Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 7, No. 2, September 2024, pp. 255 – 265
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Eligibility Study of Targeted Electricity Subsidies Using DBSCAN on 450 VA and 900 VA Households at PLN UP3 Bandung

[1]Randy Zakya Suchardy, [2]Adi Firmansyah, [3]Nugraha Priya Utama, [4]Rahman Indra Kesuma
[1,2,3,4]School of Electrical Engineering and Informatics Engineering, Bandung Institute of Technology, Indonesia
Email: [1]23523027@mahasiswa.itb.ac.id, [2]23522310@mahasiswa.itb.ac.id, [3]utama@itb.ac.id,
[4]33221026@std.stei.itb.ac.id, [5]rahman.indra@if.itera.ac.id

Article Info

Article history:
Received Dec 15th, 2023
Revised Mar 25th, 2024
Accepted Apr 25th, 2024

Keyword:
Classification
Clustering
DBScan
Decision Tree
Subsidy

ABSTRACT

PT. PLN (Persero), a State-Owned Enterprise (SOE), is mandated by Law No. 30/2007 on Energy and Law No. 30/2009 on Electricity to provide subsidy funds for the poor. The objective of this study is to analyze eligibility criteria for electricity subsidy recipients for customers using 450 VA and 900 VA power groups, to target the electricity subsidy program better. The data used is postpaid customer data from UP3 Bandung in September 2023. The variables used are the amount of electricity consumption, the number of bills, late fees, installment fees, and 50 other variables. The method used in this research is DBScan Clustering which is applied to each power group. Within each group, we analyzed two normalized versions of the dataset standard version and the minmax version. Furthermore, to assess the optimal clustering results, we integrated various metrics, including the Davies-Bouldin Index and Silhouette Score with visual assessment. After that, the best factor suggestions were sought through decision trees, by performing different decision tree classifiers for each power group, using normalized versions of cluster labels. The results showed that among the 50 features available in the raw dataset, it was successful in identifying key features, such as late fees, installment fees, electricity consumption, and bill charges to be important criteria.

Corresponding Author:
Nugraha Priya Utama,
Informatics Engineering,
Bandung Institute of Technology,
Jl. Ganesa No.10, Lb. Siliwangi, Coblong District, Bandung City, West Java 40132.
Email: utama@itb.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v2i2.26818

1. INTRODUCTION

PT. PLN (Persero), a State-Owned Enterprise (BUMN) in the electricity sector, is committed to delivering high-quality electricity services throughout the archipelago while adhering to international standards. In line with Indonesia's law regulation on Electricity, the government allocates subsidy funds for electricity to assist the economically disadvantaged [1].

Electricity Subsidy is an assistance the Government provides to consumers through an Electricity Tariff that is lower than the economic tariff. With the same amount of electricity usage, consumers who get subsidized tariffs will pay lower electricity bills than consumers who do not get subsidies. The government bears the difference between the subsidized and the economic tariff, which is then paid to PLN. The subsidy, provided to customers with 450 VA and 900 VA power, aims to lower their electricity tariff compared to the economic tariff, with the government covering the difference [1]. Data on eligible households comes from the Integrated Data of the Poor Handling Program managed by the National Team for the Acceleration of Poverty Reduction (TNP2K) [2].

However, the subsidy distribution faces challenges, with affluent households still receiving benefits. In 2015, approximately Rp 49.32 trillion (87%) of the total electricity subsidy was consumed by 450 VA and 900 VA households, but only 4.1 million out of 23 million 900 VA households were truly eligible [2]. And in 2015 - 2020, the inaccuracy of electricity subsidies was also shown by the gap in the proportion of subsidized consumption consumed by the rich group community where the household community was decal 5-10 consumed 74% of electricity subsidies while the vulnerable community was vulnerable to the poor 1- 4 only consume 26% [3].

The objective of this study is to analyze eligibility criteria for receiving electricity subsidies among 450 VA and 900 VA customers, to ensure more accurate targeting. The scope of this study includes collecting and analyzing data on postpaid household customers in UP3 Bandung for 450 VA and 900 VA tariff groups in September 2023. The study aimed to establish and implement the revised criteria within a three-month timeframe using the grouping of evaluation metrics as benchmarks [4]. Our study holds significance as it provides an opportunity for PLN to utilize our findings for the refinement of eligibility criteria pertaining to the allocation of electrical subsidies, thereby contributing to the establishment of a more equitable regulatory framework for such subsidies. To date, our investigation has revealed no prior studies aligning with the specific focus of our research. Previous works by Widiawati [5] and Hutahaean et al. [6] employed pre-existing labels denoting household eligibility, falling short in addressing the persistent challenge of inequitable distribution of subsidies.

Our study has some limitations due to a lack of detailed information about household customers in 450 VA and 900 VA tariff groups and criteria for customers receiving electricity subsidies. The study targets two main outcomes: identifying relevant factors for subsidy determination based on electricity consumption data and developing a machine learning model for categorizing eligible and ineligible customer groups.

We formulate our study into several sections. In section 1, we introduce the problem statement regarding the electricity subsidy. In section 2, we provide the discussion about the previous studies that are related to ours, the previous studies utilizing the existing dataset and eligibility criteria while our focus is on refining eligibility criteria based on past electricity usage activities. In section 3, we explain the research objectives and the comprehensive experimental design, model training, and evaluation. Lastly, in section 4 we conclude to identify the relevant factors to consider when deciding which customers should receive electricity subsidies based on their electricity consumption data from model training that we used.

2. RELATED WORKS

The prior research relevant to the current investigation includes two studies. First, Widiawati [5], the study aimed to classify household electrical subsidies customers based on their characteristics utilizing Support Vector Machine (SVM) and Naive Bayes Classifier methods. The comparison revealed that SVM yielded superior results with optimal parameters in the RBF kernel, specifically $C = 10$ and $\gamma = 1$. Notably, customers with a 450 VA category were correctly classified at 91.6%, with 8.4% predicted in the 900 VA category. Similarly, customers with a 900 VA category were classified at 81.9%, with 18.1% predicted in the 900 VA category. Second, Hutahaean et al. [6], the research aimed to classify households receiving electricity subsidies using data mining methods, specifically K-Nearest Neighbor (KNN) and SVM. The variables considered included the status of electricity subsidy recipients and various explanatory variables. Results indicated that the KNN method exhibited superior accuracy at 98.07%, showcasing a significant difference in performance compared to SVM, where KNN outperformed SVM in classification. In contrast to the two preceding studies associated with our subject, our objective differs. Unlike the prior research, which utilized the given dataset and consequently overlooked the existence of erroneous recipients resulting from the inequitable distribution of electricity subsidies, our focus is directed towards addressing this issue. Our aim is to enhance the definition of eligibility criteria for individuals to receive electricity subsidies by refining it based on their past electricity usage activities.

3. EXPERIMENT & METHODOLOGY

3.1. Experiment Objectives

Based on the problem statement in the preceding section, we identify our research objectives to produce the following solutions:

1. Criteria for determining an individual's eligibility for receiving subsidies
   The criteria will consist of a set of rules that encompass features and their corresponding values, structured in the form of a decision tree. The decision-tree structure is chosen for its interpretability, as it is commonly acknowledged as a model characterized by high interpretability [7].
2. Machine learning model
   Expanding upon the preceding point, it is important to mention that the machine learning model, in this context, will adopt the configuration of a decision-tree classifier. This signifies that the

underlying algorithm employed for the task at hand will take the shape of a decision tree, a classification model known for its ability to make decisions based on a hierarchical set of rules and conditions.

3.2. Experimental Design

Our experiment will consist of four phases: data acquisition and preprocessing, DBSCAN clustering, clustering evaluation, and decision tree learning. As the dataset lacks an eligibility label, the cluster labels generated during the DBSCAN clustering phase will serve as the subsequent labels employed in the training process of decision tree learning. We assume that these cluster labels are indicative of an individual's eligibility for receiving subsidies. The experiment pipeline can be seen in Figure 1.

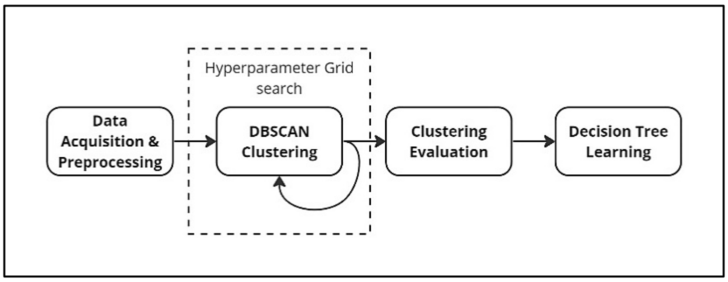


Figure 1. Experiment pipeline

3.2.1. Data Acquisition and Preprocessing

During this experiment, we used an exclusive raw dataset sourced from PLN containing 83348 rows and 50 columns. This dataset specifically encapsulates data pertaining to postpaid household customers falling under the 450 VA and 900 VA tariff groups, who are recipients of subsidies from UP3 Bandung for September 2023. As part of the data preprocessing phase, a series of judicious steps were taken, including data cleaning and feature engineering. Initially, columns deemed irrelevant, based on insights garnered from interviews with domain experts from PLN's Customer Experience Division and the second author representing PLN, were systematically eliminated. These irrelevant columns typically contained personal information about PLN's customers. Furthermore, columns exhibiting a complete absence of meaningful information—comprising solely of zeros and null values—were also excluded, given their lack of relevance to the problem at hand. In addition, a strategic feature extraction process was undertaken to enhance the dataset's utility. For instance, three distinct columns (RPBK1, RPBK2, and RPBK3) denoting late fees were amalgamated into a single feature, namely RPBK. Each of the original columns indicated varying levels of late fees, and the new feature RPBK was designed to convey whether an individual had incurred a late fee (assigned the value 1) or not (assigned the value 0). This feature extraction process served the dual purpose of reducing the dimensionality of the dataset by omitting obsolete features and introducing more meaningful and consolidated representations. Consequently, the resultant dataset now comprises 83348 rows and 18 columns, with 37700 rows attributed to the 450 VA power group and 45648 rows to the 900 VA power group. For a comprehensive understanding of each column or feature, refer to Table 1 for detailed descriptions.

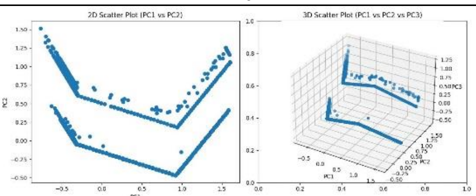
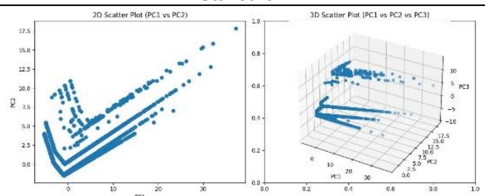
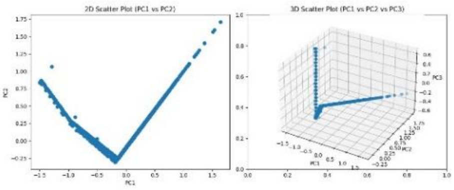
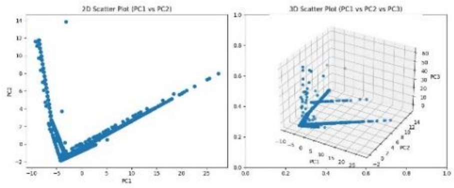
Table 1. Details of dataset

Feature Name	Description
JAMNYALA	Hours of electricity (hours)
KWHLWBP	Kwh usage outside peak load time
KWHWBP	Kwh usage at peak load time
BLOK3	Kwh usage each blok
PEMKWH	Kwh usage
RPLWBP	Outside peak load time fee
RPWBP	Peak load time fee
RPBLOK3	Fee each blok
RPBEBAN	Subscription fee
RPPTL	Electricity usage fee
RPBPJU	Public street lighting fee
RPPLN	Pln infrastructure fee
RPTAG	Billing fee
RPTAG_MAT	Billing fee with stamp duty
RPREDUKSI	(Return/reduction) fee
RPANGS	Installment fee
RPBK	Late fee
DAYA_450_900	POWER (1 for 900, 0 for 450)

3.2.2. DBSCAN Clustering

Within the dataset, crucial information regarding the eligibility of individuals to receive subsidies is notably absent. This deficiency stems from the acknowledgment made in the introduction section, where it was highlighted that a significant number of individuals who do not meet the eligibility criteria still receive subsidies. Consequently, we contend that the inclusion of information regarding subsidy entitlement is not only unnecessary but also deemed invalid. Therefore, our dataset lacks labels, prompting the initiation of a clustering process. Selecting an appropriate clustering model poses a formidable challenge, as there is no universally acclaimed framework or method applicable to all types of problems. The intricate nature of this decision necessitates a thorough investigation tailored to the specifics of the problem at hand and the characteristics of the dataset [8]. To inform our choice of a clustering method, we adopt a multi-faceted approach. Initially, we endeavor to discern the underlying structure of the dataset by transforming the high-dimensional data into a more manageable lower-dimensional space, employing Principal Component Analysis (PCA) [9]. The resulting features generated by PCA are then visualized in scatter plots, differentiating the power groups (450 and 900) and excluding the DAYA_450_900 column to facilitate individualized clustering analyses for each power group. Central to our criteria for clustering model selection is an evaluation of how well the clustering algorithm aligns with the shape of the dataset as visualized in the lower-dimensional space. Moreover, to enhance the comparability of feature values, we perform normalization using both standard normalization and minmax normalization. Comprehensive details of these visualizations are provided in Table 2.

Table 2. Visualization result

	Minmax	Standard
450 VA		
900 VA		

Upon scrutinizing the visual representation delineated in Table 2, it becomes evident that the data displays a non-convex nature. In the context of this discourse, an affine space across the real numbers, denoting the clusters, is deemed convex when, for any pair of points within the set, the entire line segment connecting them lies entirely within that set [10]. The non-convex nature observed in the dataset excludes several viable clustering methodologies, such as the density-based and hierarchical approaches. Both methods possess the capability to handle clusters with arbitrary shapes; however, the hierarchical approach tends to incur a relatively higher time complexity [11]. Consequently, our preference leans towards the density-based approach, specifically opting for DBSCAN (Density Based Spatial Clustering of Applications with Noise) as our clustering model. DBSCAN has demonstrated considerable efficacy in discerning clusters characterized by non-convex shapes. The DBSCAN algorithm hinges on two key hyperparameters, namely epsilon (ε) and minimum points. Here, epsilon defines the radius of the neighborhood, and minimum points dictates the minimum number of data points required within a neighborhood to form a cluster [12]. For a more in-depth understanding, please refer to Figure 2 for the pseudocode outlining the DBSCAN algorithm.

```
function DBSCAN(Dataset, epsilon, minPoints):
   C = 0

   for each unvisited data point P in Dataset:
      mark P as visited
      neighborPts = regionQuery(P, epsilon)

      if size(neighborPts) < minPoints:
         mark P as Noise
      else:
```

```
        C = nextCluster(C)
        expandCluster(P, neighborPts, C, epsilon, minPoints)

function expandCluster(P, neighborPts, C, epsilon, minPoints):
   add P to cluster C

   for each neighbor Q in neighborPts:
      if Q is not visited:
         mark Q as visited
         neighborPts_Q = regionQuery(Q, epsilon)

         if size(neighborPts_Q) >= minPoints:
            neighborPts = neighborPts joined with neighborPts_Q
      if Q is not yet member of any cluster
         add Q to cluster C

function regionQuery(P, epsilon):
   return all data points within distance epsilon from point P
```

Figure 2. DBSCAN pseudocode

To determine the optimal values for the hyperparameters, we undertake a grid search process, systematically combining a range of values for both epsilon and minimum points. Subsequently, we execute the DBSCAN clustering algorithm for each unique combination of hyperparameters. We conduct clustering analysis utilizing the normalized version of the dataset, employing both standard and minmax normalization. This choice is motivated by the utilization of Euclidean distance metrics to determine neighborhood relationships. The specific values assigned to each hyperparameter combination are detailed in Table 3.

Table 3. Values of hyperparameter

Hyperparameter	Values
Min Pts	[3, 5, 7, 9, 11, 13, 15]
ε	[0.2, 0.5, 0.7, 1.0, 1.2, 1.5, 1.7, 2.0]

3.2.3. Clustering Evaluation

To assess the optimal clustering outcome, we integrate various metrics, including the Davies-Bouldin Index [13] and Silhouette Score [14]. Additionally, we perform an evaluation based on visualization, leveraging the principal components generated by the PCA algorithm for low-dimensional representation. It's important to note that PCA is employed solely for visualization purposes, while the clustering process utilizes the original features. The rationale behind visualization-based assessment lies in the importance of human visual perception as the benchmark for evaluating clustering algorithms in clustering analysis [15]. The amalgamation of quantitative metrics and visual judgment forms the foundation for selecting the most favorable clustering result. Comprehensive information for each quantitative metric is available in Table 4.

Table 4. Quantitative evaluation metrics details

Davies-Bouldin Index	$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$	0 to ∞	Lower is better
Silhouette Score	$s(i) = \frac{b(i) - a(i)}{\max\big(b(i),\ a(i)\big)}$	-1 to +1	Greater is better

3.2.4. Decision Tree Learning

Following the attainment of the optimal clustering outcome, we employ a supervised learning algorithm, specifically the decision tree learning algorithm. For every version of the normalized dataset, we execute an inverse transformation to obtain a meaningful criterion for splitting. For clarity, Figure 3 illustrates the specifics of the data flow from the clustering phase to the decision tree learning phase.

In the concluding step, a post-pruning process is conducted to enhance generalization and to map labels from the clustering results into binary categories (eligible and not eligible). The primary objective of pruning is twofold: to diminish the tree's complexity and to enhance generalization performance [7]. We prune the unnecessary node that does not satisfy the domain logic and also considering the impurity value (where lower impurity values are more likely to result in pruning). Additionally, we consolidate the pruned decision trees to create the final decision tree.

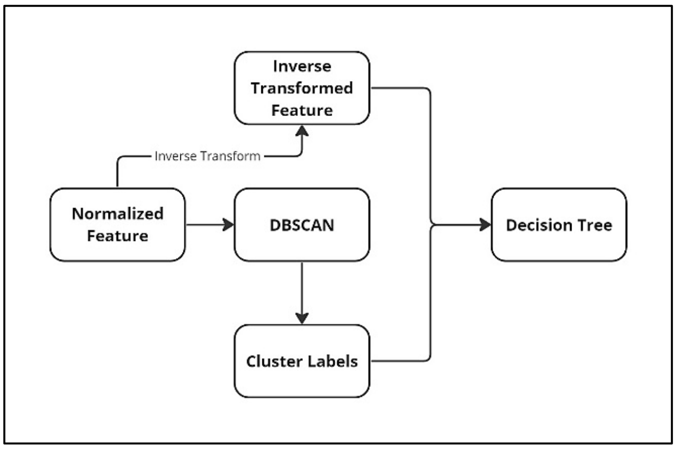


Figure 3. Data flow from clustering to decision tree learning

4. RESULTS AND ANALYSIS

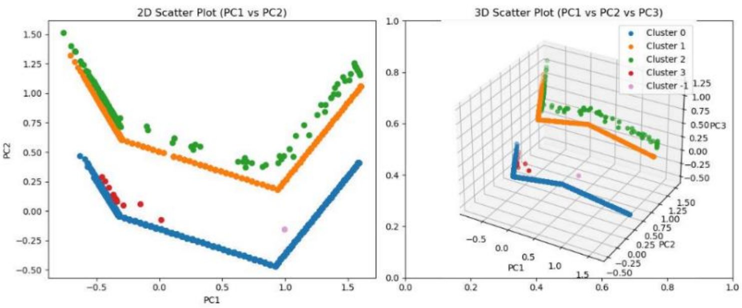
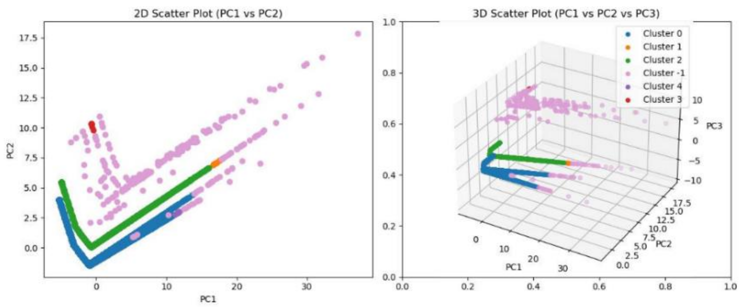
In this section, we'll go over the outcomes from the experiment, covering both the DBSCAN clustering and decision tree learning stages.

4.1. DBSCAN Clustering Result

We obtain the clustering outcomes for two distinct tariff groups: the 450 VA power group and the 900 VA power group. Within each group, we analyze two normalized versions of the dataset—namely, the standard version and the minmax version. We assess these different normalized versions independently due to potential variations in evaluation scores resulting from the different scales of values (standard and minmax versions having different value scales). Initially, we showcase the outcomes by highlighting the optimal clustering result determined solely based on evaluation metrics, excluding visual judgment. Subsequently, we reveal our preferred clustering result, considering both the evaluation metrics and visual judgment. In this study, we only show the best clustering result within each power group and normalization version, the rest can be seen in our repository. For the visualization can be seen in Table 5 and for the hyperparameters and scores can be seen in

Table **6**.

Table 5. DBSCAN result visualization

Power Group	Normalization	Visualization
450 VA	Minmax	
	Standard	

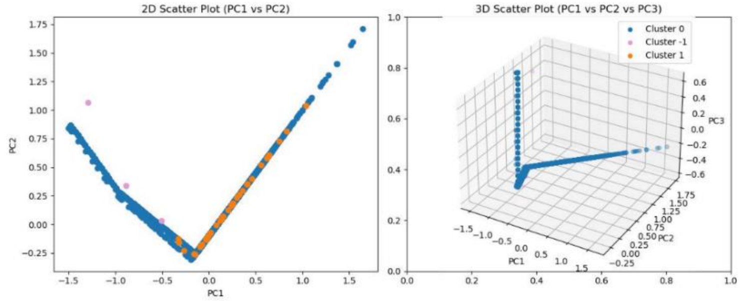
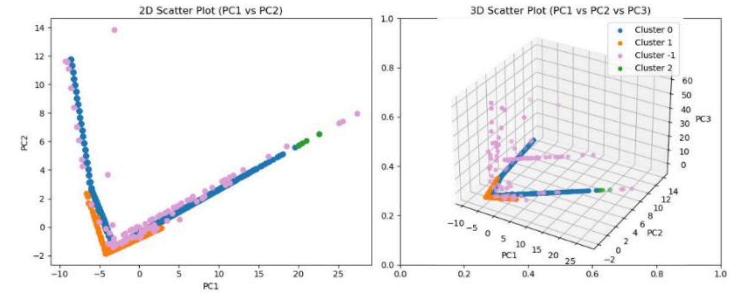
Power Group	Normalization	Visualization
900 VA	Minmax	
	Standard	

Table 6. Hyperparameter and evaluation score

Power Group	Normalization	Hyperparameter		Evaluation Score	
		Epsilon (ε)	Min Pts	DBI	Silhouette
450 VA	Minmax	0.7	9	1.1859	0.4221
	Standard	0.7	11	1.0644	0.4973
900 VA	Minmax	0.7	3	1.1134	0.6993
	Standard	1.7	7	1.4092	0.6764

1. **450-Minmax**
   In the 450 VA power group, the minmax normalized version yielded a Silhouette Score that closely resembled our preferred result. However, an issue arises where certain data points, highlighted in pink at the upper left of the 2D visualization, are misclassified. Regarding the version determined as the best by the Davies-Bouldin Index, two stick-like data points are deemed as belonging to the same cluster, contrary to our judgment, which suggests they should be separate clusters.

2. **450-Standard**
   Our preferred clustering method for the 450 VA power group, employing the standard normalization version, aligns closely with the cluster identified as the best by the Davies-Bouldin Index. However, concerning the cluster identified as the best by the Silhouette Score, there is an abundance of small-sized clusters in the upper part of the visualization. This outcome contradicts our judgment, which suggests that these data points should be considered as a single cluster.

3. **900-Minmax**
   Evaluating the clustering outcome for the 900 VA power group using the minmax normalized version proves challenging. The optimal result from quantitative evaluation metrics tends to yield a singular cluster with just one outlier. After a comprehensive examination of all clustering results, we conclude that our preferred clustering approach yields a superior outcome. In our assessment, the orange data points (cluster 1) are perceived as the ones immersed in the deep z-axis in the 3D visualization.

4. **900-Standard**
   In the case of the 900 VA power group utilizing the standard normalized dataset, the clustering outcome deemed best by the Silhouette Score closely resembles our preferred clustering approach. However, this top Silhouette Score result exhibits suboptimal clustering for the small, orange-colored cluster (cluster 1). Conversely, the result identified as the best by the Davies-Bouldin Index introduces excessive noise and substantial overlap between data points belonging to different clusters.

4.2. Decision Tree Learning Result

We generate four distinct decision tree classifiers for each power group, utilizing the normalized version of cluster labels. We illustrate the pruning process for each decision tree and subsequently combine all the pruned decision trees to construct the final tree. Figure 4 displays the decision trees prior to pruning for the 450 VA power group and Figure 6 for the 900 VA power group. The pruning steps are visually depicted in Figure 5 for the 450 VA power group and in Figure 7 for the 900 VA power group. The ultimate decision tree is showcased in Figure 8.

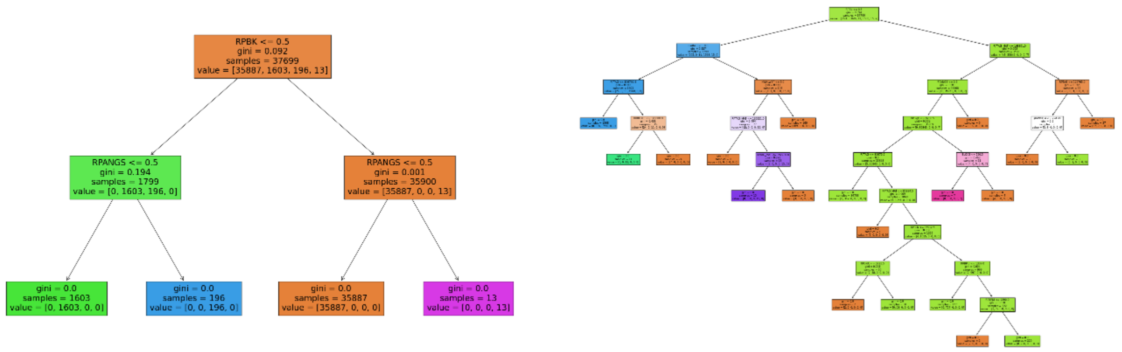


Figure 4. Decision trees for the 450 VA power group
(minmax cluster labels left, standard cluster labels right)

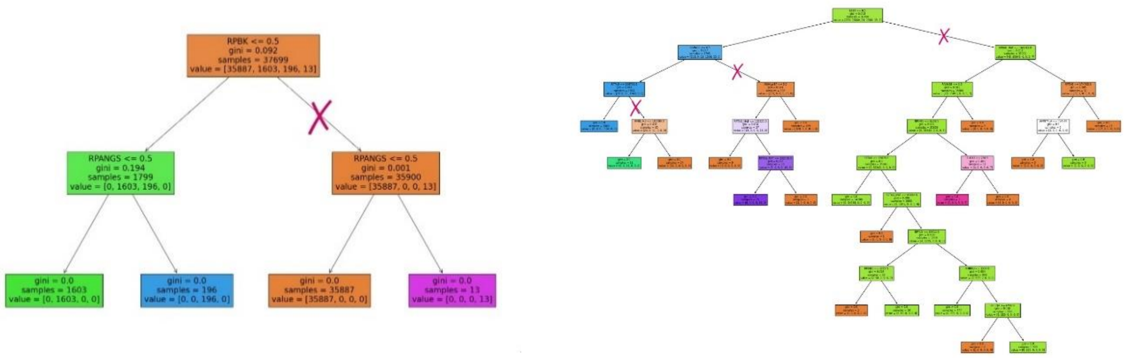


Figure 5. Pruning process for the 450 VA power group's decision trees
(minmax cluster labels left, standard cluster labels right)

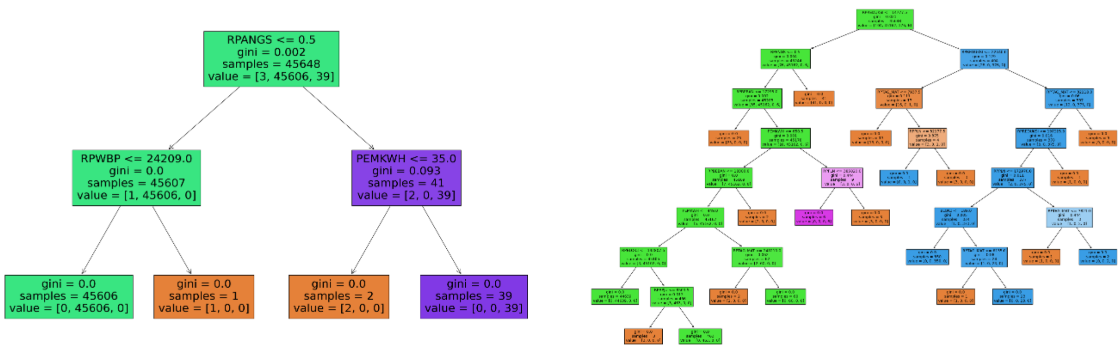


Figure 6. Decision trees for the 900 VA power group,
(minmax cluster labels left, standard cluster labels right)

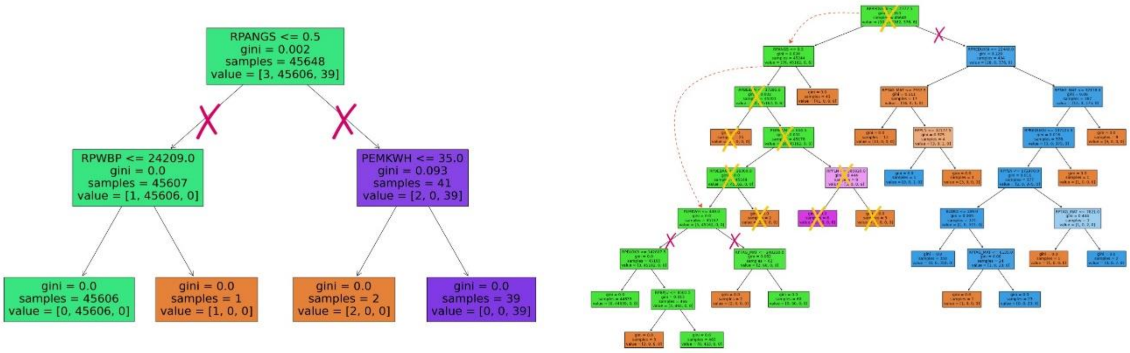


Figure 7. Pruning process for the 900 VA power group's decision trees (minmax cluster labels left, standard cluster labels right)

Our primary focus during pruning is on internal nodes exhibiting extremely low impurity, measured by the Gini index. These low impurity nodes often contain imbalanced samples, such as 40000 vs 1 or 35000 vs 13. Following the consideration of pruning low impurity internal nodes, we proceed to prune or eliminate internal nodes that do not align with domain logic. Notably, we observe similarities in the decision-making process among different trees generated from various normalized versions of cluster labels. For instance, in the 450 VA power group, both the minmax and standard versions of the decision tree begin by evaluating the value of RPBK and subsequently determine the RPANGS value. Ultimately, we consolidate the decision trees and remap the cluster labels to binary labels signifying eligibility. It's crucial to emphasize that nodes involving currency-related values like RPTAG should be adapted to the currency's value at the time of using the decision tree. Additionally, features such as PEMKWH (KWH USAGE) may need adjustments in the presence of inflation-like trends within the feature.

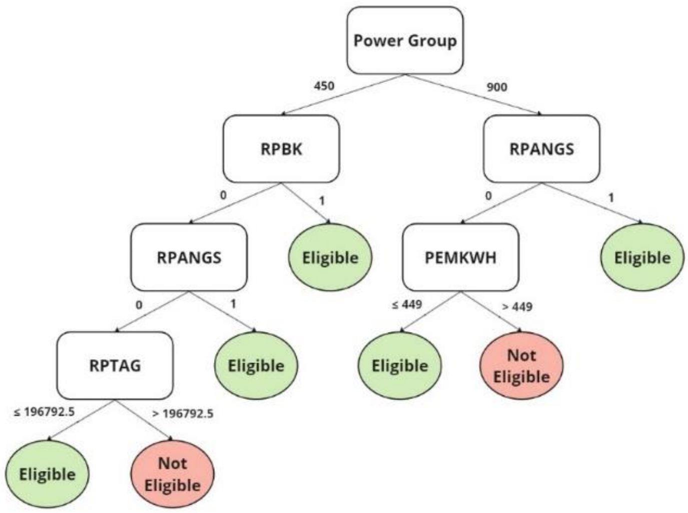


Figure 8. Final decision tree

5. CONCLUSION

In conclusion, employing post-paid household customer data allows for the redefinition of criteria determining individual eligibility for electrical subsidies. Among the 50 features available in the raw dataset, we successfully identify key features—RPBK, RPANGS, PEMKWH, and RPTAG—to serve as crucial criteria. In addition to constructing a machine learning model based on the given dataset, we ascertain both the eligibility label and novel criteria through the clustering and decision tree learning procedures. Employing a straightforward model like the decision tree facilitates domain experts in reassessing prevailing criteria and contrasting them with our proposed ones. Nonetheless, considerable prospects for improvement persist, notably through the exploration of time series data to assess customer behavior for eligibility determination. Additionally, considering external features unrelated to post-paid household data, such as general household consumption and house type, presents avenues for improvement.

REFERENCES

[1] R. K. Gobel, B. S. Laksmono, M. Huseini, and M. Siscawati, "Equity and Efficiency: An Examination of Indonesia's Energy Subsidy Policy and Pathways to Inclusive Reform," *Sustainability*, vol. 16, no. 1, p. 407, Jan. 2024, doi: 10.3390/su16010407.

[2] Tim Nasional Percepatan Penanggulangan Kemiskinan (TNP2K), *Policy Paper (Naskah Kebijakan) Transformasi Subsidi Listrik di Indonesia: Tujuan dan Usulan Mekanismenya*, 1st ed. 2021. [Online]. Available: www.tnp2k.go.id

[3] N. Noor Hanan and M. S. Fuady, "REVIEW OF ELECTRICITY SUBSIDIES IN INDONESIA 2015-2020," *Jurnal Pajak dan Keuangan Negara*, vol. 4, no. 2, pp. 529–538, 2023.

[4] I. Firman Ashari, E. Dwi Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *Journal of Applied Informatics and Computing (JAIC)*, vol. 7, no. 1, pp. 2548–6861, 2023, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[5] W. Y. Widiawati and R. M. Atok, "Analisis Klasifikasi Pelanggan Listrik Rumah Tangga Bersubsidi Kota Surabaya Menggunakan Support Vector Machine dan Naïve Bayes Classifier," *JURNAL SAINS DAN SENI ITS Vol. 7, No. 2*, vol. 7, no. 2, pp. D235–D242, 2018.

[6] Y. M. Hutahaean and A. W. Wijayanto, "Electricity Subsidy Recipient Households Classification in Gorontalo Province in 2019 using K-Nearest Neighbor and Support Vector Machine," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 10, no. 1, p. 63, Jan. 2022, doi: 10.26418/justin.v10i1.51210.

[7] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: an updated survey," *Artif Intell Rev*, vol. 56, no. 5, pp. 4765–4800, May 2023, doi: 10.1007/s10462-022-10275-5.

[8] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng Appl Artif Intell*, vol. 110, p. 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[10] V. Soltan, *Lectures on Convex Sets*. WORLD SCIENTIFIC, 2020. doi: 10.1142/11328.

[11] S. Pitafi, T. Anwar, and Z. Sharif, "A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms," *Applied Sciences*, vol. 13, no. 6, p. 3529, Mar. 2023, doi: 10.3390/app13063529.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *kdd*, 1996, [Online]. Available: www.aaai.org

[13] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans Pattern Anal Mach Intell*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[15] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggag, and H. Bensmail, "Toward Perception-Based Evaluation of Clustering Techniques for Visual Analytics," in *2019 IEEE Visualization Conference (VIS)*, IEEE, Oct. 2019, pp. 141–145. doi: 10.1109/VISUAL.2019.8933620.

BIBLIOGRAPHY OF AUTHORS

Randy Zakya Suchrady is a skilled AI Engineer and Research Assistant with a strong background in informatics. He holds a Bachelor's degree in Informatics from the Bandung Institute of Technology in Indonesia and is currently pursuing a Master's degree specializing in Intelligent System. Randy has a keen interest in artificial intelligence and has gained extensive experience in the field through various internships. He has worked as an AI Engineer Intern at GLAIR, Text AI Engineer Intern at Prosa.ai, and Python Developer at Ctscope. Randy's impressive work experience and education background make him a valuable asset to any organization in the field of artificial intelligence.

Adi Firmansyah is a professional with a Bachelor's degree in Informatics Engineering from the Electronic Engineering Polytechnic Institute of Surabaya (EEPIS). He currently resides in South Jakarta, Jakarta, and has previously worked at PT PLN (Persero). Currently, he is a work assignment employee at PT ICON Plus, a subsidiary of PLN. In addition to his professional endeavors, Firman is also pursuing his Masters in Informatics at ITB (Bandung Institute of Technology), specializing in Data Science and Artificial Intelligence. He is deeply committed to improving his knowledge and skills in the field of Informatics Engineering, with a particular focus on the application of data science and artificial intelligence, as evidenced by his academic activities and professional experience.

Nugraha P. Utama is a lecturer at the Informatics Department, School of Electrical Engineering and Informatics, Institute Technology Bandung. He has a background in computational Intelligence and Neuroscience from Tokyo Institute of Technology, and National Center of Neurology and Psychiatry, Japan. His current work focuses on artificial intelligence and computer vision in all ways of applications.

Rahman Indra Kesuma is a lecturer at Institut Teknologi Sumatera (ITERA) in the Informatics Engineering Undergraduate Program. Competence and mastery in artificial intelligence, expert systems, decision support systems, machine learning, and deep learning. He had a bachelor’s degree in computer science from Lampung University, Indonesia, and a master's degree in computer science from Gadjah Mada University in the same country. Currently, he is focusing on dissertation research in video anomaly detection and prediction with self-supervised learning at the School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia.