**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 22, No. 2, May 2021, pp. 835~841
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp835-841


# K-means method for clustering learning classes

**A D Indriyanti[1], D R Prehanto[2], T Z Vitadiar[3]**
[1,2]Department of Information System, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
[3]Department of Information System, Faculty of Information Technology, Universitas Hasyim Asy'ari, Indonesia

**Article Info**

***Article history:***

Received Mar 10, 2020
Revised Dec 4, 2020
Accepted Jan 11, 2021

***Keywords:***

Clustering
K-means
Learning classes

**ABSTRACT**

Learning class is a collection of several students in an educational institution. Every beginning of the school year the educational institution conducts a grouping class test. However, sometimes class grouping is not in accordance with the ability of students. For this reason, a system is needed to be able to see the ability of students according to the desired parameters. Determination of the weight of test scores is done using the K-means method as a grouping method. Iteration or repetition process in the K-means method is very important because the weight value is still very possible to change. Therefore, the repetition process is carried out to produce a value that does not change and is used to determine the ability level of students. The results of the class grouping test scores affect the ability of students. Application of K-means method is used in building an information system grouping student admissions in an educational institution. Acceptance of students will be grouped into 3 groups of learning classes. The results of testing the system that applies K-means method and based on data on the admission of prospective students from educational institutions have very high accuracy with an error rate of 0.074.



***Corresponding Author:***

A D Indriyanti
Department Information System
Faculty of Engineering, Universitas Negeri Surabaya
Building E1, Jl. Ketintang, Surabaya City, East Java 60231, Indonesia
Email: ariesdwi@unesa.ac.id

## 1. INTRODUCTION

Class grouping is important and implemented by educational institutions when beginning learning, this grouping aims to group students into the variables that have been determined so that they can gather students into the right group [1]. The purpose of this class grouping is to facilitate the instructor in adjusting the material presented by paying attention to the categories of student groups that have been determined, so that students have the readiness or ability to accept lessons that will be given by the instructor [2].

From several articles studied mention that class grouping is done through questionnaires and interviews with teachers in providing learning over a period of 5 to 11 years [3], this is of course the results are less precise because it should be questionnaire and interview about grouping learning classes given to students who will start the process learning is not to the instructor [4], and the determination of class groups takes a long time because they have to wait for evaluation for several years [5]. In another article also mentioned the process of class grouping is done with the questionnaire choice "a" and "b" [6] and only to measure the student's personality value [7], this is less precise if what the institution wants is the potential value or ability of students to determine the concentration of certain lessons [8].

Based on the background and problems above, the writer has a solution by creating a grouping system of learning classes using the K-means algorithm method [9]. K-means method is a very appropriate method in grouping students because of the characteristics of K-means that do the classification with clear

and precise variables[10], this is in accordance with the types of lessons that can be called by these variables [11]. K-means was also proven to be able to analyze the air-pollution constituents PM10 and BC over a period of 1 year by grouping several day profiles so as to produce characteristic models that explain weather conditions, seasons and daily human activities [12], another article also mentions K-means are used in grouping hybrid based learning optimization by using a 3 phase hybrid comparison method [13]. Where the two articles that contain K-Means can also be applied in the grouping of learning classes [14].

## 2. RESEARCH METHOD

### 2.1. Information system

The system can be interpreted as a collection or set of processes that communicate with each other, are connected, and are interdependent in doing certain goals. Information is the result of processing data that can be justified into a form that is beneficial to the recipient. Information systems can be described as a set of processes that process and present information in such a way that it benefits the recipient [15].

### 2.2. K-Means method

Classification methods are intended to find models or functions that explain or distinguish concepts or data classes to be able to estimate classes and objects whose labels are unknown [16]. In the process of grouping data or clustering data, there are two ways that are often used, namely clustering hierarchy and non clustering hierarchy [17]. The Hierarchy Method is a grouping of two or more objects that have the closest resemblance [18]. Then proceed to other objects and so on so that the cluster will produce a pattern 'tree' where there are clear levels (hierarchy) between objects, from the most similar to the least similar [19]. While the non-hierarchy method is initially set by the number of clusters (two, three, or others) first [20]. After the number of clusters is determined, then the cluster process is carried out without following a hierarchical process or randomly [21].

K-means is a data method with a non-hierarchical method that aims to partition data into one or more clusters [22]. The purpose of this data clustering method is to minimize the objective function set in the clustering process [23]. Following the stages of data clustering using the K-Means method is commonly done with the basic algorithm as follows [24]:

a) There are 3 classes which will then be determined by K-means namely classes a, b and c, where one of these classes becomes the selected cluster point. Then randomly assign pieces of data k as a starting point cluster, in this step random determination is still based on class data that has been obtained previously.
b) Euclidian Distance is used in calculating the distance between data and cluster points. The euclidean formula can be calculated by the following equation [25]:

$$\mathrm{d}\,(x, ci = \sqrt{(x1i - c1i)^2 + (x2i - c2i)^2 + \cdots + (xmi - ci)^2} \quad (1)$$

Where, d (x,c) is distance of data x to cluster center c, xmi is Data i on attribute data k, $ci$ is the center point to j on attribute k.

c) The data is placed in the nearest cluster then perform calculations from center of the cluster. If data has been assigned to closest cluster, the next process is to determine the cluster center. How to determine center of a new cluster is to find the average value in the previous centroid using the formula(2):

$$C_k = 1/ n_k \sum d_k \quad (2)$$

where, Ck is new centroid, nk is amount of data that is a member in the cluster, and $d_k$ is data in the cluster k

d) The value of the centroid as a reference in determining cluster center and data placement, if centroid value changes continuously, process of determining cluster center will be repeated [26].

Type of data that will be used in clustering is quantitative. Exam results from grade level inclusions included quantitative data types because they are numeric and can be counted.

## 3. RESULTS AND ANALYSIS

This information system framework focuses on data processing, where at the input stage there are collecting data on the results of test scores based on variables 1, 2, 3, 4, 5 as input to the processing system. Then the system will do the clustering process using the K-Means method to determine learning classes A, B, C as the output of the system in the Figure 1.

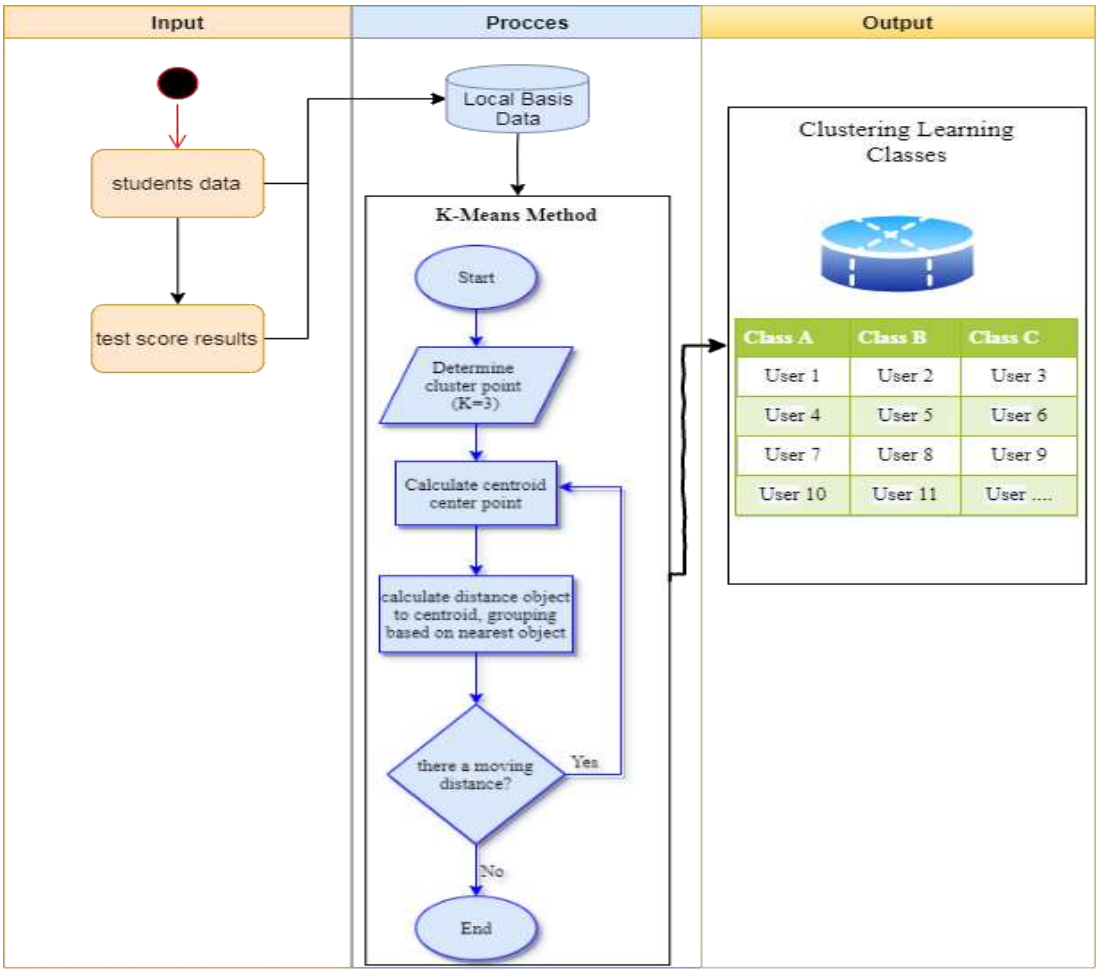


Figure 1. Framework system

The initial stage of the calculation using the K-means method is to determine the centroid point. There are 3 centroid points taken because there are 3 learning classes in this system. The class is class A, B, and C. The selected data is data 1 as centroid 1, data 4 as centroid 2, and data 10 as centroid point 3. The following are examples of data determined by the class using K-means method presented in Table 1.

The initial stage of the calculation using the k-means method is to determine the centroid point. There are 3 centroid points taken because there are 3 classes in this system. The class is class A, B, and C. The selected data is data 1 as centroid 1, data 4 as centroid 2, and data 10 as centroid 3. After determining the centroid data point in the process of the k-means method is iterating by finding the closest value from the predetermined centroid point. The following is the calculation used in the search for the closest distance to the first data.

Table 1. Data training

| No | Name | Var 1 | Var 2 | Var 3 | Var 4 | Var 5 |
|---|---|---|---|---|---|---|
| 1 | Habib | 45 | 50 | 50 | 55 | 70 |
| 2 | Aldi | 92 | 87 | 80 | 80 | 85 |
| 3 | Bima | 50 | 45 | 50 | 55 | 65 |
| 4 | Hasan | 75 | 70 | 60 | 70 | 72 |
| 5 | Elfira | 75 | 70 | 78 | 80 | 78 |
| 6 | Allisa | 80 | 80 | 75 | 85 | 85 |
| 7 | Nancy | 78 | 75 | 70 | 80 | 70 |
| 8 | Resty | 90 | 90 | 80 | 85 | 80 |
| 9 | Wulan | 80 | 85 | 75 | 78 | 80 |
| 10 | Winda | 82 | 85 | 78 | 82 | 90 |

$$d(1,1) = \sqrt{(45-45)^2 + (50-50)^2 + (50-50)^2 + (55-55)^2 + (70-70)^2}$$
$$d(1,2) = \sqrt{(45-75)^2 + (50-70)^2 + (50-60)^2 + (55-70)^2 + (70-72)^2}$$
$$d(1,3) = \sqrt{(45-82)^2 + (50-85)^2 + (50-78)^2 + (55-82)^2 + (70-90)^2}$$

In the above calculation, the first count of the santri data is as stated in Table 2.

Table 2. Initial distance

| No | Name | Centroid Point | | | Distance Value | Class |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | | |
| 1 | Habib | 0 | 40 | 67 | 0 | A |
| 2 | Aldi | 73 | 35 | 11.7 | 11.7 | C |
| 3 | Bima | 8.7 | 40 | 69 | 8.6 | A |
| 4 | Hasan | 40 | 0 | 32.6 | 0 | B |
| 5 | Elfira | 52.7 | 21 | 20 | 20 | C |
| 6 | Allisa | 62 | 27 | 8.5 | 8.4 | C |
| 7 | Nancy | 52 | 15 | 24 | 15 | B |
| 8 | Resty | 74 | 36 | 14 | 14 | C |
| 9 | Wulan | 60.8 | 24 | 11 | 11 | C |
| 10 | Winda | 67 | 32.6 | 0 | 0 | C |

In Table 2 a distance search of 10 data. After searching for values closest next step is to check the value of the centroid point is changing or no. The following is a calculation to find a new centroid point in cluster 1.

$$C1(\text{Var } 1) = \frac{1}{2}(45+50)$$
$$C1(\text{Var } 2) = \frac{1}{2}(50+45)$$
$$C1(\text{Var } 3) = \frac{1}{2}(50+50)$$
$$C1(\text{Var } 4) = \frac{1}{2}(55+55)$$
$$C1(\text{Var } 5) = \frac{1}{2}(70+65)$$

The above calculation is a calculation for finding the new first centroid point as shown in Table 3. In Table 4 is the result of calculating the search for new centroid points. If there is a change iterate again until centroid value does not change using formula like searching for iteration 1. In Table 4 shows the iteration 2.

Table 3. New centroid

| No | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 |
|---|---|---|---|---|---|
| 1 | 47.5 | 47.5 | 50 | 55 | 67,5 |
| 2 | 76,5 | 72,5 | 65 | 75 | 71 |
| 3 | 83,16667 | 82,83333 | 77,66667 | 81,66667 | 83 |

Table 4. Iteration 2

| No | Name | Centroid Point | | | Distance Value | Class |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | | |
| 1 | Habib | 4 | 46 | 64.6 | 4.3 | A |
| 2 | Aldi | 73.3 | 30 | 10 | 10.3 | C |
| 3 | Bima | 4.3 | 46 | 65.8 | 4.3 | A |
| 4 | Hasan | 40 | 7.7 | 28 | 7.7 | B |
| 5 | Elfira | 53 | 16 | 16 | 16 | B |
| 6 | Allisa | 63 | 21.5 | 6.3 | 6.3 | C |
| 7 | Nancy | 52.1 | 7.7 | 18 | 7.7 | B |
| 8 | Resty | 74.6 | 30 | 11 | 11 | C |
| 9 | Wulan | 61 | 19 | 6.6 | 6.6 | C |
| 10 | Winda | 68 | 27.6 | 7.4 | 7.4 | C |

In Table 5 is the search distance from 10 data. After searching for the closest value next step is to check the value of centroid point has changed or not. The following is a calculation to find a new centroid point in cluster 1. In Table 6 the centroid point does not change so that the iteration process is stopped, so that calculation process using the k-means method produces as in Table 7. Final result as shown in Table 8.

Table 5. New centroid 2

| No | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 |
|---|---|---|---|---|---|
| 1 | 47.5 | 47.5 | 50 | 55 | 67.5 |
| 2 | 76 | 71.66667 | 69.33333 | 76.66667 | 73.33333 |
| 3 | 84.8 | 85.4 | 77.6 | 82 | 84 |

Table 6. Iteration 3

| No | Name | Centroid Point | | | Distance Value | Class |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | | |
| 1 | Habib | 4.3 | 47.8 | 67.2 | 4.3 | A |
| 2 | Aldi | 73 | 27.4 | 8.1 | 8.1 | C |
| 3 | Bima | 4.3 | 48 | 68.5 | 4.3 | A |
| 4 | Hasan | 40 | 11.7 | 30.5 | 11.7 | B |
| 5 | Elfira | 52. | 10.5 | 19.3 | 10.6 | B |
| 6 | Allisa | 62.8 | 18 | 8.3 | 8.3 | C |
| 7 | Nancy | 52 | 6.1 | 20.3 | 20.3 | B |
| 8 | Resty | 74.6 | 27.5 | 8.9 | 8.9 | C |
| 9 | Wulan | 61 | 19 | 6.6 | 6.6 | C |
| 10 | Winda | 68 | 27.6 | 7.4 | 7.4 | C |

Table 7. Centroid tetap

| No | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 |
|---|---|---|---|---|---|
| 1 | 47.5 | 47.5 | 50 | 55 | 67.5 |
| 2 | 76 | 71.7 | 69.3 | 76.7 | 73.3 |
| 3 | 84.8 | 85.4 | 77.6 | 82 | 84 |

Table 8 Final result

| No | Name | Class |
|---|---|---|
| 1 | Habib | A |
| 2 | Aldi | C |
| 3 | Bima | A |
| 4 | Hasan | B |
| 5 | Elfira | B |
| 6 | Allisa | C |
| 7 | Nancy | B |
| 8 | Resty | C |
| 9 | Wulan | C |
| 10 | Winda | C |

The accuracy/error calculation of the K-means method system that the author has made with manual calculations is used the MSE formula. Based on the formula in this study, 54 actual data (At) were obtained. These data are data obtained from educational institutions. While the system data (St) of 52. This data was obtained from the calculation of the system that the author made. In order to obtain the results of MSE system data and actual data as follows:

$$MSE = \frac{(54-52)^2}{54} = 0{,}074074$$

After calculating the accuracy of the system, it turns out there is an error 0.074074. This error is considered small because it is close to 0. This shows that the grouping done by the system has high accuracy so that it can be applied to the educational institution system.

## 4. CONCLUSION

Based on the system that has been created and tested conclusions can be drawn that the system which is made by the K-means method can be used to determine the class of learning with based on the results of the grade level exam scores. Values data in the form of determined subjects are then calculated using the k-means method to determine class. As well as the system which is created using the PHP programming language and data is taken from the MySQL database. System which is has been made to have a very low error rate of 0.074074 so that it can perform grouping well.

## REFERENCES


[1] G S Permadi, A Kusworo, and R Gernowo, "Application Mail Tracking Using RSA Algorithm As Security Data and HOT-Fit a Model for Evaluation System," *31 in E3S Web of Conferences*, Semarang, 2018, doi: 10.1051/e3sconf/20183111007.

[2] Y Chen, *et al.,* "Coherent Clustering Method Based on Weighted Clustering of Multi-Indicator Panel Data," *IEEE Access* vol. 7, pp. 43462-43472, 2019, doi: 10.1109/ACCESS.2019.2907270.

[3] S Kumar, M Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 240-247, 2019, doi: 10.26599/BDMA.2018.9020037.

[4] A D Indriyanti, D R Prehanto, G S Permadi, C. Mashuri, T.Z. Vitadiar, "Using Fuzzy Time Series (FTS) and Linear Programming for Production Planning and Planting Pattern Scheduling Red Onion" *in E3S Web of Conferences, Semarang*, vol. 125, p. 23007, 2019, doi: 10.1051/e3sconf/201912523007.

[5] Z Wang, Z Yu, C. L. P Chen, J You, T Gu, H-S Wong, J Zhang, "Clustering by Local Gravitation," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1383-1396, 2018, doi: 10.1109/TCYB.2017.2695218.

[6] T Liu, H Li, X Zhao, "Clustering by Search in Descending Order and Automatic Find of Density Peaks," *IEEE Access*, vol. 7, pp. 133772-133780, 2019, doi: 10.1109/ACCESS.2019.2939437.

[7] W Huang, A Ribeiro, "Hierarchical Clustering Given Confidence Intervals of Metric Distances," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2600-2615, 2018, doi: 10.1109/TSP.2018.2813322.

[8] G S Permadi, T Z Vitadiar, T Kistofer, A H Mujianto, "The Decision Making Trial And Evaluation Laboratory (Dematel) And Analytic Network Process (ANP) For Learning Material Evaluation System," in *E3S Web of Conferences, Semarang*, vol. 125, 2019.

[9] H Jia and Y-M Cheung, "Subspace Clustering of Categorical and Numerical Data With an Unknown Number of Clusters," *IEEE Transactions On Neural Networks And Learning Systems*, vol. 29, no. 8, pp. 3308-3325, 2017, doi: 10.1109/TNNLS.2017.2728138.

[10] T Kistofer, G S Permadi, T Z Vitadiar, "Development of Digital System Learning Media Using Digital Learning System," *Advances in Social Science, Education and Humanities Research*, vol. 379, pp. 177-182, VEIC, 2019.

[11] M Banikhalaf, M A Khder, "A Simple and Robust Clustering Scheme for Large-Scale and Dynamic VANETs," *IEEE Access*, vol. 8, pp. 103565-103575, 2020, doi: 10.1109/ACCESS.2020.2999368.

[12] W Liang, Y Zhang, J Xu, D Lin, "Optimization of Basic Clustering for Ensemble Clustering: An Information-Theoretic Perspective," *IEEE Access*, vol. 7, pp. 179048-179062, 2019, doi: 10.1109/ACCESS.2019.2950159.

[13] V Albert, D Sano, H Nindito, "Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 2, pp. 141-150, 2016, doi: 10.21512/comtech.v7i2.2254.

[14] S G Lee, C Lee, "Developing an Improved Fingerprint Positioning Radio Map using the K-Means Clustering Algorithm," *International Conference on Information Networking (ICOIN)*, 2020, doi: 10.1109/ICOIN48656.2020.9016627.

[15] Y Zhou, R Xie, T Zhang, J Holguín-Veras, "Joint Distribution Center Location Problem for Restaurant Industry Based on Improved K-Means Algorithm With Penalty," *IEEE Access*, vol. 8, pp. 37746-37755, 2020, doi: 10.1109/ACCESS.2020.2975449.

[16] Z Wu, R Li, C Li, "Adaptive Speech Information Hiding Method Based on K-Means," *IEEE Access*, vol. 8, pp. 23308-23316, 2020, doi: 10.1109/ACCESS.2020.2970194.

[17] L Wang, S Ding, H Jia, "An Improvement of Spectral Clustering via Message Passing and Density Sensitive Similarity," *IEEE Access*, vol. 7, pp. 101054-101062, 2019, doi: 10.1109/ACCESS.2019.2929948.

[18] X Wang, C Shao, S Xu, S Zhang, W Xu, Y Gua, "Study on the Location of Private Clinics Based on K-Means Clustering Method and an Integrated Evaluation Model," *IEEE Access*, vol. 8, pp. 23069-23081, 2020, doi: 10.1109/ACCESS.2020.2967797.

[19] E Saboori, S Parsazad, A Sadeghi, "Improving the K-means algorithm using improved downhill simplex search," *2010 2nd International Conference on Software Technology and Engineering*, vol. 2, pp. V2-350, 2010, doi: 10.1109/ICSTE.2010.5608792.

[20] D R Prehanto, A D Indriyanti, K D Nuryana, S Soeryanto, A S Mubarok, "Use of Naïve Bayes classifier algorithm to detect customers' interests in buying internet token," *Journal of Physics: Conference Series*, vol. 1402, no. 6, p. 066069, *IOP Publishing*. 2019.

[21] J Hou, A Zhang, "Enhanced Dominant Sets Clustering by Cluster Expansion," *IEEE Access*, vol. 6, pp. 8916-8924, 2018, doi: 10.1109/ACCESS.2018.2808485.

[22] W'An Zhou, Q Zhou, "Deep Embedded Clustering With Adversarial Distribution Adaptation," *IEEE Access*, vol. 7, pp. 113801-113809, 2019, doi: 10.1109/ACCESS.2019.2935388.

[23] C Mashuri, A H Mujianto, H Sucipto, R Y Arsam, G S Permadi, "Production Time Optimization using Campbell Dudek Smith (CDS) Algorithm for Production Scheduling," *E3S Web of Conferences*, vol. 125, p. 23009. 2019.

[24] K P. Sinaga, M-S Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[25] N Pang, J Zhang, C Zhang, X Qin, "Parallel Hierarchical Subspace Clustering of Categorical Data," *IEEE Transactions on Computers*, vol. 68, no. 4, pp. 542-555, 2019, doi: 10.1109/TC.2018.2879332.

[26] G S Permadi, T Z Vitadiar, T Kistofer, "Sistem Evaluasi Bahan Pembelajaran Menggunakan Metode DEMATEL dan ANP," *JSINBIS (Jurnal Sistem Informasi Bisnis)*, vol. 9, no. 2, pp. 228-235, 2019, doi: 10.21456/vol9iss2pp228-235.

## BIOGRAPHIES OF AUTHORS

**Aries Dwi Indriyanti** received hir S.Kom degree from the Information Systems Department, STIKOM, Surabaya, Indonesia, at 1998, and an M.Kom degree from the Masters in Information Systems, Postgraduate School, Diponegoro University, Semarang, Indonesia in 2009. She is currently completing a doctoral program in information systems, Diponegoro University. She is currently a lecturer and researcher at Department of Information Systems, State University of Surabaya, Indonesia. His current research interests include statistics, forecasting, supply chain and big data.

**Dedy Rahman Prehanto** received her S.Kom degree from the Information Systems Department, STIKOM, Surabaya, Indonesia, at 2005, and an M.Kom degree from the Masters in Information Systems, Postgraduate School, Diponegoro University, Semarang, Indonesia in 2012. He is currently completing a doctoral program in information systems, Diponegoro University. He is currently a lecturer and researcher at Department of Information Systems, State University of Surabaya, Indonesia. His current research interests include big data, expert systems, and IoT.

**Tanhella Zein Vitadiar** received hir S.SI degree from the Information Systems Department, University of Jember, Indonesia, at 2014, and an M.Kom degree from the Masters in Information Systems, Postgraduate School, Diponegoro University, Semarang, Indonesia in 2017. He is currently a lecturer and researcher at Information systems department, Hashim Ashari University, Jombang, Indonesia. His current research interests include machine learning, expert systems, forecasting, and scheduling.