CUSTOMER SEGMENTATION USING MACHINE LEARNING & HYBRID MODEL

Prof. Rajendra Arakh[1], Raja Shoaib[2], Prof. Sweta Kriplani[3], Vanshika Yadav[4], Swechchha Agrawal[5]

Dept. of Computer Science & Engineering
Shri Ram Institute of Technology Jabalpur
M.P India

Abstract: Customer segmentation is a fundamental strategy for businesses seeking to tailor their offerings effectively to diverse consumer needs. Utilizing machine learning techniques alongside a hybrid model approach presents a promising avenue to enhance the precision and efficiency of segmentation processes. This abstract outlines a comprehensive framework for customer segmentation that integrates traditional clustering algorithms with advanced machine learning methods. The hybrid model merges both supervised and unsupervised learning approaches to optimize segmentation results, leveraging both labeled and unlabeled data to improve accuracy while accommodating the dynamic nature of customer preferences. By combining unsupervised techniques such as k-means clustering with supervised algorithms like decision trees or support vector machines, the model can identify natural groupings within the data and refine segment definitions based on domain knowledge and business goals. This adaptive capability ensures that segmentation remains relevant and actionable, enabling businesses to develop targeted marketing strategies, personalized product recommendations, and enhanced customer experiences. Ultimately, customer segmentation using a hybrid model approach empowers businesses to gain deeper insights into their customer base, foster long-term loyalty, and drive sustainable growth in competitive markets.

Keywords: predictive features, data analysis, machine learning models, artificial intelligence, K-means clustering, marketing effectiveness, hierarchical cluster, DBSCABN.

I. INTRODUCTION

Customer segmentation is a cornerstone of contemporary marketing strategies, designed to divide a diverse customer base into distinct groups with shared characteristics and behaviors. This segmentation facilitates targeted marketing campaigns, personalized communication, and improved customer satisfaction. Traditional segmentation methods, while somewhat effective, often fall short of capturing the intricate nuances of customer behavior in today's data-rich environment. Machine learning (ML) techniques offer a promising approach to refining customer segmentation by leveraging advanced algorithms to analyze vast amounts of data and extract meaningful insights. Despite the benefits of ML-based segmentation, challenges such as dealing with noisy or unstructured data and ensuring interpretability for decision-making persist. To address these issues, hybrid models that combine the strengths of both ML algorithms and traditional statistical methods have emerged as a compelling solution. This introduction provides an overview of customer segmentation using machine learning, highlights its benefits and challenges, and introduces the concept of hybrid models as a means to enhance segmentation accuracy and interpretability. By integrating ML techniques with traditional statistical approaches, hybrid models offer a balanced solution for customer segmentation, catering to the diverse needs of businesses across various industries.

2. LITERATURE REVIEW

Literature Review Customer segmentation is a pivotal element of marketing strategy, aimed at dividing a heterogeneous customer base into homogeneous groups based on shared characteristics or behaviors. Traditional segmentation methods often rely on manual categorization or simple rules, which limit their effectiveness in capturing the complexity of customer preferences and behaviors. In recent years, machine learning (ML) techniques have become prominent in customer segmentation due to their ability to handle large volumes of data and uncover intricate patterns not easily identifiable through traditional methods. Various studies have examined the application of different ML algorithms, including k-means clustering, hierarchical clustering, and Gaussian mixture models, in customer segmentation tasks. These algorithms provide flexibility and scalability, allowing businesses to identify meaningful customer segments and tailor their marketing strategies accordingly. Despite the advantages of ML-based segmentation, challenges such as data quality issues, algorithm selection, and interpretability constraints persist. In response to these challenges, researchers have proposed hybrid models that combine the strengths of ML techniques with traditional statistical methods to improve segmentation accuracy and interpretability. Hybrid models often integrate clustering algorithms with dimensionality reduction techniques or ensemble methods to enhance segmentation performance while maintaining transparency in the segmentation process. Furthermore, the literature showcases practical applications of ML-driven segmentation and hybrid models across various industries, including retail, e-commerce, banking, and telecommunications. Real-world case studies illustrate the effectiveness of these approaches in identifying distinct customer segments, predicting customer behavior, and optimizing marketing campaigns. In summary, the literature emphasizes the increasing importance of ML-driven customer segmentation and the emergence of hybrid models as a promising approach to addressing the challenges associated with traditional segmentation methods. Future research directions may focus on developing more advanced hybrid models, addressing algorithmic biases, and exploring novel applications of ML techniques in customer segmentation to further enhance marketing effectiveness and customer satisfaction.

3. METHODOLOGY

Customer segmentation using machine learning involves grouping customers into distinct segments based on similarities in their characteristics or behavior. Here's a flow chart for showing general methodology for customer segmentation using machine learning:

4. RESULTS

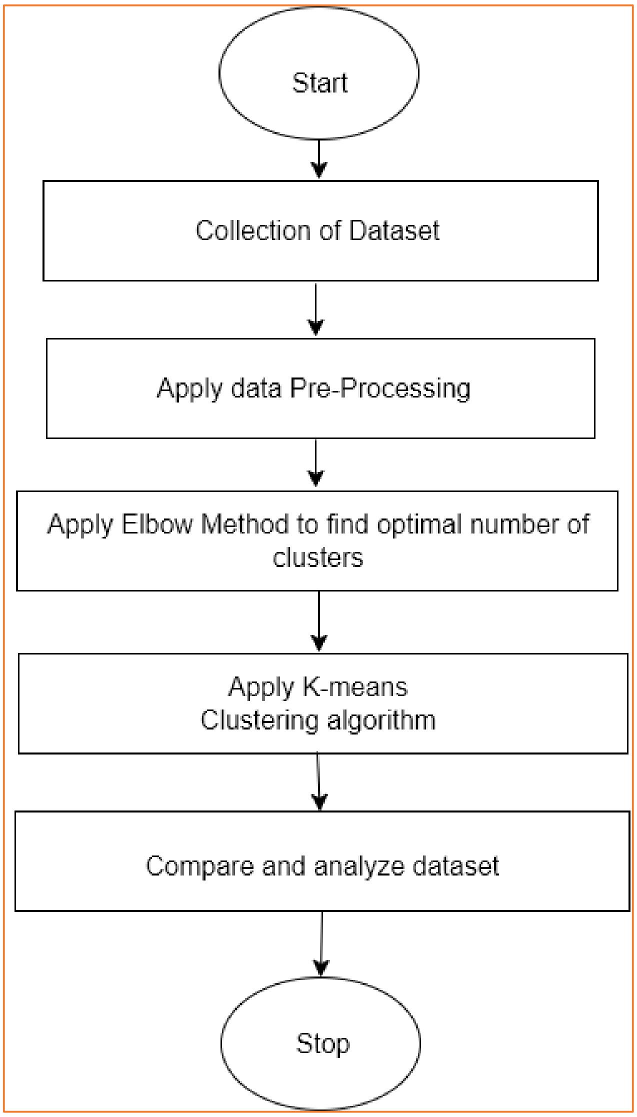

Start
Collection of Dataset
Apply data Pre-Processing
Apply Elbow Method to find optimal number of clusters
Apply K-means Clustering algorithm
Compare and analyze dataset
Stop

Figure1: Flow Chart

This section focuses on data visualization and addresses the normality issue through the Box-Cox transformation. Additionally, it presents the results obtained after applying several machine learning (ML) approaches, namely K-means and DBSCAN. Furthermore, a comparative analysis between these methods is conducted.

4.1. DATA VISUALIZATION

- The dataset is taken form Kaggle.com.
- The dataset name is Mall_Customer.csv consists of 5 columns which are customer ID, Gender, age, Annual income(k$), spending score (1-100) where gender is a categorical value and rest all features are numeric.
- The size of the dataset is (200,5) which is 200 rows and 5 columns.

Before proceeding with the clustering process, an examination was conducted to analyze the distribution of their age, gender, income, spending score values and etc. Figures 1, 2, 3, 4, 5 and 6 clearly indicate the presence of outliers within the dataset.

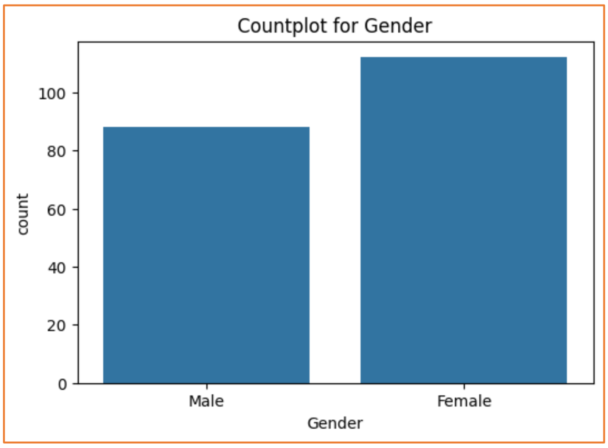


Figure 2: Distribution of Gender

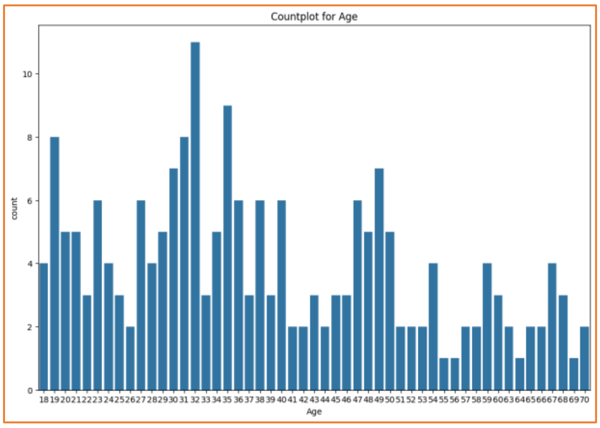


Figure 3: Distribution of Age

A mean spending score for women (51.5) is higher than men (48.5) hence women spend more than man in figure 3.

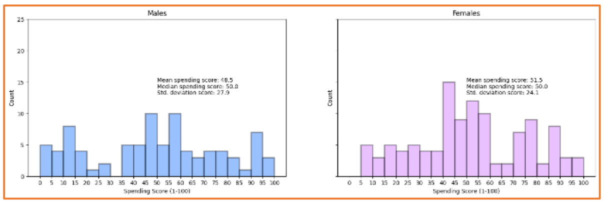


Figure 4: Spending Score of Gender

For Females the age groups of 35-45 earn the most and Female 35-40 age group overshadows the earnings of male.

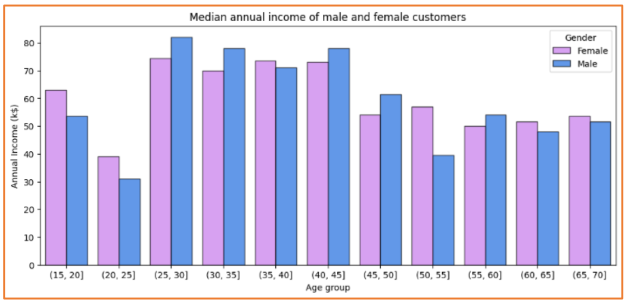


Figure 5: Median annual income of male and female

calculating Pearson's correlation between age groups and spending power.

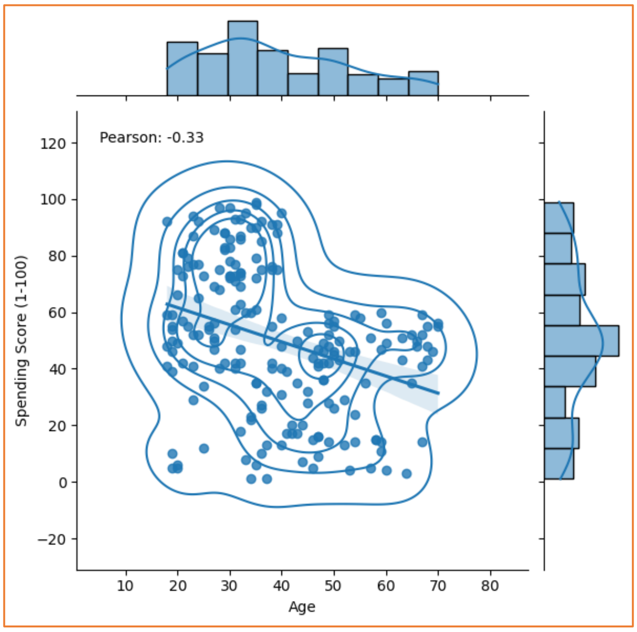


Figure 6: Calculating Pearson’s correlation

4.2. Clustering Approach using K-Means

The most well-known partitional clustering algorithm is K-Means. There are 3 main steps in K-Means algorithm (known also as Lloyd’s algorithm):

1. Split samples into initial groups by using seed points. The nearest samples to these seed point will create initial clusters.

2. Calculate samples distances to groups’ central points (centroids) and assign the nearest samples to their cluster.

The third step is to calculate newly created (updated) cluster centroids.

Then repeat steps 2 and 3 until the algorithm converges.

This is known as NP-hard problem, meaning this is a greedy algorithm and converges to the local minimum. The computational cost of Lloyd’s K-Means algorithm is O(kn), where k is a number of clusters and n is a number of samples. This is not bad when compared with other clustering algorithms. Despite converging usually to a local minimum, K-means is relatively fast and when groups are well isolated from each other it is likely that it converges to the global minimum. Because the result of cauterization depends on the initialization criteria it is common to run the analysis for various initialization points and choose the one with minimum resultant inertia. There are some improvements to the algorithm solving problem of the local minima.

In general, a user of the K-Means algorithm is required to define three main parameters:

A. INITIALIZATION CRITERIA

In scikit-learn, a clever initialization scheme is implemented: "k-means++" proposed by Arthur and Vasilevskiy. It creates initial centroids generally distant from each other increasing probability of obtaining better results. There is also a possibility to use a random point's generator. There are ongoing efforts to create the most efficient seeding method for K-Means algorithm, one of them is based on Independent Component Analysis.

B. NUMBER OF CLUSTERS

Selecting a number of clusters is the most challenging part of setting this algorithm. There are no hard mathematical criteria for this and many heuristic/simplified approaches have been developed. One of the simplest and the most popular one is the elbow method shown in this analysis. Additionally a silhouette score will be used as well. There are also other, often advanced, options for choosing the optimal number of clusters (however, not used in this notebook and not implemented in sk-learn).

C. A DISTANCE METRIC (NOT REQUIRED IN SCIKIT LEARN IMPLEMENTATION)

There are various options to calculate the distance between points. The most popular one is simply the Euclidean metric and it is the one implemented in scikit-learn. It is often called spherical k- means model. It has a drawback that it finds spherical-like groups only and tends to become inflated in highly multi-dimensional analyses ("curse of dimensionality").

There are numerous ongoing researches and variations proposed to K-Means, e.g.:

K-Medoid where the centroid is defined as the most centrally located object) K-Median where the centroid is calculated using median instead of a mean, Fuzzy C-means model.

D. SOME TAKE-AWAYS ABOUT K-MEANS:

- Euclidean distances are used
- Number of clusters has to be defined for the algorithm
- Centroid is calculated using mean distance to cluster members
- Clusters are assumed isotropic and convex
- Stochastic algorithm – results depend on the initialization criteria
- Creates groups of equal variance (minimizes inertia)
- Prone to the "curse of dimensionality"
- Can be run in parallel – so it scales well

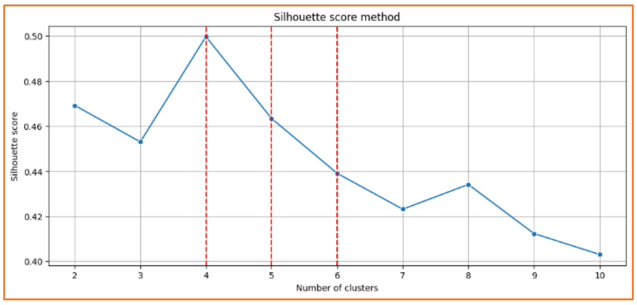
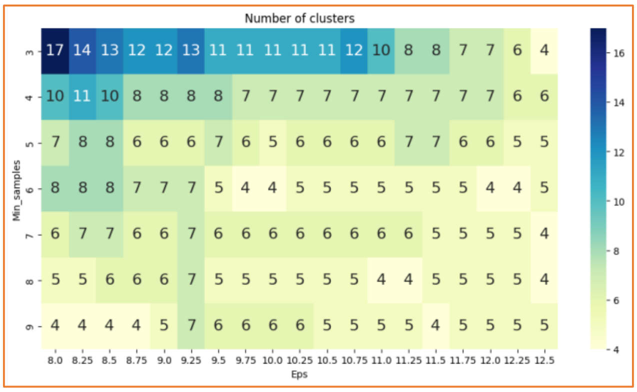


Figure 6: Calculating Pearson's correlation

E. K-MEANS ALGORITHM GENERATED THE FOLLOWING 5 CLUSTERS:

- clients with **low** annual income and **high** spending score.
- clients with **medium** annual income and **medium** spending score.
- clients with **high** annual income and **low** spending score.
- clients with **high** annual income and **high** spending score.
- clients with **low** annual income and **low** spending score.



Min_samples \ Eps	8.0	8.25	8.5	8.75	9.0	9.25	9.5	9.75	10.0	10.25	10.5	10.75	11.0	11.25	11.5	11.75	12.0	12.25	12.5
3	17	14	13	12	12	13	11	11	11	11	11	12	10	8	8	7	7	6	4
4	10	11	10	8	8	8	8	7	7	7	7	7	7	7	7	7	7	6	6
5	7	8	8	6	6	6	7	6	5	6	6	6	6	7	7	6	6	5	5
6	8	8	8	7	7	7	5	4	4	5	5	5	5	5	5	5	4	4	5
7	6	7	7	6	6	7	6	6	6	6	6	6	6	6	5	5	5	5	4
8	5	5	6	6	6	7	5	5	5	5	5	5	4	4	5	5	5	5	4
9	4	4	4	4	5	7	6	6	6	6	5	5	5	5	4	5	5	5	5

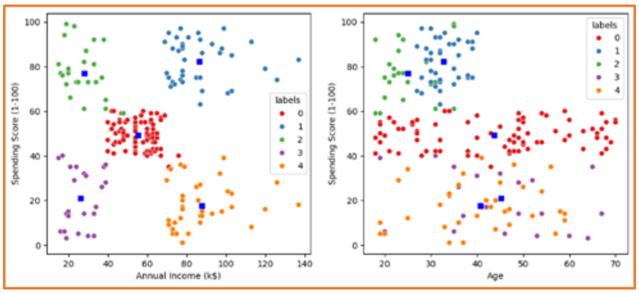


Figure 8: Spending score of Annual income and Age

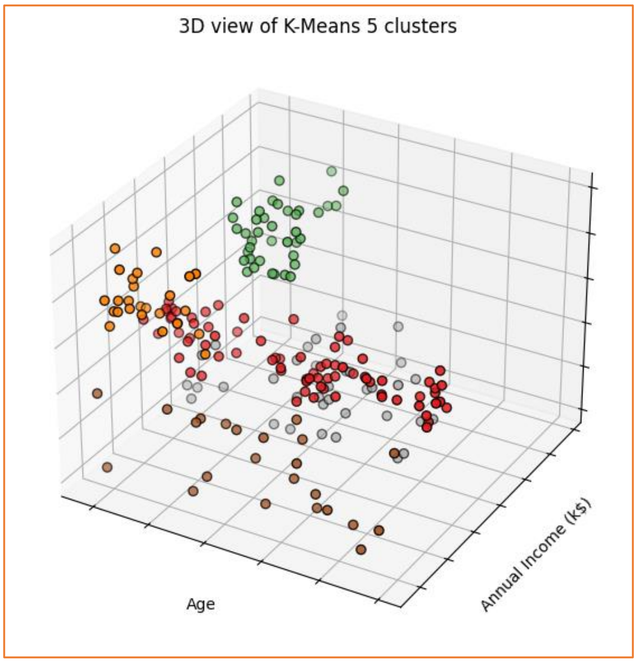


Figure 9: 3D view of K-Mean 5 cluster

4.3. Clustering Approach Using DNSCAN

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k- means which assumes that clusters are convex shaped. The central component to the DBSCAN is the concept of core samples, which are samples that are in areas of high density.

A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, **min_samples** and **eps**, which define formally what we mean when we say dense. Higher min_samples or lower eps indicate higher density necessary to form a cluster.

A “dense region” is therefore created by a minimum number of points within distance between all of them, Eps. Points which are within this distance but not close to minimum number of other points are treated as “border points”. Remaining ones are noise or outliers.

Range of clusters is between 17 to 4. Now we see which cluster has the maximum value from the heat map and choose the corresponding min and eps values.

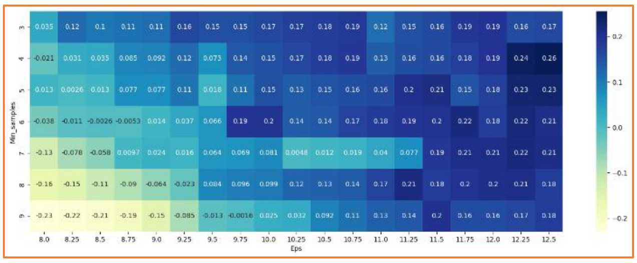


Figure 11: DBSCAN Clustering

4.4. Affinity Propagation Clustering Approach

Affinity Propagation is a clustering algorithm proposed for the first time by Brendan Frey and Delbert Dueck in 2007 ("Clustering by Passing Messages Between Data Points") is somewhat based on this. It is built around the concept of sending messages between a pair of points until it converges. These messages are a way of measuring how similar these two points are to each other and can they be exemplars of each other. In plain English, for a set of data points, a “group formation” process begins, where each sample competes with other ones in order to gain group membership. The ones with most group capital, the group leaders are called exemplars.

The algorithm finds an optimum number of clusters itself. This also implies very high time complexity cost of the order O(n^2T) where n is the number of samples and T is the number of iterations until convergence. However, a big advantage of AP is the lack of sensitivity to the initialization criteria.

The user is required to specify two parameters:

1. Preference which is a negative number and controls how many exemplars are used.
2. Damping factor which prevents numerical oscillations when updating messages.

Important points

● Contrary to K-means clustering, where convergence is determined with some threshold value, with Affinity Propagation you configure a number of iterations to complete.

● During each iteration, each sample broadcasts two types of messages to the other samples.

○ The first is called the responsibility r(i,k) – which is the “evidence that sample k should be the exemplar for sample i”.I always remember it as follows: the greater the expected group leadership of k, the greater the responsibility for the group.

○ The other type of message that is sent is the availability. This is the opposite of the responsibility: how certain i is that it should choose k as the exemplar, i.e. how available it is to join a particular group.

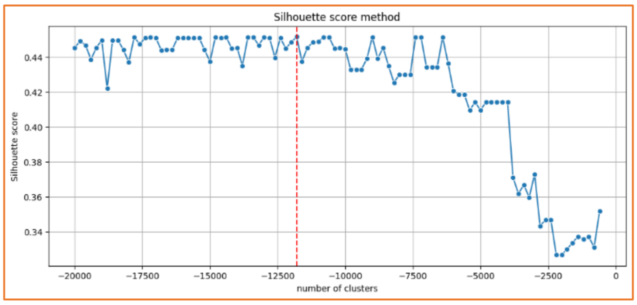


Figure 12: Silhouette score method

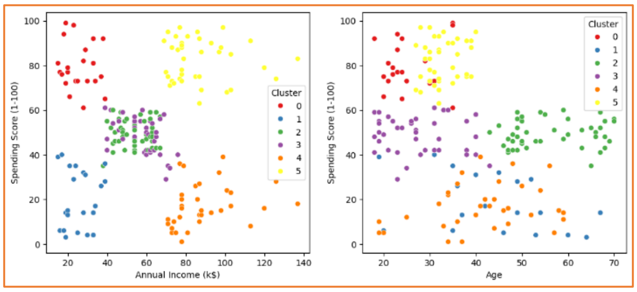


Figure 13: Spending score of Annual income and Age

4.5. HIERARCHICAL CLUSTERING APPROACH

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. This hierarchy of clusters is represented as a tree (or dendrogram).

One common approach to hierarchical clustering is agglomerative clustering, which follows a bottom-up strategy. In agglomerative clustering, each observation initially starts as its own cluster, and then pairs of clusters are successively merged together based on some similarity metric. The algorithm proceeds iteratively, continually merging the two most similar clusters until all observations belong to a single cluster or until a predefined stopping criterion is met.

In summary, hierarchical clustering is a powerful and versatile technique for uncovering structure within datasets, offering both interpretability and flexibility. Its hierarchical nature and visual representation make it a valuable tool for exploratory data analysis, pattern recognition, and classification tasks in various domains.

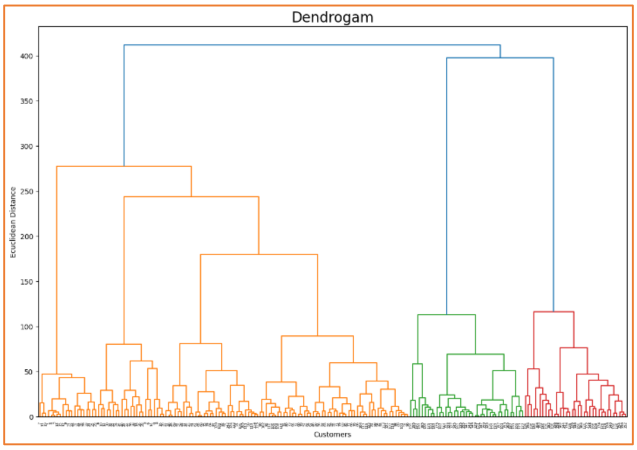


Figure 14: Dendrogram

5. FUTURE USE

A. Personalized Customer Experiences: Machine learning algorithms will continue to refine customer segmentation, enabling businesses to tailor products, services, and marketing messages to individual customer preferences and behaviors. This level of personalization can enhance customer satisfaction and loyalty.

B. Real-Time Segmentation: With advances in technology and computing power, real-time customer segmentation will become more feasible. Businesses will be able to dynamically adjust their strategies based on evolving customer behavior, leading to more timely and relevant interactions.

C. Omni-Channel Integration: Hybrid models combining machine learning with traditional statistical methods will facilitate seamless integration of customer data across multiple channels, such as online, mobile, social media, and physical stores. This holistic view of customer interactions will enable businesses to deliver consistent experiences across various touchpoints.

D. Predictive Analytics: Future models will not only segment customers based on historical data but also incorporate predictive analytics to forecast future behavior. By identifying patterns and trends, businesses can anticipate customer needs and proactively address them, thereby staying ahead of the competition.

E. Ethical Considerations: As customer data privacy concerns continue to grow, there will be a greater emphasis on ethical use of machine learning in customer segmentation. Businesses will need to ensure transparency, consent, and data security to maintain customer trust and comply with regulations.

6. Conclusion

In conclusion, the integration of machine learning techniques into customer segmentation represents a transformative evolution in how businesses understand and interact with their customer base. By combining traditional clustering algorithms with advanced machine learning methods, companies can achieve unprecedented levels of accuracy and efficiency in segmenting their clientele. This hybrid approach leverages both supervised and unsupervised learning, utilizing labeled and unlabeled data to refine segment definitions and accommodate the dynamic nature of customer preferences.

The future of customer segmentation using machine learning holds immense promise, characterized by automated, dynamic, and behavior-based segmentation driven by predictive analytics and ethical considerations. Through continuous refinement and adaptation, businesses can ensure that their segmentation strategies remain relevant and actionable in an ever-changing marketplace. Moreover, by prioritizing customer privacy and data protection, businesses can foster trust and loyalty among their clientele, ultimately leading to sustainable growth and success in competitive markets.

In embracing collaborative, data-driven approaches and addressing common pitfalls, businesses can unlock deeper insights into their customer base, enabling the development of targeted marketing strategies, personalized product recommendations, and superior customer experiences. As technology continues to advance and data sources proliferate, the potential for machine learning-powered segmentation to revolutionize customer engagement and drive business outcomes is boundless.

7. References

[1] Customer Segmentation Tutorial Python Projects K-Means Algorithm Python Training Edureka. https://www.youtube.com/watch?v=4jv1pUrG0Zk&t=1510s

[2] Machine Learning Project With Code Customer Segmentation End To End Implementation by Data Science Diaries. https://www.youtube.com/watch?v=frvailWW6Iw&t=1046s

[3] Implementing Customer Segmentation Using Machine Learning [Beginners Guide] by Dhiraj Kumar. https://neptune.ai/blog/customer-segmentation-using-machine- learning

[4] Customer Segmentation using K-means Clustering J Madhu1 , Kavita K Revanakar2 , Lavanya3 , Akash4 Department of Computer Science and Engineering Srinivas Institute of Technology, Mangalore, Karnataka, India. https://ijaem.net/issue_dcp/Customer%20Segmentation%20using% 20K%20means%20Clustering.pdf

[5] How to Perform Customer Segmentation in Python– Machine Learning Tutorialby Ibrahim Abayomi Ogunbiyi. https://www.freecodecamp.org/news/customer-segmentation- python-machine-learning/

[6] Customer Segmentation using K-Means Clustering l Machine Learning Project by Engineering WalaBhaiya. https://www.youtube.com/watch?v=he9FrAo6pms&t=1048s

[7] Customer Segmentation Using Machine Learning. https://www.javatpoint.com/customer-segmentation-using-machine-learning.

[8] Customer Segmentation Using Machine Learning Prof. Nikhil Patankar a ,1, Soham Dixit a , Akshay Bhamare a , Ashutosh Darpel a and Ritik Raina a a Dept. Of Information Technology Sanjivani College of Engineering, Kopargaon 423601 (MH), India. https://www.researchgate.net/publication/356756320_Customer_Segmentation_Using_Machine_Learning

[9] How to Use Machine Learning For Customer Segmentation by Eric Vardon. https://hawke.ai/blog/machine-learning-for-customer- segmentation/

[10] How To Solve Customer Segmentation Problem With Machine Learning by Mrinal Singh Walia. https://www.analyticsvidhya.com/blog/2021/06/how-to-solve-customer-segmentation-problem-with-machine-learning/