EDUCATIONAL DATA MINING APPLICATION FOR INCREASING QUALITY IN ENGINEERING EDUCATION

Bergen Karabulut
Faculty of Engineering, Kirikkale University, Kirikkale, Turkey
brgnkarabulut@gmail.com

Şeyma Cihan*
Faculty of Engineering, Kirikkale University, Kirikkale, Turkey
cihanseyma@gmail.com

*Corresponding Author

ABSTRACT

The laboratory courses, which have an important role in engineering education, allow students to transfer theoretical knowledge to practice and realize the differences that may arise between theory and practice. Analyzing the performance of the learners in these lectures and determining the factors affecting student performance is important in terms of increasing the success. In order to improve the quality of education, these factors need to be interpreted well and educational environments must be structured accordingly. Recently, data mining based on educational environments has been widely used in determining factors affecting student performance. Data mining facilitates operations on data to be collected from educational environments. In this respect, the use of outputs from educational data mining in preparation process for programs to improve the quality of engineering education such as the Association for Evaluation and Accreditation of Engineering Programs (ENAEE) will provide beneficial results. In this study, educational data mining application was carried out to be used in the process of increasing qualifications in engineering education within the scope of ENAEE preparation. Data mining algorithms were applied on the laboratory data set created by the Karabulut et al. (2017). The results obtained are interpreted in terms of MÜDEK preparation process.

Keywords: Educational Data Mining, Laboratory Dataset, MÜDEK, Education Quality

INTRODUCTION

Engineering and technology are of great importance both in the developed and developing countries in terms of the economic and social development of the countries. In today's ever-changing world, new strategies should be developed in engineering education to educate engineers who provide positive contribution to the society. These strategies should include support for technical, personal and professional development, assessment and evaluation, monitoring and control of factors that affect student performance (Memon et al., 2009). However, considering the factors that affect students' success and learning behaviors can be a complicated problem. This problem is also described in the literature as the "1000-factor problem". There are a number of factors that affect the academic performance of students. These; social, cultural, familial, demographic, educational infrastructure, socio-economic status, psychological profile and academic progress (Shaaleena and Paul, 2015; Desai et al., 2016).

Different methods based on the analysis of student data are used in evaluating student performance. Today, data mining is one of the most commonly used methods to investigate previously unexplored information, patterns and relationships from the data set containing student information (Baradwaj and Pal, 2012). The use of data mining techniques can find out wide range of critical and preliminary information such as rules of association, classes and clusters (Yadav and Pal, 2012).

Outputs of data mining applications can be used by different members of the education system (Romero and Ventura, 2010). Data mining research findings and patterns can be used for developing their own learning behaviors by students. Also, educators can use them for identifying students at risk and planning appropriate guidance for this group, identifying and resolving common problems. Besides, managers can use results for developing effective education policy development by managers (Kabra and Bichkar, 2011).

Data mining requires a standardized approach in the conversion of problem areas to data mining tasks, appropriate data transformation, preparation, selection of data mining model, assessment of the effectiveness of the results, and experience reporting. CRISP-DM (CRoss Industry Standard Process for Data Mining) defines a process model that provides a systematic framework for conducting data mining projects independently of both the business sector and the technology used. The CRISP-DM process model makes large data mining projects less costly, more reliable, more repeatable, more manageable and faster (Wirth and Hipp, 2000; Palaniappan and Awang, 2008). CRISP-DM consists of six main stages. These; determination of goal, understanding data preparation, modeling, evaluation, using / applying the results (Çınar and Arslan, 2008).

In this study, the academic performances of computer engineering students were analyzed by using classification methods. As a result of the application of data mining algorithms, it is thought that previously unexplored knowledge and patterns can be used to analyze student performance and to control critical factors affecting the performance of students. In addition to these, the educators can apply effective teaching approach, identify students at risk and use them in guidance areas. Besides, it is considered that the analysis results of the data set will contribute positively to the accreditation process such as MUDEK.

RELATED WORKS

In their study, Al-Radaideh et al. (2006) used ID3, C4.5 and Naive Bayes algorithms to evaluate the final performances of students taking C ++ programming course. The researchers have prepared 13 variable data sets in order to apply the data mining algorithms. The data set prepared in the study included the sociodemographic characteristics of the students, the characteristics of the educator and the performance of the student in the C ++ course. In the study, the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, which is considered as the standard of data mining applications, is used. Data were analyzed using the WEKA (Waikato Environment for Knowledge Analysis) program.

Kabra and Bichkar (2011) conducted their studies to evaluate the academic performance of engineering students. They collected the data of 346 students at the entrance stage and analyzed the data mining algorithms with j48 (J4.5-Java Application) Decision Tree in order to predict students' academic performance at the end of the first year. The researchers gathered information on students’ sociodemographic characteristics, contact information and past academic performance from the students and formed the study dataset. Data were analyzed by WEKA program. Of the 346 students in the study, 209 (60.46%) were correctly classified. In addition, it was determined that the most important factor affecting the academic success of the students was the entrance exam score.

In their study, Baradwaj and Pal (2012) performed a performance analysis with a data mining model by collecting the data of 50 students enrolled in graduate program in applied computer science. Past semester grade, Semester performance, homework performance, general proficiency, attendance to lectures, laboratory studies and final grade were included in the dataset used in the research. The researchers implemented ID3, ASSISTANT and C4.5 decision tree algorithms on the dataset.

Hajizadeh, and Ahmadzadeh (2014) used Turkey Student Evaluation data sets from the UCI (University of California, Department of Information and Computer Science-Machine Learning Repository) Machine Learning database. The researchers used the data mining techniques (Apriori) and Classification (REPTree) to examine the factors in the prevention of course repetition. The data were analyzed via the WEKA program.

Satyanarayana and Nuckowski (2016) used three different classification algorithms to evaluate the academic performances of students in the field of computer technology systems. These are; Decision Trees J48, Naive Bayes and Random Forest data mining algorithms. The researchers implemented the model on two separate sets of data. UCI Student Performance Data Set and New York City College of Technology CST Computer Introductory Course Data Set were used in the study. In addition, in order to identify effective associations on the performance of students in the study, Apriori, Filtered Associator and Tertius, which are rule-based algorithms, have been used.

In their study, Figueiredo et al. (2016) analyzed the effects of the methods applied in the chemistry laboratory on student motivation and learning behaviors with data mining algorithms. In the study 3447 students' information was collected by questionnaire method. The k-means clustering algorithm is applied on the data set via the WEKA program.

In their work, Desai et al (2016) gathered information from a 60-student web-based tool to profile the third-year students of computer engineering. The k-means clustering algorithm is used to profile the students.

Asif et al (2017) collected information from 210 students in order to evaluate the academic performance of the students enrolled in the information technology program. Data mining algorithms have been implemented through the RapidMiner program. Decision Trees, Nearest Neighborhood, Neural Networks, Naive Bayes, Random Forest algorithms were applied in the study and the results were compared in terms of classification accuracy.

In their study, Costa et al. (2017) formulated a data mining model in order to predict early probability of failure of students in the introduction to programming. The researchers applied the Naive Bayes, Decision Tree (J48), Multilayer Neural Network, SVM algorithms on the data, which consisted of information about students' sociodemographic characteristics and academic performances. Pentaho open-source software tool for preliminary analysis of data and WEKA program for data mining algorithms were used in the study. The researchers found that students with a likelihood of failure were approximately 50% to 80% correct from the first week after enrollment.

METHOD

In this study, the laboratory dataset created by Karabulut et al. (2017) to determine the factors that affect the student performance was used. This dataset contains data collected from students of Electronic Circuits and Electronic Circuits Laboratory in the Department of Computer Engineering. There are 57 attributes in the dataset and the dataset contains 140 records.

In study, 17 attributes were selected from the laboratory dataset and used in the study. Selected attributes are given in Table 1. The 17th attribute in the Table 1, which is called result, is set as the target attribute. This attribute refers to the status of the students' lecture success (values of *result*-P: passed, F: failed, NE: Not Enrolled)

Table 1. Selected attributes of laboratory dataset

#	Attribute name	Type
1	bYear	Numeric
2	sex	Binary
3	birthplace	Nominal
4	residence	Nominal
5	highSchool	Nominal
6	CGPA	Numeric
7	eduType	Binary
8	Physics1	Categorical
9	Physics2	Categorical
10	Math1	Categorical
11	Math2	Categorical
12	Electric	Categorical
13	reTakingNumber	Numeric
14	attandence	Numeric
15	labEnrollment	Categorical
16	labResult	Categorical
17	result	Categorical

There are missing values in some attributes in the laboratory dataset. Table 2 shows the attributes, which have missing value, with the missing value rates.

Table 2. Missing Values Rates of Attributes

Attribute Name	Rate of missing values
highSchool	26%
residence	1%
Physics1	1%
Physics2	1%
Math1	1%
Math2	1%
Electric	1%

In data preparation process, if the missing value rate is below a certain value, the corresponding records can be deleted. In addition, the missing values can be cleared with some operations. Generally, when missing data is cleared, mode operation is used for categorical variables and average operation is used for numerical variables. The Weka (Waikato Environment for Knowledge Analysis) Program used in this study is performing the missing values clearing process.

The classification process has been applied on the dataset that has become ready by clearing the missing values. There are various methods of classification in data mining. In this study, 4 of these methods are used and the classification methods used are;

- J48 Decision Tree
- Naive Bayes
- Multilayer Perceptron
- Zero R

One of the most common programs used in data mining applications is Weka. The Weka program was used to apply the classification methods specified in this study. This program provides convenience in analyzing data and applying various algorithms.

FINDINGS

Classification process was applied on the laboratory dataset with the help of Weka program. Multilayer Perceptron, Naive Bayes, J48 Decision Tree and Zero R classification methods are used for classification. In addition, pruning operation was applied on J48 Decision Tree. The results obtained are given in Table 3.

Table 3. Accuracy rates of applied classification methods

Classification Method	Accuracy
Pruned J48 Decision Tree	%80
Multilayer Perceptron	%76.42
Naive Bayes	%73.57
J48 Decision Tree	%72
Zero R	%62.85

When the Table 3 is examined, it is seen that the best result is obtained with the help of Pruned J48 Decision Tree. When the J48 decision tree is applied, the number of leaves is 128 and the size of the tree is 141. Pruning operation reduced the number of leaves to 5 and the size of the tree to 8. The obtained pruned J48 decision tree is given in Figure 1.

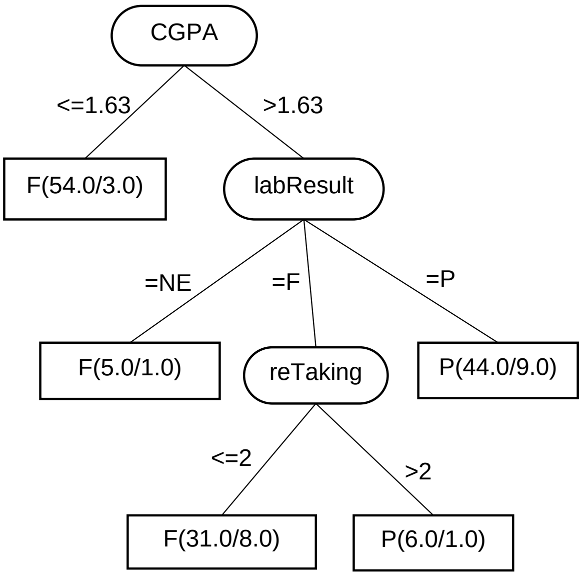

Figure 1: Pruned J48 Decision Tree

Examining Figure 1, it can be seen that a simpler and more interpretable tree is obtained with pruning operation. In addition, pruning increased the classification success. When all the methods applied are evaluated, it is seen that Pruned J48 Decision tree has the highest accuracy rate.

CONCLUSIONS

In this study, the classification process was performed on the laboratory data set. For classification, Multilayer Perceptron, Naive Bayes, J48 Decision Tree and Zero R classification methods are used. These classification methods are compared in terms of accuracy rate. Pruned J48 Decision Tree has the highest accuracy rate and is important in terms of being able to provide interpretation to educators and researchers. According to pruned J48 tree results; it is seen that CGPA, labResult and reTaking have been found to have more influence on students' performance.

The determination, evaluation and monitoring of the program and course outputs and the management of the continuous improvement process, which are among the most important criteria of the MÜDEK accreditation program, require considerable time and effort by educators and managers. For this reason, it is thought that the obtained records and results from the educational data mining process facilitate the implementation of the procedures related to the MÜDEK criteria which are mentioned above. Also, all information related to student performance gained by educational data mining provide significant contributions to accreditation process. As the final results of all activities related to determining and monitoring the factors affecting student performance with educational data mining applications have a key role in increasing the quality of engineering education.

REFERENCES

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. *In International Arab Conference on Information Technology (ACIT'2006)*, Yarmouk University, Jordan.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*.

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.

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*(6), (pp. 601-618).

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, (pp. 247-256).

Çınar, H. & Arslan, G., 2008. “Veri madenciliği ve CRISP-DM yaklaşımı”, XVII. İstatistik Araştırma Sempozyumu, (pp. 304-314), Ankara.

Desai, A., Shah, N., & Dhodi, M. (2016, August). Student profiling to improve teaching and learning: A data mining approach. *in Data Science and Engineering (ICDSE), 2016 International Conference on* (pp. 1-6). IEEE.

Figueiredo, M., Esteves, L., Neves, J., & Vicente, H. (2016). A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation. *Chemistry Education Research and Practice*, 17(1), (pp. 156-171).

Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect the students academic performance-Data Mining Approach. arXiv preprint arXiv:1409.2222.

Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, 36(11), (pp. 8-12).

Karabulut, B., Cihan, Ş., Ünver, H. M., Ergüzen, A. (2017). Creating Laboratory Dataset for Educational Data Mining”, I. International Scientific and Vocational Studies Congress, Cappadocia, Turkey.

Memon, J. A., Demirdöğen, R. E., & Chowdhry, B. S. (2009). Achievements, outcomes and proposal for global accreditation of engineering education in developing countries. *Procedia-Social and Behavioral Sciences*, 1(1), (pp. 2557-2561).

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. In Computer Systems and Applications, 2008. *AICCSA 2008. IEEE/ACS International Conference on* (pp. 108-115). IEEE.

Satyanarayana, A., & Nuckowski, M. (2016). Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance.

Shaleena, K. P., & Paul, S., “Data mining techniques for predicting student performance”, In *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*, IEEE, pp. 1-3, 2015.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, (pp. 29-39).

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.