Clustering the Students' Behavior on the e-Learning using the Density-based Algorithm

by Wiyli Yustanti

Clustering the Students' Behavior on the e-Learning using the Density-based Algorithm

2021 International Seminar on Application for Technology of Information and Communication (iSemantic) | 978-1-6654-2804-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISEMANTIC52711.2021.9573234

Yuni Yamasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yuniyamasari@unesa.ac.id

Anita Qoiriah
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
anitaqoiriah@unesa.ac.id

Naim Rochmawati
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
naimrochmawati@unesa.ac.id

Wiyli Yustanti
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
wiyliyustanti@unesa.ac.id

Atik Wintarti
Department of Mathematics
Universitas Negeri Surabaya
Surabaya, Indonesia
atikwintarti@unesa.ac.id

Tohari Ahmad
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
tohari@if.its.ac.id

***Abstract*— The corona pandemic has changed the learning method from conventional to a more flexible one, such as through the internet. Consequently, students may have less direct interaction with teachers. This condition has made it difficult for teachers to monitor the students' behavior. This research works on this problem by focussing on the clustering of students' behavior using the DBSCAN, which is a density-based algorithm. Noises generated in this process can be considered students who do the uncommon behavior when taking the e-Learning system. Further, we evaluate the resulted clusters using the silhouette index to find their quality. The experimental result shows that the DBSCAN can differentiate clusters containing noises. By taking the silhouette index, the Manhattan distance parameter is superior to that of Euclidean.**

Keywords*—*student, behavior, data mining, e-learning, DBSCAN, silhouette

I. Introduction

Nowadays, the implementation of advanced Information and Communication Technology (ICT) in education is crucial. Primarily, it happens in this corona pandemic condition. Many processes have changed their business to the online-based business, including that in the education environment. This virtual process brings consequences, like increasing daily traffic data, which requires extracting specific data.

Data mining can be used in an educational environment to obtain important information from these data [1], for example, student data. In many cases, the exploration is done to generate information that relates to students' characteristics, for example, their performance [2][3][4][5][6], attitude [7][8], and achievement[9][10].

Some previous research focuses on the students' behavior domain to analyze the patterns of doing quiz[11], the competition-driven educational game [12][13], online learning [14], student potential [15]. This student behavior has inspired us to explore further research, specifically in the domain where the interaction between teachers and students is not intensive. Our research works on a clustering task using one of the density methods, called DBSCAN, on the students' behavior domain. Here, we propose a method to detect the uncommon students' behavior based on this density algorithm in their interaction with the e-learning system. It is needed to support teachers in monitoring online learning.

Clustering is one of the tasks in data mining, mapping a set of data points, so that similar data points are grouped. Therefore, clustering algorithms search similarities or dissimilarities among data points. Clustering is an unsupervised learning technique; therefore, there is an unlabeled instance associated with data points. The algorithm finds the underlying structure of data.

This paper is arranged in following sections. The first is the introduction, which is followed by the proposed method. The experimental result is provided in the subsequent section. Finally, we conclude the research in Section 4.

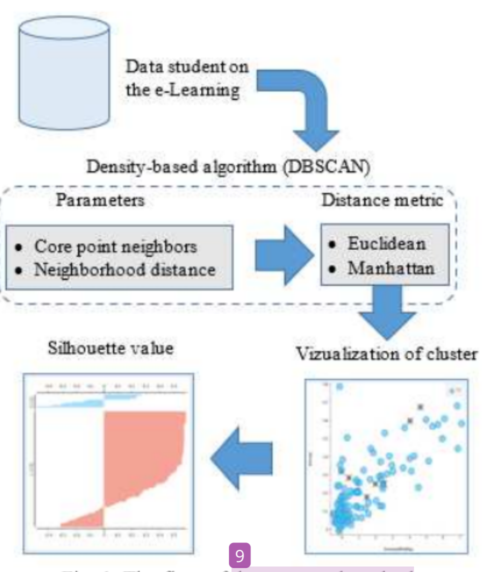

Fig. 1. The flow of the proposed method

TABLE I. STUDENT DATA

No.	Feature	Data type	Description
1	DownloadIndFar	Numeric	The number of students' activity relating to downloading the material of faraday's law of induction
2	DownloadMedMag	Numeric	The number of students' activity relating to downloading the material of magnet field
3	ForumIndFar	Numeric	The number of students' activity when they join the forum relating to the material of Faraday's law of induction
4	ForumMedMag	Numeric	The number of students' activity when they join the forum relating to the material of the magnet field
5	FailLogin	Numeric	The number of students' activity relating to success or failure in the login process
6	LearnVideo1	Numeric	The number of students' activity relating to learning material using video1
7	LearnVideo2	Numeric	The number of students' activity relating to learning material using video2
8	LearnVideo3	Numeric	The number of students' activity relating to learning material using video3
9	LearnVideo4	Numeric	The number of students' activity relating to learning material using video4
10	LearnVideo5	Numeric	The number of students' activity relating to learning material using video5
11	LearnVideo6	Numeric	The number of students' activity relating to learning material using video6
12	LearnVideo7	Numeric	The number of students' activity relating to learning material using video7
13	LearnVideo8	Numeric	The number of students' activity relating to learning material using video8
14	LearnVideo9	Numeric	The number of students' activity relating to learning material using video9
15	LearnVideo10	Numeric	The number of students' activity relating to learning material using video10
16	LearnVideo11	Numeric	The number of students' activity relating to learning material using video11
17	LearnVideo12	Numeric	The number of students' activity relating to learning material using video12
18	LearnVideo13	Numeric	The number of students' activity relating to learning material using video13
19	LearnVideo14	Numeric	The number of students' activity relating to learning material using video14
20	LearnVideo15	Numeric	The number of students' activity relating to learning material using video15
21	LearnVideo16	Numeric	The number of students' activity relating to learning material using video16
22	LearnVideo16	Numeric	The number of students' activity relating to learning material using video17
23	LearnVideo18	Numeric	The number of students' activity relating to learning material using video18
24	SuccessLogin	Numeric	The number of students' activities relating to login with status success
25	Logout	Numeric	The number of students' activities relating to the process of leaving students from the e-learning system.
26	Examination	Numeric	The number of students' activity in doing the examination
27	ExercIndFar	Numeric	The number of students' activity in doing the exercises for Faraday's law induction
28	ExcerMedMag	Numeric	The number of students' activity in doing the exercises for magnetic field
29	Sum	Numeric	The amount of the student operation after logging in until they log out of e-learning;
30	Average	Numeric	The number of students' activities after logging in before they log out of e-learning is separated by the number of activities.

II. METHOD

The process of the proposed method is provided in this section, whose stages are depicted in Fig. 1. The detail steps can be explained as follows.

Step 1: Student data

Student data are obtained from students' activities in the e-Learning system. We extract them to generate the student data consisting of various features. In this research, we use all of them for the next process.

Step 2: Clustering the student data

Fig. 1. In the clustering process, the DBSCAN is applied by exploring the data density. The method is a clustering method building an area based on density-connected. Every object of an area radius has to contain at least the number of minimum data. All objects that are not included in the cluster are considered as noise. Some parameters are specified: core point neighbors, neighborhood distance, and distance parametric, either Euclidean or Manhattan. The computation of the Density-Based Spatial Clustering of Application with Noise (DBSCAN) Algorithm is as follows:

1. Initialize parameters minpts, eps.
2. Determine the starting point or p at random.
3. Repeat steps 3-5 until all points are processed.
4. Calculate eps or all density reachable point distances with respect to p.
5. If the point that meets the eps is more than minpts then the point p is the corepoint and a cluster is formed.

TABLE II. THE CLUSTERING USING DBSCAN WITH DISTANCE PARAMETER: EUCLIDEAN

No.	Core point neighbors	Neighborhood distance	Cluster 0	Cluster 1
1	5	9.8	12	115
2	6	9.82	12	115
3	7	9.92	12	115
4	8	9.99	12	115
5	9	10	12	115
6	10	10.08	12	115
7	11	10.14	12	115
8	12	10.22	12	115
9	13	10.22	12	115
10	14	10.22	12	115
11	15	10.39	12	115
12	**16**	**10.52**	**12**	**115**
13	17	11.03	12	115
14	18	11.03	12	115
15	19	11.03	12	115
16	20	11.03	12	115
17	21	11.03	12	115
18	22	11.07	12	115
19	23	11.08	12	115
20	24	11.08	12	115
21	25	11.08	12	115
22	26	11.13	12	115
23	27	11.13	12	115
24	28	11.16	12	115
25	29	11.18	12	115

TABLE III. THE COMPOSITION OF THE CLUSTER MEMBER ON DBSCAN WITH THE EUCLIDEAN DISTANCE ON THE 12TH EXPERIMENT

Cluster	Index of the students	The number of students
0	1,6,9,20,40,73,84,85,94,100,126,127	12
1	2,3,4,5,7,8,10,11,12,13,14,15,16,17,18,19,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,74,75,76,77,78,79,80,81,82,83,86,87,88,89,90,91,92,93,95,96,97,98,99,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125	115

6. If p is a border point and no point is density reachable with respect to p, then the process continues to another point.

Step 3: Visualizing the clustering result

The clustering result of students' behavior is visualized in scatterplot form to make it easy for teachers to monitor the student behavior while using an e-learning system.

Step 4: Measuring the cluster validity

The generated clusters are evaluated to find their validity. One of the metrics for measuring is the silhouette index, whose interval value is from -1 to 1. The fewer instances in a cluster, the higher the validity of the cluster.

TABLE IV. THE CLUSTERING USING DBSCAN WITH DISTANCE PARAMETER: MANHATTAN

No.	Core point neighbors	Neighborhood distance	Cluster 0	Cluster 1
1	5	22.98	12	115
2	6	23.61	12	115
3	7	23.71	12	115
4	8	24.55	10	117
5	9	24.58	11	116
6	10	24.64	11	116
7	11	25.64	12	115
8	12	24.94	12	115
9	13	24.94	12	115
10	14	25.15	12	115
11	15	25.15	12	115
12	**16**	**25.87**	**10**	**117**
13	17	25.88	11	116
13	18	25.88	12	115
14	19	25.88	12	115
15	20	25.88	12	115
16	21	25.88	12	115
17	22	25.88	12	115
18	23	25.88	12	115
19	24	25.88	12	115
20	25	25.88	12	115
21	26	26.09	12	115
22	27	26.19	12	115
23	28	26.25	11	116
24	29	26.25	11	116
25	30	26.25	12	115

TABLE V. THE COMPOSITION OF THE CLUSTER MEMBER ON DBSCAN WITH THE MANHATTAN DISTANCE ON THE 12TH EXPERIMENT

Cluster	Index of the students	The number of students
0	1,9,20,34,73,84,85,94,100,126	10
1	2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,19,21,22,23,24,25,26,27,28,29,30,31,32,33,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,74,75,76,77,78,79,80,81,82,83,86,87,88,89,90,91,92,93,95,96,97,98,99,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125	117

III. RESULT

In this section, the experimental result of the proposed method is described. It consists of three parts: data description, the implementation results of density-based algorithm-DBSCAN, and the result's visualization.

A. Data description

The mined student data are collected from the two vocational high schools. In this case, 127 students interact with an e-learning system, and their behavior is recorded. These data, which comprise 30 features having numeric data types, are explored. The detail of the collected data is provided in Table I.

B. The execution result of the DBSCAN Algorithm

Before the execution of this algorithm is done, two parameters need to be specified: core point neighbors (corepoint) and neighborhood distance. The first parameter relates to the minimum number of point neighbors (minpts) required to create a dense region. The second parameter is the radius of a neighborhood concerning some points. Here, we explore two distance measurements, namely: Euclidean and Manhattan.

This method runs 25 times on each distance parameter with various scenarios. Therefore, the experimental results are represented in two parts. The first is that with Euclidean and Manhattan. The second is that with different compositions of the cluster members on DBSCAN comprising Euclidean and Manhattan distances.

In the first experiment of DBSCAN with Euclidean distance, we specify the core point neighbors as the number ranging between 5 and 29. For the second parameter, which is neighborhood distance, we explore the distance whose value is between 9.8 and 11.18. The clustering results using DBSCAN with the same student number in each cluster on all scenarios are depicted in Table II. The student number in clusters 0 and 1 are 12 and 115, respectively. Further, we show in Table III an example of different compositions of the cluster member in each cluster. This is the 12th experimental result. Cluster 0 indicates that object included as noise instance. In the research, the noise instances mean that students have significantly different behavior from the general students, for example, students access the e-learning during the extreme time. For Cluster 1, the member of this cluster is students interacting the system reasonably.

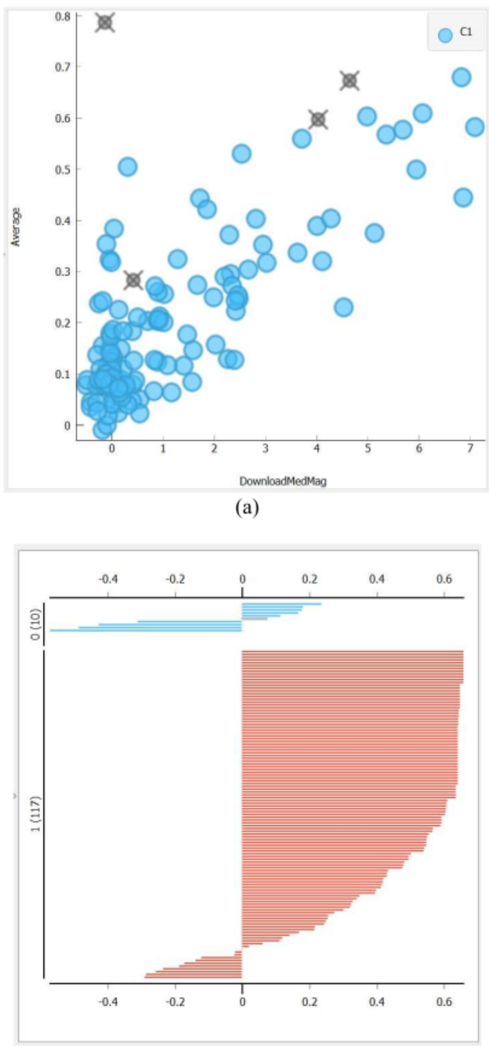

Fig. 3. The visualization is based on the Manhattan distance parameter. (a). the clustering result. (b). the cluster validity

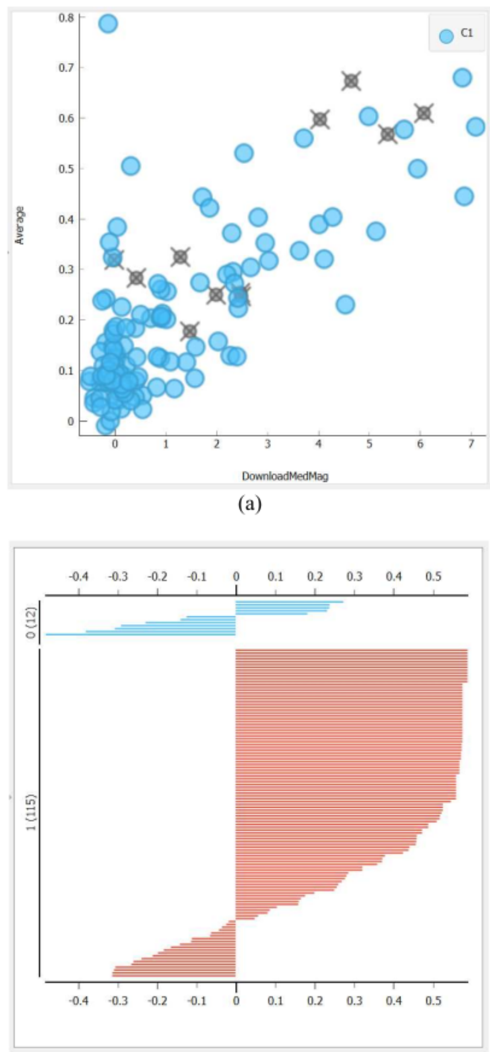

Fig. 2. The visualization is based on the Euclidean distance parameter. (a). the clustering result. (b). the cluster validity

Secondly, we determine two parameters: the core point neighbors and the neighborhood distance on the experiment of DBSCAN with Manhattan distance as follows: the first parameter with value 5-30.

In the neighborhood distance, we explore the distance whose value is from 22.5 to 26.5. The experimental results are shown in Tables IV and V. In Table IV, the result exhibits the fluctuation of student numbers in each cluster in all scenarios. Overall, the frequent student number in clusters 0 and 1 are 12 and 115, respectively. Next, in Table V, we present the cluster member's composition of the 12th experiment in each cluster.

The result presents that there are 12 students in cluster 0. which indicates that the students have uncommon behavior when they interact with the system. On contrary, there 115 students on Cluster 1 do similar behaviors or common behaviors.

C. The visualization of clustering result and cluster validity

In this sub-section, the visualization of the clustering result is presented. The visualization covers the mapping of students' behavior on the e-learning in a scatterplot graph and the silhouette value for every student on all parameters distance. This scatterplot visualizes the students' behavior mapping on the e-learning with that of the 12th experiment. Respectively, those with Euclidean and Manhattan are shown in Fig. 2(a) and Fig. 3(a). The visualization of student members of cluster 1 is illustrated by the blue circle, while that of cluster 0 is by the grey cross. The students of cluster 0 indicate that they do significantly different behaviors from most of the other student

In this graph, the downloadmedmagnet is the student behavior relating to the behavior downloading the magnetic field material from the e-learning as X-axis ; the average is the mean of all behaviors during students interact with the e-learning as Y-axis.

To validate the cluster, we compute the silhouette index for all students. Then, we visualize it in the graph as depicted respectively in Fig. 2(b) and Fig. 3(b) for the Euclidean and the Manhattan distances. This index value has ranged from -1 to 1. The less silhouette value, the higher the validity of the cluster. It means that the clustering process has better performance. The experimental result shows that the cluster validity of DBSCAN with the Manhattan distance is better than that with Euclidean distance. It is indicated by the DBSCAN-Manhattan distance having the silhouette index < 0 and less than the DBSCAN-Euclidean distance.

A clustering algorithm based on density-DBSCAN can find arbitrary shaped clusters and clusters with noise or outliers that in this paper displayed in a grey cross form. The main idea behind DBSCAN is that a point belongs to a cluster if it is close to the majority points from that cluster. In this paper, a point relates to students' behavior, so we can consider noises or outliers as students doing uncommon behaviors when interacting with the e-learning system.

IV. Conclusion

The clustering based on DBSCAN can be applied to the student's behavior on the e-Learning system to detect the students who do the uncommon behavior. This method can work optimally if the values of its parameters are specified with the appropriate values. For our research, DBSCAN reaches the optimal clustering process when its distance parameter uses Manhattan.

In the subsequent research, we would like to rank the noise level. It may be applied by using existing algorithms. Furthermore, appropriate parameters should be defined appropriately and adaptively.

References

[1] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014, doi: 10.1016/j.eswa.2013.08.042.

[2] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, p. e01250, Feb. 2019, doi: 10.1016/j.heliyon.2019.e01250.

[3] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Comput. Electr. Eng.*, vol. 66, pp. 541–556, Feb. 2018, doi: 10.1016/J.COMPELECENG.2017.03.005.

[4] L. M. Crivei, V.-S. Ionescu, and G. Czibula, "An Analysis of Supervised Learning Methods for Predicting Students' Performance in Academic Environments," *ICIC Express Lett.*, vol. 13, no. 3, pp. 181–189, 2019, doi: 10.24507/icicel.13.03.181.

[5] I. Shingari, D. Kumar, and M. Khetan, "A review of applications of data mining techniques for prediction of students' performance in higher education," *J. Stat. Manag. Syst.*, vol. 20, no. 4, pp. 713–722, Jul. 2017, doi: 10.1080/09720510.2017.1395191.

[6] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/J.COMPEDU.2017.05.007.

[7] Y. Promdee, S. Kasemvilas, N. Phangsuk, and R. Yodthasarn, "Predicting Persuasive Message for Changing Student's Attitude Using Data Mining," in *2017 International Conference on Platform Technology and Service (PlatCon)*, Feb. 2017, pp. 1–5, doi: 10.1109/PlatCon.2017.7883721.

[8] S. E. Sorour, T. Mine, K. Goda, and S. Hirokawa, "Regular Paper A Predictive Model to Evaluate Student Performance," *J. Inf. Process.*, vol. 23, pp. 192–201, 2015, doi: 10.2197/ipsjjip.23.192.

[9] J. N. Purwaningsih and Y. Suwarno, "Predicting students achievement based on motivation in vocational school using data mining approach," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, May 2016, pp. 1–5, doi: 10.1109/ICoICT.2016.7571880.

[10] N. Buniyamin, U. bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," in *2015 IEEE 7th International Conference on Engineering Education (ICEED)*, Nov. 2015, pp. 49–53, doi: 10.1109/ICEED.2015.7451491.

[11] L. Juhaňák, J. Zounek, and L. Rohlíková, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Comput. Human Behav.*, Dec. 2017, doi: 10.1016/J.CHB.2017.12.015.

[12] Z.-H. Chen, "Exploring students'behaviors in a competition-driven educational game," *Comput. Human Behav.*, vol. 35, pp. 68–74, 2014, doi: http://dx.doi.org/10.1016/j.chb.2014.02.021.

[13] M.-T. Cheng, Y.-W. Lin, and H.-C. She, "Learning through playing Virtual Age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters," *Comput. Educ.*, vol. 86, pp. 18–29, Aug. 2015, doi: 10.1016/J.COMPEDU.2015.03.007.

[14] Wang Jie, Lv Hai-yan, Cao Biao, and Zhao Yuan, "Application of educational data mining on analysis of students' online learning behavior," in *2017 2nd International Conference on Image, Vision*

and Computing (ICIVC), Jun. 2017, pp. 1011–1015, doi: 10.1109/ICIVC.2017.7984707.

[15] F. Yang and F. W. B. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Comput. Educ.*, vol. 123, pp. 97–108, Aug. 2018, doi: 10.1016/J.COMPEDU.2018.04.006.

Clustering the Students' Behavior on the e-Learning using the Density-based Algorithm

ORIGINALITY REPORT

16%	10%	13%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	2%
2	Mutiara Auliya Khadija, ST Widyawan, Ir. Lukito Edi Nugroho. "Detecting Network Intrusion by Combining DBSCAN, Principle Component Analysis and Ranker", 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019 Publication	2%
3	Submitted to University of Oklahoma Health Science Center Student Paper	2%
4	Yuni Yamasari, Anita Qoiriah, Naim Rochmawati, Wiyli Yustanti, Hapsari P. A. Tjahyaningtijas, Puput W. Rusimamto. "Combining the Unsupervised Discretization Method and the Statistical Machine Learning on the Students' Performance", 2020 Third International Conference on Vocational	2%

Education and Electrical Engineering (ICVEE), 2020
Publication

5	Galuh Ayu Paramastri, Ricardus Anggi Pramunendar. "Glaucoma Detection Through Fundus Images Using Radial Basis Function With Color and GLCM Feature Extraction", 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), 2021 Publication	1 %
6	Diksy Media Firmansyah, Tohari Ahmad. "Overlapped scheme for neighboring similarity method in video-based data hiding", 2016 IEEE Region 10 Conference (TENCON), 2016 Publication	1 %
7	"Table of Content", 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), 2021 Publication	1 %
8	Mustakim, Reza Nurul Gayatri Indah, Rice Novita, Oktaf Brillian Kharisma et al. "DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru", Journal of Physics: Conference Series, 2019 Publication	1 %

9	Yuni Yamasari, Naim Rochmawati, Ricky E. Putra, Anita Qoiriah, Asmunin, Wiyli Yustanti. "Predicting the Students Performance using Regularization-based Linear Regression", 2021 Fourth International Conference on Vocational Education and Electrical Engineering (ICVEE), 2021 Publication	1 %
10	Yuni Yamasari, Anita Qoiriah, Hapsari P. A. Tjahyaningtijas, Ricky E. Putra, Agus Prihanto, Asmunin. "Improving the Quality of the Clustering Process on Students' Performance using Feature Selection", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 2020 Publication	<1 %
11	www.coursehero.com Internet Source	<1 %
12	kinetik.umm.ac.id Internet Source	<1 %
13	www.researchgate.net Internet Source	<1 %
14	Y. Yamasari, S. M. S. Nugroho, R. Harimurti, M. H. Purnomo. "Improving the cluster validity on student's psychomotor domain using feature selection", 2018 International	<1 %

Conference on Information and Communications Technology (ICOIACT), 2018
Publication

15	www.ijimai.org Internet Source	<1 %
16	dergipark.org.tr Internet Source	<1 %
17	repository.ubaya.ac.id Internet Source	<1 %
18	repository.unair.ac.id Internet Source	<1 %
19	"Educational Data Mining", Springer Science and Business Media LLC, 2014 Publication	<1 %
20	"Recent Advances on Soft Computing and Data Mining", Springer Science and Business Media LLC, 2017 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		