Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 11 Issue 1 Year 2021 Pages 26-35

DOI: http://dx.doi.org/10.31940/matrix.v11i1.2330
URL: http://ojs.pnb.ac.id/index.php/matrix

The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI)

Yulison Herry Chrisnanto [1]****, Gunawan Abdullah*** [2]

[1,2] *Universitas Jenderal Achmad Yani, Indonesia*

*Corresponding Author: *y.chrisnanto@gmail.com*

Abstract: Education is an important thing in a person's life, because by having adequate education, one's life will be better. Education can be obtained formally through formal institutions that constructively provide a person's abilities academically. This study aims to determine student performance in terms of academic and non-academic domains at a certain time during their education using techniques in data mining (DM) which are directed towards academic data analysis. Academic performance is delivered through the Educational Data Mining (EDM) integrated data mining model, in which the techniques used include classification (ID3, SVM), clustering (k-Means, k-Medoids), association rules (Apriori) and anomaly detection (DBSCAN). The data set used is academic data in the form of study results over a certain period of time. The results of EDM can be used for analysis related to academic performance which can be used for strategic decision making in aca-demic management at higher education institutions. The results of this study indicate that the use of several techniques in data mining together can maximize the ability to analyze academic performance with the same data source and produce different analysis patterns.

Keywords: educational data mining, academic performance, educational institution, data mining technique.

History Article: Submitted 25 January 2021 | Revised 25 February 2021 | Accepted 22 March 2021

How to Cite: Y. H. Chrisnanto and G. Abdullah, “The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI),” Matrix: Jurnal Manajemen Teknologi dan Informatika, vol. 11, no. 1, pp. 26–35, 2021.

Introduction

A person's academic ability can usually be measured by a measure of intelligence known as the intelligence quotient (IQ), however, the level of IQ does not correlate with a person's level of success. There is a weak positive correlation between IQ and academic achievement. Various dimensions of emotional intelligence were found to be predictors of academic success [1].

Educational Data Mining (EDM) is an application in data mining and statistics to form information from a specific educational area, such as a school or college. EDM refers to tools and techniques for automatically extracting meaning from a large repository of learning activities in an educational environment [2]. The techniques used depend on the required data analysis. The Association Rule technique can be used to evaluate student behavior. K-Means Clustering can be used to find the best centroid in student data such as attendance, GPA, final grades, and others. Meanwhile, the Rule Based Classification (RBC) technique can also be used to extract the relationship between attributes from the dataset and class labels [2]. EDM is an analysis mechanism related to educational data mining to find out interesting patterns and knowledge in educational organizations [3].

Student academic performance will determine the level of success in completing the study according to the required academic load. The education system based on semester credit units (SKS) is designed to make it easier for students to manage the learning process independently. However, there are still many students who have not met the required academic performance. In the case study in this study, UNJANI determined a target of 80% of graduation on time, but in 2016-2019, the average graduate on time from students averaged 68.5% [4][5]. This indicates that the learning process carried out has obstacles. The ability of education managers in analyzing student performance is still limited to using the data on the accumulated results of the course scores obtained by students at a certain time.

This study aims to use data mining techniques to explore the heaps of academic information that are already owned to become knowledge in the form of analysis results that can support strategic decision making related to student academic performance. The data mining techniques used will be integrated into one data mining system for academic purposes, this system is known as Educational Data Mining (EDM).

Related Work

Educational Data Mining (EDM) is a mechanism for the use of interdisciplinary knowledge that ap-pears in research areas related to the development of methods for exploring data originating from an educational context with a computational approach to analyzing educational data, for studying educational questions [6]. Various techniques used in EDM have emerged over the last few years. Several techniques in general have similarities with the use of data mining in other domains be-sides education. There are 4 (four) main method classes that are very often used in EDM, including: (a) Prediction models, (b) Structure discovery, (c) Relationship mining, and (d) Discovery with models [7].

Using EDM in higher education institutions can improve the teaching and learning process. EDM, is useful in many different areas of higher education including identifying students who are at risk of failure, identifying prioritized learning needs for specific groups of students, increasing graduation rates, effectively assessing institutional performance, maximizing college resources, and optimizing curriculum reform [8]. Based on the results of 402 studies, it was found that certain EDM techniques and learning analytics (LA) can be used as the best way to solve certain learning problems. Implementing EDM and LA in tertiary institutions can develop strategies that are directed at students and provide the tools necessary so that they can be used by institutions for continuous improvement purposes [9].

Algorithms in data mining (DM) are used to analyze academic performance. The DM algorithm used includes clustering and classification. Clustering is used to group student learning patterns to be more efficient, classification algorithms are used to predict future student behavior with de-tailed learning information such as student scores, knowledge, achievement, motivation, and attitudes [10]. The EDM system is also used as a reminder to students, which uses classification and clustering techniques in building system intelligence. This system can be used as a consultation tool for students in the first year to reduce the academic failure rate [11]. The most widely used technique in the EDM system is the Decision Tree Algorithm in the classification with the best accuracy (99.7%), while the clusterization technique is the Expectation Maximization (EM) algorithm which is the best [12] . Another DM technique is the ensemble method used to improve classification performance. By using Bagging, Boosting and Random Forest (RF), which are generally an ensemble method that is widely used in the literature [13].

EDM has the power to use raw data effectively where it has been generated by various academic activities in higher education, as well as to create a significant impact on the academic domain that can illustrate hidden patterns and relationships between attributes used in predicting student performance, or their behavior. effectively, so that strategic decisions can be taken appropriately [14][15]. EDM is an emerging cross-disciplinary research field that deals with the development of various methods for exploring data that come from an educational context. EDM uses a computational approach to analyzing educational data to study educational questions [16].

Methodology

The system for detecting student academic performance using data mining techniques can involve two actors, namely academic managers and students as parties whose academic performance needs to be measured. The data used for the detection process involves academic systems and support systems that generate data temporally. The result of the EDM process is in the form of new knowledge that will be used by the academic manager in developing academic programs. Figure 1 below illustrates the EDM utilization model. The main data source in this system is a dataset that comes from a database on the higher education academic system.

Two actors who interact with the system include management, namely the management of education and students who receive educational services. The external part of the system is the academic program, which is the current academic system owned by the institution. The academic system in the form of academic implementation governance, can be the main data source related to academic activities, student data, curriculum data and data related to the academic process.

Educational environment is a data processing environment originating from the academic system, which is then managed with the help of a software system dedicated to managing academic data owned by the institution. Especially in this study, the academic data needed is in the form of temporal data, where academic activities are recorded temporarily through the e-learning system equipped with a temporal data design that established in previous studies. In addition to temporal data that stores academic aspects, the system is expected to record non-academic data that represents a student's EQ ability.

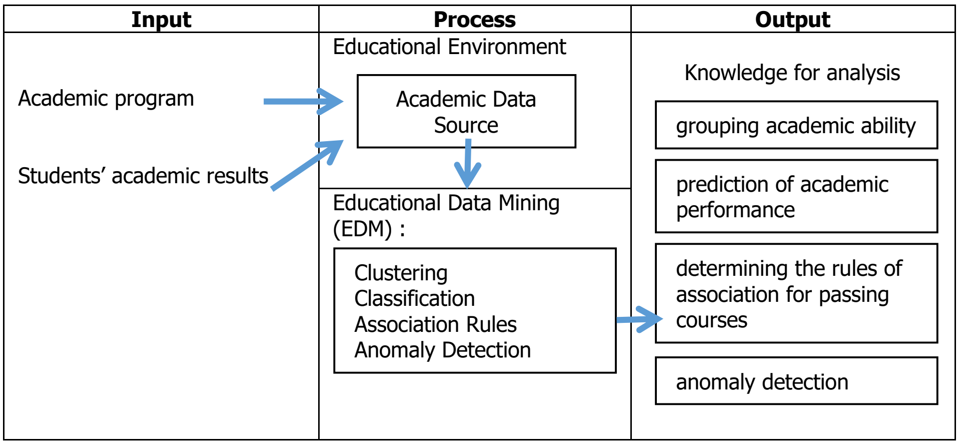


Figure 1. EDM concept

The concept of data mining is implemented using EDM, where the techniques used are analytical support in accordance with the requirements to be able to identify student academic performance. Before entering the main process, the system will pre-process, which is to prepare the dataset according to the technique to be used. The dataset used is the academic data record of students majoring in Informatics at UNJANI with the 2015/2016 and 2016/2017 Academic Years, in which these students have completed their studies. The data mining models used include: Association Mining Rules, Classification and Clustering, and Outlier's Detection. The stages carried out in this study are as follows:

1. Preparation of the main dataset: In this study, a dataset will be used in the form of scores of students who have completed the study for 8 (eight) semesters and have declared their graduation status. Value data is in the form of all course scores that have been converted into numbers. Non-academic data is data in the form of student activities that are recorded and interpret the attitudes or behaviour of students during lectures.
2. Pre-processing: namely the process of forming a final dataset that can be used for the entire EDM process. In this process, first the main dataset is prepared which comes from the data-based on the academic system that has been used. The data is transferred into a comma separator (CSV) text format. There are several datasets in the CSV format that will be used in this EDM process.
3. Determination of parameter values: Each data mining technique used requires parameters before the mining process is carried out, including the parameters for conducting the clustering process in the form of a k value which states the number of clusters to be formed. Other parameters, namely the value of support and confidence to carry out the mining process in the form of association rules formation. For the classification process, some parameters are needed to determine the depth of the decision tree to be formed.
4. Selection of data mining models: There are several datasets that are prepared through pre-processing which will be used by different mining techniques. The clustering model uses a dataset of average scores for odd semesters and even semesters. The association rule model

uses a dataset that contains the value of courses with a quality value of 'A', while the dataset used for the classification process is in the form of all courses for 8 (eight) semesters.

Results and Discussions

In accordance with the research flow that the educational data mining (EDM) concept uses a data mining model with machine, either supervised or unsupervised, learning in determining the pattern of relationships between data with one another in the form of association rules. The dataset used is then entering the pre-processing stage so that each data mining model will use a different dataset. The data mining techniques used consist of clusterization, classification, association rules and detection of outliers. EDM implementation uses Python programming tools version 3.8.5 with IDE Spyder 4.1.4 and uses RapidMiner Studio Educational 9.6.0. The use of this tool is in consideration of adequate library support and features in data mining processes.

As previously stated, the dataset used is an academic database obtained from the university's academic system. Furthermore, pre-processing is carried out to become a dataset that is ready to be mined.

Results

The first function in EDM is the application of a clustering model using the k-Means algorithm. Before clustering, it is necessary to determine the ideal k value (number of clusters) first, that is, it can use the calculation of the Sum of Square Error (SSE). The SSE results can be visualized to see the formation of an elbow that tends to be sharp one. The following is the result of the visualization of the elbow determination as shown in Figure 2.

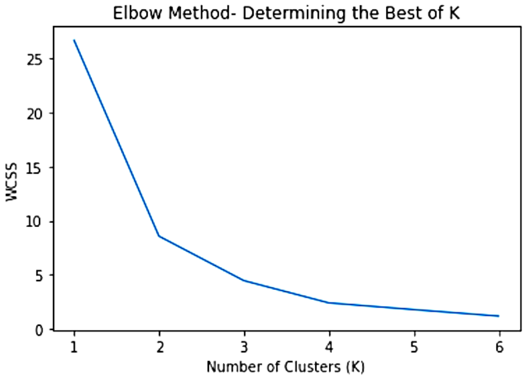


Figure 2. Elbow visualization results

By using the dataset that has been previously presented, the results of calculations with SSE obtains the ideal k value of 2 (two). This can be seen from the visualization with the formation of an elbow. The clustering process is then determined by k value of 2. By using the same dataset, the clustering process can be visualized (Figure 3) as follows.

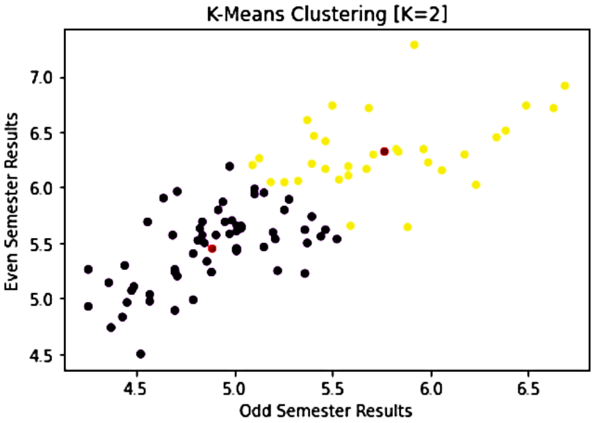


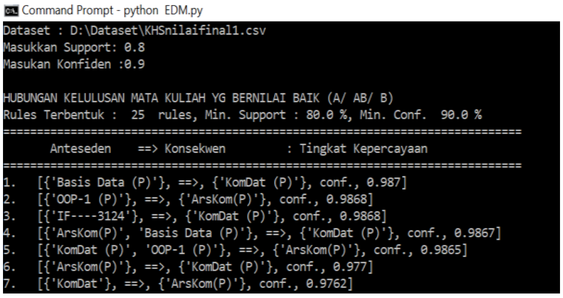
Figure 3. Clustering results with k of 2

The number of datasets used is 91 data (representing the number of students) who have completed the study for 8 (eight) semesters. Based on the cluster formed (Figure 5) it can be interpreted that good academic abilities are less when compared to moderate and less academic abilities. The horizontal axis represents the grade in the odd semester, and the vertical axis represents the even semester grade. Cluster analysis is also carried out using the k-Medoids algorithm, where this algorithm determines the data center (medoid) by calculating the closest distance (cost) between the non-medoid data objects and the randomly selected medoid candidate. It also compares the average distance of average non-medoid data objects with old medoids. The data is processed using the Davies-Bouldin Index (DBI), that when the k-Means algorithm is smaller when compared to the DBI results for k-Medoids, thus the accuracy of the clusters formed for the 2 k-Means algorithm clusters is slightly better. Table 1 shows the results of the cluster analysis.

Table 1. Cluster analysis results

Methods	**DBI**	**Avg. Dist**	**Analysis**		**Amount of data**	
			Cluster 1	**Cluster 2**	**Cluster 1**	**Cluster 2**
k-Means	0.714	0.251	31.38%	63.80%	61	30
k-Medoids	0.770	0.269	41.29%	51.85%	50	41

The next EDM process is to determine the pattern of relationships between passing courses of one another. Whether the passing of one course will be influenced by the passing of other courses. The level of relationship between one data object and another data object is determined based on the value of support (*Masukan Support* in Indonesian) and the value of confidence (*Masukan Konfiden* in Indonesian). This EDM process uses the Apriori algorithm in the association rule (AR) model. Figure 4 below shows the results of the execution of the association rule formation as follows in Indonesian.



```
Dataset : D:\Dataset\KHSnilaifinal1.csv
Masukkan Support: 0.8
Masukan Konfiden :0.9

HUBUNGAN KELULUSAN MATA KULIAH YG BERNILAI BAIK (A/ AB/ B)
Rules Terbentuk :  25  rules, Min. Support : 80.0 %, Min. Conf.  90.0 %
===========================================================================
      Anteseden    ==> Konsekwen         : Tingkat Kepercayaan
===========================================================================
1.   [{'Basis Data (P)'}, ==>, {'KomDat (P)'}, conf., 0.987]
2.   [{'OOP-1 (P)'}, ==>, {'ArsKom(P)'}, conf., 0.9868]
3.   [{'IF----3124'}, ==>, {'KomDat (P)'}, conf., 0.9868]
4.   [{'ArsKom(P)', 'Basis Data (P)'}, ==>, {'KomDat (P)'}, conf., 0.9867]
5.   [{'KomDat (P)', 'OOP-1 (P)'}, ==>, {'ArsKom(P)'}, conf., 0.9865]
6.   [{'ArsKom(P)'}, ==>, {'KomDat (P)'}, conf., 0.977]
7.   [{'KomDat'}, ==>, {'ArsKom(P)'}, conf., 0.9762]
```

Figure 4. The results of the formation of association rules with the passing grade of subject is A.

The result of association rule will be influenced by the value of support and confidence given as parameters. The higher of the confidence value, the more confidence of the rules are formed, where each antecedent will determine the consequences. With a support value of 80% and a confidence value of 90%, it obtains 25 rules consisting of courses with A grade, as shown in Figure 4 above.

Another EDM process is the classification process using the Decision Tree (DT) model, which is a model that can be used to determine the status of an unknown student's graduation based on new data objects. Classification is the process of determining a new data class based on a training dataset (past dataset). DT is a data mining model that uses a decision tree pattern, where each node in the tree is an attribute in the dataset. The initial node of DT is determined based on the entropy value. The result of the decision tree is knowledge of conditional rule patterns (if-then-else) which can be used to determine previously unknown data classes. It can be seen in Figure 5.

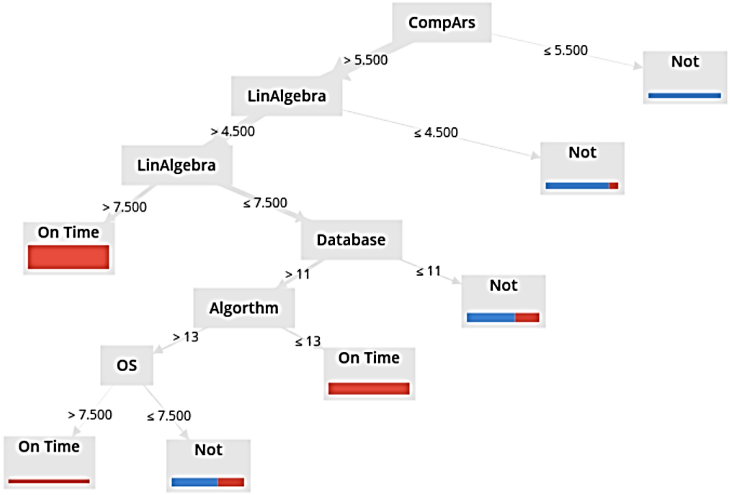


Figure 5. The results of the formation of a decision tree using ID3

Figure 5 shows the tree structure generated by the DT algorithm, where root is the initial attribute determined based on the entropy value calculation. The DT data mining model can be used to determine labels / data classes that are not yet known and will be assigned a class (test dataset). Based on the DT performance measurement, the rule formation process in the decision tree is very good, as can be seen in Table 2 and 3. While the value of AUC (optimistic): 0.993 (positive class: right), AUC (pessimistic): 0.946 (positive class: right), thus the resulting classification is said to be a very good classification.

Table 2. Decision Tree (DT) performance measurement

Measurement	
Accuracy	93.41%
Precision	92.54%
Recall	98.41%

Table 3. Confusion matrix

True	**True Not On Time**	**True On Time**	**Class Precision**
Predictions Not on time	24	3	88.89%
Predictions on time	4	60	93.75%
Class Recall	85.71%	95.24%	

Another EDM process is to determine the outliers of student grade data using the DBSCAN model, which is a model with the same way of working with clustering such as k-Means. DBSCAN determines the farthest object from the cluster formed, so that the farthest object can be identified as an outlier. In the student academic score dataset, the object farthest from the cluster can be interpreted as a data anomaly. Figure 6 below visualizes the results of the DBSCAN execution process to determine anomalous data as follows.

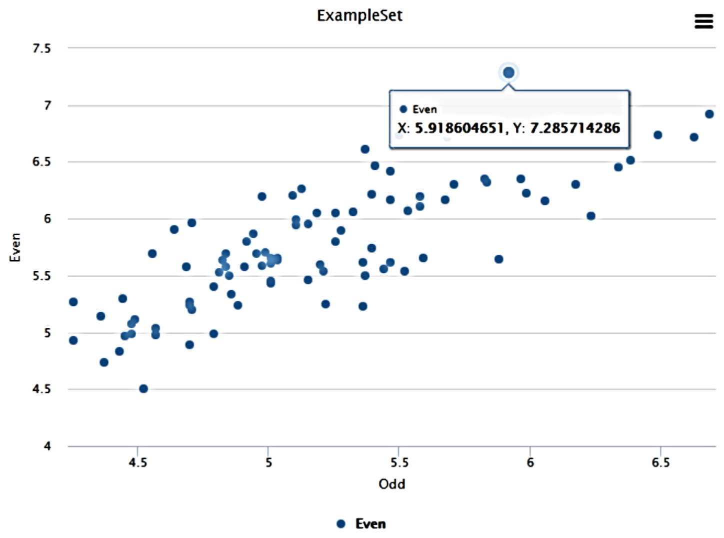


Figure 6. The results of the visualization of data anomaly determination with DBSCAN

The picture above shows that based on the same academic dataset, the farthest data can be seen from other data sets and can be interpreted as an anomaly. In this case, it can be interpreted as a student's academic ability that is more prominent when compared to other academic abilities. Based on the measurement of the performance of the formation of outliers using the DBSCAN algorithm with an epsilon value of 0.6 and main points value of 10, one data is produced that is outside the other data objects that can be identified as anomalies, namely data with x of 5.92 and y of 7.29. By changing the epsilon value, it will affect the formation of the resulting data anomaly. The following Table 4 shows the anomalous changes that can be detected which are influenced by the epsilon value as follows.

Table 4. Change in the value of epsilon in DBSCAN for min points of 10

epsilon	average	anomaly
0.2	64	27
0.3	77	14
0.4	84	7
0.5	87	4
0.6	**90**	**1**
0.7	91	0

The best data anomaly is obtained at the epsilon value of 0.6 (Table 4), where the number of anomaly formed is 1. The resulting anomaly data is outside the data object set with average density, thus it can be identified that the anomaly data is beyond average of academic scores. Based on the results of anomaly detection data (Table 4), only one data indicates that academic performance is very good or above average, where the data is the farthest data from other data sets as shown in Figure 6 and 7.

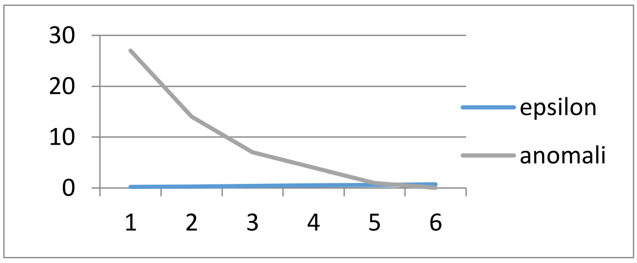


Figure 7. The epsilon and anomaly value.

Discussions

After conducting a series of experiments by applying various techniques in data mining, where the dataset used is the recording of student academic results in the UNJANI Informatics Department, with the academic years 2015/2016 and 2016/2017 having produced various information that can be interpreted as a description of student academic abilities. Table 5 shows an overview of the results of the analysis of the use of EDM to analyze student academic performance.

Table 5. The results of the academic performance analysis through the application of EDM

Data Mining Algorithms					**EDM analysis**
Clustering	Algoritma	DBI	Cluster-0	Cluster-1	
	k-Means	0.714	31.38%	63.80%	As many as 61 students have moderate academic ability As many as 30 students have good academic abilities
Classification	Algoritma	Accuracy	Precision	Recall	If the value of MK-13 courses> 2.5, it will be determined by the MK-11 course, if MK-11> 7.5, then it will be on time If the value of the MK-13 course <2.5, it will be determined by the MK-5 course, if the MK-5> 13, then it will be on time If the MK-11 course score is <7.5, the MK-9 course score must be> 11 to be able to graduate on time outside of the above provisions, it is not able to pass on time
	ID3	93.41%	92.54%	98.41%	
Outliers Detection	DBSCAN	*epsilon*	Min points	anomali	Based on the results of anomaly detection data (Table 4), only one data indicates that academic performance is very good or above average, where the data is the farthest data from other data sets.
		0.6	10	1	
Association Rules	Support	Confidence	Sum of rules		shows that passing the database course also has an effect on passing data communication courses with a confidence level of 98.7%.
	80%	90%	25 rules		

Conclusion

Educational Data Mining (EDM) has become an adequate tool to be used as a support for strategic decision making, especially in higher education institutions. EDM is a relatively new concept used for the acquisition of knowledge in the field of education through techniques in data mining. The techniques used in this study including clusterization, classification, association rules and anomaly detection.

The dataset used in this study comes from the academic data of students majoring in Informatics, Faculty of Science and Informatics UNJANI, batch 2015 and 2016, in which these students have taken all the required semesters (8 semesters) so that their graduation status is known. By using this dataset, the EDM process can be carried out in acquiring knowledge. The clustering process uses the k-Means technique using the ideal k value = 2, after first measuring the SSE (sum square of error) = 18.1388, the result of clustering is that there are 2 (two) groups of student academic patterns. The k-Means algorithm gives better results when compared to the k-Medoids algorithm. Another technique used is the classification technique with a decision tree algorithm, where the result of this technique is the formation of a conditional rule pattern (if-then-else) by determining the attribute as the initial root to calculate the entropy value. By comparing the two algorithms in the classification (ID3 and SVM) it can be seen that the ID3 algorithm has better performance, this can be influenced by the dataset used. To determine data anomalies from academic data using the DBSCAN algorithm, this algorithm can detect academic data anomalies with epsilon value of 0.6 and a min points value of 10.

The techniques in EDM can be used as a way to analyse patterns that are formed into new knowledge in the management of education in higher education using academic data sets generated through various information systems available at higher education institutions.

Acknowledgments

Thanks are conveyed to LPPM General Achmad Yani University (UNJANI) for the opportunity and concrete support that has been given so that this research can be carried out. Thank you also to the Department of Informatics, Faculty of Science and Informatics, UNJANI for providing access to academic data so that this research can use real academic data.

References

[1] N. Shipley, M. Jackson, and S. Segrest, "The effects of emotional intelligence, age, work experience, and academic performance," *Research in Higher Education Journal*, pp. 1–18, 2010.

[2] P. M. Kumari, S. A. Nabi, and P. Priyanka, "No educational data mining and its role in educational field," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, 2014.

[3] A. Abu, "Educational data mining & students' performance prediction, "*International Journal of Advanced Computer Science and Applications*," vol. 7, no. 5, pp. 212–220, 2016.

[4] Y. H. Chrisnanto and A. Kanianingsih, "Identifikasi pola kemampuan akademik menggunakan teknik association rules," in *Sentika*, 2017.

[5] Y. Chrisnanto and G. Abdillah, "Penerapan algoritma partitioning around medoids (PAM) clustering untuk melihat gambaran umum kemampuan akademik," in *Seminar Nasional Teknologi Informasi dan Komunikasi*, 2015, pp. 444–448.

[6] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, 2020.

[7] R. S. Baker and P. S. Inventado, *Educational Data Mining and Learning Analytics*. 2014. [Online]. Available: https://www.semanticscholar.org/paper/Chapter-X-%3A-Educational-Data-Mining-and-Learning-Baker-Inventado.

[8] A. Algarni, "Data Mining in Education, "*International Journal of Advanced Computer Science and Applications*," vol. 7, no. 6, 2016.

[9] H. Aldowah, H. Al-samarraie, and W. Mohamad, "Telematics and Informatics Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019.

[10] R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, "Analysis of educational data mining," in *Harmony Search and Nature Inspired Optimization Algorithms*, Springer, 2017, pp. 897–907.

[11] H. M. Nagy, W. M. Aly, and O. F. Hegazy, "An Educational Data Mining System for Advising Higher Education Students," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 7, no. 10, pp. 1266–1270, 2013.

[12] S. Hari Ganesh and A. Joy Christy, "Applications of Educational Data Mining: A survey," in *ICIIECS 2015 - 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems*, 2015.

[13] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016.

[14] H. Kaur and G. Bathla, "Student performance prediction using classification data mining techniques," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 12, pp. 93–97, 2007.

[15] A. Hicham, A. Jeghal, A. Sabri, and H. Tairi, "A survey on educational data mining [2014-2019]," in *2020 International Conference on Intelligent Systems and Computer Vision, (ISCV),* 2020, pp. 21–25,.

[16] J. D. Patón-romero, M. Teresa, M. Rodríguez, and M. Piattini, "Computer standards & interfaces application of ISO 14000 to information technology governance and management," *Computer Standards & Interfaces*, vol. 65, no. April, pp. 180–202, 2019.