Ömer Faruk Akmeşe
PhD of Computer Engineering, Senior Lecturer at the Department of Computer Technologies
University of Hitit, Çorum, Turkey
ORCID ID 0000-0002-5877-0177
ofarukakmese@hitit.edu.tr

Hakan Kör
PhD of Computer Engineering, Assistant Professor at the Department of Computer Engineering
University of Hitit, Çorum, Turkey
ORCID ID 0000-0002-8314-9585
hakankor@hitit.edu.tr

Hasan Erbay
PhD of Computer Engineering, Professor at the Department of Computer Engineering
University of Turkish Aeronautical Association, Ankara, Turkey
ORCID ID 0000-0002-7555-541X
herbay@thk.edu.tr

USE OF MACHINE LEARNING TECHNIQUES FOR THE FORECAST OF STUDENT ACHIEVEMENT IN HIGHER EDUCATION

Abstract. The machine learning method, which is a sub-branch of artificial intelligence and which makes predictions with mathematical and statistical operations, is used frequently in education as in every field of life. Nowadays, it is seen that millions of data are recorded continuously, and a large amount of data accumulation has occurred. Although data accumulation increases exponentially, the number of analysts and their capabilities to process these data are insufficient. Although we live in the information age, it is more accurate to say that we live in the data age. By using stored and accumulated data, it is becoming increasingly essential to reveal meaningful relationships and trends and to make predictions for the future. It is important to analyze the data obtained from the education process and to evaluate the success of the students and the factors affecting success. These analyses may also contribute to future training activities. In this study, a data set, including socio-demographic variables of students enrolled in distance education at Hitit University, was used. The authors estimated the success of the students with demographic and social variables such as age, gender, city, family income, family education level. The primary purpose is to provide students with information about their estimated academic achievement at the beginning of the process. Thus, at the beginning of the education process, students' success can be increased by informing the students who are predicted to be unsuccessful. Diversification and enhancement of this data may also support other decision-making mechanisms in the training process. Additionally, the factors affecting students' academic success were researched, and the students' educational outcomes were evaluated. Prediction success was compared using various machine learning algorithms. As a result of the analysis, it was determined that the Random Forest algorithm was more predictive of student achievement than others.

Keywords: machine learning in education; adult education; educational data mining.

1. INTRODUCTION

A sustainable learning approach is needed in the 21st century, where information and technology are intertwined and the variety of information sources increases day by day. Analyzing the educational process and activities can contribute to future educational activities [1], [2]. The digital age has brought about incredible changes in the production of information, consumption, adaptation, sharing, and the transformation of resources and services [3]. If the raw data is not processed and made into information, it will not have much value. Unprocessed data are metaphorically compared to crude oil [4], [5]. Unrefined data has value, but it is not useful. It gains value when similarly processed in raw data and can be valued by processing it with various data mining methods.

One of the methods commonly used in the processing of data in education is machine learning. In general, machine learning can be defined as achieving better results in the future based on past examples [6]. The focus of machine learning studies is to gain the ability to perceive intricate patterns and to take data-based rational decisions to computers. Machine learning is closely related to areas such as data mining, statistics, pattern recognition, probability theory, supervised control, and artificial intelligence.

There are a number of studies aiming to predict student success. With the regression method, the relationship between students' demographic characteristics, university entrance qualifications, aptitude test scores, first-year courses' performances, and general performances were investigated [7]. Kabakchieva has developed models to predict student success based on pre-university and university performance characteristics [8].

In today's knowledge-oriented societies, individuals have to use the information they have acquired, adapt to new ideas, cooperate with other people, and keep up with unpredictable changing situations [9]. Therefore, if it is thought that consciousness and determination can only be provided to individuals through education, education is a functional key that will lead to the transformation of nations into a modern information society.

In this study, a prediction system with machine learning has been developed by going beyond traditional graphics and descriptive statistics. In the research, the success of students with demographic and social variables such as age, gender, city, family income, family education level was estimated. Most of the variables in the study consist of factors in which students and instructors do not have intervention control. However, the main purpose here is to provide information to instructors and students at the beginning of the educational process. In other words, the data obtained from the students at the beginning of the education process is aimed to contribute to the future educational activities of the students. Thus, it may be possible to take precautions in advance for possible malfunctions in the education process. With this information, instructors can monitor students' progress and intervene early in academic problems.

2. RESEARCH METHODS

Data mining focuses on the exploration of previously unknown features using an existing data set [10]. Data mining methods are used by many researchers for prediction purposes. The classification and evaluation under data mining techniques help in the creation of training data, the classification of the estimation model as well as the testing of classification efficiency [11],[12].

In data mining, models can be divided into two main categories, predictive and descriptive. Estimated models are intended to make inferences about the future. Descriptive models enable identifying patterns in the data to guide decision-making [13].

The field of study related to the use of various algorithms for computers to be able to perform the learning function is called machine learning [14]. To solve a certain problem, imitating the problem-solving abilities of the human and equipping the system with information beforehand is within the field of machine learning [15]. Machine learning algorithms train themselves using available data. These algorithms then make predictions for possible new situations. There exist many machine learning methods. The success of these methods varies according to the data. Therefore, it is not correct to say that a certain method is suitable for every data.

2.1. Data Analysis

This study was carried out with student data obtained from Hitit University Distance Education Center. These data include variables such as gender, age, family income, students' mother's and father's profession using the online learning system. In the conceptual model of the study, independent variables were used to predict the academic success of the students. Successful ones from Ataturk's Principles and Revolution History course were determined as one group and unsuccessful ones as a separate group. In addition to descriptive statistical analysis, machine learning methods were used to predict groups. The universe of the study consisted of distance education centers located in Turkey. The sample of the study consists of randomly selected students from different units registered at Hitit University Distance Education Center. A total of 478 records were included in the study: 211 males (44.1%) and 267 females (55.9%). There were 295(61.7%) students who succeeded in the course and 183(38.3%) students who failed. Data were analyzed with exploratory data analysis and machine learning methods. In the study, data analyses were carried out using Rapidminer (9.5) packet programs, and Python (3.7) programming language.

This study was carried out with the data of the students who enrolled in the Atatürk Principles and Revolution History Course at Hitit University. The data name, data type, and definition for each record are shown in Table 1.

Table 1.

Dataset description

Features	Type	Description	Variables
Unit	Categorical	A unit where students register	Input
Address	Categorical	The city where the students live(small or big)	Input
Age	Numerical	Age of students	Input
Gender	Categorical	1(male) / 2(female)	Input
Mother's education	Categorical	Mother's education status	Input
Father's education	Categorical	Father's education status	Input
Number of siblings	Numerical	Number of student's siblings	Input
Family income	Categorical	Family income is divided into low, medium, and high.	Input
Mother's profession	Categorical	Mother's profession is identified as either "housewife" or "employee."	Input
Father's profession	Categorical	The profession of the father is identified as either civil servant or non-civil servant.	Input
High school type	Categorical	High school type is divided into vocational high schools and normal high schools.	Input
University placement score	Numerical	University placement score	Input
Exam Score	Numerical	The score of the students who enrolled in the Atatürk Principles and Revolution History course with distance education	
Status	Categorical	Student's passing or staying status 0(fail) / 1(pass)	Target

Categorical data represent types of data that can be expressed in groups Numerical data are data expressed in numbers. Target: refers to the dependent variable, Input: refers to the independent variable.

Table 1 shows many features of the data set used. According to the data set obtained from the distance education unit, weighting analysis was carried out to determine the

characteristics that are important in determining successful students. The analysis results are shown in Table 2. The resulting sample set has two properties, 'Attributes' and 'Weight'. In the process, the results of the weight of the Chi-Square Statistics operator are seen. Chi-square statistics is a nonparametric statistical technique used to determine whether the distribution of observed frequencies differs from theoretically expected frequencies. Chi-square statistics use nominal data. The calculated weights are normalized between 0 and 1. Values above 0.1 as a threshold value were included in the machine learning analysis.

Table 2.

Attribute and Weight

No	Attributes	Weight(Chi-square)
1	University Placement Score	0.532
2	Unit	0.193
3	Family Income	0.191
4	High School Type	0.154
5	Father's Profession	0.145
6	Mother's Education	0.117
7	Father's Education	0.078
8	Address	0.078
9	Number of Siblings	0.077
10	Mother's Profession	0.017
11	Age	0.016
12	Gender	0.010

Unit: A unit where students register, Family Income: Family income is divided into low, medium, and high, High School Type: High school type is divided into vocational high schools and normal high schools, Father's Profession: Father's profession is identified as either civil servant or non-civil servant, Mother's Education: Mother's education status, Father's Education: Father's education status, Address: The city where the students live(small or big), Number of Siblings: Number of student's siblings, Mother's Profession: Mother's Profession is identified as either "housewife" or "employee.", Age: Age of students, Gender: 1(male) / 2(female).

2.2.1 The architecture of the proposed system

In the study, machine learning algorithms were used and the accuracies of these algorithms were compared. Some of the data was used to test the accuracy of the model, and some of it was used as training data. The proposed architecture is shown in Figure 1.

The data collected according to the architecture seen in Figure 1 were subjected to pre-processing. In the pre-processing phase, missing and incorrect data have been cleared. The quality of the data greatly influences the outcome of the estimation. This means that pre-processing plays an important role in the model [16]. After the dependent variable was determined, the data were divided into two as training and test data. Finally, the performance of the model was evaluated.

Cross-validation is generally preferred because it allows the model to be trained with multiple training and test groups. One of the groups is used as a test set and the rest are used as a training set. Each group should be repeated in the same way to train the model. This gives you a better idea of how well the model will perform with new data and is essential for the accuracy of the model. In general, the dataset is divided into 10 equal parts. 1 of them is used for testing and 9 of them are used for training. In this study, the dataset is divided into 10 equal parts. At each iteration, test results are calculated, averaged, and performance is achieved.

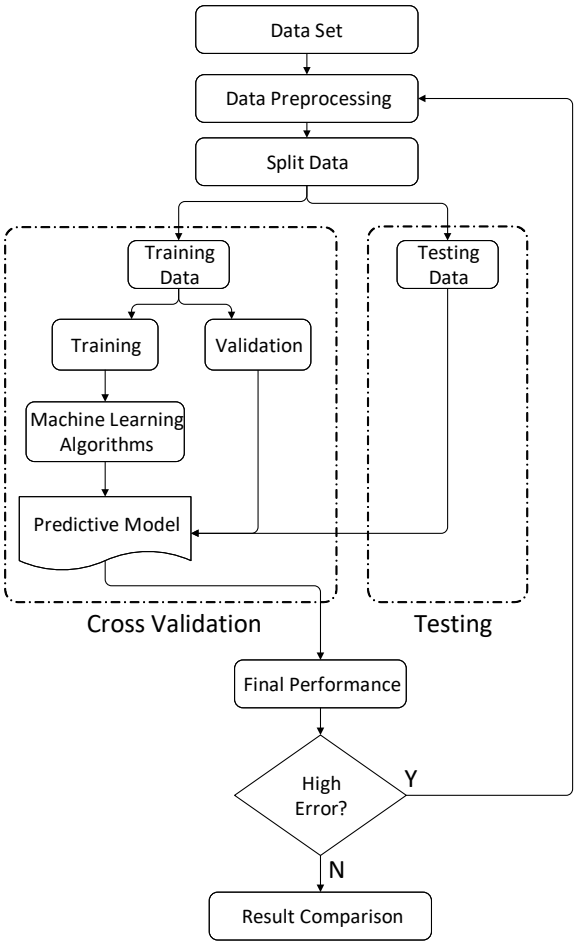

Figure 1. The architecture of the proposed system

2.2.2. Random Forest

Random forest algorithm proved to be the most effective for the purpose of predicting students' success in this study. The random forest algorithm is based on the principle of using Decision Trees and Bagging methods together and enters into Ensemble methods. Ensemble methods are a machine learning concept that trains multiple models using a learning algorithm. A basic learning algorithm should be chosen to use Bagging or Boosting, which are among these methods. In the bagging method, new trees are created by combining the properties in different ways and the most popular class is selected from the trees created [17].

Random forest is a flexible machine learning method used for regression or classification problems. In its simplest form, the random forest is the combination of a large number of decision trees to obtain a more accurate estimate.

Random forest performs training by randomly selecting a large number of different subsets from both the data set and the feature set to solve the overfitting problem that often occurs in decision trees. Thus, each of the numerous decision trees makes an estimate.

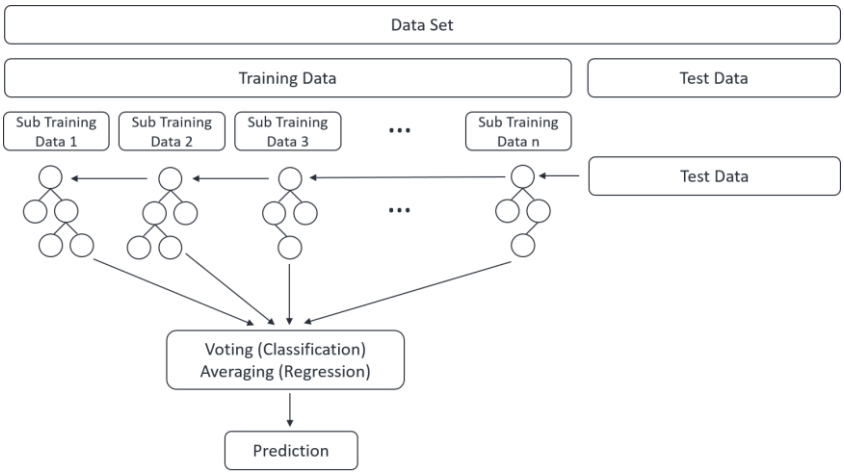

Figure 2. Random Forest

In the regression problem, while the estimates of the decision trees are averaged, the prediction with the highest number of votes in the classification problem is selected. In this model, overfitting is reduced as training is carried out on different data sets. It can also be used to identify the most important features in the data set. The model is shown in figure 2.

2.2.3. Exploratory Data Analysis

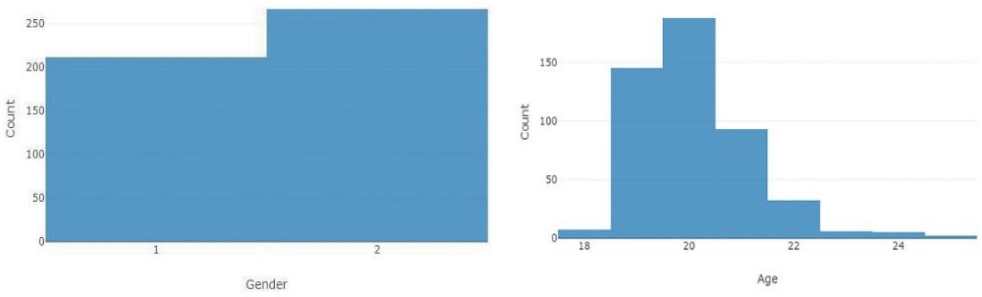

Figure 3. Gender and age histogram graphs

In the data set, there are 478 students, 211 males, and 267 females, according to Figure 3. While the age range is between 18 and 25, the density is between 19 and 21.

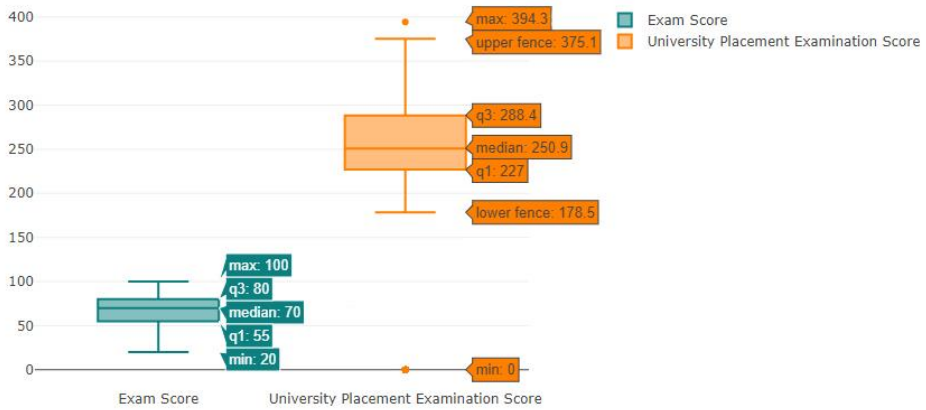

Figure 4. Exam score and university placement score box graphs

Figure 4 shows the box graph and data distribution of the grades of the Atatürk Principles and Revolution History course and the university placement scores.

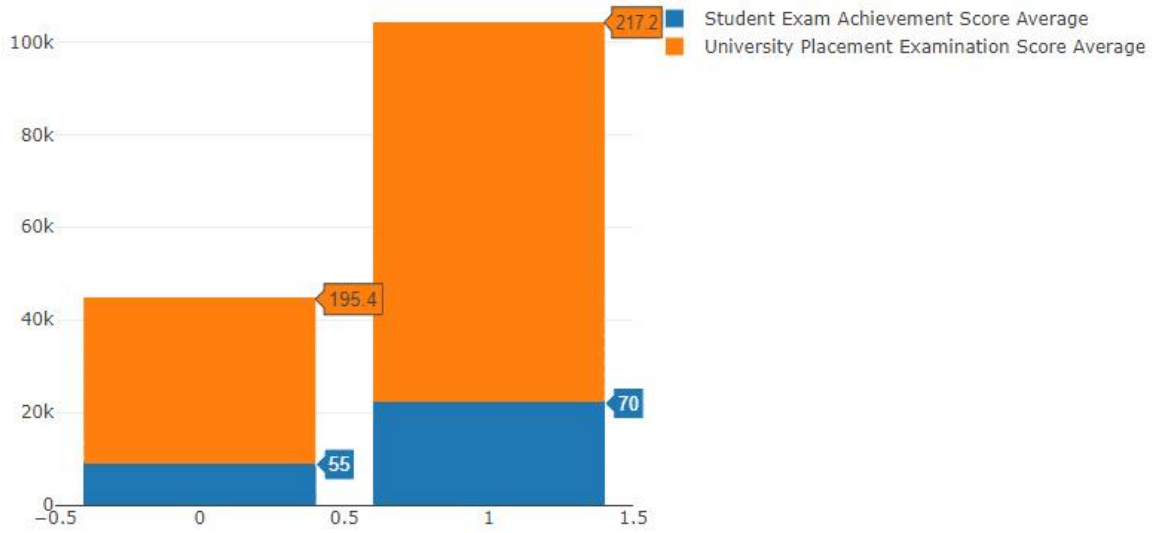

Figure 5. Grade averages of successful and unsuccessful students

According to Figure 5, the students who passed the Atatürk Principles and History of Revolution course have a mean score of 70 and a university placement score of 217.2. While the grade point average of the failed students in this course was 55, the mean score of university placement was 195.4.

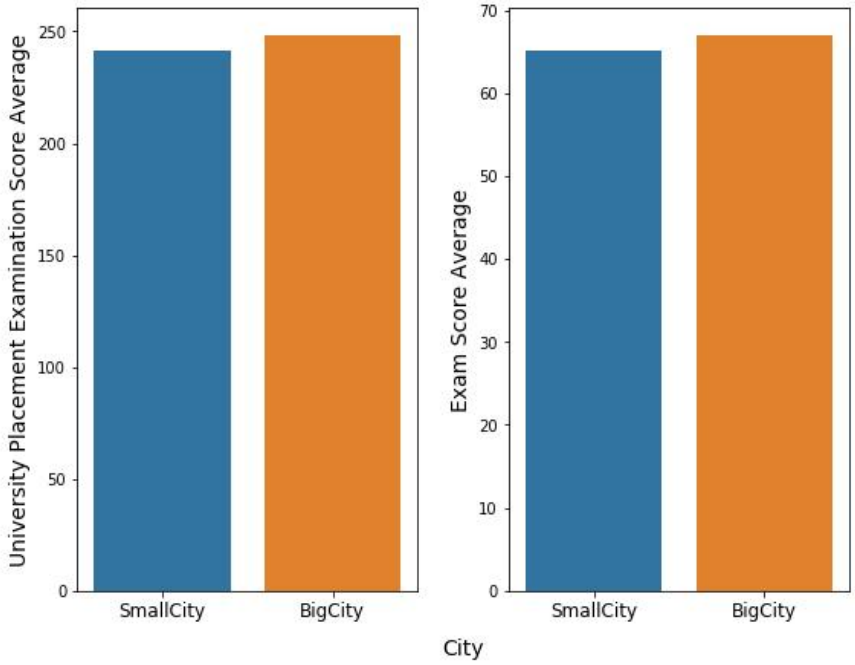

Figure 6. Student achievement by city size

Figure 6 shows that university placement point averages and grade point averages of students from metropolitan cities are higher than in small cities. City size was determined based on the city category in Turkey.

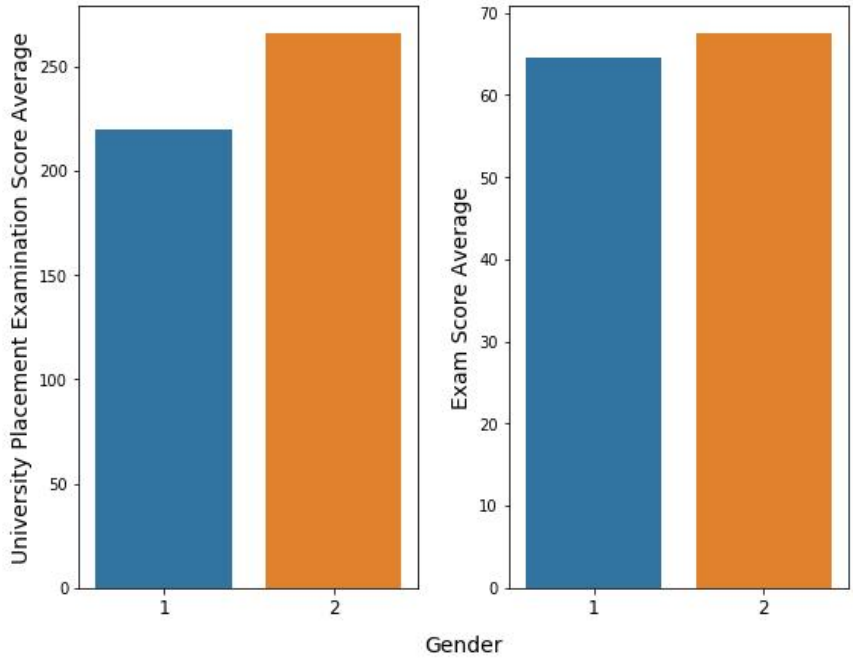

Figure 7. Student achievement by gender

According to Figure 7, it can be said that women are more successful than men. According to the data in this study, it has been observed that both the university placement score average and the exam score average of women are higher than men.

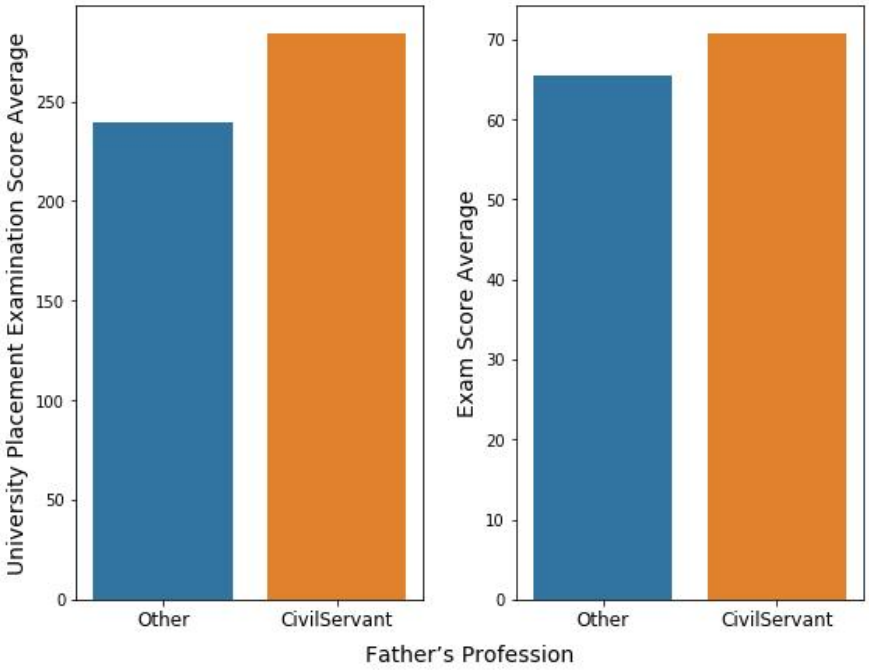

Figure 8. Student success by father's profession

According to Figure 8, the university placement point average and grade point average of the students whose father's profession is a civil servant are higher than those of other students. According to the available data, it can be said that the students whose father is a civil servant are more successful than other students.

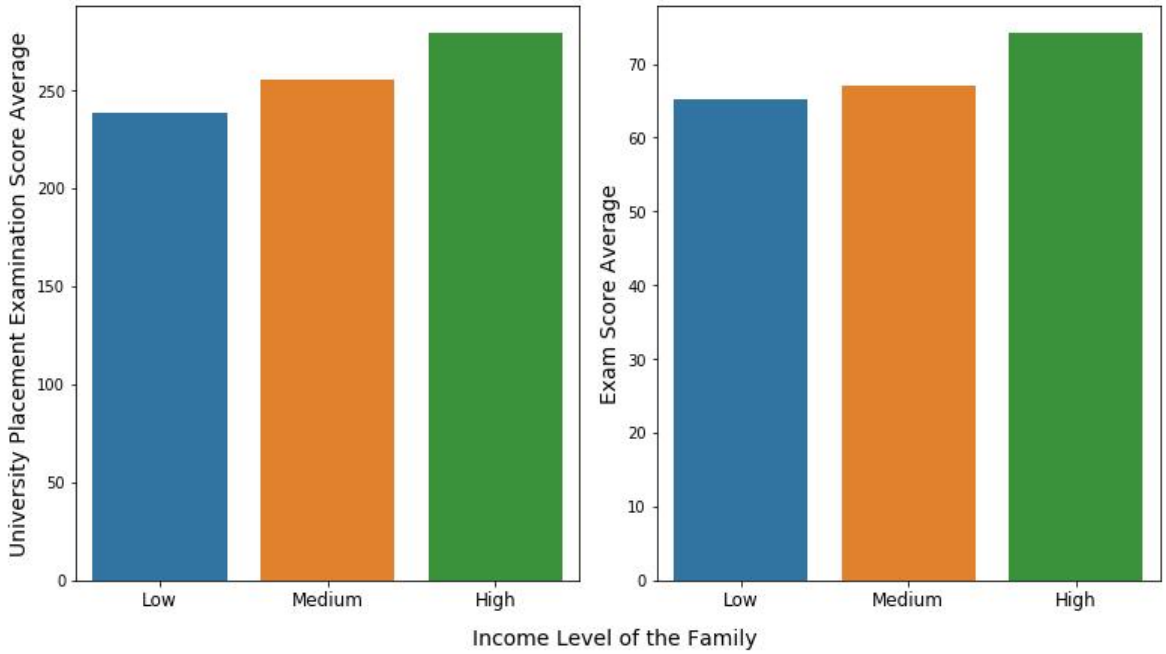

Figure 9. Student success according to family income

Figure 9 shows that, as the family income increases, university placement point averages and grade point averages increase. According to the available data, it is observed that students' academic success increased as the financial situation improved.

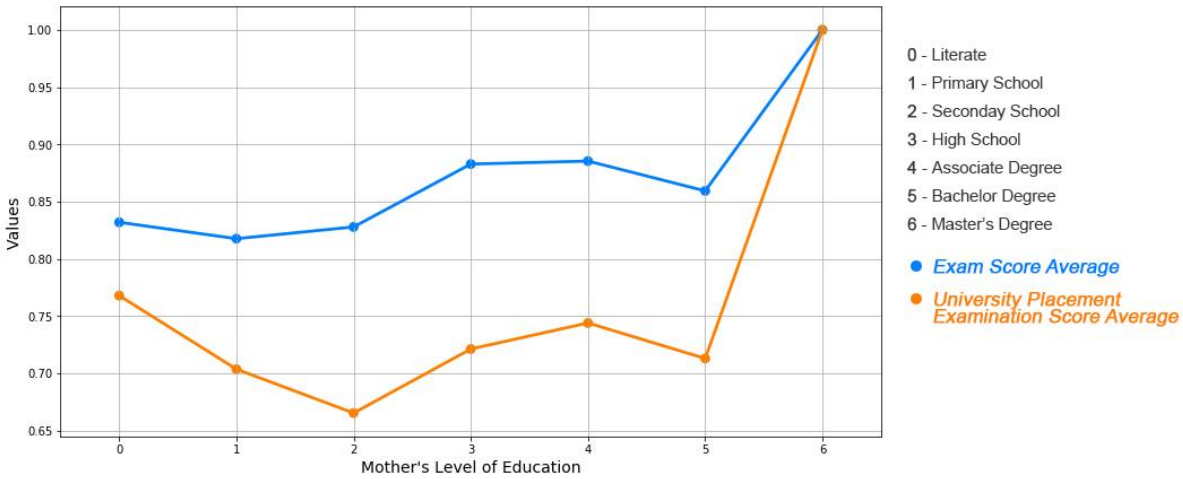

Figure 10. Student success according to mother's education level

Figure 10 shows the effect of mothers' education on student achievement. Figure 11 shows the effect of fathers' education on student achievement. Overall, student achievement is on an upward trend with the education level of the father.

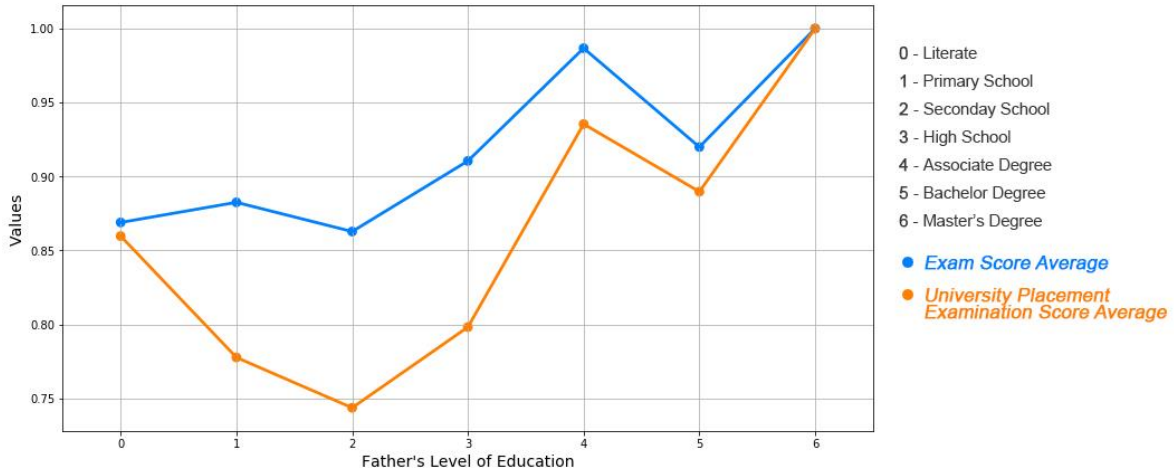

Figure 11. Student success according to father's education level

3. THE RESULTS AND DISCUSSION

During the research, 8 machine learning algorithms were tried. Among the algorithms applied in the research process, according to Figure 12, the random forest has the highest accuracy rate according to other estimation methods.

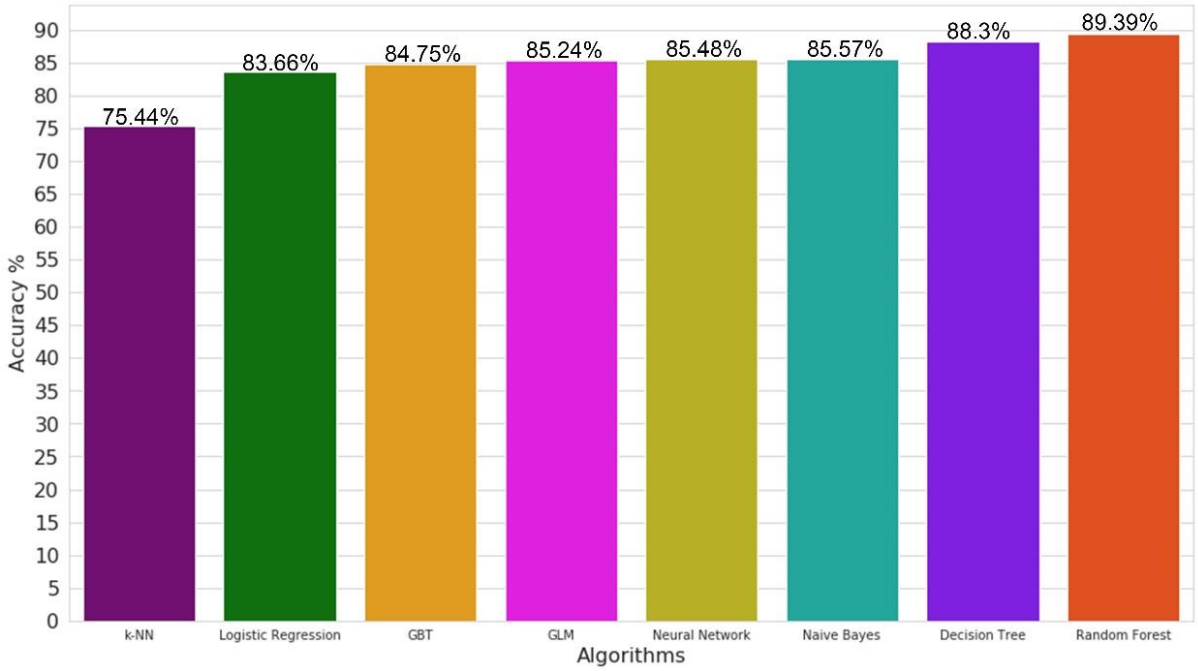

Figure 12. Accuracy percentages of algorithms

According to Figure 12, the ratio of correctly estimated samples to the number of all samples according to Table 3 is 89.39%.

This ratio represents the total accuracy. It is represented as TP true positives, TN true negatives, FN false negatives, and FP false positives.

Table 3.

Results of Random Forest

Accuracy: 89.39%+/- 5.42% (micro average: 89.34%)	**True 1**	**True 0**	**Total**	**Class Precision**
Pred. 1	182 (TP) Correct Decision	38 (FP) Type I error	220 (P')	82.73%
Pred. 0	1 (FN) Type II error	145 (TN) Correct Decision	146 (N')	99.32%
Total	183 (P)	183 (N)	366 (P+N)	
Class Recall	99.45%	79.23%		

TP: true positives, TN: true negatives, FN: false negatives, and FP: false positives. Recall: The ratio of correctly predicted positive samples to the number of samples in the true positive class. Precision: It is the ratio of correctly predicted positive samples to the number of samples estimated in the positive class.

3.1. Evaluation of Performances

While looking at the outputs of the algorithms used in the model, the performance of the obtained measurements is evaluated. For each method, true positive, false positive, true negative, and false-negative results were examined.

True Positive (TP): In the actual case, it means the correct estimation of the students who pass the course.

False Positive (FP): In reality, the students who failed the course were misidentified as successful. (Tip I Error).

True Negative (TN): Students who fail in real situations have been estimated as unsuccessful.

False Negative (FN): In reality, the students who succeeded in the course were misidentified as unsuccessful. (Tip II Error).

Accuracy: The ratio of correctly estimated samples to the number of all samples. That is, the test is the rate of total correct diagnoses.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N} \tag{1}$$

Precision: It is the ratio of correctly predicted positive samples to the number of samples estimated in the positive class.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{P'} \tag{2}$$

Sensitivity: It is the ratio of correctly predicted positive samples to the number of samples in the true positive class.

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{3}$$

Specificity: It is the ratio of correctly predicted negative samples to the number of samples in the true negative class [11].

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{TN}{N} \tag{4}$$

The purpose of data mining methods is to obtain meaningful information from the data, and the meaningful information obtained from the collected data can contribute positively to the development of the educational process [18]. The use of data mining techniques in higher education institutions has a significant and positive effect when viewed as a tool that can help to find the most appropriate solutions [19]. The use of these techniques for educational purposes is a promising area that aims to develop methods of discovering data and discovering meaningful patterns from computational training environments [20]. Therefore, the results in this process can provide invaluable support in decision-making. An example of this is the identification of groups of learners who exhibit a similar pattern of behavior, where the aim is to identify similar groups of students in terms of learning preferences, personal characteristics, and individual differences [21].

The data stored in the universities' learning and content management systems (LMS / CMS) led to the accumulation of large amounts of data related to the learning process [22]. Also, Massive open online courses (MOOCs) have attracted millions of students and offer the opportunity to apply and develop machine learning methods to improve student's learning outcomes and use the collected data [23].

Online learning environments, unlike traditional classroom environments, allow students to record the traces they leave behind while performing their learning activities. These traces can be the number of students entering and leaving the online environment, the interaction with the course materials, the answer to a question on the discussion forum. Considering the increase in the number of students enrolled in online environments, it is known that a significant amount of data related to learning processes are recorded in online databases. However, the use of these recorded data to improve the educational process is limited to simple graphs and descriptive statistics [24]. Processing this data will improve the quality of education and training activities. Besides, data mining methods used to detect hidden patterns and trends in databases in different areas have significant potential in the analysis of educational data. Data mining techniques applied in educational environments can be used in student support and feedback, evaluation of learning hypothesis, early warning systems, performance estimations, learning technologies, and future learning practices [25]. In this way, the instructors can use this information to monitor the students' development process and produce valuable information on the development of appropriate intervention methods for students who have problems. At the same time, this data can be used to automatically classify students in adaptive learning environments or to make automatic adaptations to similar student groups. It is possible to take precautions by informing students who are thought to have failed at the beginning of the process. Exploring factors beyond the control of students and teachers can be beneficial for education and training policymakers in the long term.

4. CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

In this article, student success was predicted by machine learning methods in a data set that includes socio-demographic variables of students taking distance education courses. According to these analyses, students coming from big cities are more successful than those from small towns. Women's successes are higher than men's. Students whose father’s profession is a civil servant are more successful than other students. Increased family income is also accompanied by an increase in student achievement. Furthermore, in general, student achievement tends to increase with the education level of the father. The random forest algorithm was the algorithm with the best prediction success of 89.39%. The accuracy of estimating whether the students were successful or not was considered as the essential factor. The model used for student success prediction can be helpful in future studies. In subsequent

studies, new data and variables should be added to increase the success of the model. A better accuracy rate can be obtained for estimation as the number of data increases.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] C. Dede, "Rethinking how students learn," in *Comparing frameworks for 21st century skills.*, J. A. Bellanca and R. S. Brandt, Eds. 2010, pp. 51–76.

[2] E. H. Toytok and S. Gürel, "Does Project Children's University increase academic self-efficacy in 6th graders? A weak experimental design," *Sustain.*, vol. 11, no. 3, Feb. 2019, doi: 10.3390/su11030778.

[3] E. P. Frank and N. Pharo, "Academic Librarians in Data Information Literacy Instruction: A Case Study in Meteorology."

[4] M. Palmer, "ANA Marketing Maestros: Data is the New Oil." [Online]. Available:https://ana.blogs.com/maestros/2006/11/data_is_the_new.html .Accessed on: Sep. 18, 2020.

[5] P. Rotella, "Is Data The New Oil?" [Online]. Available:https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#3919a64b7db3. Accessed on: Sep. 18, 2020.

[6] R. Schapire, "COS 511, Spring 2014: Home." [Online]. Available:https://www.cs.princeton.edu/courses/archive/spring14/cos511/ Accessed on: Sep. 18, 2020.

[7] P. Golding and O. Donaldson, "Predicting academic performance," in *Proceedings - Frontiers in Education Conference, FIE*, 2006, pp. 21–26, doi: 10.1109/FIE.2006.322661.

[8] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification," *Bulg. Acad. Sci. Cybern. Inf. Technol.* vol. 13, no. 1, 2013, doi: 10.2478/cait-2013-0006.

[9] A. Hargreaves, *Teaching in the knowledge society: Education in the age of insecurity*. Teachers College Press, 2003.

[10] C. Petit, R. Bezemer, and L. Atallah, "A review of recent advances in data analytics for post-operative patient deterioration detection," *Journal of Clinical Monitoring and Computing*, vol. 32, no. 3. Springer Netherlands, pp. 391–402, Jun. 01, 2018, doi: 10.1007/s10877-017-0054-7.

[11] O. F. Akmese, G. Dogan, H. Kor, H. Erbay, and E. Demir, "The Use of Machine Learning Approaches for the Diagnosis of Acute Appendicitis," *Emerg. Med. Int.*, vol. 2020, pp. 1–8, 2020, doi: 10.1155/2020/7306435.

[12] P. Sinha and P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," *Int. J. Eng. Res. Technol.*, vol. 4, no. 12, pp. 608–12, 2015.

[13] E. Zahn, "Informationstechnologie und Informationsmanagement," *Allg. Betriebswirtschaftslehre*, vol. 2, pp. 376–428, 2001.

[14] E. Uzun, "İnternet tabanlı bilgi erişimi destekli bir otomatik öğrenme sistemi," 2007.

[15] A. F. KOCAMAZ, "Makine Öğrenmesi Tabanlı Bir Uzman Sistem Tasarımı," 2012.

[16] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9728, pp. 420–427, doi: 10.1007/978-3-319-41561-1_31.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[18] J. Han, M. Kamber, and J. Pei, "Introduction," in *Data Mining*, 2012, pp. 1–38.

[19] A. Van Barneveld, K. E. Arnold, and J. P. Campbell, "Analytics in Higher Education : Establishing a Common Language," *researchgate.net*, no. January, pp. 1–11, 2012.

[20] J. T. Avella, M. Kebritchi, S. G. Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review," *J. Asynchronous Learn. Netw.*, vol. 20, no. 2, 2016, doi: 10.24059/olj.v20i2.790.

[21] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, *Handbook of educational data mining*. 2010.

[22] M. M. A. Tair, "Mining Educational Data to Improve Students ' Performance : A Case Study Mining Educational Data t o Improve Students ' Performance : A Case Study," *iugspace.iugaza.edu.ps*, vol. 2, no. October, 2015.

[23] K. Lee, "Large-Scale and Interpretable Collaborative Filtering for Educational Data," *ML4ED KDD Work.*, pp. 1–7, 2017.

[24] L. Ali, M. Asadi, D. Gašević, J. Jovanović, and M. Hatala, "Factors influencing beliefs for adoption of a learning analytics tool: An empirical study," *Comput. Educ.*, vol. 62, pp. 130–148, 2013, doi: 10.1016/j.compedu.2012.10.023.

[25] W. Greller and H. Drachsler, "Translating learning into numbers: A generic framework for learning analytics," *Educ. Technol. Soc.*, vol. 15, no. 3, pp. 42–57, 2012.

Text of the article was accepted by Editorial Team 21.09.2020

ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ НАВЧАЛЬНИХ ДОСЯГНЕНЬ СТУДЕНТІВ ЗАКЛАДІВ ВИЩОЇ ОСВІТИ

Омер Фарук Акмеше
кандидат комп'ютерних наук, старший викладач кафедри комп'ютерних технологій
Університет Хітіта, м. Чорум, Туреччина
ORCID ID 0000-0002-5877-0177
ofarukakmese@hitit.edu.tr

Хакан Кьор
кандидат комп'ютерних наук, доцент кафедри комп'ютерної інженерії
Університет Хітіта, м. Чорум, Туреччина
ORCID ID 0000-0002-8314-9585
hakankor@hitit.edu.tr

Хасан Ербей
кандидат комп'ютерних наук, професор кафедри комп'ютерної інженерії
Університет Турецької авіаційної асоціації, м. Анкара, Туреччина
ORCID ID 0000-0002-7555-541X
herbay@thk.edu.tr

Анотація. Метод машинного навчання, який є підрозділом штучного інтелекту і дає можливість прогнозувати за допомогою математичних і статистичних операцій, часто використовується як в освіті, так і в усіх сферах життя. Сьогодні безперервно записуються мільйони даних і відбувається їх величезне накопичення. Хоча таке накопичення даних зростає в геометричній прогресії, кількість аналітиків недостатня, а їх можливості обробки обмежені. Ми живемо у вік інформації, точніше сказати, в епоху даних. Зберігаючи та накопичуючи дані, важливо виявляти значущі взаємозв'язки і тенденції, а також робити прогнози на майбутнє. Важливо проаналізувати дані, отримані в процесі навчання, і оцінити успіхи студентів і фактори, що впливають на їх успішність. Такий аналіз може сприяти майбутньій навчальній діяльності. У цьому дослідженні використовувався набір даних, що містить соціально-демографічні змінні студентів, які навчаються дистанційно в Hitit University. Автори оцінювали успішність студентів за допомогою таких демографічних і соціальних змінних, як-от: вік, стать, місто, сімейний дохід, рівень сімейної освіти. Основна мета - надати студентам на початку навчального процесу інформацію про їх передбачувані академічні досягнення. У такий спосіб можна на початку навчального процесу шляхом інформування студентів, які передбачувано можуть відставати в навчанні, підвищити їх мотивацію до навчання. Диверсифікація і поліпшення цих даних може також підтримувати інші механізми прийняття рішень у процесі навчання. Крім того, були досліджені фактори, що впливають на академічну успішність студентів, і оцінені результати навчання студентів. Успішність прогнозів порівнювалась з використанням різних алгоритмів машинного навчання. У результаті аналізу було визначено, що алгоритм *Випадковий ліс* краще за інших передбачав успішність студентів.

Ключові слова: машинне навчання в освіті; освіта дорослих; освітній інтелектуальний аналіз даних.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УЧЕБНЫХ ДОСТИЖЕНИЙ СТУДЕНТОВ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

Омер Фарук Акмеше
кандидат компьютерных наук, старший преподаватель кафедры компьютерных технологий
Университет Хитита, г. Чорум, Турция
ORCID ID 0000-0002-5877-0177
ofarukakmese@hitit.edu.tr

Хакан Кёр
кандидат компьютерных наук, доцент кафедры компьютерной инженерии
Университет Хитита, г. Чорум, Турция
ORCID ID 0000-0002-8314-9585
hakankor@hitit.edu.tr

Хасан Эрбэй
кандидат компьютерных наук, профессор кафедры компьютерной инженерии
Университет Турецкой авиационной ассоциации, г. Анкара, Турция
ORCID ID 0000-0002-7555-541X
herbay@thk.edu.tr

Аннотация. Метод машинного обучения, который является подразделом искусственного интеллекта и дает возможность прогнозировать с помощью математических и статистических операций, часто используется как в образовании, так и во всех сферах жизни. В настоящее время очевидно, что непрерывно записываются миллионы данных и происходит их огромное накопление. Хотя такое накопление данных растет в геометрической прогрессии, количество аналитиков недостаточно и их возможности для обработки ограничены. Мы живем в век информации, точнее сказать, в эпоху данных. Сохраняя и накопляя данные, очень важно устанавливать значимые взаимосвязи и тенденции, а также делать прогнозы на будущее. Важно проанализировать данные, полученные в процессе обучения, и оценить успехи студентов и факторы, влияющие на их успеваемость. Такой анализ может способствовать будущей учебной деятельности. В этом исследовании использовался набор данных, включающий социально-демографические переменные студентов, обучающихся дистанционно в Hitit University. Авторы оценивали успеваемость студентов с помощью демографических и социальных переменных, таких как возраст, пол, город, семейный доход, уровень семейного образования. Основная цель - предоставить студентам информацию об их предполагаемых академических достижениях в начале учебного процесса. Таким образом, в начале учебного процесса успеваемость учащихся может быть повышена путем информирования студентов, которые, предполагаемо могут отставать в обучении. Диверсификация и улучшение этих данных может также поддерживать другие механизмы принятия решений в процессе обучения. Кроме того, были исследованы факторы, влияющие на академическую успеваемость студентов, и оценены результаты обучения студентов. Успешность прогнозов сравнивалась с использованием различных алгоритмов машинного обучения. В результате анализа было определено, что алгоритм *Случайный лес* лучше других предсказывал успеваемость студентов.

Ключевые слова: машинное обучение в образовании; образование взрослых; образовательный интеллектуальный анализ данных.