IJESIM
International Journal of Educational Studies in Mathematics
EDUGARDEN

Predicting Students' Performance in Mathematics through Educational Data Mining Techniques during the Transition from Primary to Lower Secondary School

Senad Orhani

University of Prishtina "Hasan Prishtina", Republic of Kosovo

ABSTRACT

The ability to predict students' performance tendency is very important to improve their learning skills. For this, Educational Data Mining (EDM) is a more active research field. It aims to find useful information from the educational data set using data extraction techniques. The most important EDM tasks for this study are to predict student performance. The overall goal of EDM is to understand how students learn and identify those aspects that can improve teaching and learning aspects. This paper reviews some existing research and identifies other future pathways based on EDM knowledge. Therefore, the purpose of this study is to describe how EDM techniques can help math teachers identify students who are most likely to fail and then take appropriate action, and change strategies for it to improve the performance of their students in this area.

ARTICLE INFO

Article History:
Received: 28.05.2022
Received in revised form: 17.08.2022
Accepted: 22.08.2022
Available online: 10.09.2022
Article Type: Review

Keywords: data mining, EDM, mathematics, performance, and students

1. Introduction

Education Data Mining in (EDM) is an interdisciplinary research field created as an application for data mining in the field of education. This field is a new discipline with high potential for every participant in the educational process (Zorić, 2020). Data extraction techniques were developed to automatically detect hidden knowledge and recognize patterns from data. It uses a variety of methods and techniques from statistics, artificial intelligence, neural networks, database systems, machine learning, pattern recognition, data visualization, knowledge acquisition, information theory, extraction of data and data analysis, to analyze data collected during teaching and learning.

Data mining has been used in various fields such as marketing, banking, telecommunications, health and medicine, industry, science and engineering, etc. (Zorić, 2020). Recently, one of these areas is the educational environment. As a result of the application of data extraction techniques in education, the field of data extraction education EDM has emerged. Educational Data Mining is defined as an evolving discipline, related to the development of methods for exploring unique types of data coming from educational settings, and the use of these methods to better understand students and the environments in which they learn (Idil, Narli and Aksoy, 2016).

EDM is the process of transforming raw data from large educational databases into useful and meaningful information, which can be used for a better understanding of students and their learning conditions, improving teaching support as well as for decision-making in education systems. Educational Data Mining is used to detect important phenomena and to resolve educational issues that occur in the context of teaching and learning. This study provides a systematic review of the

[1] Corresponding author's address: University of Prishtina "Hasan Prishtina", Republic of Kosovo
e-mail: senad.orhani@uni-pr.edu

DOI: https://doi.org/10.17278/ijesim.1122751

Educational Data Mining literature in the field of mathematics (Silva and Fonseca, 2017). EDM is an emergency discipline at the crossroads of data mining and pedagogy. On the one hand, pedagogy contributes to the internal knowledge of the learning process. On the other hand, data mining enhances information analysis and modelling techniques (Kumar and Vijayalakshmi, 2011).

Objectives can be identified to use EDM as a technology in the field of education. One of them is the pedagogical objectives, to help students improve in their achievements, to design the course content in a better way, etc. (Kumar and Vijayalakshmi, 2011).

The application of knowledge mining techniques in education systems in order to improve learning can be seen as a formative assessment technique. Data mining techniques can reveal useful information that can be used in formative assessment to help teachers establish a pedagogical basis for decisions when designing or modifying a learning environment or approach (Idil, Narli, and Aksoy, 2016).

2. Literature

EDM is an important research area (Koedinger, et al., 2016) in future analysis, modelling and decision making based on educational data. EDM is used to better understand data sets, students and their learning process. It is also used to develop practical approaches to provide useful information to students. EDM has attracted attention as a research area for researchers worldwide in recent years. Romero and Ventura (2010) studied educational data by conducting a comprehensive literature search. In their work, they cited EDM as a recursive twist involving hypothesis building, testing, and performance improvement (Romero and Ventura, 2010).

Predicting students' academic performance is one of the main topics of educational data mining (Fan, Liu, Chen and Ma, 2019; Guruler and Istanbullu, 2014). The main purpose of the forecasting approach is to help researchers extract information about a particular variable of interest from a set of other (predictive) variables, and also to explore which constructs in a data set have an important role in predicting another (Baker and Inventado, 2014).

Baradwaj and Pal (2011) conducted a survey on a group of 50 students enrolled in a specific course program for a period of 4 years, with multiple performance indicators, including "Previous Semester Grades", "Test Scores", "Seminar Performance", "Tasks", "General Ability", "Participation", "Laboratory Work", and "Final Semester Decisions". They used the ID3 decision tree algorithm to eventually build a decision tree, and if-then rules, which will ultimately help teachers as well as students better understand and predict student performance at the end of the semester (Baradwaj and Pal, 2011).

Pandey and Pal (2011) conducted data mining research using the Naïve Bayes classification to analyze, classify, and predict students as performing or non-performing. Naïve Bayes classification is a simple probability classification technique, which assumes that all attributes given in a data set are independent of each other. These authors conduct this research on a sample of enrolled student data to obtain a postgraduate degree in computer applications. The research was able to classify and predict to a certain extent the grades of students in the next year, based on their grades in the previous year. Their findings can be used to assist students in their future education in many ways (Pandey and Pal, 2011).

Bhardwaj and Pal (2012) identified their main objectives of this study as: Generating a data source of predictive variables; Identify various factors that affect the behavior and performance of student learning during the academic career; Construct a forecasting model using data mining classification techniques based on the identified predictor variables; and Evaluation of the model developed for higher education students studying at universities. They found that the most influential factor for a student's performance is his grade in high school, which tells us that those students who have performed well in their high school will undoubtedly have good results in their Bachelor studies. Furthermore, it was found that the place of residence, the teaching environment, the mother's qualifications, other student habits, the annual family income and the student's family status all contribute greatly to the student's educational performance, so can predict his / her grade or overall performance if basic personal and social knowledge has been gathered about him / her (Bhardwaj and Pal, 2012).

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding (Baradwaj and Pal, 2011).

3. Objectives of the Study

It is very important to determine the determinants / traits of students' success in the mathematics learning process. Since the features that affect student success are well known, it can be assumed that student success will increase by taking precautions for these features or by changing current conditions. Therefore, the objective of EDM is to identify and extract from the data new and potentially valuable hidden knowledge. The EDM techniques for this study aim to develop a model that can draw conclusions on:

- improving the quality of the learning process in the subject of mathematics;
- improving the successful completion of the course;
- supporting students in selecting topics from this field;
- profiling of students for additional training;
- finding problems that lead to math failure;
- the intention and orientation of students to study mathematics;
- development of curricula for the field of mathematics;
- predicting student performance;

4. Problem / Purpose of Study

Many sixth graders students are unprepared to make a successful transition from primary school to lower secondary school, and are also unprepared to face some of the challenges in the subject of mathematics, in which courses pass to an advanced level of knowledge, which can be very stressful for them. Factors influencing the proper performance of mathematics students and the quality of reasoning in mathematics remain an open topic. In fact, in the classrooms it is noticed that some students have good performance, while others do not, despite having the same conditions during the learning process. Therefore, discovering current factors in the learning and reasoning process is key to helping promote the highest levels of success in their future lives and professions.

The education system does not deal with failures in the subject of mathematics, does not warn students about the lack of basic knowledge, does not identify the weak student and does not inform the mathematics teachers. Therefore, the existing education system in our institutions is still primitive and unable to identify the most appropriate methods for resolving issues in this area.

Although, there are abundant studies in the international literature that have discussed educational data mining, and there is agreement on the importance of data extraction in the educational context; the use of data mines to improve the education system is still relatively new. Therefore, this existing research is very specific and most current research is limited in this context. The purpose of this study is to describe how EDM techniques can help math teachers identify students who are most likely to fail and then take appropriate action, and change strategies to improve performance. of their students in the field.

5. Materials and methods

Data Mining is the process of finding anomalies, patterns, and correlations within large data sets to predict results. Using a wide range of techniques, you can use this information to increase student success and performance.

Analyzing students' data and information to classify students, or to create decision trees or association rules, to make better decisions or to enhance student's performance is an interesting field of research, which mainly focuses on analyzing and understanding students' educational data that indicates their educational performance, and generates specific rules, classifications, and predictions to help students in their future educational performance (Saa, 2016).

The data mining process can be divided into these four main stages (Stedman and Hughes, 2022):

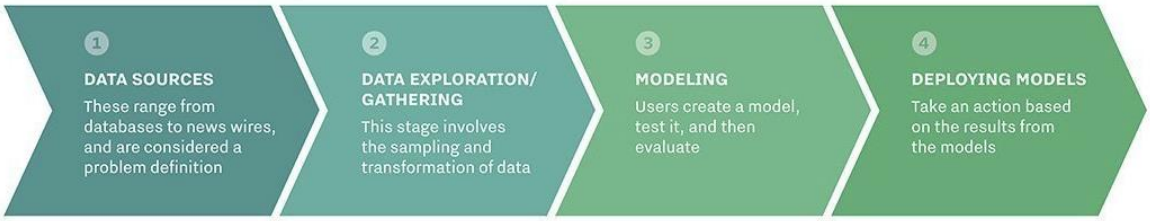
1
DATA SOURCES
These range from databases to news wires, and are considered a problem definition
2
DATA EXPLORATION/ GATHERING
This stage involves the sampling and transformation of data
3
MODELING
Users create a model, test it, and then evaluate
4
DEPLOYING MODELS
Take an action based on the results from the models

Figure 1. Four stage of data mining (Stedman and Hughes, 2022)

Data collection. Relevant data for an analytics application is identified and collected. Data can be stored in various source systems, a data warehouse or a data lake, an increasingly common repository in large data environments containing a mix of structured and unstructured data. External data sources can also be used. Where the data comes from, a data scientist often shifts it to a data lake for the remaining steps in the process.

Data preparation. This stage involves a step of series to get the data ready to be mined. It starts with exploring, profiling, and pre-processing data, followed by data cleansing work to fix errors and other data quality issues. Data transformation is also done to make data sets stable, unless a data scientist wants to analyze raw unfiltered data for a given application.

Data mining. Once the data has been prepared, a data scientist selects the appropriate data mining technique and then applies one or more algorithms to perform the mining. In machine learning applications, algorithms usually need to be trained in sample datasets to retrieve the required information before running against the complete data set.

Data analysis and interpretation. Data mining results are used to create analytical models that can help foster students' decision-making and other actions. The data scientist or another member of a data science team should also communicate the findings to teachers, often through data visualization and the use of data display techniques.

Data mining can be defined as applications of different algorithms to identify patterns and relationships in a data set. Data mining techniques can be classified as follows (Idil, Narli and Aksoy, 2016):

- Clustering
- Classification and regression
- Association rules

In clustering, the goal is to split the data into clusters, such that, there is homogeneity within clusters and heterogeneity between clusters (Siemens and Baker, 2014). In educational research, clustering procedures have been used to find patterns of effective problem-solving strategies in exploratory computer-based learning environments (He, 2013; Beal, Qu and Lee, 2006; Amershi and Conati, 2009). In regression, the goal is to develop a model that can infer or predict something about a data set. In regression analyses, a variable is identified as the predicted variable and a set of other variables as the predictors (similar to dependent and independent variables in traditional statistical analyses) (Siemens and Baker, 2014). In association rules mining, the goal is to extract rules of the form if-then, such that if some set of variable values is found, another variable will generally have a specific value (Siemens and Baker, 2014).

In the educational data mining method, predictive modelling is commonly used in predicting student performance. To construct predictive modelling, several tasks are used, which are classification, regression and categorization. The most well-known task for predicting student performance is classification. There are several sub-task algorithms that have been applied to predict student performance. Among the algorithms used are Decision Tree, Artificial Neural Networks, Naive Bayes, K-Neighbor and Vector Machine Support (Shahiri, Husain, and Rashid, 2015).

Decision Tree – Student performance appraisal is based on features derived from data recorded in a web-based education system. Examples of the data set for students from primary school to lower secondary in the subject of mathematics can be the final grades of students, the cumulative final grade in this field and the grades obtained in the preliminary tests of the subject of mathematics. All of these datasets can be studied and analyzed to find key attributes or factors that may affect student performance in the lower middle grades. Next, the appropriate data extraction algorithm will be investigated to predict student performance.

Neural network – is another popular technique used in educational data mining. The advantage of the neural network is that it has the ability to detect all possible interactions between predictor variables. The neural network can also make a complete discovery without any doubt even in complex nonlinear relationships between dependent and independent variables. Therefore, the neural network technique has been chosen as one of the best methods of prediction. This technique can be used for our study to predict student performance. Attributes analyzed by the Neural Network may be student data from grades 1-5 in the subject of mathematics, students' attitudes towards the subject of mathematics, and general academic performance in the natural sciences.

Naive Bayes Algorithm – it is also an option for researchers to make a prediction. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For our study, an excellent student in the subject of mathematics can be considered a talented student if he prefers mathematics, has high grades in mathematics in the primary school and with good feedback from the teacher. A Naive Bayesian classifier considers each of these traits to contribute independently to the probability that this student will be a successful lower secondary student in mathematics, regardless of any possible correlation between performance traits in general subjects.

K-Nearest Neighbor – This method takes less time to identify student performance as a poor student, average student, good student and excellent student. K-Nearest Neighbor where it gives a good accuracy in evaluating the detailed model for the student's progress in the lower secondary school in the subject of mathematics. For this study we can have a student from the previous year who is with similar features and with almost the same performance in the subject of mathematics. So, for this identification, we can use the KNN algorithm, as it works on a similarity measure. The KNN model will find similar features of the new data set with student performance and based on the most similar features will place it in the category as a poor student, average student, good student or excellent student in the subject of mathematics in the lower secondary school.

Support Vector Machine – is a supervised learning method used for classification. They are supervised learning models with accompanying learning algorithms that analyze data for classification and regression analysis. SVM are one of the most powerful forecasting methods, based on statistical learning frameworks. Here you can use parameters such as participation in math activities, competitions, school quizzes and check how students performed on the tests, even though they participated in such activities. Suppose a student who has performed well in competitions, math activities, also received a high average, then he / she will be labelled with high performance even though there are other students who have a higher overall average than this student.

6. Discussion and Conclusion

The way students learn and the assessment of their knowledge can help mathematics teachers to revise current curricula and pedagogical practice in general. By analyzing the data of students from primary school for the subject of mathematics, as well as by analyzing the importance of the influence of individual variables, different data models can be used as support for decision-making in the planning of activities in this area, thus contributing to a successful study and in increasing the quality and success in mathematics. The results from the EDM analysis can help principals, teachers and especially schools to communicate the results of success or failure more effectively. Educational Data Mining allows to conclude that the found characteristics are useful for teachers to assist in the teaching process, for the fact that, this will enable the identification of students who need reinforcement and feedback on the most complex topics of mathematics, which affect the reduction of failure rate in mathematics related subjects.

Predicting student performance in mathematics during the transition from primary to lower secondary allows teachers to distinguish non-engaging students based on their actions and activities in that subject. It also helps identify students in difficulty learning the subject of mathematics and this enables them to find additional measures to improve their chances of passing during their study curriculum. Therefore, this study presented different models of Educational Data Mining techniques used to predict student success during this transition. Since there are many approaches used to predict student performance, the study chose the data classification technique, introducing the algorithms of Decision Tree, Artificial Neural Networks, Naive Bayes, K-Neighbor and Support Vector Machine. In our future studies, we also intend to use results and detailed data to predict students' academic performance during the transition from primary teacher to math teacher.

Acknowledgment

I am grateful to the students of the elementary and lower secondary school "Heronjtë e Lumës" in Prizren / Kosovo that with their inspiration this research was conducted. Without their participation the search would not have been possible.

References

Amershi, S., & Conati, C. (2009). *Combining unsupervised and supervised machine learning to build user models for exploratory learning environments.* Journal of Educational Data Mining, 1(1), 71-81.

Baker, R. S., & Inventado, P. S. (2014). *Educational data mining and learning analytics," in Learning Analytics: From Research to Practice, eds J. A. Larusson and B. White.* Springer, doi: 10.1007/978-1-4614-3305-7_4.

Baradwaj, B. K., & Pal, S. (2011). *Mining Educational Data to Analyze Students' Performance.* International Journal of Advanced Computer Science and Applications, 2(6), 63-69.

Beal, C. R., Qu, L., & Lee, H. (2006). *Classifying learner engagement through integration of multiple data sources.* Paper presented at the 21st National Conference on Artificial Intelligence (AAAI-2006), Boston, MA.

Bhardwaj, B. K., & Pal, S. (2012). *Data Mining: A prediction for performance improvement using classification.* International Journal of Computer Science and Information Security, 9(4), 136-140.

Fan, Y., Liu, Y., Chen, H., & Ma, J. (2019). *Data mining-based design and implementation of college physical education performance management and analysis system.* International Journal of Emerging Technologies in Learning, 14(6), 87-97.

Guruler, H., & Istanbullu, H. (2014). *Modeling student performance in higher education using data mining.* Studies in Computational Intelligence. 524,105-124.

He, W. (2013). *Examining students' online interaction in a live video streaming environment.* Computers in Human Behavior, 29(1), 90-102.

Idil, F. H., Narli, S., & Aksoy, E. (2016). *Using Data Mining Techniques Examination of the Middle School Students' Attitude towards Mathematics in the Context of Some Variables.* International Journal of Education in Mathematics, Science and Technology, 4(3), 210-228.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., Rosé, P., & Rosé, C. (2016). *Data mining and education.* Carnegie Mellon University, 1-30.

Kumar, S. A., & Vijayalakshmi, M. N. (2011). *Efficiency of decision trees in predicting student's academic performance.* First International Conference On Computer Science, Engineering And Applications, India.

Pandey, U. K., & Pal, S. (2011). *Data Mining: A prediction of performer or underperformer using classification.* International Journal of Computer Science and Information Technologies, 2 (2), 686-690.

Romero, C., & Ventura, S. (2010). *Educational data mining: a review of the state of the art.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40, 601-618.

Saa, A. A. (2016). *Educational Data Mining & Students' Performance Prediction.* International Journal of Advanced Computer Science and Applications, 7(5), 212-220 .

Siemens, G., & Baker, R. S. (2014). *Educational data mining and learning analytics. In K.Sawyer (Ed.).* Cambridge Handbook of the Learning Sciences, 253-274.

Silva, C., & Fonseca, J. (2017). *Track2 – Artificial Intelligence in Education Distributed Artificial Intelligence In Education (DAIED) And Webbased AIED Systems.* Advances in Intelligent Systems and Computing, 1-9.

Stedman, C., & Hughes, A. (2022). *Data mining*. Retrieved from Teach Target: https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). *A Review on Predicting Student's Performance using Data Mining Techniques.* Procedia Computer Science 72, 414-422.

Zorić, A. B. (2020). *Benefits of Educational Data Mining* . Journal of International Business Research and Marketing, 6(1), 12-16.