Factors Influencing Romanian Teachers' Choice of IT Training to Avoid Issues Raised by Online Education: A Data Mining Approach

Corina Simionescu
Doctoral School of Applied Sciences and Engineering
Computers and information technology
Ștefan cel Mare University
Suceava, Romania
corina.simionescu@usm.ro

Adina-Luminița Bărîlă
Faculty of Electrical Engineering and Computer Science
Ștefan cel Mare University
Suceava, Romania
mirela.danubianu@usm.ro

Mirela Danubianu
Faculty of Electrical Engineering and Computer Science
Ștefan cel Mare University
Suceava, Romania
adina.barila@usm.ro

Bogdănel Constantin Grădinaru
Doctoral School of Applied Sciences and Engineering
Computers and information technology
Ștefan cel Mare University
Suceava, Romania
bogdan.gradinaru@usm.ro

***Abstract*— Raising the students' academic performance and improving the overall quality of education is an important goal of our society. With the exploding development of information technologies, Educational Data Mining has become a growing activity that uses the essence of Data Mining concepts to help institutions learn useful information on various academic issues. Romanian schools are facing several problems and inequalities, many of which were accentuated during the Covid-19 pandemic, when, due to restrictions on physical proximity, classes were held entirely or partly online. This paper aims to presents an experimental study on modelling teachers' attitudes towards attending IT professional development courses to deliver quality online lessons. A real data set collected through a questionnaire answered by 956 teachers is used. The final goal was that, after the best-performing model was built, the features that are significant in this approach would be analyzed. These features are influential factors in teachers' choices. We initially considered four classification techniques, after which, we refined the optimization process for Random Forest, which proved, from the first experiments, the best performance.**

Keywords—educational data mining, classification, dimensionality reduction, process parameters optimization

I. INTRODUCTION

The general development that the IT field is constantly experiencing, and in particular e-learning, online learning, or distance learning, as the area of application of information technology in education, have given rise to the field of Educational Data Mining (EDM). It is described as a multidisciplinary field - combining techniques from statistics, artificial intelligence, machine learning, neuronal networks, database systems, data visualization, and aims to discover how people learn effectively and to identify aspects that can improve the whole educational processes by exploiting educational data sets using data mining techniques [1]. Beyond learning behavior or predicting student performances [2], it is interesting to find out which aspects can influence for the better the whole complex of resources in the educational system. When we talk about educational resources, we should necessarily think about the current paradigm of education that rests on three pillars: school, family, community. Moreover, mining economic data from education system can reveal aspects related to existing material issues or can recommend their better management.

Romanian schools are facing several problems and inequalities. If part of them is known or intuited, another part can be deduced by analyzing data collected from the system by different methods. One lesson can be considered the Covid-19 pandemic when, due to restrictions on physical proximity, classes were conducted entirely or partly online. During this time, we designed and distributed three questionnaires targeting teachers, students and parents involved in the pre-university education system in Romania, in order to explore their opinions on the quality of the online learning context and its usefulness [3].

This paper aims to presents a study on modelling teachers' attitudes towards attending IT professional development courses to deliver quality online lessons.

The performed experimental research allowed to find the best model describing this attitude and to discover and quantify those factors that determine it.

Our work presents the following novel elements:

- It uses real data collected from a questionnaire implemented in Google Forms completed by 956 teachers. The respondents' sample was chosen to reflect the reality of the system.
- Empirical research on the extent to which feature selection and automatic optimization of modeling process parameters have an impact on overall model performance (accuracy, or execution time) in order to implement a framework that provides an acceptable performance / resource consumption ratio.
- After careful analysis of the literature, we found that no studies on the factors influencing teachers' choice of IT training courses, in order to prevent possible problems raised by the need to teach online, have yet been carried out.

Further, the paper is structured as follows: Section II provides an overview of the related work, Section III focuses on the methodology and describes the dataset used, the preprocessing actions and the modelling process, Section IV emphasizes the presentation and interpretation of the obtained results. The paper concludes with Section V containing

979-8-3503-4929-0/24/$31.00 ©2024 IEEE

limitations, Section VI summarizing conclusions, and finally a list of references.

II. Related Work

At the international level, during and after the COVID-19 pandemic, many studies have been conducted on online education and its challenges. They have mainly addressed issues in higher education, referring to the directions in which EDM methods have been applied during the pandemic [2], to effective techniques for assessing student satisfaction [4][5]. In [6] twenty-seven factors were found to be involved in student dropout. There are also works that explore teachers' views, but these are focused on the impact that online education has on them [7], or to explore how they design online teaching activities and online teaching processes at all levels [8].

Although in Romania, interest in knowledge discovery from educational data has increased in recent years, efforts are still modest. Oprea presented in [9] how data mining methods can be used in education, and Haisan and Bresfelean proposed an analysis of teachers' views in the Romanian education system, concerning their opinion about the way how their financial needs are covered by income, using data mining methods [10]. In [11] a parallel between educational data mining and learning analytics is drawn. It is presented a case study where Google Analytics is employed on data collected from the activities of users of its own website https://www.modinfo.ro. In [12] a machine learning-based framework for enhancing the performance of decision-making processes is presented and a case study for educational data mining is made, and in [8] starting from unsupervised learning used for mining behavioral patterns from data, the Romanian baccalaureate exam is presented as a case study. There are other works by Romanian researchers [9] [10], which address various aspects concerning educational data mining, but all the mentioned works either cover algorithms, methods or data mining techniques using public datasets [13], refer to higher education [14], or address the professor's financial motivation. In [15], the authors employ unsupervised machine learning methods to identify behavioral patterns from data sets about high school students in Romania. Two main unsupervised learning techniques are utilized: Self-Organizing Maps (SOMs) and Association Rules (AR). These techniques were applied to analyze the preferences of students in choosing optional subjects for the baccalaureate exam.

The literature lacks studies that explore online learning through other aspects, such as material or contextual factors that may influence its acceptance, the quality of online teaching or the level of outcomes. This paper is a step in this sense.

III. Working Metodology

A. Dataset description

As we mentioned above, we designed three questionnaires that targeted teachers, students and parents involved in the pre-university education system in Romania. This paper addresses the dataset resulted from 956 teachers' responses. The structure of sample respondents is presented in Fig.1.

The questionnaire contains 24 items with answers of various forms - from binomial and polynomial, to integer or free text answers, as shown in TABLE 1. The 24 questions were designed around several main axes: demographic and professional information: teaching position, teaching seniority, teaching grade, subject taught, status, gender and location of school unit, skills and use of technology, perceived advantages and disadvantages of online education, appreciation of the quality of online lessons and students' interest in these lessons, sustainability of online education as an alternative or complement to traditional education, the need for a legislative framework for online education and the need for continuing IT training.

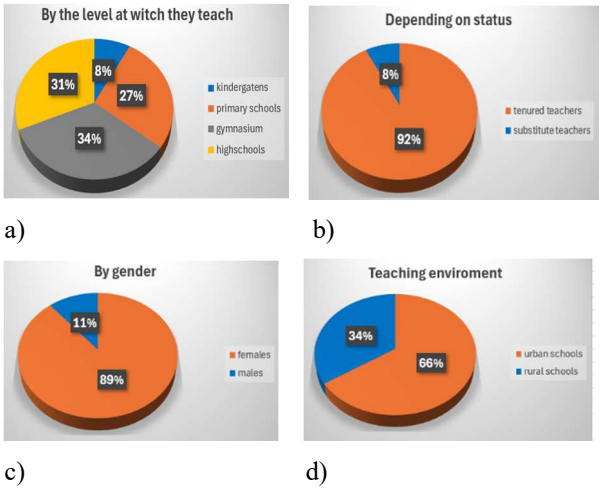


Fig. 1 Structure of respondents sample: a) by the level at which they teach, b) depending on status (tenured/substitute), c) by gender, d) by teaching environment

The questionnaire concludes with open-ended questions on the problems encountered in conducting online lessons, suggestions for improving online education and opinions on preparing the Romanian education system for alternative education.

TABLE I. DESCRIPTION OF QUESTIONAIRE FOR TEACHERS

Question	Type of answer	No. of distinct values
1. Teaching role	polynomial	7
2. Years of Teaching Experience	polynomial	6
3. Teaching Grade	polynomial	5
4. Please specify the subject(s) you teach	polynomial	17
5. Status	binomial	2
6. Please specify your gender:	binomial	2
7. The school where you have your main work is located in:	binomial	2
8. Evaluate your digital skills:	integer	10
9. Have you conduct online activities with your students?	binomial	2
10. What digital tools have you used for conducting online activities with students?	polynomial	9
11. What teaching materials have you use in conducting online lessons with students?	polynomial	6
12. What is the source of the teaching materials used?	polynomial	5
13. What are the advantages of online education?	polynomial	6
14. What are the disadvantages of online education?	polynomial	6

Question	Type of answer	No. of distinct values
15 How do you rate the online applications used?	integer	5
16. How do you assess the quality of online lessons in all aspects (interesting materials, engagement, message delivered and received, etc)?	integer	10
17. How do you rate the students' interest in online lessons?	integer	10
18. Can online school replace traditional school?	polynomial	3
19. Do you consider it necessary to have a legislative framework regulating online education?	binomial	2
20. Would participating in continuous IT training courses be helpful for better managing the software and hardware resources required by the digital school?	polynomial	3
21. If you are also a parent of school-age child/children, how do you appreciate the quality of the online classes conducted by your child/children?	integer	10
22. Please specify the issues you encountered in conducting online lessons	text	
23. Provide suggestions for future improvements of online education.	text	
24. Do you think the Romanian educational system is currently prepared to conduct an alternative form of education (such as online)?	polynomial	3

We proposed to analyze, through data mining techniques teachers' attitudes towards participating in IT continuous training courses, in order to deliver quality online lessons. We intend to build accurate models which could reveal the reasons behind their option. These reasons can be further analyzed as factor which can be used to positively influence the system. For this purpose, we considered question number 20 ("Would participation in continuing IT training courses help you to better manage the software and hardware resources needed by the digital school?") as class label.

A dataset analyze shows a strong imbalance between classes, as it is presented in Fig. 2. As a result, we did experimental research on how balancing classes by upsampling or undersampling affects the performance of the created models

B. *Data preprocessing*

Data preprocessing is a mandatory stage, and its purpose is to put the data into a form suitable for modelling.

To this end, the following operations were carried out:

- checked that all fields are filled in;
- we made changes in fields containing redundant data (e.g., we classified teachers with grade I checked that they also had doctoral studies in the category with the highest teaching grade);
- for questions answered on a scale of 1 - 10 we replaced the numbers with grades, thus:10 became E (excellent), 8-9 - FB (very good), 7 - B (good), 1-6 S (sufficient);

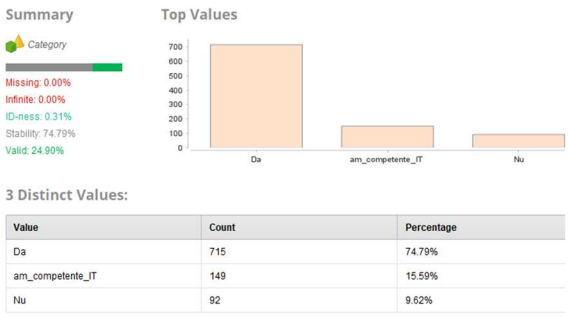


Fig. 2 Distribution of original dataset class values

- for questions with answer's values between 1 and 5 we replaced the score with qualifiers as follows: 5 became E (excellent), 4 - 3 FB (very good), 2 - B (good), 1 S (sufficient);
- to the question *Can online school replace traditional school?* for which the set answers were "*yes", "no",* and *"don't know",* we replaced "*don't know"* with "*abstain"*.
- we simplified the wording of the questions in the questionnaire (which we used in the processing as labels or attributes) by suggestive abbreviations, e.g., the question "*Do you value the digital skills you have"* was replaced by "*Level_comp_IT"*.

Additionally, using the appropriate operators provided by the RapidMiner Studio 10 environment, the highly correlated features were removed, and attributes with very many distinct values (which do not have a significant weight in the induction of classification models) as well as attributes with an increased weight of the same value were also removed.

As presented in Fig.2. class values are strongly unbalanced. As we know that this bias may affect the model, we proposed two methods of balancing them: by undersampling for the cases with the labels „Da" ("Yes") and „am_competente_IT" („I am_competent_IT") and by upsampling for those with the labels "Nu" („No") and „am_competente_IT" („I am_competent_IT").

The process for undersampling is presented in Fig. 3. We checked the parameter *balance data* of the *Sample* operator and reduced the number of cases for each of the three targeted classes to 92. To balance the data by upsampling we used the *SMOTE Upsampling* operator which uses the synthetic minority oversampling technique [16]. The process is presented in Fig. 4.

After running the two processes, two balanced datasets resulted, one with 276 cases, the other with 2145 examples.

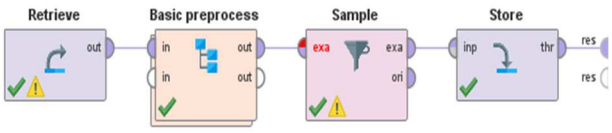
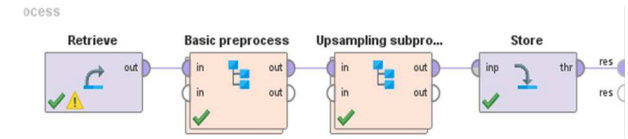
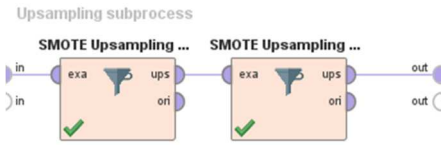


Fig. 3 Data balance through undersampling



a)



b)

Fig. 4 Data balance through upsampling: a) the whole process, b) the upsampling subprocess components

C. Data modelling

We used classification, and we considered in the first stage the following classifiers: Naïve Bayes, Decision Tree, Random Forest and SVM. We executed these processes on a system ASUS10, 12th Gen Intel(R) Core (TM) i7-12700H 2.30 GHz, in RapidMiner Studio 10.3., using the original data set and we obtained the performance values from Table II.

TABLE II. PROCESS PERFORMANCES ON ORIGINAL DATASET

Used technique	Accuracy	Runtimes [s]
Nayve Bayes	73.6%	10
Decision Tree	79.5%	9
Random Forest	**79.9%**	16
Support Vector Machine	78.8%	37

Running under the same conditions the sets obtained by the two balancing methods, we obtained the results presented in Table III and Table IV.

TABLE III. PROCESS PERFORMANCES ON UPSAMPLING DATASET

Used technique	Accuracy	Runtimes [s]
Nayve Bayes	57.7%	7
Decision Tree	62.1%	8
Random Forest	**74.7%**	2min 16 sec
Support Vector Machine	72.6%	1 min 45 sec

TABLE IV. PROCESS PERFORMANCES ON UNDERSAMPLING DATASET

Used technique	Accuracy	Runtimes [s]
Nayve Bayes	48.7%	10
Decision Tree	20.5%	9
Random Forest	34.9%	16
Support Vector Machine	**50.6%**	37

It is easy to see that for both the original dataset and the upsampled dataset, the performance is considerably better than that obtained on the undersampled balanced dataset. However, the best results do not exceed the 79.9%. threshold for accuracy, and, paradoxically, the best values are obtained on the unbalanced data set. Also, analyzing the above tables we found that the best values were obtained for Random Forest modeling. As a result, we further considered possibilities to increase the efficiency of these models, so we assessed different combinations of parameter values, but the performance did not significantly change.

According to the most used model for data mining processes, - CRISP-DM [17] these processes are highly interactive and iterative. Since in the first two iterations the results can be considered satisfactory, but not very good, we continued the experimental study in a new guideline, considering only the cases of original and upsampled datasets and the Random Forest mining technique. We addressed a framework that covered two dimensionality reduction methods. First, we performed a feature selection using wrapper methods that evaluate all possible combinations of features and select the one that produces the best results for the targeted machine learning algorithm. We considered the two possible scenarios: forward selection, and backward elimination. Through this operation we aimed to use the best combination of features, the ones that really participate in model building.

TABLE V. OUTCOMES OF FEATURE SELECTION USING WRAPPERS

Dataset	Selection method	No. of attributes (predictors) selected	Model accuracy
Original	Forward selection	**4**	**88.05 %**
	Backward elimination	23	88.05 %
Upsampled	Forward selection	10	87.20%
	Backward elimination	**14**	**87.69%**

Secondly, with the intention of identifying the best predictors, we used the method of eliminating useless attributes, that is, those attributes that cross the uselessness threshold set by parameters suitable for each type of data. The results obtained are presented in Table VI.

TABLE VI. OUTCOMES OF USELESS FEATURE REMOVAL

Dataset	Uselessness threshold	No. of attributes (predictors) selected	Model accuracy
Original	0.1	22	88.11%
	0.2	21	88.11%
	0.3	**17**	**88.11%**
	0.4	15	87.86%
	0.5	12	87.74%
Upsampled	0.1	16	85.8%
	0.2	15	85.8%
	0.3	15	85.8%
	0.4	13	80.14%
	0.5	13	80.14%

The last step that we considered as a possibility to increase the quality of the model is the use of the automatic optimization operator of the data mining process parameters.

This optimization process is based on the data sets (original and usampled) that, in the previous step, provided the best values for accuracy.

The optimization process uses an iterative execution of the modeling process using all combinations of parameters.

Briefly, the achieved results consisting in the optimal parameters, both for original and balanced datasets, are presented in Table VII.

TABLE VII. THE OPTIMAL PROCESS PARAMETERS

Data set	No. of. trees	Splitting criterion	Pre-prunning	Prun-ning	Acc
Original	**21**	**Gain_ ratio**	**T**	**F**	**89.4%**
Upsampl.	70	Gain_ ratio	F	F	86.9%

IV. DISCUSSION

The research has produced a surprising result. In all the scenarios studied the performance of the models built on the original data set was superior to those built on the balanced datasets.

If the patterns built on the undersampled set had only 276 training cases, which is a very small number, and thus the results are explainable, the same cannot be said for those based on the upsampled set.

In this case, the reason could be related to the method used for upsampling. This involves injecting into the original set synthetically generated data corresponding to the minority classes. Although this procedure prevents the model from being biased towards the majority class, on our dataset this came with a decrease in accuracy.

Another observation is related to the fact that a significant improvement in model accuracy, around 10%, was obtained by feature selection. This involves reducing the initial set to a minimum cardinality. This new data set must contain the most relevant attributes for the modeling purpose. Finally, the process parameters optimization also led to an increase in performance, but this was at a level of about 1%.

Since the best model consists of a combination of 21 classification trees, most of which have at least 6 levels and a significant number of leaves, to explain the predictions we considered the weights of the predictors in the model development. These are shown in Fig. 5.

The main factors influencing teachers' desire to improve their online teaching by enhancing their IT skills are: „The necessity of a regulated legal framework for online learning", „Environment", „Level", „Online_vs_traditional", and „Level_comp_it".

V. LIMITATION

Our study utilizes a real dataset collected online and anonymously from teachers in Romania.

An important limitation may be that the data is collected at a single point in time, and we built the model for that moment. To observe long-term trends or changes in teachers' behavior or attitudes, we will conduct a comparative analysis in future works on this data set with a dataset we collected in February 2024.

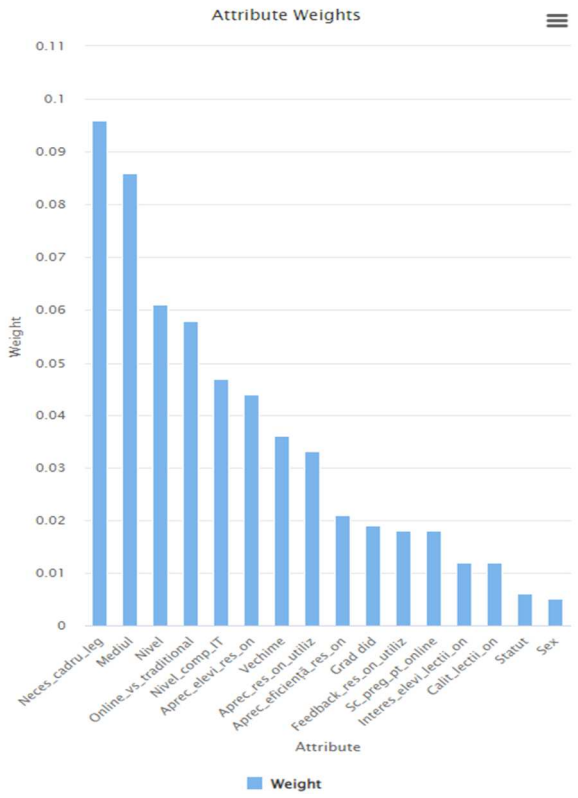


Fig. 5 The predictors' weights in the best classification model

VI. CONCLUSIONS

It is no longer a novelty that EDM is an extremely useful tool, especially now, in the information society, when education generates impressive volumes of data to explore.

However, it should be noted, that behavioral learning patterns or analysis of the educational stakeholders' sentiments are not the only factors leading to quality improvement in the field. Seen from the right perspective, educational data can provide information on other, less visible, but important aspects in the context of the complete system.

An example might be the management of educational resources, from the material to the human ones. Our research falls into this area by targeting the influencing factors of options that obviously lead to an increase in the quality of human resources.

As technology evolves rapidly, it is crucial for teachers to be continuously trained to be able to teach using new technologies and online teaching methods. This requires constant funding and support from the state.

Following, we will address the most important of the influencing factors found in our model.

Developing a robust legal framework for IT training programs and online education in Romania has significant political implications, which could influence the educational system and the job market in the long term.

Creating a legal framework for online education and IT programs should ensure equal access for teachers to

educational resources, regardless of geographic location or socioeconomic status. This involves government investments in digital infrastructure, particularly in rural or disadvantaged areas, to reduce disparities in access to education. With the increasing number of online educational platforms, it is essential to have clear regulations regarding the protection of personal data of students and teachers. Policies must comply with GDPR standards to ensure the security and confidentiality of information.

We believe that promoting IT education and digital skills is essential, generally, to prepare the workforce for the digital economy. A solid legal framework in this regard could attract more foreign investment, promoting Romania as a regional technology hub. The government could encourage partnerships between universities, private companies, and research institutes to foster innovation in digital education and IT. These collaborations could lead to the development of new technologies and pedagogical methods that position Romania as a leader in digital education.

By addressing these aspects in the legislative framework, politicians can play a crucial role in shaping Romania's future educational and economic landscape, turning challenges into opportunities for all citizens. This initiative requires close collaboration between various government agencies, the private sector, and academic communities, emphasizing the importance of a holistic and well-coordinated approach.

REFERENCES

[1] Ryan S. Baker, "Educational Data Mining: An Advance for Intelligent Systems in Education", Columbia University, Editor: Judy Kay, University of Sydney, may 2014

[2] Romero, C., Ventura, S., "Educational data mining: a review of the state of the art". IEEE Trans. Syst., Man, Cybern., Part C. (Appl. Rev.) 40 (6), 601–618, 2010.

[3] Simionescu, C., Danubianu, M., & Marcu, D. "Analysis of online education romanian schools due to Covid-19 pandemics and areas of improvement". In ICERI2020 Proceedings (pp. 3523-3529). IATED 2020

[4] Dasari, H. R., & Prasad, G. N. R. "An Effective Data Mining Method for Determining Higher Education Students' Satisfaction with Online Learning During the COVID-19", IJIRT, 10(10), 2024

[5] Abdelkader, H. E., Gad, A. G., Abohany, A. A., & Sorour, S. E. An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19. IEEE Access, 10, 6286-6303, 2022

[6] A. Urbina-Nájera, J. Camino-Hampshrire, and R. C. Barboza, "University dropout:Prevention patterns through the application of educational data mining", Journal of Educational Research, Assessment and Evaluation, vol. 26, no. 1, pp. 1–19, 2020. https://doi.org/10.7203/relieve.26.1.16061

[7] Nambiar, D. "The impact of online learning during COVID-19: students' and teachers' perspective". The international journal of Indian psychology, 8(2), 783-793, 2020

[8] Wu, S. Y., "How teachers conduct online teaching during the COVID-19 pandemic: A case study of Taiwan". In Frontiers in Education (Vol. 6, p. 675434). Frontiers Media SA., 2021

[9] Oprea, C., & Zaharia, M., "Using data mining methods in knowledge management in educational field". Fascicle of Management and Technological Engineering, 10, 2011

[10] Hăisan, A. A., & Bresfelean, V. P. "A data mining examination on the Romanian Educational System-teachers' viewpoint". Int J Mathematical Models Methods Appl Sci, 7(3), 277-285, 2013

[11] Marcu, D., & Danubianu, M., "Learning Analytics or Educational Data Mining? This is the Question". BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 10((Special Issue 2), 1-14., 2019 https://lumenpublishing.com/journals/index.php/brain/article/view/2388

[12] G. Czibula, G. Ciubotariu, M. -I. Maier and H. Lisei, „IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining," in IEEE Access, vol. 10, pp. 80651-80666, 2022, doi: 10.1109/ACCESS.2022.3195531.

[13] Vultureanu-Albişi, A., & Bădică, C., Improving students' performance by interpretable explanations using ensemble tree-based approaches. In 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 215-220. IEEE., 2021

[14] Crivei, L. M., Ionescu, V. S., & Czibula, G. "An analysis of supervised learning methods for predicting students' performance in academic environments". ICIC Express Lett., 13(3), 181-189 2019

[15] Mariana-Ioana MAIER, Gabriela CZIBULA, Lavinia-Ruth DELEAN, "Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania", Studies in Informatics and Control, ISSN 1220-1766, vol. 32(2), pp. 73-84, 2023. https://doi.org/10.24846/v32i2y202307

[16] N. V. Chawla, K. W. Bowyer,L. O. Hall,W. P. Kegelmeyer , "SMOTE: Synthetic Minority Over-sampling Technique", Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002, https://doi.org/10.1613/jair.953.

[17] Wirth, R., & Hipp, J., "CRISP-DM: Towards a standard process model for data mining". In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,Vol. 1, pp. 29-39, 2000