reduces the relative distance between data points.
6.As a result, calculating the appropriate number of clusters for our clustering job requires an approach that is more accurate, rigorous, and dependable. At this moment, the silhouette score is used.

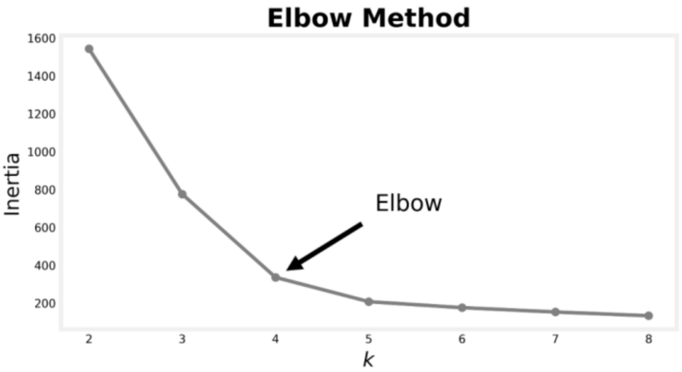


Figure.4. Elbow plot

VIII. OUTLINE OF PROPOSED MODEL

The average silhouette coefficient across all dataset occurrences is used to calculate the silhouette score. The silhouette coefficient, which ranges from -1 to 1, measures how close points in one cluster are to points in nearby clusters. [16].

The silhouette coefficient is calculated as follows:

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

Where an is the mean distance between instances in the same cluster (i.e., mean intra-cluster distance) and b is the mean distance between clusters closest to each other (i.e., mean nearest-cluster distance) (i.e., the mean distance to the instances in the nearest cluster, other than the instance own cluster).
When b > a, the silhouette coefficient is closer to +1, indicating that the instance is most likely at the cluster's core.
In the meanwhile, if b = a, the silhouette coefficient is 0, indicating that the instance is on the edge of two clusters.
Finally, if a >> b, the instance is near to another cluster center, implying that it was most likely assigned to the wrong cluster.

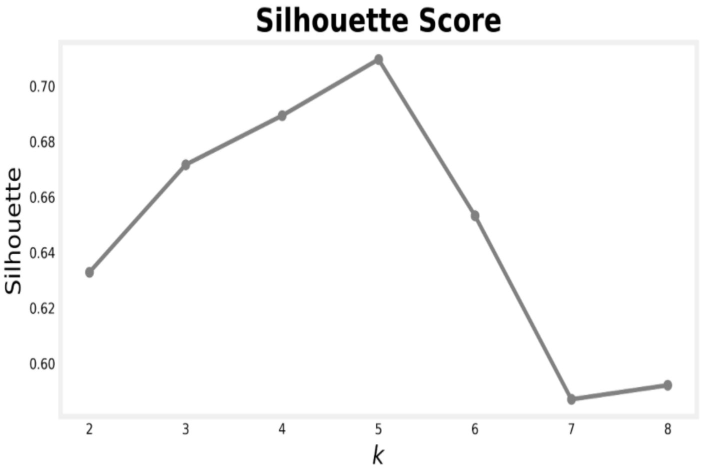


Figure. 5. Silhouette plot

IX. EXPERIMENTAL SETUP

LIBRARIES USED:

Scikit-learn: It is a free Python machine learning software, sometimes known as sklearn. It is meant to interact with the Python numerical and scientific libraries NumPy and SciPy, and features support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering algorithms.

Seaborn: Seaborn is a matplotlib-based open-source Python library. It's used for exploratory data analysis and data visualization. Seaborn makes using data frames and the Pandas library a breeze. The graphs that are generated may also be readily changed.

NumPy (numerical Python):is a package that contains multidimensional array objects and tools for manipulating them. NumPy is a Python library that allows us to perform mathematical and logical operations on arrays.
NumPy is widely used in combination with SciPy and Matplotlib (Scientific Python) (plotting library). This combination is frequently used as a substitute for MATLAB, a prominent technical computing platform. The Python counterpart to MATLAB, on the other hand, is today regarded as a more contemporary and comprehensive programming language.

Pandas: is a Python toolkit for data science, data analysis, and machine learning that is open-source. It is based on NumPy, a multi-dimensional arrays-supporting library. Pandas, being one of the most widely used data manipulation tools, works well with a variety of other Python data science modules.

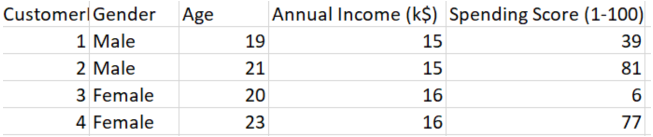
Matplotlib: For 2D array charts, Matplotlib is a superb Python visualization library. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the entire SciPy stack. The ability to show vast volumes of data in simple images is one of the most essential advantages of visualization. Line, bar, scatter, histogram, and more graphs are available in Matplotlib.

The Platform that was utilized was:

Jupyter Notebook is a server-client program that allows you to edit and run notebook documents, code, and data using a web browser. The Jupyter Notebook App can be operated locally on a PC with no internet connection (as described in this article) or remotely on a server with internet access. Users can build and organize processes in data science, scientific computing, computational journalism, and machine learning using the versatile interface.

X. METHODOLOGY

A shopping center store provided the dataset for clustering using the K-means algorithm. Five attributes and 200 tuples make up the data set, which represents the information of 200 consumers. The characteristics in the data collection are CustomerId, gender, age, yearly income (k$), and spending score on a scale of (1-100).



CustomerI	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77

Table 1. Dataset

To begin, we'll need to figure out what kind of data we'll be working with (see table 1 for the dataset). We employ a straightforward yet comprehensive dataset that contains customer ID, gender, age, yearly income, and purchase score. The worth of the client's shopping or spending at the mall is represented by an expenditure score that ranges from 1 to 100. (The higher the number, the greater the amount spent.) The dataset's structure has been correctly displayed, and there are no null values.

If a dataset contains null values, duplicates, or other noisy data, data cleaning must be performed. Data cleansing ensures that information is reliable, usable, and available for analysis.
When we have the data, we may visualize it by comparing the annual income and spending score, which is gender-specific. According to the study, there are five different types of plots that illustrate groups of customers who engage in the following activities, as well as customer behaviors linked to yearly income and expenditure scores:
1.Score of High Income/Low Spending
2.Low Income- A high score for spending
3.A high score for spending-despite Low Income
4.Average Income- Average Spending Score
5.High Income- High Spending Score.

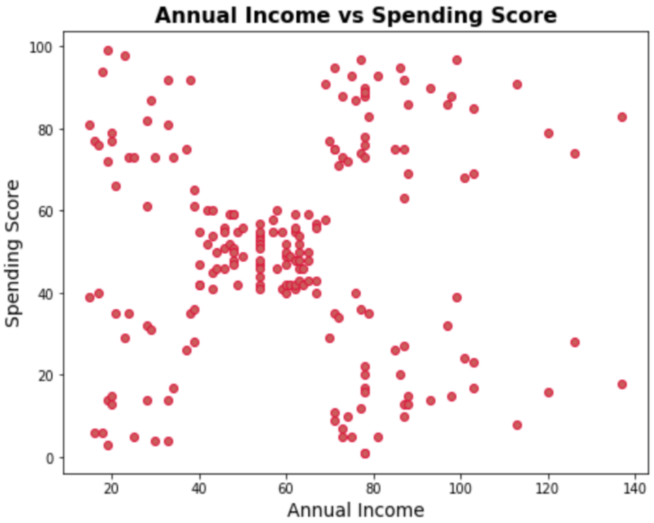
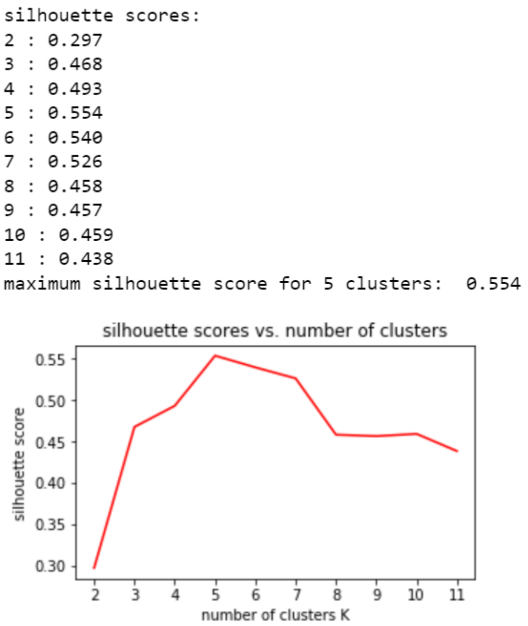


Figure.6. Annual Income vs Spending Score

We can now build a K-means model based on the fact that there are a lot of groups, but not in great detail. The silhouette coefficient approach is used to do Clustering using k-means for a range of k clusters (let's say 1 to 10) and estimate the sum of square distances from each point to its assigned center for each value. Decide on the number of clusters that will give you the best silhouette score. This defines how the silhouette score is calculated. We noticed that once K=5 is reached, there is no rapid movement in WCSS (Within

Cluster Sum of Squares). And, given the number of clusters we have now, K=5 will be the correct number of clusters. 7. Refer to the illustration.



```
silhouette scores:
2 : 0.297
3 : 0.468
4 : 0.493
5 : 0.554
6 : 0.540
7 : 0.526
8 : 0.458
9 : 0.457
10 : 0.459
11 : 0.438
maximum silhouette score for 5 clusters:  0.554
```

Figure.7. Silhouette approach result.

We can divide the plot into various groups, determine cluster can be prioritized, and then assign a label to each using the method stated above. The K-means approach can be used to decide which of the five clusters should be targeted, namely clients with Moderate Income- Moderate Spending Score, High Income- High Spending Score, and Low Income- High Spending Score. The required consumers have been located, as shown in Figure 8.

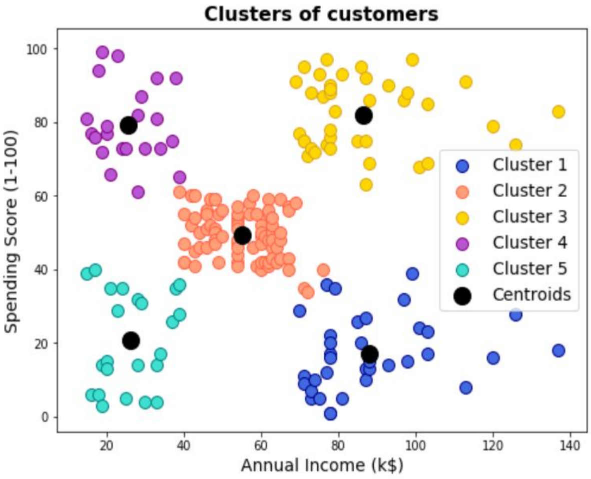


Figure.8. Final cluster of customers

XI. EXPERIMENT RESULTS

Mall shoppers can be divided into five groups depending on their yearly earnings and spending habits. To begin, the yellow group refers to people who have high incomes and high spending scores; this is an excellent example of a mall or retail center being a good target. Because these are the most profitable customers. This person could be a frequent shopper at a mall, where they could be readily apprehended by mall security.

The blue group, on the other hand, consists of those who have a lot of money but spend very little. This is an intriguing case due to the multiple causes for the development of such a club. Assume they're people who enjoy shopping but are dissatisfied with the mall's current offerings or facilities. These are good targets as well, but we'll need to figure out why they're spending so little. The manager of the department or the mall's authority could design or build a facility to entice these groups to come in and have their needs met.

Based on the facts we know, they have average earnings and expenditures, as illustrated by the orange group. We can assume they're folks who don't always buy items but have a strong desire to spend despite their financial limits. As a manager, I aim to avoid marketing strategies that target this population as much as possible because they aren't a substantial source of revenue for the mall. They might, however, employ a range of data analysis techniques to help them increase their spending.

There's the violet-colored group, which includes people with low income but high spending scores; despite their low income, people in this group enjoy or are interested in spending money. This is also possible if customers are happy with the mall's services and thus feel compelled to spend money because they are happy with the service.

The green group, fifth, had low annual incomes and poor spending habits. It's also reasonable that they have a limited budget and would cut corners wherever possible, even if what they're doing is a sensible and great decision in light of their circumstances. The folks in this cluster should be given the lowest priority by a mall manager.

By analyzing the data, we can predict customer behavior based on their Annual Income and Spending Score. This cluster analysis may be applied to a number of consumer marketing methods. We'd want to keep our target clientele, who have a high income and a high spending score, because they deliver the biggest profit margin. Because of their lifestyle demands for a high income and low spending score, customers will be lured to the Mall Supermarket because of the great variety of things available. Less Income Less Spending Scores can obtain more promotions, and they will be tempted to spend by receiving offers and discounts on a frequent basis. A cluster analysis may be used to establish what kind of things clients wish to consume, allowing for the development of more targeted marketing efforts. The people in clusters 3 and 4 are the potential clients in this situation.

XII. CONCLUSION

This study demonstrates that client segmentation in shopping malls is achievable despite the fact that this form of machine learning application is highly useful in the market, a manager can concentrate all of his or her attention on each cluster that has been discovered and meet all of their requirements. Mall managers must be able to understand what customers require and, more importantly, how to meet those needs. analyze their purchasing habits, and establish frequent encounters with customers that make them feel comfortable in order to satisfy their demands.

REFERENCES

[1] “Customer segmentation based on survival character,” IEEE, Jul. 2003.

[2] “Customer Segmentation Using K Means Clustering,” Towards Data Science, Apr. 2019.

[3] Ruhul Reddy, “Who’s who: Understanding your business with customer segmentation,” INTERCOM.

[4] Kristen Baker, “The Ultimate Guide to Customer Segmentation: How to Organize Your Customers to Grow Better,” Hunspot.

[5] Tim Ehrens, “customer segmentation,” TechTarget.

[6] V.Vijilesh, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 05, May 2021.

[7] Expert Systems with Applications, vol. 100, Feb. 2018, "Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data."

[8] “Cluster analysis.”, Wikipedia.

[9] "CUSTOMER SEGMENTATION USING MACHINE LEARNING," IJCRT, AMAN BANDUNI and ILAVENDHAN A, vol. 05, 2018.

[10] Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury, “Customer Segmentation using K-means Clustering,” IEEE, Jul. 2019.

[11] I. S. N. Chinedu, S. O. C. Kalu, E. & C. E. D. U. of U. U. A. S. O. C. Kalu, E. & C. E. D. U. of U. U. A. S. O. C. Kalu, E. & C "Application of the K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," vol. 4, no. 10, 2015, by Pascal Ezenkwu, International Journal of Advanced Research in Artificial Intelligence.

[12] Author Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning [Beginners Guide],” neptuneblog, Dec. 13, 2021.

[13] K. Maheswari, “Finding Best Possible Number of Clusters using K-Means Algorithm,” International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 1S4, Dec. 2019.

[14] AMAN BANDUNI and ILAVENDHAN A, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” IJCRT, vol. 05, 2018.

[15] Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning”, July 10th 2021.

[16] Ashutosh Bhardwaj, “Silhouette Coefficient Validating clustering techniques,” Towards Data Science, May 26, 2020.