ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 2, pp. 137-152
DOI: 10.33168/JSMS.2022.0206

Measuring Students' Academic Performances: A Data Clustering Approach

Ahmed Abdulmutallab Ahmed, Mohammed Alhanshaly, Osman A. Nasr,

Saeed Al-Maliki

King Khalid University, Saudi Arabia

oanassr@kku.edu.sa

Abstract. This research aims to measure students' performance in semester activities to improve academic advising performance using the data Clustering methodology. It aims to leverage machine learning features to form homogeneous counselling groups that contain a small number of students. During semester activities, these groups are distinguished by their academic performances and help faculty members solve students' performance issues. Emphasis was placed on students' performance in the introductory programming course as it is one of the main courses that have the most substantial impact on students' academic performance. In this study, we chose some of the features available from the academic student data file in the fourth semester. In this semester, the intensive counselling process begins and falls under the responsibility of the department's academic advisors. We also chose to apply Expectation-Maximization (EM) algorithm using the WEKA open-source package. The experiment was designed on five scenarios, and the results were analyzed in each scenario. We found out that the best scenario is to divide the students into seven clusters, which gave homogeneous groups depending on the mean and standard deviation scales and considering the effort and the available number of faculty members.

Keywords: Data clustering, EM, programming, academic advisor, class activities, academic performance, WEKA, King Khalid University

1. Introduction

Academic advising is one of the duties and responsibilities of a faculty member that complements his primary educational, research, and administrative tasks. It includes (i) supervising a group of students academically, (ii) registering students each semester according to the educational path and courses offered, (iii) contributing to solving academic and social problems, and many others. Faculty members are ideal for advising students due to their frequent interactions with students and knowledge of academic programs (Hutson 2013). Thus, they play a critical role in student academic success.

Academic advising is necessary for all students regardless of their academic levels and assists in their academic success. It can directly affect students' persistence and probability of graduating or indirectly affect grades, intentions, or satisfaction with the student role (Pascarella et al., 2005). However, it is still challenging for faculty members to advise students because no mechanism is followed to distribute students into advising groups. Also, developing the requisite skills and knowledge to provide all kinds of advising necessary to students is time-consuming and challenging (Smith et al., 2006). Therefore, we can summarize the study problems that lead to our research as follows:

- The random distribution of students into advising groups
- The variation in students' academic performance within the same group
- The variation in students' general academic performance in different semester activities
- The ratio between the student numbers and faculty members(advisors) in the department
- The effort of a faculty member in the academic advising process
- The classrooms do not accommodate many students for the collective academic advising process under the conditions of the Corona pandemic and the imposed health restrictions.

The objective of this study is in four folds: (i) to facilitate the academic advising process, (ii) to analyze and evaluate students' overall academic performance, (iii) to create homogeneous groups based on students' academic performance, and (iv) to share the academic advising efforts among faculty members using machine learning. Machine learning is a branch of artificial intelligence (AI) that precisely use data and learning algorithms to imitate how humans learn and gradually improves its accuracy (Silver 2016; Mohamed et al., 2018). It plays a significant rule in data science for providing intelligent data analytics. It applies several methods using algorithms, for example, statistical methods; here algorithms are trained to make classifications, predictions, and uncovering critical insights within data mining projects (Wong).

Unsupervised learning or unsupervised machine learning is one of the machine learning categories, which applies algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns and group objects so that objects

in one group will be similar and different from those in another group (Ariouat et al., 2016). This technique of machine learning has an ability to discover similarities and dissimilarities measure that makes it an ideal solution for Educational Data Mining (EDM) problem, such as exploring, analyzing, and retrieving data in an educational setup. It is also used for dimensionality reduction, Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) which has two common approaches. Similarly, there are some more examples of unsupervised learning algorithms, such as neural networks, k-means clustering, and probabilistic clustering methods (Paul et al., 2020).

The presented study uses the Expectation-Maximization (EM) algorithm, which is a clustering method, and gives the mean and standard deviation in its output forms. The study's has applied the EM algorithm to evaluate and measure the homogeneity of a group in this study.

The rest of the paper is organized into seven sections: Section 2 presents related work. Section 3 consists of our research methods. Section 4 presents results and discussions. Finally, section 5 shows the work's conclusion.

2. Literature Review

Clustering is one of the most common exploratory data analysis techniques used today to understand the data structure and its classification. It classifies data objects into subgroups (clusters) so that objects within a subgroup (cluster) have high similarities in comparison to one another but are dissimilar to objects in other clusters (Deepika et al., 2018). In other words, it leads to the discovery of previously unknown groups within the data. Unlike supervised learning, clustering is an unsupervised learning method since the data points are unlabeled, and cannot compare the output of the clustering algorithm to the actual labels to evaluate its performance. The study's approach is to investigate the effective data classification technique by grouping the data points into distinct subgroups. The data objects are grouped or clustered based on a similarity measure such as correlation-based distance or Euclidean-based distance. The use of similarity measure types is application-specific. Noticeably, there are several clustering methods, and each may generate different results on the same data set.

Clustering concept is widely used in various applications, such as business intelligence, the web, security, and image pattern recognition. For example, a business intelligence application, customers are classified and clustered based on their spending behaviours. This clustering helps the business establishment to target such customers and market the product of the customers' choice.

An emerging and active area of research is process mining, which considers different ways to understand to understand students' habits and the factors that influence their performance. It applies several models, such as spaghetti models, which are used for the students' general behavior, but these models too large and

complex to use. A two-steps approach was used to improve mining in the educational process – creating clusters based on employability indicators and obtaining clusters using the AXOR algorithms, which is used for obtaining, refine results from the first step (Ariouat et al., 2016; Paul et al., 2020). In context to educational data mining, hidden knowledge can be obtained by discovering relationships between student learning characteristics and behavior. Instructional data modeling mining enables educators with future teaching assets that enhance students' learning and it helps academic stakeholders to improve teaching quality standards and reduce students' failure rate (Deepika et al., 2018; Wong et al., 2020).

Educational data mining is a beneficial approach in case of addressing students' risks, creating clusters based on priority learning needs, and help improving result performance. By extracting required information, accessing educational data set helps to analyze students' academic performances. On the other hand, students' academic performance depends on various factors such as psychological, personality, demographic, educational background, academic progress, and other environment variables. Mostly, these variables are related in a complicated nonlinear way (Almarabeh 2017; Naser et al., 2015; Nagy et al., 2013; Abawajy 2019).

3. Research Method

In this study, we present a technique to create homogenous clusters based on academic activities, such as assessments (mid exams, lab exam, and assignments) to improve academic advising services (Han et al., 2011; Romero et al., 2008; Barros et al., 2000; Bartko et al., 1976; Cohen 1960; Florin 2011; Chaitra et al., 2022). Fig. 1 demonstrates the research methodology and study's approach.

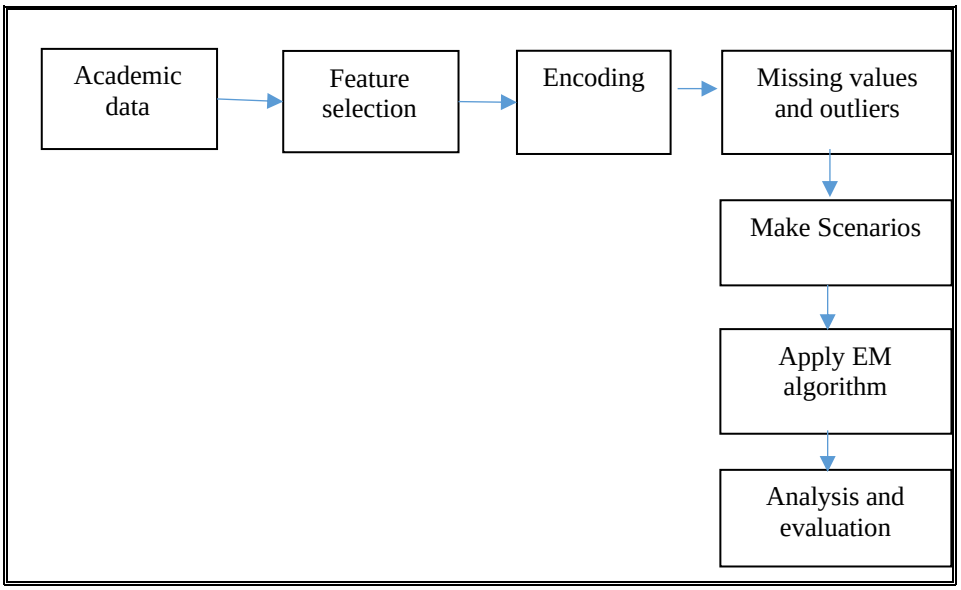

Fig. 1: Research method

3.1. Experiment design and tool

We used WEKA, an open-source data analysis tool, for the experiments. It is an effective tool and supports many learning algorithms used in data mining applications (Ian et al., 2005; Gao et al., 2017; Jiawei et al., 2012; Kalyani et al.,; Husain et al., 2021). We formulated different scenarios to analyze the results of the EM algorithm.

1. Scenario I: divide the set of students into three clusters
2. Scenario II: divide the set of students into four clusters
3. Scenario III: divide the set of students into five clusters
4. Scenario IV: divide the set of students into six clusters
5. Scenario V: divide the set of students into seven clusters

The EM algorithm was applied to all these scenarios and the results of each scenario were analyzed concerning the arithmetic mean and standard deviation of the academic performances in each group and compared these results with the study objective.

3.2. EM algorithm

Expectation-Maximization (EM) is a classic algorithm developed in the 60s and 70s with diverse applications. For example, it can be used as an unsupervised clustering algorithm and extends to NLP applications like Latent Dirichlet Allocation[1], the Baum–Welch algorithm for Hidden Markov Models, and medical imaging. As an optimization procedure, it is an alternative to gradient descent with the significant advantage: in many circumstances, the updates can be computed analytically. It is also a flexible framework for thinking about optimization (Yoon et al., 2020; Kass 2020; Michael et al., 2004; Quilan 1986; Remco 2012).

Like k-means clustering, we start with a random guess for two distributions/clusters are and then proceed to improve iteratively by alternating two steps:

1. (Expectation) Assign each data point to a cluster probabilistically. In this case, we compute the probability of the red and yellow clusters, respectively.

2. (Maximization) Update the parameters for each cluster (weighted mean location and variance-covariance matrix) based on the points in the cluster (weighted by their probability assigned in the first step).

The Expectation-Maximization algorithm provides solution to the problem. Our plan is:

1. Start with an arbitrary initial choice of parameters.
2. (Expectation) Form an estimate of Δ.
3. (Maximization) Compute the maximum-likelihood estimators to update our parameter estimate.
4. Repeat steps 2 and 3 to convergence.

3.3. Data collection and preprocessing

The dataset was obtained from the Deanship of Admission and Registration, which contains various types of student related data. The quantitative characteristics were selected to represent the semester activities. The dataset contains 146 records and five quantitative features, as shown in Table 1.

Table 1: Selected quantitative features

Index	Feature name	Description
1	MID I	Student's mark in Mid I exam
2	MID II	Student's mark in Mid II exam
3	LAB	Student's marks in Lab exam
4	HW	Student's total mark in HomeWorks

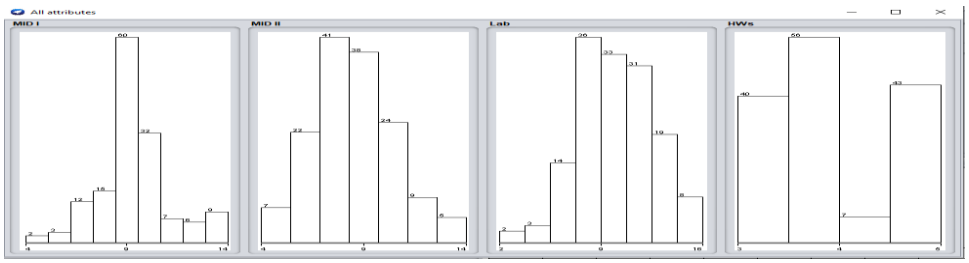

Fig. 2: Statistical metrics of class's activities

4. Result and Discussion

We assumed a specific number of clusters based on each scenario's (see Section 4.1) requirements. We analyzed the characteristics of the clusters and commented on the mean and standard deviation of students' performance in-class activities and the general trend of performance.

4.1. Scenario I

For this scenario, we distributed students into three groups (clusters), as shown in Fig. 2, and thus nominated three faculty members for academic advising.

It is clear from the algorithm results that the average performance of the three groups varies in all quarterly activities with minor differences of less than 2.5, indicating homogeneity in performance as shown in fig 4. Also, we found that the general academic performance for group 0 is above the mean, which ranges between (8.9-10.6) in all activities. Around 66% students fall into this group, as shown in Table 2.

As for group 1, we found that the degree of homogeneity is significant and the standard deviation between the scores of students in this group is less than 2.0. However, the general performance of the students of this group is weak in all

activities, so guidance and effort should be high and intense for the students of this group.

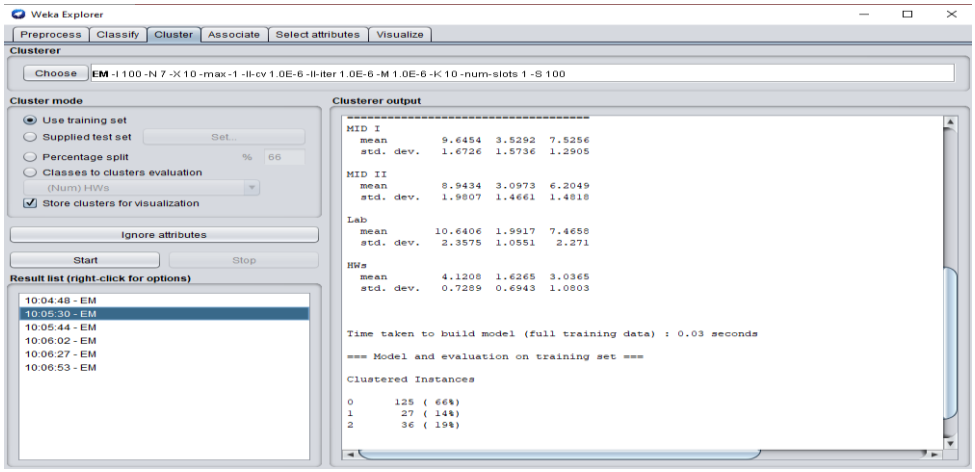

Fig. 3: Scenario II result

This group represents 27% of the total students. On the other hand, we found that the deviation reaches 2.3 in group 2 and that its general performance is considered a medium in all activities.

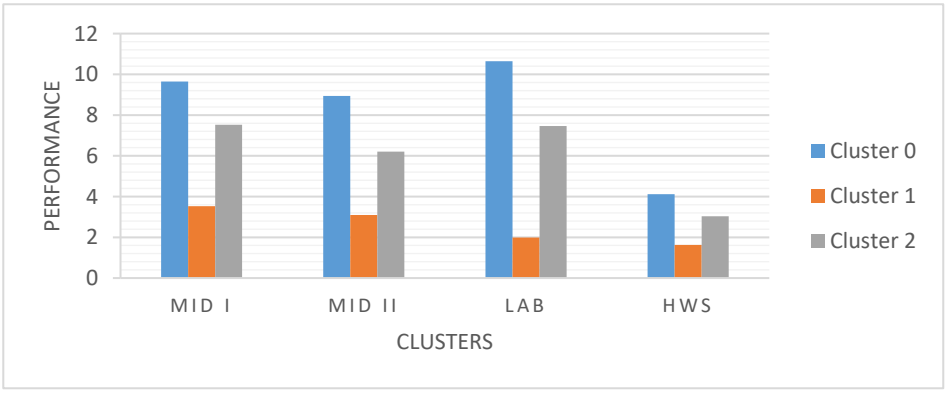

Fig 4: Activities distribution in each cluster

We can reject this scenario because of the large numbers of students are in group 0, as shown in Table 2. It represents 66% of students, which exhausts the faculty member, makes the guidance process difficult, and the absence of classrooms can accommodate this number of students.

Table 2: Number of students in each cluster

Clustered Instances		
0	125	%66
1	27	%14
2	36	%19

4.2. Scenario II

In this scenario, we distributed students into four groups (clusters) and thus nominated four teaching faculty members for academic advising.

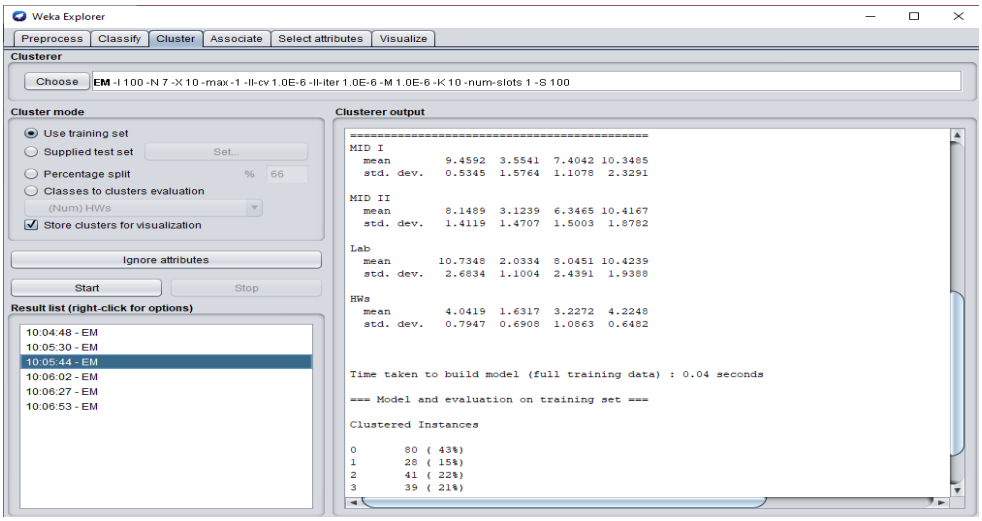

Fig. 5: Scenario II result

We can read from the results from Figure 5 that dividing students into four gives us homogeneous groups with a standard deviation lies between (0.6 -2.6) in all activities. Also, the general performance of groups 0 and 3 are considered above the mean, while the academic performance of students in group 1 is weak and in group 2 is middle.

Fig. 6: Activities distribution in each cluster

We can reject this scenario because of the large numbers of students are in group 0, as shown in table 3. It represents 43% and group 3 represents 21% and with numbers 80 and 39, respectively, making academic advising difficult due to many students and adherence to the required precautions.

Table 3: Number of students in each cluster

Clustered Instances		
0	80	%43
1	28	%15
2	41	%22
3	39	%21

4.3. Scenario III

In this scenario, we distributed students into five groups (clusters) and thus nominated five teaching faculty members for academic advising.

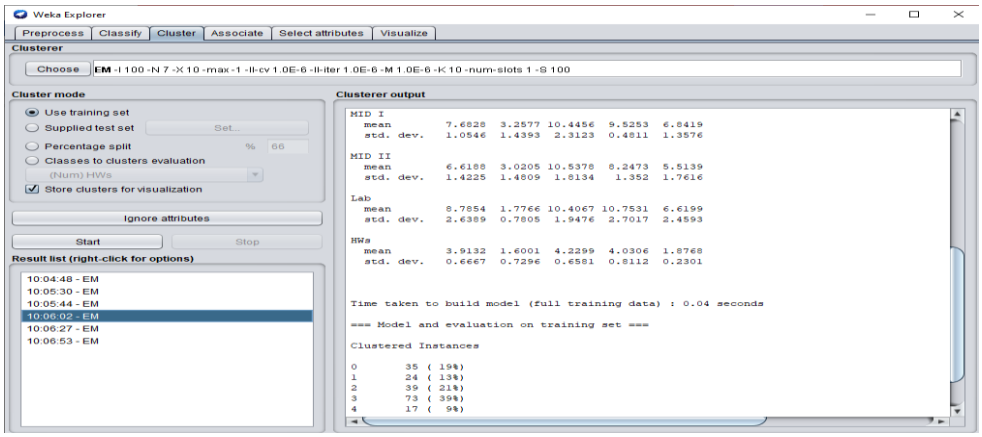

Fig. 7: Scenario III results

Dividing students into five clusters helped to create semi-homogeneous groups. However, by looking at the numbers of students in Table 4, we found that the number of students is large, and the faculty member may not be able to follow up and improve students' academic performance. Thus, we can look at the results of the rest of the scenarios.

Table 4: Number of students in each cluster

Clustered Instances		
0	35	%19
1	24	%13
2	39	%21
3	73	%39
4	17	%9

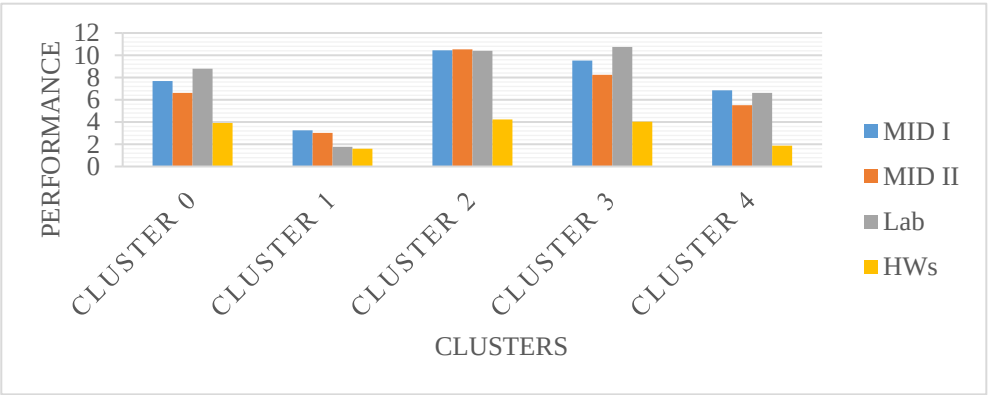

Fig. 8: Activities distribution in each cluster

4.4. Scenario IV

In this scenario, we distributed students into six groups (clusters) and thus nominated six teaching faculty members for academic advising.

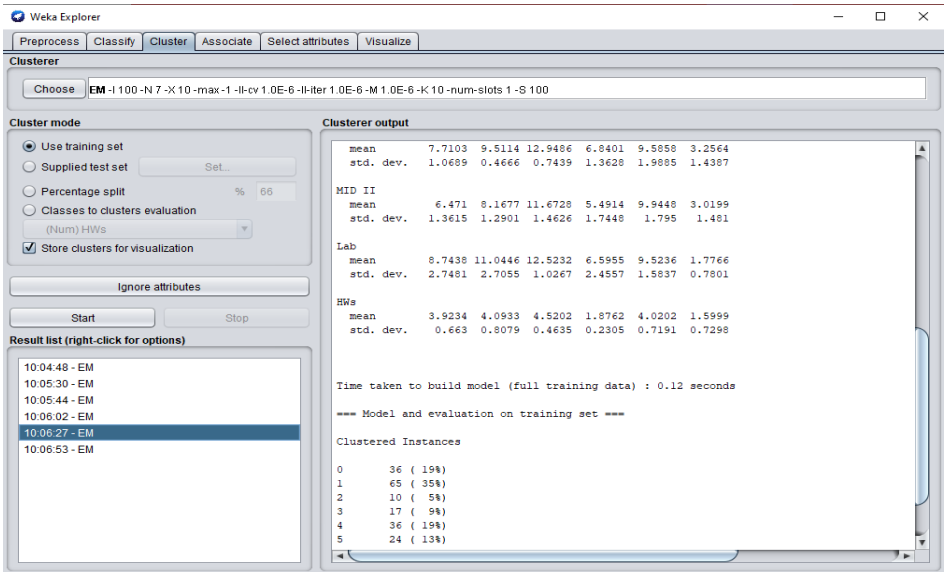

Fig. 9: Scenario IV results

The results show that the groups' average is close in some groups and for all activities and that; the standard deviation is relatively small in most of the semester activities, which creates groups with an acceptable degree of homogeneity. Also, as shown in Table 4, the number of students in each cluster is appropriate except in cluster 1, as the number of students reached 65, which consider being a large number. Although this number of clusters is suited to the faculty members available, we cannot accept this scenario due to covid-19 and classrooms that do not accommodate many students for collective academic advising.

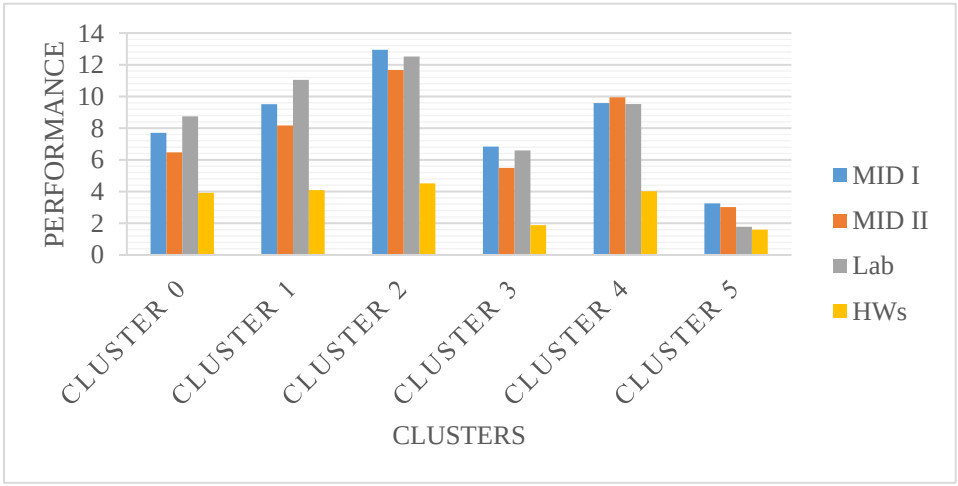

Fig. 9: Activities distribution in each cluster

Table 5: Number of students in each cluster

Clustered Instances		
0	36	%19
1	65	%35
2	10	%5
3	17	%9
4	36	%19
5	24	%13

4.5. Scenario V

In this scenario, we distributed students into seven groups (clusters) and thus nominated seven teaching faculty members for academic advising.

This division is appropriate for the number of students in all clusters. Also, the homogeneity is acceptable to a degree of deviation in all activities, enabling the application of health protocols, helping the faculty members control and manage the group quickly, and improving students' academic performance. Therefore, this scenario can be adopted because it suits the number of faculty members available.

5. Conclusion

Academic advising is one of the pillars of academic services offer to students during their study period in an institution. It is sensitive and tiring process for both students and teachers. Attention to this aspect is necessary, and it is one of the responsibilities of faculty members and they should try commensurate with the importance of this aspect.

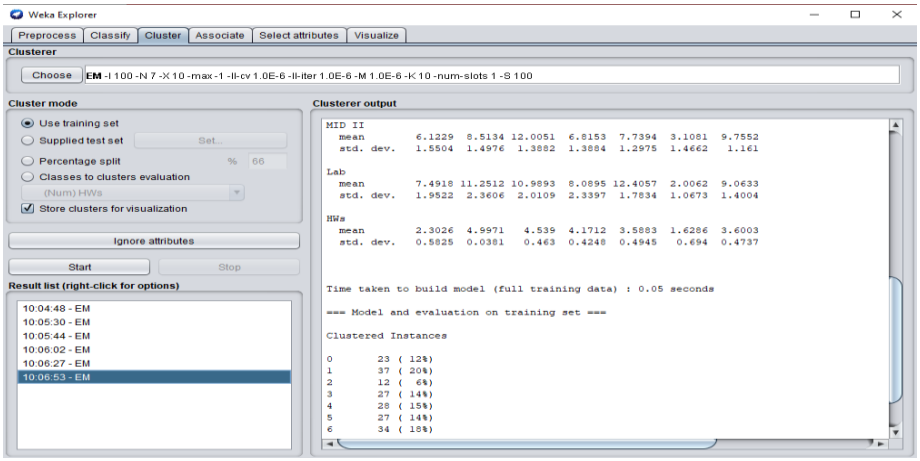

Fig. 10: Scenario V results

Fig. 11: Activities distribution in each cluster

Table 6: Number of students in each cluster

Clustered Instances		
0	23	%12
1	37	%20
2	12	%6
3	27	%14
4	28	%15
5	27	%14
6	34	%18

Other aspect is academic counselling process, which are affected by the academic performances. If a faculty member succeeds means, he should have an efficient platform for offering academic services. Additionally, Psychological and Social aspects also require effort from a faculty member, but for a limited group of students, to rely on modern technologies and use them to facilitate and develop the counselling process. Analyzing and evaluating the general performance of students is a necessary and continuous process. It gives indications of the student's activities general performance. It can be used to evaluate the faculty's performance based on students' academic performance and identify weaknesses and strengths in performance or the general trend of students' performance.

References

Abawajy, G. (2019). Learning evaluation methods in university based on data mining. *Asia-Pacific Journal of Educational Management Research*,4(3), 21-32. https://doi.org/10.21742/AJEMR.2019.4.3.03.

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *Int. J. Mod. Educ. Comput. Sci.*, 9(8), 9–15.

Ariouat, H., Cairns, A., Barkaoui, K., Akoka, J. & Khelifa, N. (2016). A two-step clustering approach for improving educational process model discovery. *In 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 38-43. https://doi.org/10.1109/wetice.2016.18.

Bartko, J. J. & Carpenter, W. T. (1976). On the methods and theory of reliability. *J NervMent Dis*, 163, 307-317.

Barros, B. & Verdejo. (2000). Analyzing student interaction processes in order to improve collaboration: the degree approach. *International Journal of Artificial Intelligence in Education*, 11, 221-241.

Chaitra, H. K. & Suneetha, K. R. (2022). Applying spectral clustering algorithm to group users by interest. *Journal of System and Management Sciences*, 12(1), 363–382. https://doi.org/10.33168/JSMS.2022.0125.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Deepika, K. & Sathvanaravana, N. (2018). Analyze and predicting the student academic performance using data mining tools. *Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 76-81, https://doi.org/10.1109/ICCONS.2018.8663197.

Florin, G. (2011). Data mining: Concepts, models and techniques. *Springer-Verlag. Berlin Heidelberg*.

Gao, X., Qian, J. & Gu, K. (2017). Research on urban green transformation based on grey theory and fuzzy clustering, *Journal of System and Management Sciences*, 7(2), 29-52.

Hussain, S., Atallah, R., Kamsin, A. & Hazarika, J. (2018). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. *In Computer Science Online Conference*, 196-211. https://doi.org/10.1007/978-3-319-91192-2_21.

Han, J., Kamber, M. & Pei, J. (2011). Data mining: Concepts and techniques 3rd edition. San Diego, CA, USA: Elsevier Science.

Hutson, B. (2013). Faculty development to support academic advising: rationale, components and strategies of support. *The journal of faculty development*, 27(3), 5.

Ian, H. W. & Eibe, F. (2005). Data mining: Practical machine learning tools and techniques. *Second Edition. Elsevier Inc*. San Francisco: USA.

Jiawei, H., Micheline, K. & Jian P. (2012). Data mining: concepts and techniques. *Third edition. Elsevier Inc: USA.*

Mohamed, O., Alistair, M., & David, T. (2018). Analysis and research based on international education performance index report. *Asia-Pacific Journal of Educational Management Research*, 3(2), 1-12, doi:10.21742/AJEMR.2018.3.2.01.

Nagy, H., Aly, W. & O. Hegazy. (2013). An educational data mining system for advising higher education students. *World Acad. Sci. Eng. Technol. Int. J. Inf. Sci. Eng*, 7(10), 175-179, 2013.

Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R. & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, 8(2), 221-228.

Paul, R., Gaftandzhieva, S., Kausar, S., Hussain, S., Doneva, R., & Baruah, A. (2020). Exploring student academic performance using data mining tools. *International Journal of Emerging Technologies in Learning (iJET)*, 15(08), 195–209. https://doi.org/10.3991/ijet.v15i08.12557.

Pascarella, E. T. & P. T. (2005). Terenzini, how college affects students: a third decade of research, 2nd ed., 2. San Francisco: Jossey-Bass.

Romero, N. M., Irisarri, M., Roth, P., Cauerhff, A., Samakovlis, C., & Wappner, P. (2008). Regulation of the Drosophila hypoxia-inducible factor-alpha Sima by CRM1-dependent nuclear export. *Mol. Cell. Biol*. 28(10), 3410-3423.

Smith, C. L. & Allen, J. M. Essential functions of academic advising: what students want and get. *NACADA Journal*, 26(1), 56-66.

Silver, D. (2016) Mastering the game of go with deep neural networks and tree search. Nature (London), 529(7587), 484-489.

Wong, J. C. F. & Yip, T. C. Y. (2020). measuring students' academic performance through educational data mining. *International Journal of Information and Education Technology*, 10(11), 797-804. https://doi.org/10.18178/ijiet.2020.10.11.1461.

Wong, Y. K. Machine learning algorithms using big data analysis. *International Journal of Computer Science Trends and Technology (IJCST)*, 9, [Online]. Available: www.ijcstjournal.org.

Kalyani, G. & Jaya, A. Lakshmi, Performance assessment of different classification techniques for intrusion detection. *Journal of Computer Engineering (IOSRJCE)*.

Hussein, A., Ahmad, F. K. & Kamaruddin, S. S. (2021). cluster analysis on covid-19 outbreak sentiments from twitter data using k-means algorithm, *Journal of System and Management Sciences*, 11(4), 167-189, DOI:10.33168/JSMS.2021.0409.

Yoon, J. & Joung, S. (2020). A big data based cosmetic recommendation algorithm. *Journal of System and Management Sciences*, 10(2), 40-52.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

Michael, J. A. B. & Gordon, S. L. Data mining techniques for marketing, sales, and customer relationship management. *Second edition. Wiley Publishing, Inc*. Indianapolis, Indiana: USA.

Quinlan, J. R. (1986). Introduction of decision trees. *Machine Learning*, 1, 81-106.

Remco, R., Eibe, F., Richard, K., Mark, H., Peter, R., Alex, S. & David, S. (2012). WAIKATO, Weka manual for version 3-6-8. Hamilton: New Zeland.