Implementation Of The K-Means Cluster Algorithm In Rice Production Mapping And As A Decision Support For Agricultural Function Transition

Wahyu Wijaya Widiyanto[1], Fendy Nugroho[2], Kusrini [3]

[123]Department of Informatics, University AMIKOM Yogyakarta, Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta, Central Java, 55283, Indonesia
E-mail: wahyu.widiyanto@students.amikom.ac.id[1],
fendy@amikom.ac.id[2], kusrini@amikom.ac.id[3]

Abstract

According to the Food and Agriculture Organization of the United Nations (FAO) in 2014 Indonesia ranked 3rd with a total rice production of 70.6 million tons, but it still remains a rice importing country. As one of the districts known as rice barn, Sukoharjo is targeted to continue to increase crop productivity every year to keep up with the growing population, so it is necessary to know areas with less optimal yields, and minimize changes in agricultural land use change. A mapping method for harvest results is needed to group data in each region based on the similarity of harvest data. In data mining, clustering techniques are known that can be used to map harvest productivity data based on their similarity. This study applies clustering techniques using the K-Means algorithm to map rice harvest productivity data by dividing data into 3 groups, namely many, medium, and less. The research method used is SDLC (Software Development Life Cycle) with a waterfall model. The K-Means algorithm is implemented using website-based programming to map harvest productivity data using attributes of planting area and rice production. The results of the mapping are visualized into a recommendation of agricultural land clustering and agricultural products as well as one of the decision makers in the transfer of agricultural functions so that sub-districts that have a lot of productivity, are moderate and lacking based on the characteristics of the data.

Keywords:*K-Means Cluster, Agriculture, SDLC, Website, Recommendations*

1. INTRODUCTION

Decision support systems are part of an information system that is used to support a decision based on predetermined criteria (Tomaž, 2006).

Decision support system is a tool used by managers to help take a decision based on data and criteria that already exist (Meiningsih, 2003).

Decision support systems can be interpreted as a part of knowledge-based information management systems and are used by organizations, institutions, and agencies to support decision making (Wijang, 2006). Rice is one of the most important staple foods in the world, especially in the Asian Continent, where rice is the staple food for the majority of the population, in the lower middle class. Based on data obtained from the FAO (Food and Agriculture Organization of the United Nations) organization in 2014, Indonesia ranked 3rd with a total production of 70.6 million tons (Riaddy, 2019). Even though Indonesia is one of the most rice-producing countries in the world, Indonesia is still an importer of rice. This situation is caused by Indonesia having the largest per capita rice consumption in the world, which is around 140 kilograms of rice per year. For countries that are still developing, such as Indonesia, the demand for infrastructure development in the form of roads, settlements and industrial estates also contributes to the demand for land. As a result, many paddy fields, especially those that are close to urban areas, switch functions to these uses (BPS, 2016). Sukoharjo Regency is known as the national rice granary area and was recorded by BPS (Central Bureau of Statistics) as the 11th rice producing region in Central Java Province from 29 districts (Kusrini, 2009). As a national rice barn, Sukoharjo Regency is targeted to continue to increase rice production every year. Increasing the amount of rice production in Sukoharjo Regency continues to be carried out to offset the

growing number of people. In order to meet the needs of rice, Sukoharjo District Agriculture Service seeks to continue to optimize the results of rice farming. A method is needed to map crop yields based on land area and crop production in each sub-district. The aim is to find out areas with rice yields that have not been optimal. So that it needs to get effective attention and handling because it relates to the policy making of the distribution of aid carried out by the Agriculture Service of Sukoharjo Regency. The policies taken must certainly have relevance and be supported by knowledge that comes from available data. In the world of computer science, data mining is widely known as a data extraction technique to find a hidden pattern in order to produce a new knowledge in a data set (Wijang, 2006). In particular, the data mining of Sukoharjo Regency is known as the national rice barn area and was recorded by BPS (Central Bureau of Statistics) as the 11th rice producer region in Central Java Province from 29 districts (Meiningsih, 2003). As a national rice barn, Sukoharjo Regency is targeted to continue to increase rice production every year. Increasing the amount of rice production in Sukoharjo Regency continues to be carried out to offset the growing number of people. In order to meet the needs of rice, Sukoharjo District Agriculture Service seeks to continue to optimize the results of rice farming. A method is needed to map crop yields based on land area and crop production in each sub-district. The aim is to find out areas with rice yields that have not been optimal. So that it needs to get effective attention and handling because it relates to the policy making of the distribution of aid carried out by the Agriculture Service of Sukoharjo Regency. The policies taken must certainly have relevance and be supported by knowledge that comes from available data. In the world of computer science, data mining is widely known as a data extraction technique to find a hidden pattern in order to produce a new knowledge in a data set (Wijang, 2006). Specifically data mining has its own method based on the purpose of utilizing data sets namely estimation, prediction, classification, clustering, and association.

One technique that can be used for the purpose of mapping a data is clustering. Clustering is a technique in data mining that functions to group data (grouping) based on its resemblance to a cluster. Each cluster has a set of data similar to other data in one cluster, but not similar to the data in another cluster (Riaddy, 2019). There are several algorithms that can be used, one of which is popularly used for clustering a dataset is KMeans. K-Means algorithm is a non-hierarchical clustering method that has relatively fast computing time. Based on comparative analysis between K-Means and Fuzzy C-Means (FCM) conducted by Suomi G. and Sanjay Kumar D., the results prove that the K-Means algorithm is faster with the elapsed time of 0.433755 seconds compared to the FCM algorithm that has elapsed time is 0.781679 seconds (BPS, 2016). Previously, research had been carried out relating to the utilization of data mining to process data sets of rice production in Indonesia, such as the use of clustering techniques to map the potential of rice plants from a dataset. However, the clustered data is presented only in the form of a data grouping table (Kusrini, 2009) (J. Han, 2012) so that research is needed in this regard. Based on the explanation, the researcher intends to use data from the Office Sukoharjo Regency Agriculture for clustering using the K-Means algorithm and as a way of making decisions in the transition of agricultural land functions based on rice production.

Drug clustering process is defined as a process that is carried out to break down a set of data and objects so that they become classes. The clustering process is a process that is carried out without supervision so that the data is broken down based on the calculation of the distance (Alfa, 2014)

k-means clustering is a method that functions to process data analysis and data modeling processes without any supervision or supervision process. This method works by doing the process of breaking data based on the group of data with the closest value, so that later groups will form several data partitions. In the process of grouping data there are several steps carried out by this method, among others, as follows (Suparto, 2010).

a. The initial stage in the grouping process is to determine how many partitions to make that can be called a value of k.

b. The next stage is to determine the centroid value (initial cluster center) obtained from the random data value.
c. Then do the calculation of the distance of each existing data to the centroid. The calculation process uses the ecludian distance method, the distance calculation process is done to determine the closest distance from all data with the centroid. The equation used to do the calculation is as follows:

$$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \quad (1)$$

d. Then the data classification process is done with the centroid. The classification process is done to determine the smallest distance.
e. The next step is to recalculate the centroid value based on the previous cluster average value. The equations used to do the calculations are as follows.

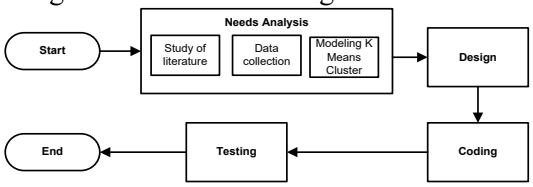
$$C_k = \frac{1}{n_k} \Sigma d_i \quad (2)$$

f. Repeat the calculation from step 2 to step 5, the new looping process will be stopped if the members of each cluster do not change.

If step 6 has been completed, then the average value of the cluster center in the last interaction will be used as a parameter for the data grouping process.

2. METHODS

The methodology carried out in this study uses SDLC (Software Development Life Cycle) with the waterfall model, starting with a needs analysis on the implementation of the k-means algorithm for mapping rice crop productivity. The method used in the needs analysis includes the study of literature, data collection. To implement the k-means algorithm, program code (programming) is compiled as the application of web-based system design using the programming languages PHP, HTML, CSS and Javascript. Furthermore, a black box testing, white box testing and a comparison of the test results of the application of the K-Means algorithm in applications with Rapidminer data mining tools to ensure the correctness of the results of clustering is carried out. The flow of system design can be seen in the figure. 1 below:



Figur. 1 Stage of Research Methodology

a. Needs Analysis

The need analysis phase is carried out to start the initial stage of software design using the SDLC (System Development Life Cycle) method using the waterfal model. The method carried out in the needs analysis consists of 3 stages, namely:

a) Literature Studies Search literature information from books, journals, theses and theses to find information that is relevant to the research being conducted
b) Data Collection This stage is collecting data and information related to the topic of research using observation and interview techniques.
c) K-Means Modeling This stage is done by data mining modeling for the application of K-Means algorithm. In this phase, K-Means pseudo-code algorithm will be created as a flow model in the program.

b. Design

In this section, the design process of software design is done using modeling. At this stage 4 UML diagrams are used, namely use case diagrams, activity diagrams, sequence diagrams to represent functions and flows of software, and class diagrams to represent database designs to be built.

c. Coding

This stage is the process of designing a data clustering system using the programming languages PHP, HTML, CSS, JavaScript, and supporting frameworks such as RaphaelJS, Jquery, and Bootstrap to improve the system that is designed both in terms of appearance or performance.

d. Testing

Evaluation of clustering results is done by comparing the results of the cluster in the

Rapidminer application (data mining tools) with the application being built.

In Rapidminer's data mining tools, a trial process is carried out with the same dataset to compare the results of clustering between applications built with Rapidminer's data mining tools for clustering

3. RESULTS AND DISCUSSION

a. Results of Need Analysis

1) Data Collection:
The results of observations and interviews obtained data on rice crop production in Sukoharjo Regency, which was spread in 12 Subdistricts from 2013 to 2016 with attributes taken as input data in the study as follows:
   a) Name of region / sub-district
   b) The area of land / paddy fields planted with rice (ha) per sub-district.
   c) Amount of rice harvest production (Ton) per sub-district.
   d) Productivity of rice harvest (Kw / Ha) per district.
   e) Year of the rice harvest period (Year).

The system for clustering rice production data is intended for 2 types of users according to the member structure in the Sukoharjo Regency Agriculture and Forestry Service in the Food Section, namely:

a. Administrators are types of users who have the highest access rights to manage master user data and manage user access rights. Administrators can be represented by the IT department who is the system administrator.
b. Analysts are the types of users whose task is to cluster the rice harvest data and provide analysis of cluster results from the resulting clusters. There are three clusters that are used to classify rice yield data, and each cluster result will be determined by the user analyst based on the analysis carried out. Analysts can be represented by the statistics department in charge of processing crop data at the Agriculture Service.

Dataset merupakan sekumpulan data yang didapatkan dari objek, karakteristik maupun sifat yang disimpan dan dikumpulkan berdasarkan kedekatan dan kemiripannya (Felicia, 2014).

Dataset juga dapat diartikan sebagai gabungan kumpulan dari beberapa data yang disimpan dalam sebuah basis data (Alfa, 2014) (Suparto, 2010).

Dataset bisa diartikan sebagai sebuah himpunan yang menyimpan data-data dengan sifat, karakter, bentuk, pola serta kebiasaan yang sama[12]

Dalam penelitian ini digunakan dataset data dummy tentang produksi hasil pertanian di daerah pertanian Sukoharjo Jawa Tengah.

2) K-Means Modeling: K-Means modeling is modeled based on the K-Means algorithm flowchart in figure 2.

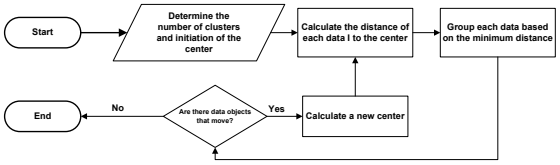
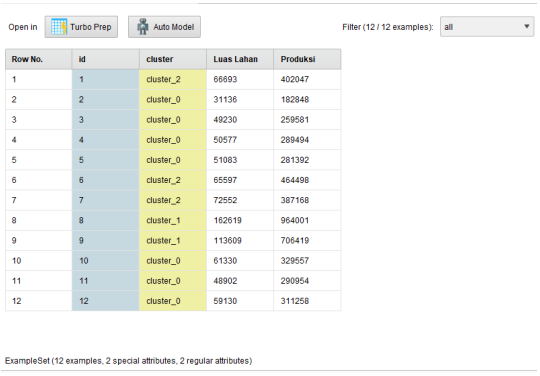


Figure 2. Flowchart of K-Means Algorithm

3) Testing

In order for the results of the research to be carried out to have good accuracy, two scenarios were carried out in the dataset test, namely with rapidminer data mining tools and the design prototype of the researcher. The results of clustering on the system built can be seen in Figure 3 below. In Rapidminer's data mining tools, the results of experiments with the same dataset to compare the results of clustering between applications built with Rapidminer's data mining tools using the clustering process design can be seen in Figure 4 and Figure 5.

The setting of the k-means parameter on Rapidminer is adjusted to the number of clusters and the distance calculation formula in the application that is built which is k = 3 with the formula for calculating distance is the euclidean distance.



Open in Turbo Prep Auto Model Filter (12 / 12 examples): all

Row No.	id	cluster	Luas Lahan	Produksi
1	1	cluster_2	66693	402047
2	2	cluster_0	31136	182848
3	3	cluster_0	49230	259581
4	4	cluster_0	50577	289494
5	5	cluster_0	51083	281392
6	6	cluster_2	65597	464498
7	7	cluster_2	72552	387168
8	8	cluster_1	162619	964001
9	9	cluster_1	113609	706419
10	10	cluster_0	61330	329557
11	11	cluster_0	48902	290954
12	12	cluster_0	59130	311258

ExampleSet (12 examples, 2 special attributes, 2 regular attributes)

Figure 4. Rapidminer clustering

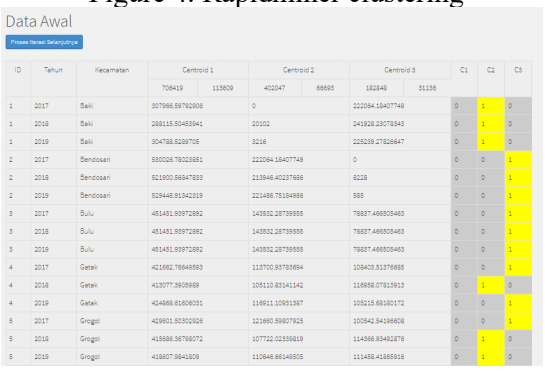


Figure 5. Prototype clustering (researcher application)

The results of clustering using Rapidminer tools on the dataset consisting of 12 data with 2 special attributes consisting of sub-districts and clusters, and 2 regular attributes used for the clustering calculation process are shown in Figure 6.

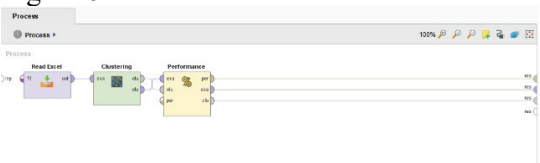


Figure 6. Rapidminer Process Design

Grouping each data using the Cartesian diagram found in Rapidminer tools is shown in Figure 7.

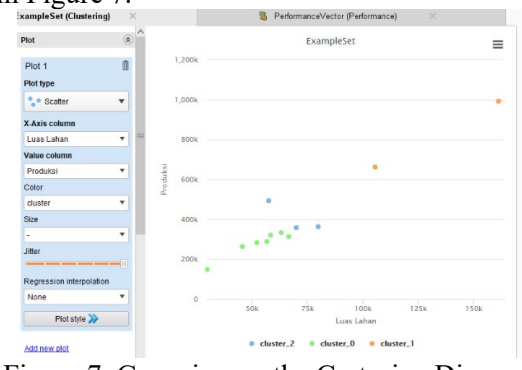


Figure 7. Grouping on the Cartesian Diagram

b. Discussion

The implementation of the K-Means algorithm for mapping rice harvest productivity in Sukoharjo Regency uses a web-based application, can group data and represent the results of grouping harvest productivity data into a base map that has color attributes based on the values of cluster members in each region. The number of clisters used in the application designed is 3 clusters (k = 3). This is based on the purpose of grouping, namely to find out areas with productivity that are less than the target, according to the target, and more than the target. Information obtained from the results of data clustering has been carried out, namely knowing areas that have crop productivity that is less than the target, on target, and exceeds the target. So that conclusions can be drawn to give a decision that the agricultural land of the cluster can be converted or not based on the parameters shown in table 1 below.

Table 1. Recommended Table of Decision Parameters

Criteria	Results		
	Many	**Medium**	**Less**
Production result	Every year increases	Decrease	Every year decreases
Land Condition	Fertile in accordance with the standard processing of agricultural land	It should be observed because the soil conditions have decreased	Infertile and not according to standards for agricultural land
Recommendation	Production Level	Increase agricultural production and alternative preparation for plantation land	Change the function of land from productive land to green open land

The initial appearance of the system prototype is shown in Figure 8.

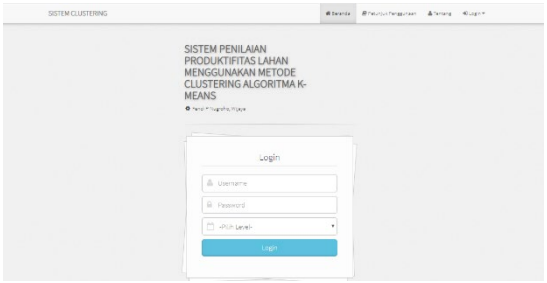


Figure 8. Login page

The prototype in this study was made using two accesses, namely admin and analysis (discussed in the sub-chapter of the results of discussion analysis). For admin access, data input consists of input data from districts, sub-districts, years, and the amount of agricultural production shown in Figure 9 and Figure 10.

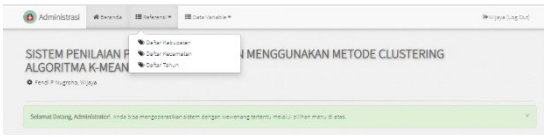


Figure 9. Admin Access to the Research Prototype

Figure 10. Admin Access Research Prototype

As for analysts access has access to generate centroids, i-terasi K-means cluster, and see the results of recommendations based on searches by city / sub-district, shown in figure 11.

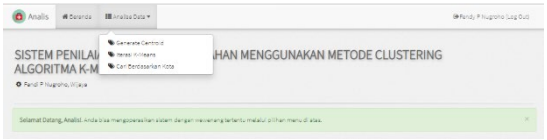


Figure 11. Analyst Access to Research Prototype

The results of cluster evaluation by comparing with Rapidminer's data mining tools with cluster 3 results in grouping data with the same cluster members. The results of the search prototype recommendations based on city / sub-district can be seen in figure 12.

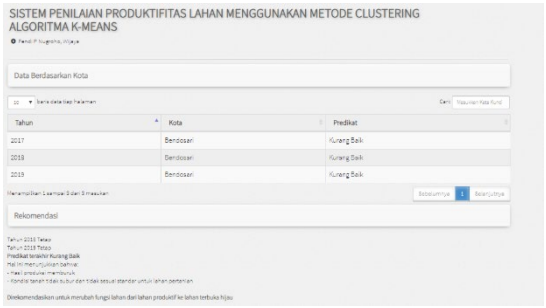


Figure 12. Results of Recommendations

4. CONCLUSION

Based on the results of the trial and analysis explained, data on paddy harvest productivity in Sukoharjo Regency can be mapped using data mining grouping techniques into 3 groups consisting of harvest productivity exceeding the target, according to target, and less than the target with a percentage increase in agricultural production and as a reference for the conversion of agricultural land based on clustering. The application of the k-means cluster algorithm can implemented using the SDLC software development method with the waterfall model on web-based programming. The results of comparison of applications built with Rapidminer's data mining tools show the results of grouping with cluster members same.

5. REFERENCES

Alfa Saleh, Klasifikasi Metode Naive Bayes Dalam Data Mining Untuk Menentukan Konsentrasi Siswa (Studi Kasus Di Mas PAB 2 Medan), KeTIK (Konferensi Nasional Pengembangan Teknologi Informasi dan Komunikasi)

BPS, ”Produksi Padi Menurut Kabupaten/Kota di Jawa Barat (Ton),” [Online]. Available: http://jabar.bps.go.id/linkTabelStatis/view/id/135. [Accessed 23 November 2016].

Ghosh, S. and S. K. Debey, ”Comparative Analysis of K-Means and Fuzzy C-Means Algorithms,” vol. 4, no 35-39, 2013.

J. Han, M. Kamber och J. Pei, Data Mining: Concepts and echniques, Waltham: Elsevier, Inc, 2012.

Kusrini and E. T. Luthfi, Algoritma Data Mining, Yogyakarta: Andi Offset, 2009.

L., Felicia, ”Penerapan Metode Clustering Dengan K- Means Untuk Memetakan Potensi Tanaman Padi Di Kota Semarang,” 2014.

Meiningsih, Siti. Pengembangan Sistem Informasi Manajemen LIPI Terpadu (Simlita). Prosiding Pemaparan Hasil Litbang IPT 2003.

Riaddy, D., ”Ini 5 Negara Penghasil Beras Terbesar didunia,” [Online]. Available: http://bisniskeuangan.kompas.com/read /2015/09/02/0951000. [Accessed 22 Juni 2019].

Suparto Darudiato, 2010, Perancangan Data Warehouse Penjualan Untuk Mendukung Kebutuhan Informasi Eksekutif Cemerlang Skin Care, Seminar Nasional Informatika 2010 (semnasIF 2010),”UPN Veteran” Yogyakarta

Tomaž Seljak and Aleš Bošnjak. Researchers’ bibliographies in COBISS.SI. Information Services & Use 26 (2006) 303–308

Wijang Widhiarso. Pemetaan Teknologi Informasi/Sistem Informasi untuk Mendukung Strategi Fungsional Perusahaan (Studi Kasus: Bisnis Jasa Gadai PT. XZY). Seminar Nasional Aplikasi Teknologi Informasi 2006. Yogyakarta, 17 Juni 2006