APPLICATION OF THE DICISION TREE ALGORITHM FOR CLASSIFICATION HOUSEHOLD ELECTRIC ENERGY CONSUMPTION USES RAPID MINER

Ubeitul Maltuf[a1], Zaehol Fatah[a2]
[a]Information Systems Study Program, Faculty of Science and Technology, Ibrahimy Universitas, East Java, Indonesia-68374
e-mail: [1]ubeitulmaltuf@gmail.com , [2]zaeholfatah@gmail.com

Abstrak

The application of the Decision Tree algorithm for classification of household electrical energy consumption using RapidMiner is an innovative approach in energy data analysis. Decision Tree is an effective machine learning technique for modeling the relationship between features and labels in a dataset. By utilizing electricity consumption data that includes variables such as equipment type, time of use, and household characteristics, this algorithm can identify significant patterns in energy consumption. This research aims to implement Decision Tree in RapidMiner to classify energy consumption levels (low, medium, high) in households. The research process includes data collection, preprocessing, data separation into training and test sets, and training models. The performance evaluation model is carried out to measure accuracy and effectiveness classification. The results show that the Decision Tree is able to provide accurate predictions and is useful in understanding the factors that influence energy consumption. It is hoped that this research can provide guidance for the public and policy makers in managing energy consumption more efficiently and sustainably.

Kata kunci: Implementation, Decision Tree Algorithm, Classification, Energy Consumption, Household, RapidMiner, Machine Learning, Preprocessing, Training Set, Test Set, Performance Evaluation, Energy Efficiency, Sustainable.

Abstract

The application of the Decision Tree algorithm for classification of household electrical energy consumption using RapidMiner is an innovative approach in energy data analysis. Decision Tree is an effective machine learning technique for modeling the relationship between features and labels in a dataset. By utilizing electricity consumption data that includes variables such as equipment type, time of use, and household characteristics, this algorithm can identify significant patterns in energy consumption. This research aims to implement Decision Tree in RapidMiner to classify energy consumption levels (low, medium, high) in households. The research process includes data collection, preprocessing, data separation into training and test sets, and training models. The performance evaluation model is carried out to measure accuracy and effectiveness classification. The results show that the Decision Tree is able to provide accurate predictions and is useful in understanding the factors that influence energy consumption. It is hoped that this research can provide guidance for the public and policy makers in managing energy consumption more efficiently and sustainably.

Keywords : Implementation, Decision Tree Algorithm, Classification, Energy Consumption, Household, RapidMiner, Machine Learning, Preprocessing, Training Set, Test Set, Performance Evaluation, Energy Efficiency, Sustainable.

INTRODUCTION

Electrical energy consumption in households is an important aspect in sustainable energy resource management. In many countries, increasing household energy demand is a major challenge that needs to be faced, especially in the context of sustainability and energy efficiency. Therefore, a deep understanding of energy consumption patterns is necessary to develop effective energy saving strategies. One method that can be used to analyze and predict energy consumption is the Decision Tree algorithm. Decision Tree is a machine learning technique that is able to model complex relationships between input and output variables in a way that is easy to understand. This method automatically divides data into categories based on relevant features, so that it can be used for classification or regression. The main advantage of Decision Tree is its ability to provide clear interpretation and intuitive visualization, making it an attractive choice for energy data analysis Loh, 2011 [1]. In this context, RapidMiner as a powerful data analysis tool offers various features for implementing algorithms Decision Tree efficiently. RapidMiner allows users to preprocess data, build models, and evaluate their performance in one platform

integrated. Through this research, we aim to apply the Decision Tree algorithm in RapidMiner to classify household electricity consumption levels into low, medium and high categories, as well as to explore the factors that influence these consumption patterns. Thus, it is hoped that this research can contribute to the understanding of electrical energy management in households, as well as become a basis for developing better policies for energy savings.

METHOD

Types of research

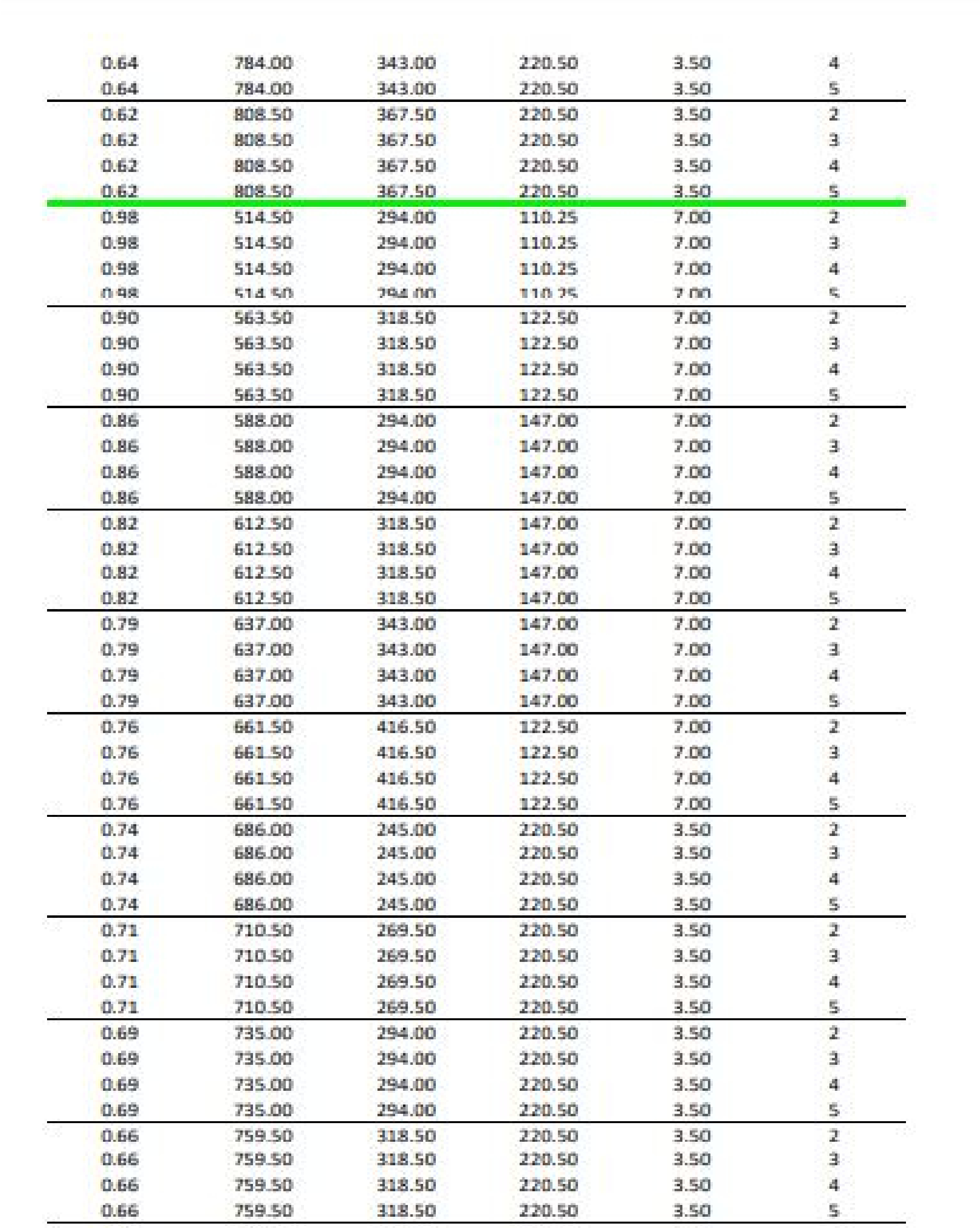
Applied research aims to solve practical problems by applying existing theories and methods in real situations. In this context, the research focuses on the application of the Decision Tree algorithm for the classification of household electrical energy consumption, which is an important issue in efficient energy management. A quantitative approach is used to collect and analyze numerical data. Data is obtained through surveys involving structured questions related to energy use, so that analysis can be carried out statistically to find patterns and relationships between variables. This research design is exploratory, where the research aims to explore and understand energy consumption patterns in households. By using the Decision Tree algorithm, this research attempts to identify significant factors that influence energy consumption levels. This research can also be considered a case study, because the analysis was carried out on a certain number of households. This allows researchers to gain in-depth insight into energy consumption patterns in the local context and provide appropriate recommendations for management

Method of collecting data

Data on household electrical energy consumption can be obtained from various sources, such as household electricity meters. Surveys or questionnaires filled out by homeowners regarding the use of electrical appliances. Data from electricity supply companies. Primary Data Collection Design a questionnaire that includes questions regarding the type of electrical equipment used (for example, lights, AC, refrigerator, etc.) Duration of use of electrical equipment per day. Number of family members. Electricity usage habits (e.g. peak usage times) Direct Observations Carry out direct

observations in selected homes to obtain accurate data on electricity usage.Pengumpulan Data Sekunder Use data from secondary sources such as statistics from government agencies regarding energy consumption. Annual reports from electricity companies. Previous research relevant to the topic Data Processing and Cleaning Once the data is collected, perform data cleaning to remove duplicates,

errors, or irrelevant data. Transform data into a format suitable for analysis, such as changing categorical variables to numeric if necessary Splitting the Dataset Divide the dataset into two parts training data (training) and test data (testing) to test the accuracy of the model being built Implementation in RapidMiner import dataset into RapidMiner Use the Decision Tree operator to build a classification model. Evaluate the model using test data and calculate metrics such as accuracy, precision, recall, and F1-score. There is a picture feature about the data (Application of the Decision Tree Algorithm to Classify Household Electrical Energy Consumption Using Rapid Miner. This is an image data about the Application of the Decision Tree Algorithm to Classify Household Electrical Energy Consumption Using Rapid Miner taken from the dataset



0.64	784.00	343.00	220.50	3.50	4
0.64	784.00	343.00	220.50	3.50	5
0.62	808.50	367.50	220.50	3.50	2
0.62	808.50	367.50	220.50	3.50	3
0.62	808.50	367.50	220.50	3.50	4
0.62	808.50	367.50	220.50	3.50	5
0.98	514.50	294.00	110.25	7.00	2
0.98	514.50	294.00	110.25	7.00	3
0.98	514.50	294.00	110.25	7.00	4
0.98	514.50	294.00	110.25	7.00	5
0.90	563.50	318.50	122.50	7.00	2
0.90	563.50	318.50	122.50	7.00	3
0.90	563.50	318.50	122.50	7.00	4
0.90	563.50	318.50	122.50	7.00	5
0.86	588.00	294.00	147.00	7.00	2
0.86	588.00	294.00	147.00	7.00	3
0.86	588.00	294.00	147.00	7.00	4
0.86	588.00	294.00	147.00	7.00	5
0.82	612.50	318.50	147.00	7.00	2
0.82	612.50	318.50	147.00	7.00	3
0.82	612.50	318.50	147.00	7.00	4
0.82	612.50	318.50	147.00	7.00	5
0.79	637.00	343.00	147.00	7.00	2
0.79	637.00	343.00	147.00	7.00	3
0.79	637.00	343.00	147.00	7.00	4
0.79	637.00	343.00	147.00	7.00	5
0.76	661.50	416.50	122.50	7.00	2
0.76	661.50	416.50	122.50	7.00	3
0.76	661.50	416.50	122.50	7.00	4
0.76	661.50	416.50	122.50	7.00	5
0.74	686.00	245.00	220.50	3.50	2
0.74	686.00	245.00	220.50	3.50	3
0.74	686.00	245.00	220.50	3.50	4
0.74	686.00	245.00	220.50	3.50	5
0.71	710.50	269.50	220.50	3.50	2
0.71	710.50	269.50	220.50	3.50	3
0.71	710.50	269.50	220.50	3.50	4
0.71	710.50	269.50	220.50	3.50	5
0.69	735.00	294.00	220.50	3.50	2
0.69	735.00	294.00	220.50	3.50	3
0.69	735.00	294.00	220.50	3.50	4
0.69	735.00	294.00	220.50	3.50	5
0.66	759.50	318.50	220.50	3.50	2
0.66	759.50	318.50	220.50	3.50	3
0.66	759.50	318.50	220.50	3.50	4
0.66	759.50	318.50	220.50	3.50	5

Figure 1. Household Electrical Energy Consumption Classification Dataset

Data Mining

Data mining is the process of discovering significant correlations, patterns and trends by analyzing large amounts of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical methods. Data mining aims to uncover interesting patterns and information from large amounts of data. The term data mining is often associated with other terms such as knowledge discovery or pattern recognition. The term knowledge discovery is considered appropriate because the main goal of data mining is to uncover hidden knowledge in data. Meanwhile, the term pattern recognition is suitable because this process focuses on discovering patterns stored in large data sets. Teknologi et al. 2021). [2]

Classification

Classification is the process of assessing data to place it into one of the available classes. In this process, a model is created based on training data which is then used to classify new data. Classification can be interpreted as a learning or training process for a target function that maps each set of attributes (features) to one of the existing classes. The goal of a classification system is to classify all data accurately, but its performance is not always 100% perfect. Therefore, measurements are needed to evaluate the performance of the classification system, which is usually carried out using the Arrohman and Fatah 2024 confusion matrix [3]

Decision Tree Algorithm

This algorithm is also used in random forests to train on different subsets of training data and achieve more accurate results. Apart from making decisions easier and providing a clear picture of the reasons why decisions were made, decision trees are also very useful for data mining. Advantages of Decision Trees Because of their simple structure, decision trees are one of the fastest methods for identifying significant variables and the relationship between two variables. Besides that,

a. Advantages of Decision Trees

It is flexible because it can be used for classification and regression tasks, and is suitable for handling various types of data such as discrete data, continuous data, and categorical data. Easy to understand, especially for people without an analytical background, because decision trees follow the same process as do humans when making decisions in the real world. The hierarchical nature of decision trees makes it easy for analysts to see which attributes are most important. Helps analysts think of all possible solutions to solve a problem. Decision trees require less data cleaning to correct errors or inconsistencies compared to other algorithms.

b. Disadvantages of Decision Trees

Decision trees have many layers so they seem complex and prone to overfitting, that is, the algorithm does not provide accurate predictions for new training data. Not fully supported by Scikit-learn, a popular Python-based machine learning library. Not ideal for large datasets because decision trees take a long time to train and are more expensive, and complexity can increase. Example of a Decision Tree Without realizing it, decision trees are often used in everyday life for both simple and complex decision making. The image below provides an illustration of the Decision Tree algorithm according to this definition.

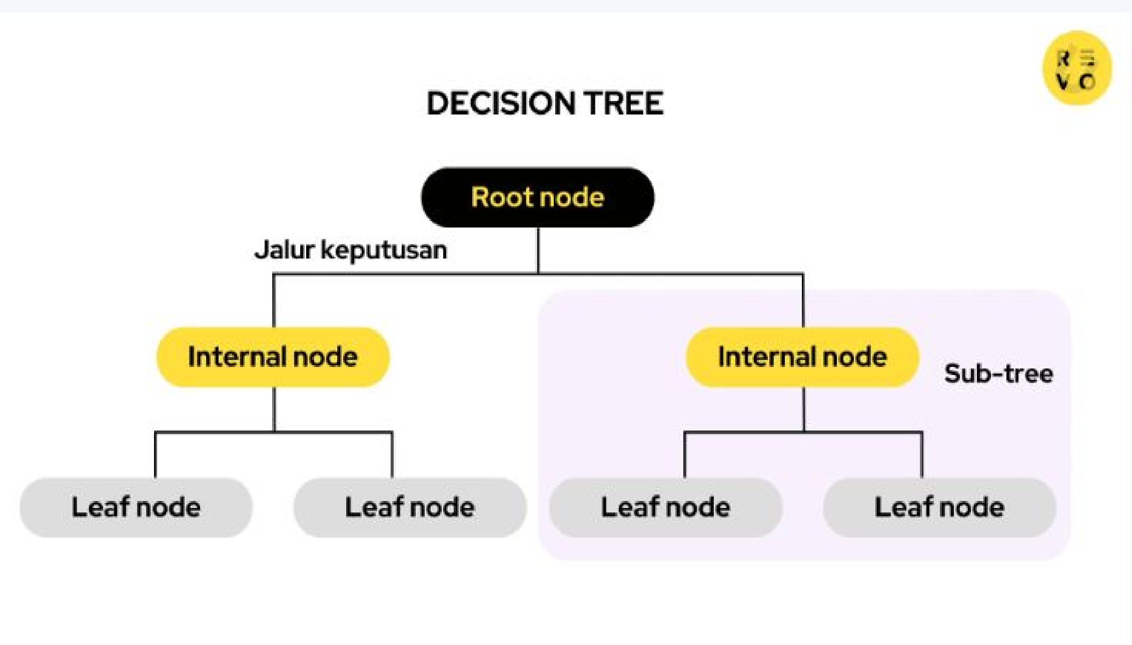


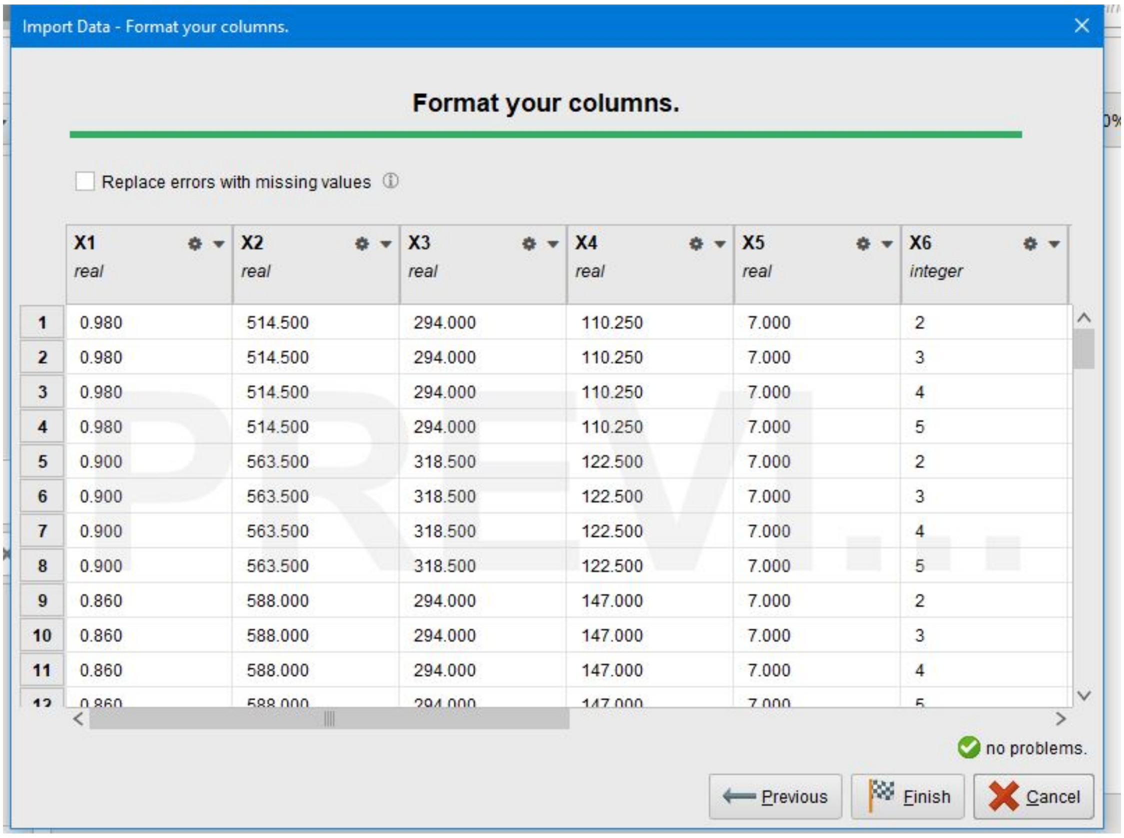
Figure 2. Decision Tree Algorithm Scheme

Data Quality Risk Factors

Description: Incomplete, inaccurate, or inconsistent data can lead to bad models. Missing data or outliers can affect classification results. Mitigation: Perform data cleaning and processing before analysis. Overfitting Decision Tree models that are too complex can learn too much from training data, so they do not generalize well to test data. Using pruning techniques to simplify the model and reduce tree complexity. Selecting irrelevant or less informative features can reduce model performance. Too many features can also cause noise. feature analysis and selection of relevant features before building the model. If the data is not representative of the larger population, the resulting model will be biased and inaccurate. Use appropriate sampling techniques and ensure the representativeness of the data. Electrical energy consumption can be influenced by many external factors such as weather, day of the week, and special events, which may not be represented in the data. Collect additional data that includes external variables. Electricity usage patterns can change over time, which can make the model less relevant if not updated. Conduct modeling regularly with the latest data to ensure accuracy. Errors in RapidMiner usage or model configuration may result in incorrect results. Ensure a good understanding of the tools and techniques used, and validate the results.

RESULTS AND DISCUSSION

This research was carried out using one of the classification models, namely Random Forest. This model was implemented using RapidMiner version 10.3 software to simplify the data analysis process. The dataset used in this research was obtained from the Kaggle site. This data contains relevant attributes for classifying individuals based on risk for the Classification of Household Electrical Energy Consumption Using Rapid Miner with categories of individuals who are obese or not, which can be used as the main parameter. The process of using visualization in RapidMiner is carried out through the following steps



Import Data - Format your columns.

Format your columns.

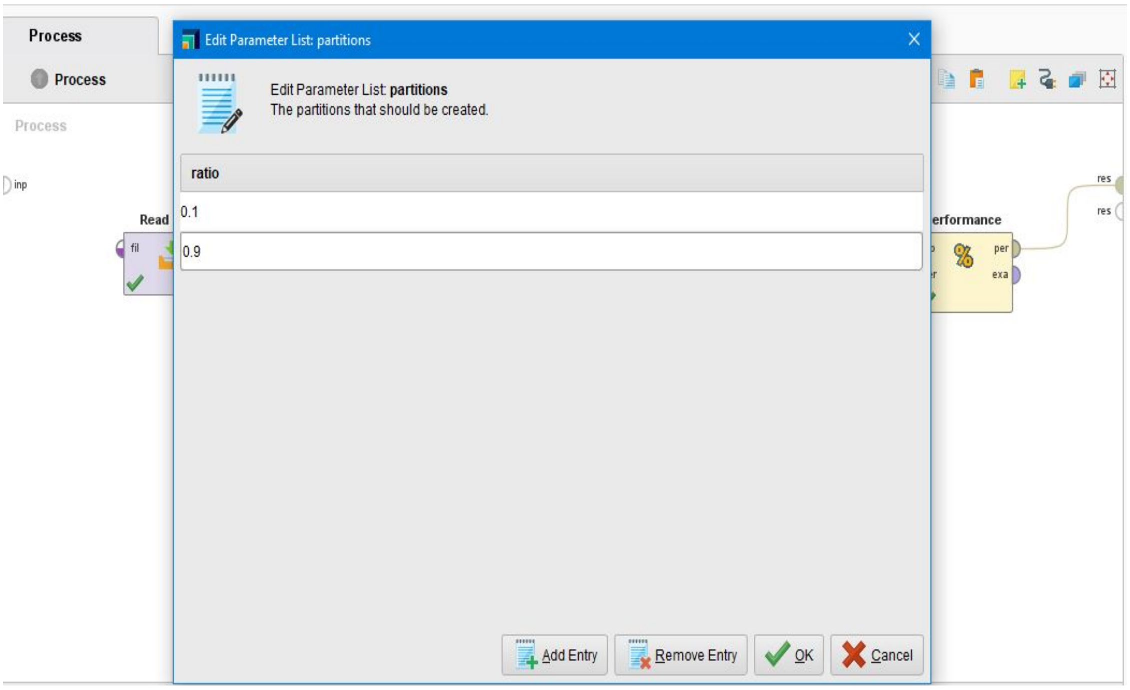
☐ Replace errors with missing values

	X1 real	X2 real	X3 real	X4 real	X5 real	X6 integer
1	0.980	514.500	294.000	110.250	7.000	2
2	0.980	514.500	294.000	110.250	7.000	3
3	0.980	514.500	294.000	110.250	7.000	4
4	0.980	514.500	294.000	110.250	7.000	5
5	0.900	563.500	318.500	122.500	7.000	2
6	0.900	563.500	318.500	122.500	7.000	3
7	0.900	563.500	318.500	122.500	7.000	4
8	0.900	563.500	318.500	122.500	7.000	5
9	0.860	588.000	294.000	147.000	7.000	2
10	0.860	588.000	294.000	147.000	7.000	3
11	0.860	588.000	294.000	147.000	7.000	4
12	0.860	588.000	294.000	147.000	7.000	5

✓ no problems.

Figure 3. Data Dialing Operator

The Read Excel operator is used to load a dataset saved in Excel format. In this section, we also determine the attributes that will be used as labels in the dataset used.



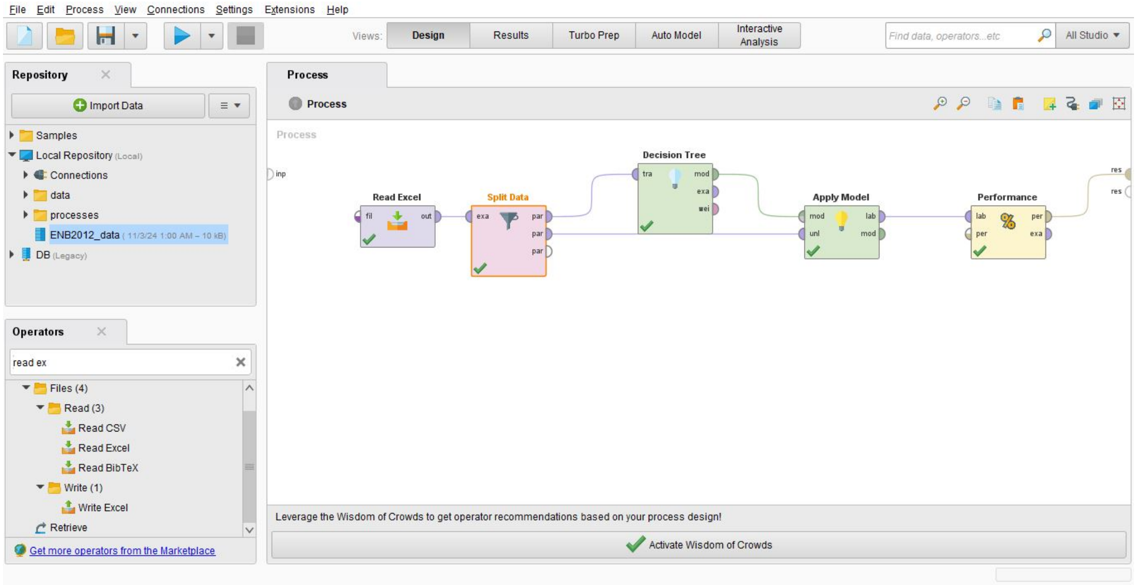
Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.1
0.9

Figure 4. Split Data

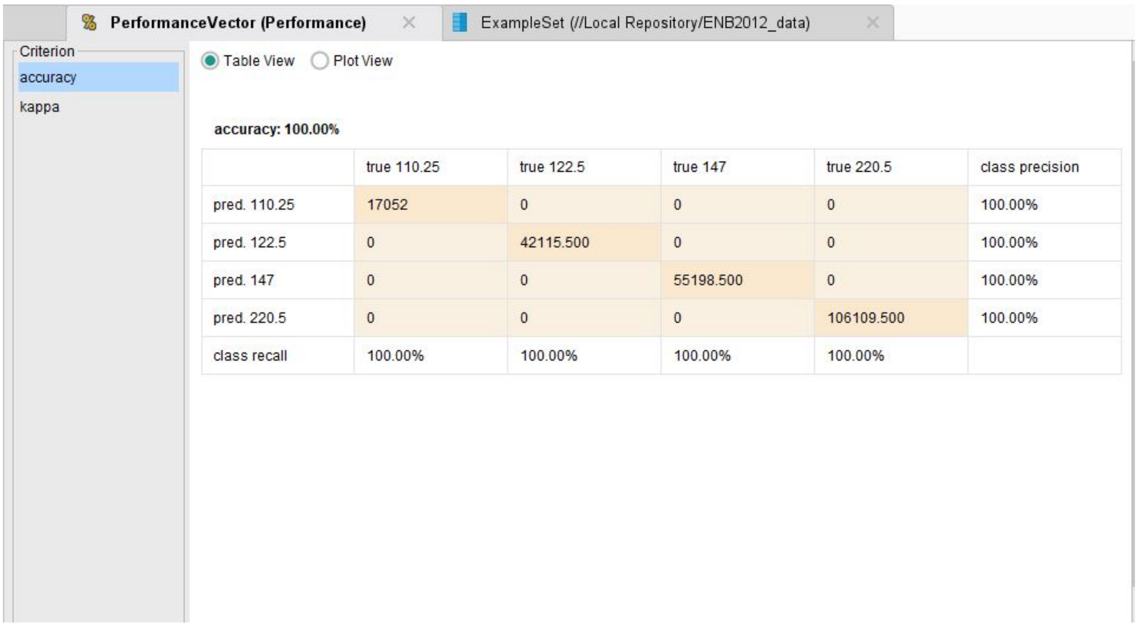
Split data functions to separate the dataset into training data and test data. Oktafiani, Hermawan, and Avianto 2023. [4] This step is very important to avoid overfitting and evaluate the model's ability to identify risk factors for energy consumption classification. household electricity. The Split Data operator can be used to split the dataset, for example 100% for training data and 10% for test data, to ensure that the Dicision Tree model can generalize well to new data, not just the data used for training.



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 5. Decision Tree Model

Implementation of data mining with the Decision Tree algorithm using the RapidMiner application version 10.3 is carried out through several steps as shown in the image above. The next stage is to connect all the operators that have been prepared. The Apply Model operator is an important component used to apply a trained model to a new dataset, allowing predictions or classifications to be made on unknown data based on a model built from the training data. After training a Decision Tree model using the training data, the Apply Model operator can be used to apply the model on test data to predict risk for classification of household electrical energy consumption. The Performance Operator is used to evaluate the performance of models that have been trained and implemented. It is an important tool for assessing the effectiveness of models in predicting or classifying risk factors for classification of household electrical energy consumption, by providing various metrics such as accuracy, precision, and recall to assess the quality of predictions produced by the model after being applied to test data Al-Giffary and Martanto 2024. [5]


PerformanceVector (Performance)
ExampleSet (//Local Repository/ENB2012_data)
Criterion
accuracy
kappa
● Table View
○ Plot View

accuracy: 100.00%

	true 110.25	true 122.5	true 147	true 220.5	class precision
pred. 110.25	17052	0	0	0	100.00%
pred. 122.5	0	42115.500	0	0	100.00%
pred. 147	0	0	55198.500	0	100.00%
pred. 220.5	0	0	0	106109.500	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	

Figure 6. Accuracy results

Based on the image above, the application of the Decision Tree algorithm in analyzing risk factors for the classification of household electrical energy consumption produces an accuracy value of 100.00%. From the displayed confusion matrix, we can see the distribution of predicted and actual values for various classes. For example, in the class "true 110 25," there are 17052 correct predictions. The evaluation results also show the precision and recall values for each class. The highest precision was achieved in the "true 2205" class with 100% recall, while the precision was found in the "true 122.5" class of 100.00%. Performance evaluation using these various metrics provides an idea of how well the Decision Model is at predicting household electrical energy consumption classification risk factors among the analyzed individuals. By considering these results, this model shows good ability in classifying data, although there are several classes that require improvement in terms of precision and recall which are quite accurate and have high accuracy in risk analysis for classification of household electrical energy consumption.

CONCLUSION

The results of the study show that several variables have a greater influence on the risk of classification of household electrical energy consumption and the existence of a history of classification of household electrical energy consumption in the family appears as a factor that increases the risk of household electrical energy. Based on analysis using the Decision Tree algorithm, it can be concluded that this model is effective and efficient in identifying the main factors that influence the risk of Household Consumption Classification. These results provide valuable insight for Household Electrical Energy Consumption policy makers and practitioners to design programs to make it even better. Although the results of this study provide strong evidence regarding the effectiveness of Decision Tree in analyzing risk factors for classifying household electrical energy consumption, this research still has several limitations. One of them is the limited size of the dataset used, which may affect the generalization of the results. Therefore, it is recommended to conduct a follow-up study using a larger and more varied dataset, and involving other variables that can contribute to the classification of household electrical energy consumption, such as household psychological and socio-economic factors. Overall, this research confirms that the Decision Tree algorithm is a useful and reliable tool for analyzing health data, especially in identifying risk factors for the classification of household electrical energy consumption. These findings can be used as a basis for efforts to prevent the classification of household electrical energy consumption so that it is more secure.

BIBLIOGRAPHY

[1] Loh, W. Y. (2011). "Classification and Regression Trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23.

[2] Teknologi, Jurnal et al. 2021. “Analisis Data Mining Untuk Clustering Kasus): 100–108.

[3] Arrohman, Supri, and Zaehol Fatah. 2024. “Gudang Jurnal Multidisiplin Ilmu Prediksi Diabetes Menggunakan Algoritma Klasifikasi K-Nearest Neighbors (K-NN) Pada Perempuan Indian Pima.” 2: 220–26.

[4] Oktafiani, Rian, Arief Hermawan, and Donny Avianto. 2023. “Pengaruh Komposisi Split Data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning.” Jurnal Sains dan Informatika 9(April): 19–28.

[5] Al-Giffary, Farhan Rizky, and Martanto Martanto. 2024. “Klasifikasi Kelulusan Siswa Tahun 2024 Menggunakan Metode Decision Tree (Studi Kasus Sma Islam Alazhar 5 Cirebon).” Jurnal Manajamen Informatika Jayakarta 4(2): 195.

[6] RapidMiner. (n.d.). "RapidMiner: Data Science Platform." Retrieved from https://rapidminer.com U.S. Energy Information Administration. (2020). "Residential Energy Consumption Survey (RECS)." Retrieved from https://www.eia.gov/consumption/residential

[7] Creswell, J. W. (2014). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications.

[8] Kothari, C. R. (2004). Research Methodology: Methods and Techniques. New Age International. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). Classification and Regression Trees. Wadsworth and Brooks/Cole.RapidMiner. (n.d.). "RapidMiner: Data Science Platform." Retrieved from https://rapidminer.com

[9] U.S. Energy Information Administration. (2020). "Residential Energy Consumption Survey (RECS)." Retrieved from https://www.eia.gov/consumption/residential Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann. Articles:

[10] Quinlan, J.R. (1986). "Induction of Decision Trees". Machine Learning, 1(1), 81-106. Liu, H., & Motoda, H. (2008). Computational Methods of Feature Selection. CRC Press.