JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation Using Machine Learning for a Shopping Mall Customers

Samreen Siddiqui [1], Dr. Sunita Soni [2] and Dr. Sumit Sar [3]
[1]M. Tech Scholar, [2]Professor, [3]Associate Professor
Computer Science and Engineering
Bhilai Institute of Technology, Durg

***Abstract*— A tremendous amount of data is gathered every day in the world in which we live. It is imperative that such data be analyzed. In this highly innovative era of fierce competition to surpass everyone, the company plan needs to consider the present environment. Modern businesses are built on innovative ideas because there are so many prospective customers who aren't sure what to buy or not buy. Divide consumers who may be relevant for advertising according to factors like gender, age, interests, and other purchasing patterns is known as customer segmentation. Any organization's primary goal is to identify its core customers and understand how their buyers behave and utilize its products. Additionally, each consumer may utilize an organization's goods in a unique way. We're trying to solve the issue of listing this organization's buyers to describe the constructive actions and methods such customers use the company's products for. In addition, companies who work in this industry are unable to pinpoint the possible customers in the target market. In order to find the hidden patterns in the data and make better decisions, machine learning is used in this work. The customer segmentation process employing the clustering technique determines which consumer segment to target.**

One common method in unsupervised machine learning is customer segmentation. We have suggested a solution in this research that makes use of K-Means clustering, a powerful method for dataset clustering. With the elbow method, the ideal clusters are found. After visualizing the data, the strategy is to identify the important characteristics that may be used to categorize the clients and derive some conclusions. Created clusters assist the business in focusing on certain clients and promoting material to them on social media platforms and marketing campaigns that truly interest them.

***Index Terms*— Machine learning, Customer segmentation, K-means algorithm, Elbow Method.**

1. INTRODUCTION

Since many businesses today operate online, online marketing is becoming increasingly important to retain customers. However, in the process of doing this, treating every customer the same and targeting them all with the same marketing strategy is not only ineffective, but it also irritates customers by ignoring their uniqueness. As a result, customer segmentation is growing in popularity and has emerged as an effective solution to this ongoing issue.

Over time, the proliferation of knowledge mining approaches to extract essential and strategic information concealed in the datasets of the companies has been facilitated by the availability of large-scale historical data and the growing competitiveness among firms. Data filtering is the process of removing logical features from a dataset and presenting them in a way that is comprehensible to people in order to aid in decision-making. Techniques for processing data set apart fields such as data systems, artificial intelligence, machine learning, and statistics. Applications for data processing encompass a wide range of fields, such as finance, biology, meteorology, fraud detection, and consumer segmentation. Divide a company's client base according to behavioral (product categories ordered, annual income) and demographic (age, gender, and marital status) factors. This is known as customer segmentation. Behavioral features are a superior method for customer segmentation because they focus on individuality and allow us to properly segment based on preferences that may differ among age groups, whereas demographic variables do not highlight the individuality of customers.

The current study focuses on data mining for the purpose of identifying client segments in the commercial company. The process of classifying

customers into groups known as customer segments based on particular business-specific criteria is known as customer division. Each customer segment is made up of customers who have comparable market characteristics.

These differences include supporting factors that will have an indirect or direct impact on the market or business, such as product preferences or expectations, location, behavior, and so on. The ability for a company to tailor marketing strategies that are suitable for each customer segment within its customer base, as well as the ability to support riskier business decisions such as debt relationships with clients, are just a few examples of the importance of customer segmentation. Accurately predicting customer decline, identifying products associated with particular components and supply chain management, exposing the exchange and interdependence between customers, products, or customers and products, identifying high-risk customers, considering additional marketing research questions and providing solutions, and so on.

According to [1], data tuples are regarded as objects by clustering algorithms. The data objects are divided into groups, or clusters, according to how similar or different the things in a cluster are from the objects in other clusters. A customer's base can be divided into multiple groups, or customer segments, based on shared features. This process is known as customer segmentation. Based on similarities in several aspects that are pertinent to marketing, including age, gender, interests, and various purchasing habits, the population is divided into segments. The primary objective of this work is to identify consumer categories through data mining using the K-means clustering algorithm, a partitioning method. The elbow technique is used to determine the optimal clusters.

1.1 RESEARCH OBJECTIVES

Here, the objective is to pinpoint client groups in order to take the appropriate steps to increase the business's revenue. For example, a certain subset of consumers may earn a lot of money yet have a low spending score (the amount they spend in the mall). Using tactics like better marketing, customer feedback, and product development, we can convert these kinds of consumers into functional customers, whose spending score is excessive, based on their rating. The suggested approach locates clusters depending on several parameters in order to identify such clients.

The most important problem for any commercial organization is statistics. We can carry out various operations to determine customer interests with the aid of grouped or ungrouped facts. Data mining can be useful to extract statistics from the database in a human-readable format. However, it's possible that we won't identify the true beneficiaries across the board in the dataset. The primary factors used in customer segmentation are the consumer's age, demographics, spend score, income, gender, and so forth.

This allows us to deduce things like which products generate enormous earnings for the business or which age group shops the most, among other things. As a result, the corporation primarily targets the demographics and goods that have the greatest potential to bring in the most money.

2. LITERATURE REVIEW

[2] A strategy is suggested to use machine learning techniques called NEM, LiRM, and LoRM to separate the client group into two groups: premium and standard. Reference [3]. Vishal Singh, Tanupriya Choudhury, Suraj Bahuguna, and Tushar Kansal. "Determining Customer Segmentation Through K-means Clustering," International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).2018, In this work, customer segmentation on Telecom consumers is achieved through the use of data such as age, interest, etc., through the cluster analysis approach.

A. *Customer Classification*

As businesses strive to improve their operations, draw in new clients, and satisfy their current clientele with their goods and services, the commercial world has grown increasingly competitive over time. [4] Determining and defining each customer's demands and wants is an incredibly meticulous task. The rationale is that different consumers have different requirements, wants, and preferences based on factors including size, taste, geography, and product attributes. It also turns out that it is not a good idea to treat every customer the same. This study's foundation is the idea of consumer or market segmentation, which divides consumers into smaller groups or segments and encourages members of each subcategory to avoid engaging in identical behavior. It appears that the process of splitting the market into various target sectors is called customer segmentation.

Customer segmentation is a tactic for breaking the market up into homogenous segments, claims [5]. Based on many factors such as geographic location, economic status, demographics, and behavioral patterns, the data used in the customer segmentation technique is used to form customer groups. By using the customer segmentation technique, a firm can demonstrate a deeper understanding of the demands of the client, make better use of its marketing budget, and obtain a competitive advantage over other

businesses. It also supports the development of a more robust brand strategy, the identification of client retention, new market potential, and increased marketing efficacy for a firm.B. *Data Repository*

One can evaluate results and identify relevant solutions by collecting and comparing data against specific system improvements. This is known as data collection. [6] Data gathering is a necessary component of research in all fields, including the social and physical sciences, the humanities, and business. The goal of any data collection is to provide high-quality evidence that directs analysis and yields clear and substantial answers to the provided questions.

C. *Clustering the data*

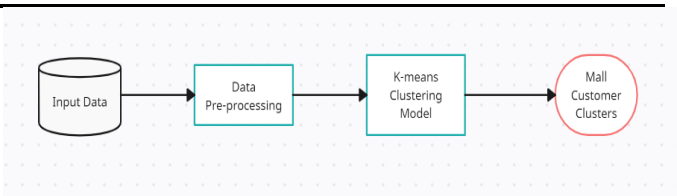
The process of organizing data according to certain traits or attributes is called clustering. Depending on the given circumstance, a variety of algorithms can be applied to datasets.[7] However, no single algorithm can be used to solve every clustering issue, therefore selecting the right clustering strategies becomes essential.

D. *K Means*

K-means suggests that a particular classification algorithm is one of the most popular ones. This clustering technique, called Centro, assigns each data point to one of the overlapping groups that have already undergone K-algorithm pre-sorting. Groups that correspond with hidden patterns in the data are created in order to aid in identifying the execution process. K-means can be assembled in a variety of methods; here, we'll employ the elbow technique[8].

3. METHODOLOGY

The data set from a retail mall was gathered and utilized to apply the Kmeans and clustering algorithms. The 200 tuples in the data collection, which represents the information of 200 clients, are composed of 5 attributes. CustomerId, Gender, Age, Annual income (k$), and Spending score on a scale of (1–100) are among the attributes in the data collection. Initially, investigated the dataset and discovered some information that may help our model. Subsequently, eliminated duplicate records, handled null or NAN values, and sanitized the dataset. Next, we visualized our data and extracted significant information from it using visualization tools. To prepare the data for fitting the model, preprocessing was done, which included feature scaling. Prior to putting the K-means method into practice, K-means clustering model was constructed, and a scatter plot was used to show the output groupings.



4. TECHNICAL INTRODUCTION

k-means Clustering Algorithm:

It is a kind of unsupervised method designed to address clustering issues Its method categorizes a given data set in an easy-to-understand way by using a predefined number of clusters—let's say k. In comparison to peer groups, data points inside a cluster are varied and homogeneous.

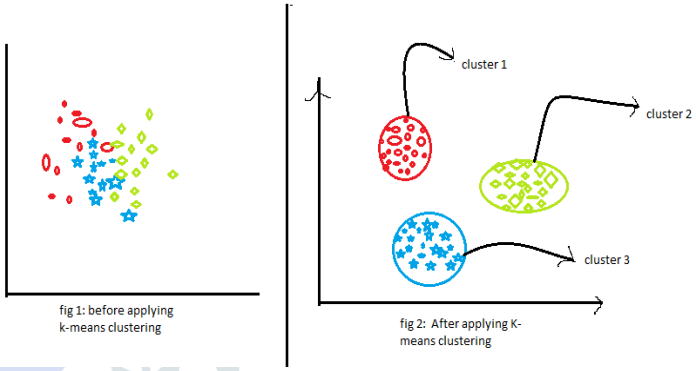


fig 1: before applying k-means clustering

fig 2: After applying K-means clustering

The mechanism of the k-means clustering algorithm is to group data points according to their similarities into clusters. The algorithm takes input data points, their attributes, and their data set. Based on similarities, the algorithm creates K clusters. Kmeans employs the Euclidean distance measuring method to calculate the similarity.

The steps listed below are how K-means creates a cluster:

1. Initialization: First, the data space is initialized by extracting the k centroids after the K has been determined.

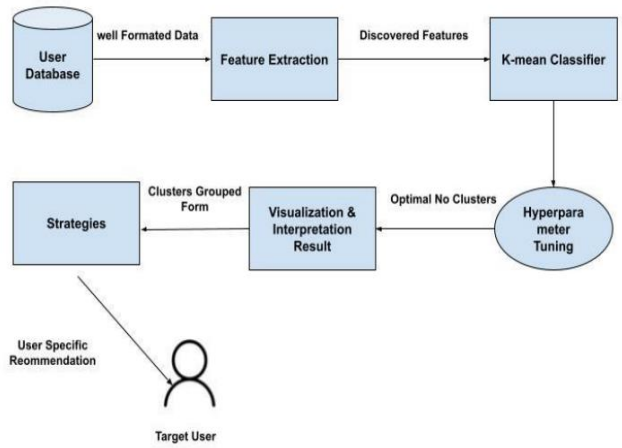
2. Secondly, the objects are assigned to the centroids: Each object in the data is assigned to the nearest centrroid.

3. Updating the Centroids: For each group, the new centroids are updated using the average position of the items.

4. Cluster Assignment: The cluster assignment process is now complete. After the assignment is finished, a procedure is put in

place to automatically modify the mean value for every cluster in the data.

5. PROPOSED MODEL



A) Data Gathering:
The required data is first taken from the database, formatted (all NA values are removed), and made ready for processing.

B) Feature Extraction:
Selects features that improve the correctness of the model; for this example, the features are the annual income and the spending score for efficient analysis.

C) K-means Classifier:
Following that, K denotes the classifier's performance of clustering in relation to the characteristics supplied to it.

D) Hyper Parameter Tuning:
We used hyperparameter tweaking, which is accomplished using the Elbow technique, to determine the optimal number of clusters while building groups.

E) Data Visualization:
The marketing team can develop new techniques for more effectively targeting clients in figure 3 using the clusters that have been developed.

6. ANALYSIS AND VISUALIZE THE DATASET

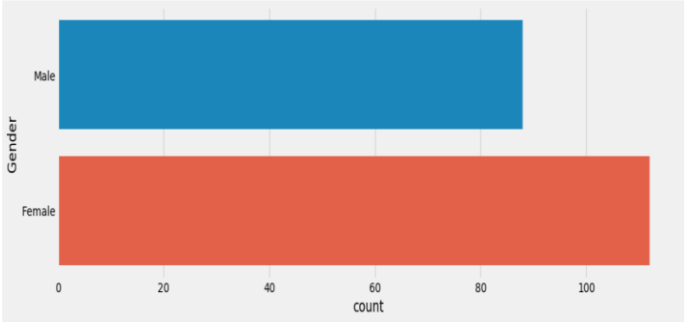
Within this analysis, we comprise that the hypothetical consumer data in the mall customer dataset is an intriguing dataset. You can get a taste of what it's like to run your own store.

Consequently, we must categorize the clients into different categories based on the data.
The characteristics we have here are as follows:

1. Customer_ID: This is the special ID that each customer is assigned.
2. Gender: The gender of the client.
3. Age: The age of the client.
4. Annual Income ($): This is the annual income of the customer.
5. Spending Score(1-100): This is the rating, out of 100, that a consumer receives from mall management depending on how much money they spend and how they behave.

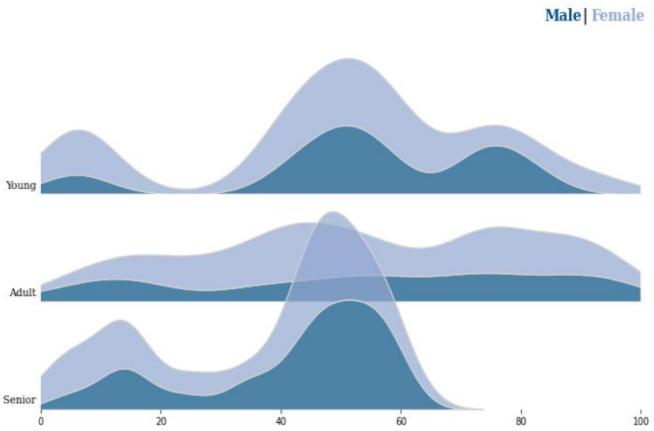
Visualize our dataset:

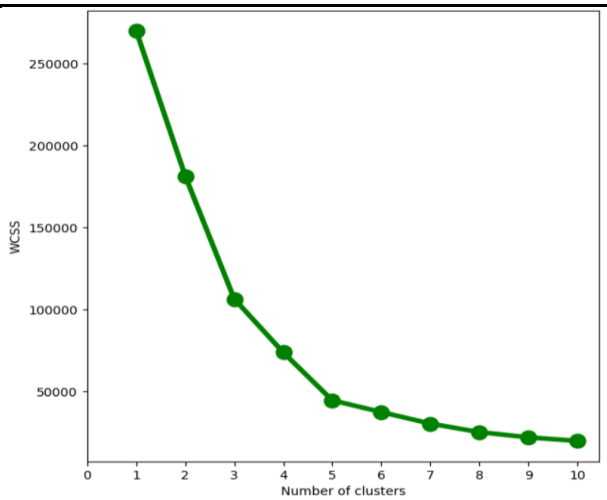
Visualize the gender of customers



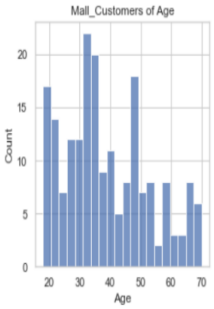
In this instance, it is evident that women shop more than men do.

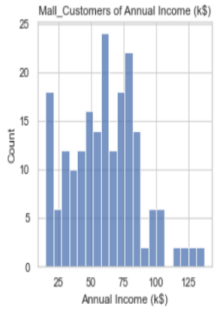
Now we will visualize the dataset to understand the spending score distribution by gender and age range we notice that Interestingly, Seniors have no higher spending scores than 60.

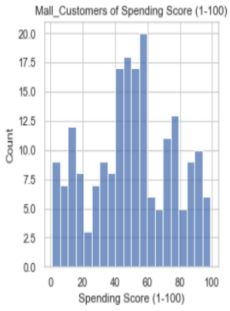




We will now use Seaborn and Matplotlib to show the dataset in order to see how the columns relate to one another. This indicates that compared to persons in other age groups, those in the 20–40 age range shop more. Additionally, compared to others, those with annual incomes between $50,000 and $1,000,000 shop more.







ELBOW METHOD:

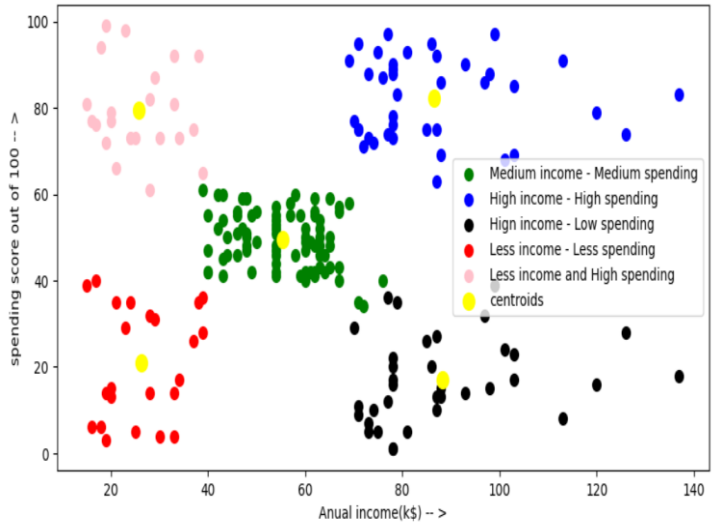
A greater number of clusters can help reduce the overall within-cluster variation within each cluster, which is the basis for the elbow technique. This is due to the fact that having more clusters makes it possible to identify more specialized groups of related data elements. To specify which clusters are ideal The clustering algorithm is first run using a range of values for k. Change k to reflect one to ten clusters in order to do this. The total sum of squares within the cluster is then ascertained. After that, we show the intra-cluster sum of squares based on the number of clusters. The graphic shows how many clusters are approximately required for our model. The graph's bend can be used to determine the ideal cluster locations.

The figure shows that the number of clusters to be chosen is equal to five, despite the curve's slope not being steep enough after it.

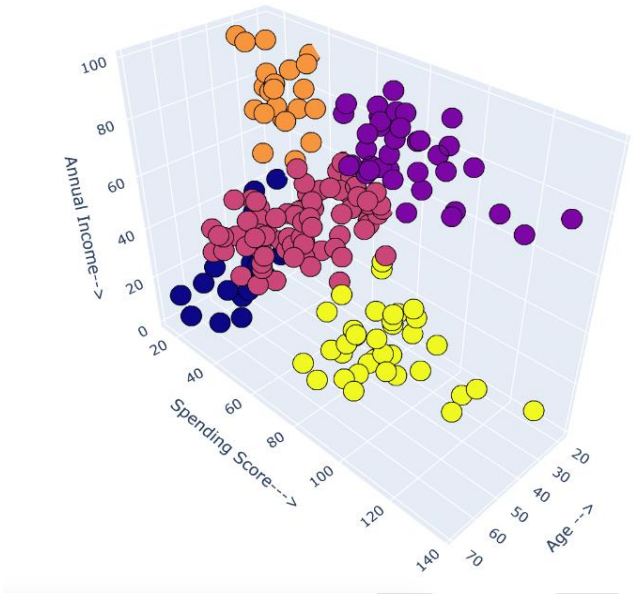
7. RESULT

Following data analysis and customer classification based on attributes like spending score and annual income, we were able to group customers into clusters. From these clusters, the marketing team created recommendations tailored to each individual client.

A spending score vs. annual income curve is used to visualize the data (clusters). Now let's examine the model's output.



Following data analysis and customer feature classification based on age, spending score, and annual income, we were able to create the 3D cluster representation seen in Figure below.



- Cluster 1(Green): Young adults with medium annual incomes and medium spending scores make up this category.
- Cluster 2(Blue): Middle-aged persons with high expenditure and annual income scores make up this category.
- Cluster 3(Black): Middle-aged adults with low expenditure scores and large annual incomes make up this category.
- Cluster 4(Red): Older folks with low annual incomes and spending scores make up this category.
- Cluster 5(Pink): Young adults with high expenditure scores and low annual incomes make up this category.

The following is how we may develop a marketing strategy based on the preceding Cluster Chart Result:

- According to preliminary data, there are more female clients than male customers. A marketing campaign aimed at male consumers could be developed.
- The yearly incomes of Clusters 1 and 3 are medium and high, respectively, yet their spending scores are low or medium. We may thus provide certain deals to entice customers to make more purchases.
- Customers in clusters 2 and 5, who have high spending scores, may receive special treatment or be eligible for loyalty programs.
- Spending score is low for Cluster 3, despite their large yearly income. These are adults in their middle years. To determine their demands, wants, and requirements in order to encourage them to make more purchases, we could conduct marketing research.

8. CONCLUSION

This study shows how the k-Means clustering approach was applied for client segmentation using data collected from an online shop. Our method has separated customers into five groups that are mutually incompatible. Applying additional data mining techniques will be helpful, and the insights that are obtained are beneficial for business wings' decision-making.

The following are the clustered found using K-Means clustering-

Cluster 1 displays the clients with average expenditure and salary scores, allowing us to classify them as Standard clients.
Cluster 2 displays the high-earning and high-spending customers. The mall owner may find that these customers are also the most profitable, allowing us to classify them as target customers.
Cluster 3 indicates that they have a high income but modest spending so we can classify the consumer as careful customers.
Cluster 4 can be categorized as sensible clients because it has both low income and moderate spending.
Cluster 5 displays the low-income consumers who spend a lot of money, making them fit the definition of careless consumers.

Customer segmentation is carried out using the company's customer data, and the K-means clustering machine learning technique is employed to split clients according to characteristics such as annual income and total spending. This study also demonstrates that segmenting customers based on behavioral characteristics is a better way to address the current customer segmentation issue, and the K-means clustering algorithm is recognized as a suitable option for this strategy.

FRENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques", Third Edition.

[2] Sukru Ozan, "A Case Study on Customer Segmentation by using Machine Learning Methods", IEEE, Year: 2018.

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

[4] Stone and Bob, Successful Direct Marketing Methods[M]. 4 ed. NTC Business Books. 1989:10ã110.

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[6] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.

[7] Management Science: A Comparative Research on the Methods of Customer Segmentation Based on Consumption Behavior. 2003.2, Vol.16.Google Scholar.

[8] Aman Banduni, Prof Ilavendhan "Customer Segmentation Using Machine Learning", IJCRT, Year: 2022.

[9] I. Pranata and Geo Skinner. "Segmenting and selecting customers through clustering selection & analysis". In: Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on. IEEE. 2015, pp.303–308.

[10] Minghua Han. "Customer segmentation model based on shopping consumer behavior analysis". International Symposium on. IEEE. 2018,pp.914–917.

[11] Ardabili, S.F.; Mosavi, A.; Várkonyi-Kóczy, A.R. Systematic Review of Deep Learning and AI Models in Biofuels Research. In Proceedings of the 18th International Conference on Global Research and Education,Budapest,Hungary,4–7Sept2019.

[12] Jiang, T., Tuzhilin, A., March 2009. Improving personalization solutions through optimal segmentation of customer bases. IEEE Trans. Knowledge Data Eng. 21 (3), 305-320. https://doi.org/10.1109/TKDE.2008.163N.

[13] He X., Li, C., 2016. The research and application of customer segmentation on e-commerce websites. In: 2016 6th International Conference on Digital Home (ICDH), Guangzhou, pp. 203–208. Doi: 10.1109/ICDH.2016.050.

[14] Hung, P. D., Lien, N. T. T., & Ngoc, N. D. (2019, March). Customer segmentation using hierarchical agglomerative clustering. In Proceedings of the 2019 2nd International Conference on Information Science and Systems (pp. 33-37).

[15] Zadoo, Ankita, et al. "A review on Churn Prediction and Customer Segmentation using Machine Learning." 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-ITCON). Vol. 1. IEEE, 2022.

[16] K. Torizuka, H. Oi, F. Saitoh and S. Ishizu, "Benefit Segmentation of Online Customer Reviews Using Random Forest," 2018 IEEE International Conference on Industrial Engineering nd Engineering Management (IEEM), Bagkok, Thailand, 2018, pp. 487-491, doi: 10.1109/IEEM.2018.8607697.

[17] Shaik, Anjaneyulu Babu, and Sujatha Srinivasan. "A brief survey on random forest ensembles in classification model." International Conference on Innovative Computing and Communications. Springer, Singapore, 2019.

[18] Marcus, Claudio. "A practical yet meaningful approach to customer segmentation." Journal of consumer marketing (1998).

[19] Yogita Rani and Dr. Harish Rohil "A Study of Hierarchical Clustering Algorithm", IJICT, Year: 2013.

[20] Omar Kettani, Faycal Ramdani, Benaissa Tadili "An Agglomerative Clustering Method for Large Data Sets", IJCA, Year: 2014.

[21] Rivedi, A., Rai, P., DuVall, S. L., and Daume III, H. (2010, October). ´Exploiting tag and word correlations for improved webpage clustering in Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 3-12). ACM.

[22] Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6).