IJISCS (International Journal of Information System and Computer Science)

Full Paper
eISSN 2598-246X
pISSN 2598-0793

THE COMPARISON USING EXPECTATION-MAXIMIZATION ALGORITHM AND C4.5 ALGORITHM TO PREDICT THE RESULT OF BIOGAS PRODUCTION AS A POWER PLANT AT PT BUDI STARCH & SWEETENER (BSSW)

Eriska Vivian Astuti[1], Nurmayanti[2], Rima Mawarni[3], Asep Afandi[4], Aris Munandar[5]
[1,3,4,5]Information System, ITBA-DCC PSDKU Kotabumi, Kotabumi
[2]Technology Computer, ITBA-DCC PSDKU Kotabumi, Kotabumi

*Corresponding author
Email:
eriskavivianastuti@gmail.com[1]
Nurmayanti89@mail.com[2]
rima@dcc.ac.id[3]
asep@dcc.ac.id[4]
Arismunandar@gmail.com[5]

Article history:
Received November 26, 2022
Revised December 10, 2022
Accepted December 28, 2022

Keywords:
Biogas,
waste,
data mining,
C4.5 Algorithm

Abstract
Biogas is the result of the development of alternative energy that has formed through the decomposition of organic matter through an anaerobic fermentation process (without oxygen) that produces gas in the form of methane gas (CH4) which has burned. Biogas is a kind of renewable energy because it has a high methane content and calorific value. Methane has one carbon in each chain, which can produce combustion that is more environmentally friendly when compared to fuels that have long carbon chains using specific calculation techniques or methods, a data mining process has been carried out to locate interesting patterns or information in selected data to manipulate the data into more valuable information by extracting significant patterns from the database.

1.0 INTRODUCTION

Biogas is the end product of anaerobic digestion or degradation of organic matter synthesized by anaerobic microorganisms in an airtight or oxygen-free environment. The majority of biogas is produced through anaerobic digestion (AD), making it a renewable energy source similar to solar or wind power. Although biogas production makes use of well-established technology, its commercial application is still limited due to the space requirement for purification before transportation or use[1]. Biogas production is increasingly in demand because the world's fuel reserves have experienced a fairly high risk. Anaerobic digesters to produce highly useful products from urban solid waste, which is silently abundant, lack research on design considerations that can lead to process optimization[2]. Metals and some recalcitrant organic compounds, many of which are toxic, are among the waste's components that are typically difficult to break down and accumulate into residues. While this is one of the most important considerations and should be divided into several accounts when implementing processing, it is a big topic and will not be reviewed here[3]. Bioenergy is energy that comes from biomass. Biomass has changed into natural material to get from living things like plants and creatures (earthbound and amphibian). Waste comes in many forms, including agricultural and forestry wastes; waste from homes, businesses, and factories; in additional to specific energy crops like jatropha, which are examples of bioenergy raw materials[4].

Biomass has transformed through fermentation, pyrolysis, gasification, or combustion into various bioenergy steals of the technology that has been converted will produce streams with different streams and concentrations of CO2 so that they can experience carbon capture. We use a system-wide optimization model that has worked to determine the conversion technology like the carbon capture rate, sequestration credits biogas formed from the biomass process (i.e., manure or any agricultural result by-products) produces methane-rich gas, resulting in a minimum break-even fuel cost for various capacities and usages[5]. As a result, methane accounts for only about 60%. The separation process typically also removes the purpose contaminant, CO2 (40 percent), and the biogas network that is already in place can be used to export the remaining methane[6]. The majority of biogas produced at AD-Plants or landfills consists of methane (CH2) and carbon dioxide (CO2), with oxygen (O2), nitrogen (N2), saturated or halogenated carbohydrates, and hydrogen (H2) occasionally present[7].

The mechanism that methane-producing biocathodes use to produce transported methane at low cathodic potentials (750 mV versus SHE) uses hydrogen as an intermediate[8]. This prediction has been done to reduce the level of uncertainty and try to make a better estimate of what will happen in the future. In this review, the expectations of biogas creation will be done by applying the idea of information mining to handle information. The term "data mining" also refers to a series of procedures used to manually extract unknown information with explicit added value from a database[9]. Bioenergy has produced from biomass in the formed organic materials, including their respective wastes and residues. Biogas is naturally distracted by biogenic materials. Under these conditions, all forms of biomass are under anaerobic conditions[10]. The first model gives rise to a power plant using distributed farm-scale biogas power, which features digestion along with the animal so that more and more bioenergy crops are usually from a single farm. Large-scale crops are attached to large pig farms or dairy farms. The second model includes factories on a large scale, which digest animal manure and factory waste in the form of tapioca collected from several farms along with organic residues from industry and municipalities[11].

Organic residues make up the majority of pulp and paper mill biomass. Sawdust, various types of sludge, paper waste, and FA are all examples of waste, as black waste is the purpose of waste[12]. Cementation, bituminization, and vitrification are three of the various methods that can be used to compact radioactive waste. The oldest method of compacting radioactive waste is cementation, and its technical stability and ease of use have been demonstrated[13]. As part of the certification and continual evaluation process for the new facility for making electricity, the obtained results also serve as the foundation for the extensive use of the model in this simulation. In addition, it demonstrates the innovative compliance results of a dynamic PV power plant simulation model that describes in detail the steps required to certify a power plant in accordance with the new Spanish grid code and accurately represents a PV power plant[14].

The EM calculation isn't expressly a calculation but a recipe for a calculation. Either the M-step or the E-step is obvious in many examples. The majority of statisticians immediately turn to other algorithms, such as structured scoring algorithms, when faced with such a dilemma. As a result, statisticians are currently beginning to develop strategies for overcoming the lack of an explicit solution for either the E-step or the M-step problems [15]. The meaning of C4.5 algorithm is a decision tree-like classification model in the same way that a tree structure has internal nodes (not leaves) that describe the attributes of each branch. Each leaf describes the entire data class.[16]

2.0 THEORETICAL

2.1 Biogas

Biogas is a combination of the results of the development of alternative energy that has been formed through a series of decomposition processes of organic matter through an anaerobic fermentation process (without oxygen) to produce gas in the form of methane gas (CH4), the results of which has burned. Biogas was developed for household and industrial needs. Not only a very high anaerobic process cycle in biogas formation but also determined by factors that affect the growth of microorganisms, such as the amount of temperature or temperature, pH levels, salinity, more ions, nutrient richness, inhibition, and the amount of toxicity in the several processes, and solids concentration[17]. A crucial component of biogas is methane. Methane gas is useful for fuel mixtures because it has a fairly high calorific value, and has around 4,800 to 6,700 Kcal/m3[18].

2.2 Tapioca Liquid Waste

Tapioca liquid waste comes from sources of waste from a production process, such as livestock, waste, and industrial liquid waste. These wastes can be in the form of solid, liquid, or gaseous waste, if not handled properly, will have a low impact[19]. Tapioca industrial liquid waste has generated from the washing and extraction process, which contains many organic materials such as starch, protein fiber, and sugar[20]. Management that is much more digestible, lower transportation costs for solid fractions, and the prevention of uncontrollable decomposition processes are just a few of the many benefits of solid-liquid separation[21].

2.3 The Data Mining Process

The goal of data mining—also known as the "data mining process"—is to discover interesting patterns in a large amount of data in order to generate knowledge[22]. Some of the main kinds of data mining methods that could be developed recently, Data visualization, guided mining of metarules, paternal matching, as well as a generalization, characterization, classification, clustering, association, and evolution will all be discussed in this section. Global information systems, relational, transactional, object-oriented, spatial, and active databases, as well as methods for extracting knowledge from these databases, were also examined. Applications for possible data mining and a few research questions will soon be discussed[23]. The main things related to the meaning of data mining are as follows:

1. As an automated method for processing existing data, data mining can be beneficial
2. The quantity of data that needs to be processed can be very large or very small
3. It is possible for the amount of data that needs to be processed to be very small or very large[24].

2.4 Knowledge Discovery in Database

The entire non-trivial process of finding and identifying patterns in the data, as well as the ways in which the patterns are valid, novel, useful to researchers, and easy to understand, is referred to as Knowledge Discovery in Databases, or KKD. As a framed of a progression of cycles, information mining could be isolated into a few models, and furthermore the phases of the cycles as outlined in Figure 1. These stages are interactive, with direct user participation or access to a knowledge base[25]. The following process of KDD stages could be seen in Figure 1. The stages of the KDD process consist of:

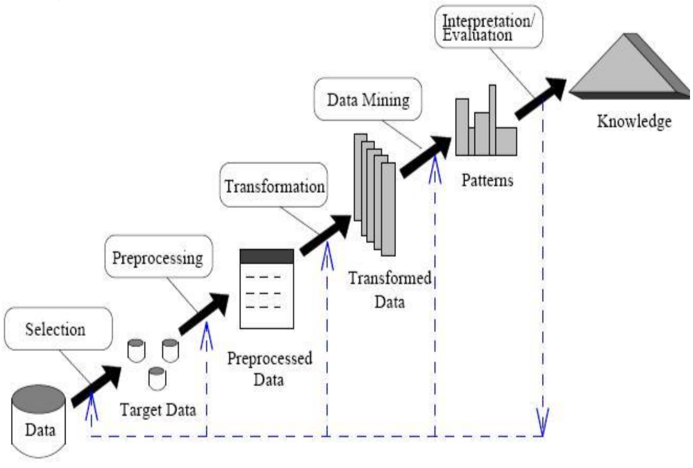


Figure 1 Knowledge Discovery in Database

1. The Data Selection Process

The process of selecting data from a set of data must be completed before the stages of information extraction begin. the data that comes out of the part would be used in the data mining processes, and it would be saved in a file that is different from the real database.

2. Cleaning and Pre-Processing Data

Data cleaning, also known as the cleaning process, involves correcting data errors like typographical errors, checking for inconsistent data, and removing duplicate data.

3. The Transformation Data

Transformation is the one the of processes to change the selected data into something new and simpler so that it can be used in the data mining process. Determines the kind or pattern of data that will be immediately searched for in the database case and the nature of this data design process.

4. The Data Mining Review

The converting the selected data into a result is called data meaning. This method involves calculating the data and displaying the information that would be searched immediately in the database case.

2.5 Interpretation / Evaluation

The final stage of the KDD process involves looking at patterns or information that contradicts facts or existing hypotheses.

2.6 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) method can facilitate maximizing every probability function that appears in some estimation problems in statistical models. In the classical EM paradigm, one can interactively maximize the conditional logarithmic probability of a complete data space, which is very difficult to observe. That is the same as maximizing the unsolvable probability function for whether the data has been measured or incomplete. The EM algorithm builds all parameters as a whole together, which has two drawbacks:

1) On progressive convergence
2) difficult maximization steps due to coupling when a fineness penalty is used[26].

2.7 The C4.5 Algorithm Classification

Algorithm Classification of the C4.5 algorithm is simple using a decision tree. This decision tree was constructed in practice by recursively dividing a set of data until each part contained data from the same data class. [27].

3.0 METHODOLOGY

3.1 Organization Review

Budi Starch and Sweetener/BSSW Tbk (BUDI), formerly known as Budi Acid Jaya Tbk, was founded on January 15, 1979, and it began trading in January 1981. Budi's headquarters are on floors 8 to 9 of Jalan HR Rasuna Said Kavling C6, in Jakarta. In the meantime, BUDI has factories in Surabaya, Lampung, Jambi, and Subang. Sungai Budi's business group includes Budi Starch & Sweetener (BSSW) Tbk. The following shareholders hold at least 5% of Budi Starch & Sweetener Tbk's shares: PT Budi Delta Swakarya and PT Sungai Budi (25.03%).

3.2 Data on Gas Production and Electricity Results

The following is data on gas production and electricity results as follows:

Table 1 2018 Gas Production Results and Electricity Results Data

Gas Production (Nm3)	Electricity Yield (MWh)	Y	X1	X2	DIST 1	DIST 2	CLASS
0.184	0.213	2	-0.095514689	-0.056743583	26.77080809	74.88103941	1
0.173	0.240	2	-0.098254462	-0.050909667	26.71785132	74.88977512	1
0.239	0.344	2	-0.083118993	-0.031102095	26.48513505	74.52439257	1
0.184	0.283	2	-0.095514689	-0.042306907	26.62501138	74.7929894	1
0.208	0.296	2	-0.089869725	-0.039836102	26.58776395	74.6865957	1
0.097	0.148	2	-0.121424381	-0.07292823	26.99219347	75.40042189	1
0.154	0.229	2	-0.103262795	-0.05323937	26.75241693	74.98513858	1
0.129	0.235	2	-0.110523509	-0.051961138	26.75557381	75.09505837	1
0.137	0.242	2	-0.108103328	-0.050492434	26.73536681	75.04684367	1
0.161	0.285	2	-0.101373159	-0.041923347	26.63401388	74.88554471	1
0.157	0.291	2	-0.102446178	-0.04078027	26.62485331	74.89597872	1
0.152	0.303	2	-0.103813067	-0.038526451	26.60515209	74.90442923	1

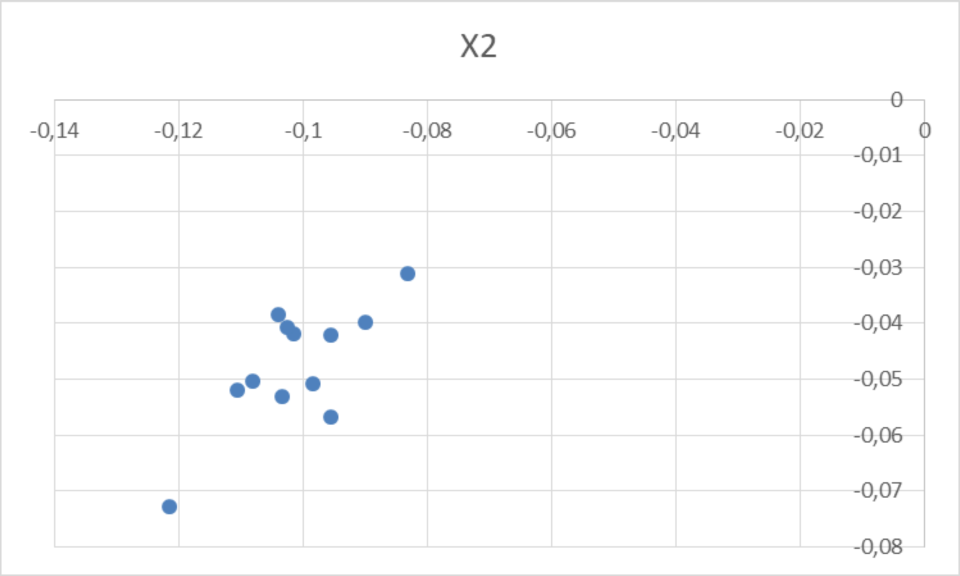


Figure 2 Graph Results of X1 and X2 in 2018

The result of the graph above show that in 2018 there was a significant increase starting at 0,03 and ending at 0,072. At the point above there was an increase and there was a slight decrease in production in 2018. The increase occurred due to several factors, such as the needed for gas and electricity is increasing, so that producers increase their production.

Table 2 Gas Production Results and Electricity Results Data 2019

Gas Productio n (Nm3)	Electricity Yield (MWh)	Y	X1	X2	DIST 1	DIST 2	CLAS S
0.179	0.282	2	-0.064504426	-0.042229906	26.55725209	74.29139444	1
0.161	0.250	2	-0.069344228	-0.048865303	26.63453793	74.40989591	1
0.173	0.267	2	-0.066081591	-0.045289988	26.59148103	74.33546334	1
0.208	0.337	2	-0.057309866	-0.031605194	26.43495498	74.11087234	1
0.219	0.348	2	-0.054739096	-0.029569348	26.40904238	74.05711234	1
0.148	0.240	2	-0.07306338	-0.051028574	26.66435467	74.48312769	1
0.214	0.352	2	-0.055898073	-0.028835	26.40410219	74.07133522	1
0.216	0.365	2	-0.055432621	-0.026468536	26.37932396	74.0495065	1
0.176	0.280	2	-0.065288753	-0.042633222	26.56298994	74.30649959	1
0.182	0.312	2	-0.063728328	-0.036332868	26.49616671	74.24303183	1
0.168	0.290	2	-0.067422715	-0.040630161	26.54734367	74.32874104	1
0.169	0.285	2	-2.973149530	-5.059885319	116.9872098	185.3717619	1

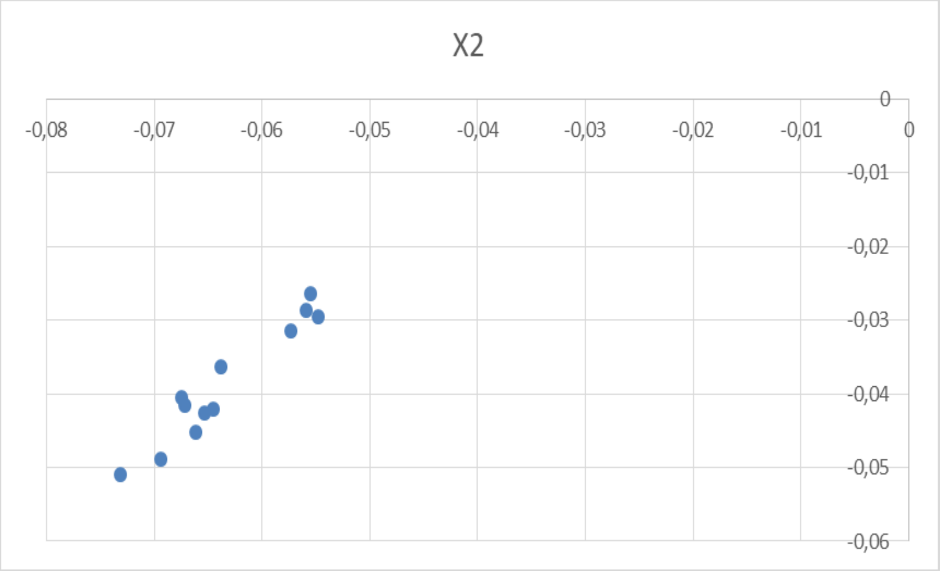


Figure 3 Graph Results of X1 and X2 in 2019

The result of the graph above show that in 2019 there was a significant increase starting at 0,25 and ending at 0,051. At the point above there was an increase and there was a slight decrease in production in 2019. The increase occurred due to several factors, such as the needed for gas and electricity is increasing, so that producers increase their production. These result are different from 2018.

Table 3 Gas Production Results and Electricity Results Data 2020

Gas Production (Nm3)	Electricity Yield (MWh)	Y	X1	X2	DIST 1	DIST 2	CLASS
0.182	0.299	2	-0.800837186	-0.510005336	33.60317338	89.77487264	1
0.196	0.302	2	-0.755276467	-0.5024181	33.35760043	88.92179857	1
0.197	0.323	2	-0.752099502	-0.450206978	32.77460876	88.50317388	1
0.206	0.335	2	-0.723933648	-0.421010247	32.35929932	87.81032941	1
0.189	0.323	2	-0.777796865	-0.450206978	32.86531779	88.953646	1
0.255	0.435	2	-0.58177717	-0.190019468	29.43832109	83.82312359	1
0.265	0.475	2	-0.554645293	-0.101181965	28.43897942	82.79928566	1
0.214	0.363	2	-0.699504472	-0.354397182	31.55788463	86.93335851	1
0.229	0.394	2	-0.655086856	-0.28263959	30.64559353	85.68625116	1
0.232	0.408	2	-0.64640306	-0.250769619	30.28122463	85.327789	1
0.221	0.383	2	-0.678561858	-0.30789777	30.99134865	86.25962358	1
0.223	0.394	2	-0.672648476	-0.28263959	30.70403396	85.99055426	1

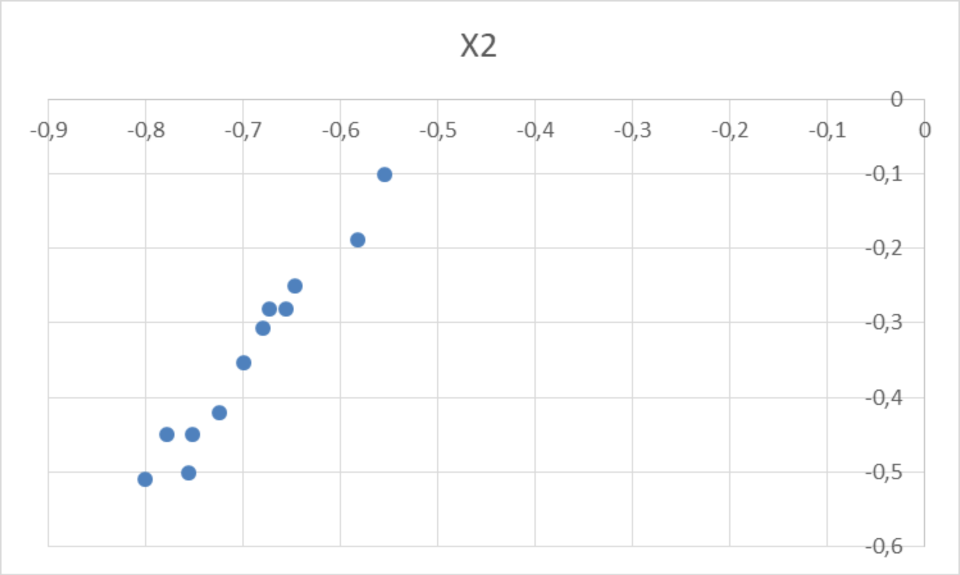


Figure 4 Graph Results of X1 and X2 Year 2020

The result of the graph above show that in 2020 there was a significant increase starting at 0,1 and ending at 0,51. Gas and electricity production in 2020 showed the best results compared to previous years. At the point above there was an increase and there was a slight decrease in production in 2020. The increase occurred due to several factors, such as the needed for gas and electricity is increasing, so that producers increase their production. These result are different from 2018 and 2019.

3.3 Mathematical Formulation

Expectation-Maximization Algorithm

Then from the table above that, it can be calculated using the Expectation-Maximization Algorithm first as follows:

1) The stage begins with taking testing data.
2) Then from the table above can be calculated using the Expectation-Maximization Algorithm and the C4.5 Algorithm, for how it works is as follows:
   1. Finding the probability value

$$p(x) = \sum_{i=1}^{k} \pi i f i(x) \tag{1}$$

2. Determine the log-likelihood value

$$\ln p\ (x;\ \pi,\mu,\Sigma)\ = \sum_{i=1}^{N} \ln \{ \sum_{k=1}^{K} \pi\ k\ N\ (xi\ |\mu k, \sigma k)\} \tag{2}$$

- Finding the value Y = IF (x > stdev, 1, 2)
- Find the values X1 and X2 = IF (Y=1, NORMINV (x, mean, stdev), NORMINV (x, mean, stdev))
- Find the value of DIST = (X1 * MEANx)2 + (X2 * MEANx)2
- Specifies CLASS = IF(DIST 1 < DIST 2, 1, 2)

3. Reestimate parameters using Gamma
4. Evaluate log-likelihood

$$\ln p\ (x;\ \pi,\mu,\Sigma)\ = \sum_{i=1}^{N} \ln \{ \sum_{k=1}^{K} \pi\ k\ N\ (xi\ |\mu k, \sigma k)\} \tag{3}$$

5. Check for convergence, if any converge back to point 3 and if not exit the loop.
6. Done

C4.5 Algorithm

The following are the steps for calculating the Entropy and Gain values for each criterion that has High and Low descriptions.

1. Entropy Calculation

The first step of the C4.5 algorithm is to find the entropy value. First, determine the total Entropy value in the case. With the following formula:

$$\text{Entropy}\ (S) = \sum_{i=1}^{n} -\text{pi} * \log_2 \text{pi} \tag{4}$$

Information:
S-: Case set
A-: Attribute
n-: Number of partitions S
pi-: Proportion of Si to S Gain Calculation

The next step after calculating the entropy value is to calculate the gain value to determine the root of the decision tree with the following formula:

$$\text{Gain}\ (S,A) = \text{entropy}(S) - \textstyle\sum_{i}^{n} = 1\ \frac{Si}{S} * \text{entropy (Si)} \tag{5}$$

Information:
S-: case set
A-: attribute
n-: number of partition attributes A
|Si|-: number of cases on partition i
|S|-: number of cases in S
The following is the process of the C4.5 Algorithm:

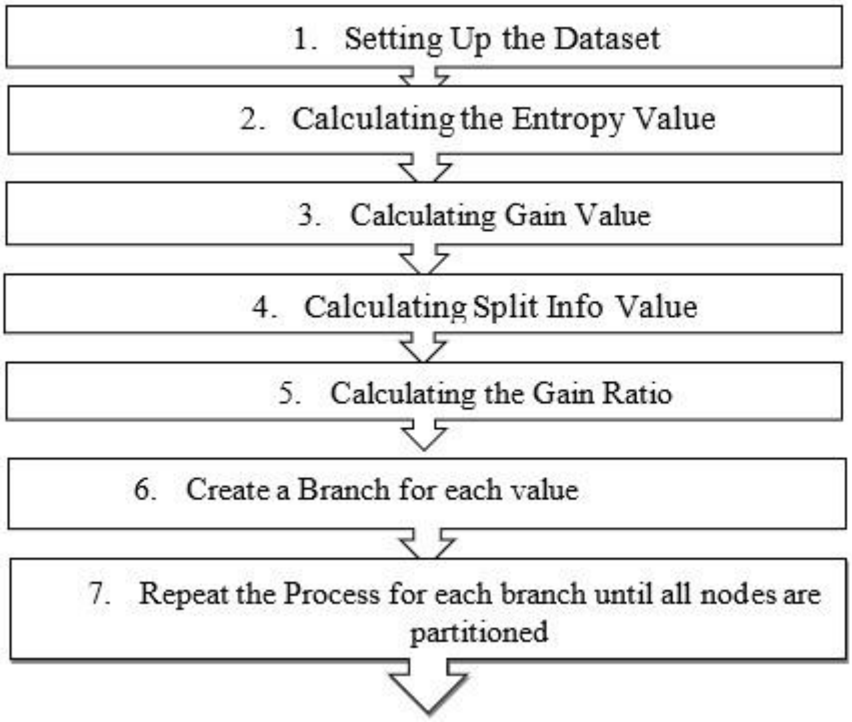

1. Setting Up the Dataset
2. Calculating the Entropy Value
3. Calculating Gain Value
4. Calculating Split Info Value
5. Calculating the Gain Ratio
6. Create a Branch for each value
7. Repeat the Process for each branch until all nodes are partitioned

Figure 5 Process Algorithm C4.5

Counting the number of cases

= (-(high number/number of months)*log2 (high number/number of months) + (-(low number/number of months)*log2 (low number/number of months))

= (-(6/12)*$\log_2$ (6/12) + (-(6/12)*$\log_2$ (6/12))

= ((-0.5)*(-0.30103) + (-0.5)*(-0.30103))

= 0.150515 + 0.150515

= 0.30103

Next, calculate the entropy for each criterion based on the number of cases per subset of criteria.

a. Calculation of criteria Gas Production (Nm3)(A1)

Entropy ">=260" (3)

= (-(0/0)*$\log_2$ (0/0) + (-(0/0)*$\log_2$ (0/0))

= ((0)*(0) + (0)*(0))

= 0 + 0= 0

Entropy ">=209" (2)

= (-(1/1)*$\log_2$ (1/1) + (-(0/0)*$\log_2$ (0/0))

= ((-1)*(0) + (0)*(0))

= 0 + 0 = 0

Entropy ">=97" (1)

= (-(5/11)*$\log_2$ (5/11) + (-(6/11)*$\log_2$ (6/11))

= (-0.44554)*(-0.34242)) + ((-0.545454)*(-0.26324)

= 0.155634 + 0.069295 = 0.29923291

b. Calculation of criteria Electricity Yield (MWh)(A2)

Entropy ">=475" (3)

= (-(0/0)*$\log_2$ (0/0) + (-(0/0)*$\log_2$ (0/0))

= ((0)*(0) + (0)*(0))

= 0 + 0= 0

Entropy ">=366.233" (2)

= (-(0/0)*$\log_2$ (0/0) + (-(0/0)*$\log_2$ (0/0))

= ((0)*(0) + (0)*(0))

= 0 + 0 = 0

Entropy ">=147.9" (1)

$= (-(6/12)*\log_2 (6/12) + (-(6/12)*\log_2 (6/12))$
$= (-0.5)*(-0.30103)) + ((-0.5)*(-0.30103)$
$= 0.150515 + 0.150515 = 0.30103$

Gain value calculation

After the entropy calculation for each subset of criteria is complete, then calculate the gain value:

a. Gain Gas Production (Nm3) (A1)
$= 0.301029996 – (0/0)* (0) – (1/12)*(0) – (11/12)*(0.29923291)$
$= 0.026733161$

b. Gain Electricity Yield (MWh) (A2)
$= 0.301029996 – (0/0)* (0) – (0/0)*(0) – (6/12)*(0.30103) – (6/12)*(0.30103)$
$= 0$

Table 4 Calculation results in Microsoft Excel

Knot	Amount	Collectibility		Entropy	Gain Information
		Tall	Low		
Number	12	6	6	0.30103	
A1Gas Production					0.0267332
>=260	0	0	0	0	
>=209	1	1	0	0	
>=97	11	5	6	0.2992329	
A2Electrical Yield					0
>=475	0	0	0	0	
>=366.233	0	0	0	0	
>=147.9	12	6	6	0.30103	

Determining the Decision Tree (decision tree)

From the results of the gain ang entropy values above, a decision tree is obtained to provide answers to the calculations above. The results is starting from A1, namely high gas production results of more than equal to 209 Nm3 and then continued with A2, namely electricity results that have the highest calculated value more than equal to 147,9 MWh and the lowest is less than 147,9 MWh. This figure is the decision tree of the gain and production entropy calculations.

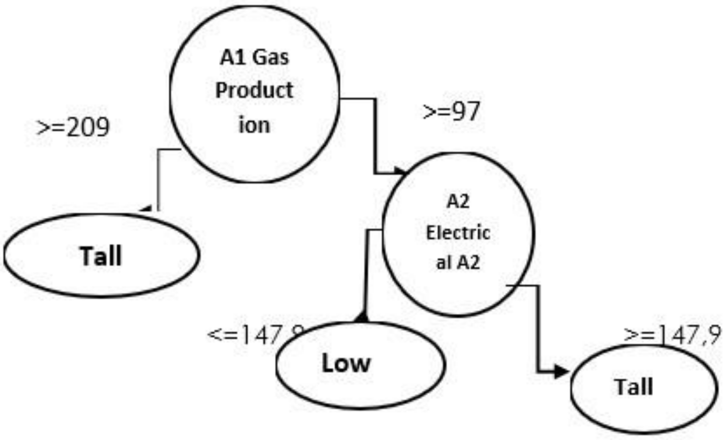


Figure 6 Decision Tree node 1

3.4 Accuracy

In determining the percentage of accuracy of the processed data, the formula used is as follows:
the percentage of accuracy

$$= \frac{The\ data\ on\ the\ number\ of\ correctly\ predicted\ outcomes}{The\ number\ of\ predictions\ made} \times 100\%$$

Calculation of Data Testing with a total of 12 Data of Gas Production Results and Electricity Results 2018 and presentation of accuracy as follows:

$$Percentage\ of\ Accuracy = \frac{12}{12} \times 100\% \ = \ 100\%$$

Table 5 Confusion Tabel

Class		
Prediction	Tall	Low
Tall	6	0
Low	0	6

Based on the above calculations, it has been concluded that the accuracy of the testing data, which amounted to 12 data points, accuracy results in 100%. based on the decision tree on the testing data above, the criteria that most influence electrical results can be PT. Budi Starch & Sweetener (BSSW), Terusan Nunyai Subdistrict, Central Lampung Regency, Lampung Province shows that the Information Gain on Criterion A1 (Gas Production (Nm3)) is 0.0267332 greater than the other criteria.

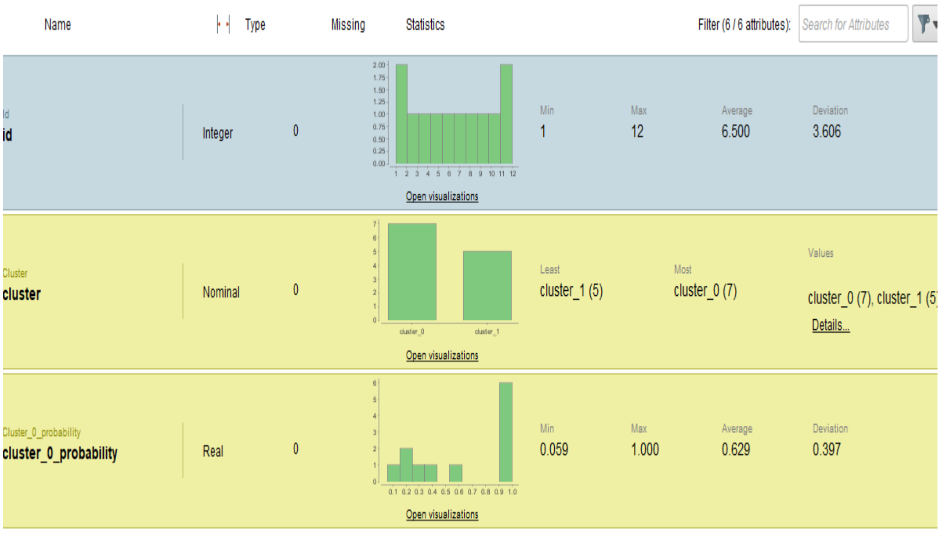
4.0 RESULTANTS

4.1 Expectation-Maximization Algorithm

The following is the calculation of the Expectation-Maximization Algorithm :

4.1.1 Statistics 1 in Data processing

After being calculated using the rapidminer application the above results are obtained. How to read it from the right, namely the Min value is 1, the Max value is 12. Then the average shows a value of 6000 (total gas and electricity) then the standard deviation is 3606. Then below it there is a cluster number 1 which is ranked 5th and explained from cluster 0-7. Then the real Min value is 0,059, Max is 1000, Average is 0,629, and deviation is 0,397.



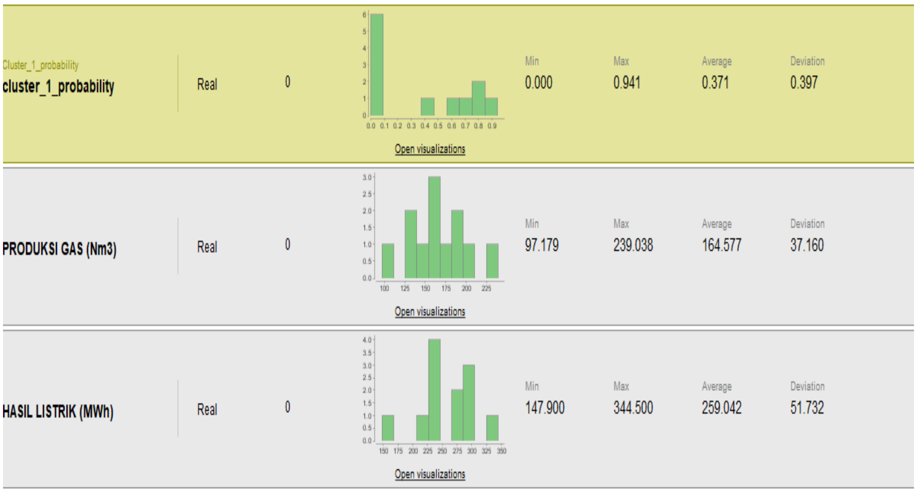
Filter (6 / 6 attributes): Search for Attributes

Name	Type	Missing	Statistics				
Id id	Integer	0	Open visualizations	Min 1	Max 12	Average 6.500	Deviation 3.606
Cluster cluster	Nominal	0	Open visualizations	Least cluster_1 (5)	Most cluster_0 (7)	Values cluster_0 (7), cluster_1 (5 Details...	
Cluster_0_probability cluster_0_probability	Real	0	Open visualizations	Min 0.059	Max 1.000	Average 0.629	Deviation 0.397

Figure 7 Stats 1

4.1.2 Statistics 2 in Data processing

After being calculated using the rapidminer application the above results are obtained. How to read it from the right, namely cluster1 probability the Min value is 0, the Max value is 0,941. Then the average shows a value of 0,371 (total gas and electricity) then the standard deviation is 0,397. Then below it there is a calculation of gas production of gas production (Nm3) which has a value is 97,179, Max is 239,038, Average is 164,577, and deviation is 37,160. Then for the calculation of the electricity yoeld (MWh) the Min value is 147,900, the Max value is 344,500, the Average value is 259,042. And the deviation value is 51,732. So we can see that the electricity is higher than the gas output.



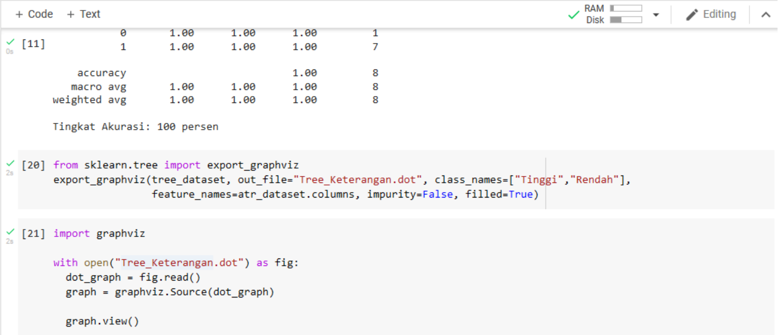
Cluster_1_probability cluster_1_probability	Real	0	Open visualizations	Min 0.000	Max 0.941	Average 0.371	Deviation 0.397
PRODUKSI GAS (Nm3)	Real	0	Open visualizations	Min 97.179	Max 239.038	Average 164.577	Deviation 37.160
HASIL LISTRIK (MWh)	Real	0	Open visualizations	Min 147.900	Max 344.500	Average 259.042	Deviation 51.732

Figure 8 Stats 2

4.2 C4.5 Algorithm

The following is the calculation of the C4.5 Algorithm:

4.2.1 Sample Accuracy



```
              0       1.00      1.00      1.00         1
[11]          1       1.00      1.00      1.00         7

       accuracy                           1.00         8
      macro avg       1.00      1.00      1.00         8
   weighted avg       1.00      1.00      1.00         8

   Tingkat Akurasi: 100 persen
```

```
[20] from sklearn.tree import export_graphviz
     export_graphviz(tree_dataset, out_file="Tree_Keterangan.dot", class_names=["Tinggi","Rendah"],
                     feature_names=atr_dataset.columns, impurity=False, filled=True)
```

```
[21] import graphviz

     with open("Tree_Keterangan.dot") as fig:
       dot_graph = fig.read()
       graph = graphviz.Source(dot_graph)

       graph.view()
```

Figure 9 100% accuracy

The image above shows the calculation of the C4.5 algorithm. The accuracy obtained is 100%. Starting from opening the application, writing the coding on the google colabs and changing it in the data transformation process, then eliminating some data that is not needed, then selecting the data class so that the results of electricity and gas can be known.

4.2.2 Determine the decision tree

There is a slight difference between calculations using microsoft excel and google. collabs. research namely in making decision trees. Here we can see that there are 28 sample data with low class there are 7 data and high class are 21 data. It is show that electricity dominates production at PT Budi Starch and Sweetener (BSSW). The decision tree shows that each branch has an explicit value to give the decicion that is expected to give the best result. The results above show that the electricity yield is higher than the other criteria.

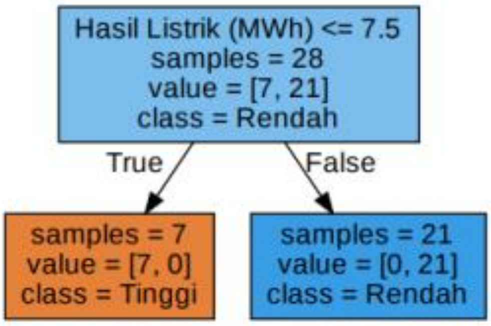


Figure 10 Decision Tree

4.3 Comparison Between Algorithms

Table 6 Comparison Results

Year	Expectation-Maximization Algorithm	C4.5 Algorithm
2018	Based on the group MEAN 1 X1 value is 0.00184, MEAN 1 X2 is 0.00323, MEAN 2 X1 is -0.002, MEAN 2 X2 is -0.003, STDEV is 0.68, ALPHA is 0.5.	The accuracy of the Testing data which amounts to 12 data has an accuracy of 100% and information gain Criteria A1 (Gas Production (Nm3)) is 0.0267332
2019	Based on the group MEAN 1 X1 has a value of 0.00215, MEAN 1 X2 has a value of 0.00455, MEAN 2 X1 has a value of -0.002, MEAN 2 X2 has a value of -0.046, STDEV a value of 0.88, ALPHA 0.5.	The accuracy of the Testing data which amounts to 12 data has an accuracy of 100% and information gain Criteria A1 (Gas Production (Nm3)) is 0.0231399
2020	Based on the group MEAN 1 X1 has a value of 0.0036, MEAN 1 X2 has a value of 0.0051, MEAN 2 X1 has a value of -0.002, MEAN 2 X2 has a value of -0.046, STDEV has a value of 0.065, ALPHA 0.5.	The accuracy of the Testing data, which amounts to 12 data has an accuracy of 100%, and information Gain on Criterion A1 (Gas Production (Nm3)) and A2 (Electricity) Yield (MWh) is 0.

5.0 CONCLUSION

The implementation of the Expectation-Maximization algorithm is to be carried out using Microsoft Excel and Rapid Miner 9.9 applications. The result is a more detailed and easy-to-understand explanation of the entire formula. Meanwhile, the information system buildings based on the theory and application of the C4.5 algorithm can all be ready to use. It should use Google Colab research and Microsoft Excel. Google Colab Research is a software that is on Google too different from Rapid Miner. So, every calculation is done online with more accurate results. With the results showing that in 2018, 2019, and 2020 there has been an increase in gas production and electricity production, the company expected to be able to optimize gas and electricity production in the coming year. In the meantime, Google's Colab Research and Microsoft Excel were utilized for the C4.5 algorithm implementation. the translation of the entire formula and simpler prediction results are the outcomes of this. The use of the EM Algorithm and the C4.5 Algorithm to determine the results of PT's biogas production Producers can predict whether the company should increase or decrease biogas production3 because BSSW has a significant influence. The use of the EM Algorithm and the C4.5 Algorithm to determine the results of PT's biogas production Producers can predict whether the company should increase or decrease biogas production because BSSW has a significant impact.

REFERENCES

[1] XY Chen, H. Vinh-thang, AA Ramirez, D. Rodrigue, and S. Kaliaguine, *RSC Advances Membrane gas separation technologies for biogas upgrading*, pp. 24399–24448, 2015, doi:10.1039/c5ra00666j.

[2] AH Igoni, *Designs of anaerobic digesters for producing biogas from municipal*, doi:10.1016/j.apenergy.2007.07.013.

[3] W. Management, *The anaerobic digestion of solid organic waste*, DOI: 10.1016/j.wasman.2011.03.021.

[4] AD Singh, B. Gajera, and AK Sarma, *Appraising the availability of biomass residues in India and their bioenergy potential,* Waste Manag., **152**(July), pp. 38–47, 2022, doi:10.1016/j.wasman.2022.08.001.

[5] CH Geissler and CT Maravelias, *Environmental Science capture strategies: present and future*, pp. 2679–2689, 2022, doi:10.1039/d2ee00625a.

[6] C Dannesboe Reaction, *Chemistry & Engineering experimental results from a reactor at full-scale* †, pp. 183–189, 2020, doi:10.1039/c9re00351g.

[7] M. Perssonet al., *Biogas Upgrading to Vehicle Fuel Standards and Grid Injection*, submitted for publication.

[8] P. Batlle-Vilanova, *RSC Advances Deciphering the Electron Transfer Mechanisms for Biogas Upgrading to Biomethane Within a Mixed Culture Biocathode*, pp. 52243–52251, 2015, doi:10.1039/c5ra09039c.

[9] T. Syahputra, J. Halim, And K. War-Angin, *Application of Data Mining in Predicting the*

Passing Level of Midwives Competency Test (Ukom) at Medan Senior Stikes Using Multiple Linear Regression Methods, Science and Computers (Saintikom), **17**(1), Pp. 1–7, 2018.

[10] D. Nofriansyah and I. Mariami, *Implementation of Data Mining for Grouping Books at the New Indonesia Nurul Islam Foundation Library Using the K-Means Clustering Method*, **1**(1), pp. 1–12, 2021.

[11] WM Budzianowski And I. Chasiak, *The Expansion Of Biogas Fueled Power Plants In Germany During The 2001 – 2010 Decade: Main Sustainable Conclusions For Poland*, Journal Of Power Technologies, **91**(2), Pp. 102–113, 2011.

[12] S. Arita, D. Kristianti, and L. Nurul, *South African Journal of Chemical Engineering Effectiveness of biomass-based fly ash in pulp and paper liquid waste treatment,* South African J. Chem. Eng., **41**(February), pp. 79–84, 2022, doi:10.1016/j.sajce.2022.05.004.

[13] J. Shon, H. Lee, T. Kim, G. Kim, and H. Jeon, *Evaluation of cementation of intermediate level liquid waste produced from fission 99 Mo production process and disposal feasibility of cement waste form*, Nucl. eng. Technol., **54**(9), pp. 3235–3241, 2022, doi:10.1016/j.net.2022.03.033.

[14] M. Martínez-lavín, R. Villena-Ruiz, and A. Honrubia-Escribano, *Evaluation of the latest Spanish grid code requirements from a PV power plant perspective,* Energy Reports, **8**, pp. 8589–8604, 2022, doi:10.1016/j.egyr.2022.06.078.

[15] A. Arbor And FR May, *A Gradient Algorithm Locally Equivalent To The Em Algorithm, Journal Of The Royal Statistical Society: Series B(Methodological),* **2**, Pp. 425–437, 1995, Doi: 10.1111/J.2517-6161.1995.Tb02037.X.

[16] S. Haryati, A. Sudarsono, And E. Suryana, *Implementation of Data Mining to Predict Student Study Period Using the C4.5 Algorithm (Case Study: Universitas Dehasen Bengkulu), J. Media Infotama*, **11**(2), Pp. 130–138, 2015.

[17] Dhaniswara Kindergarten, *Effect of Pre-treatment of Organic Waste on Anaerobic Biogas Production*, Journal Of Research And Technology, **3**(2), 2017.

[18] Y. Kurniati, A. Rahmat, BI Malianto, D. Nandayani, W. Sri, And W. Pratiwi, *Review of Optimum Condition Analysis in the Biogas Production Process*, Engineering 14 **14**(2), Pp. 272–281, 2021.

[19] AC Adityawarman, *Simple Cattle Waste Treatment in Pattalassang Village, Sinjai Regency, South Sulawesi, Journal of Animal Production Science and Technology*, **03**(3), Pp. 171–177, 2015.

[20] BK Government, P. Region, P. Java, B. Department, E. Resources, And F. Economics, *Internalization of Tapioca Small Medium Industrial Liquid Waste (Ikm) through Biogas WTP for Power Generation, Journal of Agricultural and Environmental Policy Minutes Formulation of Strategic Studies in Agriculture and the Environment,* **4**(1), Pp. 73–88, 2017.

[21] Carol Sze KL, Chong L., Christian VS, Guneet K., Xiaofeng Y., *Waste Valorisation*, ed. **1**, Willey, 2020.

[22] S. Hendrian, *Data Mining Classification Algorithm To Predict,* **11**(3), pp. 266–274, 2018.

[23] IT Julian too, D. Kurniadi, MR Nashrulloh, And A. Mulyani, *Comparison Of Data Mining Algorithm For Forecasting Bitcoin Crypto Currency Trends Comparison Of Data Mining Algorithm For Forecasting Trends,* Journal of Informatics Engineering, **3**(2), Pp. 245–248, 2022.

[24] M. Syahril, K. Erwansyah, And M. Yeti, *Application of Data Mining to Determine School Equipment Sales Patterns on the Wiggle Brand Using the Apriori Algorithm,* Journal of Information Systems and Computer Systems Technology, **3**(1), Pp. 118–136, 2020.

[25] I. Fitri, P. Ginting, And D. Saripurna, *Application of Data Mining in Determining Patterns of Stock Availability of Goods Based on Consumer Demand at Chykes Minimarket Using the Apriori Algorithm, Scientific Journal (Journal of Informatics and Computer Management Science)*, **20**(1), Pp. 28–37, 2021.

[26] JA Fessler And A. Hero, *Space-Alternating Generalized Expectation-Maximization Algorithm*, IEEE Xplore, **42**(10), 1994.

[27] N. Yahya et al., *For Prediction of New Student Admission Activities (Case Study: Stikubank University Semarang), Inspiration: Journal of Information and Communication Technology*, 2014, pp. 978–979, 2019.