Engineering and Technology Journal e-ISSN: 2456-3358

Volume 05 Issue 11 November-2020, Page No.-759-764

DOI:10.47191/etj/v5i11.02, I.F. – 6.127

Fixed Voters Clustering to Determine the Level of Beginner Voters using Data Mining Techniques

Ummul Hairah[1], Edy Budiman[2]

[1,2]Department of Informatics, Mulawarman University, Samarinda, East-Kalimantan, Indonesia

ABSTRACT: Data mining clustering technique is used to classify the level of beginner voters using the K-Means method. Fixed voter clusters are used for decision making for stakeholders regarding information on beginner voters in each district and sub-district. The error calculation method is used to measure the level of error value for each distance calculation used, the distance calculation method used ie Euclidean, Manhattan, and Minkowski Distance with the Means Square Error (MSE) approach to measure the level of the error value. The calculation results show that the lowest error occurs in the calculation of the Minkowski Distance model 3 cluster, where the error rate is 11%, while the highest error rate occurs in the calculation of the Manhattan Distance model 5 cluster, which is 38%.

KEYWORDS: voters, cluster, k-means, euclidean, manhattan, minkowski

I. INTRODUCTION

Political participation is an important aspect in a democratic country structure as well as a characteristic of political modernization[1]. In countries where the modernization process has generally been going well, the level of citizen participation usually increases. Political modernization can be related to both political and government aspects[2].

An election to directly elect a government leader in a country or a region is a very crucial moment for a country, for this it must be accompanied by a high level of people's political participation[3]. In this case the desired participation is not just using the right to vote, but most importantly how the right to vote can be implemented with rational choices in order to provide the best leader for the country.

Students or adults are a large enough community and are counted enough as the basis of votes in every election. Community of students or adults who are joining for the first time elections are called Beginner voters or they will vote for the first time because their age has just entered the voting age, are those who are Indonesian citizens who are 17 years of age and or more or have or have been married who have the right to vote, and are not previously voters because of the provisions of the Election Law.

The number of beginner voters in Indonesia cannot be underestimated. the Beginner voters who participate in each election are around 36 million people or the equivalent of 19-20% of the total number of voters. This number is very significant because it is equivalent to 20% of the total national voting power. With 20% of the votes it will allow a new party to pass the electoral threshold at the election. With a figure of 20% it can also run for President and Vice President. Because the requirements to nominate as President and Vice President only get five percent of the total votes, and with 20% of the vote, it could become the third largest political force in Indonesia.

This study aims to determine how much the level of beginner voters based on districts and sub-districts is divided into 3 categories, ie small, medium and large. For this reason, a data mining method approach is used to cluster these categories using the data mining techniques. The number of datasets used was 549,626 fixed voters data with 9 attributes; Date of Birth, Gender, Beginner, districts and sub-districts, collected from Election Supervisory Committee in one of the capitals of the Indonesian Province.

This type of quantitative research, where data is collected, is recorded, compiled, and presented in tabular form[4],[5], which is then measured in statistical values to prove the truth of the theory. This research was conducted to classify the fixed voter list data to determine how many beginner voters in a sub-districts by clustering beginner voters between 17 and 20 years of age.

II. METHODOLOGY

An overview of the beginner voter data mining cluster analysis process is presented in Figure 1.

A. Data Collection

This study uses secondary data collected from the results of the observation of the number of datasets of 549,626 fixed voters and their attributes from the Election Supervisory Committee in one of the capitals of the Indonesian Province.

B. CRISP-DM Analysis Model

One such process that has become a standard and popular, the 'Cross-Industry Standard Process for Data Mining' - or CRISP-DM - was proposed in the mid-1990s by a consortium of European companies to become a non-proprietary methodology standard for DM[6].

Figure 1, describes the data mining development life cycle of the process proposed in this study, which is a six sequential stage starting with a good understanding of the business and the need for a DM project and ending with a 'deployment' of a solution that satisfies specific business needs[7],[8].

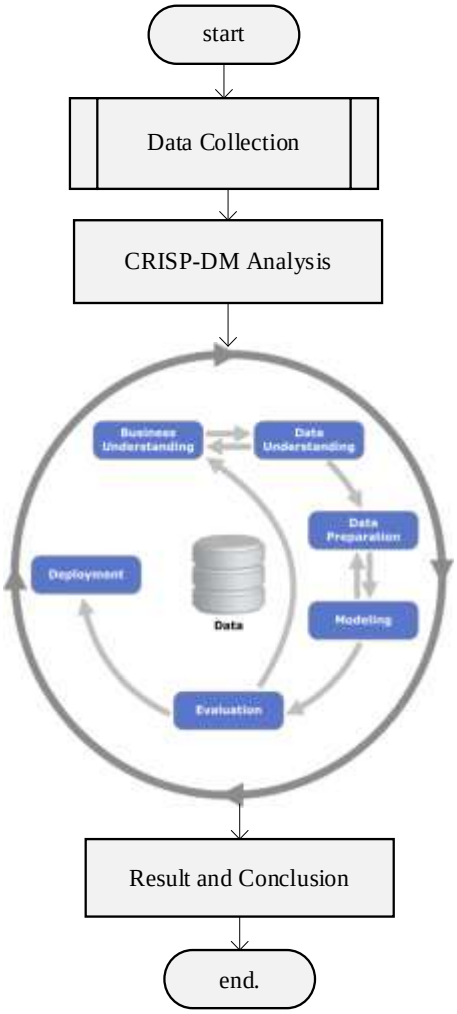

Figure 1: The beginner voter data mining cluster analysis process

1). **Business Understanding**: this study applies data mining techniques using the K-Means method to obtain the best clustering in determining the number of beginner voters and disabilities based on region (district and sub-district) from the fixed voter dataset.

2). **Data Understanding**: In this stage, data on the Fixed voters list from the Election Supervisory Committee in one of the Provincial Capitals in Indonesia are obtained. The following attributes are contained in the data obtained and presented in **Table 1**.

Table 1. Data Description of Fixed Voter Attribute List

Attribute	Descriptions
Family Card Number	Fixed voter family card number
Family ID	Fixed voter family ID
Name	Name of fixed voter
Place of birth	Place of birth of the fixed Voter
Date of birth	Date of birth of fixed voters
Marital status	Voter marital status
Gender	Gender of fixed voters
Address	Voter address
Disability	Natural voter status
Districts	Districts voter
Subdistricts	Subdistricts voter

3). **Data Preparation**: This stage includes all activities to build the final dataset (data to be processed at the modeling stage) from raw data. This stage was repeated several times which included selecting tables, records, and data attributes, including the process of cleaning and transforming data to be used as input in the modeling stage. After the basic process is carried out through the data transformation stage, the attributes that will be used to determine the number of beginner voters and disabilities are obtained in **Table 2**.

Table 2. Data attribute used

Attribute	Scala	Descriptions
Date of birth	Nominal	Date of birth of fixed voters
Address	Nominal	Voter address
Districts	Nominal	Districts voter
Subdistricts	Nominal	Subdistricts voter

4). **Modeling**: The algorithm used in this study is the K-Means algorithm for clustering in determining the number of beginner voters and disabilities. The process of the K-Means algorithm refers to B Boehmke et al[9].

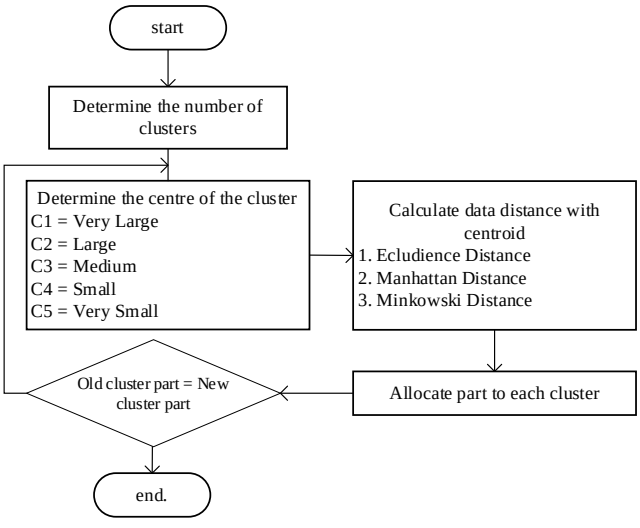

Figure 2: K-Means Algorithm Flowchart

Calculating the average centroid of the data in each cluster with the initial centroid (24^{th}, 26^{th}, 58^{th} data), calculating the

distance of each centroid to the cluster using Euclidean Distance[10], Manhattan Distance, Minkowski Distance, the formula refers to [11], [12] the equation:

Euclidean $$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Manhattan $$d = \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

Minkowski $$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (3)$$

d is the distance between x and y, x is the cluster centre data, y is data in attributes and p is power.

5). **Evaluation**: This stage tests the initial data into variable data, measuring the error rate of the model using the MSE method[13], [14].

$$E = \sum_{i=1}^{n} \frac{(x_t - s_t)^2}{n} \quad (4)$$

III. RESULT AND DISCUSSION

This study used a dataset of 549,626 data, analyzed using the K-Means method. The purpose of the analysis is to determine the level of beginner voters based on districts and sub-districts.

A. Result: Cleaning and Transformation Data

The cleaning and transformation process obtained 59 datasets of data. Then after all the processes proceed to the data normalization stage so that the vulnerability between each data is not too far away, using the MIN-MAX method for data normalization[15], the results are seen in Table 3.

Table 3. Normalization Data

Item	Districts	Sub-districts	Beginner
1	A	1a	1.08
2		2a	1.15
3		3a	0.92
4		4a	1.04
5		5a	0.84
⋮	⋮	⋮	⋮
54	J	1j	1.07
55		2j	0.96
56		3j	0.99
57		4j	1.19
58		5j	1.80
59		6j	1.17

Data allocation into the cluster center is randomly selected data to be used as the cluster center by calculating the average, minimum, and maximum value of the beginner and attributes[16], the data that be made into the cluster center are the data 24th (0.8), data 26th (1.03), and data 58th (1.80).

B. Result: Calculating the Distances

The results of calculating the distance of each centroid to the cluster using 3 distance calculations, ie Euclidean, Manhattan and Minkowski Distance refers to Archana Singh et al[17]. The calculation results of the distance of each centroid distance are presented in the Table 4.

Table 4. The distance of each centroid results

Cluster	C1	C2	C3
C2		**1.14**	
C2		**0.92**	
C2			**0.84**
C3	⋮	⋮	⋮
C1	**1.8**		
C2		**1.17**	

The data in Euclidean distance is data that has normalized using the Min-Max method, then the calculation of the Euclideans distance is based on equation (1). Example calculation for Centroid 1 for 1st data:

$$d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$
$$= \sqrt{(1{,}08 - 1{,}80)^2 + (0{,}85 - 1{,}80)^2}$$
$$= 1{,}1876$$

Example for Manhattan distance equation (2):

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

$$= |1{,}08\text{-}1{,}80| + |0{,}85\text{-}1{,}80| = 1{,}67$$

Example for Minkowski distance equation(3):

$$d = \sqrt{|x1 - x2\ h\ |h + |y1 - y2|h}$$
$$= 3\sqrt{|1{,}08 - 1{,}80|3 - |0{,}85 - 1{,}80|3}$$
$$= 1{,}0675$$

K-Means calculation using the help of RapidMiner and sklearn (pandas profiling) as shown in Figure 3 is the K-Means algorithm modeling design and the results of the K-Means modeling[18].

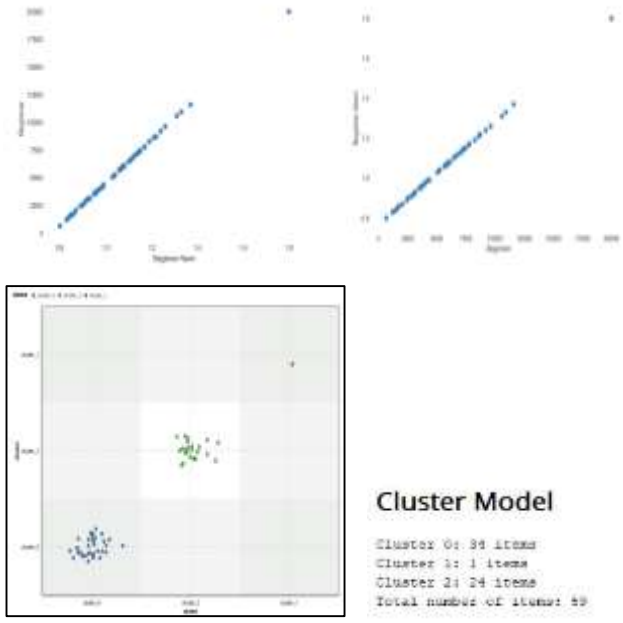

Figure 3: Scatter graph K-Means modeling results

Figure 3 shows the results of forming a model with 59 data and 2 label attributes (district and sub-district). K-Means

modeling results show that the data allocated to cluster 1 is 34 data, cluster 2 is 1 item, cluster 3 is 24 data.

C. Result: Evaluation and Error Value

In evaluation, will conduct a modelling test using Rapidminer software, the cluster distance performance test and calculating the amount of error value (Means Square Error). Model formation and testing the cluster model using Rapidminer are shown in Figure 4.

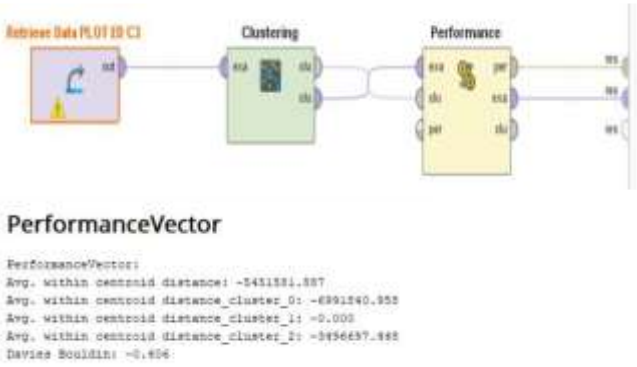

Figure 4: Design Performance test with the ED method

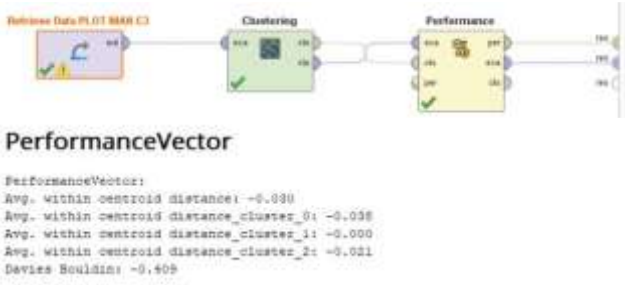

Figure 5: Design Performance test with the MAN method

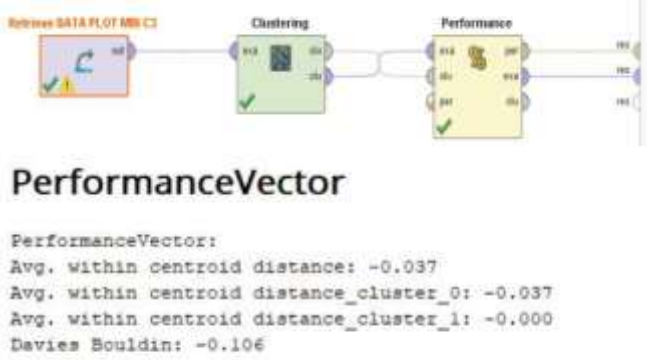

Figure 6: Design Performance test with the MIN method

To find out the performance of the modelling used in this research, the error value is calculated with various numbers of clusters. The author tested the model using the MSE (Means Squared Error) method with the number of clusters consisting of 3 clusters, 4 clusters, and 5 clusters. The data used to calculate the error value is the result of calculations between cluster distances using 3 distance calculation methods, namely Euclidean, Manhattan, Minkowski. The results of the calculation of the distance between cluster centres are presented in Table 5.

Table 5. Jarak antar pusat cluster ED, MAN and MIN

Distance	istance between centres		
Euclidean ED3	C1	C2	1.0691
	C1	C3	1.3056
	C2	C3	0.2570
Manhattan	C1	C2	1.4916
	C1	C3	1.8335
	C2	C3	0.3419
Minkowski	C1	C2	0.95886
	C1	C3	1.1637
	C2	C3	0.2397

From the table of distances between cluster centers of K-Means modeling, then the error value is calculated with equation (4), the results of which are presented in Table 6. The calculation of the 3 methods of calculating the distance above, obtained the error value of each method of calculating the distance.

Table 6. MSE Value Result

Distance	Cluster	MSE
Euclidean	C3	21%
	C4	18%
	C5	18%
Manhattan	C3	29%
	C4	37%
	C5	38%
Minkowski	C3	11%
	C4	14%
	C5	16%

With the 3 cluster model, the error value obtained from each method is 21% for Euclideans, 29% for Manhattan, and 11% for Minkowski. Furthermore, the 4 cluster model for the error value obtained for each method is 18% for Euclideans, 37% for Manhattan, and 14% for Minkowski. Finally, the 5 cluster model for each distance calculation method, namely 18% for Euclideans, 38% for Manhattan, and 16% for Minkowski, so based on the results of the clustering analysis that the author conducted, the number of sub-districts that have the highest number of first-time voters is only 1 sub-district, for moderate beginner voters is 34 districts and for relatively small number of beginner voters, namely 24 sub-districts.

D. Discussion

Based on the calculation results, starting from the initial normalized data using the min-max method to performing cluster calculations using 3 distance calculation methods, namely Euclideans Distance, Manhattan Distance, and Minkowski Distance, the cluster results are slightly different from each distance calculation. where the difference in clusters is what affects the magnitude of an error rate in each method of calculating the distance (Distance), where the error calculation is done using the MSE (Means Square Error) method, which is when calculating the distance between the

cluster centers using 2 forecasts, here the author uses 2 forecasts so that the cluster center distance can be calculated, considering the cluster calculation used by the author is a 3 cluster model, then which is chosen, namely 2 forecast. By looking at the error rate in each distance calculation, it can be concluded that the lowest error occurs in the Minkowski Distance model 3 cluster calculation, where the error rate is 11%, while the highest error rate occurs in the calculation of the Manhattan Distance model 5 cluster, which is 38%, so the best cluster calculations happen at Minkowski Distance by 11%, because the smaller the error rate, the better the calculation. As for Modeling in the District, there is no error in each of the distance calculation methods so that it can be ascertained that any method used to calculate the above sub-districts is very effective because each method has no difference in error values.

CONCLUSIONS

Based on the research results described above, it can the following conclusions are drawn:

The author implements the K-Means Clutsering method with a model of 3 clusters, 4 clusters, and 5 clusters to calculate sub-district modeling while for the calculation of sub-district modeling only uses a 3 cluster model with the data used, namely the Fixed Voters List (DPT), here the author wants looking for the level of beginner voters in each sub-district and sub-district, the distance calculation used is Euclideans Distance, Manhattan Distance, Minkowski Distance, and to calculate the error value of each method, the writer uses the Means Square Error (MSE) method.

Testing the K-Means 3 cluster model with the MSE method on 3 distance calculation methods, namely Euclideans, Manhattan, and Minkowski, the results for Euclideans are 21%, for Manhattan it is 29%, and for Minkowski it is 11%. Likewise with testing the K-Means 4 cluster model where the results of the calculation error are 18% for Euclideans, 37% for Manhattan, and 14% for Minkowski. After the results of the calculation of the cluster 3 and cluster 4 models are obtained, then proceed to the calculation of the 5 cluster model where the percentage magnitude is 18% for Euclideans Distance, 38% for Manhattan Distance, and for Minkowski Distance by 16%.

Looking at the error rate in each distance calculation above, it can be concluded that the lowest error occurs in the calculation of the Minkowski Distance model 3 cluster, where the error rate is 11%, while for the high error rate occurs in the calculation of Manhattan Distance model 5 cluster, which is 38%, so the best cluster calculation occurs at Minkowski Distance by 11%, because the smaller the error rate, the better the calculation. For modeling the clustering calculation for the district there is no difference in the error value so that it can be ascertained that whichever method is used will not affect the cluster.

The work of analysis refers to the results of this study, in the future it is necessary to reconsider in determining the variables that have a significant effect in determining the level of beginner voters in each sub-district and exploration or optimization of other cluster methods.

ACKNOWLEDGEMENT

The author's team would like to thank the Election Supervisory Committee, Institution and other contributors for the citizen's data and attributes (a dataset of permanent voters and novice voters), and the Department of Informatics, Faculty of Engineering, Mulawarman University for their financial assistance support.

REFERENCES

1. P. Parvin, "Democracy Without Participation: A New Politics for a Disengaged Era," *Res Publica*, 2018, doi: 10.1007/s11158-017-9382-1.
2. J. M. Stonecash, "Democracy for Realists: Why Elections do not Produce Responsive Government," *The Forum*, 2017, doi: 10.1515/for-2017-0024.
3. M. S. Alelaimat, "Factors affecting political participation (Jordanian universities students' voting: field study 2017-2018)," *Review of Economics and Political Science*, vol. ahead-of-p, no. ahead-of-print, 2019, doi: 10.1108/reps-05-2019-0072.
4. M. B. H. Ibrahim, M. T. Jufri, S. N. Alam, Zakaria, M. A. Akbar, and E. Budiman, "Statistical Analysis of Performance Goals Effect to Lecturer Work Achievement in Higher Education," 2018, doi: 10.1109/EIConCIT.2018.8878571.
5. D. E. McNabb and D. E. McNabb, "Fundamentals of Quantitative Research," in *Research Methods for Public Administration and Nonprofit Management*, 2018.
6. R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000.
7. M. Hancock, "The Data Mining Process," in *Practical Data Mining*, 2011.
8. F. Nurul Auliah, A. Lawi, E. Budiman, and S. Astuti Thamrin, "Selection of Informative Genes to Classify Type 2 Diabetes Mellitus using Support Vector Machine," Institute of Electrical and Electronics Engineers Inc., Apr. 2019. doi: 10.1109/ICCED46541.2019.9161111.
9. B. Boehmke, B. Greenwell, B. Boehmke, and B. Greenwell, "K-means Clustering," in *Hands-On Machine Learning with R*, 2020.
10. E. Maria, E. Budiman, Haviluddin, and M. Taruk, "Measure distance locating nearest public facilities using Haversine and Euclidean Methods," in *Journal of Physics: Conference Series*, 2020, vol.

1450, no. 1, pp. 131–134, doi: 10.1088/1742-6596/1450/1/012080.

11. A. Deshpande, A. Louis, and A. Singh, "On Euclidean k-means clustering with α-center proximity," 2020.
12. I. Dabbura, "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks," *Towards Data Science*, 2018.
13. J. Fürnkranz *et al.*, "Mean Squared Error," in *Encyclopedia of Machine Learning*, 2011.
14. F. Aziz, A. Lawi, and E. Budiman, "Increasing Accuracy of Ensemble Logistics Regression Classifier by Estimating the Newton Raphson Parameter in Credit Scoring," 2019, doi: 10.1109/ICCED46541.2019.9161078.
15. N. Dengen, Haviluddin, L. Andriyani, M. Wati, E. Budiman, and F. Alameka, "Medicine Stock Forecasting Using Least Square Method," 2018, doi: 10.1109/EIConCIT.2018.8878563.
16. Haviluddin, N. Dengen, E. Budiman, M. Wati, and U. Hairah, "Student Academic Evaluation using Naïve Bayes Classifier Algorithm," 2018, doi: 10.1109/EIConCIT.2018.8878626.
17. A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *International Journal of Computer Applications*, 2013, doi: 10.5120/11430-6785.
18. E. Budiman, Haviluddin, N. Dengan, A. H. Kridalaksana, M. Wati, and Purnawansyah, *Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation*, vol. 488. 2018.