Proceedings of the workshops on ACeSYRI 2023

June 20-22, 2023
Žilina, Slovakia

September 26-28, 2023
Nancy, France

Co-funded by the Erasmus+ Programme of the European Union

ACeSYRI

Co-funded by the Erasmus+ Programme of the European Union

Table of Contents

ANALYSIS OF EDUCATIONAL DATA USING DATA-MINING TOOLS

Lylia Kurmasheva[1], Vitaly Levashenko[2], Ildar Kurmashev[1], and Valentina Kulikova[1]

[1]*M.Kozybaev North Kazakhstan University, Petropavl, Kazakhstan, lb_kurmasheva@mail.ru*
[2]*Dept. of Informatics, University of Zilina, Zilina, Slovakia*

Abstract- The changes currently taking place in the world impose qualitatively new requirements for systems and their management. Process management in educational institutions requires reliable information, for this purpose a large amount of data is collected and analyzed. The paper presents the results of a study to identify factors that influence the learning outcome of students, that is, the probability of completing a full course of OP using regression analysis and building a neural network. The study was conducted on the basis of a regional university of Kazakhstan.

Keywords- data analysis, management in education, regression analysis, neural networks.

I. INTRODUCTION

The education system has always been considered a conservative industry, and is currently crossing the threshold of fundamental changes. For example, many believed that "digital learning" is just a fashionable topic that will pass over time, but now we are seeing the widespread development of online learning platforms. Microsoft has published data that more than 30 million people in 249 countries gained access to new digital skills during the COVID-19 pandemic. More than 10,000 Kazakhstanis are among the participants of the program [1]. School graduates have become more likely to choose online courses for obtaining a profession [2]. In developing countries, many experts believe that digital learning will lead to mass education among the population, because the costs of creating online platforms, mobile applications are significantly lower than the construction of university educational buildings, dormitories [3]. People have increasingly begun to prefer online courses rather than studying in educational institutions. The Digital Kazakhstan program notes that "Digitalization is significantly ahead of the existing system of production requirements for the composition of professions employed in the labor market. The lack of an operational link between the labor market and the education system can simultaneously lead to the training of no longer in demand personnel and the release of personnel in disappearing professions" [4]. Based on the above, it can be argued about the impending crisis of modern universities. It is obvious that not all universities will remain relevant and in demand in the world community, but only those who responded to global challenges in time by studying the behavior of applicants studying under changing conditions. Thus, the analysis of educational data is always relevant in scientific and practical terms.

Currently, many universities invest in predictive analytics, which is provided by data generated by educational activities on the Internet, survey results, statistical indicators, and also relies on access to external consulting experts. Mathematical modeling, machine learning and artificial intelligence are increasingly being used to analyze educational data, optimize services and support students. Many universities are introducing chatbots based on artificial intelligence - to assist students when applying for admission, settling in dormitories, getting advice on various issues, etc., thereby reducing costs, as well as to obtain data on students' problems, their preferences and suggestions. For example, the project "Digital Admissions Committee" was implemented at the M.Kozybayev National Research University, within the framework of which a telegram channel (@skgu_pk_2020) and an official chatbot of the admissions committee (@Inform_PKBot) were created. Around the clock, applicants could get the information they were interested in. The London-based Edge company has invested about 3.4 million euros in the

innovative educational technology platform Springpod for student recruitment. The online platform allows employers and educational service providers to orient themselves, inform and attract students at the stage of career choice [5]. Such tools are created with the aim of the best career guidance for applicants, because many universities face the problem of large deductions of students after the end of the first year. A lot of works are devoted to this problem. For example, in [6] the authors use the predictive properties of "balanced" decision trees in identifying students who are at risk for academic performance. In [7] a methodology for predicting student performance based on cluster analysis methods is proposed. In [8] the influence of family income on the results of entrance exams, the choice of university and strategies for preparing for entrance exams is investigated. The work is based on conducting a survey and analyzing its results.

II. MATERIALS AND METHODS

During the study, statistical data for 2006-2017 for a regional university of the Republic of Kazakhstan and data taken from the website of the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan were used as initial data [9] (Table I). The sample size for students was 14,877 students.

TABLE I
Data for analysis

X_1	Place of residence	X_1=1 if the place of residence is Petropavlovsk, X_1=0 otherwise
X_2	The average score of the school certificate	The arithmetic mean of all grades given in the document on education
X_3	Characteristics of the form of educating	X_3=1, if the student is studying on a budget basis, X_3=0 - for students on a commercial basis
X_4	GPA	Average academic performance at the university
X_5	The percentage of the population with incomes below the subsistence minimum, as a percentage	Indicator for Kazakhstan
X_6	Unemployment rate, as a percentage	Indicator for Kazakhstan
Y	the result of the educating	Y=1 if the student was expelled, Y=0 if the student graduated

Multiple correlation and regression analysis was applied to analyze the factors affecting the student's probability of graduating from university, and a neural network was built and trained. When building a neural network, the Levenberg-Marquardt method was used, according to which learning occurs "by epochs". In this case, the network error is considered for the entire training epoch and the network parameters change when all the elements of the training set have already been presented to the network [10].

III. RESULTS OF WORK

Having constructed a matrix of paired correlation coefficients in the Statistica package, it was revealed that some of the regressors are collinear with each other (X_5 and X_6,), i.e. there is a suspicion of the presence of multicollinearity. To eliminate the multicollinearity of regressors, a step-by-step regression analysis was performed, as a result of which the number of factors decreased. All parameters of the resulting model are statistically significant, the equation is significant ($p - value = 0.3 \cdot 10^{-8}$):

$$Y = 0.061 \cdot X_2 + 0.052 \cdot X_3 - 0.63 \cdot X_4 - 0.08 \cdot X_5 + \varepsilon \quad (1)$$

The coefficient of determination ($R^2 = 0.81$) proves the adequacy of the model to the process under study. The analysis of random residuals indicates the absence of autocorrelation (the Darbin-Watson criterion dw = 2.09 at critical values 0.79 and 1.44), and also checked the normality of their distribution ($p - value = 0.91$).

When building a neural network, the available statistical data were divided into 3 subsamples: training, control and test (the division was chosen as follows: 70% of the data was

used for training, 15% of the data was used for verification, 15% of the data was used during testing). The neural network training results were determined by the root-mean-square error and the value of the correlation coefficient between the network outputs and the target values (Table II).

TABLE II
Values of neural network training results by samples

	Root-mean-square error	Correlation coefficient
Training sample	0,061	0,692
Control sample	0,056	0,728
Test sample	0,065	0,638

According to the graph (Fig.1), we can notice a decrease in error during the learning process at each epoch. The training was stopped when the error on the test set stopped decreasing (at the 12th epoch).

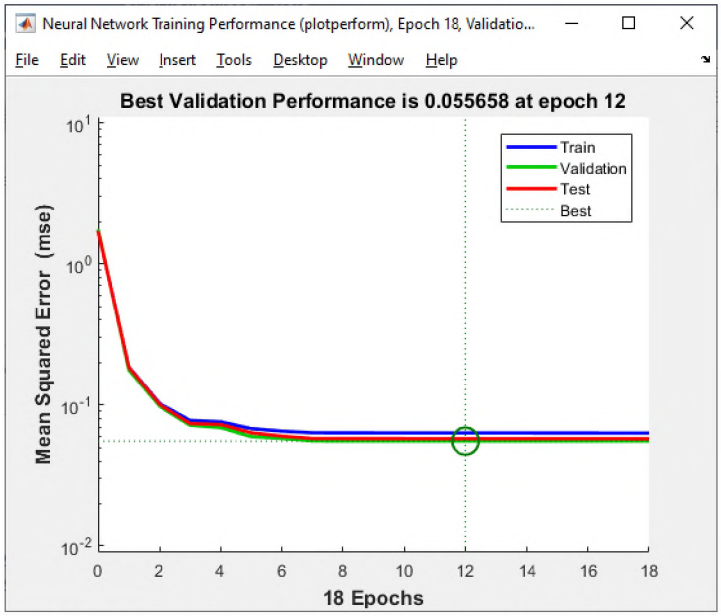

Fig. 1 Graph of the effectiveness of network training

IV. CONCLUSION

Based on the results of the data analysis, the following conclusions were made: with a decrease in the average academic achievement score (GPA), the probability of a student's expulsion will increase, this factor has the greatest impact on the learning outcome. Other factors also have an impact on the learning outcome, but their impact was the least significant. The study used a limited amount of data, in the case of a large number of data, it was possible to obtain more accurate and interesting conclusions.

Conducting this kind of research is necessary to provide the administrative management of the university with information about the current state, as well as to identify problems and develop concrete actions based on the results of the analysis to solve them.

ACKNOWLEDGMENT

This research work was realized in the project Erasmus+ *Advanced Centre for PhD Students and Young Researchers in Informatics* (ACeSYRI), reg.no. 610166-EPP-1-2019-1-SK-EPPKA2-CBHE-JP.

Co-funded by the Erasmus+ Programme of the European Union

REFERENCES

[1] Media holding "Atameken business" (07.04.21). 30 million people worldwide have mastered digital skills during the COVID-19 pandemic. https://inbusiness.kz/ru/news/30-millionov-chelovek-po-vsemu-miru-osvoili-cifrovye-navyki-v-techenie-pandemii-covid-19

[2] E.Yerohina, "School graduates have become more likely to choose online courses to get a profession" (24.04.2023). https://skillbox.ru/media/education/vypuskniki-shkol-stali-chashche-vybirat-onlaynkursy-dlya-polucheniya-professii/

[3] D.Kochergin, E.Zhernov, "Experience of digitalization of higher education in the USA", The international scientific and practical conference "Vocational education youth employment: XXI century. Training of personnel for the digital economy", 2 (34), 2019.

[4] Resolution of the Government of the Republic of Kazakhstan dated December 12, 2017 No. 827 on approval of the State Program "Digital Kazakhstan".

[5] R.Vignesh (May 17, 2023), "London-based edtech startup Springpod bags €3.4M from Edge to help young people choose right career", Silicon Canals. https://siliconcanals.com/news/startups/london-springpod-bags-3-4m-from-edge/

[6] N. Nakaryakova, S.Rusakov, O.Rusakova, " Prediction of the risk group (by academic performance) among first course students by using the decision tree method," Applied mathematics and management issues, no.4, pp. 121-136, October 2020

[7] V.Shevchenko, "Prediction of the student performance based on cluster analysis methods", Bulletin of the Kharkiv National Automobile and Road University, no.68, pp.15-18, 2015.

[8] M.Yudkevich, I. Prakhov, "Effect of Family Income on USE Performance and the Choice of University," Voprosy obrazovaniya / Educational Studies Moscow, National Research University Higher School of Economics, issue 1, pp. 126-147, 2012.

[9] Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. https://stat.gov.kz/region/264023/dynamic

[10] D. Binu, B.R. Rajakumar, "Artificial Intelligence in Data Mining", Academic Press, Pages 109-131, 2021.