Science for Education Today

2020, том 10, № 6 http://sciforedu.ru ISSN 2658-6762

DOI: 10.15293/2658-6762.2006.08

УДК 378+004.89

Оценка влияния электронного обучения и социальных параметров на успеваемость студентов

Д. А. Петрусевич (Москва, Россия)

Проблема и цель. *В статье исследуется проблема оценки успеваемости студентов в современной ситуации. Цель статьи – оценить влияние внедрения элементов электронного обучения и некоторых социальных параметров на успеваемость студентов.*

Методология. *Исследование основано на методах машинного обучения, при помощи которых становится возможным оценивать проблемы образовательной системы, поведения студентов и действий администрации образовательных учреждений высшего образования. В работе используются методы математического анализа данных и математической статистики. Автор использует алгоритмы и методы анализа данных, основанные на классификации: решающее дерево, логистическая регрессия и т.д. В целях повышения точности классификации применяются ансамбли классификаторов (градиентный бустинг и случайный лес).*

Результаты. *В центре внимания исследования автора анализ нескольких наборов данных (Students' Performance in Portugal, E-learning Student Reactions и Students' Academic Performance), посвящённых учёту успеваемости студентов нескольких высших и средних учебных заведений в разных странах.*

В результате проведенного исследования были выявлены и обобщены статистические взаимосвязи, существующие между социальными параметрами студентов и их успеваемостью; а также выявлены степени влияния применения онлайн или смешанного формата обучения на показатели успеваемости студентов.

В рамках представленного исследования удалось показать, что методы математической статистики и анализа данных позволяют выявить взаимосвязи в данных, посвящённых успеваемости студентов, выявить неявные зависимости, получить новые актуальные результаты, которые могут быть важны для администрации вузов.

Заключение. *В заключении автором обобщаются результаты проведенной оценки влияния внедрения элементов электронного обучения и некоторых социальных параметров на успеваемость студентов.*

Ключевые слова: *кластеризация студентов, смешанное обучение, оценка успеваемости, цифровизация образования, цифровые технологии в образовании, корреляция признаков, повышение успеваемости.*

Петрусевич Денис Андреевич – кандидат физико-математических наук, доцент кафедры высшей математики, Российский технологический университет (МИРЭА).
E-mail: petrdenis@mail.ru

СПИСОК ЛИТЕРАТУРЫ

1. Amrieh E. A., Hamtini T., Aljarah I. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods // International Journal of Database Theory and Application. – 2016. – Vol. 9 (8). – P. 119–136. DOI: https://doi.org/10.14257/ijdta.2016.9.8.13
2. Андрианова Е. Г., Головин С. А., Зыков С. В., Лесько С. А., Чукалина Е. Р. Обзор современных моделей и методов анализа временных рядов динамики процессов в социальных, экономических и социотехнических системах // Российский технологический журнал. – 2020. – Т. 8, № 4. – С. 7–45. DOI: https://doi.org/10.32362/2500-316X-2020-8-4-7-45 URL: https://elibrary.ru/item.asp?id=43756167
3. Li Y., Allen J., Casillas A. Relating psychological and social factors to academic performance: A longitudinal investigation of high-poverty middle school students // Journal of Adolescence. – 2017. – Vol. 56. – P. 179–189. DOI: https://doi.org/10.1016/j.adolescence.2017.02.007
4. Gimenez G., Martín-Oro Á., Sanaú J. The effect of districts' social development on student performance // Studies in Educational Evaluation. – 2018. – Vol. 58. – P. 80–96. DOI: https://doi.org/10.1016/j.stueduc.2018.05.009
5. Law K. M. Y., Geng S., Li T. Student enrollment, motivation and learning performance in a blended learning environment: The mediating effects of social, teaching, and cognitive presence // Computers & Education. – 2019. – Vol. 136. – P. 1–12. DOI: https://doi.org/10.1016/j.compedu.2019.02.021
6. Salameh W., Sathakathulla A. The Impact of Social-Economic Factors on Students' English Language Performance in EFL Classrooms in Dubai // English Language and Literature Studies. – 2018. – Vol. 8 (4). – P. 110. DOI: https://doi.org/10.5539/ells.v8n4p110
7. Mushtaq B., Jyotsna J. Effect of Socio Economic Status on Academic Performance of Secondary School Students // The International Journal of Indian Psychology. – 2016. – Vol. 3 (4). – P. 56. DOI: https://doi.org/10.13140/RG.2.2.19730.71369
8. Ishizaka A., Lokman B., Tasiou M. A Stochastic Multi-criteria Divisive Hierarchical Clustering Algorithm // Omega. – 2020. – P. 102370. DOI: https://doi.org/10.1016/j.omega.2020.102370
9. Анфёров М. А. Генетический алгоритм кластеризации // Российский технологический журнал. – 2019. – Vol. 7 (6). – P. 134–150. DOI: https://doi.org/10.32362/2500-316X-2019-7-6-134-150 URL: https://www.elibrary.ru/item.asp?id=42347089
10. Asrial M., Habibi A., Mukminin A., Hadisaputra P. Science teachers' integration of digital resources in education: A survey in rural areas of one Indonesian province // Heliyon. – 2020. – Vol. 6 (8). – P. e04631. DOI: https://doi.org/10.1016/j.heliyon.2020.e04631
11. Sarkisov S. S., Lomonosova N. V., Zolkina A. V., Sarkisov T. S. Integration of digital technologies in mining and metallurgy industries // Tsvetnye Metally. – 2020. – Vol. 2020. – P. 7–14. DOI: http://dx.doi.org/10.17580/tsm.2020.03.01
12. Lomonosova N. V., Zolkina A. V. Digital learning resources: Enhancing efficiency within blended higher education // Novosibirsk State Pedagogical University Bulletin. – 2018. – Vol. 8 (6). – P. 121–137. DOI: http://dx.doi.org/10.15293/2226-3365.1806.08 URL: https://www.elibrary.ru/item.asp?id=36655296
13. Деменкова Т. А., Томашевская В. С., Ширинкин И. С. Мобильные приложения для задач дистанционного обучения // Российский технологический журнал. – 2018. – Vol. 6 (1). – P. 5–19. URL: https://www.elibrary.ru/item.asp?id=32466033
14. Alkhowailed M. S., Rasheed Z., Shariq A., Elzainy A., Sadik A.E., Alkhamiss A., Alsolai A. M., Alduraibi S. K., Alduraibi A., Alamro A., Alhomaidan H. T., Al Abdulmonem W. Digitalization

plan in medical education during COVID-19 lockdown // Informatics in Medicine Unlocked. – 2020. – Vol. 20. – P. 100432. DOI: https://doi.org/10.1016/j.imu.2020.100432

15. Andrade H. L. A Critical Review of Research on Student Self-Assessment // Frontiers in Education. – 2019. – Vol. 4. – P. 87. DOI: https://doi.org/10.3389/feduc.2019.00087
16. Aricò F. R., Lancaster S. J. Facilitating active learning and enhancing student self-assessment skills // International Review of Economics Education. – 2018. – Vol. 29. – P. 6–13. DOI: https://doi.org/10.1016/j.iree.2018.06.002
17. Piper K., Morphet J., Bonnamy J. Improving student-centered feedback through self-assessment // Nurse Education Today. – 2019. – Vol. 83. – P. 104193. DOI: https://doi.org/10.1016/j.nedt.2019.08.011
18. Panadero E., Brown G. L., Strijbos J.-W. The future of student self-assessment: a review of known unknowns and potential directions // Educational Psychology Review. – 2016. – Vol. 28 (4). – P. 803–830. DOI: https://doi.org/10.1007/s10648-015-9350-2
19. Sharma R., Amit J., Gupta N. Garg S., Batta M., Dhir S. Impact of self-assessment by students on their learning // International Journal of Applied and Basic Medical Research. – 2016. – Vol. 6 (3). – P. 226. DOI: https://doi.org/10.4103/2229-516X.186961
20. Erkens M., Bodemer D. Improving collaborative learning: Guiding knowledge exchange through the provision of information about learning partners and learning contents // Computers & Education. – 2019. – Vol. 128. – P. 452–472. DOI: https://doi.org/10.1016/j.compedu.2018.10.009
21. Liao C.-W., Chen C.-H. & Shih S.-J. The interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behavior patterns in a digital game-based learning environment // Computers & Education. – 2019. – Vol. 133. – P. 43–55. DOI: https://doi.org/10.1016/j.compedu.2019.01.013
22. Hernández-Sellés N., Muñoz-Carril P.-C., González-Sanmamed M. Computer-supported collaborative learning: an analysis of the relationship between interaction, emotional support and online collaborative tools // Computers & Education. – 2019. – Vol. 138. – P. 1–12. DOI: https://doi.org/10.1016/j.compedu.2019.04.012
23. Díaz-Ramírez J. Gamification in engineering education – An empirical assessment on learning and game performance // Heliyon. – 2020. – Vol. 6 (9). – P. e04972. DOI: https://doi.org/10.1016/j.heliyon.2020.e04972
24. Золкина А. В., Ломоносова Н. В., Петрусевич Д. А. Оценка востребованности применения геймификации как инструмента повышения эффективности образовательного процесса // Science for Education Today. – 2020. – Т. 10, № 3. – С. 127–143. DOI: http://dx.doi.org/10.15293/2658-6762.2003.07
25. Landers R. N., Landers A. K. An empirical test of the theory of gamified learning. The effect of leaderboards on time-on-task and academic performance // Simulation & Gaming. – 2015. – Vol. 45 (6). – P. 769–785. DOI: http://dx.doi.org/10.1177/1046878114563662
26. Rastrollo-Guerrero J. L., Gómez-Pulido J. A., Durán-Domínguez A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review // Applied Sciences. – 2020. – Vol. 10 (3). – P. 1042. DOI: https://doi.org/10.3390/app10031042
27. Asif R., Merceron A., Ali S. A., Haider N. G. Analyzing undergraduate students' performance using educational data mining // Computers & Education. – 2017. – Vol. 113. – P. 177–194. DOI: https://doi.org/10.1016/j.compedu.2017.05.007
28. Fernandes E., Holanda M., Victorino M., Borges V., Carvalho R., Erven G. V. Educational data mining: Predictive analysis of academic performance of public school students in the capital of

Brazil // Journal of Business Research. – 2019. – Vol. 94. – P. 335–343. DOI: https://doi.org/10.1016/j.jbusres.2018.02.012

29. Yang F., Li F. W. B. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining // Computers & Education. – 2018. – Vol. 123. – P. 97–108. DOI: https://doi.org/10.1016/j.compedu.2018.04.006
30. Ripoll V., Godino-Ojer M., Calzada J. Teaching Chemical Engineering to Biotechnology students in the time of COVID-19: assessment of the adaptation to digitalization // Education for Chemical Engineers. – 2020. – Vol. 34. – P. 94–105. DOI: https://doi.org/10.1016/j.ece.2020.11.005
31. Mishra L., Gupta T., Shree A. Online Teaching-Learning in Higher Education during Lockdown Period of COVID-19 Pandemic // International Journal of Educational Research Open. – 2020. – P. 1000012. DOI: https://doi.org/10.1016/j.ijedro.2020.100012

Science for Education Today

2020, Vol. 10, No. 6 http://en.sciforedu.ru/ ISSN 2658-6762

DOI: 10.15293/2658-6762.2006.08

Denis Andreevich Petrusevich
Candidate of Physical and Mathematical Sciences, PhD, Associate Professor,
Higher Mathematics Department,
Russian Technological University (RTU MIREA), Moscow, Russian Federation.
ORCID ID: http://orcid.org/0000-0001-5325-6198
E-mail: petrdenis@mail.ru

The impact of e-learning and social parameters on students' academic performance

Abstract

Introduction. *The article examines the problem of assessing students' academic performance in the current situation.*

The purpose of the paper is to evaluate the influence of e-learning and some social and behavioral parameters on students' academic performance.

Materials and Methods. *The author employed the machine learning procedures in order to identify and assess the current problems of the educational system, students' behavior, and universities' policy. Methods of mathematical analysis and statistics as well as ensemble methods (gradient boosting and the random forest algorithms) were used in order to achieve high accuracy of the research.*

Results. *The author conducted the analysis of the following datasets devoted to academic performance at higher and secondary educational institutions in a number of countries: Students' Performance in Portugal, E-learning Student Reactions and Students' Academic Performance.*

The purposes of the current study were to identify statistical correlations between social parameters of students and the level of their academic performance and to understand how academic performance is determined by the implementation of online learning and blended learning.

The research findings suggest that mathematical statistics and data analysis methods allow to identify correlations between students' performance data and reveal hidden relationships which can be important for university staff.

Conclusions. *In conclusion, the author summarizes the results of evaluating the impact of the introduction of e-learning elements and some social parameters on students' academic performance.*

Keywords

Clustering students; Blended learning; Academic performance evaluation; Digitalization of education; Digital technologies in education; Correlation of features; performance improvement.

Introduction

Data science methods applied in a large variety of domains of knowledge allow to find hidden dependencies, divide data into new set of clusters that can better explain structure of information, construct classifiers and so on. These methods are used to solve a lot of different tasks and problems of pedagogical science and dependencies between social parameters of students and their development and performance appear in the scope of this science. In this research there are four datasets that have been under investigation. The Students' Performance

dataset[1,2,3] has got two parts including grades at math and Portuguese language exams. There are values of social and behavioral features of each student. Thus, it's possible to construct and test statistical hypotheses on dependencies between quality of student's life and his or her performance. Though, it's difficult to explain these logical dependencies and it's possible to treat these results as mutual dependencies on a set of unobservable values. At the same time, these logical connections are the best result nowadays. They allow constructing hypotheses on changes in social structure that could lead to students' performance improvement.

The university staff can analyze clusters containing their students, handle each cluster separately improving performance in each of them with own methods. At the same time dependencies between students' social parameters and performance can be treated by government in order to improve development of citizens.

The E-learning Student Reactions dataset[4] contains information that potentially can reveal dependencies between students' performance in traditional forms of study and their achievements inside of the e-learning system that supports collaborative learning. Dependencies between traditional grades and emodji-based reactions on messages in e-learning system can confirm that e-learning process leads to gaining knowledge by students in the way they like or it can show that reactions on their behaviour in this system don't correlate with traditional grades.

The Students' Academic Performance dataset [1][5,6] can be used to test dependencies between grades gained by students in traditional forms of learning process and their activity inside of the e-learning system.

These analysis experiments are important in new educational conditions. The Covid-19 pandemic makes universities all over the world move to distant or blended forms of study. Analysis of students' behaviour inside of such systems can reveal how such transformation affects quality of learning process. It's also important to understand whether students evaluating help from their mates in learning process construct the grades of the same structure as their teachers or their reactions correlate with teachers' grades.

Methods

The research is based on analysis and generalization of papers and books concerning the main theme. There are two forms of students' grades. The first one is an integer value in some diapason. In other experiments the grades are transformed into a parameter that can take value from a small set. Such problems are usually considered as classification tasks and they are solved by with the decision tree, logistic regression classifiers are ensembles (the gradient boosting and the random forest classifiers). These methods are based on mathematical statistics. It

[1] *Student Grade Prediction*. URL: https://www.kaggle.com/dipam7/student-grade-prediction

[2] *Student Performance Data Set*. URL: https://www.kaggle.com/larsen0966/student-performance-data-set

[3] Cortez P., Silva P. Using Data Mining to Predict Secondary School Student Performance. *Proceedings of the 5th Future Business TEChnology Conference* (FUBUTEC 2008) EUROSIS. 2008. pp. 5–12.

[4] *E-learning Student Reactions*. URL: https://www.kaggle.com/marlonferrari/elearning-student-reactions

[5] Amrieh E. A., Hamtini T., Aljarah I. Preprocessing and analyzing educational data set using X-API for improving student's performance. *Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE Jordan Conference 2015. pp. 1–5. DOI: https://doi.org/10.1109/AFECT.2015.7360581

[6] *Students' Academic Performance Data set*. URL: https://www.kaggle.com/aljarah/xAPI-Edu-Data

should be mentioned that statistical results have to be explained by researchers accurately and conclusions have to be made to allow governments or university staffs to correct learning process.

Results

The Portuguese language and math final grades are main values that show level of quality of educational process. Results of students of two colleges have been explored: Gabriel Pereira and Mousinho da Silveira school (*parameter school*). Features in the dataset can be divided logically into three groups: binary parameters, integer values in small diapason and integer values in large diapason. The first group contains *activities* (extra-curricular activities), *address* (urban or rural), *famsup* (is there educational support from student's family?), *higher* (does this student intend to get higher education?), *internet* (is there access to internet at home?), *nursery* (attended nursery school), *paid* (does the student take part in additional paid classes?), *Pstatus* (do parents live together or apart?), *romantic* (has this student got a relationship?), *schoolsup* (extra educational support), *sex*.

The second type of features consists of integers in small diapason: *age* is between 15 and 22 years; *famrel* (quality of relationships in the family), *Medu* and *Fedu* (level of parent's education) *Mjob* and *Fjob* (type of parent's job), *freetime* and *goout* (how often does this student has got free time and can go out?), *Dalc* and *Walc* (daily and weekend level of alcohol consumption), *health* status have got 5 possible levels. Variables *reason* (reason to choose certain school), *traveltime* and *studytime* (how much time does student spend to get to school, doing homework?), *failures* (number of past class failures) have got 4 levels. Student's guardian (mother, father, other) is explained with the *guardian* parameter. Features *absences* and *G1*, *G2*, *G3* grades can also be treated as factor variables but there's a lot of levels. Grades belong to diapason between 0 and 20. More thorough explanation of the features is presented in the paper of F. Unal[7]. Influence of social features on students' performance is also discussed in [3–5].

The grades can be transformed into parameters showing whether they belong to some diapason[7]. Thus this problem can be transformed into a classification task. But another way is to construct regression models explaining the grade variables by means of all other parameters that are contained in the dataset. In the present work the second way is used and the *G3* grade is used as an explained value. The *G1*, *G2* values are intermediate grades that form the final grade. So, there's logical dependence between them and only one of them should be used in linear models.

All parameters of the second group are transformed with one-hot encoding technique (or dummy variables are created)[8,9].

Number of absences is transformed according to expression (1):

$$x' = \frac{x - \mu}{\sigma}. \qquad (1)$$

In the expression (1) x is a source value, x' is a transformed version of this parameter, μ is mean value of x and σ is its standard deviation[8, 9].

[7] Unal F. Data Mining – Methods, Aplications and Systems. *Data Mining for Students Performance Prediction in Education. IntechOpen*, 2020. DOI: https://doi.org/10.5772/intechopen.91449

[8] James G., Witten D., Hastie T., Tibshirani R. *An introduction to statistical learning with applications in R.* Springer-Verlag, New York, USA, 2015. 426 p. DOI: https://doi.org/10.1007/978-1-4614-7138-7

[9] Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning*. Springer-Verlag, New York, USA, 2009. 533 p.

2020, Vol. 10, No. 6 http://en.sciforedu.ru/ ISSN 2658-6762

Dependencies between social and economic factors and digitalization and quality of education are investigated in papers [3; 5–7] that analyze situation in different countries and various universities.

The investigated dataset containing the Portuguese language grades have been clustered into five groups. The clustering technique allows dividing dataset into several clusters depending on inner structure of the data. Number of clusters is obtained with the agglomerative clustering techniques [8; 9]. The Euclidean metrics is used to measure distance between objects and the Ward's method is implemented in order to measure distance between clusters. The structure of the dataset in these terms is presented at the figure 1. The Y axis shows the distance between clusters and the X axis shows the objects inside the dataset united into clusters.

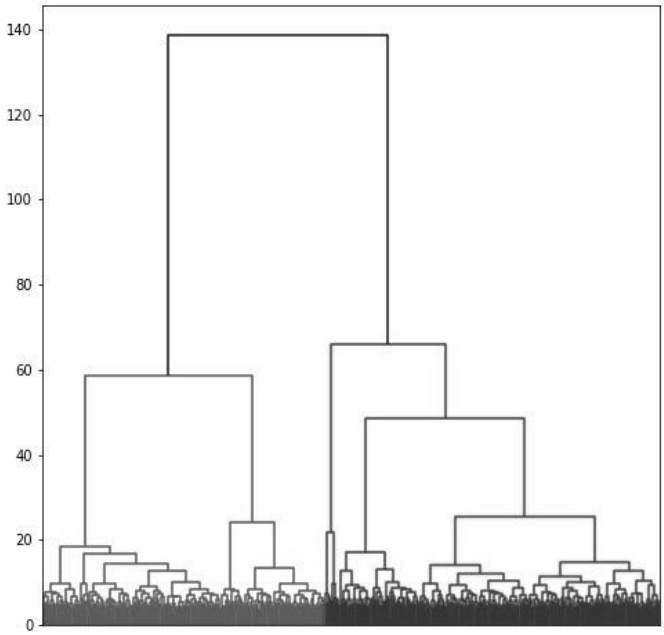

Fig. 1. Result of the agglomerative clustering method implemented at the Students' Performance dataset (the Portuguese language)

Number of clusters is obtained with setting horizontal line at the figure at some level. Number of intersections with hierarchical structure equals to number of clusters.

Thus, there are five clusters. One of them is quite small. To explain the hierarchical structure of clusters mean values of final grades are counted inside of each one. "Parent" clusters are divided into a set of smaller ones that are considered as "children". This action can be repeated until there's only one object in each cluster. But such approach can't lead to generality. So, this division must be stopped at some level that is obtained analyzing the figure 1.

The structure of clusters is presented in the table 1. Quantity of objects in clusters, mean value of the final grade in each cluster and "parent" clusters are shown in the columns.

Table 1

Mean values of students' grades in each cluster (the Portuguese language)

Name of the cluster	Quantity of objects	Mean value of the final grade	Parent cluster
C1	368	9.30	dataset
C2	281	13.89	dataset
C11	16	0.06	C1
C12	265	9.86	C1
C121	48	7.71	C12
C122	217	10.33	C12
C21	190	12.38	C2
C22	178	15.51	C2

The first cluster C1 contains students with lower *G3* grade. There are 336 students of overall 649 ones. In the second one the grade is higher. The majority of students in the first cluster belong to the Mousinho da Silveira school. They've got more absences. These students are worse students by academic activities. Among them there are more males. More often they live in rural areas than in urban ones. Level of their parents' education is lower. Mothers of the students in the first cluster work at home more often. There are more students in the second cluster whose parents have got higher education.

The distance from school to home is higher in the 1st cluster. But difference isn't large. In the first cluster there are more students that have to pay 15–30 minutes. In the second one more students pay 15 minutes or less to get to school.

Students in the first cluster usually pay less time to learning process. The majority of the students in the 2nd cluster C2 have got study time between 5 and 10 hours.

The agglomerative clustering algorithm has been implemented to the both clusters. The C1 cluster containing students with lower grades is divided into three parts: C121, C122, C11. There's a small cluster C11 that contains only 16 records with extremely low grades. There are no general conclusions on this cluster so considering the other part C12 of the 1st cluster allows to construct two clusters: C121, C122.

The C122 cluster with better academic results is going to be discussed. The number of absences is less, the grades are higher than in the C121 cluster. In this cluster the majority of students pass their exams and in the other cluster usually there are 1 or 2 past class failures. The family educational support is higher and students tend to get higher education though in common their parents' level of education is lower. More students answer that their health level is very good.

In each cluster linear regression model has been constructed. The *G3* grade variable was explained with all other parameters in the dataset. Thus, full model contains all parameters of the dataset (except grades) as regressors. There's a lot of insignificant parameters. After that there has been an attempt to construct reduced models containing only significant parameters. All insignificant regressors have been removed from a model. After that regressors that have become insignificant are also removed from the model.

In the cluster C121 the determination coefficient R^2 value [10, 11] of the full model is 89.7 % and the adjusted coefficient R^2_{adj} = 22.6 %. It means that there's a lot of other factors that aren't included into this dataset which could explain behaviour of students. All terms in the regression are insignificant though the determination coefficient value is high and one can conclude that parameters explain dynamics of the *G3* value well.

In the cluster C122 the determination coefficient R^2 value of the full model is 41.5 % and the adjusted coefficient R^2_{adj} = 17.5 %. It means that there is a lot of other factors that aren't included into this dataset which could explain behaviour of students. But still there are significant terms that need to be considered thoroughly.

The insignificant terms have been removed from the model. The expression (2) shows the result of this process:

$$G_3 = 10.20 + 0.72\,paid + 0.65\,failures0 - 0.47Dalc3. \quad (2)$$

Here *failures0* is equal to 1 for students that have got zero failures and *Dalc3 = 1* for students with medium daily alcohol consumption. So, alcohol consumption decreases the grade and students without failed exams usually have got higher rades.

R^2 value of the reduced model is about 11.5%. Thus, there are factors which aren't reflected in this model.

The cluster C2 with higher grades can also be divided into two parts. There are fewer absences in the part with higher grades C22. So, this part can be explained as a part of students having better academic results. At the same time their parents have got higher education more often. So, these students tend to be forced to get higher education in a social way. The other part gets extra educational support more often.

Let's consider the part with higher grades, i.e. the C22 cluster. The determination coefficient R^2 value of the full model is 68.3 % and the adjusted coefficient R^2_{adj} = 18.7 % and also there are a few significant terms that need to be considered thoroughly. The reduced model structure is shown in the expression (3):

$$G_3 = 16.36\,failures0 + 0.48Fedu2 - 0.44romantic. \quad (3)$$

Here *failures0 = 1* for students that with zero past class failures, *Fedu2 = 1* for students whose fathers have got 5^{th} – 9^{th} grade education, *romantic = 1* for students with relationships. Thus, relationships decrease final grades but in common this effect "costs" about 1 point. Also, if there are no past class failures the grade should be higher. In this model R^2 = 7.9 %.

In the cluster C21 the determination coefficient R^2 value of the full model is 51.6 % and the adjusted coefficient R^2_{adj} = 22.6 %. These values can be considered as quite small but the reduced model can be constructed. It's shown in the expression (4):

[10] James G., Witten D., Hastie T., Tibshirani R. *An introduction to statistical learning with applications in R.* Springer-Verlag, New York, USA, 2015. 426 p. DOI: https://doi.org/10.1007/978-1-4614-7138-7

[11] Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning.* Springer-Verlag, New York, USA, 2009. 533 p.

$$G_3 = 13.31 - 0.42 school\,\text{sup} + 0.39 age3 + 0.53\, freetime2 + 0.36\, freetime4 + 0.72\, freetime5 - 0.35 health3. \quad (4)$$

Here *schoolsup = 1* for students with extra educational support, *age3* is 1 for students of the 3rd course, *freetime2 = 1* for students with less quantity of free time than average level, *freetime4 = 1* or *freetime5 = 1* means that there's a lot of free time after lessons. *health3 = 1* if current health status is normal. In this model R^2 = 15.5 %.

Full model that has been constructed for the whole dataset has got R^2 = 43.3 %, R^2_{adj} = 35.9 %. The reduced model has got values of determination coefficients: R^2 = 25.3 %, R^2_{adj} = 24.4 %. Thus, full models in clusters are better by value of the coefficient of determination. But the reduced models look worse: values of the R^2 coefficient is better in the model constructed for the whole dataset.

$$G_3 = 10.39 - 1.79 school - 0.73 sex - 1.4 school\,\text{sup} + 3.01 higher + 0.53\, freetime2 + 0.86 studytime3 + 0.62 goout2 - 2.69 Dalc4 - 0.78 health5. \quad (5)$$

Here *higher* denotes whether a student wants to take higher education, *freetime2 = 1* for students with less quantity of free time than average level, *studytime3=1* if this student pays amount of time close to average level to do homework, he or she goes out less (*goout2 = 1*), *Dalc4 = 1* means that the student consumes alcohol more often than other students, *health5 = 1* for students with excellent health. In the expression (5) students of different schools are differentiated, there are differences between males and females, students with relatively low grades are supported. Desire to take higher education correlates with high level of grades. Students with high grades go out less often than other students. Alcohol consumption is lower.

Overall results of the linear regression model for the whole dataset and models constructed in each cluster are presented in the table 2. The models constructed for special clusters have got better R^2 values everywhere except cluster C122. At the same time the best reduced model has been constructed for the whole dataset. In this task there are too few records to construct good models. According to low values of determination coefficients it's possible to conclude that there are unobservable parameters or that students' grades can't be explained only by their means.

Table 2

Comparison of linear regressions constructed for each cluster and for the whole dataset

Name of the cluster	R^2, R^2_{adj} of the full model, %	R^2 of the reduced model, %
The whole dataset	43.3, 35.9	25.3
C121	89.7, 22.6	–
C122	41.5, 17.5	11.5
C21	51.6, 22.6	15.5
C22	68.4, 18.7	7.9

2020, Vol. 10, No. 6 http://en.sciforedu.ru/ ISSN 2658-6762

The same technique has also been implemented in order to analyze the second part of this dataset which contains the same features and grades at the math exams. Scheme of agglomerative clustering implemented to this dataset is shown at the figure 2.

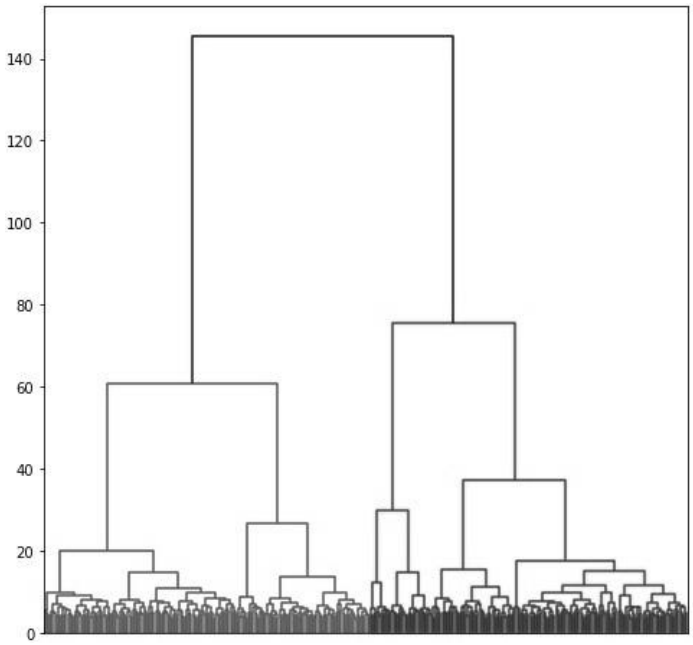

Figure 2. Result of the agglomerative clustering method implemented at the Students' Performance dataset (math)

The dataset has also been clustered. There are four major parts that can be treated with use of mean values of final grades. Their structure is presented in the table 3. Also value of determination coefficient of the full linear model explaining behavior of the grade *G3* variable is shown in the last column.

Table 3

Mean values of students' grades in each cluster (maths)

Name of the cluster	Quantity of objects	Mean value of the final grade	Parent cluster	R^2 of the full model, %
C1	195	7.03	dataset	–
C2	200	13.72	dataset	–
C11	38	0.00	C1	–
C12	157	8.72	C1	46.8
C21	118	12.18	C2	59.8
C22	82	15.94	C2	66.7

The linear regression model in the C11 cluster hasn't been built because of low number of instances in it. It's also noticeable that mean value of the grade is zero. The R^2 of the full linear model constructed for the whole dataset is equal to 18.2 % and is less than the same values calculated for each cluster individually (except the C11 cluster). In all clusters determination

coefficients of reduced models are very low and they aren't presented in the paper. To show dependencies one can glance at the linear model constructed for the whole dataset in the form of the expression (6):

$$G_3 = 6.41 + 0.93\,famsize + 1.30sex - 1.10romantic + 0.54absences + 4.03\,failures0 + 1.86studytime3. \quad (6)$$

Here *failures0* shows if a student has got no failed exams, *studytime3=1* for students who pay approximately average value of time to study. Thus, family size correlates with grades positively. Students with higher grades have got relationships less often, there are less failures. Positive correlation between grades and absences can be explained with confidence of students that they know enough to pass exams. But they don't want to move to clusters with higher grades (for students with low grades).

One can explain differences between clusters considered in the table 3.

The C2 cluster contains almost a half of the whole dataset and records about students with higher grades. School support is lower for these students, there are less absences but less freetime. More often parents of the students in the C1 cluster have got low level of education and in the second cluster there are students with parents having higher education more often. Students in the second cluster haven't got any failures at exams much more often.

The cluster C2 with higher grades are divided into two almost equal by size parts. Also the cluster C22 with higher grades is considered. There are less absences and romantic relationships are less often. More students have got access to internet. They live in urban areas. Their parents have got higher education more often and at the same time there are more teachers among them. It's noticeable that there are less students with no failures but more students with 1 failed exam than in the C21 cluster. Low alcohol consumption is noticed more often in the C22 cluster but students in the C21 cluster evaluate health status as normal more often.

It's difficult to construct logically "strong" conclusions on dependencies between social parameters and grades. In all parts of society there are students with high grades and with low ones. One part pays a lot of time to study, the second one does only part of their work to have some medium level and has got a lot of free time. One of the most interesting cluster is the C122. There are students from the lower half by performance but they've got a chance to get to the higher part and usually they work to achieve it. Also, their parents encourage and support them in this desire. That's man cluster that can be handled with university staff to achieve higher overall performance.

Dependencies of social and economic features are more thoroughly investigated in [2–7].

The dataset about students' behavior in e-learning system has been investigated afterwards. The e-learning system has been very useful addition to standard courses. But during the Covid-19 pandemic universities had only two possible ways: either to stop learning process, or to use distant learning techniques with support of e-learning systems. Thus, research of online, distant and blended learning systems are extremely actual and valuable [10–14]. During the investigated experiment students could discuss complex topics in forum of such system. The main subject is discussion of algorithms. Students' comments could be evaluated by them with possible marks: "helpful", "nice_code", "creative", "amazing", "collaborative" and also

"confused", "bad". Each record contains information on amount of such votes on student's posts and academic grades. So, it's possible to compare students' activity in the system and their final grades. In this approach students evaluate the most interesting answers and it can be considered as some kind of self-evaluation. This question has been in scope of view of pedagogical research worldwide[12] [15–19].

The final grade were divided into five parts: evaluation of critical thinking and problem solving skills inside of the first mark *sk1*, creativity and innovations in the second one *sk2*, constant and self-learning skills in the third part *sk3*, collaboration and self-direction skills *sk4* in the fourth grade and social and cultural responsibility in the last part of the grade *sk5*. All parameters have been handled according to expression (1). Among them there's also time spent by students in the system.

Amount of marks "helpful", "nice_code", "creative", "amazing", "collaborative" correlate with each other and with academic grades. Thus, these "positive" votes are usually set together at some posts. Though amount of "confused" and "bad" posts don't correlate. Five considered "positive" marks have got some level of correlation with final grades. But correlation among grades and also correlation among "positive" votes are stronger. In the table 4 there are correlation levels of pairs of features of the considered dataset. To save space only a part of them is presented: only one grade and five "positive" marks. The grades correlate with each other, "bad_posts" and "confused_posts" don't correlate with other marks. Collaborative learning is investigated in a lot of papers [20–22].

Table 4

Pair correlations between features of the E-learning Students Reactions dataset

	"helpful"	*"nice_code"*	*"collaborative"*	*"creative"*	*"amazing"*	*timeonline*	*sk1*
"helpful"		0.92	0.90	0.92	0.89	0.57	0.44
"nice_code"	0.92		0.87	0.89	0.86	0.61	0.44
"collaborative"	0.90	0.87		0.94	0.94	0.69	0.60
"creative"	0.92	0.89	0.94		0.95	0.69	0.58
"amazing"	0.89	0.86	0.94	0.95		0.69	0.58
timeonline	0.57	0.61	0.69	0.69	0.69		0.63
sk1	0.44	0.44	0.60	0.58	0.58	0.63	

In the table 4 it's clear that "positive" posts correlate with each other. Also, number of these marks correlate with time spent in the system but this logical connection is slightly weaker. And finally one can say that these marks set by students have got some level of correlation with final grades *sk1*,…, *sk5*.

Linear regression models explaining number of "positive" posts variable with "negative" posts, time spent in the system and final grades have been tested. The final grade *sk4*

[12] Brown G. T., Harris L. R. Student self-assessment in Sage. In: (ed.) J. H. McMillan *Handbook of Research on Classroom Assessment.* Los Angeles, CA, Sage, USA, 2013. pp. 367–393. DOI: https://doi.org/10.4135/9781452218649.n21

evaluates "collaborative" skills. That's why only "collaborative_posts" variable remains in the dataset and all other "positive" marks have been removed from the data because of correlation. Because of the same mathematical problems only the *sk4* remains in the dataset and all other grades have been removed. "bad_posts" and "confused_posts" variables don't correlate with the grade and they have been removed from the model. After that integer degrees of the "collaborative_posts" *cp* variable have been consequently added to the model. Thus, a part of the Taylor series is used to explain the grade variable. One can see this result in the formula (7). All degrees have been handled according to expression (1).

$$sk_4 = 0.27timeonline + 4.87cp - 22.45cp^2 + 43.17cp^3 - 25.06cp^4. \quad (7)$$

The regression (7) has got level of R^2 coefficients: R^2 = 61.3 %, R^2_{adj} = 58.4 %. Thus, there's correlation and dependence between students' evaluation of comments and academical grades. But these grades aren't expressed with each other directly.

Thus, one can see that the grades set by teachers have got correlation with emodji-based reactions of students that were supposed to evaluate the same value. The final grades correlate with appropriate students' marks and there's positive correlation with time spent in the system. This result has been expected and it confirms the model.

Thorough research devoted to implementation of gamefication in educational process can be found in [21; 23–25].

The Students' Academic Performance dataset contains information about activities of students of the Jordan university. There are several subjects: math, IT, foreign languages. There are students from several countries from the Middle East, Europe and America. The information gathered in the dataset is obtained from the learning management system: *VisitedResources* (how many times does the student visit online resources recommended in materials of some discipline?), *AnnouncementsView* (how many times does the student read announcements in online content of the discipline) and *Discussion* (how many times does the student participate in discussion groups?). There's also information on student's activity during traditional practice and lectures: *RaisedHands* (how many times does the student take part in discussion during traditional lessons?), *StudentAbsenceDays* (how many days was this student absent?). Authors of the dataset placed there information about nationality and relationships with family but they are out of special consideration in the paper.

This dataset can be investigated in order to describe connections between students' achievements in traditional form and their behaviour and grades in learning management system. Some parts of such analysis have been made handling another arrays of data in [9–10].

Grades of students have been transformed into three possible values: high (90-100 points), medium (70 – 90) and low (less than 70) ones. Thus, it's possible construct classifiers into these groups. Structure of such classifiers can be treated as a set of main features that correlate with medium and high grades and can be influenced in order to improve students' achievements. The decision tree algorithm implementation leads to

construction of classification trees [13]. Their structure is itself an explanation of the classification process and also feature importance values can be built.

The classification trees that have got an accuracy of 70 % (weighted F_1 value[14,15]) have been built. The boosting and bagging techniques allow constructing more accurate classifiers but their behavior can't be explain in a simple way.

The most important features for classification by grade are: *StudentAbsenceDays* (27 %), *VisitedResources* (20 %), *RaisedHands* (10 %), *AnnouncementsView* (6.5 %) and *Discussion* (5.4 %). Thus, grade diapasons correlate with students' behaviour during classic lessons and lectures and with their activity in an electronic learning system.

There are papers devoted to students' performance analysis [26–29] investigating other datasets and posing other questions for analysis.

Discussion, Conclusions

The E-learning Student Reactions and Students' Academic Performance datasets have been considered in this research. The students' performance inside of the e-learning systems and in traditional form has been investigated. There's correlation between behaviour during traditional lessons and study in such systems. Students' self-evaluation grades and teachers' grades also correlate. Thus, blended learning systems and online systems are one of appropriate ways of education development that shouldn't make it worse.

Also, dependencies between social and behavioral parameters and students' grades have been considered in the Students' Performance in Portugal dataset analysis. Records of students have been clustered into a few groups. It's possible to say that students of various have got some common social features though such conclusions must be made very carefully. The cluster C122 contains students from upper part of the lower half of students by grades. Expression (2) shows that students that would like to get better results have got extra paid lessons and they've got no failed exams. These students are the best potential goal for university staff to achieve higher average performance. At the same time students with lower grades have got school support in the cluster C21 containing lower part of the higher half of students by grade. Students with higher grades have got less free time. They're usually encouraged by their parents to study well and to take higher education.

Digitalization of education is especially actual nowadays when a lot of courses worldwide are supported and held in online mode due to struggle against Covid-19 pandemic [14; 30; 31].

REFERENCES

1. Amrieh E. A., Hamtini T., Aljarah I. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 2016, vol. 9 (8), pp. 119–136. DOI: https://doi.org/10.14257/ijdta.2016.9.8.13

[13] Breiman L., Freidman J. H., Olshen R. A., Stone C. J. *Classification And Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, USA. 358 p. ISBN: 0534980548

[14] James G., Witten D., Hastie T., Tibshirani R. *An introduction to statistical learning with applications in R.* Springer-Verlag, New York, USA, 2015. 426 p. DOI: https://doi.org/10.1007/978-1-4614-7138-7

[15] Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning*. Springer-Verlag, New York, USA, 2009. 533 p.

2. Andrianova E. G., Golovin S. A., Zykov S. V., Lesko S. A., Chukalina E. R. Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems. *Russian Technological Journal*, 2020, vol. 8 (1), pp. 7–45. (In Russian) DOI: https://doi.org/10.32362/2500-316X-2020-8-4-7-45 URL: https://elibrary.ru/item.asp?id=43756167
3. Li Y., Allen J., Casillas A. Relating psychological and social factors to academic performance: A longitudinal investigation of high-poverty middle school students. *Journal of Adolescence*, 2017, vol. 56, pp. 179–189. DOI: https://doi.org/10.1016/j.adolescence.2017.02.007
4. Gimenez G., Martín-Oro Á., Sanaú J. The effect of districts' social development on student performance. *Studies in Educational Evaluation*, 2018, vol. 58, pp. 80–96. DOI: https://doi.org/10.1016/j.stueduc.2018.05.009
5. Law K. M. Y., Geng S., Li T. Student enrollment, motivation and learning performance in a blended learning environment: The mediating effects of social, teaching, and cognitive presence. *Computers & Education*, 2019, vol. 136, pp. 1–12. DOI: https://doi.org/10.1016/j.compedu.2019.02.021
6. Salameh W., Sathakathulla A. The Impact of Social-Economic Factors on Students' English Language Performance in EFL Classrooms in Dubai. *English Language and Literature Studies*, 2018, vol. 8 (4), pp. 110. DOI: https://doi.org/10.5539/ells.v8n4p110
7. Mushtaq B., Jyotsna J. Effect of Socio Economic Status on Academic Performance of Secondary School Students. *The International Journal of Indian Psychology*, 2016, vol. 3 (4), pp. 56. DOI: https://doi.org/10.13140/RG.2.2.19730.71369
8. Ishizaka A., Lokman B., Tasiou M. A Stochastic Multi-criteria Divisive Hierarchical Clustering Algorithm. *Omega*, 2020, pp. 102370. DOI: https://doi.org/10.1016/j.omega.2020.102370
9. Anfyorov M. A. Genetic clustering algorithm. *Russian Technological Journal*, 2019, vol. 7 (6), pp. 134–150 (In Russian) DOI: https://doi.org/10.32362/2500-316X-2019-7-6-134-150 URL: https://www.elibrary.ru/item.asp?id=42347089
10. Asrial M., Habibi A., Mukminin A., Hadisaputra P. Science teachers' integration of digital resources in education: A survey in rural areas of one Indonesian province. *Heliyon*, 2020, vol. 6 (8), pp. e04631. DOI: https://doi.org/10.1016/j.heliyon.2020.e04631
11. Sarkisov S. S., Lomonosova N. V., Zolkina A. V., Sarkisov T. S. Integration of digital technologies in mining and metallurgy industries. *Tsvetnye Metally*, 2020, vol. 2020, pp. 7–14. DOI: http://dx.doi.org/10.17580/tsm.2020.03.01
12. Lomonosova N. V., Zolkina A. V. Digital learning resources: Enhancing efficiency within blended higher education. *Novosibirsk State Pedagogical University Bulletin*, 2018, vol. 8 (6), pp. 121–137. DOI: http://dx.doi.org/10.15293/2226-3365.1806.08 URL: https://www.elibrary.ru/item.asp?id=36655296
13. Demenkova T. A., Tomashevskaya V. S., Shirinkin I. S. Mobile applications for tasks of distance learning. *Russian Technological Journal*, 2018, vol. 6 (1), pp. 5–19. (In Russian) URL: https://elibrary.ru/item.asp?id=32466033
14. Alkhowailed M. S., Rasheed Z., Shariq A., Elzainy A., Sadik A.E., Alkhamiss A., Alsolai A. M., Alduraibi S. K., Alduraibi A., Alamro A., Alhomaidan H. T., Al Abdulmonem W. Digitalization plan in medical education during COVID-19 lockdown. *Informatics in Medicine Unlocked*, 2020, vol. 20, pp. 100432. DOI: https://doi.org/10.1016/j.imu.2020.100432
15. Andrade H. L. A Critical Review of Research on Student Self-Assessment. *Frontiers in Education*, 2019, vol. 4, pp. 87. DOI: https://doi.org/10.3389/feduc.2019.00087

16. Aricò F. R., Lancaster S. J. Facilitating active learning and enhancing student self-assessment skills. *International Review of Economics Education*, 2018, vol. 29, pp. 6–13. DOI: https://doi.org/10.1016/j.iree.2018.06.002
17. Piper K., Morphet J., Bonnamy J. Improving student-centered feedback through self-assessment. *Nurse Education Today*, 2019, vol. 83, pp. 104193. DOI: https://doi.org/10.1016/j.nedt.2019.08.011
18. Panadero E., Brown G. L., Strijbos J.-W. The future of student self-assessment: a review of known unknowns and potential directions. *Educational Psychology Review*, 2016, vol. 28 (4), pp. 803–830. DOI: https://doi.org/10.1007/s10648-015-9350-2
19. Sharma R., Amit J., Gupta N. Garg S., Batta M., Dhir S. Impact of self-assessment by students on their learning. *International Journal of Applied and Basic Medical Research*, 2016, vol. 6 (3), pp. 226. DOI: https://doi.org/10.4103/2229-516X.186961
20. Erkens M., Bodemer D. Improving collaborative learning: Guiding knowledge exchange through the provision of information about learning partners and learning contents. *Computers & Education*, 2019, vol. 128, pp. 452–472. DOI: https://doi.org/10.1016/j.compedu.2018.10.009
21. Liao C.-W., Chen C.-H. & Shih S.-J. The interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behavior patterns in a digital game-based learning environment. *Computers & Education*, 2019, vol. 133, pp. 43–55. DOI: https://doi.org/10.1016/j.compedu.2019.01.013
22. Hernández-Sellés N., Muñoz-Carril P.-C., González-Sanmamed M. Computer-supported collaborative learning: an analysis of the relationship between interaction, emotional support and online collaborative tools. *Computers & Education*, 2019, vol. 138, pp. 1–12. DOI: https://doi.org/10.1016/j.compedu.2019.04.012
23. Díaz-Ramírez J. Gamification in engineering education – An empirical assessment on learning and game performance. *Heliyon*, 2020, vol. 6 (9), pp. e04972. DOI: https://doi.org/10.1016/j.heliyon.2020.e04972
24. Zolkina A. V., Lomonosova N. V., Petrusevich D. A. Gamification as a tool of enhancing teaching and learning effectiveness in higher education: Needs analysis. *Science for Education Today*, 2020, vol. 10 (3), pp. 127–143. (In Russian) DOI: http://dx.doi.org/10.15293/2658-6762.2003.07
25. Landers R. N., Landers A. K. An empirical test of the theory of gamified learning. The effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming*, 2015, vol. 45 (6), pp. 769–785. DOI: http://dx.doi.org/10.1177/1046878114563662
26. Rastrollo-Guerrero J. L., Gómez-Pulido J. A., Durán-Domínguez A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, 2020, vol. 10 (3), pp. 1042. DOI: https://doi.org/10.3390/app10031042
27. Asif R., Merceron A., Ali S. A., Haider N. G. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 2017, vol. 113, pp. 177–194. DOI: https://doi.org/10.1016/j.compedu.2017.05.007
28. Fernandes E., Holanda M., Victorino M., Borges V., Carvalho R., Erven G. V. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 2019, vol. 94, pp. 335–343. DOI: https://doi.org/10.1016/j.jbusres.2018.02.012
29. Yang F., Li F. W. B. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, 2018, vol. 123, pp. 97–108. DOI: https://doi.org/10.1016/j.compedu.2018.04.006

30. Ripoll V., Godino-Ojer M., Calzada J. Teaching Chemical Engineering to Biotechnology students in the time of COVID-19: assessment of the adaptation to digitalization. *Education for Chemical Engineers*, 2020, vol. 34, pp. 94–105. DOI: https://doi.org/10.1016/j.ece.2020.11.005
31. Mishra L., Gupta T., Shree A. Online Teaching-Learning in Higher Education during Lockdown Period of COVID-19 Pandemic. *International Journal of Educational Research Open*, 2020, pp. 1000012. DOI: https://doi.org/10.1016/j.ijedro.2020.100012

Submitted: 10 September 2020 Accepted: 10 November 2020 Published: 31 December 2020