**Clustering model of electricity load profile using K-means clustering: A case study of electricity companies in Indonesia**

Radit Rahmadhan1 and Meditya Wasesa2

1Institute Technology Bandung

**Abstract.**

The increasing number of electricity users in Indonesia does not necessarily mean positive growth for the state-owned electricity provider in Indonesia, i.e., PT. PLN Persero. Understanding customer segmentation and customer preferences is very important to increase customer satisfaction.

**Keyword**: Customer Relationship Management, Machine Learning, Clustering

# Introduction

The increase in electricity consumption in Indonesia continues to increase from 2015 to 2020 around 98.89%, business customers dominate electricity consumption [1]. PT. PLN Persero is the only electricity provider in Indonesia that provides higher power for the entire region, including the West Sumatra region, due to the increasing electricity consumption with a focus on business customers. This is because it has the potential to increase the company's revenue. However, according to customer information, blackouts often occur in West Sumatra and even occur four times a month.

Based on the results of data analysis that has been carried out, power outages cause the average electricity usage time for business customers to be under 50 hours. Based on the information from the Manager of Commerce of PLN for the West Sumatra Region, the incident was since customers used a lot of peak-load electricity usage time instead of using electricity outside of peak hours, even though PT. At that time, customers rarely used it. Based on these problems, PT. PLN Persero must understand the characteristics of customers so that the use of electricity at times outside the peak load is more optimal by customer segmentation.

Customer segmentation is one way to better understand customer preferences. According to previous research, customer segmentation refers to grouping customers into more specific ones to predict future customer actions or behavior [2]. Customer segmentation is used to predict customer characteristics in buying or using facilities provided by the company by mapping customer characteristics to increase sales or use facilities. That customers use and build relationships and enhance customer commitment to building a solid business[3][4].

Based on previous research on customer segmentation, electricity consumption is still small. This study uses the clustering method to create customer segmentation. Clustering is part of data segmentation to group large amounts of data into several groups with the same characteristics in each group[5]. Clustering is also widely used to understand customer behavior to increase company profitability. The clustering model used in this study is an efficient K-Means Clustering approach to evaluate customer differences in using electricity consumption. The main factor in the selection of this model is the handling of several large data sets, such as data owned by PT. PLN Persero. The dataset used is customer transaction data of PT. PLN Persero from 2019 to 2020. The data points that will be predicted are installed power at the customer, peak load electricity usage time, peak load electricity usage time. We want to break down the group of business customers with high potential in peak load electricity consumption into several dimensions. A validation method is needed to determine the best number of clusters to determine the clustering measurements made. The Elbow method is used for the correct number of groups by looking at the SSE value by looking at the sloping point of the curve that has been determined[5][6].

This study aims to develop a predictive model of electricity consumption by examining the energy consumption patterns of business customers using the data described previously. We divide consumption behavior into two parts: customers who use peak load electricity and that outside of peak loads, using the K-Means Clustering model in grouping and determining the best number of clusters using the Elbow method. We categorize customers based on their average electricity usage per month. These findings can help companies identify potential customers using higher peak loads which help optimize the power provided by PT. PLN Persero.

# Literature Review

Previous studies in customer segmentation have explored various dimensions of customer clustering problems [6]– [10]. Many of them use the marketing context as a case study. Exploring customer grouping using the K-Means clustering model by considering the specified product preferences and predicting customer behavior in buying products offered by the company[6].

A study of the marketing context [7]assigns each customer attribute as a dimension and sets each customer as a particle. All customers in the company can form a multidimensional space, defined as the customer attribute space. Mapping relationships between customer attributes and conceptual categories can be constructed using analytical methods or sample learning methods. The analysis method analyzes the attribute character of each conception category that must be possessed, then forms a mapping of the relationship between the attribute space and the conception space. However, many mapping relationships between attribute spaces and conception spaces are unclear, so it is necessary to use sample learning methods to establish mapping relationships. Later this method will also be applied in customer grouping to find customers who use peak load electricity consumption [6]– [12].

A context study of electricity consumption [4] using experimental data by installing 4000 intelligent meters in several homes in Ireland with existing methods used to classify household electricity use, in general, can be divided into four categories, statistics, manipulation, time series, and clustering. Statistical methods have been widely used in the unregulated power market to form a standard load PC. A typical load PC is used for settlement purposes and estimates the amount and Time of Use (ToU) of electricity used. A series of PCs are manufactured for different market segments (e.g., residential, commercial, industrial) and derived on an average for all customers within a customer class.

Another study in marketing that combines the LFRM, CLV, and K-Means models explains Customer Lifetime Value (CLV) [8] in each customer segment. The grouping uses the K-Means Clustering method based on the LRFM (Length, Recency, Frequency, Monetary) model. The cluster formation process uses the Elbow and SSE methods with the best clusters = 2 clusters. The CLV value is generated from the multiplication of the LRFM normalization results, and the LFRM weight values ​​are then added up and performed on each cluster formed. Based on the LRFM matrix, this cluster has a high loyalty value, with the LRFM symbol being a loyal customer (the best segment with a high customer loyalty value). Based on the LRFM symbol, companies can create strategies to retain customers and acquire loyal customers with high profitability. In determining the number of clusters using the elbow method as the best number of clusters[8]– [10].

Another study used the Placket-Luce (PL) probabilistic ranking model. Each cluster is represented as a composite of Voronoi cells defined by prototypes and assigned a set of PL label scores that rank the cluster-specific labels. The unknown PL cluster parameters and prototype positions were determined using supervised learning techniques. Cluster membership and ranking for a new instance are determined by its leading members. The proposed algorithm is empirically based on the synthesis of scale and real-life data. The OT-based method is superior to the heuristic-based supervised clustering approach. The proposed PL-based algorithm is also tasked with predicting label rank. The results show that it is highly competitive with ranking algorithms and partially accurate on ranking data [11].

Table 1 provides an overview of previous studies that analyzed marketing topics using transaction data or customer history.

Table 1 Reviewed Studies on Customer Segmentation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Article | Business Context | Dataset | Segmentation  Features | Techniques |
| Vucetic et al,2018 | Marketing | Restaurant customer feedback data period January 1, 2016, to December 31, 2016 | Customer name, type of food ordered, food review, gender | K-Means using PL based algorithm |
| Aziz et al, 2019 | Small Medium Enterprise | SME customers are all pulse server operators AR-Pulsabiz Malang Indonesia period January 1, 2018, to June 30,2018 | ID Customer, Length, Recency, Frequency, Monetary | K-Means Clustering and LRFM Model |
| Ye Jingyi, 2021 | E- Commerce | Online Retail Data Set period 12 January 2010 and 12 September 2011 from UK | Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country, Total Price | K-Means Clustering |
| Sano et al, 2021 | Marketing | Transaction Supermarket data January 1, 2017, to December 31, 2018 | ID Customer, Product Name, Length, Recency, Frequency, Monetary | AEF, RFM, k-means, Ward method, FCM, and the decision tree |
| Antony et al,2019 | RFM Analysis | Sales data of a pharmacy in Palembang period January 2015 until December 2015 | ID Customer, Product Name, Length, Recency, Frequency, Monetary | K-Means Method and eight indexes of validity to determine the optimal number of clusters namely Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Rutkowski Index, Hubert Index, Ball-Hall Index, and Krakowski-Lai Index |
| Puh et al, 2020 | Food Retailing | Questionnaire data consisting of 500 consumers in Croatia in 2020 | Demographic characteristics (Age, Gender, Education, Occupation, Monthly Income in HRK), Product, Frequency, Percentage | Latent Class Model |
| Abdi et al, 2018 | Telecom Company | Customers of a telecom company period January 1,2017 to December 31, 2017 | Socio-demographic attributes (Region, Age, Marital, Address, Income, Education, Employment, Retire, Gender), Behavioral Attributes (Hours of Usage (Longmont, Tollmon, Equipmon, Cardmon, Wiremon), Selected Service (Multiline, Voice, Pager, Internet, Call Id, Call wait, Forward, Confer, Call card, Wireless, Churn)) | K-Means Clustering, Neural network, |
| McLoughlin et al, 2014 | Electricity | Experimental data by installing smart meters to more than 4000 residences in Ireland, period January 1, 2009 to December 31, 2010 | Dwelling type, No. of bedrooms, Age, Social Class, Electronic Type | k-means, k-medoid and Self Organizing Maps (SOM) |
| Li et al, 2012 | Transportation | Historical data from the vehicle sharing platform database at the university detailed data of all customers from November 30, 2015, to November 30, 2017 | User ID, driving mileage, points, discounts and 29 other attributes. The Variable are used User id, current miles, cost, car id | K-Means Method and KLRFMD model |
| Marisa et al, 2019 | Small Medium Enterprise | The sale of electric pulses period January 1, 2016, to December 31, 2017 | Name Customer, Average of transaction/ week, Payment System | K-Means Clustering |
| Chindyana et al ,2021 | Tourism | Google review rating in 2020 | Id Customer, Gender, Place, Review | K-Means Method and DBSCAN Method |
| Zhao et al,2021 | E-commerce | The customer transaction online customer company in the UK period January 1, 2016, to December 31,2017 | Invoice number, quantity, price, address, and zip code. | K-Means Clustering and RFM Model |

Based on the literature, previous research has mostly done customer segmentation, the context is more towards marketing and to map customer behavior in the future because it affects the company's marketing strategy. In the context of electricity consumption, previous researchers rarely used customer segmentation techniques. In this study, we conducted clustering using the K-Means Clustering method with validation of the number of clusters using the Elbow method and evaluation using ANNOVA to determine customers who use electricity at peak load times.

# Method

Figure 1 shows the framework in this study. The framework is adapted from standard methods for building predictive analytical models[13]. There are five stages: data collection, data cleaning, selecting relevant predictor variables, determining potential predictive methods, evaluating, validating, choosing the best predictive model, and finally reporting the research results.



Figure 1 Research Framework

**3.1 Data Collection**

This study, we used data from PT. PLN Persero. The data taken is only 1 area because the fields for each region are the same. The data taken by PLN is the West Sumatra zone. Our research uses customer transaction data from January 2019 to December 2020, consisting of 16,504,228 and 107 data variables. Table 2 describes the data that has been taken from 2 years. Some records were removed from the data set because they showed illogical conclusions, i.e., duplicate records or missing values.

Table 2 Descriptive Statistics of The Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Year** | **Row** | **Variable** |
| Customer Transactions history | 2019 | 7,945,689 | 107 |
| Customer Transactions history | 2020 | 8,558,539 | 107 |

* 1. **Data Cleaning**

Customer transaction data collected for two years will be selected based on the potential for prediction. The data has 107 variables and 16,504,228 rows. The first data cleaning removes variables that do not have data variations so that it becomes 49 variables and 16,504,228 rows. The second data cleaning removes variables that do not affect prediction so that it becomes 31 variables and 16,504,228 rows. The third data cleaning removes variables with a bit of variation in data to become 18 variables and 16,504,228 rows.

The four data cleaning variables were chosen with the condition of the customer service area in Padang because of the high number of customers in this area than in other areas. They then chose a group focused on general customers. There were 15 variables and 1,187,934 data. The fifth data cleaning due to data outliers was selected to focus on business customers. This has the potential to be predicted to have exciting data variants and discard two variables due to slight variations so that it becomes 13 variables and 508,934 rows. Table 3 shows the process of data cleaning, and Table 4 shows the results of data cleaning.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Period | Filter | | | | |
| Variable | Rates  (Variable) | Group Code  (Variable) | Customer Service Area  (Variable) | Row |
| January 2019 - December 2020 | 49 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 31 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 27 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 18 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 15 | - | 0 | Padang | 1,187,934 |
| January 2019 - December 2020 | 13 | Business | 0 | Padang | 508,934 |

Table 3 The Process of Data Cleaning

Table 4 The Result of Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Numeric/Nominal | Data Type | Variable Description |
| ID Customer | Nominal | Integer | Identity of the customer |
| Customer Service Unit | Factor | String | Customer Service Units or service branches provided by the company which are located in 4 customer service centers namely Belanti, Painan, Indarung, Pariaman, Lubuk Basung, Lubuk Sikaping, Koto tuo, Baso, Sijunjung, Sungai Rumbai, Kayu Aro, Sawah Lunto, Batusangkar, Lintau, Lima Puluh Kota and others |
| Data Entry Date |  | Date | Admin enters data per 1 month |
| Rates | Factor | String | B1 means a business that uses electricity from 450 kwh to 5500 kwh, B2 means a business that uses electricity from 6600 to 200 thousand kwh, B3 means a business that uses 200 thousand kwh of electrical power and above |
| Power | Factor | Integer | Power used by customers such as 450 kwh,900 kwh,1300 kwh, 2200 kwh,3300 kwh, 7700 kwh,15400 kwh,132000 kwh,110000 kwh and others |
| Meter Code | Factor | String | M means analogue meter and E means digital meter |
| Flash time |  | Double | Electricity usage time by customer |
| Total KWH |  | Integer | The total of peak load kwh usage and peak external load kwh used by customers |
| KWH Off - Load |  | Integer | KWH used at peak external load by customers |
| KWH Peak Load |  | Integer | KWH used at peak load by customers |
| Discount |  | Double | Discounts given by the company based on the provisions of the company such as using unused kwh by the company or because of a natural disaster |
| Peak Offload Fee |  | Double | Payments made when using Peak Offload |
| Peak Load Fee |  | Double | Payments made when using Peak Load |
| Total Cost |  | Double | The total cost paid by the customer |

* 1. **Choice of Variable**

Power, Meter Code, Flash time, Total KWH, Discount, Peak Offload Fee, Peak Load Fee, Total Cost because this predictor variable is not the one that has the potential to be included in the clustering model. The Power variable shows the electrical power installed by the customer. The KWH Off-Load variable shows customers using electricity from 5 pm to 6 am. The KWH Peak Load variable shows indicators of customers using electricity from 6 am to 5 pm. We grouped based on these three variables to determine the number of customers using peak load times and off-peak load times. Table 5 shows detailed information about the predictor variables.

Table 5 The Result of Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Numeric/Nominal | Data Type | Variable Description |
| Power | Factor | Integer | Power used by customers such as 450 kwh,900 kwh,1300 kwh, 2200 kwh,3300 kwh, 7700 kwh,15400 kwh,132000 kwh,110000 kwh and others |
| KWH Off - Load |  | Integer | KWH used at peak external load by customers |
| KWH Peak Load |  | Integer | KWH used at peak load by customers |

* 1. **Choice of Potential Method**

This research aims to develop a prediction model with customer segmentation or clustering that can provide accurate predictions of customers who have the potential to use electricity at peak or off-peak loads. However, this research is still investigating a clustering model and its ease of implementation. We use the K-Means Clustering model to group customers.

**3.4.1 K-Means**

Commonly, K-means is one of the well-known unsupervised learning techniques for cluster analysis[5]. Cluster analysis is used to aggregate or divide the data set into several clusters according to the similarity value. The situation in this model is used because this algorithm has simplicity and ease of use, and users can determine the number of clusters themselves. This number of clusters (k) needs to be determined by validation[14]. It starts with randomly generated centroids and iteratively computes new centroids to converge to the last group. The steps in the k-means model are explained as follows[8].

Step 1: Determine the number of clusters with validation

Step 2: Each data point in the data set will be assigned to the nearest centroid, and then a   
 new centroid is generated.

Step 3: To recalculate a new cluster by assigning all data points to the nearest centroid,   
 and then a new group is created.

Step 4: The process will be repeated between step 2 and step 3 until the stopping criteria   
 are met.

In Step 4, the points used as the result of clustering refer to table 6 which is the result of clustering from k = 2 to 10.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Dissimilarity***  ***Ratio*** | K | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  |  |  |  |  |  |  |  |  |  |

* 1. **Evaluation and Validation**

Evaluated the prediction model is performance described earlier, we use the elbow method[15], which is a method used to determine the optimal number of clusters, by looking at the percentage comparison of the number of clusters that will form an angle on the curve[15]– [18]. This method is used in cluster analysis to interpret and perform the correct number of clusters by looking at the SSE value. If the value of the first cluster with the weight of the second cluster forms an angle on the curve or the most significant decreasing value, the cluster value is the best[19]. This method is a visual method that looks at the total intra-cluster variation or the total Within-Clusters Sum of Squares (WSS) function of the number of clusters[20]. The larger the number of clusters k, the smaller the WSES value or vice versa. In this study, determine the best number of clusters[7], [16], [18], [20].

* 1. **Model Use and Reporting**

The performance of each cluster (k) and the predicted scores will be compared. The best prediction model with the best predictive ANOVA score will be selected and used to help decision-makers to determine the correct number of customer groupings in a better way.

# Finding and Discussion

The research aims to break down the behavior of electricity-using customers into 2, namely customers who use electricity at peak load times and customers who use electricity during peak off-load times by using an unsupervised machine learning model. The model creates a model automatically from the training data set. A non-learning algorithm tries to identify and build patterns that can be grouped based on pre-selected predictor variables. Based on the design created, the first step is to determine the number of clustering using the elbow method to get the best number of clusters (k).Figure 1 shows the number of clusters based on the results of the predictor variables previously described using the Elbow method. The correct number of clusters is determined by looking at the line graph when it starts to slope. From Figure 1, the chart begins to drop at point 3 and point 4.

Chart, line chart, scatter chart

Description automatically generated

Figure 1 The Number of clusters of K

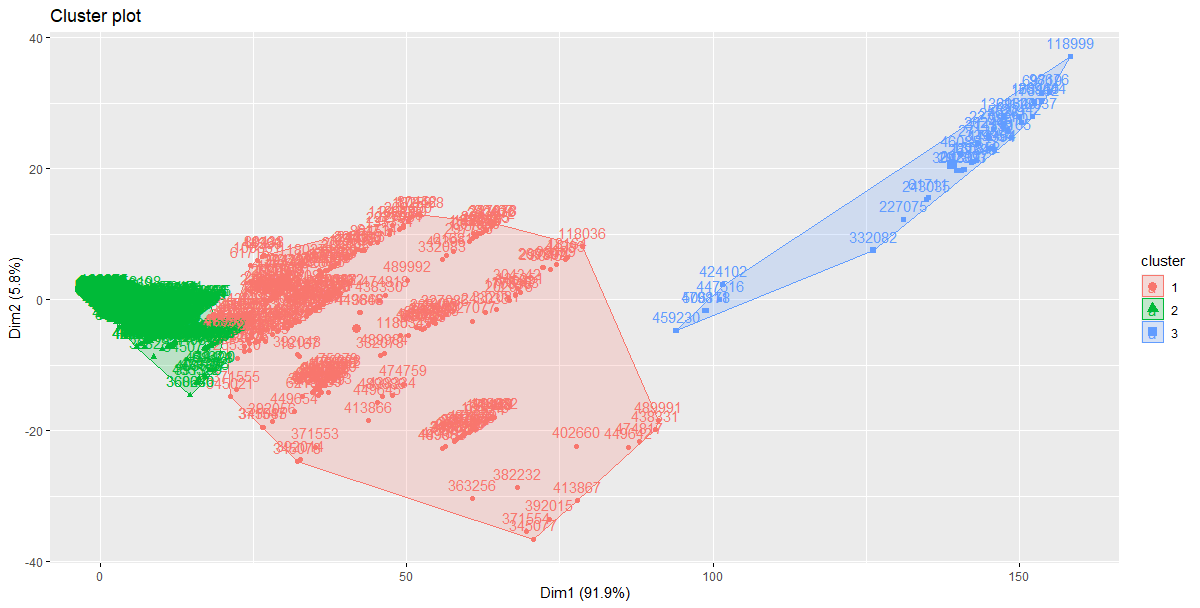
From the analysis that has been carried out using the Elbow Method at point 3 and point 4 using the K-Means clustering model, it was found that the best usage grouping in the electricity consumption sector is at point 3; The results of the analysis show that at point 4 there are outliers (groups at the dark green point) in the distribution. The results of the study of the results of the k-means are Figure 3 and Figure 4 can be seen below. 

Figure 2 Cluster result of k = 3

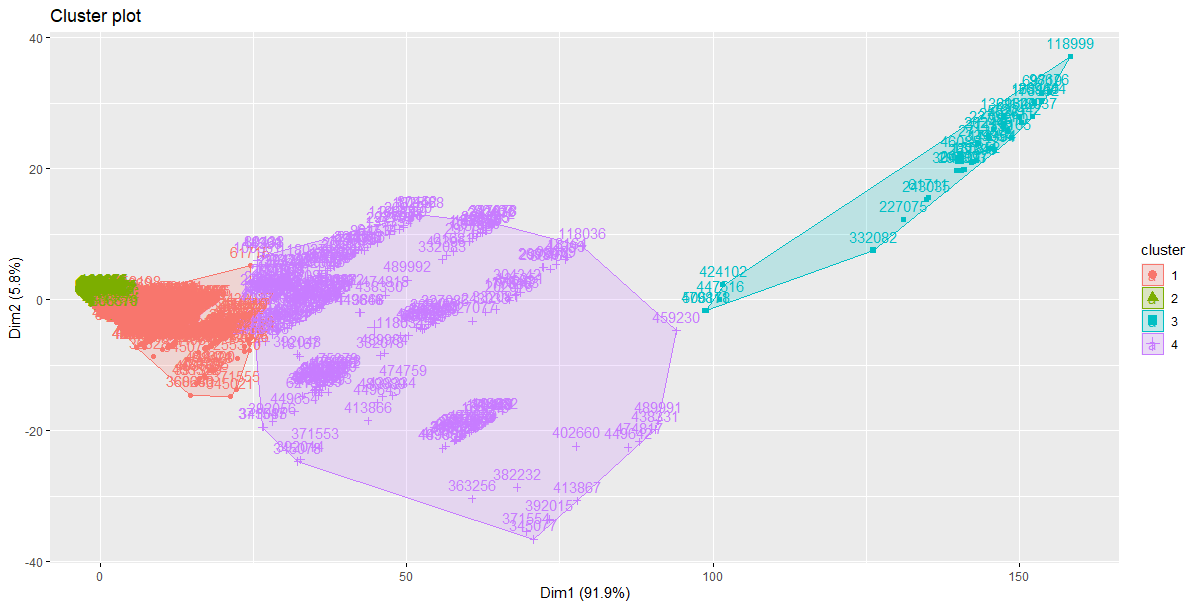


Figure 3 Cluster result of k = 4

# Conclusion

# References

[1] Katadata, “Konsumsi Listrik Nasional Terus Meningkat,” *www.databook.com*, Jan. 09, 2020. https://databoks.katadata.co.id/datapublish/2020/01/10/konsumsi-listrik-nasional-terus-meningkat (accessed Jan. 04, 2022).

[2] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[3] J. Ye, “Analysis on E-commerce Order Cancellations Using Market Segmentation Approach,” in *ACM International Conference Proceeding Series*, Jan. 2021, pp. 33–40. doi: 10.1145/3450588.3450596.

[4] F. McLoughlin, A. Duffy, and M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Applied Energy*, vol. 141, pp. 190–199, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.039.

[5] R. Bapna, P. Goes, A. Gupta, and Y. Jin, “USER HETEROGENEITY AND ITS IMPACT ON ELECTRONIC AUCTION MARKET DESIGN: AN EMPIRICAL EXPLORATION 1,” 2004.

[6] Z. J. Lee, C. Y. Lee, L. Y. Chang, and N. Sano, “Clustering and classification based on distributed automatic feature engineering for customer segmentation,” *Symmetry*, vol. 13, no. 9, Sep. 2021, doi: 10.3390/sym13091557.

[7] F. Marisa, S. Sakinah Syed Ahmad, Z. Izzah Mohd Yusof, T. Mohammad Akhriza, W. Purnomowati, and R. Kumar Pandey, “The Analyze of Relationship between Revenue and Customer Payment Methods in Small Medium Enterprise Based on Clustering K-Means,” in *Journal of Physics: Conference Series*, Jul. 2021, vol. 1908, no. 1. doi: 10.1088/1742-6596/1908/1/012021.

[8] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[9] S. Janardhanan and R. Muthalagu, “Market segmentation for profit maximization using machine learning algorithms,” in *Journal of Physics: Conference Series*, Dec. 2020, vol. 1706, no. 1. doi: 10.1088/1742-6596/1706/1/012160.

[10] F. Abdi and S. Abolmakarem, “Customer Behavior Mining Framework (CBMF) using clustering and classification techniques,” *Journal of Industrial Engineering International*, vol. 15, pp. 1–18, Dec. 2019, doi: 10.1007/s40092-018-0285-3.

[11] M. Grbovic, N. Djuric, and S. Vucetic, “Supervised Clustering of Label Ranking Data.” [Online]. Available: https://epubs.siam.org/page/terms

[12] S. L. Lo, R. Chiong, and D. Cornforth, “Ranking of high-value social audiences on Twitter,” *Decision Support Systems*, vol. 85, pp. 34–48, May 2016, doi: 10.1016/j.dss.2016.02.010.

[13] G. Shmueli and O. R. Koppius, “PREDICTIVE ANALYTICS IN INFORMATION SYSTEMS RESEARCH 1.”

[14] R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470–477, 2019, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[15] P. Bholowalia and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” 2014.

[16] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[17] C. D. Rumiarti and I. Budi, “SEGMENTASI PELANGGAN PADA CUSTOMER RELATIONSHIP MANAGEMENT DI PERUSAHAAN RITEL: STUDI KASUS PT GRAMEDIA ASRI MEDIA.”

[18] “Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco.” [Online]. Available: https://www.researchgate.net/publication/284899160

[19] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, “Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty,” *Expert Systems with Applications*, vol. 37, no. 7, pp. 5259–5264, Jul. 2010, doi: 10.1016/j.eswa.2009.12.070.

[20] IEEE Power & Energy Society, IEEE Industry Applications Society, and Institute of Electrical and Electronics Engineers, *2018 IEEE PES/IAS PowerAfrica : 28-29 June 2018.*