Applying Data Mining for Anomalies Detection on the Academic Performance of Student

Yuni Yamasari*, Anita Qoiriah, Naim Rochmawati, Aditya Prapanca, Agus Prihanto, I Made Suartana, Ricky Eka Putra

Department of Informatics, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: yuniyamasari@unesa.ac.id

ABSTRACT

One of the crucial problems in online learning is the difficulty of monitoring student academic performance by teachers. However, research to find a solution to this problem is not much. On the other hand, the change in the educational paradigm during the corona pandemic resulted in large amounts of data stacks, especially student data. Data mining is often used to identify patterns in large datasets that can be used to train AI models. So, Data mining can be applied to student data to find knowledge or information that can be used to create a better educational environment. Therefore, our research focuses on the application of data mining to overcome the difficulties of monitoring student performance. The method used is based on density. Furthermore, our research detects an anomaly that occurs in student academic performance which produces information about students whose academic performance is different from the majority of other students. This knowledge is very important for teachers to prevent student failure in achieving academic performance.

Keywords: *Data Mining, Student, Academic Performance, Anomalies Detection.*

1. INTRODUCTION

The corona pandemic has accelerated the application of ICT in the education sector, one of which is the learning process turning into online learning. This raises a crucial problem, namely the difficulty of teachers in monitoring student performance due to reduced interaction between teachers and students. Monitoring student performance is very important so that students avoid failure in achieving their academic performance.

On the other hand, changes in online business processes in educational institutions have many consequences, such as an increase in online transaction data every day. This condition stimulates researchers to perform data mining to obtain important information from the data which is called Data Mining [1], [2]. Specifically for student data, in many cases, exploration is carried out to produce information related to student characteristics, for example, performance. [3]-[9], attitude or behaviour [10]-[12], and achievement [13], [14].

Based on the description of previous research, the domain of student academic performance is quite popular. [15], This shows that this domain is quite important for the success of students in their academic achievement and is of great interest to researchers. However, previous research on student academic performance has not focused on monitoring student academic performance with the student performance anomaly detection approach.

For this reason, our paper performs anomaly detection of student performance working on grouping tasks using one of the data density data mining methods, called DBSCAN. Here, we apply this method to student performance data. This research is needed to support teachers in monitoring student academic performance in online learning.

This method is a clustering method which is one of the tasks in data mining. This method maps a set of data points to group similar data points. Therefore, the clustering algorithm looks for similarities or dissimilarities between data points. This grouping is an unsupervised learning technique. Accordingly, our student data is not labeled where students are associated with data points.

A. Kusumastuti et al. (eds.), *5th Vocational Education International Conference (VEIC 2023)*,
Advances in Social Science, Education and Humanities Research 813,
https://doi.org/10.2991/978-2-38476-198-2_105

Finally, this paper is organized into the following sections. The first is the introduction, which is followed by the proposed materials and methods. The experimental results are presented in the next section. Finally, we conclude the research in Section 4.

2. METHOD

2.1 Student Academic Performance Data

The student academic performance data used in this study are student data when they take basic programming courses. There are 115 students involved in this study. The features of student academic performance data consist of 6, namely: presence, part, assignment, midterm Exam, final Exams, and final Score which is explained in detail in our research [16].

2.2 Proposed Method

The architecture of the proposed method is shown in Figure 1. There are several steps to be taken to detect these student academic performance anomalies, namely:

Step 1: Collecting student academic performance data which is explained in detail in the previous sub-chapter.

Step 2: Application of one of the existing methods in data mining, namely: DBSCAN. This method is based on data density. There are several parameters that we must set for this method to run. These parameters are the core point neighbour and neighbourhood distance. A core point neighbour is defined as a point that has several neighbours (including itself) within a certain predetermined radius. These neighbours are called core point neighbours or core point neighbours. The core point and its neighbours form a cluster or data group. Meanwhile, neighbourhood distance is defined as a certain distance or radius that is used to determine the neighbours of a point in the data space. In other words, neighbourhood distance is the maximum distance that is determined to determine whether a point will be considered a neighbour of another point.

Step 3: Exploration of distance measures. Our research explores 2 measures of distance, namely: Euclidean and Manhattan. The formula of Euclidean for 2 points (x_1, y_1) and (x_2, y_2)

$$E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

The formula of Manhattan M for 2 points (x_1, y_1) and (x_2, y_2)

$$M = |x_2 - x_1| + |y_2 - y_1| \tag{2}$$

Step 4: Results analysis. The test results were analyzed to find out student performance anomalies from parameter setting and distance exploration in the previous step. Many students are categorized as an anomaly by existing methods.

Step 5: Validity analysis. This step is intended to find out which detection results are the most valid from several trial scenarios that are carried out. This study uses the silhouette index size which refers to [17].

Step 6: Results visualization. This result visualizes to make it easier for users, in this case, teachers, to find out whether there are students who deviate from their academic performance because they are different from the majority of students, which in this case is referred to as an anomaly.

Step 7: Anomaly detection results. The final step of this architecture is that we present the results of the detection of academic performance anomalies. These results are to support teachers in taking action as early as possible so that the students concerned do not fail in their academics.

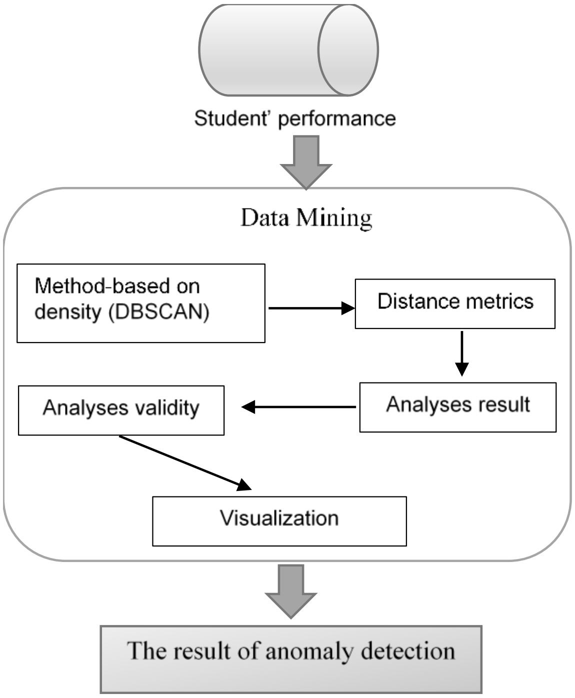


Figure 1. The Proposed Method.

3 RESULT AND DISCUSSION

The results of the execution of the proposed architecture, analysis of the results, and visualization are discussed in this chapter. As explained in the previous chapter, this study conducted trials by exploring this density-based method with 2 distance measures, namely Euclidean and Manhattan. The test results are presented in Table 1 and Table 2 for the Euclidean and Manhattan distances respectively. The experiment was carried out with Core point neighbours from 2 to 13. For neighbours' distance, the value is determined using the value in the first "valley" in the graph based on research [18] so that the results of applying the method are more optimal. The test results in Table 1 show that the most optimal range of values for the neighbourhood distance parameter is 1.34 to 1.66. Then, the number of students whose performance anomalies were detected was in the range of 7-9 students. For the size of the Manhattan distance, the value range of the neighbourhood distance parameter is 2.08 to 3.45. The number of students whose performance anomaly is detected is 8-10 students. In both distance measures, there is a tendency that the higher the value of the core point neighbours and neighbourhood distance, the more students whose performance is detected by an anomaly.

Furthermore, on the Euclidean distance measure, students with the most anomaly performance are 9 students in the scenario of core point neighbours = 13 and neighbourhood distance = 1.662. The neighbourhood distance value is determined with this value because this value is the most optimal based on the first valley recommendation as shown in Figure 2. (a). In contrast, the number of students with the least anomaly performance in the test scenario core point neighbours = 2 and neighbourhood distance = 1.34 is 7 students. Of course, the determination of the value of the neighbourhood distance is also based on the value of the first valley so that the detection results obtained are optimal.

Table 1. The experiment result using Euclidean distance.

Core Point Neighbors	Neighborhood Distance	The Student Number of the Anomaly	Index Of Student
2	1.34	7	1,11,15,17,44,56,107
3	1.344	7	1,11,15,17,44,56,107
4	1.392	7	1.11,15,17,40.54,107
5	1.548	8	1,11,15,17,40,44,56,107
6	1.561	9	1,11,15,17,40,44,54,56,107
7	1.584	9	1,11,15,17,40,44,54,56,107
8	1.594	9	1,11,15,17,40,44,54,56,107
9	1.596	9	1,11,15,17,40,44,54,56,107
10	1.63	9	1,11,15,17,40,44,54,56,107
11	1.63	9	1,11,15,17,40,44,54,56,107
12	1.657	9	1,11,15,17,40,44,54,56,107
13	1.662	9	1,11,15,17,40,44,54,56,107

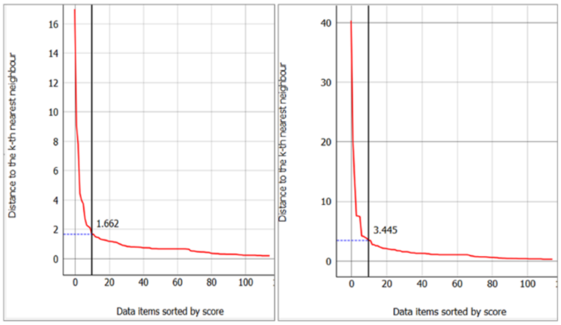
Table 2. The experiment result using the Manhattan distance.

Core Point Neighbors	Neighborhood Distance	The Student Number of the Anomaly	Index Of Student
2	2.081	8	1,11,17,40,44,54,56,107
3	2.122	9	1,11,15,17,40,44,54,56,107
4	2.731	9	1,11,15,17,40,44,54,56,107
5	2.979	9	1,11,15,17,40,44,54,56,107
6	3.113	9	1,11,15,17,40,44,54,56,107
7	3.146	9	1,11,15,17,40,44,54,56,107
8	3.313	9	1,11,15,17,40,44,54,56,107
9	3.313	9	1,11,15,17,40,44,54,56,107
10	3.35	10	1,11,15,17,40,44,49,54,56,107
11	3.37	10	1,11,15,17,40,44,49,54,56,107
12	3.396	10	1,11,15,17,40,44,49,54,56,107
13	3.445	10	1,11,15,17,40,44,49,54,56,107

Next, the visualization is carried out in the form of a scatter plot for the scenario with the most anomaly detection results. The detection results are based on the Euclidean and Manhattan distance parameters as presented in Figure 3. (a) and 3. (b) respectively. The two figures show that students who are detected as an anomaly are symbolized by a grey circle and students whose academic performance is normal are symbolized by a blue circle. For the Euclidean distance parameter, there were 9 students whose academic performance detected an anomaly. Meanwhile, the Manhattan distance parameter detected an anomaly of 10 students. However, the visualization does not display all the anomalies due to the limited dimensions. This scatterplot is presented using only 2 dimensions.

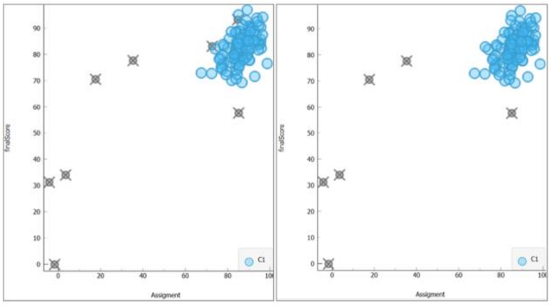
Lastly, this study validates the existing model using the silhouette index technique. The results of the execution of this silhouette index are visualized in the form of a silhouette plot as presented in Figures 4. (a) and 4. (b) This silhouette plot is presented on the two distance parameters, namely: Euclidean and Manhattan. Based on research [17], the fewer instances where the silhouette value is less than 0, the more valid the modelling is. Based on the two figures, the number of students whose silhouette score is less than 0 in the model with the Euclidean distance parameter is less than in the model with the Manhattan distance parameter. This shows that the detection results for the model built with the Euclidean distance parameter are more valid than the model built with the Manhattan distance parameter.

Of course, the results of this detection are very helpful for teachers in monitoring student academic performance. In addition, teachers can immediately take appropriate action against students with anomalies to prevent academic failure.



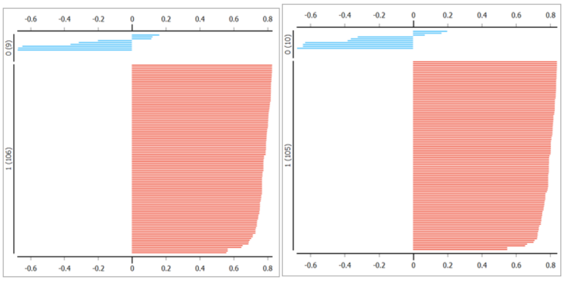
(a) (b)

Figure 2. Parameter of neighborhood distance, (a). Euclidean Distance, (b). Manhattan Distance.



(a). (b)

Figure 3. Visualization of detection results in scatter plot form, (a). Euclidean Distance, (b). Manhattan.



(a). (b)

Figure 4. The visualization of silhouette value (a). Euclidean Distance, (b). Manhattan.

4. CONCLUSION

The conclusions that can be drawn from this research are as follows: (1) detection of student academic performance can be done with an anomaly detection approach based on the Data Mining algorithm; (2) selection of the right algorithm parameters can produce optimal detection performance.

REFERENCE

[1] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining. Pearson Addison Wesley, 2005.

[2] J. Han, J. M. Kamber and Pei, Data Mining Concepts and Techniques. USA: Elsevier, 2012.

[3] O. D. Oyerinde and C. P. Chia, Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression, Int. J. Comput. Appl., vol. 157, 2017, pp. 37–44. DOI: 10.5120/IJCA2017912671.

[4] Y. Yamasari, A. Qoiriah, H. P. A. Tjahyaningtijas, R. E. Putra, A. Prihanto, and Asmunin, Improving the Quality of the Clustering Process on Students' Performance using Feature Selection, International Seminar on Application for Technology of Information and Communication (iSemantic), 2020, pp. 454–458. DOI: 10.1109/iSemantic50169.2020.9234249.

[5] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, Analyzing Undergraduate Students' Performance Using Educational Data Mining, Comput. Educ., vol. 113, 2017, pp. 177–194. DOI: 10.1016/j.compedu.2017.05.007.

[6] 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, 2019, p. e01250. DOI: 10.1016/j.heliyon.2019.e01250.

[7] A. M. Shahiri, W. Husain and N. A. Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Comput. Sci., vol. 72, 2015, pp. 414–422. DOI: 10.1016/J.PROCS.2015.12.157.

[8] B. Guo, R. Zhang, G. Xu, C. Shi and L. Yang, Predicting Students Performance in Educational Data Mining, International Symposium on Educational Technology (ISET), 2016, pp. 125-128. DOI: 10.1109/ISET.2015.33.

[9] N. Tomasevic, N. Gvozdenovic and S. Vranes, An Overview and Comparison of Supervised Data Mining Techniques for Student Exam Performance Prediction, Comput. Educ., vol. 143, 2020, p. 103676. DOI: 10.1016/j.compedu.2019.103676.

[10] L. Jia, H. N. H. Cheng, S. Liu, W.-C. Chang, Y. Chen and J. Sun, Integrating Clustering and

Sequential Analysis to Explore Students' Behaviors in an Online Chinese Reading Assessment System, in 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2017, pp. 719–724. DOI: 10.1109/IIAI-AAI.2017.55.

[11] M. Jovanovic, M. Vukicevic, M. Milovanovic and M. Minovic, Using Data Mining on Student Behavior and Cognitive Style Data for Improving E-Learning Systems: a Case Study, Int. J. Comput. Intell. Syst., vol. 5, 2012, pp. 597–610. DOI: 10.1080/18756891.2012.696923.

[12] Y. Yamasari, A. Qoiriah, N. Rochmawati, K. Yoshimoto, R. Aydin Ahmad and O. Virgantara Putra, Detecting Students' Behavior on the E-Learning System Using SVM Kernels-Based Ensemble Learning Algorithm, Int. J. Intell. Eng. Syst., vol. 16 (1), 2023, pp. 142-153, DOI: 10.22266/ijies2023.0228.13.

[13] B. Sen and E. Ucar, Evaluating the Achievements of Computer Engineering Department of Distance Education Students with Data Mining Methods, Procedia Technol., vol. 1, 2012, pp. 262–267. DOI: 10.1016/j.protcy.2012.02.053.

[14] T. Fang, S. Huang, Y. Zhou and H. Zhang, Multi-model Stacking Ensemble Learning for Student Achievement Prediction, Proc. - Int. Symp. Parallel Archit. Algorithms Program. PAAP, 2021, pp. 136–140. DOI: 10.1109/PAAP54281.2021.9720454.

[15] A. Peña-Ayala, Educational Data Mining: a Survey and a Data Mining-Based Analysis of Recent Works, Expert Syst. Appl., vol. 41 (4), 2014, pp. 1432–1462. DOI: 10.1016/j.eswa.2013.08.042.

[16] Y. Yamasari, N. Rochmawati, R. E. Putra, A. Qoiriah, Asmunin and W. Yustanti, Predicting the Students Performance using Regularization-based Linear Regression, Int. Conf. Vocat. Educ. Electr. Eng., 2021, pp. 1–5. DOI: 10.1109/ICVEE54186.2021.9649704.

[17] P. J. Rousseeuw, Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, J. Comput. Appl. Math., vol. 20, 1987, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[18] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD-96 Proceedings, 1996, pp. 226-231.