

The practical approach in Customers segmentation by using the K-Means Algorithm

E.Y.L Nandapala *, K.P.N Jayasena[†]

Department of Computing and Information Systems

Sabaragamuwa University of Sri Lanka

Email: *yashodhalakmini31@gmail.com, [†]pubudu@appsc.sab.ac.lk

Abstract—The partnership between businesses and consumers is increasingly crucial with technological growth. It is necessary to manage this relationship for the company's future growth. The communication mechanism between companies and customers is called Customer Relationship Management (CRM). The CRM plays a significant role in the business sector. Moreover businesses may classify client attitudes, characteristics, etc following the appropriate CRM method. By using this knowledge, companies can determine which consumers are the most profitable. This can be done with the method called customer segmentation. The Customer segmentation is a breakdown of the broad database of customers into subparts. The subparts members have some common features, although, in other subparts, these features are different. Most scientists consider various ways of segmenting customers with multiple clustering algorithms in different fields with the support of data mining. In the data analytical field, there are many techniques associated with the segmentation process. To order to identify the target consumer audience, companies use traditional market segmentation. It ensures that marketers will target their marketing strategies on the most likely to attract customers. Also, businesses may adopt more effective marketing strategies following the successful customer segmentation process and thereby reduce the uncertainty associated with an investment. Here, the k-mean was used as an algorithm for the segmentation process to segment the customers. In this study, determine the six clusters base on the annual income and spending score. Also, identify six clusters according to the two principal components. Following the successful clustering process, lead the organizations to make accurate decisions and organizations can deliver the new product and services, and also can do some changes for existing product services according to the customer requirements through the identification of customers correctly.

Keywords: Customer Relationship Management, K-Mean Algorithm, Customer Segmentation, Data Mining,RFM

I. INTRODUCTION

Through the advancement of technology, business-customer relationships have become more important. Managing this connection is critical for the future growth of the business. The management process between companies and customers is called Customer Relationship Management (CRM). In the business sector, CRM act as a significant role [1]. However, Using CRM process, companies can identify the customer's behaviors, characteristics, etc [2], can identify Who are the most profitable customers [3]. Customer segmentation is a breakdown of the large database of customers into sub-parts. Members in the sub-parts have similar characteristics and those features are dissimilar to other members in other sub-parts [4]. The Customer segmentation is separated in connection

with demographic (age, sex, religion, size of the family, and more), compliance (consumption, spending, and more), geographic (Where and when they work), and psychographic (Social, lifestyle, and more). The segmentation process has many benefits to it. Others are, more customer service, better connection with customers, best performing economies, improve productivity, improves distribution channel, sets the brand identity, lead to optimizing prices, etc. So, an Insurance company can easily find the details about their customers as they wish, using these customer segmentation processes [5]. Not only an Insurance company, but Customer segmentation can also be used for various sectors. A few sectors are the telecommunication sector, financial sector, sanitation sector, e-commerce sector, and marketing sector, etc [6]. Some researchers with the support of k-means clustering and Kohonen's self-organizing map, show how the customer segmentation support for financial institution [7]. Within the technology development, the telecom market also developed. However, the customer's requirements are also changing and they find new services from the telecommunication sector. To provide a new service that needs to identify the customer requirements correctly. Because of that some researchers are propose clustering techniques under the k-means clustering algorithm and multi-variable quantum shuffled frog leaping algorithm [8]. To obtain the highest marketing advantages in the e-commerce website field can be used as the three-dimensional customer segmentation model [9].

By Using the customer segmentation with a clustering algorithm can obtain a successful result from the segmentation process. Following the accurate method for customer segmentation is very critical because of the accuracy of the method directly affects the success of the segmentation results. Machine learning is among the Artificial Intelligence sub-areas. Machine learning has the ability to identifying the patterns from the data, based on its experiences. Results of that ability, machine learning can predict the results accurately without programming. For machine learning, the primary assumption is to implement the algorithms which can bring the data as input, and with the support of the statistical hypothesis, machine learning algorithms can predict the output results by updating the output through the arrival of the new data. Machine learning allows for the study of large amounts of data. Machine learning has two types of algorithms. Supervised machine learning is one type and another type is unsupervised machine learning algorithms [10]. From the list of better

clustering algorithms, under unsupervised learning, K-mean clustering is most suitable for cluster the dataset. Maintaining consumer loyalty and customer attention span are major problems facing the business sector today. With the support of the k-means algorithms, can easily find who is the most profitable customer based on the respective clusters [11].

The main problem addressed in this paper is finding the best way that can be used to segment the customer dataset efficiently and effectively. The segmentation makes it easier for management. Any organization can select what type of consumers they required to be considered and targeted. The collection of client classes that will be categorized as best clients, mediocre clients, average clients, and other clients [12]. K-mean is used as the clustering algorithm for cluster the dataset and the main objective in this paper is to find the efficient way within K-mean clustering for the segment the dataset accurately.

The way this paper was organized as follows. Section 2 presents the related work and Section 3 presents the methodology and experimental design. Section 4 presents results and discussion. Section 5, presents the conclusion of this paper.

II. RELATED WORK

A vital number of papers have been studied relating segment the customers in several areas. A broad range of methodologies has been recognized as well as objectives. It seems that methods used to produce segmentation predictions fundamentally within the support of different technologies. Here presents the related work for the customer segmentation within Kmean clustering.

Many researchers used the customer segmentation process for various areas. The main priority in the banking sector is to improve customer segmentation and take it into account in the design and distribution of new products. Introduce the loyalty program, that includes the issuance of various types of cards for such customers in this paper [3]. Internet banking customers are growing rapidly. Based on Internet banking data, customer segmentation can be applied. Build clustering models on customer profile data based on their usage of Internet Banking in XYZ bank [13]. By applying the data mining process, the online shop can identify customers, Therefore customers can get special facilities in the appropriate marketing strategies according to their needs [14]. With the rapid development of the telecom market, telecom customers show different characteristics. In that market, a customer segmentation method can be used as the proposed cluster algorithm in this paper [8]. In this paper [15] discussed the usage pattern-based customer segmentation method is proposed to measure the heterogeneity of customers by the smartphone's usage pattern dissimilarity . For customer segmentation, many researchers use various methods. K-Means Algorithm is used for the segmentation process in many papers. TF-IDF is used as a Document Preprocessing method, K-Means as a clustering method, and the elbow method used to optimize the number of clusters [16]. Many researchers compare the different algorithms with the k-means algorithm. Using the K-Means

method and K-Medoids method based on their RFM (Recency, Frequency, Monetary) score of internet banking transaction and compare the performance of both clustering methods in customer segmentation [13]. Different clustering algorithms will be compared, specifically centroid-based clustering K-Means, CLARA, and PAM with Fuzzy CMeans clustering [17].

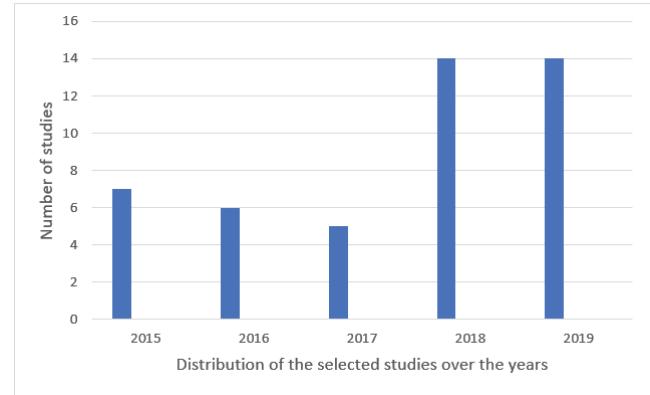


Fig. 1. Customer Segmentation with timewise

In here Figure 1, visualize the 46 studies over the year between 2015 to 2019. According to figure 1, the number of researches is increasing within the period (near to 2018 and 2019) under the customer segmentation field.

III. METHODOLOGY AND EXPERIMENTAL DESIGN

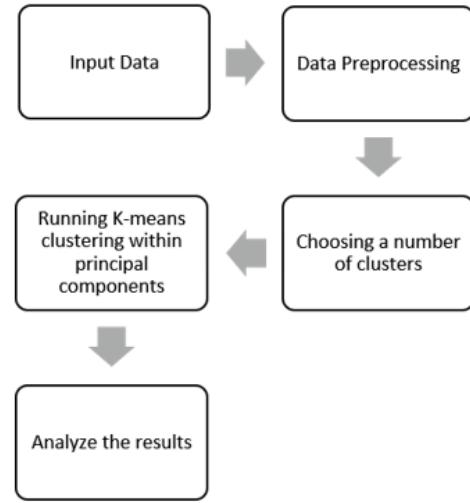


Fig. 2. Research method

Figure 2 illustrates the research method used in this research. It consists of five steps followed as Data acquisition, Data Pre-processing. Data analyzing using K-means within

principal component, choosing a number of clusters and analyze the results.

A. Data Preprocessing

In the data pre-processing step, first checked the null values from the dataset. And also, critical to identify the outliers in the dataset. Then remove all null values and outliers and this step called as Data Preprocessing step.

B. Choosing several clusters

We used K-means clustering algorithm for cluster the dataset.

The K-means clustering algorithm has several steps. The first step is to determine how many clusters that need to be generated. Then randomly select the k objects and that objects be the initial clusters. After that, assign data points into the centroid. The distance between data points and centroid is measured based on the Euclidean distance [18]. According to the results of those measurements, the points are assigned to the nearest centroid. Next, update each centroid by calculating the value of the mean of every data point that stays in the cluster. Then again calculate the distances and assign data points into the new centroid. Like that, these previous two steps are done iteratively. The stopping condition for this process is the centroid remains the same, the data points assigned to the specific centroid remain the same and the fixed number of iterations is reached [13]. Here summarized the 4 main steps in the k-mean: Select initial centroids at random, assign each object to the cluster with the nearest centroid, compute each centroid as the mean of the objects assigned to it, Repeat the previous two steps until no change.

This paper used three types of methods to determine how many clusters that need to be specified.

1) *Elbow Method*: Decrease the value of the intra-cluster is the main purpose of the k-mean algorithm. The number of clusters(k) that determined using k-means, can be used as the input for this elbow method. Then measure the square number of intra-clusters. Intra-cluster distance is the distance between each data point to the corresponding centroid in each cluster center. Here need to measure the square number of intra-cluster for every number of clusters(k) that is determined in the beginning as input [19]. After measuring all values can be a plot that values according to the k values. The point that has the curvature in this plot, represents the optimum number of clusters [20].

2) *Silhouette Method*: With the support of the silhouette method, can appraise the validity of the number of clusters accurately and efficiently. According to the silhouette method, the silhouette width has the value range. That range is between minus one to a positive one [21]. If the cluster is not classified well, then the silhouette width close to minus one according to that range of value. And also, the cluster is classified correctly then the silhouette width close to a positive one. If the value is equal to zero, the data points assigned to the cluster are not suitable for the corresponding cluster. That points are most suitable for another cluster. The maximum average value for

the silhouette width represents the better cluster process and according to that value, the optimal number of clusters (k) can be represented.

3) *Gap Statistic Method*: In this method, the accumulation of the intra-cluster variance for the number of clusters(k), compares within their expected value under the null reference distribution. The optimal number of clusters(k) are displayed in the point that has a maximum gap statistic [19]. For any clustering algorithm, this method can be used.

C. Running k-mean clustering within principal components

Here running the k-means clustering results within the support of the principal components. As a result of this, can visualize the data clusters.

D. Analyze the results

In this step, analyze the cluster results within visualization and recognize the way that arrangement of the dataset.

IV. RESULTS AND DISCUSSION

Visualizing the dataset according to various features using different types of plots. First, visualize the gender using the bar plot. Figure 3 shows that the female count is higher than the male count.

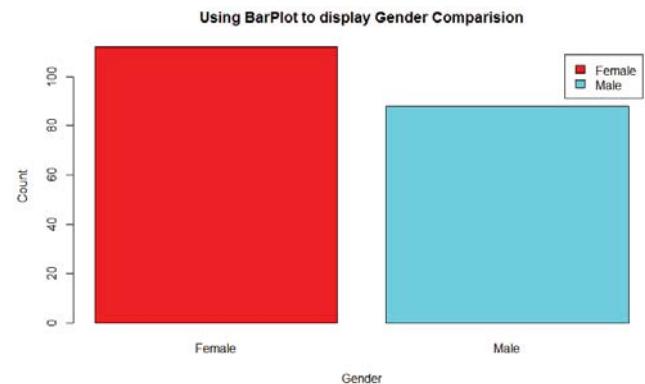


Fig. 3. Gender comparison

Then visualize age class using the histogram. Figure 4 shows the highest age range in the dataset. According to that, the 30-35 age class has the highest frequency.

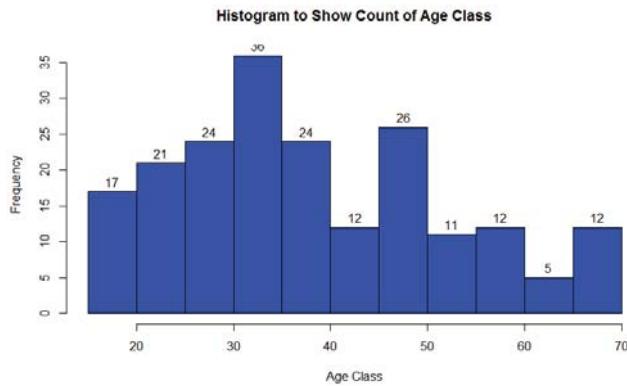


Fig. 4. Count of age class

After analyzing the age using boxplot and figure 5 shows the highest age class clearly.

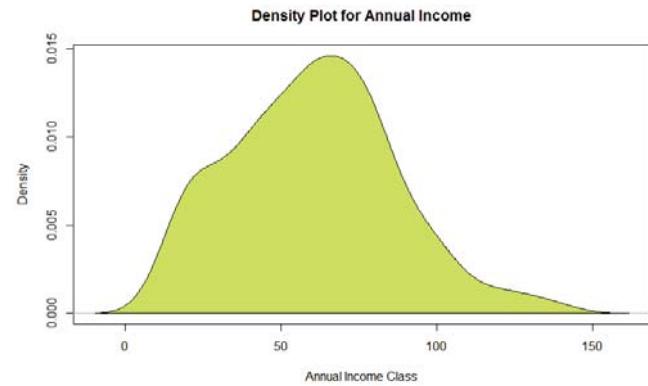


Fig. 7. Density plot for annual income

Visualizing the spending score here used boxplot and figure 8 shows that plot.

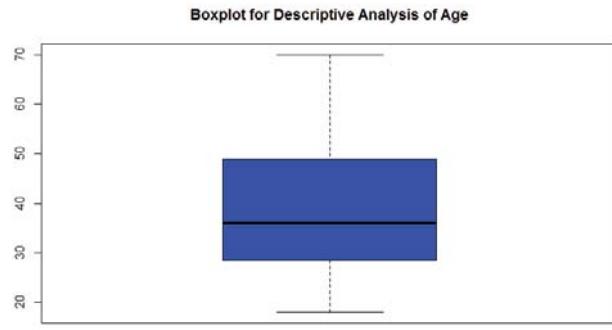


Fig. 5. Analysis of age

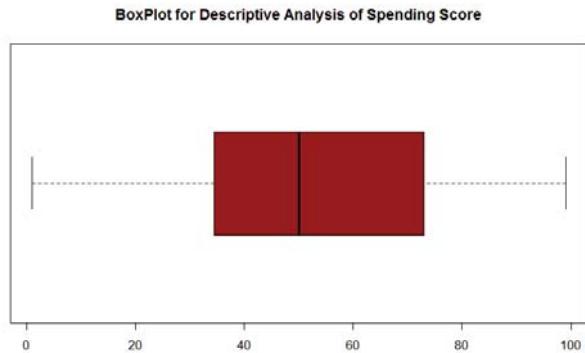


Fig. 8. Spending score

Then used a histogram to visualize the annual income.

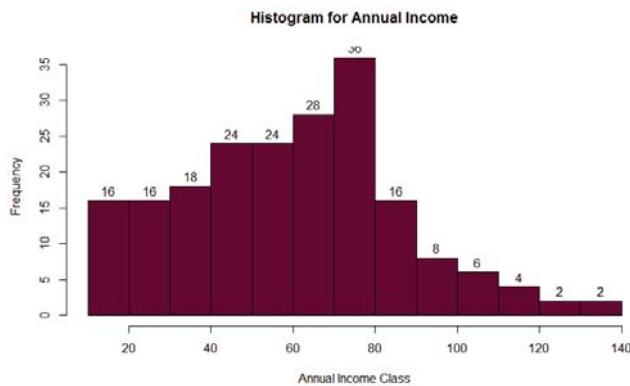


Fig. 6. Annual income

Figure 6 shows 70-80 income classes have the highest frequency and also we can visualize it using a density plot. Figure 7 represents the density plot.

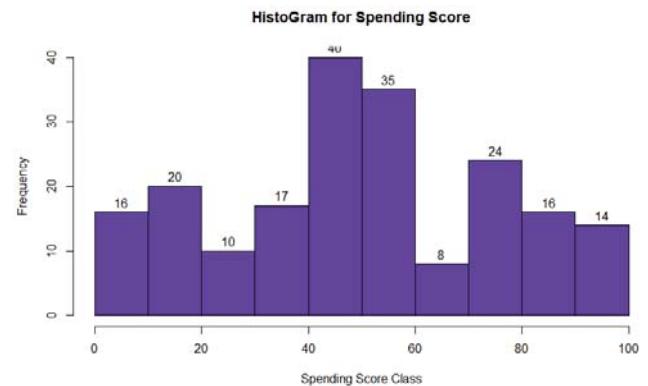


Fig. 9. Histogram for Spending score

According to figure 8 and figure 9, we can say the 40-50 spending score class has the highest frequency than other score classes. After this, used the elbow method to optimize k.

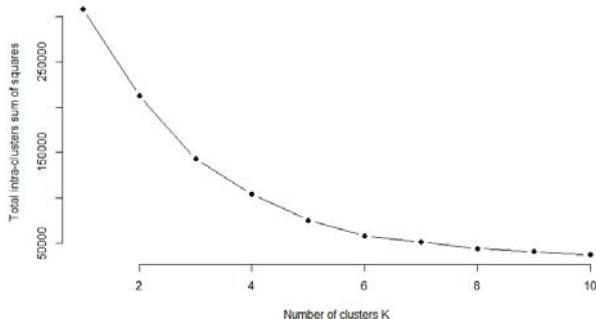


Fig. 10. Elbow method

Figure 10 shows the optimal value for the k, according to an elbow method is 4. Then used the average silhouette method to optimize the k value. Here we used 2-10 values for k respectively. Figure 11 display the variation of the k in the silhouette plot when the k value is 2.

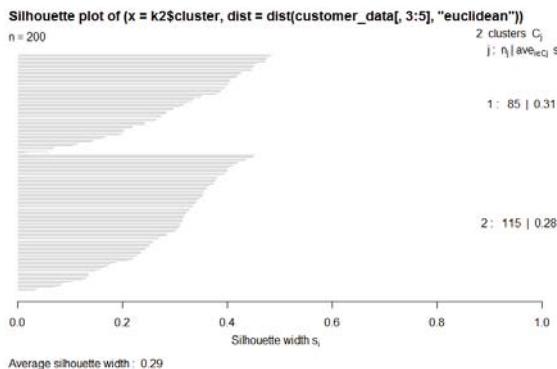


Fig. 11. The silhouette method plot when k=2

Finally, find the optimal k value using the silhouette plot. Figure 12 shows the variation of the k value.

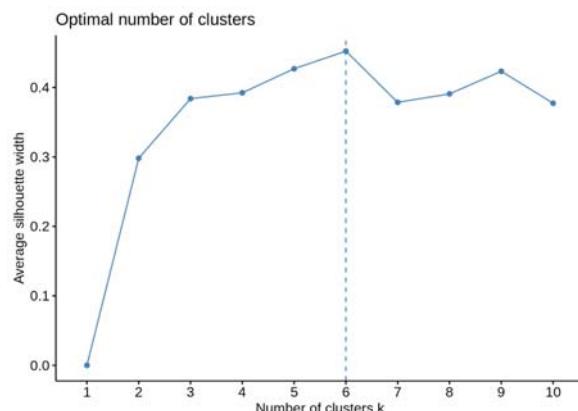


Fig. 12. K value in the silhouette plot

Then used the gap static method to find the optimal k value. Figure 13 shows the variation of the k value according to the gap static method.

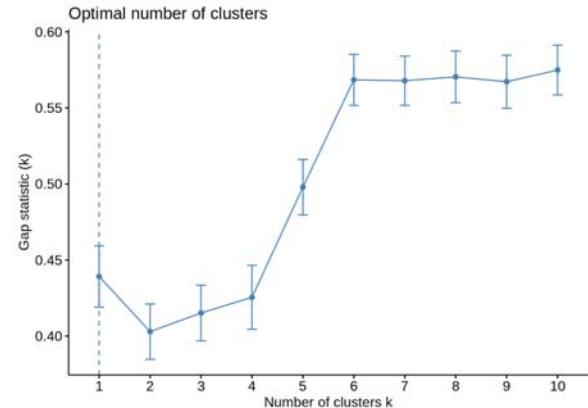


Fig. 13. K values in gap static method

According to the above three methods (Elbow, Silhouette, Gap Statistic methods), the best value for k is 6. Then visualize the clustering results with principal components.

After that visualize the clusters. According to figure 14, cluster 1 represents high annual revenue customer data as well as high annual spending. Cluster 2 represents high annual sales and low annual spending. Low annual revenue as well as low annual income spending is represented in cluster 3. Consumer data within a medium wage and total annual compensation spending represent by cluster 4. Cluster 6 represents the same data as cluster 4. The 5th class reflects low annual sales but high annual spending.

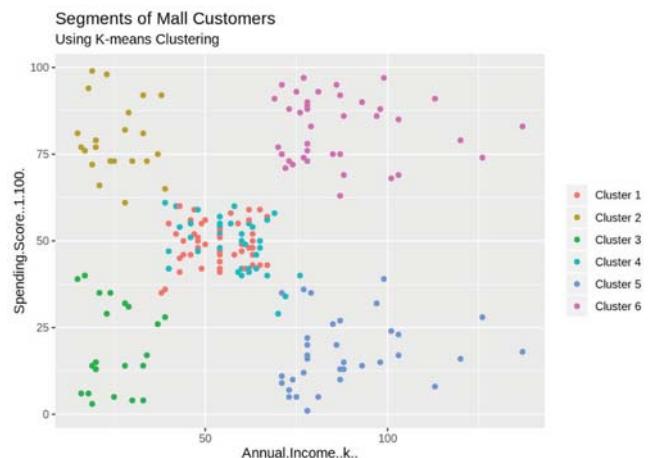


Fig. 14. Clusters

After that visualize the clusters according to the two principal components.

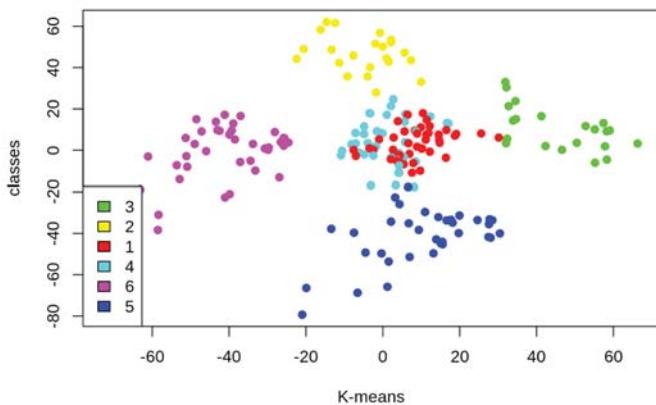


Fig. 15. Clusters according to principal components

According to figure 15, such two clusters (clusters 1 and 4) consist of low score PCA1 and low score PCA2 customers. Cluster 2 involves clients with a maximum PCA2 and clients within minimum annual sales expenditure. Clients who have maximum PCA1 income within minimum PCA2, represented by cluster 3. The moderate value of PCA1 within the minimum score of PCA2 consists of cluster 5. Cluster 6 shows the clients who have maximum PCA2 within clients who have minimum PCA1.

V. CONCLUSION

The support of the accurate clustering process we can efficiently to identify the customers very clearly and determine the features deeply. And also, following the successful clustering process, lead the organizations to make accurate decisions and organizations can deliver the new product and services, and also can do some changes for existing product services according to the customer requirements through the identification of customers correctly. Within the way that introduces in this paper, can be used as an accurate segmentation process for the segment of the dataset in any sector.

As future work, we would like to carry out experimentation for real-world workload traces and improve the algorithm. We would also like to carry out experimentation of the proposed algorithm for workflow planning.

REFERENCES

- [1] A. Sharma, H. Jit, and S. Sidhu, "Customer Relationship Management Using Clustering And Classification Technique," vol. 20, no. 5, pp. 67–72, 2018.
- [2] O. Sharma, "A New Activation Function for Deep Neural Network," *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, pp. 84–86, 2019.
- [3] V. Mihova and V. Pavlov, "A customer segmentation approach in commercial banks," *AIP Conference Proceedings*, vol. 2025, no. October 2018, 2018.
- [4] M. Hosseini and M. Shabani, "New approach to customer segmentation based on changes in customer value," no. November, 2018.
- [5] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," *Procedia Computer Science*, vol. 144, pp. 277–290, 2018. [Online]. Available: <https://doi.org/10.1016/j.procs.2018.10.529>
- [6] J. N. Sari, L. E. Nugroho, R. Ferdiana, and P. I. Santosa, "Review on Customer Segmentation Technique on Ecommerce," no. July 2018, 2016.
- [7] P. Gončarovs, "Using Data Analytics for Customers Segmentation: Experimental Study at a Financial Institution," *59th International Scientific Conference on Information Technology and Management Science of Riga Technical University, ITMS 2018 - Proceedings*, pp. 1–5, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8552951>
- [8] C. Cheng, X. Cheng, M. Yuan, C. Song, L. Xu, H. Ye, and T. Zhang, "A novel cluster algorithm for telecom customer segmentation," in *2016 16th International Symposium on Communications and Information Technologies, ISCIT 2016*, 2016, pp. 324–329. [Online]. Available: <https://ieeexplore.ieee.org/document/7751644>
- [9] X. He and C. Li, "The Research and Application of Customer Segmentation on E-Commerce Websites," *Proceedings - 2016 International Conference on Digital Home, ICDH 2016*, pp. 203–208, 2017. [Online]. Available: <https://ieeexplore.ieee.org/document/8039109>
- [10] M. M. A. Grau, M. Tajtakova, and D. A. Aranda, "Machine learning methods for the market segmentation of the performing arts audiences," *International Journal of Business Environment*, vol. 2, no. 3, p. 356, 2009.
- [11] K. Bhade, V. Gulalkari, N. Harwani, and S. N. Dhage, "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization," *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8494019>
- [12] J. Panuš, H. Jonášová, K. Kantorová, M. Doležalová, and K. Horáčková, "Customer segmentation utilization for differentiated approach," pp. 227–233, 2016.
- [13] M. Aryuni and E. Miranda, "Customer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering," *2018 International Conference on Information Management and Technology (ICIMTech)*, no. September, pp. 1–9, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8528086>
- [14] L. Zahrotun, "Implementation of Data Mining Technique for Customer Relationship Management (CRM) on Online Shop Tokodiapers .com With Fuzzy C-Means Clustering," pp. 299–303, 2017. [Online]. Available: <https://ieeexplore.ieee.org/document/8285515>
- [15] H. Chen, L. Zhang, X. Chu, and B. Yan, "Smartphone customer segmentation based on the usage pattern," *Advanced Engineering Informatics*, vol. 42, no. October, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1474034619305737>
- [16] D. Marutho, S. H. Handaka, and E. Wijaya, "The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News," *2018 International Seminar on Application for Technology of Information and Communication*, pp. 533–538, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8549751>
- [17] A. Maulina, Isti, "Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering," *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, no. March, pp. 1–6, 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8887739>
- [18] K. Maheswari, "Finding Best Possible Number of Clusters using K-Means Algorithm," no. 1, pp. 533–538, 2019.
- [19] C. Yuan, "Research on K-Value Selection Method of K-Means Clustering Algorithm," 2019.
- [20] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using K-means Clustering," *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 135–139, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8769171>
- [21] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 2321–2782, 2013.