Accredited Rank SINTA 2
Decree of the Director General of Higher Education, Research and Technology, Number: 158/E/KPT/2021
validity period from Volume 5 Number 2 Year 2021 to Volume 10 Number 1 Year 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 6 (2022) 905 - 910 ISSN Media Elektronik: 2580-0760

K-Means Clustering Algorithm Approach in Clustering Data on Cocoa Production Results in the Sumatra Region

Mawaddah Harahap[1], Arief Wahyu Dwi Ramadhanu Zamili [2], Muhammad Arie Arvansyah[3], Erwin Fransiscus Saragih[4], Selwa Rajen[5], Amir Mahmud Husein[6]
[1,2,3,4,5]Department of Computer Science, Faculty of Technology and Computer Science, University of Prima Indonesia
mawaddah@unprimdn.ac.id

Abstract

Cocoa agricultural production in Indonesia is currently very low while demand continues to increase every year, so it is very important to build a model that can categorize cocoa farming data. The main objective of this research is to analyze agricultural data using data mining techniques that specifically use the K-Means Clustering algorithm, and Gaussian Mixture Models. In this research, we used quantitative research because it measure number-based data. The results of cocoa production so far still depend on land area, then the number of cocoa trees has a significant effect on the amount of production so it is very important for the government and researchers to develop technologies that can increase cocoa production yields where the demand for cocoa is currently very high in demand worldwide because it can classify the cocoa quality from good quality to poor quality. Based on testing the K-Means Clustering and Gaussian Mixture Model algorithms on data on cocoa production in four provinces, namely North Sumatra, West Sumatra, Lampung and Aceh which were optimized by the Silhouette method, it produced cluster values of 2, 3 and 4. second with a value of 59.8%.

Keywords: *K- Means Clustering Algorithm, Gaussian Mixture Model, Data Mapping, Cocoa Farm*

1. Introduction

Indonesia is one of the Cocoa agricultural production countries, but the growth of Cocoa agricultural production has decreased by an average of 2.01% in the last ten years from the top five cocoa producing countries. [1]. Meanwhile, demand for the product is growing at 3% per year and the top five cocoa producing countries produce more than 95% of the world's cocoa demand. The low growth of cocoa production and productivity of cocoa farms in Indonesia is based on several reasons such as limited knowledge of farmers in the cultivation and management of cocoa plants, marketing problems and others. [1]. Therefore, it is very important to understand the trend of cocoa production especially clustering data of farmers and their production so as to produce accurate information that can be useful for decision making.

In recent years, data mining methods are one of the methods that are widely used to extract useful information from a set of data, one of which is cluster algorithms such as K-Means Clustering, *dbscan clustering*, *hierarchical clustering*, fuzzy c means and others. Cluster analysis is one of the algorithms used to group databases so that data in a cluster is similar, and as different as possible from data in other clusters. Clustering allows for in-depth interpretation with many implications about which data should be targeted with specific data that is most likely to be of interest. [2]. Partition clustering algorithms, such as K means assign objects into k (a predetermined number of clusters) clusters, and reallocate objects iteratively to improve the quality of the clustering results. [3].

Hierarchical clustering algorithms assign objects in tree-structured clusters, i.e., a cluster can have representatives of data points from lower-level clusters. The idea of Density-based clustering method is that for each point of a cluster, the neighborhood of a given distance unit contains at least a minimum number of points, i.e., the density in the neighborhood must reach some threshold. The idea of a density-based clustering algorithm is that, for any point of a cluster, the neighborhood of a given distance unit must contain at least a minimum number of points. The application of cluster analysis is widely applied in various fields such as customer data clustering [2], [4]-[6], crop productivity mapping [7], agricultural data [3], palm oil production results [8], fruit yield grouping [9], grouping

Accepted: 29-06-2022 | Received: 23-10-2022 | Published: 27-12-2022

port [10] and others. Data mining clustering techniques such as K-Means is one of the algorithms that are widely applied by many researchers including [11] using clustering algorithms for data grouping, data mapping, data classification, and so on. [12]-[15].

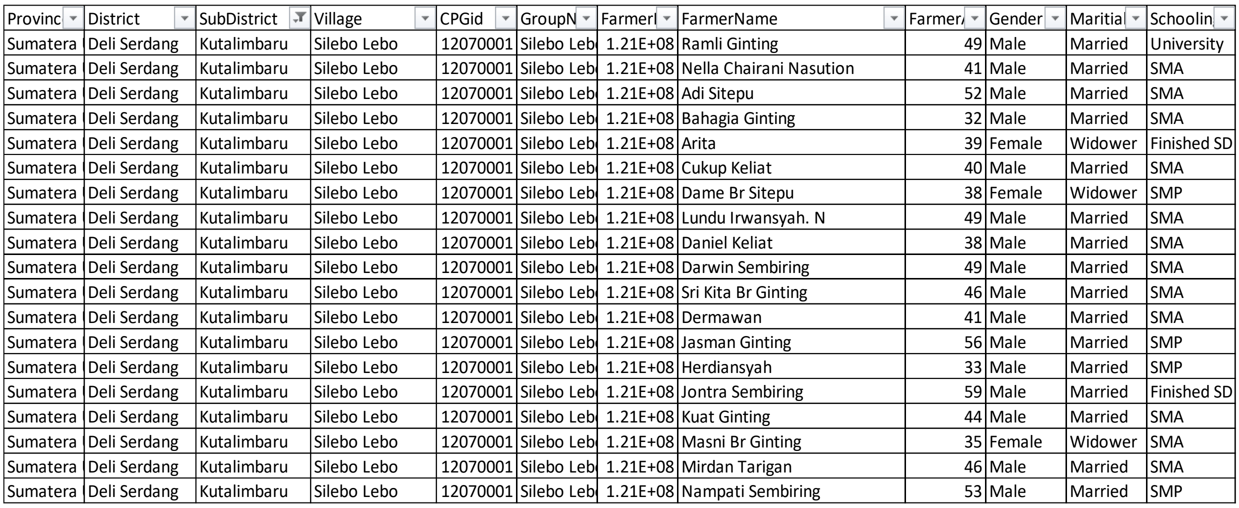
In this paper, we aim to apply cluster algorithms in the field of agriculture. [16]−[18] specifically applied to mapping cocoa farming data in four provinces in Indonesia, namely North Sumatra, West Sumatra and Aceh. The source dataset is a set of survey data that contains information about farmers, land and production. K-Means Clustering and Gaussian Mixture Models clustering algorithm approaches are used to cluster cocoa production data. [19]-[21].

2. Research Methods

This research uses quantitative research because it uses number-based measurements. Quantitative research is an investigation of social problems based on testing a theory consisting of variables, measured by numbers, and analyzed by statistical procedures to determine whether the predictive generalization of the theory is true. The dataset is obtained from surveys and interviews of cocoa farmers in 4 (four) regions of Sumatra, namely Aceh Province, North Sumatra Province, West Sumatra Province and Lampung Province.

The data used is data from the Swisscontact program, namely the database of Sustainable Cocoa Production Program (SCPP) farmers in the Sumatra Region Batch I 2017 with details of data for North Sumatra province totaling 1,492, West Sumatra Province totaling 4,594, Aceh Province totaling 4480, and Lampung Province totaling 2,007 with a total of 12,573. Table 1. is a partial dataset of the data used.

Table 1. Part of the dataset



Provinc	District	SubDistrict	Village	CPGid	GroupN	FarmerI	FarmerName	Farmer/	Gender	Maritial	Schoolin
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Ramli Ginting	49	Male	Married	University
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Nella Chairani Nasution	41	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Adi Sitepu	52	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Bahagia Ginting	32	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Arita	39	Female	Widower	Finished SD
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Cukup Keliat	40	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Dame Br Sitepu	38	Female	Widower	SMP
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Lundu Irwansyah. N	49	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Daniel Keliat	38	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Darwin Sembiring	49	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Sri Kita Br Ginting	46	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Dermawan	41	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Jasman Ginting	56	Male	Married	SMP
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Herdiansyah	33	Male	Married	SMP
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Jontra Sembiring	59	Male	Married	Finished SD
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Kuat Ginting	44	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Masni Br Ginting	35	Female	Widower	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Mirdan Tarigan	46	Male	Married	SMA
Sumatera	Deli Serdang	Kutalimbaru	Silebo Lebo	12070001	Silebo Leb	1.21E+08	Nampati Sembiring	53	Male	Married	SMP

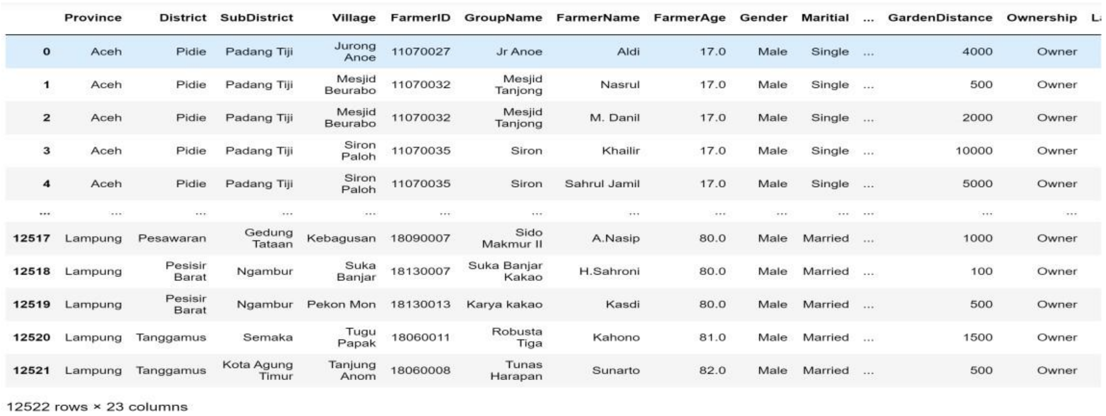
Research is conducted to obtain information that has a relationship with processing on the dataset. The stages in this study begin with collecting datasets/training data from SCPP data in the Sumatra region Batch I 2017, followed by a *pre-processing* stage which includes data cleaning which includes missing values, smooth noise data, identifying and removing outliers, resolving inconsistencies. Data integration process from several databases and data transformation in the form of normalization and aggression.

Next, the *Principal Component Analysis* stage is carried out, the point is to make the dataset simpler by the linear transformation method so that a new coordinate system with maximum variation is formed. Followed by selecting a subset of data that is relevant to the problem from the existing set of features, without transforming and combining all features to improve prediction capabilities followed by a pre-process that gets raw features so that the right amount of data is not always large.

The data that has been obtained is then segmented in order to separate and analyze the subset of data based on these data segments. the last step is to evaluate, which is to display the information patterns generated from the data maining process in a form that is easily understood by interested parties.

3. Result and Discussion

In this section we describe the research results of analyzing and clustering Cocoa production data in Indonesia specifically in the provinces of North Sumatra, West Sumatra, Lampung and Aceh. We will observe the statistical description of the data set, consider the relevance of each feature, and select a few sample data points from the data set that we will track throughout this project.



	Province	District	SubDistrict	Village	FarmerID	GroupName	FarmerName	FarmerAge	Gender	Maritial	...	GardenDistance	Ownership	L
0	Aceh	Pidie	Padang Tiji	Jurong Anoe	11070027	Jr Anoe	Aldi	17.0	Male	Single	...	4000	Owner	
1	Aceh	Pidie	Padang Tiji	Mesjid Beurabo	11070032	Mesjid Tanjong	Nasrul	17.0	Male	Single	...	500	Owner	
2	Aceh	Pidie	Padang Tiji	Mesjid Beurabo	11070032	Mesjid Tanjong	M. Danil	17.0	Male	Single	...	2000	Owner	
3	Aceh	Pidie	Padang Tiji	Siron Paloh	11070035	Siron	Khailir	17.0	Male	Single	...	10000	Owner	
4	Aceh	Pidie	Padang Tiji	Siron Paloh	11070035	Siron	Sahrul Jamil	17.0	Male	Single	...	5000	Owner	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
12517	Lampung	Pesawaran	Gedung Tataan	Kebagusan	18090007	Sido Makmur II	A.Nasip	80.0	Male	Married	...	1000	Owner	
12518	Lampung	Pesisir Barat	Ngambur	Suka Banjar	18130007	Suka Banjar Kakao	H.Sahroni	80.0	Male	Married	...	100	Owner	
12519	Lampung	Pesisir Barat	Ngambur	Pekon Mon	18130013	Karya kakao	Kasdi	80.0	Male	Married	...	500	Owner	
12520	Lampung	Tanggamus	Semaka	Tugu Papak	18060011	Robusta Tiga	Kahono	81.0	Male	Married	...	1500	Owner	
12521	Lampung	Tanggamus	Kota Agung Timur	Tanjung Anom	18060008	Tunas Harapan	Sunarto	82.0	Male	Married	...	500	Owner	

12522 rows × 23 columns

Figure 1. Data Set

Figure 1 is a collection of cocoa production data from four provinces in Indonesia, namely North Sumatra, West Sumatra, Lampung and Aceh. This data set of cocoa production results is the result of survey data in 2017. There are 12,522 rows and 23 columns of the total. Furthermore, an analysis is carried out to get missing data or null values, this is very important so that no errors occur during the clustering process. In the data set there is 0.1% missing data in the Productivity column and 3.5% in the ShadeTreesNr column, the data will be deleted. After cleaning the missing data, the total data for further analysis is 12,053.

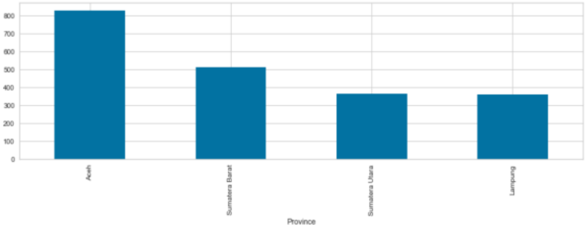


Figure 2. Production Results

In Figure 2. it can be seen the results of Cocoa production in four provinces where Aceh province produces the largest Cocoa production every year, then Sumatra Goods, North Sumatra and lastly Lampung. Cocoa production is highly influenced by the area of plantations in each province as shown in Figure 3. Aceh province has the most extensive plantations among the three provinces so it is natural to produce the highest cocoa production.

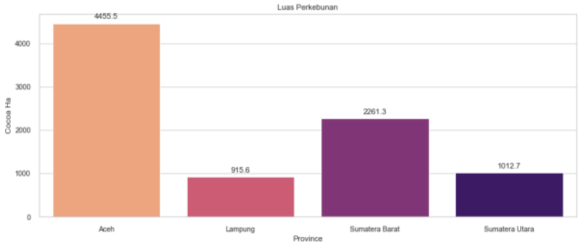


Figure 3. Cocoa Plantation Area

In Figure 3, it can be seen that the plantation area still dominates cocoa production in each province, this is something that naturally occurs but in the current era of technological development, it can utilize technology to increase production with limited plantation area. With this condition, it can be concluded that cocoa farmers in Indonesia currently still use traditional techniques. The results of the data analysis presented are still many that have not been described, due to data limitations only available in 2017 so that analysis of the development of cocoa production every year cannot be done.

Next, the results of clustering cocoa production data in the four provinces will be described where the K-Means Clustering and Gaussian Mixture Model (GMM) algorithms are applied to the dataset. Both algorithms will be evaluated using the Silhouette method in determining the optimal cluster points, then applied to both algorithms. Before applying the clustering algorithm, the first stage is carried out various models to determine the optimal features. From the results of several experiments from the data set, there are 9 features that have an impact on data analysis, namely 'FarmerID', 'CacaoAge', 'GardenDistance', 'Cocoa Ha', 'Production', 'Productivity', 'Trees', 'Trees Ha' and 'Tree_Productivity'. From the nine features, the correlation of each variable will be analyzed in the form of a heatmap. A correlation heatmap is a graphical representation of a correlation matrix that represents the correlation between different variables with values of -1 to 1, the closer to 1 the better the correlation between the two variables.

In Figure 4. is the result of the correlation heatmap between nine variables, from the results of the figure it can be seen that the variables Cocoa Ha, Production, Productivity, Trees and Tree_Productivity produce a correlation value closest to number 1 of each variable so that in this study the 5 variables will be used as feature models in the K-Means Clustering and GMM algortima.

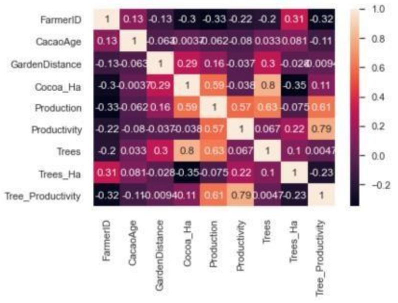
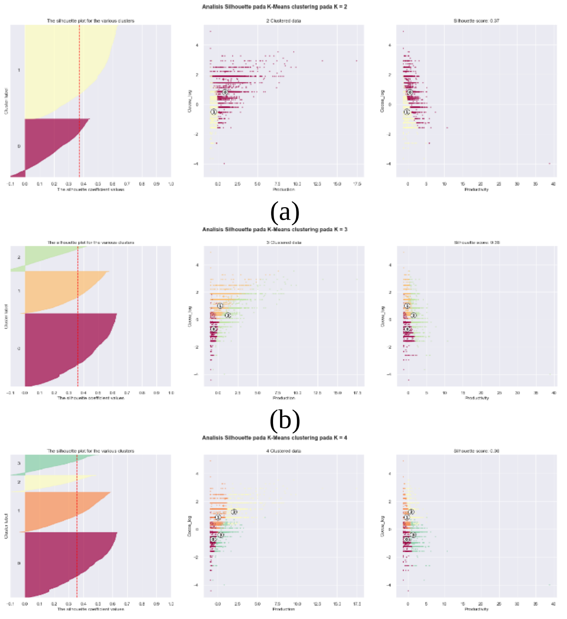


Figure 4. Correlation Heatmap

In the K-Means Clustering algorithm, the first step before applying cocoa production data to the K-Means Clustering algorithm is to analyze the determination of the K value using the Silhouette method.



(a)

(b)

(c)

Figure 5. Silhouette Method Results (a) K=2, (b) K=3 dan (c) K=3

In Figure 5. it can be seen the results of the Silhouette method analysis of the K value, where for the value of K = 2 produces the highest score with a value of 0.37, then a score of 0.36 for each K = 3 and K = 4. Although the value of K 2 is the highest value, the difference is not too much different from the value of K = 3 as well as the value of K = 4, so in this study all three values will be applied. The results of the K-Means Clustering algorithm for clustering cocoa production data in the four provinces are shown in Figure 6.

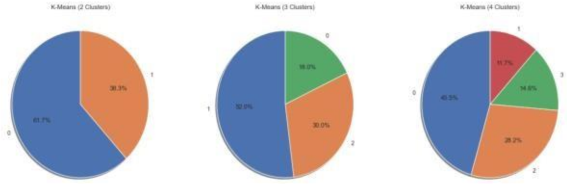
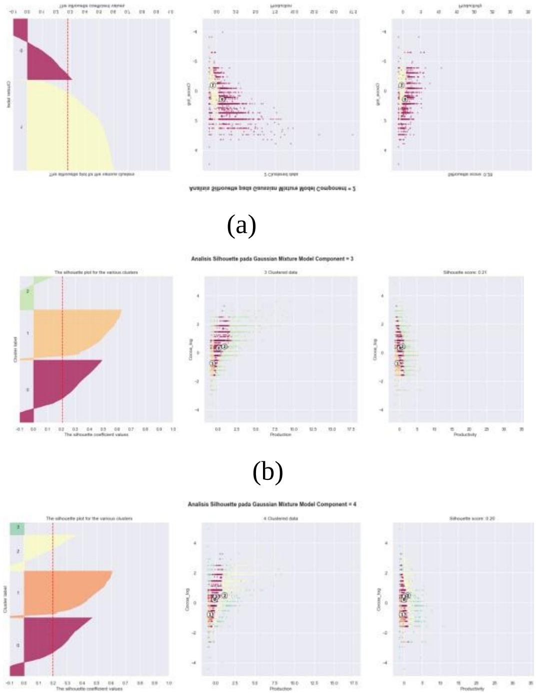


Figure 6. K-Means Clustering Result

Figure 6 illustrates the results of clustering using the K-Means Clustering algorithm at values of K=2, K=3 andK=4. From the figure, it can be seen that for cluster0 results in 61.7% and cluster1 38.3% at K=2 where this value indicates that the level of plantation area is still highly dependent on production, then the rest is the number of cocoa farmers. Furthermore, the value of K=3 results in 18% cluster0, 52% cluster1 and 30% cluster2 where these results indicate that the number oftrees depends on cocoa yield. Finally, the value of K=4resulted in 45.5%, 11.7%, 14.6% and 28.2% in each cluster. The final results of the clustering data are presented in the appendix.

The application of the Gaussian Mixture Model (GMM) on cocoa production data of the four provinces, as described in the previous section that the GMM model will be evaluated using the Silhouette method in determining the optimal component value. The results of the Silhouette method analysis can be seen in Figure 7.



(a)

(b)

(c)

Figure 7. Results of Silhouette Analysis on GMM (a) component =2, (b) component =3 and (c) component =3

Figure 7 illustrates the results of the Silhouette method analysis on the GMM algorithm to determine the optimal component value. From this result it can be seen that the value of component =2 produces the highest score of 0.28, then components 3 and 4. This result is not too different from the K-Means Clustering algorithm where the clustering results on the data produce 2 and 4 groups. Next, the GMM clustering

DOI: https://doi.org/10.29207/resti.v6i6.4199

results will be attached to the Cocoa production data, as shown in Figure 8.

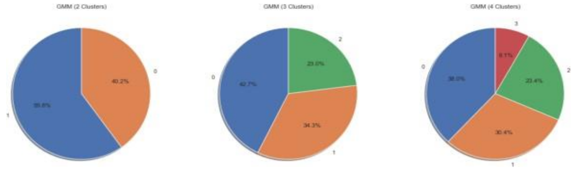


Figure 8. GMM Clustering Results

In Figure 8. is the result of clustering using the GMM algorithm on cocoa production, it can be seen that the results have differences with the K-Means Clustering algorithm? For GMM algorithm with component 2 value results in 59.8% for cluster0 and 40.2% in cluster1, then for component 3 results in cluster 0 value 42.7%, cluster1 34.3% and cluster2 23%. Finally, the GMM results with component 3 produced 38%, 30.4%, 23.4% and 6.1% in each cluster.

Based on the results of testing the K-Means Clustering algorithm and the Gaussian Mixture Model on cocoa production data in four provinces namely North Sumatra, West Sumatra, Lampung and Aceh optimized by the Silhouette method resulted in cluster values 2, 3 and 4 where from these results it can be concluded that cocoa production so far still depends on land area, then the number of cocoa trees has a significant effect on the amount of production so it is very important for the government and researchers to develop technology that can increase cocoa production where cocoa needs are currently very high demand worldwide.

4. Conclusions

Based on the results of testing the K-Means Clustering algorithm and Gaussian Mixture Model on cocoa production data in four provinces, namely North Sumatra, West Sumatra, Lampung and Aceh, a conclusion can be drawn, namely based on the results of testing the K-Means Clustering algorithm and Gaussian Mixture Model on cocoa production data in four provinces, namely North Sumatra, West Sumatra, Lampung and Aceh which are optimized by the Silhouette method to produce cluster values 2, 3 and 4. For the Gaussian Mixture Model (GMM) algorithm with component 2 values resulted in 59.8% for cluster 0 and 40.2% in cluster1, then for component 3 resulted in cluster 0 values of 42.7%, cluster1 34.3% and cluster 2 23%. Finally, the GMM results with component 3 resulted in 38%, 30.4%, 23.4% and 6.1% in each cluster.

Cocoa production so far is still dependent on land area, then the number of cocoa trees has a significant effect on the amount of production so it is very important for the government and researchers to develop technology that can increase cocoa production where cocoa demand is currently very high worldwide.

Reference

[1] I. M. Fahmid, H. Harun, M. M. Fahmid, Saadah, and N. Busthanul, "Competitiveness, production, and productivity of cocoa in Indonesia," IOP Conf. Ser. Earth Environ. Sci., vol. 157, no. 1, 2018, doi: 10.1088/1755-1315/157/1/012067.

[2] S. G. Carbajal, "Customer segmentation through path reconstruction," Sensors, vol. 21, no. 6, pp. 1–17, Mar. 2021, doi: 10.3390/s21062007.

[3] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," J. Big Data, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0077-4.

[4] A. Camero, G. Luque, Y. Bravo, and E. Alba, "Customer segmentation based on the electricity demand signature: The andalusian case," Energies, vol. 11, no. 7, p. 1788, Jul. 2018, doi: 10.3390/en11071788.

[5] A. Savic, G. Bjelobaba, S. Janicijevic, and H. Stefanovic, "An Application of PCA Based K-Means Clustering for Customer Segmentation in One Luxury Goods Company," UBT Int. Conf., 2019: https://knowledgecenter.ubt-uni.net/conference/2019/events/189.

[6] M. P. Fernandes, J. L. Viegas, S. M. Vieira, and J. M. C. Sousa, "Segmentation of residential gas consumers using clustering analysis," Energies, vol. 10, no. 12, p. 2047, Dec. 2017, doi: 10.3390/en10122047.

[7] M. R. Ridlo, S. Defiyanti, and A. Primajaya, "Implementasi Algoritme K-Means Untuk Pemetaan Produktivitas Panen Padi Di Kabupaten Karawang," Citee 2017, pp. 426–433, 2017.

[8] D. F. Pasaribu, I. S. Damanik, E. Irawan, Suhada, and H. S. Tambunan, "Memanfaatkan Algoritma K-Means Dalam Memetakan Potensi Hasil Produksi Kelapa Sawit PTPN IV Marihat," BIOS J. Teknol. Inf. dan Rekayasa Komput., vol. 2, no. 1, pp. 11–20, 2021, doi: 10.37148/bios.v2i1.17.

[9] M. A. W. K. MURTI, "Penerapan Metode K-Means Clustering Untuk Mengelompokan Potensi Produksi Buah – Buahan Di Provinsi Daerah Istimewa Yogyakarta," Skripsi, 2017.

[10] Hazman, A., Elektro, F. T., & Telkom, U. (2019). Penerapan Metode K-Means Clustering Untuk Mengelompokkan Data Pelabuhan Dan Bongkar Muat Barang Di Indonesia Application of K-Means Clustering Method for Grouping Port and Unload Loading Goods Data in Indonesia. 6(1), 1450–1454.

[11] G. Maciejewski, S. Mokrysz, and Ł. Wróblewski, "Segmentation of coffee consumers using sustainable values: Cluster analysis on the Polish coffee market," Sustain., vol. 11, no. 3, pp. 1–20, 2019, doi: 10.3390/su11030613.

[12] Hadinata, E. (2016). Simulasi Sistem Pendukung Keputusan Menggunakan Metode Klustering Algoritma Fuzzy C-Means. *Jurnal Ilmiah Media Sisfo*, *10*(1), 401–409. http://ejournal.stikom.db.ac.id/index.php/mediasisfo/article/view/186.

[13] Stefanović, H., Veselinović, R., Bjelobaba, G., & Savić, A. (2018). *An Adaptive Car Number Plate Image Segmentation Using K-Means Clustering*. 74–78. https://doi.org/10.15308/sinteza-2018-74-78

[14] Seta, P. T., & Hartomo, K. D. (2020). Mapping Land Suitability for Sugar Cane Production Using K-means Algorithm with Leaflets Library to Support Food Sovereignty in Central Java. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *6*(1), 15–25. https://doi.org/10.23917/khif.v6i1.9027

[15] Priambodo, Y. A., & Prasetyo, S. Y. J. (2018). Pemetaan Penyebaran Guru di Provinsi Banten dengan Menggunakan Metode Spatial Clustering K-Means (Studi kasus : Wilayah Provinsi Banten). *Indonesian Journal of Computing and Modeling*, *1*(1), 18–27. https://doi.org/10.24246/j.icm.2018.v1.i1.p18-27.

[16] Sari, B. N., & Primajaya, A. (2019). Penerapan Clustering Dbscan Untuk Pertanian Padi Di Kabupaten Karawang. *Jurnal Informatika Dan Komputer*, *4*(1), 28–34. www.mapcoordinates.net/en.

[17] Tamaela, J., Sediyono, E., & Setiawan, A. (2017). Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-means Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara. *Jurnal Buana Informatika*, *8*(3), 151–160. https://doi.org/10.24002/jbi.v8i3.1317.

[18] Alkhairi, P., & Windarto, A. P. (2019). Penerapan K-Means Cluster pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara. *Seminar Nasional Teknologi Komputer & Sains*, 762–767.

[19] Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, *78*(December 2015), 507–512. https://doi.org/10.1016/j.procs.2016.02.095.

[20] Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016.

[21] Kirana, M. C., & Etisa, L. M. (2017). Penerapan Mixture Model Kepada Aplikasi Helpdesk Berbasis Web. *NJCA (Nusantara Journal of Computers and Its Applications)*, *2*(1),17–23. https://doi.org/10.36564/njca.v2i1.26.