Technological Applied Sciences
ISSN: 1308 7223 (NWSATAS)
ID: 2018.13.4.2A0161

Status : Original Study
Received: December 2017
Accepted: October 2018

Bergen Karabulut
Şeyma Cihan
Halil Murat Ünver
Atilla Ergüzen
Kırıkkale University, Kırıkkale-Turkey
brgnkarabulut@gmail.com; cihanseyma@gmail.com;
hmunver@hotmail.com; aerguzen@yahoo.com

DOI	http://dx.doi.org/10.12739/NWSA.2018.13.4.2A0161	
ORCID ID	0000-0003-0755-1289	0000-0001-6267-2441
	0000-0001-9959-8425	0000-0003-4562-2578
CORRESPONDING AUTHOR	Şeyma Cihan	

ELECTROLAB: A NEW DATASET FOR EDUCATIONAL DATA MINING

ABSTRACT

Engineering education prepares students for life by presenting theoretical and practical knowledge together. A common method is applying laboratory experiments for practicing theoretical knowledge by students. The objective of the laboratory experiments is to gain student the ability of transferring theoretical knowledge to practice and see the differences between theory and practice. However; classical evaluation of laboratory courses has some difficulties in terms of assessing complex input factors related to students. Educational data mining, which has been widely used recently, allows evaluations for student performance to be made easier. Implementing educational data mining for laboratory lesson can be important contributions to the determination of the factors affecting student performance and the structuring of training methods accordingly. In this study, Electronic Circuits Laboratory Course, which is the practice of Electronic Circuits Course as a basic course of Computer Engineering education, were examined. A laboratory data set called ELECTROLAB was created by collecting data from these courses. The first phases of CRISP, the standard for data mining operations, have been implemented on this data set. The data set was prepared and the attributes in the data set were analyzed according to these phases. In the study, R programming language and Weka program were used. The data set created by this study and the analysis process will be the source of data mining methods to be applied in future studies. In this way, it will be possible to determine the factors that affect the student performance and to make studies to increase the success.

Keywords: Data Mining, Educational Data Mining, Laboratory Dataset, Student Performance, CRISP-DM

1. INTRODUCTION

Engineering education has great importance for the technological and economic development of the countries. Developing new strategies in engineering education is needed to catch up with rapidly evolving and changing technology and to be able to train useful engineers to the society. These strategies were described by Memon at al. [1] as supporting personal growth and professional development, measuring and evaluating, monitoring and controlling factors affecting student performance. However, in literature evaluation of the factors affecting student success and performance was described as a difficult

How to Cite:
Karabulut, B., Cihan, Ş., Ünver, H.M., and Ergüzen, A., (2018). Electrolab: A New Dataset for Educational Data Mining, Technological Applied Sciences (NWSATAS), 13(4): 318-328, DOI:10.12739/NWSA.2018.13.4.2A0161.

problem and was even called a “1000 factor problem”. Factors affecting the academic success of the students were summarized as social, cultural, familial, demographic, educational infrastructure, socioeconomic status, psychological profile and academic progress [2 and 3]. In addition, engineering education has a structure that provides theory and practice together. This increases the factors affecting performance and requires evaluations to consider the relationship between theory and practice. The practice of the knowledge learned in the theoretical courses is usually provided by laboratory courses. Laboratory courses allow students to practice the knowledge learned in the theoretical lectures and see the differences that may arise between theoretical and practical. To be able to train effective engineers, it is important to increase the effectiveness of laboratory lessons and to allow the student to practice more. 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 [4]. The use of data mining techniques can find out wide range of critical and preliminary information such as rules of association, classes and clusters [5].

The results of data mining applications can be used by different members of the education system [6]. 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 [7]. 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 [8][9]. CRISP-DM consists of six main stages. These are; determination of goal, understanding data preparation, modeling, evaluation, using/applying the results [10].

In this study, a laboratory data set was created for use in educational data mining. The purpose of this dataset is investigation of theoretical and practical lessons in terms of academic performance, weekly change of achievement in practice lessons, factors affecting course success, student perception, and effect of student perceptions and ideas on performance. The courses of Electronic Circuits and Electronic Circuits Laboratory which are one of the basic courses of computer engineering are investigated. These lessons were followed up till the end of one semester and the obtained data were transformed into a data set. Preliminary analyzes were performed with the help of Weka program and R programming language on the generated data set. It is thought that the educational data mining study to be done on this dataset will be able to determine the factors that affect students’ performance and to control these factors. Besides, it is considered that the analysis results of the data set will contribute positively to the accreditation process such as MÜDEK.

2. RESEARCH SIGNIFICANCE

It is very important to be able to determine the factors that affect the success and performance of the student in the laboratory environments that is the most basic parts of the engineering education and provide the opportunity to practice the student. In addition, analyzing the data obtained from these environments and shaping the laboratory training according to the obtained results will increase the contribution of the practices. However, it is a very difficult problem to determine the factors affecting student success and performance, which is described as a 1000 factor problem in the literature. Educational data mining, which has been widely used recently, facilitates this problem by analyzing data obtained from educational environments. In this direction, educational data mining studies that deal with laboratory environments are needed. For this reason, in this study, one of the core courses of Computer Engineering and the laboratory course belonging to this course were discussed and a laboratory data set was created for use in educational data mining.

There are several studies in the literature which have obtained important findings about educational data mining. In this respect, some of the studies that show the significance of the topic and achieve significant results are mentioned below. In their study, Al-Radaideh et al. [11] evaluated the final performances of students who took C++ programming courses. For this purpose, they constructed a data set consisting of 13 variables including the sociodemographic characteristics of the students, the characteristics of the educator and the student’ performance of the C ++ course. They used ID3, C4.5 and Naive Bayes algorithms in the study. It has been carried out in accordance with the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, which is regarded as the standard of data mining applications. Variable analyzes were conducted through the WEKA (Waikato Environment for Knowledge Analysis) program.

Kabra and Bichkar [7] conducted a study to evaluate the academic performance of engineering students and collected the data of 346 students in the entry phase. They used the j48 Decision Tree algorithm, which is one of the data mining algorithms. With these analyzes, the students' academic performance was predicted at the end of the first year. The researchers gathered information about sociodemographic characteristics, communication knowledge and past academic achievements of students and formed a data set for the study. This data set was converted into student. ARRF file format and analyzed by WEKA program. In the study, 209 of 346 students were classified correctly. In addition, it was determined that the most important factor affecting the academic success of the students was the entrance examination score. In their study, Baradwaj and Pal [4] performed a performance analysis with a data mining model by collecting the data of 50 students enrolled in the master's program in applied computer field. The dataset used in the research includes students' grade average of the previous term, term grade average, seminar performance, homework performance, general proficiency, attendance to lectures, laboratory studies and end-of-term grades.

Hajizadeh and Ahmadzadeh [12] studied effective factors in preventing the repetition of the course. In their study, they used the Turkey Student Evaluation dataset, which has 5820 samples and 33 variables, on the UCI Machine Learning Repository web page and applied the Data mining techniques of Extraction of Association Rules (Apriori) and Classification (REPTree). The data were analyzed via the WEKA program. Satyanarayana and Nuckowski [13] used three different classification algorithms to evaluate the academic performances of students in the field of computer systems technology. These algorithms are the data mining algorithms Decision Trees J48, Naive Bayes and Random Forest. The researchers implemented the model they developed on two separate datasets. 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, they used Apriori, Filtered Associator and Tertius, which are rule-based algorithms, to define the association rules affecting the performance of the students.

Figueiredo et al. [14] 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. Survey form; the characteristics of the students, the characteristics of the training method applied in the laboratory and the students' thoughts on the effects on the learning of laboratory studies. The k-means clustering algorithm is applied on the data set via the WEKA program. Desai et al. [3] conducted a study to profile the third-year students of computer engineering. In the study, information was collected from 60 students with the aid of a web-based tool. The k-means clustering algorithm is used to profile the students. Asif et al. [15] collected information of 210 students to evaluate the academic performance of the students enrolled in the Informatics Technologies program and applied data mining algorithms through 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. [16] formulated a data mining model to predict early probability of failure of students in the introduction to programming. In the study, data was collected from 262 students enrolled in distance education, and 161 students receiving in-class education on campus. 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. Researchers have found that students who are likely to fail the algorithms of the applied data mining can detect about 50%-80% correctly from the first week after enrollment.

In the study of Almarabeh [17], Naive Bayes, BayesNet, ID3, C4.5 (J48), and Neural Network (NLP) classification algorithms were used on the dataset of 225 students records. This dataset contains the following attributes: midterm marks, assignment performance, attendance, seminar performance, lab experiments, project performance, workshop and final marks. The best performance in the study was obtained from the Bayesian Network classification algorithm. Olaniyi et al. [18] conducted a study to estimate the end of semester examination performances of 284 students enrolled in the Internet Technology and Programming course in the computer engineering program on the data set consisting of previous semester marks, class test grade, assignment, attendance, lab work, and end semester marks by using BFTree, J48 and CART decision tree algorithms. The best performance in the study was obtained from the BFTree algorithm with 67.07% accuracy rate. In their work, Sarra et al. [19], used the Bayesian Profile Regression model, an educational data mining tool, to profile the students at risk for academic failure. In the study, an online questionnaire was applied to students via e-mail and a dataset consisting of 561 students was created. By applying this approach on the dataset, 9 typical student profiles have been identified.

Robbiano et al. [20] used Spearman correlation analysis and Principle Component Analysis (PCA) methods on a data set of 803 students in their work to analyze the performances of electrical and computer engineering students in core technical courses. In the correlation analysis, it was found that the highest correlations were a strong positive correlation between the individual lectures and the cumulative general grade average. The cumulative grade point average can account for about 60% of the total variance of individual courses. However, when the relationship between individual lessons is examined while excluding the cumulative general grade average, it is determined that the highest correlations are a strong positive correlation in the middle order among the sequential courses covering the same topic. In the study, it was determined that prerequisite courses provided a significant benefit in compulsory courses. Khan and Ghosh [21] conducted a study to determine the relationship between teaching quality and student performance on a dataset of 9072 student records collected from academic online system. In the study, only the performances of the students enrolled in the theoretical courses in the traditional classroom environment were analyzed by using association rules approach.

3. EXPERIMENTAL STUDY

In this study, Electronic Circuits Laboratory Course, which is the practice of Electronic Circuits Course as a basic course of Computer Engineering education, were examined. Data collection studies were carried out in Kırıkkale University Department of Computer Engineering during the fall semester of 2016-2017 academic years. At the end of the semester, a questionnaire was applied to the students who took these courses and some data for the data set were obtained in this way. Besides, during the semester, data for student performance were collected from the laboratory. Also, in this study, the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, which is considered as the standard of data mining applications, is used for determination of purpose, understanding data and preparation. Diagram of CRISP-DM process was shown in Figure 1.

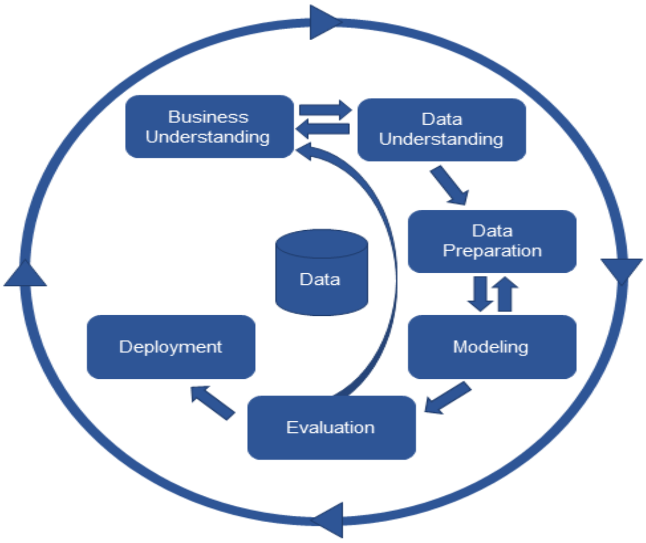

Figure 1. CRISP-DM

3.1. Survey

A survey was applied to the students who enrolled Electronic Circuits courses in the fall semester of 2016-2016 academic year. Some questions were directed to the students within this survey. The information received from the student with the help of this survey is:

- **Socio-demographic characteristics:** Information such as the age, gender, year of birth of the student was collected within this scope.
- **Level of perception about courses:** The difficulty perceived by the lesson for the Electronic Circuits and Electronic Circuits Laboratory lessons was asked. In this context, 5 different categories were presented to students and they were asked to choose between Very Difficult, Difficult, Medium, Easy and Very Easy.
- **Success status of related courses:** In this context, the lessons learned in previous semesters related to Electronic Circuits course were determined and the success status of the students was taken into consideration. The courses related to Electronic Circuits course are taken as Mathematics, Physics and Electric Circuits by taking the opinions of the instructors of these courses. Within the scope of the survey, the students were asked about their past performance from these related courses. This question is addressed in two categories as passed or failed.
- **Retaking of course:** It has been asked how many times the student has retake the prescribed course.
- **Student opinion:** The opinions of the students belonging to the laboratory course were asked. It has been asked that how much the experiments and simulation exercises have contributed to the course and whether they want simulation exercises.
- **General performance:** Cumulative Grade Point Average was collected to assess the overall success of the student.
- **Other Effective Factors:** The place information where the student stayed throughout the student life was taken. In addition, status of access to the internet in the place where the student stayed was asked.

3.2. Laboratory Information

The performance of the student in the course of the Electronics Circuits Laboratory was rated in the range 0-10. These scores are given weekly and added to the data set. Besides, it is desirable for the student to set up a practice in the simulation environment that will be done in the laboratory before coming to the laboratory every week and to examine the results. On a weekly basis, information was collected and recorded whether the student did the simulation exercises. In addition, information about how long the student worked before coming to the laboratory weekly was asked and kept on a weekly basis. Also, the attendance information on laboratory courses has also been taken into consideration. At the end of the semester, the grade taken from the laboratory course and the information about the student's status of the course were taken.

3.3. Electronic Circuits Course Information

During the 14-week course period, students' attendance information related to this course was taken. Also, at the end of the semester, the student's status of success and grading were taken.

4. RESULTS AND DISCUSSION

Literature studies on student performance show that the number of attributes is kept smaller. In addition, it is seen that the current period performance of student is not taken into consideration and generally the information of the student's past performance is handled. In addition, it has been observed that the students' perception of the course and the idea of the teaching methods are not taken into consideration. In the light of these observations, a more comprehensive data set was created in this study. However, the data related to current term performance of student was taken into account. In addition, the student's perception of the courses and their views towards the teaching methods are also taken into consideration.

Table 1. Attributes of dataset

#	Attribute name	Type
1	bYeard	Numerical
2	sex	Categorical
3	birthplace	Categorical
4	residence	Categorical
5	highSchool	Categorical
6	CGPA	Numerical
7	eduType	Categorical
8	Physics1	Categorical
9	Physics2	Categorical
10	Math1	Categorical
11	Math2	Categorical
12	Elektric	Categorical
13	reTakingNumber	Numerical
14	accommodation	Categorical
15	internet	Categorical
16	labDifficulty	Categorical
17	lectureDifficulty	Categorical
18	labContribution	Categorical
19	simContribution	Categorical
20	simHomework	Categorical
21	attandence	Numerical
22	labEnrollment	Categorical
23	labAttandence	Numerical
24	G1	Numerical
25	S1	Categorical
26	G2	Numerical
27	S2	Categorical
28	G3	Numerical
28	S3	Categorical
30	G4	Numerical
31	S4	Categorical
32	G5	Numerical
33	S5	Categorical
34	G6	Numerical
35	S6	Categorical
36	G7	Numerical
37	S7	Categorical
38	G8	Numerical
39	S8	Categorical
40	G9	Numerical
41	S9	Categorical
42	G10	Numerical
43	S10	Categorical
44	grade	Categorical
45	result	Categorical
46	labGrade	Categorical
47	labResult	Categorical

In this way, a comprehensive data set was created by combining the features that the literature studies did not take into consideration. As a result of the study, the data set with the attribute name, possible value and type is created in Table 1. The created dataset is in csv format and ready to use. The generated dataset has 47 attributes and 140 records.

Using Weka and R programming language created dataset was examined and missing values were detected. Rates of missing values were presented in Table 2.

Table 2. Missing value rates of attributes

Attribute Name	Rate of Missing Values
labContribution	28%
simContribution	28%
simHomework	27%
highSchool	26%
accommodation	26%
internet	26%
labDifficulty	26%
lectureDifficulty	26%
residence	1%
Physics1	1%
Physics2	1%
Math1	1%
Math2	1%
Electric	1%

The summarized statistics of the numerical variables in the ELECTROLAB dataset are given in Table 3. In addition, bar graphs of some categorical variables in the dataset were generated. The bar graphs are given in Figure 2.

Table 3. Summary of numeric attributes

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
bYear	1980	1994	1995	1995	1997	1998
CGPA	0.320	1.218	1.930	1.783	2.275	3.340
retaking	1.000	1.000	1.000	1.764	2.000	6.000
attendence	0.000	0.000	8.00	5.65	10.00	13.000
labAttendence	0.000	0.000	7.500	5.186	9.000	11.000
G1	0.000	0.000	0.000	4.3	10.0	10.000
G2	0.000	0.000	0.000	3.429	7.250	10.000
G3	0.000	0.000	0.000	3.157	7.000	10.000
G4	0.000	0.000	0.000	2.957	6.000	10.000
G5	0.000	0.000	0.000	2.279	5.000	10.000
G6	0.000	0.000	1.000	3.736	8.000	10.000
G7	0.000	0.000	0.000	2.436	5.000	10.000
G8	0.000	0.000	0.000	1.286	2.000	8.000
G9	0.000	0.000	0.000	2.95	5.000	10.000
G10	0.000	0.000	0.000	1.743	2.000	10.000

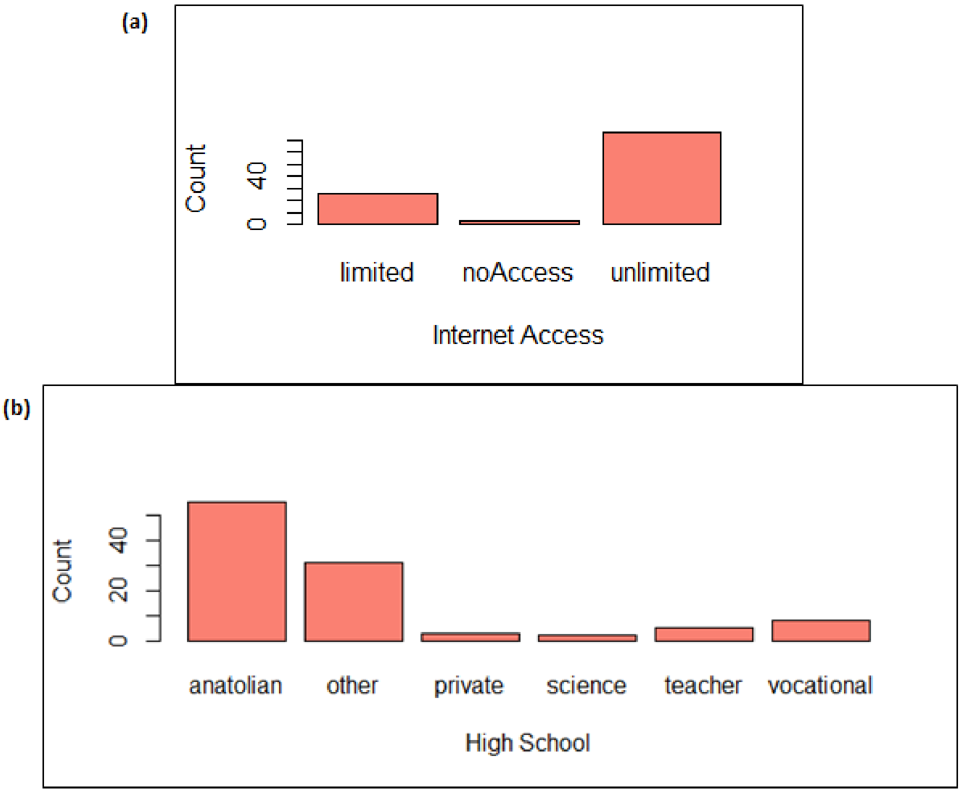

Figure 2. Bar graphs, a) internet; b) high school

5. CONCLUSIONS AND RECOMMENDATIONS

Educational data mining is used to analyze data obtained from educational environments and to extract meaningful and useful information from these data. The meaningful and useful information obtained from the data is transferred to the educational environments again. It is given theoretical and practical together in engineering education. In this case, the influence of the theoretical and practical courses and the determination of the factors affecting the performance in these courses are very important to provide a better education. In this study, a laboratory data set called as ELECTROLAB was created for use in educational data mining. The generated dataset has a total of 140 samples and 47 attributes. In terms of the features it covers, a comprehensive data set with student opinions and perceptions was created. Analyzes on this dataset can provide important findings for student performance. Besides, the results of the study on this data set can be used in studies to improve the quality of engineering education such as MÜDEK.

NOTICE

This study was presented as an oral presentation at the I. International Scientific and Vocational Studies Congress (BILMES 2017) in Nevşehir/Ürgüp between 5-8 October 2017.

REFERENCES

[1] Memon, J.A., Demirdöğen, R.E., and Chowdhry, B.S., (2009). Achievements, Outcomes and Proposal for Global Accreditation of Engineering Education in Developing Countries. Procedia-Social and Behavioral Sciences, Volume:1, Number:1, pp:2557-2561.

[2] Shaleena, K.P. and Paul, S., (2015). Data mining Techniques for Predicting Student Performance. In Engineering and Technology

(ICETECH), 2015 IEEE International Conference on, IEEE, pp:1-3.

[3] Desai, A., Shah, N., and Dhodi, M., (2016). Student Profiling to Improve Teaching and Learning: A Data Mining Approach. In Data Science and Engineering (ICDSE), 2016 International Conference on.IEEE. pp:1-6.

[4] Baradwaj, B.K. and Pal, S., (2012). Mining Educational Data to Analyze Students' Performance. arXiv preprint arXiv:1201.3417.

[5] Yadav, S.K. and Pal, S., (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students Using Classification. arXiv preprint arXiv:1203.3832.

[6] Romero, C. and 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), Volume:40, Number:6, pp:601-618.

[7] Kabra, R.R., and Bichkar, R.S., (2011). Performance Prediction of Engineering Students Using Decision Trees. International Journal of Computer Applications, Volume:36, Number:11, pp:8-12.

[8] Wirth, R., and 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.

[9] Palaniappan, S. and 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, IEEE, pp:108-115.

[10] Çınar, H. and Arslan, G., (2008). Veri Madenciliği ve CRISP-DM Yaklaşımı, XVII. İstatistik Araştırma Sempozyumu, Ankara, pp:304-314.

[11] Al-Radaideh, Q.A., Al-Shawakfa, E.M., and Al-Najjar, M.I., (2006). Mining Student Data Using Decision Trees. In International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.

[12] Hajizadeh, N. and Ahmadzadeh, M., (2014). Analysis of Factors That Affect the Students’ Academic Performance-Data Mining Approach. arXiv preprint arXiv:1409.2222.

[13] Satyanarayana, A. and Nuckowski, M., (2016). Data Mining Using Ensemble Classifiers for Improved Prediction of Student Academic Performance.

[14] Figueiredo, M., Esteves, L., Neves, J., and 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, Volume:17, No:1, pp:156-171.

[15] Asif, R., Merceron, A., Ali, S.A., and Haider, N.G., (2017). Analyzing Undergraduate Students' Performance Using Educational Data Mining. Computers & Education.

[16] Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., and 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, Volume:73, pp:247-256.

[17] Almarabeh, H., (2017). Analysis of Students' Performance by Using Different Data Mining Classifiers. International Journal of Modern Education and Computer Science, Volume:9, No:8.

[18] Olaniyi, A.S., Kayode, S.Y., Abiola, H.M., Tosin, S.I.T., and Babatunde, A.N., (2017). Student’s Performance Analysis Using Decision Tree Algorithms. Annals. Computer Science Series, Volume:15, No:1.

[19] Sarra, A., Fontanella, L., and Di Zio, S., (2018). Identifying Students at Risk of Academic Failure within the Educational Data Mining Framework. Social Indicators Research, pp:1-20.

[20] Robbiano, C., Maciejewski, A.A., and Chong, E.K., (2018, June). Board 77: Work in Progress: An Analysis of Correlations in Student Performance in Core Technical Courses at a Large Public Research Institution’s Electrical and Computer Engineering Department. In 2018 ASEE Annual Conference & Exposition.

[21] Khan, A. and Ghosh, S.K., (2018). Data Mining Based Analysis to Explore the Effect of Teaching on Student Performance. Education and Information Technologies, pp:1-21.