Jurnal Teknik Informatika C.I.T Medicom, 15 (2) (2023) 73-83

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Comparison of distance metric in k-mean algorithm for clustering wheat grain datasheet

Suraya[1], Muhammad Sholeh[2], Dina Andayati[3]

[1)]Computer Systems Engineering Study Program, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia

[2)]Informatics Study Program, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia

[3)]Digital Business Study Program, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia

Article Info

Article history

Received : Apr 18, 2023
Revised : Apr 23, 2023
Accepted : Apr 27, 2023

Kata Kunci:

Datasheet;
Distance matrix.
Clustering;
K-means.

Abstract

One of the data mining models is clustering, clustering models can be used to create groupings of data. Clustering is done by creating groups of data that are close to each other. The research was conducted by clustering wheat seed datasheets. The wheat grain datasheet contains various types of wheat data. The purpose of this research is to create a clustering model. The algorithm used is the K-means algorithm and a comparison is made with several distance Metric algorithms. The datasheet used was tested with the K-means algorithm and tested the clustering value (k) ranging from k = 2 to k = 6. Comparison of clustering results with K-means is also done by comparing with distance metric algorithms, namely Euclidean distance, Manhattan distance, and Chebychev distance. All testing processes are evaluated, and the evaluation is done to select many good groupings. The evaluation process is carried out using the Davis-Bouldin method. The results of the grouping that has been done, each seen Davis Bouldin evaluation. The evaluation value of Davis Bouldin is sought from the smallest value and if the evaluation result is negative, the value is solved. The research method used is Knowledge Discovery in Database (KDD). The results showed that the same datasheet and using the K-means algorithm and the same evaluation resulted in different evaluation values. The Euclidian, Manhattan, and Chebychev algorithms produce the best k value of 2, The conclusion of the wheat seed datasheet clustering research produces a value of k = 2.

Corresponding Author:

Suraya,
Computer Systems Engineering Study Program, Faculty of Information Technology and Business,
Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia
Jl. Kalisahak No.28 Kompleks Balapan, Yogyakarta, Indonesia, 55222
Email : mymuhash@gmail.com

1. Introduction

Data grouping is one of the most important processes in the data processing. Data grouping makes it easier to process data. The data grouping process in data mining can use a clustering model. Apart from clustering models, other models in the data mining process are regression, forecasting, classification, and associatio [1][2].

The method used to group data into several groups or clusters is the clustering model. The clustering results will produce data groups that are similar or have maximum similarity characteristics.

Research related to data mining with various models and using various datasheets have been studied by several researchers. Research related to data mining including regression models is carried out[3][4][5][6][7]. Classification models are performed [8][9][7]. Research related to clustering methods and using the K-means algorithm using various datasheets have been conducted. These studies include [10][11][12][13][14][15][16].

Bastian conducted clustering research with human infectious disease datasheets. The data clustering process uses the K-means r algorithm and the datasheet is data processed from Puskesmas data in Majalengka Regency. Clustering research using Rapid miner implementation and using the K-Means method was conducted by [11]. Datasheet processed from BPS (Central Bureau of Statistics). The case study processed data on measles immunization in children under five years old. based on Province. The results produced 3 clusters with categories of high cluster level (C1), medium cluster level (C2), and low cluster level (C3).

The grouping of Covid 19 distribution levels was researched by [14] . There are about 16,284 data processed. The result of the clustering is the level of distribution of Covid 19 per province. The distribution is grouped in several clusters so that it can be seen which areas have the highest number of cases and the least.

Implementation of clustering mode data mining methods can use various applications such as Python, Rapid Miner, R, Orange, and others. Implementation of clustering models using Python is done [17][18][19][20]. Clustering research using Rapid Miner was researched by [21][22][23]. The use of Rapid Miner in the process of creating clustering models is done Amanda [22]. Rapid Miner is used to implementing product analysis. The clustering results produce groupings of products that are often used and those that are rarely used.

Research that compares the K Means algorithm by comparing distance metrics includes [24][25]. Religia [24], conducted research with the village potential datasheet in 2014. The processed data totaled 74093. The data clustering process is carried out using the k- means algorithm and the calculation of the closest distance from data to a centroid point by comparing the Manhattan, Euclidean, and Chebychev distance calculation methods.

According to Nishom [25], in the research conducted, data grouping using the K-Means algorithm has weaknesses, including the problem of the level of accuracy between points with other points. One way that can be done is by comparing using Euclidean distance, Manhattan distance, and Minkowski distance. The datasheet used is the data used to determine the status of disparity in teacher needs, especially teachers in Tegal City.

Based on the discussion above, one of the data mining models that can be done is clustering data. The problem in the research is how to produce a clustering of wheat seed data. This clustering aims to distinguish wheat seeds that have good, medium, or poor quality. The process to produce many clusterings is done using clustering techniques.

The process to overcome these problems is done using the K-means algorithm and to get maximum results, the existing datasheet is carried out with the process of measuring the similarity and proximity of a point to another point. The process of grouping data is done by looking at the proximity of distance between one point and the point. The algorithm used to compare the closest points uses Euclidean distance, Manhattan distance, and Chebychev distance. The results of clustering are done by testing many groupings ranging from groupings 2 to 7. The best results of clustering are tested by evaluating the model using Davis Bouldin,

The benefits of this research provide the results of wheat seed clustering and users can determine the clustering results by giving category names, such as categories with good, ordinary, or bad seed composition.

2. Research Methods

Research methodology using *Knowledge Discovery in Database* (KDD) [26]. The stages of the KDD process are presented in Figure 1.

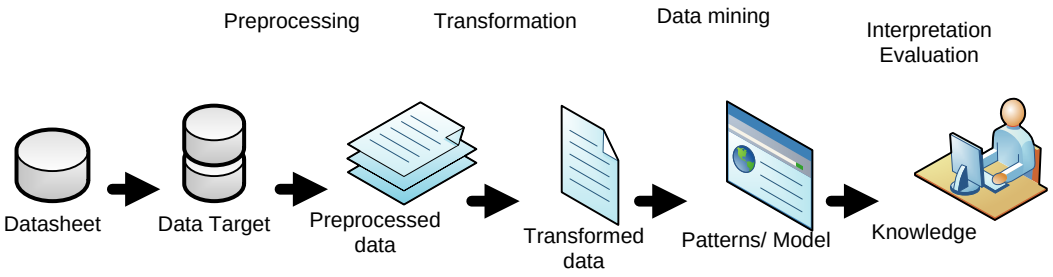

Figure 1 Process stages in research with *Knowledge Discovery in* Database

The KDD process as shown in Figure 1, starting from the determination of the datasheet and the process of generating the model, must first stop the target data, data processing, data transformation, model creation, and model evaluation.

2.1. Datasheet.

Datasheet processed from http://archive.ics.uci.edu/ml/datasets/seeds . The datasheet consists of 210 data and 8 columns, namely Id, area A, perimeter, compactness, length of the kernel, width of the kernel, asymmetry coefficient, length of kernel groove. All attribute types are number types. Figure 2, is an example of a wheat germ datasheet used in the research. The implementation process of the wheat grain datasheet clustering model using Rapid Miner [27].

ID	area	perimeter	compactness	lengthOfKer...	widthOfKern...	asymmetry...	lengthOfKer...
1	15.260	14.840	0.871	5.763	3.312	2.221	5.220
2	14.880	14.570	0.881	5.554	3.333	1.018	4.956
3	14.290	14.090	0.905	5.291	3.337	2.699	4.825
4	13.840	13.940	0.895	5.324	3.379	2.259	4.805
5	16.140	14.990	0.903	5.658	3.562	1.355	5.175
6	14.380	14.210	0.895	5.386	3.312	2.462	4.956
7	14.690	14.490	0.880	5.563	3.259	3.586	5.219
8	14.110	14.100	0.891	5.420	3.302	2.700	5
9	16.630	15.460	0.875	6.053	3.465	2.040	5.877
10	16.440	15.250	0.888	5.884	3.505	1.969	5.533
11	15.260	14.850	0.870	5.714	3.242	4.543	5.314
12	14.030	14.160	0.880	5.438	3.201	1.717	5.001
13	13.890	14.020	0.888	5.439	3.199	3.986	4.738

Figure 2, datasheet of wheat grains (seeds)

The example datasheet shown in Figure 2, is the data content of wheat kernels processed with the clustering model. According to the K-Menas algorithm, all the contents of the dataset are in integer form and no attribute is a label.

2.2. K-Means Algorithm

In the clustering model, the K-Means algorithm is one of the algorithms that can be used in clustering datasheets. The datasheet used does not yet have a label that can distinguish a grouping [28][29][30].

2.3. Distance metric

The distance metric is a method to measure the similarity and proximity of a point to another point. The data grouping process is done by looking at the proximity of the distance between one point and another. The distance metric method can affect the results of making data grouping.

2.4. Euclidean distance.

Euclidean distance is an algorithm used to measure the distance measured between two vectors by calculating the square root of the sum of the squared differences between them [31].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^{\ 2}} \tag{1}$$

For example, suppose there are two points in the 3D dimension (x, y, z), namely A(1, 2, 3) and B(4, 5, 6). Then the calculation of the Euclidean Distance between the two points is as follows:

d(A,B) = √((1-4)^2 + (2-5)^2 + (3-6)^2)
= √((-3)^2 + (-3)^2 + (-3)^2)
= √(9 + 9 + 9)
= √27
= 5.196

So, the Euclidean Distance between points A and B in the 3D dimension is 5,196.

2.5. Manhattan distance

Manhattan distance is a measurement metric commonly used to calculate the distance between two data points in a grid-like path [31].

$$d(x,y) = \sum_{i=1}^{n} |\, x_i - y_i| \tag{2}$$

Suppose there are two points in 2D dimension (x, y), namely A(1, 2) and B(4, 5). Then the calculation of Manhattan Distance between the two points is as follows:

d(A,B) = |1-4| + |2-5|
= |-3| + |-3|
= 6

So, the Manhattan Distance between points A and B in 2D dimension is 6.

2.6. Chebychev distance

The maximum distance value or *Chebychev distance is* calculated by performing the calculation of the absolute result of the difference between a pair of objects. [26].

$$d(i,j) = \lim_{n\to\infty}\left(\sum_{f=1}^{p}|x_{if} - x_{jf}|^{\ h}\right)^{1/h} \tag{3}$$

Suppose there are two points in 2D dimension (x, y), namely A(1, 2) and B(4, 5). Then the calculation of Chebyshev Distance between the two points is as follows:

d(A,B) = max(|1-4|, |2-5|)
= max(3, 3)
= 3

So, the Chebyshev Distance between points A and B in 2D dimension is 3.

2.7. Davis bouldin.

The davies-bouldin index (DBI) is one of the methods that can be used in the evaluation process. [32].

$$DBI = \frac{1}{K}\sum_{i=1}^{k} \max_{i \neq j}(R_{i,j}) \quad (4)$$

$$SSW = \frac{1}{N}\sum_{i=1}^{n} || x_i - c_{pi} ||^2 \quad (5)$$

where

$$SSB = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq 1}^{n} || C_i - C_j ||^2 \quad (6)$$

The following is an example of calculating the Davis Bouldin Index for two clusters in the dataset:

Table 1.
Calculating the Davis Bouldin Index for two clusters

No	X1	X2
1	2	3
2	3	2
3	2	2
4	4	5
5	5	4
6	5	5

Suppose the clustering results in two clusters, that is:
Cluster 1: {1, 2, 3}
Cluster 2: {4, 5, 6}
Then, we can calculate the DBI value by:
Calculate the centroid of each cluster.
Centroid of Cluster 1: (2.33, 2.33).
Cluster 2 centroid: (4.67, 4.67).
Calculate the distance between each cluster centroid to the other cluster centroids.
Distance between centroids of Cluster 1 and Cluster 2: sqrt((2.33-4.67)^2 + (2.33-4.67)^2) = 3.32
Calculate the Rij value for each cluster
R1 = (S1 + S2) / d(C1, C2) = ((0.47 + 0.23 + 0.47) / 3) / 3.32 = 0.38
R2 = (S4 + S5 + S6) / d(C1, C2) = ((0.47 + 0.23 + 0.47) / 3) / 3.32 = 0.38
where S is the match value within the cluster, and d is the distance between the cluster centres.
Calculate the DBI value
DBI = (R1 + R2) / k = (0.38 + 0.38) / 2 = 0.38
where k is the number of clusters.
So, the Davis Bouldin Index (DBI) value for the clustering result with two clusters on the dataset is 0.38. The lower the DBI value, the better the clustering quality.

3. Results and Discussion

3.1. Model making

The K-Means algorithm is used in making clustering models [13],, before making the model, the data cleaning and data checking processes are carried out. Data checking is done by selecting the attributes used. Figure 3 Data checking process and selection of attributes used.

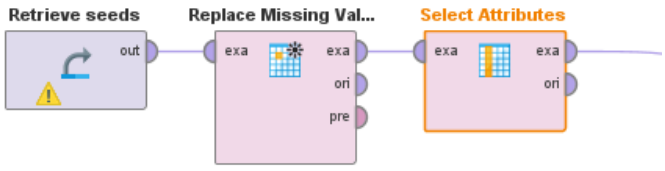

Figure 3. The process of checking the data and selecting the attributes used.

The process in Figure 7, uses the replace missing value operator which is used to replace empty attribute data with the average value, and the select attribute operator to determine which attributes are not used.

The datasheet is cleaned mainly from empty data and checking for data that deviates from the average (outliers). Another process is the selection of attributes used in the formation of clustering. Data grouping is done using the K means algorithm and testing is done with groupings ranging from k = 2 to k = 7. At this stage, the process of calculating different distances is carried out.

3.2. Comparison of model results

The results of the selected clustering model are compared by evaluating the results using the Davis-Bouldin method. The process of creating a clustering model will select the best number of groupings. To get the best k (clustering) value, the modeling process is done by comparing distance measurements using the Euclidean, Manhattan, and ChebychevD algorithms [22]. The clustering process was carried out by making groupings ranging from 2 groupings to 7 groupings. The results of 7 groupings were evaluated using Davies Bouldin. Figure 4 clustering process from k=1 to k=7 using Euclidian distance metric. Figure 5, the clustering process from the value of k = 1 to k = 7 using the Manhattan Metric distance metric. Figure 6, the clustering process from the value of k = 1 to k = 7 using the ChebychevD Metric distance metric.

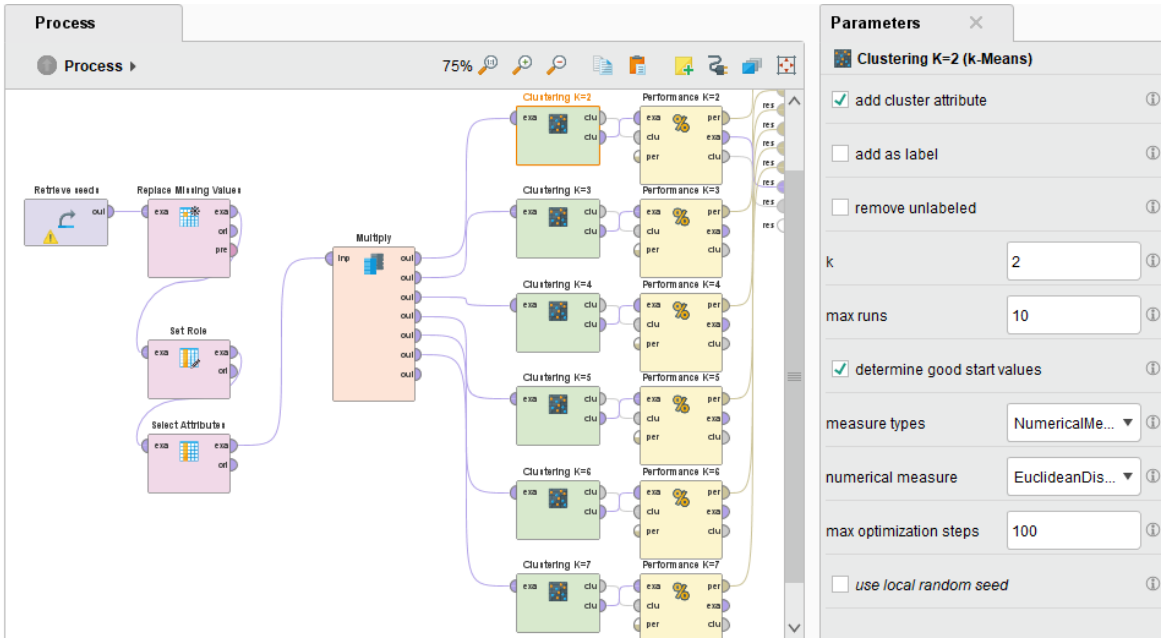

Figure 4. Clustering process from k=1 to k=7 using Euclidian distance metric.

The results of the clustering process in Figure 4, the use of the Chebychev distance metric. Performed by selecting the K-mens operator and in the parameters selecting the numerical measure with Euclidian distance.

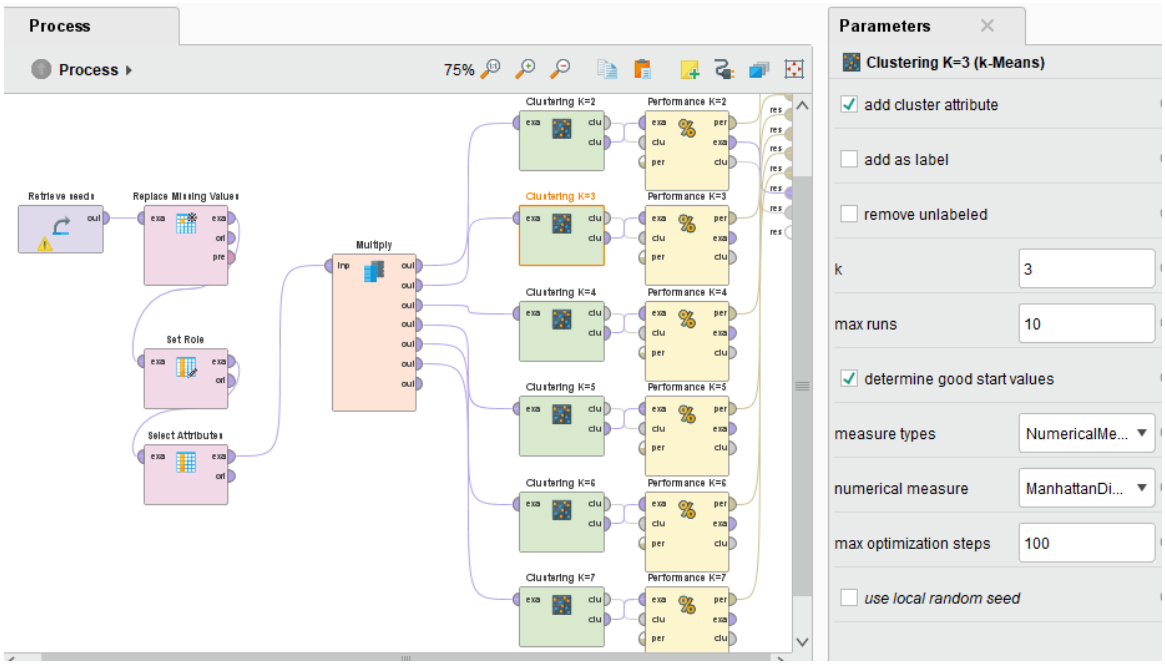

Figure 5. Clustering process from k=1 to k=7 using Manhattan distance metric.

The results of the clustering process in Figure 5, the use of the Chebychev distance metric. Performed by selecting the K-mens operator and in the parameters selecting the numerical measure with Manhattan distance metric.

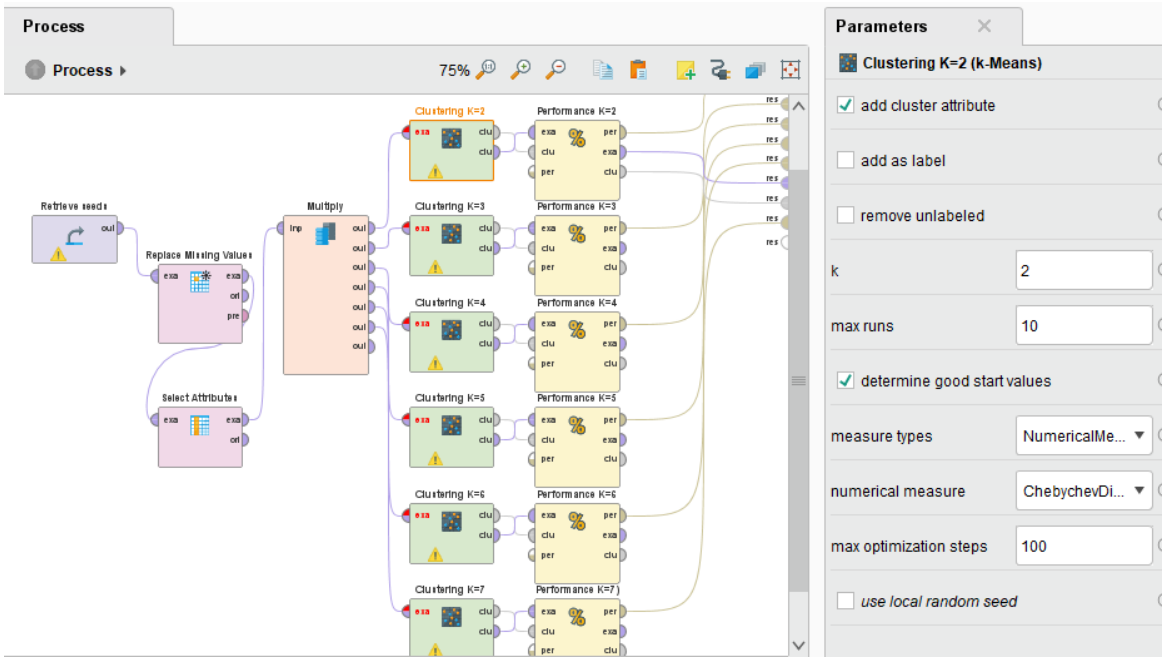

Figure 6. Clustering process from k=1 to k=7 using Chebychev distance metric.

The results of the clustering process in Figure 6, the use of the Chebychev distance metric. Performed by selecting the K-mens operator and in the parameters selecting the numerical measure with Chebychev

3.3. Davies Bouldin Results

The results of the process of creating a clustering model with 3 distance metrics are compared with the Davies Boldin value for each K value. Many clusters are selected based on the smallest Davies Bouldin value. Table 2 evaluation comparison with Davies Bouldin for each K value and distance metric method.

Table 2.
Comparison of evaluation with Davis Bouldin

Many K	Euclidean	Manhattan	ChebychevD
2	0.689	0.667	0.666
3	0.753	0.768	0.753
4	0.891	0.895	0.895
5	0.915	1.213	0.939
6	0.918	0.963	0.920
7	0.951	0.965	0.945

Based on table 2, the results of the Davies Bouldin calculation from the value of k = 2 to K = 7. The selection of the number of groupings is the value of the evaluation calculation results with the smallest value [25]. The calculation results show the value of K = 2 for each distance metric is the smallest value and the best result is grouping as much as 2.

3.4. Data grouping results

Although the recommended number of groupings for each distance metric method is the same, namely 2 groups, the therapy grouping results on the datasheet are different. The clustering results of each distance metric method are presented in Table 3.

Table 3.
Clustering results of each distance metric algorithm

Distance Metric	Clustering Results
Manhattan Distance	Cluster 0: 76 data Cluster 1: 134 data
Euclidean Distance	Cluster 0: 127 data Cluster 1: 83 data
Chebychev Distance	Cluster 0: 75 data Cluster 1: 135 data

Based on Table 3, the results of clustering with the Manhattan distance algorithm, group 0 has as much as 76 data and group 1 as much as 134. Grouping with the Euclidean Distance algorithm, resulting in grouping 0 as much as 127 data 83. Grouping Chebychev Distance data, group 0 as much as 75 data and group 1 as much as 135 data.

The results of data clustering can be presented in the form of data and visualization. Figure 7, clustering results using the Euclidean distance metric distance algorithm.

id	cluster	area	perimeter	compactness	lengthOfKer...	widthOfKern...	asymmetry...
1	cluster_0	15.260	14.840	0.871	5.763	3.312	2.221
2	cluster_0	14.880	14.570	0.881	5.554	3.333	1.018
3	cluster_0	14.290	14.090	0.905	5.291	3.337	2.699
4	cluster_0	13.840	13.940	0.895	5.324	3.379	2.259
6	cluster_0	14.380	14.210	0.895	5.386	3.312	2.462
7	cluster_0	14.690	14.490	0.880	5.563	3.259	3.586
8	cluster_0	14.110	14.100	0.891	5.420	3.302	2.700
11	cluster_0	15.260	14.850	0.870	5.714	3.242	4.543
12	cluster_0	14.030	14.160	0.880	5.438	3.201	1.717
13	cluster_0	13.890	14.020	0.888	5.439	3.199	3.986
14	cluster_0	13.780	14.060	0.876	5.479	3.156	3.136

Figure 7. Example of cluster 0 data with Euclidean distance metric

Figure 7, an example of data entering cluster 0 from the data in Figure 7, data numbers 1,2,3,4,6,7,8 are included in cluster 0, and data numbers 5,9,10 are included in cluster 1. The clustering results in cluster 1 are presented in Figure 8.

id	cluster ↓	area	perimeter	compactness	lengthOfKer...	widthOfKern...	asymmetry...
5	cluster_1	16.140	14.990	0.903	5.658	3.562	1.355
9	cluster_1	16.630	15.460	0.875	6.053	3.465	2.040
10	cluster_1	16.440	15.250	0.888	5.884	3.505	1.969
18	cluster_1	15.690	14.750	0.906	5.527	3.514	1.599
23	cluster_1	15.880	14.900	0.899	5.618	3.507	0.765
26	cluster_1	16.190	15.160	0.885	5.833	3.421	0.903
32	cluster_1	15.490	14.940	0.872	5.757	3.371	3.412
36	cluster_1	16.120	15	0.900	5.709	3.485	2.270
37	cluster_1	16.200	15.270	0.873	5.826	3.464	2.823
38	cluster_1	17.080	15.380	0.908	5.832	3.683	2.956
44	cluster_1	15.500	14.860	0.882	5.877	3.396	4.711

Figure 8. Example of cluster 1 data with Euclidean distance metric

The results in Figure 8, show that the data in numbers 5,9,10,18,23 are included in cluster 1. In addition to presenting in the form of data, the data presentation process can be displayed in the form of graphs. The results of clustering with Euclidean distance metric in graph form are presented in Figure 9.

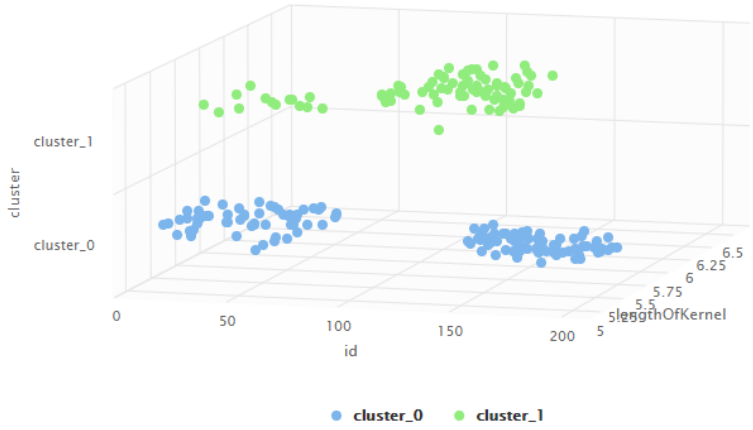

Figure 9 . Visualization of datasheet clustering hail with Euclidean distance metric.

The visualisation results as presented in Figure 9 show the clustering of wheat seed data that is close to each other. The data visualisation at the top is a grouping for data that belongs to cluster 1 and the data grouping visualisation at the bottom belongs to cluster 0.

The results of research using distance metric calculations using the Euclidean, Manhattan and Chebychev algorithms produce the same K value K = 2. The difference between the 3 distance algorithms is in the grouping of data. Thus the proposal from the grouping can be divided into a collection of good and bad seeds.

4. Conclusions

The data mining clustering model is used to perform groupings of a datasheet. The more data to be clustered, the more difficult the clustering process. The clustering process can be done using data mining. The clustering process on the wheat seed datasheet produces a K value of 2. This K = 2 value shows the results of clustering wheat seed data are 2 groups. The results of research using distance

metric calculations using the Euclidean, Manhattan, and Chebychev algorithms produce the same K value of K=2. The difference between the 3 distance algorithms is in the grouping of data, the results of grouping with the Manhattan distance and Chebychev distance algorithms, a lot of data is grouped almost the same while the Euclidean distance grouping results have quite a lot of differences. This research contributes to the development of clustering models in data mining because it helps users in choosing the most suitable clustering algorithm for comparison use cases with several distance metrics. The implications obtained include the use of clustering on datasets, especially Seeds datasheets, which can provide options for determining good wheat seeds. Limitations in the study include the data used in a public datasheet and the recommendation of the research results is the grouping of wheat seeds using the Euclidean Distance metric distance algorithm. Suggestions for further research are to develop the system by adding features to the datasheet used and using the latest machine learning systems and methods.

References

[1] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Yogyakarta: Penerbit Andi, 2020.

[2] D. Cielen, A. D. B. Meysman, and M. Ali, *Introducing Data Science*. New York: Manning Publications, 2016.

[3] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, no. 102817, pp. 13–24, 2022, doi: https://doi.org/10.1016/j.ssresearch.2022.102817.

[4] Ekka Pujo Ariesanto Akhmad, "Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran," *J. Apl. Pelayaran dan Kepelabuhanan*, vol. 10, no. 2, pp. 120–131, 2020, doi: 10.30649/japk.v10i2.83.

[5] S. M. A. A, O. M. E. E, S. M. A. A, and T. F. Sarnaghi, "Determination of households benefits from subsidies by using data mining approaches," *J. Inf. Technol. Polit.*, vol. 19, no. 3, pp. 1–20, 2022, doi: https://doi.org/10.1080/19331681.2022.2097974.

[6] Y. L. Jin Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *J. Evidence-Based Med. Online Libr.*, vol. 13, no. 1, pp. 57–69, 2020, doi: https://doi.org/10.1111/jebm.12373.

[7] D. K. Sharma, S. Lohana, S. Arora, A. Dixit, M. Tiwari, and T. Tiwari, "E-Commerce product comparison portal for classification of customer data based on data mining," in *Materials Today: Proceedings*, 2022, vol. 51, no. 1, pp. 166–171, doi: https://doi.org/10.1016/j.matpr.2021.05.068.

[8] K. Deepika and N. Sathyanarayana, "Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques," *Int. J. Comput. Eng. Res.*, vol. 8, no. 9, pp. 28–38, 2018.

[9] A. O. Oyedeji, A. M. Salami, O. Folorunsho, and O. R. Abolade, "Analysis and Prediction of Student Academic Performance Using Machine Learning," *J. Inf. Technol. Comput. Eng.*, vol. 4, no. 1, pp. 10–15, 2020, doi: https://doi.org/10.25077/jitce.4.01.10-15.2020.

[10] A. Bastian, H. Sujadi, and G. Febrianto, "Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka)," *J. Sist. Inf. (Journal Inf. Syst.*, vol. 14, no. 1, pp. 26–32, 2018, doi: https://doi.org/10.21609/jsi.v14i1.566.

[11] A. Fotouhi and M. Montazeri-Gh, "Tehran driving cycle development using the K-means clustering method," *Sci. Iran.*, vol. 20, no. 2, pp. 286–293, 2013, doi: https://doi.org/10.1016/j.scient.2013.04.001.

[12] W.-J. Son and I.-S. Cho, "Analysis of Trends in Mega-Sized Container Ships Using the K-Means Clustering Algorithm," *Appl. Sci.*, vol. 12, no. 4, pp. 10–17, 2022, doi: https://doi.org/10.3390/app12042115.

[13] J. Vijay and J. Subhashin, "An efficient brain tumor detection methodology using K-means clustering algoriftnn," in *2013 International Conference on Communication and Signal Processing-IEEE Xplore*, 2013, pp. 653–657, doi: doi: 10.1109/iccsp.2013.6577136.

[14] T. Hardiani, "Analisis Clustering Kasus Covid 19 Di Indonesia Menggunakan Algoritma K-Means," *Janapati*, vol. 11, no. 2, pp. 156–165, 2022, doi: https://doi.org/10.23887/janapati.v11i2.45376.

[15] A. Lia Hananto *et al.*, "Analysis of Drug Data Mining with Clustering Technique Using K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1908, no. 1, 2021, doi: 10.1088/1742-6596/1908/1/012024.

[16] K. Rahayu, L. Novianti, and M. Kusnandar, "Implementation Data Mining with K-Means Algorithm for Clustering Distribution Rabies Case Area in Palembang City," *J. Phys. Conf. Ser.*, vol. 1500, no. 1, 2020, doi: 10.1088/1742-6596/1500/1/012121.

[17] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, no. 114060, p. 166, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114060.

[18] S. Kunnakorntammanop, N. Thepwuttisathaphon, and S. Thaicharoen, "An Experience Report on

Building a Big Data Analytics Framework Using Cloudera CDH and RapidMiner Radoop with a Cluster of Commodity Computers," *Springer-Soft Comput. Data Sci.*, vol. 1100, no. August, pp. 208–222, 2019, doi: DOI: 10.1007/978-981-15-0399-3_17.

[19] Caro Fuchs, S. Spolaor, M. S. Nobile, and U. Kaymak, "pyFUME: a Python Package for Fuzzy Model Estimation," *IEEE Int. Conf. Fuzzy Syst.*, vol. 19911264, no. August, pp. 43–47, 2020, doi: 10.1109/FUZZ48607.2020.9177565.

[20] B. Suharjo, "Application of K-Means Cluster and Spatial Statistics using Python to Analyze the Indicators of Indonesia Information Technology," *Digit. Zo. J. Teknol. Inf. Komun.*, vol. 12, no. 1, pp. 11–18, 2021.

[21] J.-J. Beunza *et al.*, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," *J. Biomed. Inform.*, vol. 97, no. 103257, pp. 1–6, 2019, doi: https://doi.org/10.1016/j.jbi.2019.103257.

[22] Rienna Oktarina and Junita, "Determine the clustering of cities in Indonesia for disaster management using K-Means by excel and RapidMiner," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 794, no. 012094, pp. 1–10, 2020, doi: DOI 10.1088/1755-1315/794/1/012094.

[23] R. W. Sari, H. Dedy, I. dan Gunawan, and W. P. Agus, "Aplikasi RapidMiner dalam Pengelompokkan Kasus Penyakit AIDS berdasarkan Provinsi dengan Data Mining K-means Clustering," *Reg. Dev. Ind. Heal. Sci. Technol. Art Life*, pp. 59–69, 2018.

[24] Y. Religia, "Metode Manhattan, Euclidean Dan Chebyshev Pada Algoritma K-Means Untuk Pengelompokan Status Desa," Universitas Dian Nuswantoro Semarang, Semarang, 2016.

[25] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.

[26] G. Bonaccorso, *Machine Learning Algorithm*. Birmingham: Packt Publishing Ltd, 2017.

[27] A. Chisholm, *Exploring Data with RapidMiner*. Birmingham: Packt Publishing Ltd, 2013.

[28] E. Mohamed and T. Celik, "Early detection of failures from vehicle equipment data using K-means clustering design," *Comput. Electr. Eng.*, vol. 103, no. 108351, pp. 1–10, 2022, doi: https://doi.org/10.1016/j.compeleceng.2022.108351.

[29] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," *Knowledge-Based Syst.*, vol. 203, no. 106167, pp. 1–10, 2020, doi: https://doi.org/10.1016/j.knosys.2020.106167.

[30] H. Cuesta and S. Kumar, *Practical Data Analysis Second Edition*. Birmingham: Packt Publishing Ltd, 2016.

[31] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics*. Bantul: Penebar Media Pustaka, 2018.

[32] S. Ozdemir, *Three Principles of Data Science*. Birmingham: Packt Publishing Ltd, 2017.