2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) | 978-1-6654-0690-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/UEMCON53757.2021.9666609

Explainable Customer Segmentation Using K-means Clustering

Riyo Hayat Khan*, Dibyo Fabian Dofadar† and Md. Golam Rabiul Alam‡
Department of Computer Science and Engineering, BRAC University
Dhaka, Bangladesh
Email: *riyo.hayat.khan@g.bracu.ac.bd, †dibyo.fabian.dofadar@g.bracu.ac.bd, ‡rabiul.alam@bracu.ac.bd

***Abstract*—Explainable AI has gained popularity in recent years, but the application of it in unsupervised learning is still a few. In this research, explainability was integrated with clustering, an unsupervised method. Customer segmentation is one of the most important aspects in the competitive business world. The most common approach for customer segmentation is clustering, however, assignments of the clusters often can be hard to interpret. To make the cluster assignments more interpretable, a decision tree based explainability was implemented for customer segmentation in this research for small and large datasets. Using the Elbow Method and Silhouette Score, an optimal number of clusters were found, then ExKMC algorithm was implemented for both datasets.**

***Index Terms*—Customer Segmentation, K-means clustering, Explainability, Iterative Mistake Minimization, Unsupervised Learning**

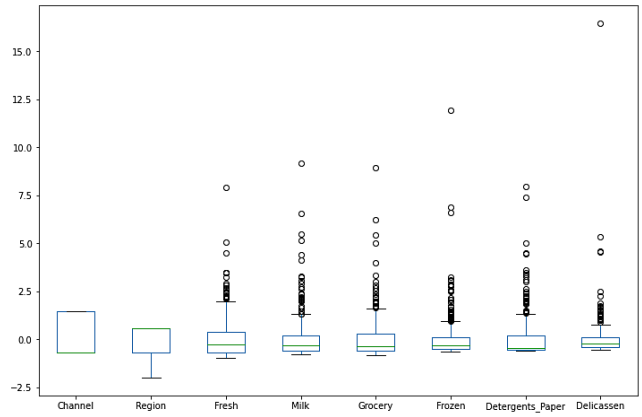
I. Introduction

Customer segmentation is the way of grouping customers based on similar characteristics such as age, location, income, education, gender, spending habits, purchasing behavior and so on [10] [14], so that organizations can distribute a customer base efficiently for promoting their business. It plays a vital role in customer relationship management [10] [14] since it can be used to plan marketing strategies like targeting appropriate audiences for product recommendations. The marketers can study the customer base, identify their desires and understand their thought process, with which they can determine who will be the target customers and how to attract them with the right kind of marketing. For decades, customer profitability and business relationship between companies and customers has been the center of analysis for companies and academics [1]. To build a sound relationship with the customers and to identify their needs, a major step is to conduct a thorough analysis of the customer base.

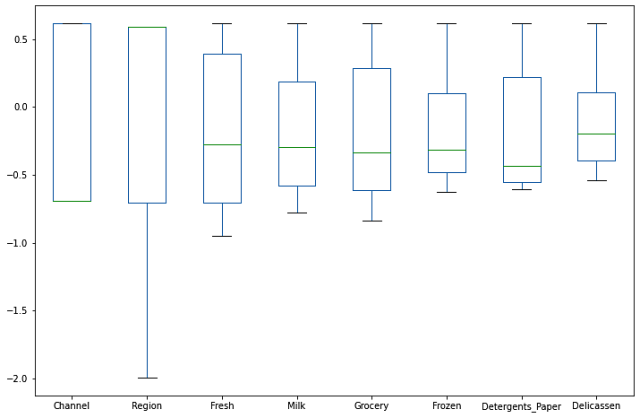
Integration of machine learning such as clustering, a way of learning unsupervised data [6], with customer segmentation has been widely experimented by many researchers. The aim of clustering is to improve the similarity within the clusters as much as possible and at the same time maximize the dissimilarity between the clusters [15]. With the use of K-means clustering, which is one of the most popular unsupervised ML algorithms in customer segmentation, the group of customers can be clustered in groups according to various characteristics and features. Many researchers have used K-means clustering [1] [7] [13] [14] [15] for solving customer segmentation problems since it is one of the simplest clustering algorithms. Other than K-means clustering, various kinds of clustering such as Hierarchical clustering, Density based clustering, Affinity Propagation clustering has also been used for customer segmentation. A comparative analysis of these algorithms has been conducted in [10]. In the analysis, K-means algorithm has shown higher computation speed, less clustering time and capability of handling dynamic data. However, there is a drawback, which is, the optimal number of clusters is hard to obtain. This results in medium level clustering efficiency. For the other three algorithms, computation speed is low and clustering time is high. Hierarchical Agglomerative Clustering (HAC) has been implemented for small datasets in [8]. In [7], the authors have implemented centroid based (K-Means) and density based (Density Based Spatial Clustering of Applications with Noise) clustering.

Although there are several clustering algorithms, the assignments of final clusters can be difficult to comprehend as the clusters may be generated using all of the features of the data. This makes it hard for users to do common characteristics analysis in order to find out why a certain data point is categorized into a particular cluster. To overcome this limitation, explainable AI can be integrated with the model. Explainable AI refers to a set of processes, methods and models that make the predictions of ML models interpretable to humans [9]. Although explainable AI is very popular, most of its implementation has been focused on supervised learning and there has been less work based on explainable methods for unsupervised learning [6].

In case of unsupervised learning, use of decision trees for explainable clustering has been presented in several papers [2] [3] [5] [16]. For a binary threshold tree, each node has a pair of features and threshold which recursively partitions the dataset. The labels on the leaves refer to the clusters. This delivers more information than traditional clustering algorithms for large, high-dimensional datasets [6]. In [11], the authors have used small decision trees to partition a dataset into clusters, with Explainable K-Means and. K-

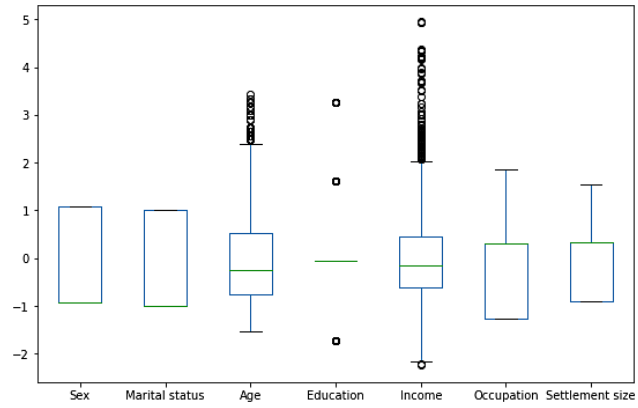


(a) Before Winsorization

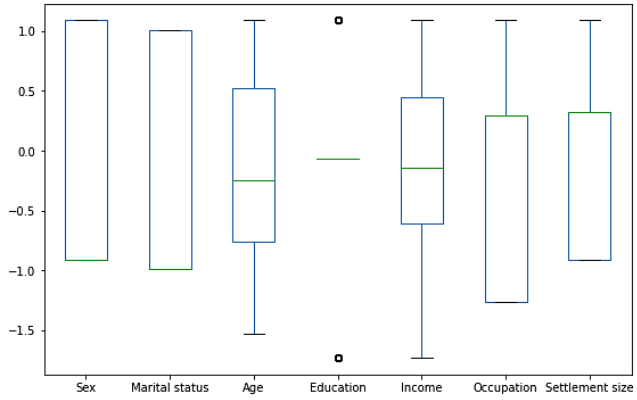


(b) After Winsorization

Fig. 1: Boxplot of Dataset 1



(a) Before Winsorization



(b) After Winsorization

Fig. 2: Boxplot of Dataset 2

Medians Clustering. The algorithm proposed in the paper is called IMM (Iterative Mistake Minimization). It generates a threshold tree which consists of k number of leaves (same as the number of clusters). In order to work with more than k leaves in the threshold tree, the authors have presented an extended version of IMM, naming it Explainable K-means Clustering (ExKMC) [6]. The algorithm initially uses IMM to build a threshold tree, then expands the generated tree to have $k' > k$ leaves. As more thresholds are added, flexibility in the data partition increases.

In this research, the ExKMC algorithm has been implemented for customer segmentation to make the model explainable. Section II describes the datasets and its pre-processing, section III includes the implementation details of this research, section IV visualizes the experimental results with explanation and section V concludes the paper with future works.

II. Data Description & Pre-processing

In this research, the implementation works were done on two datasets. The authors have considered these datasets as small and large datasets respectively. The first dataset [4] refers to the clients of a wholesale distributor where the annual spending in monetary units on different product categories are included. There are 440 non-null data points and 8 attributes. The attributes are: 'Channel', 'Region', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents Paper' and 'Delicassen'. Before implementing the algorithm, some pre-processing steps were followed. First of all, all the features were standardized and checked for outliers. Fig. 1(a) shows the boxplot of the standard scaled data. As this is a small dataset, it was decided not to trim the outliers. To handle this case, winsorization was applied to all the features. This technique helped to limit the upper extreme values by 15% which assisted to deal with the outliers. Fig. 1(b) shows the boxplot of the data after winsorizing.

The same pre-processing steps were applied to the large dataset [12]. This dataset contains the information about the purchasing behavior of two thousand individual customers from a particular area. There are 2000 non null data points along with 8 attributes. The attributes are : 'ID', 'Sex', 'Marital status', 'Age', 'Education', 'Income', 'Occupation'

and 'Settlement size'. Firstly, the column named 'ID' was dropped. Then, similar to the first dataset, the features were standardized and winsorization was applied to the scaled data for handling the outliers. This time 1% of the lower extreme values and 15% of the upper extreme values were transformed during winsorization. Fig. 2(a) and Fig. 2(b) show the boxplot for the features after standard scaling and winsorization respectively for the second dataset.

III. Implementation Details

A. Algorithm Description

The algorithm used in this research was K-means clustering. It is used for partitioning the data points into k disjoint subsets in order to minimize the sum-of-squares criterion. This algorithm tries to find the similarities between data points and groups them into clusters. This whole procedure works in the following steps:

1) Select the optimal number of clusters (k)
2) Initialize the k centroids (cluster centers)
3) Assign each datapoint to the nearest centroids with the use of Squared Euclidean Distance
4) Recompute the centroids by taking the mean of all the data points assigned to that centroid's cluster
5) Repeat steps 3 and 4 until stopping criteria is met

The algorithm is very simple to implement and adapts the new examples very frequently. It is also scalable and faster to large datasets.

B. Finding Optimal Number of Clusters

To implement the K-means clustering algorithm, one of the major challenges is finding out the right values of k. Random values of k can work sometimes, but in most cases it does not guarantee the optimal results of the models. The following two methods are widely used to find the optimal number of clusters:

1) **Elbow Method:** This approach calculates the Within Sum of Square (WSS) values among different ranges of cluster values. As the value of k increases, corresponding WSS starts to decrease. The optimal number of clusters can be decided after analyzing the WSS vs number of clusters graph. In the graph, at some value of k, the WSS curve will decrease abruptly and create a region similar to the elbow of a person. That particular value will be considered as the optimal value of the clusters.
2) **Silhouette Analysis:** This approach calculates the Silhouette score of every data point for each k value. This coefficient value is determined with the average distance of points to the currently assigned cluster and the average distance of points to the next closest cluster. After calculating the Silhouette score for every k value, a graph can be plotted. In the graph, the value of k which gives the maximum Silhouette score value is considered as the optimal value.

C. Implementing Explainability to the Model

There are limited works of XAI approaches for unsupervised learning algorithms. Fortunately, for K-means clustering algorithms an effective decision tree based approach has been introduced recently. Focusing on the explainability and accuracy tradeoff, IMM algorithm was introduced at first and then it was expanded with another algorithm called ExKMC. Both the algorithms assist to provide proper insight of the important features within the data and summarize the whole decision making process with a decision tree with less mistakes. The three basic approach for IMM as follows:

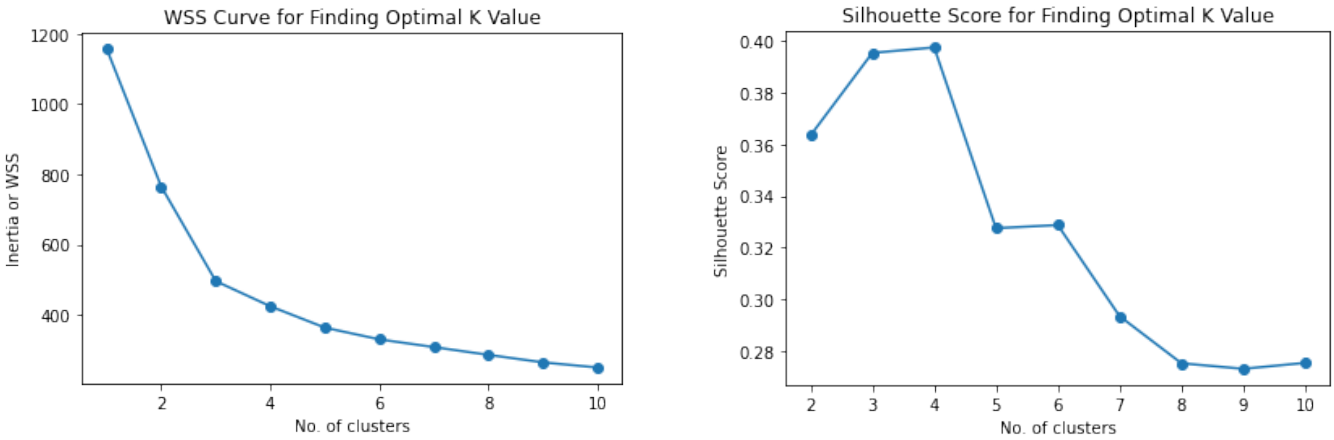
1) Build the model using clustering algorithm
2) Label each example according to the assigned cluster
3) Apply mistake minimization on the labeled examples iteratively and generate a decision tree with the least minimized mistakes

The main drawback of this algorithm is that the mistakes are not fully minimized because the value of leaf node of the tree is simply the total number of clusters. This produces a smaller tree where it is difficult to minimize the number of mistakes. When a data point is separated from its center, it is considered as a mistake. As a tree is built from top to bottom iteratively, with each step, nodes with best features and threshold is chosen that minimizes the "mistake" parameter. To overcome this problem, this approach was expanded to the ExKMC algorithm. In this modified algorithm, an additional parameter (k') is defined. This parameter denotes the maximum number of leaf nodes of the tree where this time the value must be greater than the optimal value of the cluster. As the number of the leaf nodes increases, the cost corresponding to mistake minimization becomes non-increasing. For this reason, it can generate a larger threshold tree with more accurate clustering assignments and also adds flexibility to determine the complex decisions of the model.

IV. Results & Analysis

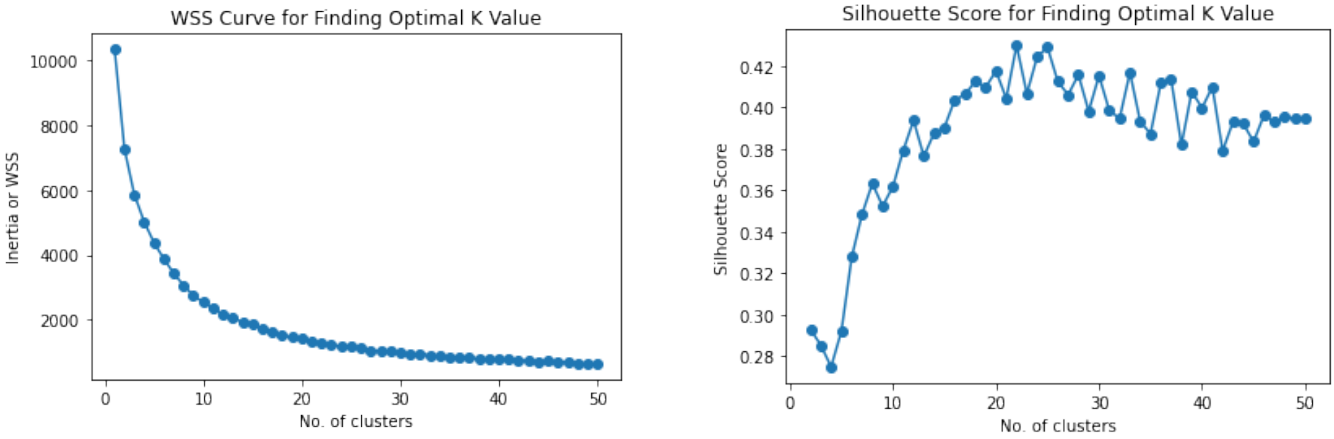
The optimal number of clusters were found applying the two previously mentioned methods to the datasets. For the first dataset, the optimal number of clusters were searched up to 10 clusters. Fig. 3 shows the WSS curve and Silhouette Score for the dataset. The elbow of the curve is clearly observable when the number of clusters are 3. On the other hand, the maximum Silhouette scores can be seen as 4. So, after analyzing the two graphs, the optimal number of clusters for the small dataset are 3 and 4. Similarly, for the large dataset, the optimal number of clusters were searched within 50 clusters. Fig. 4 shows the WSS curve and Silhouette Score for the dataset. From the figure, it is hard to interpret the optimal number of clusters as the elbow of the curve is not clearly visible. But in the case of Silhouette scores, the maximum value can be seen when the number of clusters are 22. So, in this case the optimal value is 22.

After finding out the optimal cluster values for both datasets, the two customer segmentation models were created, and after that, both the models were ready for intro-



(a) Elbow Method (b) Silhouette Analysis

Fig. 3: Elbow Method & Silhouette Analysis Applied on Dataset 1



(a) Elbow Method (b) Silhouette Analysis

Fig. 4: Elbow Method & Silhouette Analysis Applied on Dataset 2

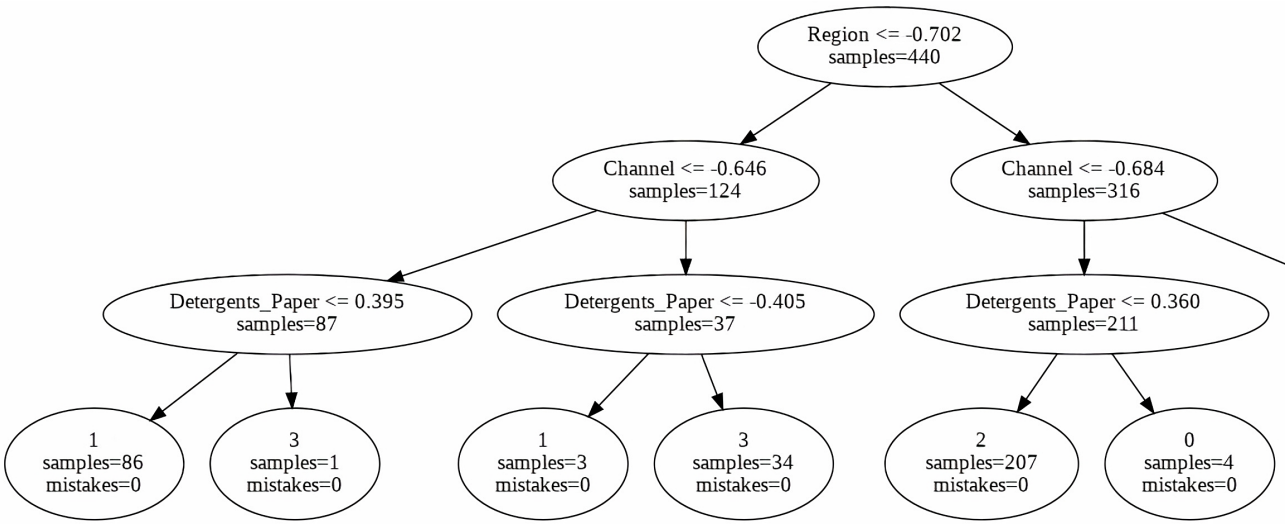


Fig. 5: Generated Tree (partial) for Dataset 1 when k = 4

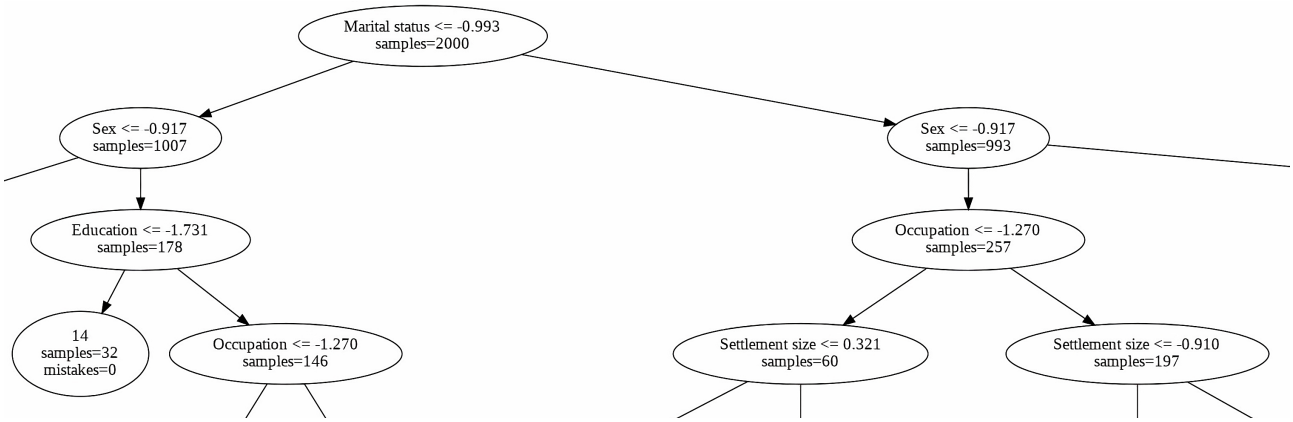


Fig. 6: Generated Tree (partial) for Dataset 2 when k = 22

ducing the XAI approach. With the help of ExKMC algorithm, trees for corresponding optimal cluster values were generated. For both cases, the maximum number of leaves was considered the optimal number of clusters times two ($k' \leq 2k$). Visualizing the trees, the relationship between the clusters and the important features can be easily observed. Fig 5. shows the partial generated tree when the number of clusters is 4 for the small dataset. From the tree, it is observable that the features 'Region', 'Channel', 'Detergents Paper' and 'Milk' play vital roles in determining the clusters and add explainability about why a datapoint belongs to one unique particular cluster. Here, out of 440 instances, the tree could successfully assign each and every datapoints to a cluster with zero total mistakes in the leaf nodes. Similarly, Fig. 6 shows the partial snapshot of the generated tree for the second dataset when the optimal number of clusters is 22. In this bigger tree, all the features are somehow associated with determining cluster members for a particular cluster. Here, out of 2000 instances, 43 datapoints were not assigned to any particular cluster which were considered total number of mistakes in this case. The rest of the instances were labeled perfectly with unique cluster assignments.

V. Conclusion & Future Works

In this paper, the XAI method with unsupervised learning was implemented for customer segmentation with the use of ExKMC algorithm. Since clustering via trees is explainable by design, it is easier for users to interpret the reason behind the decisions taken by the machine learning model. For future works, other clustering algorithms can be experimented on different datasets to obtain more satisfactory results. A comparative analysis of their performances can be done in order to figure out what approach should be used in which situation. Furthermore, hybridization of different clustering techniques can be experimented to make the clustering more efficient and less time consuming. More robust explainable approaches for unsupervised learning can be introduced to improve the model's overall performance which can lead to better customer satisfaction.

References

[1] Abidar, Lahcen, Dounia Zaidouni, and Abdeslam Ennouaary. "Customer Segmentation With Machine Learning: New Strategy For Targeted Actions." Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications. 2020.

[2] Badih Ghattas, Pierre Michel, and Laurent Boyer. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. Pattern Recognition, 67:177–185, 2017.

[3] Bing Liu, Yiyuan Xia, and Philip S Yu. Clustering via decision tree construction. In Foundations and Advances in Data Mining, pages 97–124. Springer, 2005.

[4] Cardoso, Margarida. (2014). Wholesale customers. UCI Machine Learning Repository.

[5] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering via optimal trees. arXiv preprint arXiv:1812.00539, 2018.

[6] Frost, Nave, Michal Moshkovitz, and Cyrus Rashtchian. "ExKMC: Expanding Explainable k-Means Clustering." arXiv preprint arXiv:2006.02399 (2020).

[7] Hossain, ASM Shahadat. "Customer segmentation using centroid based and density based clustering algorithms." 2017 3rd International Conference on Electrical Information and Communication Technology (EICT). IEEE, 2017.

[8] Hung, Phan Duy, Nguyen Thi Thuy Lien, and Nguyen Duc Ngoc. "Customer segmentation using hierarchical agglomerative clustering." Proceedings of the 2019 2nd International Conference on Information Science and Systems. 2019.

[9] M. T. Ribeiro et al., ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.

[10] Monil, Patel, et al. "Customer Segmentation Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology (IJRASET) 8.6 (2020): 2104-2108

[11] Moshkovitz, Michal, et al. "Explainable k-means and k-medians clustering." International Conference on Machine Learning. PMLR, 2020.

[12] Sharma, D. (2021, May). Customer Clustering, Version 1. Retrieved September 6, 2021 from https://www.kaggle.com/dev0914sharma/customer-clustering

[13] Shen, Boyu. "E-commerce Customer Segmentation via Unsupervised Machine Learning." The 2nd International Conference on Computing and Data Science. 2021.

[14] Shirole, Rahul, Laxmiputra Salokhe, and Saraswati Jadhav. "Customer Segmentation using RFM Model and K-Means Clustering." (2021).

[15] Pradana, Musthofa Galih, and Hoang Thi Ha. "Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering." Journal of Applied Data Sciences 2.1 (2021): 19-25.

[16] Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. Interpretable clustering using unsupervised binary trees. Advances in Data Analysis and Classification, 7(2):125–145, 2013.