Adaptive Recommendation System in Education Data Mining using Knowledge Discovery for Academic Predictive Analysis : Systematic Literature Review

Siti Ummi Masruroh
Department of Informatics
UIN Syarif Hidayatullah
Jakarta, Indonesia
ummi.masruroh@uinjkt.ac.id

Dede Rosyada
Department of Islamic Education
UIN Syarif Hidayatullah
Jakarta, Indonesia
dede.rosyada@uinjkt.ac.id

Zulkifli
Department of Sociology
UIN Syarif Hidayatullah
Jakarta, Indonesia
zulkifli@uinjkt.ac.id

Sururin
Department of Islamic Education
UIN Syarif Hidayatullah
Jakarta, Indonesia
sururin@uinjkt.ac.id

Nanda Alivia Rizqy Vitalaya
Department of Informatics
UIN Syarif Hidayatullah
Jakarta, Indonesia
nanda.vitalaya17@mhs.uinjkt.ac.id

***Abstract*—Recent studies of predictive data mining for academics typically produce datasets and methods that allow them to explore data from the education system and use these methods to better understand students and the settings in which they learn. Many published datasets and methods on predictive data mining for academics are complex and different so that a comprehensive picture of the status of predictive data mining research is lost. The systematic review aims to identify and analyze trends, topics, datasets, and methods and answer research questions in the field of predictive data mining for academics between 2017 and 2020. Based on data search, 14 related studies were selected to be further identified. The analysis of the selected studies shows that eight topics are the focus and trend topics, namely classification, recommender systems, educational data mining, knowledge discovery databases, student prediction, academic performance, predictive analysis, and predictive analysis. The use of datasets in the selected studies shows that 100% of studies use private datasets and 0% of studies use public datasets. Of the twenty-one methods, five of the most widely used methods in the field of predictive data mining in education were identified. Research identifies no strong consensus on which algorithm performs best when the study is viewed individually. Therefore, prediction studies in the field of data mining for academics try to be as optimal as possible in choosing algorithmic modeling to produce more optimal predictions.**

Keywords— Educational data mining, Systematic literature review, Predictive analysis, Adaptive recommendation system

I. INTRODUCTION

Educational Data Mining (EDM) is a growing multidisciplinary research field, exploring data that comes from the education system and using these methods to better understand students, and the settings in which they learn [1][2]. EDM can be used for scientific inquiry and system evaluation, define student model parameters, inform domain models, create diagnostic models, generate reports and alerts for instructors, students, and other stakeholders, and recommend resources and activities [3].

Some studies in the EDM field are devoted to the statistical analysis of log data (journals listing user and program actions) of learning management systems and to looking for relationships between measured values and traditional indicators for educational systems in particular academic performance [4]. EDM creates new opportunities to analyze, collect, visualize and present student data that is considered highly relevant for an intelligent education system [4].

Much predictive data mining for education studies that examine published datasets, methods, and frameworks are different and complex so that the overall picture of predictive data mining for education research currently available is lost. The systematic review aims to detect literature relevant to research statements between 2017 and 2020.

This paper is composed of 4 parts. Introduction in section 1, research methodology in section 2, results and answers in section 3, and conclusions in section 4.

II. METHODOLOGY

A. Review Method

A systematic approach was taken to review the literature on educational data mining[5]. SLR is a step of identifying, evaluating and interpreting all available research to examine the extent to which empirical evidence can support or contradict theoretical hypotheses and can assist in the creation of new hypotheses [6]. A systematic review seeks to identify and report studies that do not support the research hypothesis and identify and report similar studies to support the research [6]. There are 3 stages of SLR as shown in Figure 1, namely Plan Review, Conduct Review, and Document Review.

Fig. 1. Systematic Literature Review Process

B. *Research Questions*

The research question (RQ) is the most important part of a systematic review [6] and is determined to keep the review focused on the research objectives.

TABLE I. RESEARCH QUESTION ON LITERATURE REVIEW

ID	Research question	Motivation
RQ1	Which journal is the most significant academic data mining prediction journal?	Identify the most significant journals in the field of predictive data mining for academics
RQ2	What types of research topics do researchers in the field of predictive data mining choose for academics?	Identify research topics and trends in the field of predictive data mining for academics
RQ3	What types of datasets are most used in the field of predictive data mining for academics?	Identify datasets commonly used in the field of predictive data mining for academic purposes
RQ4	What types of methods are used in the field of predictive data mining for academic purposes?	Identify opportunities and trends for predictive data mining methods for academics
RQ5	What types of methods are most often used for predictive data mining for academics?	Identify the most frequently used methods for predictive data mining for academics
RQ6	Which method performs best when used for data mining prediction for academics?	Identify the best method of predictive data mining for academics
RQ7	What methods are proposed in data mining prediction for academics?	Identification proposed methods of data mining for academics

After assessing the quality of learning, extraction was carried out on methods and datasets for predictive data mining for academics. Then, an analysis of the methods and datasets was carried out to determine which ones were and were not included in the prediction of data mining for academics (RQ2 to RQ5). RQ2 to RQ5 are the main research questions whereas RQ1 helps in evaluating the main study. RQ1 provides a summary of other research areas in predictive data mining for academics.

Figure 2 shows the rationale map of the systems literature review. Identifying the methods and datasets used in predictive data mining for academics is the main objective of writing this systematic literature review.

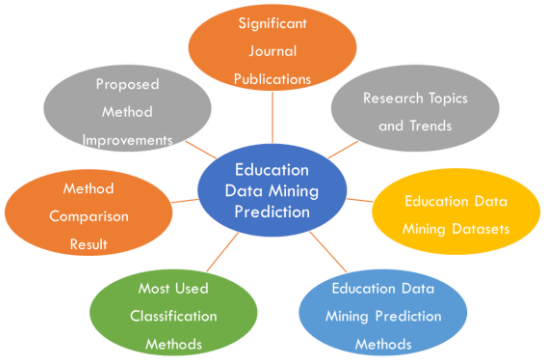

Fig. 2. Basic Mind Map of the SLR on Education Data Mining Prediction

C. *Search Strategy*

The search process consists of several activities, namely selecting the digital library. Before searching, to improve in finding relevant articles a database of appropriate keywords is needed. Popular literature databases were searched to obtain the maximum possible study. Below is a list of digital databases in use:

- IEEE Xplore (ieeexplore.ieee.org)
- Springer (springerlink.com)
- Elsevier (elsevier.com)
- ScienceDirect (sciencedirect.com)
- MDPI (mdpi.com)
- Scopus (scopus.com)

D. *Data Extraction*

Selected studies are extracted to collect answers to research questions. There are 3 properties used to answer the research questions which are shown in Table 2.

TABLE II. DATA EXTRACTION PROPERTIES MAPPED TO RESEARCH QUESTION

Property	Research Question
Researches and Publications	RQ1
Research Trends and Topics	RQ2
Education Data Mining Datasets	RQ3
Education Data Mining Methods	RQ4, RQ5, RQ6, RQ7

E. *Study Quality Assessment and Data Synthesis*

Assessments of the quality of studies can be used to guide the direction of interpretation and determine the strength of conclusions. Data synthesis aims to gather

evidence to answer research questions. A piece of evidence may have little strength, but if all the available evidence is gathered it can be of great strength to support research. Several visualization media, such as pie charts, bar charts, and tables are used to improve the distribution presentation of the Educational Data Mining method and its accuracy.

F. Threats to Validity

The aim of this review is to analyze studies on academic predictions using educational data mining. In this review, searches were not based on a manual reading of all paper titles published in the journal. This suggests that this review has excluded some papers in educational data mining from some conference proceedings or journals. This review also uses conference proceedings because based on the experience reports most of the studies published in conference proceedings.

III. RESEARCH RESULTS

A. Significant Journal Publications

In this literature review, an analysis of predictive performance in the field of data mining education was carried out on 14 studies. The distribution of the research is presented to show how the researcher's interest in the field of data mining education has changed over time. The distribution of the studies over the years is shown in Figure 3.

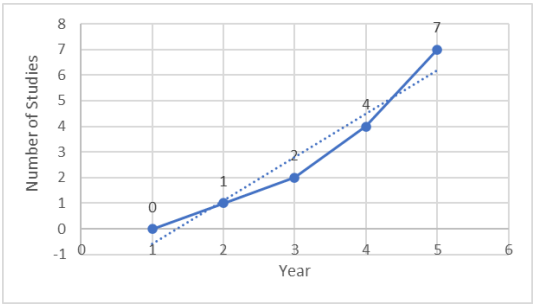

Fig. 3. Distribution of Selected Studies over the Years

According to the main study conducted, the publication of predictive data mining journals for academics is shown in Figure 4.

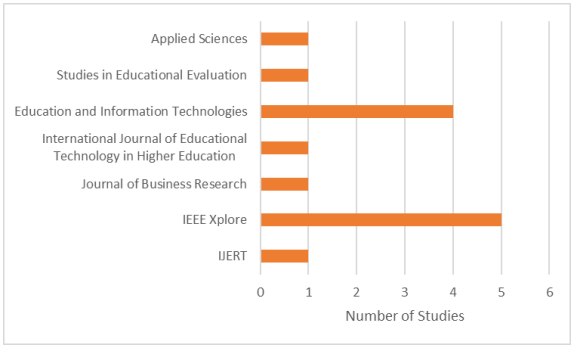

Fig. 4. Journal Publications and Distribution of Selected Studies

The Scimago Journal Rank (SJR) value and the Q1 category of educational data mining journals are shown in Table 3 which is arranged based on its SJR value.

TABLE III. SCIMAGO JOURNAL RANK (SJR) OF SELECTED JOURNALS

No	Title	Journal Publications	SJR	Q Category
1	Educational Data Mining: Predictive Analysis Of Academic Performance Of Public School Students In The Capital Of Brazil	Journal of Business Research	1.87	Q1 in Marketing
2	Prediction of Student's performance by modeling small dataset size	International Journal of Educational Technology in Higher Education	1.07	Q1 in Computer Science Applications
3	Adaptive recommendation system using machine learning algorithms for predicting student's best academic program	Education and Information Technologies	0.78	Q1 in Education
4	Personalization of study material based on predicted final grades using multi-criteria user-collaborative filtering recommender system	Education and Information Technologies	0.78	Q1 in Education
5	Data mining approach to predicting the performance of first year student in a university using the admission requirements	Education and Information Technologies	0.78	Q1 in Education
6	Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment	Studies in Educational Evaluation	0.78	Q1 in Education
7	Behind The Scenes Of Educational Data Mining	Education and Information Technologies	0.78	Q1 in Education
8	Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review	Applied Sciences	0.42	Q1 in Engineering (miscellaneous)

The significance of the international journals selected by researchers is shown in Table 4.

TABLE IV. INTERNATIONAL OF SELECTED JOURNALS

No	Title	Journal Publications
1	College Recommendation System	International Journal of Engineering Research & Technology (IJERT)
2	Using Educational Data Mining Techniques to Predict Student Performance	international conference on electrical and computing technologies and applications (IEEE)
3	Using Classification Data Mining Techniques for Students Performance Prediction	The 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering
4	An Autonomous Courses Recommender System	2020 International Conference in Mathematics, Computer

	For Undergraduate Using Machine Learning Techniques	Engineering and Computer Science
5	Educational Data Mining: Current Problems and Solutions	IEEE Xplore
6	A Survey on Educational Data Mining [2014-2019]	IEEE Xplore

B. Research Topics in the Educational Data Mining

Educational Data Mining is a field that uses statistical algorithms, machine learning, and data mining (DM) of various types of educational data [6]. Based on the analysis that has been done, 8 topics are the focus in the field of Educational Data Mining:

1. Grouping individual items based on the characteristics of the topic (Classification)
2. Create a recommendation system based on student trends (Recommender System)
3. Applying data mining techniques for educational data analysis (Educational Data Mining)
4. Find patterns and relationships from large data sets (Knowledge Discovery Database)
5. Find patterns and relationships from large data sets (Student Prediction)
6. Knowing the level of educational performance (Academic Performance)
7. Analyze data to predict something (Predictive Analysis)
8. Processing of data obtained in a particular field of study (Learning Analytics)

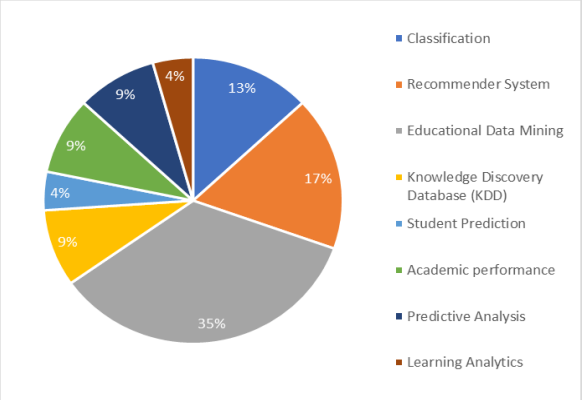

Fig. 5. Distribution of Research Topics

Figure 5 shows the distribution of research topics on predictive data mining for academics. Research studies show that the most discussed topics from 14 journals are Educational Data Mining (35%), Recommender Systems (17%), Classification (13%), Knowledge Discovery (9%), Academic Performance (9%), Predictive Analysis (9%), Student Prediction (4%) and Learning Analytics (4%). So, it can be concluded that most of the authors use Educational Data Mining as their research topic.

C. Datasets Used for Educational Data Mining

Dataset is a collection of data that is used for specific learning purposes [7]. In this literature review, based on 14 studies that have been analyzed, almost all of the datasets used are Private (100%) while Public (0%) is shown in Figure 6. Where private school-owned datasets are not distributed as public data sets.

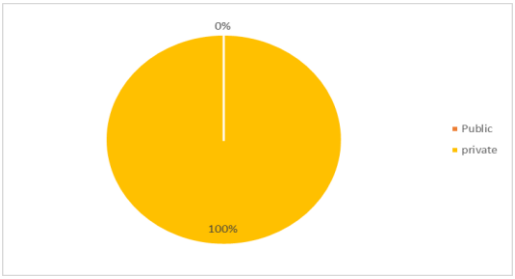

Fig. 6. Total Distribution in Datasets

Figure 7 presents the number of datasets used over the past few years based on the 14 studies selected. The use of private datasets has increased from 2017 to 2019.

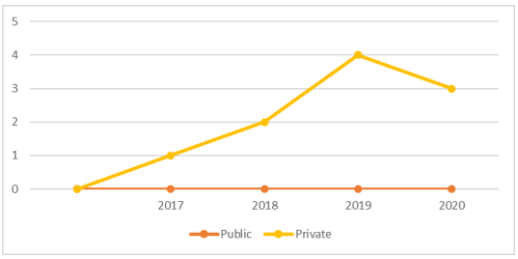

Fig. 7. Distribution of Private and Public Dataset

D. Methods Used in Educational Data Mining

In Figure 8, it is presented that twenty-one methods have been applied and proposed as the best method for predictive data mining for academics.

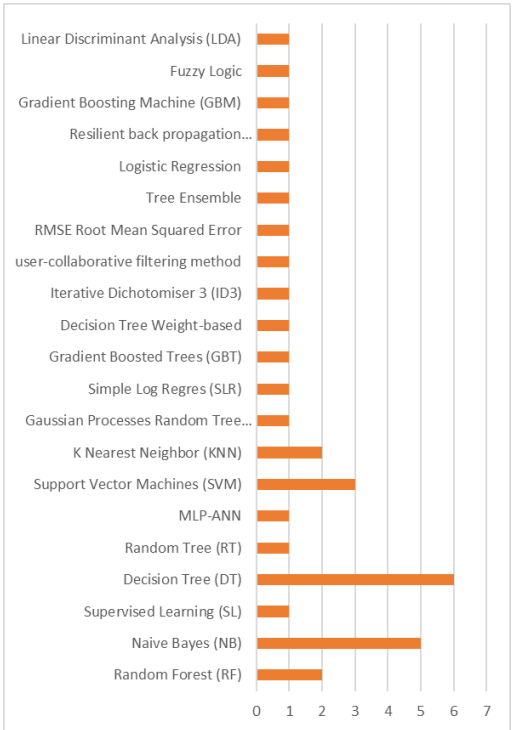

Fig. 8. Methods Used in Educational Data Mining

E. *Most Used Methods in Educational Data Mining*

Of the twenty methods presented in Figure 8 in section D, there are 5 educational data mining methods that are most widely applied in this research. These methods include:

1. K-Nearest Neighbor (k-NN)
2. Support Vector Machine (SVM)
3. Decision Tree (DT)
4. Naïve Bayes (NB)
5. Random Forest (RF)

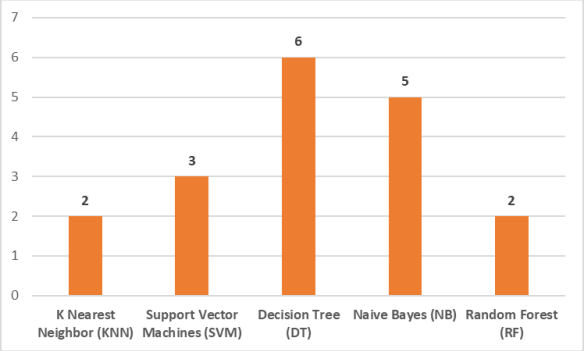

Fig. 9. Most Used Methods in Educational Mining

Decision tree and Naïve Bayes are the two most commonly used methods. Both methods have adopted 52% of the studies that have been selected in Figure 10.

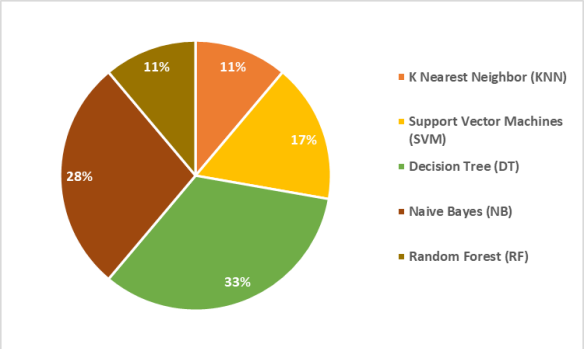

Fig. 10. Distribution of Studies over Type of Methods

F. *Method Perform Best for Educational Data Mining*

Many studies in predictive data mining for academics report the comparative performance of the modeling algorithms it uses, and there is no strong consensus on which algorithms perform best when these studies are viewed individually.

On research by [8], NB and SVM were found to have the highest recommendation accuracy among all classifiers used (KNN and DT), where both have 99.94%, followed by KNN with an accuracy of 99.8 7%, and DT is the smallest but still with an acceptable accuracy of 98.01%. Meanwhile, if we take into account the cognitive characteristics of students, the prediction accuracy increases using DT [9][10]. In addition, KNN is the best prediction system algorithm compared to RF and SVM in one of the departments at Al-Azhar University [11]. However, in the study by [12], the highest accuracy was achieved when Random Forest was used to handling omitted values.

Studies show that although most of the clustering algorithms are stated in terms of optimal criteria, there is generally no guarantee that the optimal solution has been obtained [9]. In addition, no single learning algorithm provides the best results for all faculty departments [11]. Each algorithm used has the best performance for each department or faculty [11]. Therefore, each department or faculty has different best learning algorithms.

G. *Proposed Method Improvements for Educational Data Mining*

Based on the reviews that have been done, the research proposed for data mining education is an adaptive recommendation system using a knowledge discovery database and implementing the most algorithms used in EDM are the Decision Tree and Naive Bayes.

Adaptive recommendation system uses machine learning. The results of the study analyzed the most algorithms used by researchers, namely the Decision Tree ([14], [10], [13], [15], [16]), Naïve Bayes ([9], [14], [17], [18], [19]), Support Vector Machines ([12], [17], [18]), K Nearest Neighbor ([17], [20]) and Random Forest ([15], [18]) . In addition, studies that apply educational data mining are [13], [16], [19], [21] - [23]. So that the proposed research will use the Decision Tree and Naïve Bayes algorithms for academic prediction analysis.

IV. Conclusion And Future Works

The existence of this literature review aims to identify and analyze the topics and methods used in predictive analysis of educational data mining. This systematic review is used as an assessment process, identifying and interpreting existing studies to provide answers to research questions.

Analysis of the selected studies shows that predictive data mining for academics currently focuses on eight topics, namely classification, recommender systems, educational data mining, knowledge discovery databases, student prediction, academic performance, predictive analysis, and predictive analysis.

Twenty-one different methods are applied to data mining prediction for academics. Of the twenty-one methods, there are five methods that are most often used, including K-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), and Random Forest (RF).

Studies in predictive data mining for academics show no strong consensus on which algorithms perform best when these studies are viewed individually. Although most of the clustering algorithms are stated in terms of optimal criteria, there is no guarantee that an optimal solution has been obtained [9]. In addition, there is no single learning algorithm that gives the best results for all faculty departments, where each algorithm used has the best performance for each department or faculty [11]. Therefore, predictive studies in the field of data mining for academics try to be as optimal as possible in choosing algorithmic modeling to produce more optimal predictions.

REFERENCES

[1] T. Calders and M. Pechenizkiy, "Introduction to the special section on curriculum," *Waikato J. Educ.*, vol. 14, no. 1, 2012, doi: 10.15663/wje.v14i1.244.

[2] R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009 : A Review and Future Visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–16, 2009.

[3] O. Scheuer and B. M. Mclaren, "Educational Data Mining," pp. 1075–1079, 2012.

[4] Y. S. Mitrofanova, A. A. Sherstobitova, and O. A. Filippova, *Modeling smart learning processes based on educational data mining tools*, vol. 144. Springer Singapore, 2019.

[5] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, vol. 1, no. 1, pp. 1–16, 2007, doi: 10.3923/jse.2007.1.12.

[6] Kitchenham, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *IEEJ Trans. Ind. Appl.*, vol. 126, no. 5, pp. 589–598, 2007, doi: 10.1541/ieejias.126.589.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[8] Elsevier, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Oliver Walter, 2019.

[9] M. Isma'Il, U. Haruna, G. Aliyu, I. Abdulmumin, and S. Adamu, "An Autonomous Courses Recommender System for Undergraduate Using Machine Learning Techniques," *2020 Int. Conf. Math. Comput. Eng. Comput. Sci. ICMCECS 2020*, 2020, doi: 10.1109/ICMCECS47690.2020.240882.

[10] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Appl. Sci.*, vol. 10, no. 3, 2020, doi: 10.3390/app10031042.

[11] S. Sultana, S. Khan, and M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," *Int. J. Electr. Eng. Educ.*, vol. 54, no. 2, pp. 105–118, 2017, doi: 10.1177/0020720916688484.

[12] M. Ezz and A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program," *Educ. Inf. Technol.*, vol. 25, no. 4, pp. 2733–2746, 2020, doi: 10.1007/s10639-019-10049-7.

[13] N. Ketui, W. Wisomka, and K. Homjun, "Using classification data mining techniques for students performance prediction," *ECTI DAMT-NCON 2019 - 4th Int. Conf. Digit. Arts, Media Technol. 2nd ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng.*, pp. 359–363, 2019, doi: 10.1109/ECTI-NCON.2019.8692227.

[14] V. Jain, M. Gupta, J. Kevadia, and P. K. Shinde, "College Recommendation System," *Int. J. Eng. Res. Technol.*, vol. 5, no. 01, pp. 1–3, 2017.

[15] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527–1543, 2019, doi: 10.1007/s10639-018-9839-7.

[16] F. Martínez-Abad, A. Gamazo, and M. J. Rodríguez-Conde, "Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment," *Stud. Educ. Eval.*, vol. 66, no. December 2019, 2020, doi: 10.1016/j.stueduc.2020.100875.

[17] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0160-3.

[18] B. Al Breiki, N. Zaki, and E. A. Mohamed, "Using Educational Data Mining Techniques to Predict Student Performance," *2019 Int. Conf. Electr. Comput. Technol. Appl. ICECTA 2019*, 2019, doi: 10.1109/ICECTA48151.2019.8959676.

[19] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, no. February, pp. 335–343, 2019, doi: 10.1016/j.jbusres.2018.02.012.

[20] D. F. Murad, Y. Heryadi, S. M. Isa, and W. Budiharto, "Personalization of study material based on predicted final grades using multi-criteria user-collaborative filtering recommender system," *Educ. Inf. Technol.*, vol. 25, no. 6, pp. 5655–5668, 2020, doi: 10.1007/s10639-020-10238-9.

[21] S. Kovalev, A. Kolodenkova, and E. Muntyan, "Educational Data Mining: Current Problems and Solutions," *2020 5th Int. Conf. Inf. Technol. Eng. Educ. Inforino 2020 - Proc.*, 2020, doi: 10.1109/Inforino48376.2020.9111699.

[22] A. Hicham, A. Jeghal, A. Sabri, and H. Tairi, "A Survey on Educational Data Mining [2014-2019]," *2020 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2020*, 2020, doi: 10.1109/ISCV49265.2020.9204013.

[23] Y. Feldman-Maggor, S. Barhoom, R. Blonder, and I. Tuvi-Arad, "Behind the scenes of educational data mining," *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 1455–1470, 2021, doi: 10.1007/s10639-020-10309-x.