e-ISSN: 2614-8404
p-ISSN : 2776-3234

Implementation of the K-Means Clustering Algorithm in Determining Productive Oil Palm Blocks at Pt Arta Prigel

Yesi Pitaloka Anggriani[1*)], Alfis Arif[2], Febriansyah[3]
[1,2,3]Program Studi Teknik Informatika, Institut Teknologi Pagar Alam
Email: [1] yesipitalokaanggriani@gmail.com, [2]abangarif@gmail.com, [3]Febriansyahh1213@gmail.com

Abstract – The purpose of this study is to implement the K-Means Clustering method to determine the patterns of productive oil palm production based on their blocks at Pt Arta Prigel. The research is motivated by issues within the oil palm blocks, such as the absence of productive block summaries, insufficient plantation land analysis, and erroneous decision-making. The development method utilizes CRISP-DM, with data spanning 2 years from October 2021 to October 2023. From the 1275 production records, after cleaning, 1015 records remain. Filtering the initial 51 blocks results in 37 blocks for the years 2021 and 2022, and 46 blocks for the year 2023. After clustering, the production outcomes for the year 2021 are as follows: cluster_0 has 34 blocks, cluster_1 has 10 blocks. For the year 2022, cluster_0 has 24 blocks, cluster_1 has 37 blocks. In the year 2023, cluster_0 has 44 blocks, cluster_1 has 27 blocks. The testing method employs the silhouette coefficient, and the silhouette score testing results indicate the formation of 2 clusters (K=2) with a value of 0.62, the results obtained from testing with 2 clusters indicate that the formed clusters are accurate. The findings of this study include patterns, graphs, and production tables generated using the K-Means Clustering method at Pt Arta Prigel.

Keywords – K-Means Clustering, Rapid Miner, palm, Silhouette Coefficient

I. INTRODUCTION

Technology, a single word that plays a pivotal role in the current development of human life. The proliferation of advancements and sophistication in technology at present brings about highly significant changes [1]. Big data is a collection of datasets in very large amounts [2] [3] Data mining, or data extraction, involves the analysis of data from various perspectives to transform it into valuable information that can be utilized to enhance profitability [4]. Data processed through data mining techniques will generate new scientific knowledge from old data, the results obtained from this data processing can be used to make decisions in the future [5]. The k-means algorithm is a method in unsupervised learning, it is utilized to cluster data into several groups [6]. K-Means is one of the widely used clustering algorithms for partitioning data into clusters [7]. Clustering is a machine learning method used to group data into appropriate clusters or clusters [8]. Referring to categorization such as notes, observations, or attention and forming classes of objects that have similarities. A cluster is a set of records that are similar and different from records in other clusters [6].

The application of the k-means algorithm is highly effective in addressing issues in clustering and managing big data to generate new information. Utilizing the k-means algorithm to cluster COVID-19 cases provides valuable insights for decision-making regarding the COVID-19 pandemic in Indonesia. [9]. The k-means algorithm can be employed as a tool for customer segmentation and marketing strategy development in retail stores [10], and the k-means algorithm can assist farmers in optimizing rubber production and improving resource utilization efficiency [5].

PT Arta Prigel is a palm oil plantation company that has been in commercial operation since 1983. The entire palm oil plantation operating area obtained the Right to Cultivate (HGU) in 2006 for a period of 35 years. PT Arta Prigel is located at the complete address of BBIP Palm Group, Padang Lengkuas Village, Lahat Sub-district, Lahat District, South Sumatra, Postal Code 31461.

Prior to this research, there had been previous research conducted by Pulungan et al in (2019) entitled "Implementation of K-Means Clustering Algorithm in Determining the Most Productive Palm Oil Plantation Blocks", the K-Means algorithm can assist companies in clustering productive and non-productive palm oil plantation blocks into 2 clusters: high cluster for the most productive blocks and low cluster for non-productive blocks. It was found that there are 14 highly productive palm oil plantation blocks and 26 non-productive ones.

Then next, the research conducted by Phitaloka in 2022, titled "Application of K-Means Clustering on the Effectiveness of Nutmeg Plantation", utilized the K-means algorithm for clustering the effectiveness of nutmeg plantation production in the Tanggamus region. This study employed 419 training data records with 4 features to create clusters using Weka 3.6.13. The k-means clustering method was employed to determine 2 clusters. This research demonstrates that this algorithm can assist in determining the effectiveness of nutmeg plantations.

Rapid Miner utilizes object-oriented methods within the Java hierarchy and can be employed across nearly all platforms .[11] The Silhouette Coefficient is an evaluation or validation method for clustering algorithms that measures the quality of formed clusters. It validates clusters by combining separation and cohesion methods and serves as a testing method for cluster quality used to determine the ownership level of each object within a cluster [12] [13] [14].

II. RESEARCH METHODOLOGY

This research employs data mining technique namely k-means clustering, and the CRISP-DM (Cross Industry

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Standard Process for Data Mining) development method, which is a standard data mining processing developed to ensure that existing data undergo structured and clearly defined stages efficiently. In addition to applying models in the data mining process, the selection of algorithms significantly influences the comparison of the performance of these data mining methods [15].

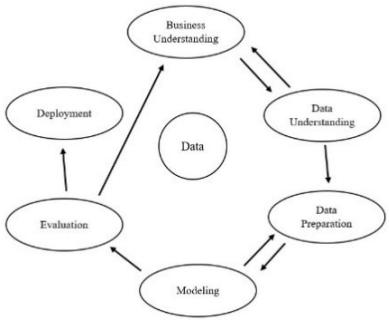

Fig 1 Metode Crisp-Dm

A. Business Understanding

In this stage, several tasks need to be carried out, namely understanding the needs and objectives from a business perspective, translating knowledge into the form of defining problems in data mining, and then determining plans and strategies to achieve the goals of data mining.

B. Data Understanding

In this stage, the first step to be elucidated is data collection, followed by data description, and evaluating data quality.

Table 1 Dataset used

No	Blocks	Ha	Results	Month	Year
1	1	21.09	5.772	10	2021
2	1	21.09	5.502	11	2021
3	1	21.09	9.120	12	2021
4	2	37.40	5.694	10	2021
5	2	37.40	7.921	11	2021
6	2	37.40	16.413	12	2021
7	13	39.97	8.899	10	2021
8	13	39.97	8.898	11	2021
9	13	39.97	11.131	12	2021
10	14	46.07	7.752	10	2021
……	….	…..	…..	……	……
1015	51	20.12	17.322	10	2023

C. Data Preparation

In this stage, constructing the dataset from raw data is undertaken. Several tasks need to be performed, including data cleaning, data selection, records, attributes, and data transformation, which will serve as inputs in the modeling stage.

D. Modeling

This stage will involve direct implementation of Machine Learning in determining techniques, tools, and data mining algorithms. The modeling approach adopted in this research is clustering method using the k-means algorithm.

1) Prepare the dataset
2) Determine the Number of Clusters
3) Cluster Center Points
4) Calculate Data Distances
5) Cluster Centers Update
6) Repeat steps 3 to 5

E. Evaluation

In this stage, the performance level of patterns generated by the algorithm is examined. The focus is on ensuring that the resulting model adheres to the standard k-means clustering and completes each stage without omission. The testing phase is conducted using the silhouette coefficient method.

F. Deployment

In this final stage, the creation of reports and journal articles is carried out using the model generated from the preceding stages. From the established patterns, it is possible to identify which blocks are productive, moderately productive, and unproductive.

III. RESULTS AND DISCUSSION

The results obtained from this research using the CRISP-DM development method can be seen as follows.

1. Business Understanding

In this stage, the process of determining objectives and search environments is carried out. The purpose of this search is to group production results based on blocks. The production data management process is still conducted manually or in a simple manner, where Excel is still used, hence there is a lack of detailed explanation regarding productive blocks.

2. Data Understanding

In the second stage, the data understanding process takes place. The data is obtained from the KrEstate division of Pt Arta Prigel, consisting of production results over 2 years from October 2021 to October 2023, totaling 1275 records and 5 attributes: Block, Hectares, Production, and Total. The data categories are received in Excel format, and upon obtaining the data, cleaning and selection processes of the data and its attributes are necessary.

3. Data Preparation

In the third stage, the data management process is conducted by directly implementing it on RapidMiner, as specified below.

3.1 Data Selection

The selection process for gathering information about the data must be carried out prior to commencing data mining. Attribute selection is performed, and some data within attributes are converted to facilitate the data mining process. The data selection process in this stage involves 5 attributes obtained from KrEstate Pt Arta Prigel, including Division, Block, Land Area, Production, and Total. The selection process is conducted in Excel using filters. After the data filtering process, the researcher utilizes Block, Hectares, Yield, Month, and Year. Initially, there were 51 blocks, which were then reduced to 37 blocks through

e-ISSN: 2614-8404
p-ISSN : 2776-3234

selection. The production records, initially 1275, are cleaned, resulting in 1015 data points that constitute the dataset.

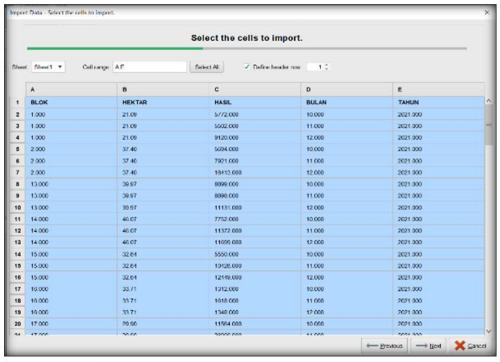

Fig 2 data selection process 2021

3.2 Data Processing

The processing process encompasses data cleaning and transformation. In this stage, the researcher has ensured that there are no more empty data. Figure 2 illustrates that if the attributes used have no empty data, with a count of 0 in the dataset for each month, they can be used for the next stage.

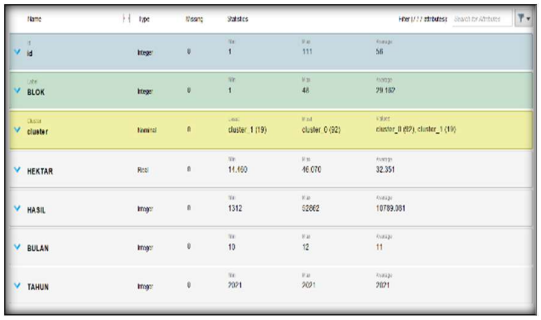

Fig 3 data processing process

3.3 Data Transformation

In the final stage, the process involves transforming the selected data so that it is suitable for the data mining process. The data transformation process conducted indicates the selected attributes processed in RapidMiner. From 1275 initial data points, they are transformed into 1015 data points, comprising 111 data points for the year 2021, 444 data points for the year 2022, and 460 data points for the year 2023. From 51 blocks, they are transformed into 37 blocks for the year 2021, 37 blocks for the year 2022, and 46 blocks for the year 2023. The production output data is transformed based on the predetermined criteria as follows.

Table 2 value criteria

Kriteria	Nilai
Tidak Produktif	C0
Produktif	C1

4. Modeling

The fourth process is where the model is utilized with the algorithm used in the clustering method, namely the k-means clustering model operator, which takes objects from the input port and sends copies to the output.

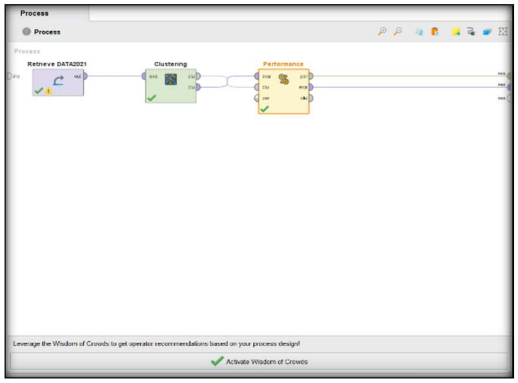

Fig 4 k-means algorithm model process

Next, in the modeling stage, the process of the k-means algorithm's steps is conducted, starting with the first step.

a) Determine the Number of Clusters
After the data has been collected, the determination of the number of clusters begins. Initially, experiments are conducted to ascertain the number of clusters that appear at the smallest cluster centroid.

Attribute	Cluster_0	Cluster_1
Hektar	31.513	36.413
Hasil	7296.750	27699.316
Bulan	10.957	11.211
Tahun	2021	2021

Figure 5 cluster experiment 2021

Attribute	Cluster_0	Cluster_1
Hektar	37.211	31.167
Hasil	31883.540	10832.036
Bulan	8.253	6.073
Tahun	2022	2022

Figure 6 cluster experiment 2022

Attribute	Cluster_0	Cluster_1
Hektar	30.266	32.070
Hasil	14701.136	35936.248
Bulan	5.150	6.479
Tahun	2023	2023

Fig 7 cluster experiment 2023

b) Random Centroid Points
The centroid points are randomly determined based on several clustering experiments in RapidMiner. Among these three trials, the smallest centroid point is chosen to obtain the best result.

c) Data Distance to Centroid

e-ISSN: 2614-8404
p-ISSN : 2776-3234

After importing the dataset and the k-means algorithm into Rapid Miner, the data process is initiated by clicking "run" to display the data distance to the data center. Among these three trials, the smallest centroid point is chosen to obtain the best result. The data distance values for cluster_0 range from 0 to 10.00 for unproductive clusters, cluster_1 ranges from 10.00 to 32.00 for productive clusters.

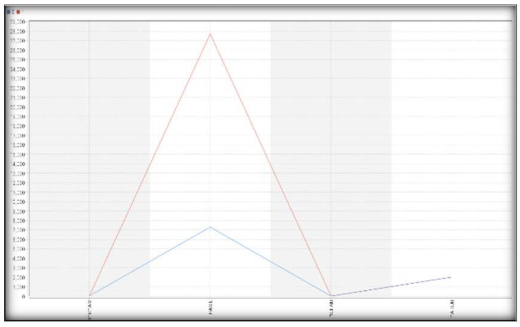

Fig 8 data distance to centroid 2021

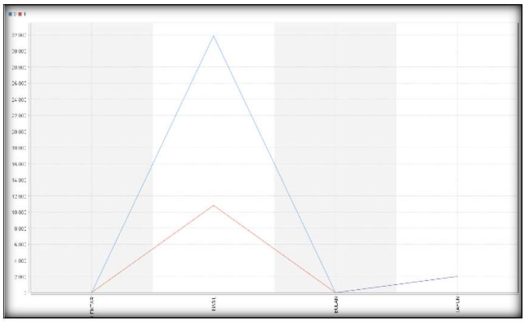

Fig 9 data distance to centroid 2022

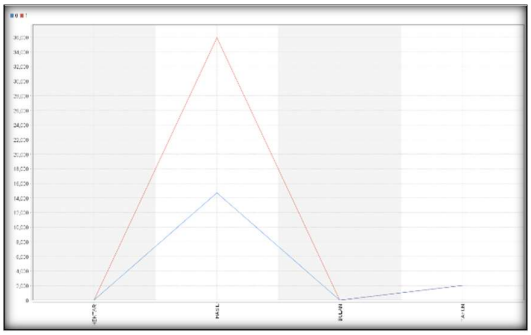

Fig 10 data distance to centroid 2023

d) Update Centroid Values
After several trials, the centroids of the 3 clusters remain unchanged. If the centroids shift, the iteration is repeated, but if there is no change, the repetition is stopped, and the results for each group are obtained. Based on the RapidMiner calculations, patterns are obtained, which will later be implemented in Python. The pattern used to group cluster data based on distance calculations is as follows
a. If C0 < C1 and C0 < C2, then cluster_0 is labeled as unproductive.
b. If C1 < C0 and C1 < C2, then cluster_1 is labeled as productive.
From the production data obtained through RapidMiner, using the k-means clustering method, the pattern of production results per block is derived as follows: cluster_0 indicates an unproductive level, cluster_1 indicates a productive level.

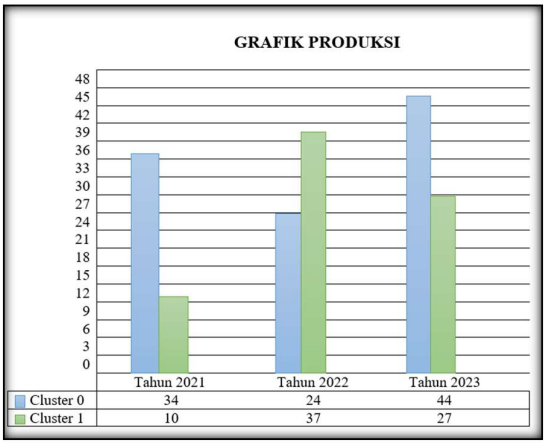

Fig 11 annual graphic results

5. Evaluation
Testing was conducted using the Silhouette Coefficient on the formed clusters with all 1015 records. The results obtained from the silhouette score after several attempts revealed that the suitable number of clusters is 2 (K=2), with a quality value of 0.62.

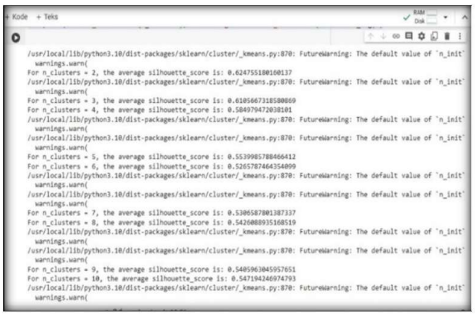

Fig 12 Test Results With Silhouette Coefficient

6. Deployment
In the deployment phase, which is the final stage, the obtained knowledge or information regarding the production patterns of Pt Arta Prigel over 2 years is presented, revealing the presence of 2 clusters. The resulting pattern for the year 2021 shows cluster_0 with 34 blocks, followed by cluster_1 with 10 blocks. In the year 2022, cluster_0 comprises 24 blocks, cluster_1 has 37 blocks. Lastly, in the year 2023, cluster_0 encompasses 44 blocks, cluster_1 contains 27 blocks. From the clustering process, it is observed that the level of productive production shows an increase in the years 2023, while for the production results at Pt Arta Prigel, the unproductive production level is more dominant.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

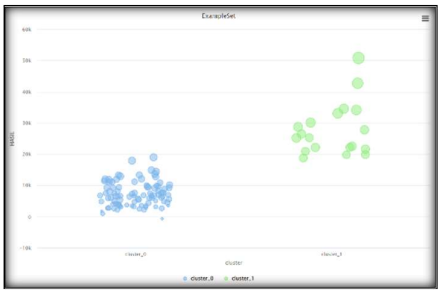

Fig 13 partitional clustering pattern results in 2021

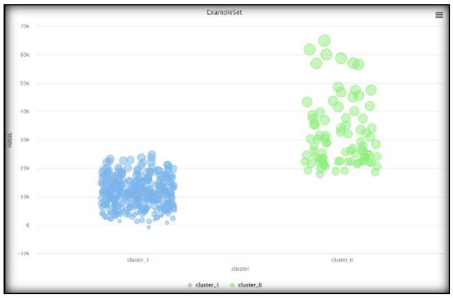

Fig 14 partitional clustering pattern results in 2022

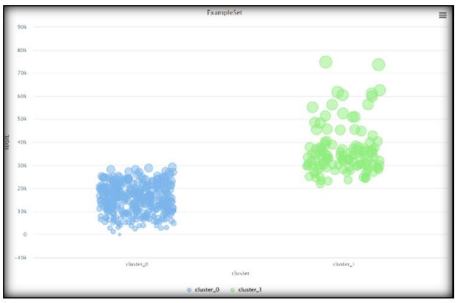

Fig 15 partitional clustering pattern results in 2023

IV. CONCLUSION

Based on the research conducted, the following results were obtained: This study yielded 2 Clustering Patterns of production results at Pt Arta Prigel, namely cluster_0 labeled as Unproductive, cluster_1 labeled as Productive. From the clustering results of Pt Arta Prigel's palm oil production, the production patterns for 2 years with a total of 51 blocks can be observed. It is known that the production for the year 2021 includes 34 blocks for cluster_0, 10 blocks for cluster_1. Furthermore, the production for the year 2022 includes 24 blocks for cluster_0, 37 blocks for cluster_1. For the year 2023, the production includes 44 blocks for cluster_0, 27 blocks for cluster_1. Then, the results of testing using the Silhouette Coefficient on the Google Colab application with the Python programming language to calculate the silhouette score obtained the suitable number of clusters as K=2 with a silhouette score value of 0.62, the results obtained from testing with 2 clusters indicate that the formed clusters are accurate. The value obtained in the silhouette score testing indicates that the quality of the clusters is appropriate. From these results, several insights are expected to be beneficial for Pt Arta Prigel, particularly for the Manager and Assistant Manager of the plantation, to support decision-making and further actions.

REFERENCES

[1] P. Amita Tri Prasasti and C. Dewi, "Pengembangan Assesment of Inovation Learning Berbasis Revolusi Industri 4.0. untuk Guru Sekolah Dasar," *J. Ilm. Sekol. Dasar*, vol. 4, no. 1, p. 66, 2020, doi: 10.23887/jisd.v4i1.24280.

[2] R. T. Aldisa, P. Maulana, and M. A. Abdullah, "Penerapan Big Data Analytic Terhadap Strategi Pemasaran Job Portal di Indonesia dengan Karakteristik Big Data 5V," vol. 3, pp. 267–272, 2022, doi: 10.30865/json.v3i3.3905.

[3] U. Dirgantara and M. 2022 Suryadarma, "Revolusi Industri 4.0: Big Data, Implementasi Pada Berbagai Sektor Industri (Bagian 2)," no. Bagian 2.

[4] R. I. O. Limabri, F. Putrawansyah, and A. Arif, "Penerapan Data Mining Untuk Mengklasifikasi Nasabah Bank Sumsel Babel Menggunakan Algoritma C4. 5," *Escaf*, pp. 1101–1108, 2023.

[5] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara," pp. 762–767, 2019.

[6] & Z. 2021 Zulfa, Auliya, Permata, "Analisis Data Mining Untuk Clustering Kasus COVID-19 di Provinsi Lampung Dengan Algoritma K-Means," vol. 2, no. 2, pp. 100–108, 2021.

[7] F. Febriansyah and S. Muntari, "Penerapan Algoritma K-Means untuk Klasterisasi Penduduk Miskin pada Kota Pagar Alam," vol. 8, no. 1, pp. 66–77, 2023.

[8] M. R. Nahjan, N. Heryana, A. Voutama, F. I. Komputer, U. S. Karawang, and R. Miner, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," vol. 7, no. 1, pp. 101–104, 2023.

[9] N. Dwitri, J. A. Tampubolon, S. Prayoga, and P. P. P. A. N. W. F. I. R. H. Zer, "Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Indonesia," vol. 4, no. 1, pp. 128–132, 2020.

[10] R. Supardi and I. Kanedi, "Implementasi Metode Algoritma K-Means Clustering Pada Toko Eidelweis," vol. 4, no. 2, pp. 270–277, 2020.

[11] S. Rokhanah, A. Hermawan, and D. Avianto, "Pengaruh Principal Component Analysis Pada Naïve Bayes dan K-Nearest Neighbor Untuk Prediksi Dini Diabetes Melitus Menggunakan Rapidminer," vol. 11, no. 1, 2023.

[12] M. Astiningrum, M. Mentari, and Y. G. Maranatha, "Mutu Buah Salak Menggunakan Pengolahan Citra Digital," pp. 205–210.

[13] N. Afdhaliah, "Perbandingan kinerja algoritma ward dan algoritma k-means dengan uji silhouette coefficient," 2020.

[14] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Bencana Kebakaran Menggunakan Model CRISP-DM," vol. 12, no. 1, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[15] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," vol. 5, no. 2, 2021.