Journal of Infrastructure, Policy and Development 2024, 8(8), 4165.
https://doi.org/10.24294/jipd.v8i8.4165

Article

Introducing machine learning to analyze factors aimed at successful development of the individual social qualities

Tatyana Panfilova[1,2], Vadim Tynchenko[2,3], Vladimir Nelyub[2,4], Anna Glinscaya[2,5,*], Alexey Borodulin[2], Andrei Gantimurov[2], Yadviga Tynchenko[1,2]

[1] Department of Technological Machines and Equipment of Oil and Gas Complex, School of Petroleum and Natural Gas Engineering, Siberian Federal University, 660041 Krasnoyarsk, Russia
[2] Artificial Intelligence Technology Scientific and Education Center, Bauman Moscow State Technical University, 105005 Moscow, Russia
[3] Information-Control Systems Department, Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
[4] Scientific Department, Far Eastern Federal University, 690922 Vladivostok, Russia
[5] Department of Information Economic Systems, Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
*** Corresponding author:** Anna Glinscaya, anna_glinskaja@rambler.ru

CITATION

Panfilova T, Tynchenko V, Nelyub V, et al. (2024). Introducing machine learning to analyze factors aimed at successful development of the individual social qualities. Journal of Infrastructure, Policy and Development. 8(8): 4165. https://doi.org/10.24294/jipd.v8i8.4165

ARTICLE INFO

Received: 10 January 2024
Accepted: 26 April 2024
Available online: 6 August 2024

Abstract: The paper considers an important problem of the successful development of social qualities in an individual using machine learning methods. Social qualities play an important role in forming personal and professional lives, and their development is becoming relevant in modern society. The paper presents an overview of modern research in social psychology and machine learning; besides, it describes the data analysis method to identify factors influencing success in the development of social qualities. By analyzing large amounts of data collected from various sources, the authors of the paper use machine learning algorithms, such as Kohonen maps, decision tree and neural networks, to identify relationships between different variables, including education, environment, personal characteristics, and the development of social skills. Experiments were conducted to analyze the considered datasets, which included the introduction of methods to find dependencies between the input and output parameters. Machine learning introduction to find factors influencing the development of individual social qualities has varying dependence accuracy. The study results could be useful for both practical purposes and further scientific research in social psychology and machine learning. The paper represents an important contribution to understanding the factors that contribute to the successful development of individual social skills and could be useful in the development of programs and interventions in this area. The main objective of the research was to study the functionalities of the machine learning algorithms and various models to predict the students's success in learning.

Keywords: education; learning opportunities; quality job; socio-economic development; social policies

1. Introduction

Throughout a person's life, the personal success concept changes with the increasing gained experience, which is demonstrated by the personal formation. As of today, any unambiguous concept of success is missing, everyone interprets it in the own way, but for most people this is the main goal of life based on achieving the success, no matter whether it is material wealth, or legacy left by a person. After achieving a set goal, a new one arises; otherwise, procrastination appears.

If the success concept is considered from a psychological point of view, then it becomes satisfaction arising after solving a problem. From the pedagogical point of view, success is a kind of simulating a certain situation, where a student is provided with the opportunity to achieve results. Today, most people are trying to achieve success in work, learning, and personal life in different ways. Besides, researchers found out that success should be permanent for a person to feel the ongoing satisfaction. Sustained success includes positive development, prosperity, well-being, and high performance (Uusiautti et al., 2021). With an increase in the number of achievements in learning and mastering new skills and abilities, the scope of realizing the own needs for career advancement is widening. The increasing career dynamics makes development and application of various career resources important in the successful career development (Haenggli et al., 2020).

Importance of education in the people's lives is changing, as they grow older. Education allows a person to share information, as well as acquired knowledge and skills within the society. In addition, a person is provided with examples of highly moral behavior and opportunity to master them during the learning process (Huffman, 1974). Education level depends on the teacher's qualification and the education system itself, which is a subject to borrowing from the other most successful countries (Harris, et al. 2011). The amount of information perceived by a student depends on the teacher's ability to present the material. The ability to learn depends on the personal qualities. Education effectiveness also depends on the student himself and on his personal qualities. Personal qualities are influencing not only in psychology, but also in the pedagogy, since attention, perception and memory are important in the learning process and affect speed, volume and quality of the acquired information (Berezan, 2019). This study was conducted to demonstrate the possibility of introducing machine learning as the auxiliary tool to analyze success factors in development of the individual social qualities in education for further use.

2. Review of literature

Modern conditions are characterized by intensified fight for the intellectual resources, which are one of the main advantages and serve as the basis for the enterprise well-being. To obtain higher education, modern youth carefully selects educational institution and area that is most interesting (Maringe, 2006). Success in learning creates competitive and in-demand employee in the labor market capable of self-development (Shafranov-Kutsev et al., 2019).

Many studies are devoted to identifying factors that influence the student's learning success. Each author identifies different factors influencing success in his opinion. For example, Smirnov (2004) assesses success in learning and refers to a number of factors. They include personal mental state, which depends both on the family where a person was raised and on himself; amount of knowledge acquired and mastered before entering the university; presence of the developed system of self-organization; purpose of choosing an educational institution; form of training; basis for admission; level of the educational process organization in an educational institution; qualification of the teaching staff and service personnel. Hodges (2018) examines relationships between the student and his teacher, classmates, parents and other persons, as well as the student's mental state with his mental characteristics.

Social-cognitive characteristics were discussed as the socioeconomic success predictors in the adultery demonstrated in education, professional status, and income.

Based on work related to the task value and choice in the real life (Sansone et al., 2000), the paper examined a situation, where potential relationships were independent of the major childhood factors, namely intelligence, parental socioeconomic status, and the adults' educational level. The study was conducted with the residents of Germany. The survey was performed with an interval of 20 years, which made it possible to more accurately identify relationship between childhood factors and further success. The childhood factors influence were moderated by education in the highly structured education systems such as Germany, since education was the strongest predictor of the socioeconomic success. Besides, the socio-cognitive characteristics, especially self-esteem, vocational aspirations, and academic self-concept, predicted the learning success. In addition, the socio-cognitive characteristics influence varied significantly depending on intelligence and the socio-economic background. However, professional aspirations, academic self-esteem in the academic direction, and academic interests demonstrated unique effects predicting not only the education, but also further socio-economic success of the adults (Becker et al., 2021; Volchenko et al., 2021).

Particular attention was paid to academic and social participation of the higher education students, particularly through their extracurricular activities and attraction to the decision-making processes, which were perceived as the critical mechanisms in their persistence to acquire higher education (Sá, 2023).

Sometimes a failure with a student, namely the unsatisfactory grades, has a detrimental effect on the mental state, after that a student becomes prone to dropping out. Several researchers differentiate between the student dropouts based on their social background, career choices, and institutional experiences (Gairin et al., 2014; Godor, 2017; Markos et al., 2019; Pusztai et al., 2019; Váradi et al., 2019; Zając et al., 2019). A Central European Research Project conducted in 2018 differentiated four factors that caused students to drop out of their studies using the cluster analysis (Kovács et al., 2019). The first factor was the missing opportunity to study on the paid basis due to impossibility to pay for the studies (Pusztai et al., 2019). The second factor included difficulties in learning due to the lack of support from the educational institution (Godor, 2017). The third factor was disillusionment with the higher education institution, or disappointment with the outdated materials and teaching style (Gairin et al., 2014). The fourth type was a hopeless situation appearing due to academic and financial difficulties (Hatos et al., 2019).

Scientists and teachers are still looking for new relationships between various factors in the student dropout. Some of them think that this is a step backwards, for example, sociologist Steven Seidman believes that dropping out is interruption in study interfering with the student's constant advancement. The others, on the contrary, say that expulsion provides an opportunity to find oneself and a new path in life, open the other opportunities and become successful independently without obtaining higher education at the educational institution. It is worth noting that a greater number of expulsions occur due to poor academic performance, lack of the ability to pay for education, as well as the very quality and conditions of the education provided (Pusztai et al., 2022).

It is practically impossible to compare countries or even institutions, not only because of differences in the educational policies, but also because of differences in methods for measuring the dropouts. It follows that factors influencing success in one educational institution may not have the same relationships in another educational institution.

The increased use of technology in education leads to the daily generation of large amounts of data that are becoming the focus of attention around the world. Obtaining educational data helps to create methods to extract interpretable, useful and new information, which can lead to a better understanding of students and the environments in which they learn (Algarni, 2016). The study of studies on data analysis in education revealed that some attributes were not taken into account, which entailed inaccuracy of the results obtained. The most important attributes to correctly categorize a student are: attendance, family size, disability, teacher competence, use of social media, psychological factors and teaching methods. Supplementing existing databases with relevant information will help to eliminate the inconsistencies in findings identified in past studies (Issah et al., 2023). Many researchers do not stop at the results obtained and seek to study student performance using a method that works with a dynamic environment (Ting et al., 2013).

3. Methodology

Introduction of the artificial neural networks is one of the most effective intelligent data mining approaches used to solve various types of problems in forecasting, classification and clustering. Increasing complexity of tasks is caused by increasing technological complexity of the processes being simulated and multifactorial search for effective models, as well as by the increasing volumes of data processed in constructing the regression models and requirements to accuracy of such models. These are the reasons in searching for approaches to increase efficiency of the regression and neural network models and the other models types. In this regard, combinations of the classifier models called the ensembles are used (Bukhtoyarov et al., 2023). This approach to the artificial neural networks was subsequently significantly developed and successfully used to solve a wide range of practical problems. They include recognition (Irvine et al., 2019; Li et al., 2018), medical diagnostics (Alzubi et al., 2019; Khan et al., 2022), seismic signal classification (Jia et al., 2021), data storage network load forecasting (Masich et al., 2022), acoustic signals classification (Barantsov et al., 2023), damage detection with the unmanned aerial vehicles (Romanova, 2022) and many others (Masich et al., 2022; Mikhalev et al., 2022). Different machine learning methods based on different ideas and assumptions solve the classification problems in recognition. Method selection in a particular task is determined by the expected classification accuracy and the situation under consideration, as well as by the purpose, for which the machine generates a recognition solution (Masich et al., 2022).

Methods of predictive analysis and value management were introduced in the study applied to group the similar data together, aa well as forecast making it possible to obtain a prediction on the time series values for the number of samples that corresponds to the given forecast horizon. This research uses an analytical platform that provides the basis for creating the end-to-end application solutions. Three algorithms were tested for each dataset: Kohonen maps, decision trees and neural networks. According to the best models, which have the lower error percentage, an explanatory analysis was presented to explain the results obtained.

A decision tree is a tree structure similar to a flowchart (Priyam et al., 2013). This algorithm contains automatic selection of features to the vertices from a set of features and construction of decision rules in the form understandable to an expert. It predicts a target variable based on data obtained through its characteristic variables, which in turn measure the likely correspondence distribution to which a certain class belongs. The self-organizing Kohonen map is a method to project a multidimensional space into the two-dimensional space and consists of two layers: the input and the output. Kohonen maps make it possible to cluster objects into the simultaneously forming clusters, as well as predict and identify features and patterns in the large datasets (Kingsford et al., 2008). A neural network is used to simulate the nonlinear system that allows finding the complex data dependencies. They help to construct a model predicting the target value. To increase the constructed model accuracy, the number of neurons is increased (Schreck et al., 2009).

Data mining is the process of extracting information from some data. Identification of patterns in the data is done by using different methods. The datasets used were processed using the following methods:

1) Cluster Analysis. The method consists of dividing the data into groups so that objects within one group are similar to each other in some way;
2) Associative rules. This transformation is used to find the relationship between variables.

Datasets were taken for analysis from a website containing the open sources: Student Grade Prediction, Factors about Students Performance Affecting and Higher Education Students Performance Evaluation.

Information from the Student Grade Prediction dataset was obtained from a survey conducted with students in mathematics and Portuguese language in a high school. The database contained a lot of social, gender and academic information about the students. The final grade was based on the test results of students in subjects such as mathematics and Portuguese. In addition, data quality and impact of various indicators on the academic performance were considered. It was assumed that different factors would be significant in different schools. In order to analyze more accurately through statistical analysis, a preponderance of data from Gabriel Pereira School and a small amount of data from Musinho da Silveira School were observed, which led to these data being excluded from further analysis.

The database consisted of 30 indicators for 396 students and 3 assessments Information about the indicators and their description is presented in **Table 1**. To simplify understanding, a graph in percentage correlation has been displayed. The bars of the graph are colored depending on the modulus of the correlation percentage (0—blue; 100—red), and the frame of the bar depending on the sign of the correlation value. If the correlation is negative, it means that the parameter has an inverse dependence on the output parameter (when one parameter increases, the other decreases). If the correlation is positive—the parameters have direct dependence.

Table 1. Information dataset "Student Grade Prediction".

Indicator	Coding	Description
school—student's school	(GP; MS)	GP—Gabriel Pereira; MS—Musinho da Silveira
sex—student's gender	{0; 1}	0—Female; 1—Male
Age	{15-22}	
address—student's home address type	{0; 1}	0—urban; 1—rural
famsize—family size	{0; 1}	0—three or less than three family members; 1—is greater than 3 family members
pstatus—parental cohabitation status	{0; 1}	0—together; 1—separately
medu—mother's education	{0; 1; 2; 3; 4}	0—lack of education; 4—higher education
fedu—father's education	{0; 1; 2; 3; 4}	0—lack of education; 4—higher education
mjob—mother's job	{0; 1; 2; 3; 4}	0—teacher; 1—healthcare; 2—civil services (for example, administrative or police); 3—housewife; 4—other
fjob—father's job	{0; 1; 2; 3; 4}	0—teacher; 1—healthcare; 2—civil services (for example, administrative or police); 3—housewife; 4—other
reason—reason for choosing this school	{0; 1; 2; 3}	0—proximity to home; 1—school reputation; 2—course preferences; 3—other
guardian—student's guardian	{0; 1; 2}	0—mother; 1—father; 2—other
traveltime—travel time from home to school	{1; 2; 3; 4}	1—less than 15 minutes; 2—from 15 to 30 minutes; 3—from 30 minutes up to 1 hour; 4—more than 1 hour
studytime—weekly study time	{1; 2; 3; 4}	1—less than 2 hours; 2—from 2 to 5 hours; 3—from 5 to 10 201 hours; 4—more than 10 hours)
failures—number of past failures in the class	{0; 1; 2; 3}	
schoolsup—additional educational support	{0; 1}	0—no; 1—yes
famsup—family educational support	{0; 1}	0—no; 1—yes
paid—additional paid classes in mathematics or Portuguese	{0; 1}	0—no; 1—yes
activities—extracurricular activities	{0; 1}	0—no; 1—yes
nursery—kindergarten 206 attendance	{0; 1}	0—no; 1—yes

Table 1. (*Continued*).

Indicator	Coding	Description
higher—desire to get higher education	{0; 1}	0—no; 1—yes
internet—Internet access at home	{0; 1}	0—no; 1—yes
romantic—romantic relationship present	{0; 1}	0—no; 1—yes
famrel—quality of family relationships	{1; 2; 3; 4; 5}	1—very bad; 5—excellent
freetime—availability of free time after school	{1; 2; 3; 4; 5}	1—very bad; 5—excellent
goout—strolling with friends	{1; 2; 3; 4; 5}	1—very bad; 5—excellent
dalc—alcohol consumption during the working day	{1; 2; 3; 4; 5}	1—very bad; 5—excellent
walc—alcohol consumption on the weekends	{1; 2; 3; 4; 5}	1—very bad; 5—excellent
health—current state of health {	{1; 2; 3; 4; 5}	1—very bad; 5—very good
absences—number of classes missed at school	{0; 93}	
G1—grade for the first lesson, mathematics	{0; 20}	
G2—grade for the second lesson, Portuguese	{0; 20}	
G3—final score	{0; 20}	

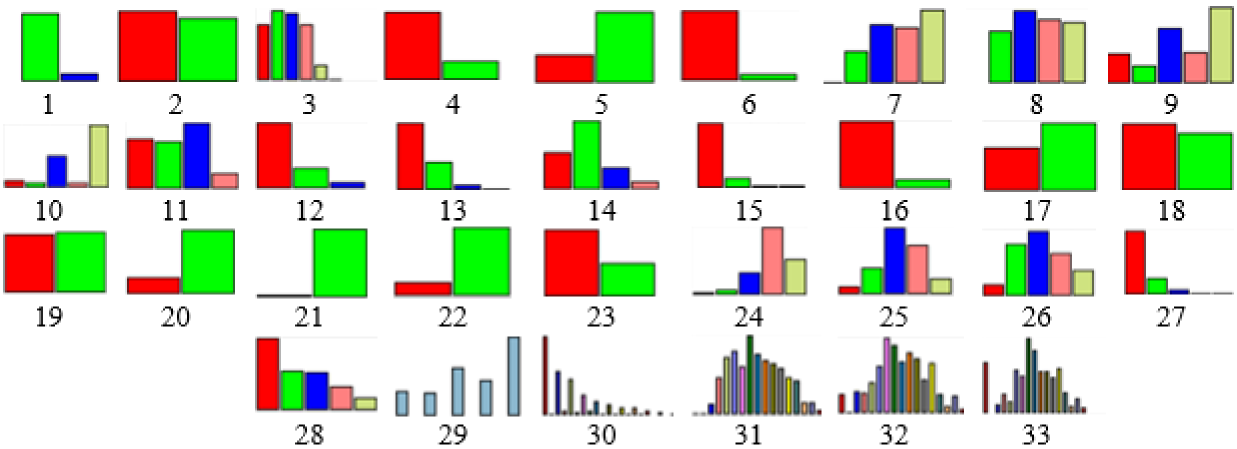


Figure 1. Statistic data display from sample: (1)—student’s school; (2)—student’s gender; (3)—student’s age; (4)—student’s home address type; (5)—family size; (6)—parents’ cohabitation status; (7)—mother’s education; (8)—father’s education; (9)—mother’s occupation; (10)—father’s occupation; (11)—this school selection cause; (12)—student’s guardian; (13)—time from home to school; (14)—weekly learning time; (15)—number of past failures in class; (16)—additional education support; (17)—family education support; (18)—additional paid classes in mathematics and Portuguese; (19)—extracurricular activities; (20)—kindergarten attendance; (21)—desire to receive higher education; (22)—access to Internet at home; (23)—romantic relations; (24)—family relations quality; (25)—free time available after school; (26)—scrolling with friends; (27)—alcohol consumption during a working day; (28)—alcohol consumption in the weekends; (29)—current state of health; (30)—number of missed classes at school; (31)—points for the first lesson, mathematics; (32)—points for the second lesson, Portuguese; (33)—total score.

Based on the initial data, the statistical data display was generated in **Figure 1**. The statistics allow us to display the distribution of the data available in the set, for example, the attribute representing the school of the learner has 2 values Gabriel Pereira and Musinho da Silveira. **Figure 1** shows a significant predominance of students from the Gabriel Pereira school. The gender of the trainees has a slight predominance towards the male gender.

The Factors about students performance affecting dataset was formed from a survey of the US college students. It consists of 100 records and 18 attributes. Information about the indicators and their description is presented in **Table 2**.

Table 2. Information dataset "Factors about students performance affecting".

Indicator	Coding	Description
Scores for 3 marks for 2 assessments in subjects (reading, mathematics, science)		
Race	{1; 2; 3}	1—European; 2—Afro-American; 3—Latino-American
Male—student's gender	{0; 1}	0—Female; 1—Male
ParEd—parents' education		
HoursKids—time spent with parents (in hours)		
GradestImport—importance of scoring for the parents	{0; 1}	0—marks are not important; 1—marks are important
Attend—attending extracurricular activities		
HourStudy—time spent studying at home		
Sports	{0; 1}	1—the student goes in for sports
Success criterion—consisting of the sum of marks for the second assessment in the reading, mathematics and science		

Based on the initial data, the statistical data display was generated in **Figure 2**. In **Figure 2**, from Charts 15–17 and the resulting Chart 7, it can be seen that of the three nationalities of students, Afro-American is less than European and Latino-American, and it can also be seen that the majority of US college students participate in sports.

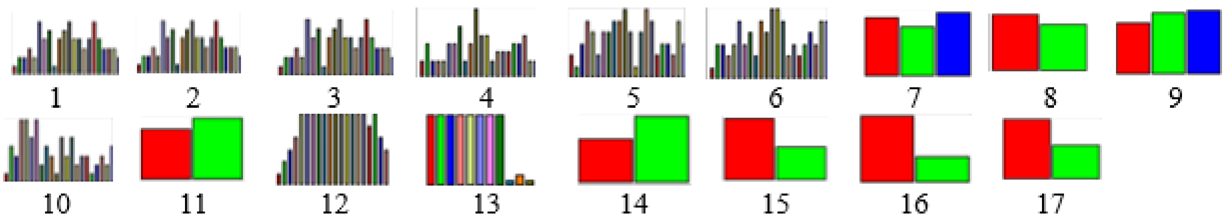


Figure 2. Statistic data display from sample: (1)—reading2; (2)—mathematics2; (3)—science2; (4)—reading1; (5)—mathematics1; (6)—science1; (7)—race; (8)—gender; (9)—parents' education; (10)—time with parents; (11)—scoring importance; (12)—extracurricular activities; (13)—learning time at home; (14)—sports; (15)—European; (16)—Afro-American; (17)—Latino-American.

The Higher education students performance evaluation dataset consists of data obtained from students of the Engineering Faculty and the Pedagogical Sciences

Faculty. It includes 145 rows and 32 factors. Information about the indicators and their description is presented in **Table 3**.

Table 3. Information dataset "Higher education students performance evaluation".

Indicator	Coding	Description
Student Age	{1; 2; 3}	0—from 18 to 21; 1—from 22 up to 25; 2—from 26 and above
Sex gender	{1; 2}	1—female
Additional work	{1; 2}	1—yes; 2—no
Regular artistic or sports activity	{1; 2}	1—yes; 2—no
Graduated high-school type	{1;2;3}	1—private; 2—public; 3—other
scholarship type	{1;2;3;4;5}	1—none; 2—25%; 3—50%; 4—75%; 5—full
Do you have a partner	{1; 2}	1—yes; 2—no
Total salary if available	{1; 2; 3; 4; 5}	1—$135-200; 2—$201-270; 3—$271-340; 4—$341-410; 5—$400 and above
Transportation to the university	{1; 2; 3; 4}	1—bus; 2—private car or taxi; 3—bicycle; 4—other
Mother's education	{1; 2; 3; 4; 5; 6}	1—primary school; 2—secondary school; 3—secondary education; 4—university; 5—master's degree; 6—Ph.D.
Father's education	{1; 2; 3; 4; 5; 6}	1—primary school; 2—secondary school; 3—secondary education; 4—university; 5—master's degree; 6—Ph.D.
Number of sisters/brothers	{1; 2; 3; 4; 5}	1—1 brother or sister; 5—5 and above
Parental status	{1; 2; 3}	1—married; 2—divorced; 3—one of them or both died
Mother's occupation	{1; 2; 3; 4; 5; 6}	1—pensioner; 2—housewife; 3—civil servant; 4—private employee; 5—self-employed; 6—other

Table 3. (*Continued*).

Indicator	Coding	Description
Father's occupation	{1; 2; 3; 4; 5}	1—pensioner; 2—civil servant; 3—private employee; 4—self-employed; 5—other
Weekly study hours	{1; 2; 3; 4; 5}	1—none; 2—less than 5 hours; 3—6- 10 hours; 4—11-20 hours; 5—more than 20 hours
Reading frequency (non-scientific books/journals)	{1; 2; 3}	1—none; 2—sometimes; 3—often
Reading frequency (scientific books/journals)	{1; 2; 3}	1—none; 2—sometimes; 3—often
Impact of your projects/activities (attending seminars and conferences)	{1; 2}	1—yes; 2—no
Impact of your projects/activities (projects/activities influence on your success)	{1; 2; 3}	1—positive; 2—negative, 3—neutral
Attendance to classes	{1; 2; 3}	1—always; 2—sometimes; 3—never
Preparation to midterm exams 1 (with whom)	{1; 2; 3}	1—alone; 2—friends; 3—not applicable
Preparation to midterm exams 2 (time spent)	{1; 2; 3}	1—the closest date to exams; 2—regularly throughout the semester; 3—never
Taking notes in classes	{1; 2; 3}	1—never; 2—sometimes; 3—always
Cumulative grade point average in the last semester r	{1; 2; 3; 4; 5}	1—less than 2.00; 2—from 2.00 to 2.49; 3—from 2.5 262 to 2.99; 4—from 3.00 to 3.49; 5—from 3.5 and above
Expected Cumulative grade point average in the graduation	{1; 2; 3; 4; 5}	1—less than 2.00; 2—from 2.00 to 2.49; 3—264 from 2.5 to 2.99; 4—from 3.00 to 3.49; 5—from 3.5 and above
course		
OUTPUT Grade		

Based on the initial data, the statistical data display was generated in **Figure 3**. The age of students allows us to notice that there are fewer students older than 26 years old than students from 18 to 25 years old. There are more married families among students than divorced families or families with deceased members.

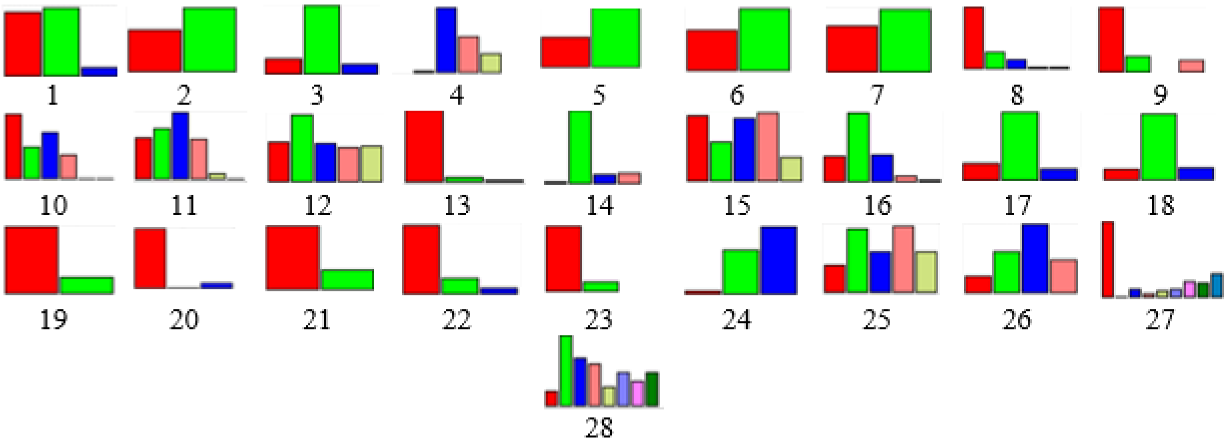


Figure 3. Statistic data display from sample: (1)—student's age; (2)—gender; (3)—regular artistic and sports activities; (4)—secondary school graduation type; (5)—scholarship type; (6)—job; (7)—partner; (8)—salary amount; (9)—transport to university; (10)—mother's education; (11)—father's education; (12)—number of brothers and sisters; (13)—parents' status; (14)—mother's occupation; (15)—father's occupation; (16)—weekly learning hours; (17)—reading periodicity (nonscientific books/magazines); (18)—reading periodicity (scientific books/journals); (19)—seminars/conferences attendance; (20)—projects/events influence on your success; (21)—classes attendance; preparation to the interim exam (with whom); (22)—classes attendance; (23)—preparation to the interim exam (time spent); (24)—taking notes in class; (25)—average score in the last semester; (26)—expected score in the graduate class; (27)—course; (28)—total scoring.

Before starting to work with the datasets, they require preprocessing to achieve data normalization. The next step involved setting the parameters separately for each method. After completing the steps to set up the data processing, learning and visualization of the results followed.

4. Results

Correlation analysis was used to determine the output parameter dependence on the input parameter for each dataset. The correlation parameter values were classified into weak (less than 0.29), moderate (0.3–0.49), medium (0.5–0.69) and strong (0.7 or more) (Solutions, 2019). The correlation analysis demonstrates that scoring data in the first and second training periods were strongly depending on the output parameter. After this analysis, it was decided to remove attributes that had a weak dependence (less than 0.05).

4.1. Student grade prediction dataset analysis

The data correlation analysis revealed, to what extent the output parameter depended on the input factors (**Table 3**). **Table 3** shows that attributes G1 and G2 have high dependence on the output parameter, as they directly depend on the output parameter. The other parameters have a dependence less than the average, which indicates a weak dependence on the influence of other attributes.

Models would be built using datasets with the different output parameters (number of classes). The assessment system is changing to the five-point system from 18 to 20 points (from 100% to 90% of the correctly completed work)—5, from 14 to 17 points (from 89% to 70%)—4, from 10 to 13 points (from 69% to 50%)—3, below—2, and to the two-point system from 20 to 11 (50% of work completed correctly)—2, from 10 to 0—1. Besides, marks by period and the number of missed classes were not taken into account in building the models, since they were obviously depending on the success output parameter; the same was with the attributes having weak dependence of less than 0.05%.

Let us start with description of the Decision Tree method. This method was applied taking into account results of the correlation analysis, see **Table 4**. To describe the results, the display methods Significance of Results in **Table 5** and Decision Tree in **Figures 4–6** were chosen.

Table 4. Correlation analysis.

№	Attribute name	Correlation, %	Percentage graph
1	School	−0.042	
2	Gender	0.103	
3	Age	−0.162	
4	Home address type	−0.106	
5	Family size	−0.081	
6	Parents’ habitation	0.058	
7	Mother’s education	0.217	
8	Father’s education	0.152	
9	Mother’s occupation	−0.146	
10	Father’s occupation	−0.091	
11	School selection cause	−0.009	
12	Guardian	−0.054	
13	Time from home to school	−0.117	
14	Weekly learning time	0.098	
15	Number of past failures in class	−0.360	
16	Additional education support	−0.083	
17	Family education support	−0.039	
18	Add. paid classes in mathematics	0.102	
19	Extracurricular activities	0.016	
20	Kindergarten attendance	0.052	
21	Desire to receive higher education	0.182	
22	Access to Internet at home	0.098	
23	Romantic relations	−0.130	
24	Family relations quality	0.051	
25	Free time after studies	0.011	
26	Strolling with friends	−0.133	
27	Alcohol consumption during working week	−0.055	

Table 4. (*Continued*).

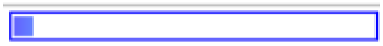
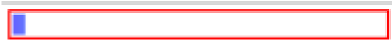
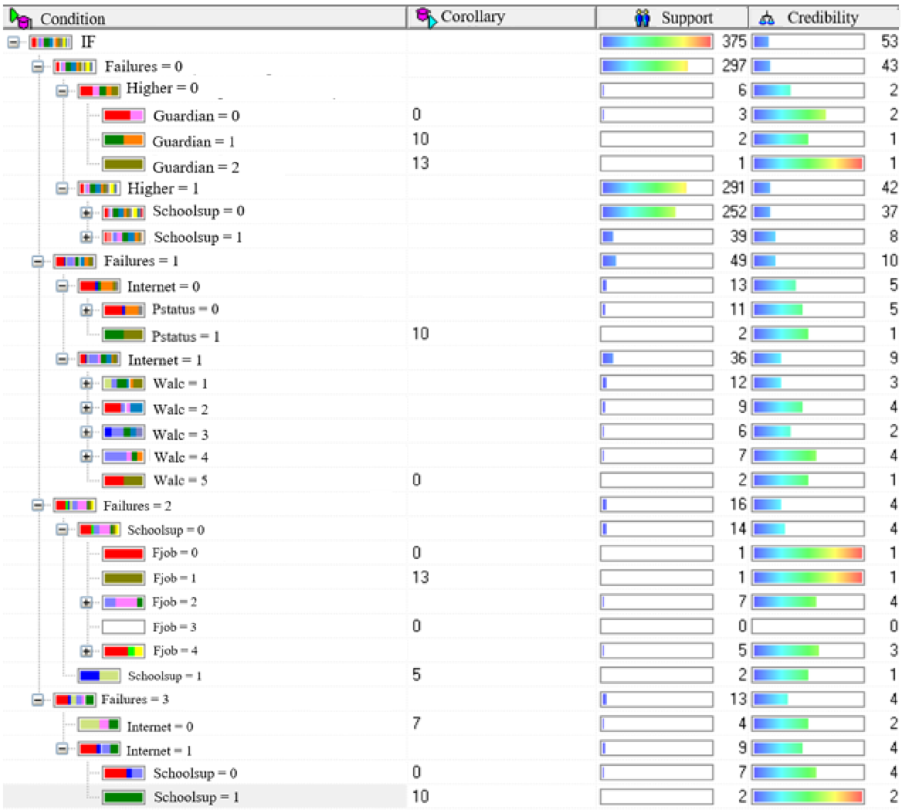
№	Attribute name	Correlation, %	Percentage graph
28	Alcohol consumption in the weekends	−0.052	
29	Current health status	−0.061	
30	Number of missed classes	0.034	
31	Scoring, first period	0.801	
32	Scoring, second period	0.905	

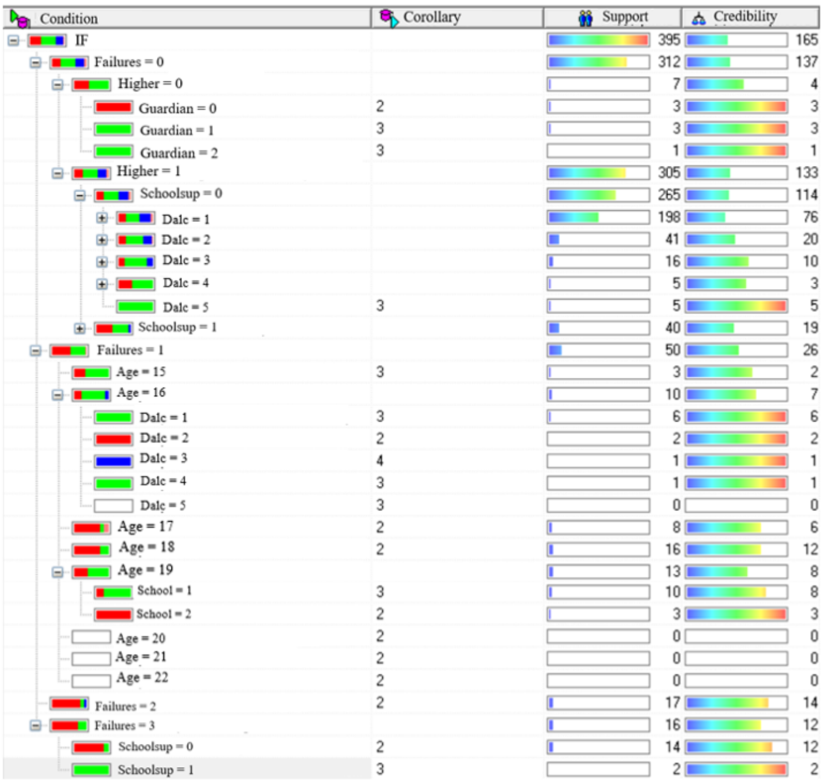
Table 5. Attribute significance.

Attribute	Significance, %		
	0–20	1–5	1–2
Mother's education	10.245	3.510	0
Number of past failures in class	7.582	12.158	42.073
Mother's occupation	7.027	0	8.728
Weekly learning time	6.575	3.441	0
Additional paid classes in mathematics	5.794	4.540	7.838
Alcohol consumption in the weekends	5.510	4.422	0
Father's occupation	4.553	5.944	0
Access to Internet at home	4.471	2.597	0
Gender	4.280	3.818	0
Alcohol consumption during the working week	4.240	7.012	0
Home address type	4.195	1.745	0
Parents' habitation	4.109	1.527	0
Additional education support	3.966	4.965	0
Current health status	3.913	3.094	0
Kindergarten attendance	3.759	3.133	3.267
Family size	3.118	2.422	0
Family relations quality	3.086	3.967	0
School	2.893	2.369	9.277
Guardian	2.493	4.565	6.239
Age	1.711	3.569	0
Father's education	1.658	5.491	0
Strolling with friends	1.164	3.883	8.992
Romantic relations	1.106	2.285	0
Family education support	0.964	4.478	0
Desire to receive higher education	0.887	1.027	13.586
Time from home to school	0.698	4.036	0



Condition	Corollary	Support	Credibility
IF		375	53
Failures = 0		297	43
Higher = 0		6	2
Guardian = 0	0	3	2
Guardian = 1	10	2	1
Guardian = 2	13	1	1
Higher = 1		291	42
Schoolsup = 0		252	37
Schoolsup = 1		39	8
Failures = 1		49	10
Internet = 0		13	5
Pstatus = 0		11	5
Pstatus = 1	10	2	1
Internet = 1		36	9
Walc = 1		12	3
Walc = 2		9	4
Walc = 3		6	2
Walc = 4		7	4
Walc = 5	0	2	1
Failures = 2		16	4
Schoolsup = 0		14	4
Fjob = 0	0	1	1
Fjob = 1	13	1	1
Fjob = 2		7	4
Fjob = 3	0	0	0
Fjob = 4		5	3
Schoolsup = 1	5	2	1
Failures = 3		13	4
Internet = 0	7	4	2
Internet = 1		9	4
Schoolsup = 0	0	7	4
Schoolsup = 1	10	2	2

Figure 4. Decision Tree on the 20-points scale.



Condition	Corollary	Support	Credibility
IF		395	165
Failures = 0		312	137
Higher = 0		7	4
Guardian = 0	2	3	3
Guardian = 1	3	3	3
Guardian = 2	3	1	1
Higher = 1		305	133
Schoolsup = 0		265	114
Dalc = 1		198	76
Dalc = 2		41	20
Dalc = 3		16	10
Dalc = 4		5	3
Dalc = 5	3	5	5
Schoolsup = 1		40	19
Failures = 1		50	26
Age = 15	3	3	2
Age = 16		10	7
Dalc = 1	3	6	6
Dalc = 2	2	2	2
Dalc = 3	4	1	1
Dalc = 4	3	1	1
Dalc = 5	3	0	0
Age = 17	2	8	6
Age = 18	2	16	12
Age = 19		13	8
School = 1	3	10	8
School = 2	2	3	3
Age = 20	2	0	0
Age = 21	2	0	0
Age = 22	2	0	0
Failures = 2	2	17	14
Failures = 3		16	12
Schoolsup = 0	2	14	12
Schoolsup = 1	3	2	2

Figure 5. Decision Tree on the 5-points scale.

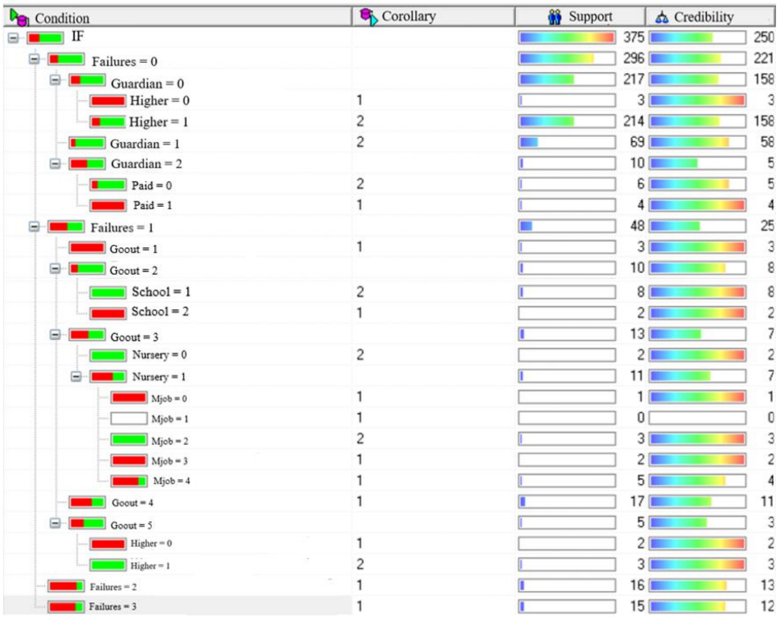


Figure 6. Decision Tree on the 2–points scale.

The Attribute Significance visualizer makes it possible to determine the output variables significance. Based on the table, a conclusion could be made on the following attributes significance. Mother's education was significant from the output parameter by 10.245% with the possible error of 39.75% in the twenty-point system of calculating the output parameter. The number of past failures was dependent by 42.073% with the probability of 21.27% in assessment using the two-point calculation system and significant by 12.158% with the error of 23.04% in the five-point calculation system. Desire to obtain higher education was significant by 13.586% with the error of 13.586%. Next, the resulting models could be applied to test sets taking into account the probable error percentage.

The next method in data processing is the Kohonen maps, also constructed taking into account the correlation analysis, see **Table 4**. Examining the resulting maps, uneven division into clusters was noticed, which made it difficult to describe the results obtained. Therefore, the method is not considered for this dataset.

Application of the neural networks to each output parameter and taking into account the correlation analysis result in **Table 4** provided with a different error percentage. However, only for the output parameter dividing the students into 2 classes, the error percentage is less than 50%. Analysis of such a neural network is presented in **Figures 7** and **8**.

The neural network made it possible to build a model capable of predicting the student's success with an error of 32.91% in the two-point output parameter calculation system. Using a test set in this model increased the error rate to 40%. By changing the neural network setting parameters, the task of reducing the error percentage was not fulfilled.

Based on the constructed models, the errors were registered in the Summary **Table 6**.

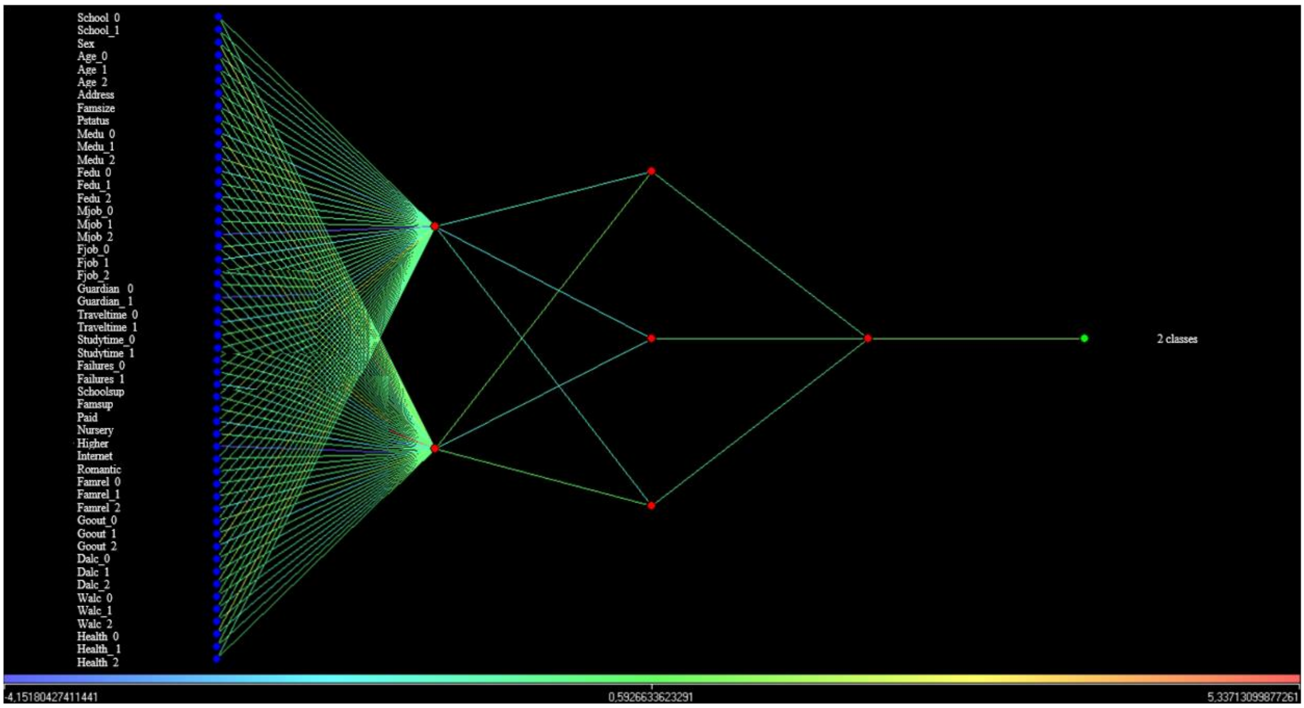


Figure 7. Neural network graph for the output parameter with two classes.

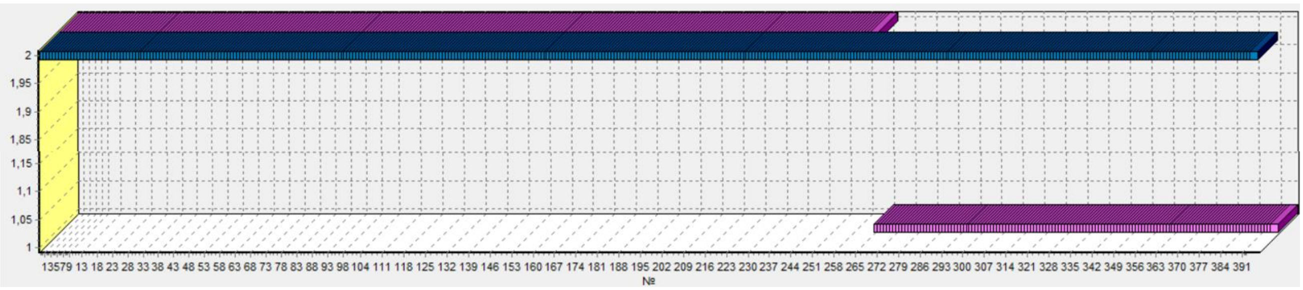


Figure 8. Neural network diagram for the output parameter with two classes.

Table 6. Error information.

	Kohonen map			Decision Tree			Neural networks		
Calculus system	**0–20**	**1–5**	**1–2**	**0–20**	**1–5**	**1–2**	**0–20**	**1–5**	**1–2**
Error percentage	52.91%	35.95%	21.27%	39.75%	23.04%	21.27%	99.49%	70.38%	32.91%
Incorrect classification	209	142	84	157	91	84	393	278	130
Correct classification	186	253	311	238	304	311	2	117	265
Rows, total					395				

Based on the table, it could be concluded that the Decision Tree method made it possible to forecast better than with the other methods. The most significant factor was the number of past failures in the class. The fewer failures a student had, the more successful he was. Mother's education and desire to obtain higher education were having less pronounced relationship dependence on the success criteria.

4.2. Factors about student performances affecting the dataset analysis

Let us conduct the data correlation analysis to identify significance of the factors in **Table 7**.

Table 7. Data correlation analysis.

№	Attribute name	Correlation, %	Percentage graph
1	Reading 2	0.998	
2	Mathematics 2	0.998	
3	Science 2	0.998	
4	Reading 1	0.992	
5	Mathematics 1	0.990	
6	Science 1	0.994	
7	Race	−0.427	
8	Gender	0.044	
9	Parents' education	0.844	
10	Time with parents	0.964	
11	Scoring importance for the parents	0.065	
12	Extracurricular activities	0.964	
13	Learning time at home	0.294	
14	Sports	−0.632	
15	European	0.402	
16	Afro-American	0.036	
17	Latino-American	−0.389	

Using correlation, factors were identified that depend directly on the success criterion, as well as factors depending inversely on the success criterion. Factors depending directly included subjects in two assessments (reading, mathematics and science); parents' education; time with parents; and extracurricular activities. The inversely dependent factors were race and sports. Gender and scoring grade importance were not the significant factors for parents dependent on the success criterion. Taking into account the correlation analysis carried out, Kohonen maps, neural network and the Decision Tree were formatted depicting relationship between the input and output factors without accounting for grades in subjects. Besides, the grading system was changed to a ten-point system: from 20 to 19—10 points, 17–18—9 points, 15–16—8 points, 13–14—7 points, 11–12—6 points, 9–10—5 points, 7–8—4 points, 5–6—3 points, 3–4—2 points, 1–2—1 point; a five-point system from 200 to 180 points—5, from 179 to 140 points—4, from 139 to 100 points—3. Then, 22 and the two-point system from 200 to 100—2, from 99 to 0—1 followed.

The Decision Tree method is applied to the generated correlation analysis. To describe results, the Significance of Results display method, see **Table 8**, and the Decision Tree method, **Figures 9–11**, were selected.

Table 8. Attribute significance.

Attribute	Significance, %		
	0–10	**1–5**	**1–2**
Race	0%	0%	0%
Gender	0%	0%	0%
Parents' education	56.734%	56.919%	100%
Time with parents	27.471%	0%	0%
Scoring importance for the parents	1.291%	0%	0%
Extracurricular activities	0%	0%	0%
Learning time at home	0%	0%	0%
Sports	14.504%	43.081	0%

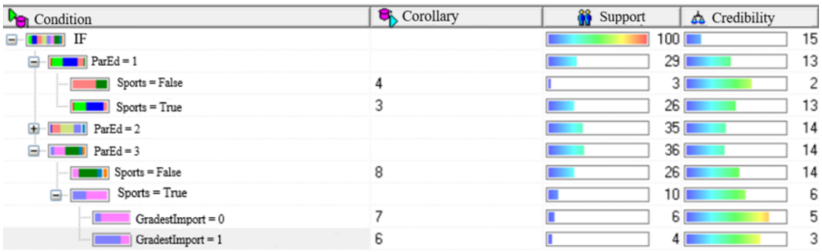


Figure 9. Decision Tree on the 10-points scale.

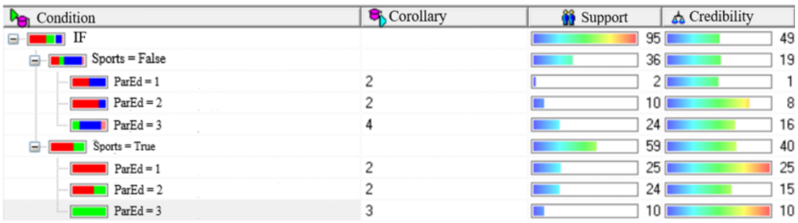


Figure 10. Decision Tree on the 5-points scale.

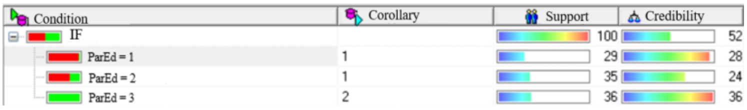


Figure 11. Decision Tree on the 2-points scale.

The Attribute Significance visualizer makes it possible to determine significance of the output variables.

Based on **Table 8**, a conclusion could be made on significance of the following attributes: parental education for ten-, five- and two-points assessment systems with significance of 56.734%, 56.919% and 100%, respectively; sports with significance of 14.504% for the two-points and 43.081% for the five-points output parameters, as well as the time with parents with significance of 27.471% for the ten-points output attribute.

The results obtained taking into account the correlation analysis and **Table 8** using the self-organizing Kohonen maps make it possible to monitor relationship of the clusters in **Figures 12–14**.

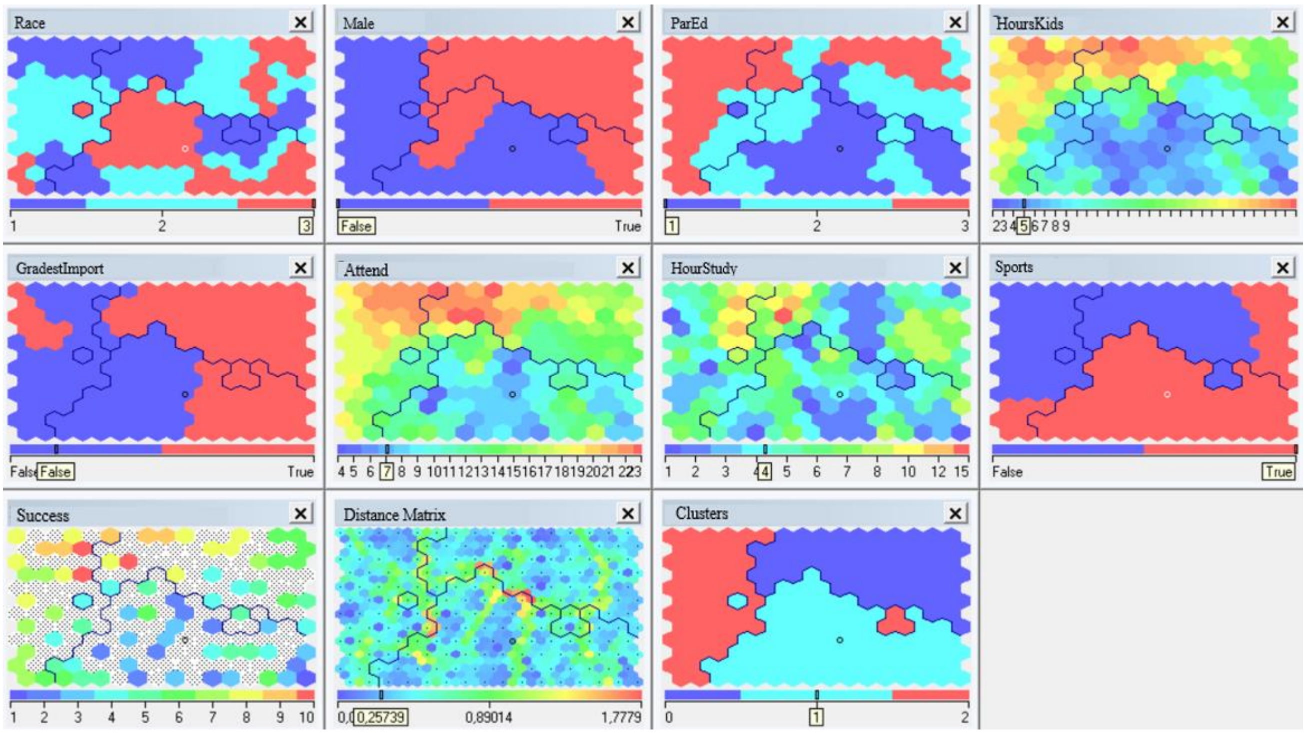


Figure 12. Kohonen maps on the 10-points calculus system.

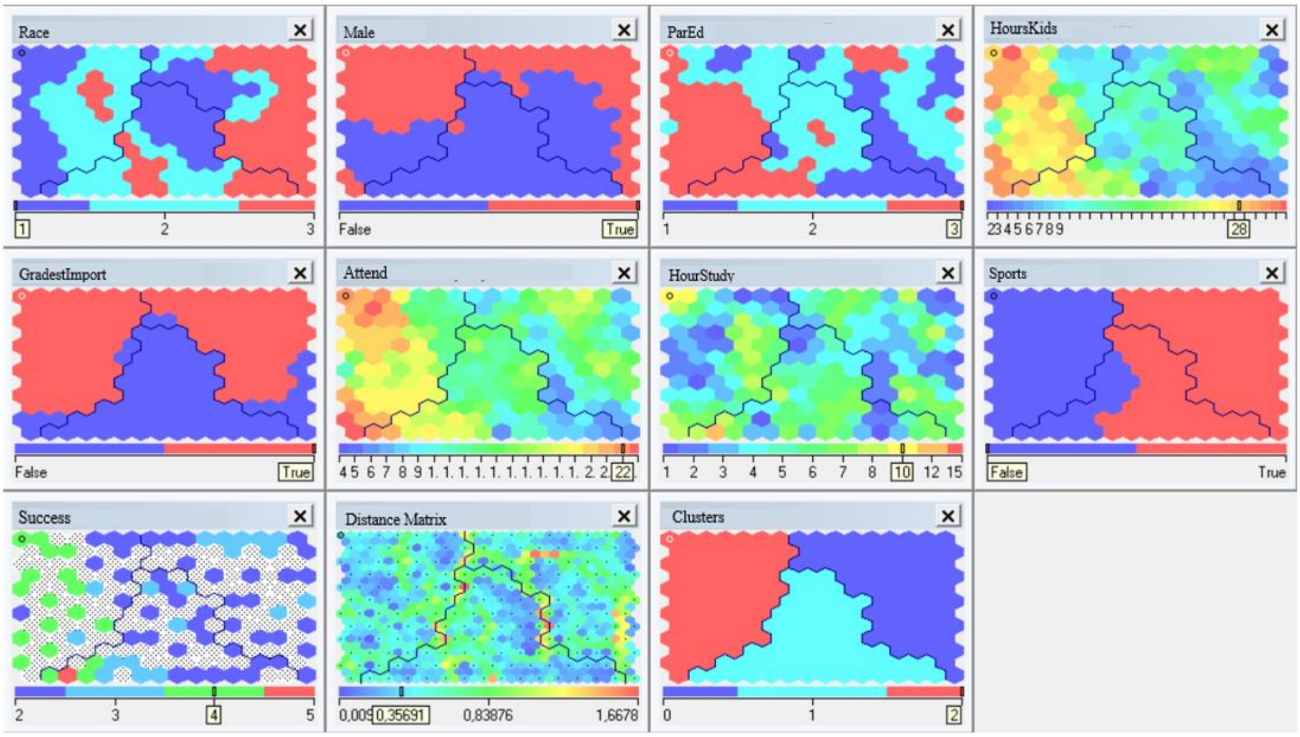


Figure 13. Kohonen maps on the 5-points calculus system.

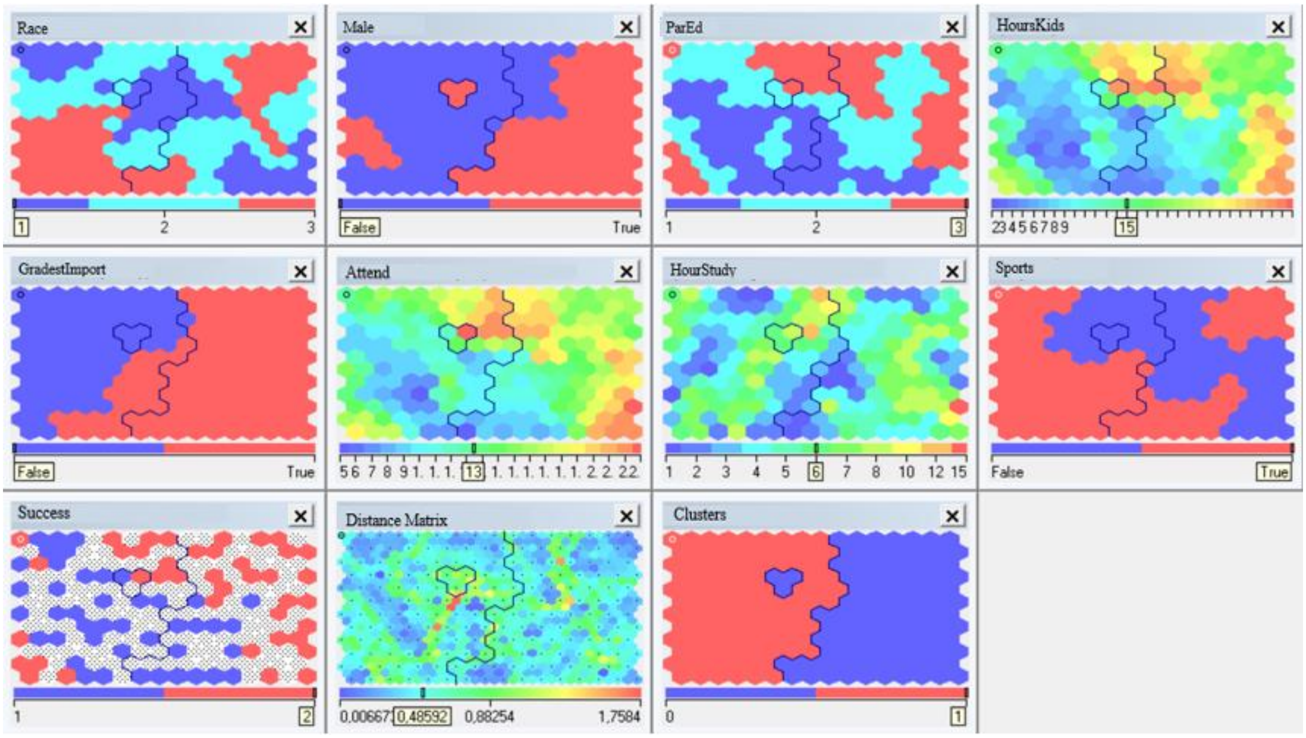


Figure 14. Kohonen maps on the 2-points calculus system.

As a result, clusters were formed depending on the number of output parameter classes that determined the student's success or failure. Division was based on the success criterion value. The race field value was insignificant (correlation analysis showed its insignificance), but the Kohonen maps could show that there appeared a dependence on the success criterion, and it consisted in the fact that a person of a certain race was successful to a certain extent. Detailed cluster description is presented in **Table 9**.

Table 9. Attribute description for each type of students.

Output parameter values	**0–10**			**1–5**			**1–2**	
Number of classes	**1**	**0**	**2**	**1**	**0**	**2**	**0**	**1**
Number of students	45	30	25	37	35	28	53	47
Sports	100%	97.7%	100%	65.3%	100%	100%	99.0%	99.4%
Parents' education	100%	99.1%	100%	5.4%	98.3%	99.1%	58.1%	62.5%
Extracurricular activities	94.1%	45.5%	96.5%	21.1%	82.3%	99.3%	18.8%	29.2%
Time with parents	92.7%	47.7%	89%	2.3%	51.6%	97.4%	1.5%	3.6%
Race	93.7%	59.1%	91.1%	39.4%	99.9%	99.6%	28.0%	30..9%
Learning time at home	2.8%	10.9%	24.4%	39.8%	4.4%	23.0%	4.3%	6.4%

Almost all factors appear significant in dividing the students by the level of their success, and the least significant factor was the time spent studying at home.

As a result of analyzing the received cards, portraits of students were compiled based on the success criterion value, see **Table 10**.

Table 10. Student's portrait.

	Unsuccessful	**Average successful**	**Most successful**
Performance	Unsatisfactory	Satisfactory	Good and excellent
Race	Latino-American	Afro-American	European
Parents' education	1	2	3
Time with parents	8	16	28
Extracurricular activities	Not attending or attending most rarely	Attends, but rarely	Attends

Neural network for the output parameter with 10 classes will not be described, because the error was 87%. Description of the neural network for the output parameter with 5 classes is presented in **Figures 15** and **16**.

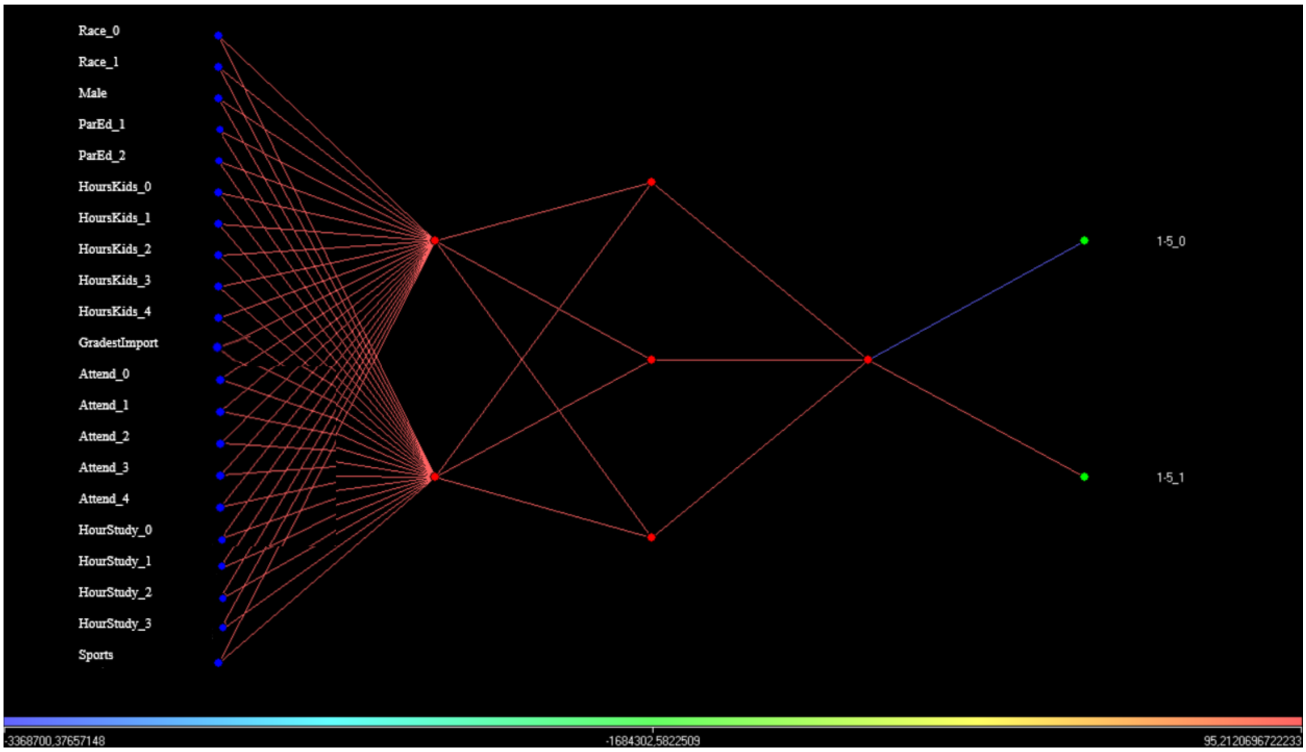


Figure 15. Neural network graph for the output parameter with 5 classes.

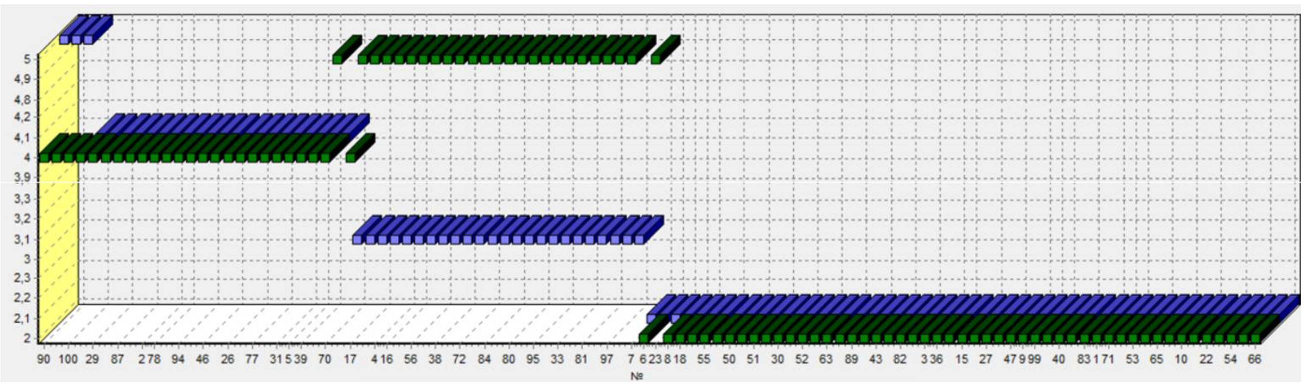


Figure 16. Neural network diagram for the output parameter with 5 classes.

The neural network was used to build a model making it possible to predict the student's success with the 29% error. Using the test subset in this model increased the error rate to 60%.

The dataset model with the output parameter having 2 classes is presented in **Figures 17** and **18**.

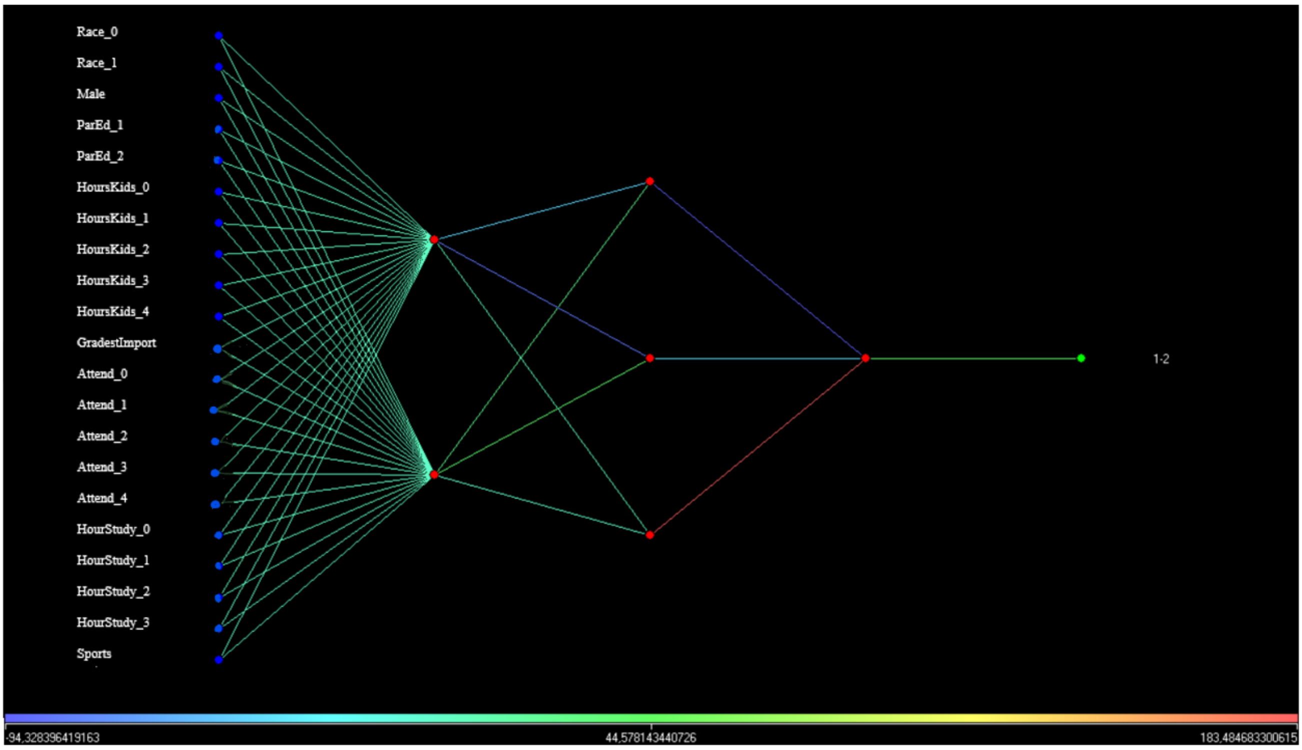


Figure 17. Neural network graph for the output parameter with 2 classes.

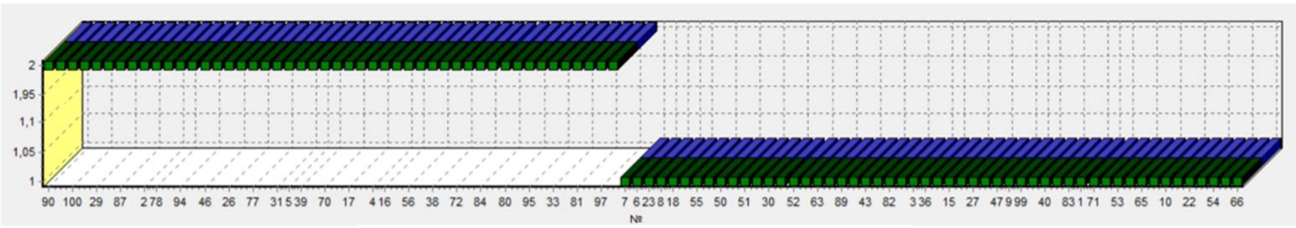


Figure 18. Neural network diagram for the output parameter with 2 classes.

The constructed model is able to accurately predict the student's success. Using this model in the test set was not resulting in any errors.

Errors for each applied method are presented in **Table 11**.

Table 11. Error information.

	Kohonen map			**Decision Tree**			**Neural networks**		
Calculus system	**0–10**	**1–5**	**1–2**	**0–10**	**1–5**	**1–2**	**0–10**	**1–5**	**1–2**
Error percentage	13%	4%	2%	35%	20%	12%	87%	29%	0%
Incorrect classification	9	4	2	35	20	12	87	29	0
Correct classification	91	96	98	65	80	86	13	71	100
Total rows	100								

Error analysis made it possible to identify methods that best described the dataset.

Each of the applied methods predicted or classified data with the acceptable error, except for the neural network for the 10 classes' output parameter (**Table 12**).

Table 12. Summary table, US college.

Attribute	Significance, %			
	0–10	**1–5**	**1–2**	**Correll.**
Race	0%	0%	0%	−0.427
Gender	0%	0%	0%	0.044
Parents' education	56.734%	56.919%	100%	0.844
Time with parents	27.471%	0%	0%	0.964
Scoring importance for parents	1.291%	0%	0%	0.065
Extracurricular activities	0%	0%	0%	0.964
Learning time at home	0%	0%	0%	0.294
Sports	14.504%	43.081	0%	−0.632

The most significant factor is education of parents; the higher it is, the greater is success of a student. Time spent with the parents and extracurricular activities are significant according to the correlation, but constructed decision trees demonstrated that the attributes were not significant.

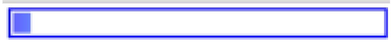
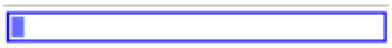
4.3. Higher education student performances evaluation dataset

The dataset was processed using the correlation analysis, see **Table 13**.

Table 13. Data correlation analysis.

No.	Attribute name	Correlation, %	Percentage graph
1	Age	−0.095	
2	Gender	0.336	
3	Secondary school graduation type	0.105	
4	Scholarship type	0.024	
5	Job	0.167	
6	Regular artistic and sports events	−0.063	
7	Partner	−0.052	
8	Salary amount	−0.166	
9	Transport to university	−0.156	
10	Mother's education	0.066	
11	Father's education	0.064	
12	Number of brothers and sisters	0.084	
13	Parents' status	0.066	
14	Mother's occupation	−0.031	

Table 13. (*Continued*).

No.	Attribute name	Correlation, %	Percentage graph
15	Father's occupation	−0.044	
16	Weekly learning hours	−0.033	
17	Reading periodicity (nonscientific books/magazines)	0.196	
18	Reading periodicity (scientific books/journals)	0.003	
19	Seminar/conference attendance	−0.185	
20	Projects/events influence	−0.203	
21	Classes attendance	−0.140	
22	Preparation to the interim exam (with whom)	0.015	
23	Preparation to the interim exam (time)	0.074	
24	Notes in class	0.045	
25	Average score in the last semester	0.315	
26	Expected score in the graduate class	0.249	
27	Course	0.142	

The Table shows that the factors have weak dependence on the output field. Let us change the assessment system to the five-points system: from 7 to 6 points—5, from 5 to 4 points—4, from 3 to 2 points—3, then 2 and the two-point system from 7 to 4—2, from 3 to 0—1. Besides, let us not take into account the GPA for the last semester in building the models.

The Decision Tree was constructed taking into account the correlation analysis, see **Table 12**. Its description is presented using the Attribute Significance visualizers in **Table 14** and in the form of a Decision Tree in **Figures 19–21**.

Table 14. Attribute significance.

Attribute	Significance, %		
	0–7	**1–5**	**1–2**
Course	44.853%	68.249%	95.786%
Mother's education	8.014%	6.714%	0%
Projects/events influence	6.243%	0%	0%
Gender	6.026%	0%	0%
Classes attendance	4.337%	0%	0%
Regular artistic and sports activities	4.260%	5.650%	0%
Age	3.919%	6.508%	4.214%
Reading periodicity (nonscientific books/magazines)	3.775%	2.296%	0%
Partner	3.598%	2.852%	0%
Daily learning hours	3.584%	4.533%	0%
Secondary school graduation type	3.325%	3.198%	0%
Transport to university	2.523%	0%	0%

Table 14. (*Continued*).

Attribute	Significance, %		
	0–7	**1–5**	**1–2**
Parents' status	2.057%	0%	0%
Job	1.980%	0%	0%
Father's education	0.753%	0%	0%
Salary amount	0.753%	0%	0%
Expected score in the graduate class	0%	0%	0%
Seminar/conference attendance	0%	05	0%
Number of brothers and sisters	0%	0%	0%
Preparation to the interim exam (time)	0%	0%	0%

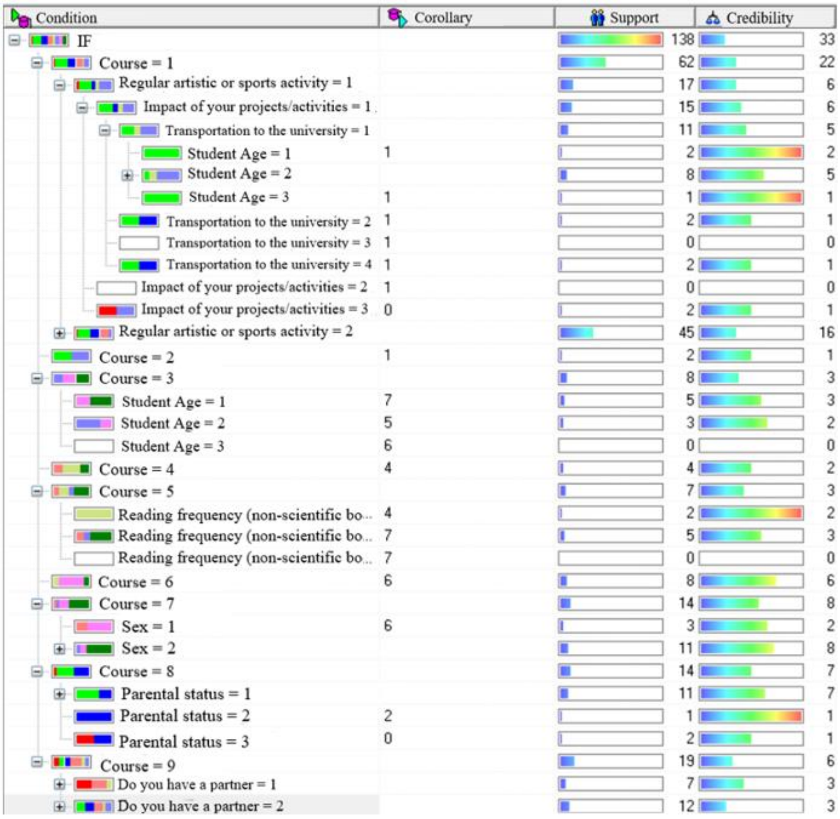


Figure 19. Decision Tree on the 7-points calculus system.

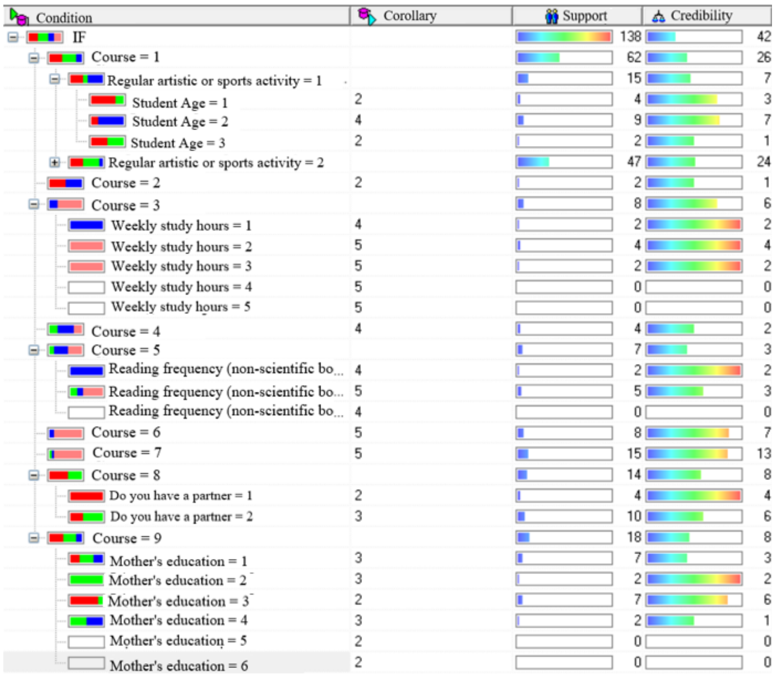


Figure 20. Decision Tree on the 5-points calculus system.

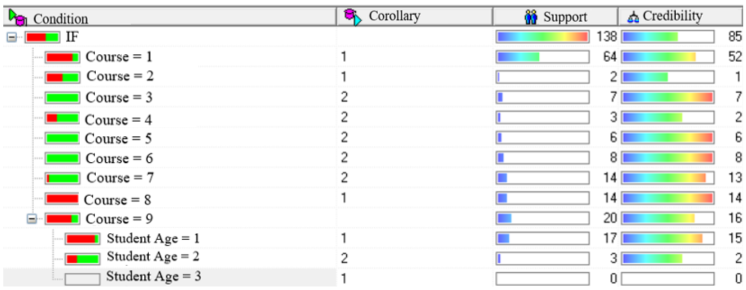


Figure 21. Decision Tree on the 2-points calculus system.

Table 14 shows that the "course" attribute is of great importance under the success criterion. Mother's education, influence of projects, events, gender, regular artistic or sports events and age have weak dependence. The remaining attributes do not have any dependence on the output parameter.

Detailed examination of the Kohonen maps made it possible to identify uneven division into clusters. With alteration in the number of clusters, division according to the success value was not changing (successful students fall into different clusters, which should not be the case).

Neural networks were built taking into account the correlation analysis carried out, see **Table 13**, for the output parameter with 7 and 5 classes and had a higher error percentage, and their consideration would not lead to reliable results. For the 2 classes, the neural network is presented in **Figures 22** and **23**.

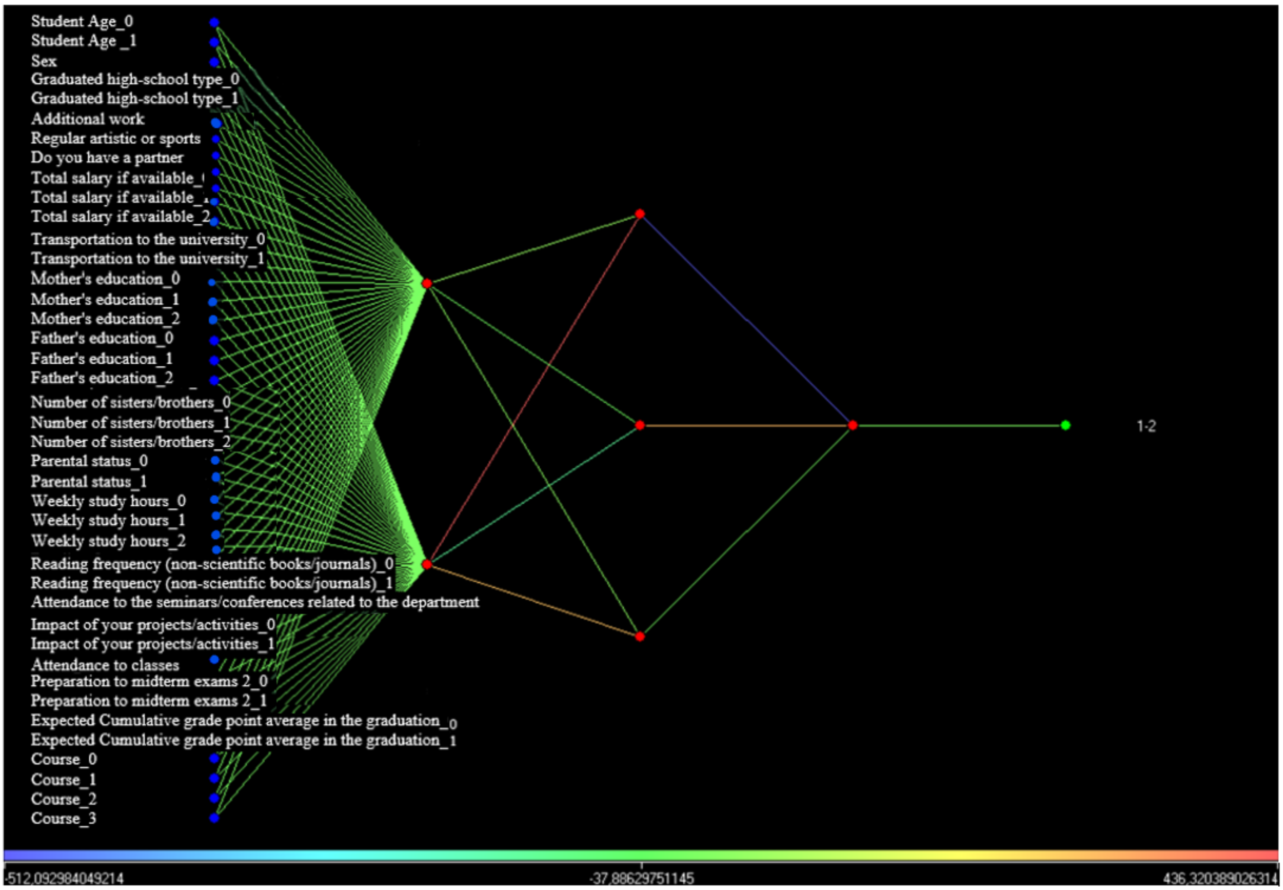


Figure 22. Neural network graph for the output parameter with two classes.

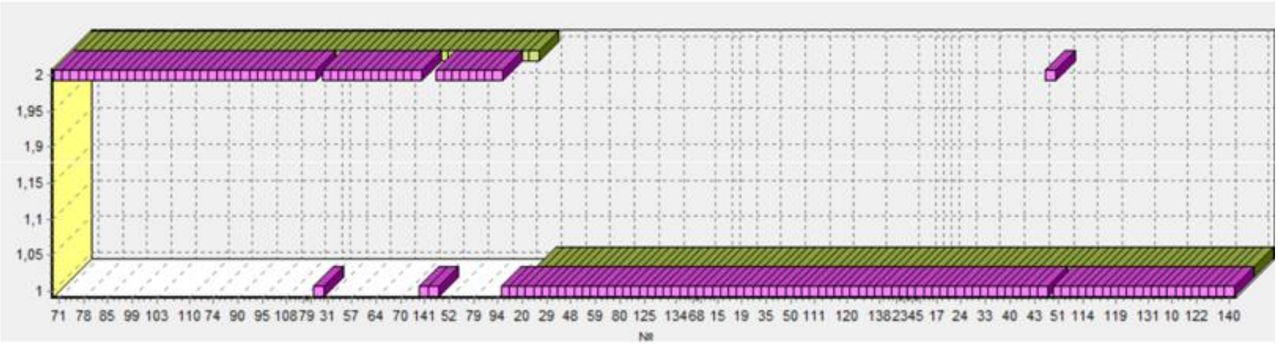


Figure 23. Neural network diagram for the output parameter with two classes.

Using a neural network, the model was built predicting student's success with the error of 4.14%.

Errors in the models construction are presented in **Table 15**.

Table 15. Error information.

	Kohonen map			**Decision Tree**			**Neural networks**		
Calculus system	**0–7**	**1–5**	**1–2**	**0–7**	**1–5**	**1–2**	**0–7**	**1–5**	**1–2**
Error percentage	23.45%	21.38%	11.03%	31.03%	32.41%	13.79%	88.28%	55.86%	4.14%
Incorrect classification	34	31	16	45	47	20	128	81	139
Correct classification	111	114	129	100	98	125	17	64	6
Total rows	145								

The Decision Tree method appears better than the other methods (in this study) in predicting the student's success (**Table 16**). The "Course" attribute is strongly dependent on the success criterion in the Decision Tree models, but is weak in the correlation analysis.

Table 16. Attribute significance.

Attribute	Significance, %			
	0–7	**1–5**	**1–2**	**Correll.**
Course	44.853%	68.249%	95.786%	0.142
Mother's education	8.014%	6.714%	0%	0.066
Projects/events influence	6.243%	0%	0%	−0.203
Gender	6.026%	0%	0%	0.336
Classes attendance	4.337%	0%	0%	−0.14
Regular artistic and sports activities	4.260%	5.650%	0%	−0.063
Age	3.919%	6.508%	4.214%	−0.095
Reading periodicity (nonscientific books/magazines)	3.775%	2.296%	0%	0.196
Partner	3.598%	2.852%	0%	−0.052
Weekly learning hours	3.584%	4.533%	0%	−0.033
Secondary school graduation type	3.325%	3.198%	0%	0.105
Transport to university	2.523%	0%	0%	−0.156
Parents' status	2.057%	0%	0%	0.066
Work	1.980%	0%	0%	0.167
Father's education	0.753%	0%	0%	0.064
Salary amount	0.753%	0%	0%	−0.166
Expected score in the graduate class	0%	0%	0%	0.249
Seminar/conference attendance	0%	05	0%	−0.185
Number of brothers and sisters	0%	0%	0%	0.084
Preparation to the interim exam (time)	0%	0%	0%	0.074

5. Discussion

The article examined three datasets: Student Grade Prediction, Factors about Students Performance Affecting and Higher Education Students Performance Evaluation containing information on school students Gabriel Pereira and Musinho da Silveira, US college students and students of the Engineering Faculty and Faculty of the Pedagogical Sciences, respectively. Analyzing students in the mathematics and Portuguese language courses in high school made it possible to reveal that the success criterion output parameter depended on the number of past failures, desire to obtain higher education and the mother's education. In regard to the US college students, the success influencing factors included parental education, time spent with the parents, and sports. As far as students of the Faculty of Engineering and the Faculty of Pedagogical Sciences were concerned, only the course field value was significant (the older the course, the better was the student's performance).

It is worth noting significant difference in the considered datasets. It lies both in the education level (school, college, university) and in the parameters under consideration. Such differences are making the datasets in question unique and require a special approach in their analysis and interpretation of the research results, which, in turn, are applicable only in particular organization and at different training levels.

To obtain the results, the data was preprocessed to obtain more correct results. For this purpose, data rows that had omissions and errors were removed. Also, the values of each attribute were normalized to the same units and scale.

Despite this, applying the investigated methods to the data analysis, it was possible to notice the appearance of large errors in the obtained results. Their appearance may be due to the choice of an inappropriate method of data analysis, as well as incorrect parameter settings, which, in turn, is limited by the environment in which the analysis was conducted. The "Neural Networks" method is more universal, as it can work with any type of data, with a large number of input parameters and does not depend on the size of the data set.

The Decision Tree method on the platform used misses the ability to fine-tune display of the received data. Kohonen maps and neural networks also do not have enough settings to more accurately identify any dependencies between the factors. Thus, not every method was applied to the datasets.

Each of the considered datasets has its own set of criteria influencing both the success in learning and development of the social qualities. For example, parental education in the US college dataset is significant, while in other datasets it is not. Likewise, sports have the varying importance degree depending on the educational institution. The attribute is significant in the US college, but at the US university it is insignificant. These results stress importance of the tailored machine learning methods in the specific context and demonstrate difficulties in analyzing the social factors that influence the student's success.

Using the methods under consideration made it possible to analyze success factors in the individual social qualities development. To obtain more accurate results, a more detailed study of the datasets content and expanding the datasets are required, as well as an increase in variations of the analysis method settings.

6. Conclusion

The main objective of this study was to identify the success criterion dependences on the individual social qualities and other factors. Success criterion was determined only by the aggregate scoring value in the subjects, which introduced the own characteristics into this study. When considering each dataset, the factors that had more significant dependence on the output factor were taken into account, without also considering the obvious attributes such as grades in the disciplines and the number of missed classes. Decision Tree, Kohonen Maps and Neural Networks methods were used in clustering and forecasting the problems.

For each of the datasets, as expected, its maximally dependent attributes are identified in relation to the others. The analysis of the Student Grade Prediction data and its derived dependencies suggests that if a student has had learning failures in the past, the student will also continue to have failures at higher grades. In such a case it is worthwhile to introduce additional educational hours for such students, so that in the future there will be less such failures and the learner will become more successful.

Also the presence of a role model in the environment (namely mother's education) and further desires to achieve more are directly related to the success criterion in the present time.

Analysis of the "Factors about student performance affecting" data have a similar relationship with the "Student Grade Prediction" data set regarding the influence of parental education on student success. Also, time spent with parents and playing sports also affect student success. During the time students spend interacting with their parents, they gain new knowledge as well as memorize and share what they have learned new. These interactions have a positive impact on students. Exercise in turn increases cognitive function, improves psychological health, develops self-discipline, the ability to set goals and achieve them, which has a positive impact on academic success.

Analysis of the Higher education student performance evaluation data only revealed the effect of course on learning success. The higher the course, the more successful the student's learning. This is facilitated by the accumulation of knowledge and skills, increased motivation and autonomy, as well as the selection of more capable students, as not all continue their studies until the senior year.

This research is planned to be further developed by using other datasets to identify new patterns, as well as applying other methods of analysis with more extensive possibilities for tuning the resulting models. Continued research aims to further the understanding of the factors that influence learning success and the development of social traits, and to develop more accurate and efficient data analysis methods to investigate them. This will not only increase theoretical knowledge, but also provide a basis for the possible development of practical recommendations and strategies in the field of education and social development.

Author contributions: Conceptualization, VT and TP; methodology, AB; software, AG (Anna Glinscaya); validation, AG (Andrei Gantimurov) and VN; formal analysis, YT; investigation, AG (Andrei Gantimurov); resources, VT; data curation, AB; writing—original draft preparation, TP; writing—review and editing, YT; visualization, VN; supervision, AG (Anna Glinscaya); project administration, VT; funding acquisition, TP. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

Algarni, A. (2016). Data mining in education. International Journal of Advanced Computer Science and Applications, 7(6).

ALzubi, J. A., Bharathikannan, B., Tanwar, S., et al. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing, 80, 579-591.

Barantsov, I. A., Pnev, A. B., Koshelev, K. I., et al. (2023). Classification of Acoustic Influences Registered with Phase-Sensitive OTDR Using Pattern Recognition Methods. Sensors, 23(2), 582.

Becker, M., & Tetzner, J. (2021). On the relations of sociocognitive childhood characteristics, education, and socioeconomic success in adulthood. Contemporary Educational Psychology, 67, 102024.

Berezan, V. (2019). The impact of information and communication technologies on professional formation and development of social educator. Available online: http://dspace.pnpu.edu.ua/bitstream/123456789/21668/1/The%20impact%20of%20information.pdf. (accessed on 2 August

2024).
Bukhtoyarov, V. V., Tynchenko, V. S., Nelyub, V. A., et al. (2023). A Study on a Probabilistic Method for Designing Artificial Neural Networks for the Formation of Intelligent Technology Assemblies with High Variability. Electronics, 12(1), 215.
Gairin, J., Triado, X. M., Feixas, M., et al. (2014). Student dropout rates in Catalan universities: profile and motives for disengagement. Quality in Higher Education, 20(2), 165-182.
Godor, B. P. (2017). Academic fatalism: Applying Durkheim's fatalistic suicide typology to student drop-out and the climate of higher education. Interchange, 48(3), 257-269.
Haenggli, M., & Hirschi, A. (2020). Career adaptability and career success in the context of a broader career resources framework. Journal of vocational behavior, 119, 103414.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. Journal of public economics, 95(7-8), 798-812.
Hatos, A., & Pop, A. (2019). Commitment to the goal of completing studies in higher education: Dropout risk of the students from social science specialization from three Romanian public universities. Journal of Adult Learning, Knowledge and Innovation, 3(1), 12-19.
Hodges, L. C. (2018). Contemporary issues in group learning in undergraduate science classrooms: A perspective from student engagement. CBE—Life Sciences Education, 17(2), es3.
Huffman, W. E. (1974). Decision making: The role of education. American Journal of Agricultural Economics, 56(1), 85-97.
Irvine, N., Nugent, C., Zhang, S., et al. (2019). Neural network ensembles for sensor-based human activity recognition within smart environments. Sensors, 20(1), 216.
Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision Analytics Journal, 100204.
Jia, D. W., & Wu, Z. Y. (2021). Seismic fragility analysis of RC frame-shear wall structure under multidimensional performance limit state based on ensemble neural network. Engineering Structures, 246, 112975.
Khan, R. U., Almakdi, S., Alshehri, M., et al. (2022). Probabilistic Approach to COVID-19 Data Analysis and Forecasting Future Outbreaks Using a Multi-Layer Perceptron Neural Network. Diagnostics, 12(10), 2539.
Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? Nature biotechnology, 26(9), 1011-1013.
Kovács, K., Ceglédi, T., Csók, et al. (2019). Students who dropped out 2018 (Hungarian). Hungarian Educational Research Journal, 12(11), 804.
Li, S., Yao, Y., Hu, J., et al. (2018). An ensemble stacked convolutional neural network model for environmental event sound recognition. Applied Sciences, 8(7), 1152.
Maringe, F. (2006). University and course choice: Implications for positioning, recruitment and marketing. International journal of educational management, 20(6), 466-479.
Markos, V., Kocsis, Z., & Dusa, Á. R. (2019). Different Forms of Civil Activity and Employment in Hungary and Abroad, and the Development of Student Drop-out. Central European Journal of Educational Research, 1(1), 41-54.
Masich, I. S., Tyncheko, V. S., Nelyub, V. A., et al. (2022). Paired Patterns in Logical Analysis of Data for Decision Support in Recognition. Computation, 10(10), 185.
Masich, I. S., Tynchenko, V. S., Nelyub, V. A et al. (2022). Prediction of Critical Filling of a Storage Area Network by Machine Learning Methods. Electronics, 11(24), 4150.
Mikhalev, A. S., Tynchenko, V. S., Nelyub, V. A., et al. (2022). The Orb-Weaving Spider Algorithm for Training of Recurrent Neural Networks. Symmetry, 14(10), 2036.
Priyam, A., Abhijeeta, G. R., Rathee, A., et al. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.
Pusztai, G., & Kocsis, Z. (2019). Combining and balancing work and study on the eastern border of Europe. Social Sciences, 8(6), 193.
Pusztai, G., Fényes, H., & Kovács, K. (2022). Factors influencing the chance of dropout or being at risk of dropout in higher education. Education Sciences, 12(11), 804.
Pusztai, G., Fényes, H., Szigeti, F., et al. (2019). Dropped-out Students and the Decision to Drop-out in Hungary. Central European Journal of Educational Research, 1(1), 31-40.
Romanova, I. K. (2022). Applying intelligent data analysis technologies for detecting damages to UAVs. In: AIP Conference Proceedings. AIP Publishing.

Sá, M. J. (2023). Student Academic and Social Engagement in the Life of the Academy—A Lever for Retention and Persistence in Higher Education. Education Sciences, 13(3), 269.

Sansone, C., & Harackiewicz, J. M. (2000). Intrinsic and extrinsic motivation: The search for optimal motivation and performance. Elsevier.

Schreck, T., Bernard, J., Von Landesberger, T., et al. (2009). Visual cluster analysis of trajectory data with interactive kohonen maps. Information Visualization, 8(1), 14-29.

Shafranov-Kutsev, G. F., & Efimova, G. Z. (2019). The place of professional education system in formation of graduates'competitiveness. The Education and science journal, 21(4), 139-161.

Smirnov, S. D. (2004). Psychological factors of successful study of university students. Vestnik Moskovskogo universiteta, 1, 10-35.

Solutions, S. (2019). Correlation. Pearson, Kendall, Spearman.

Ting, C. Y., Cheah, W. N., & Ho, C. C. (2013). Student engagement modeling using bayesian networks. In: Proceedings of 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE. pp. 2939-2944.

Uusiautti, S., & Hyvärinen, S. (2021). Defining the new concept of sustainable success–A state-of-the-art analysis on the phenomenon. New Ideas in Psychology, 60, 100819.

Váradi, J., Demeter-Karászi, Z., & Kovács, K. (2019). The connection between extracurricular, leisure time activities, religiosity and the reasons for drop-out. Central European Journal of Educational Research, 1(1), 55-67.

Volchenko, T. V., Ruzhanskaya, L. S., & Fokeev, M. A. (2021). Non-cognitive skills of employees and their influence on voluntary turnover. Upravlenets, 12(2), 87-101.

Zając, T. Z., & Komendant-Brodowska, A. (2019). Premeditated, dismissed and disenchanted: Higher education dropouts in Poland. Tertiary Education and Management, 25, 1-16.