An advanced explainable and interpretable ML-based framework for educational data mining

Conference Paper · June 2023

CITATIONS
2

READS
197

4 authors:

Ioannis Livieris
University of Patras
104 PUBLICATIONS **2,405** CITATIONS
SEE PROFILE

Nikos Karacapilidis
University of Patras
255 PUBLICATIONS **3,684** CITATIONS
SEE PROFILE

George Domalis
University of Patras
7 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Dimitris Tsakalidis
Novelcore
12 PUBLICATIONS **46** CITATIONS
SEE PROFILE

An advanced explainable and interpretable ML-based framework for educational data mining

Ioannis E. Livieris[1] [0000−0002−3996−3301], Nikos Karacapilidis[2] [0000−0002−6581−6831], Georgios Domalis[1] [0000−0002−3449−2029], and Dimitris Tsakalidis[1] [0000−0001−7185−102X]

[1] Novelcore, GR 10436, Greece
{livieris,domalis,tsakalidis}@novelcore.eu
[2] Industrial Management and Information Systems Lab, University of Patras, 26504, Greece
karacap@upatras.gr

Abstract. During the last two decades, the adoption of machine learning techniques for addressing various challenging issues in the educational domain has gained much popularity. Nevertheless, there is still a lack of research on developing AI systems that focus on the interpretability and explainability of the associated models and algorithms, thus being able to present the data analysis results in a human understandable way. In this work, we propose a new explainable framework for predicting students' performance, which provides accurate, reliable and interpretable results. Our framework builds on the recently proposed NGBoost algorithm for the development of an efficient prediction model, as well as on the LIME and SHAP methods for providing local and global explanations, respectively. The use cases presented in this paper demonstrate the applicability of our framework and give insights about the recommendations that can be provided to educators and students.

Keywords: Educational data mining, machine learning, recommendation, explainability, NGBoost, LIME, SHAP

1 Introduction

Educational Data Mining (EDM) is a field of research that builds on Machine Learning (ML) and other data analysis techniques to analyze educational data for gaining insights into the learning process. Its primary goal is to identify patterns, relationships and factors that affect the outcomes of the learning process, aiming to predict the learners' future performance and accordingly use this information to improve educational practices and policies. EDM has been already applied in a wide range of educational contexts and has a great potential to enhance student learning, teacher effectiveness, as well as the overall institutional performance [7, 8, 14, 16, 18].

In recent years, there is an increasing trend in EDM to provide meaningful explanations to educators, administrators and students about how the underlying ML models arrive at their predictions or decisions. Explainability is considered important for a variety of reasons. Firstly, as EDM models become more complex and powerful, it is difficult to understand how they make decisions or predictions. By ensuring that these ML models are transparent and explainable, educational stakeholders are able to better

understand how they work and have more confidence in their results, thus leading to more informed decision-making. Secondly, explainability is important for ethical reasons; as ML models are used more frequently in education, it is of major importance to ensure that they are fair and unbiased, and that they do not perpetuate or amplify existing biases in the associated data.

In the context under consideration, feature importance methods have been broadly used to measure the contribution of each feature on an EDM model's predictions. Local Interpretable Model-agnostic Explanations [17] (LIME) and SHapley Additive exPlanations (SHAP) [15] are two permutation-based model-agnostic techniques, which probably constitute the most widely used feature importance methods [7, 8, 10, 16]. Both of them are able to randomly sample from the marginal distribution considering unrealistic instances that are not present in the training data. In simple words, they focus on extrapolating in the areas where the model was trained for measuring each features effects on the predictions. The major difference between these methods is that the former is based on a simple model for creating a local explanation around a prediction, while the latter is based on game theory to measure the magnitude of feature attributions.

This work aims to contribute to the EDM field by proposing an advanced framework for predicting students' performance, which ensures the provision of accurate, reliable and explainable predictions. Our primary goal is to develop an efficient model for predicting students' performance and simultaneously provide human-interpretable explanations of individual predictions, which in turn give insights into how the model is making its decisions. Our prediction model is based on one of the most efficient ML algorithms, namely Natural Gradient Boosting for Probabilistic Prediction (NGBoost); as demonstrated in our experiments, NGBoost outperforms traditional state-of-the-art prediction algorithms.

An attractive advantage of the proposed approach is that it is able to provide both local and global explainability, providing users with a complete picture as well as with human-interpretable insights about why a particular decision was made by the model. In simple words, it is able to provide a description of the model's mechanisms and concepts, while simultaneously being able to explain each individual prediction made. This is achieved through the utilization of both the SHAP and the LIME methods, which ensure that the importance scores are fair and unbiased as well as a flexible, fast and reliable interpretation of each single prediction. Through two representative use cases, we also showcase the applicability of the proposed framework and the possible recommendations that could be provided to educators and students.

The remainder of this paper is as follows: Section 2 presents a brief survey of recent studies concerning the application of ML models for predicting the students' performance. Section 3 describes in detail the proposed framework, while Section 4 comments on the data used in this research. Section 5 reports on our experimental results and describes the two cases used for the application of the proposed framework. Finally, Section 6 sketches concluding remarks and proposes future research directions.

2 Related work

During the last two decades, the adoption of ML techniques for providing useful insights about the learning process and students' behavior has gained much popularity. An interesting application concerns the successful prediction of students' performance, especially of students that are at risk of failing, which has a significant impact to diverse educational stakeholders including teachers and students, but also the educational institute. In this context, the provision of accurate and explainable predictions is essential for effectively conducting the necessary pedagogical interventions to enhance students' performance. A number of interesting related studies have been already proposed in recent years, whose findings and limitations are outlined below.

Tampakas et al. [18] proposed a two-level classification scheme for predicting students' graduation time within the first two years of their studies. The proposed scheme has two major features: (i) identification of students which are likely to fail to graduate, and (ii) prediction of students' expected graduation time. The presented numerical experiments showed the superiority of the proposed approach compared to the traditional ones, and reveal that it is possible to accurately predict students' graduation time within the first two years of their studies. However, this approach does not provide any explainable feedback about its predictions.

Hue et al. [8] implemented personalized system intervention using a ML model to predict student performance and explained its predictions via a SHAP method. The proposed approach was evaluated in a self-paced, self-guided online learning system for college-level topics, which provided personalized interventions to enhance a learning behavior. A randomized controlled trial of 37 expert-system condition and 36 explanation condition participants showed that similar learning and topic-choosing behavior between conditions. Based on their findings, the authors stated that XAI-informed interventions facilitated student learning to a similar degree as expert-system interventions. However, a limitation of this work is the relatively small pilot and experimentation data size.

Ramaswami et al. [16] presented a generic predictive model for identifying students at risk across a variety of courses in a blended learning environment. The numerical experiments included the evaluation of several state-of-the-art ML algorithm and showed that the CatBoost algorithm demonstrates the best overall performance. In this work, the authors used the SHAP method to estimate model behaviour without providing any recommendations or any explainable framework for assisting the education process.

Guleria and Sood [7] proposed a new framework for students' career counseling based on ML and AI techniques. White and black box models were trained on an educational dataset for predicting future students' placement status (binary classification task). The numerical experiments include the performance evaluation of several state-of-the-art white and black box models. Additionally, the authors provided some global explainability insights using the SHAP method. Two limitations of this work are that the used dataset contains a limited scope of attributes and sample size and the fact that the proposed framework does not provide any local explainability feedback.

3 The proposed framework

Aiming to overcome the above mentioned limitations, this work proposes an advanced framework for predicting students' performance, which provides accurate, reliable and explainable predictions. The proposed framework builds on the recently proposed NGBoost algorithm for the development of an efficient prediction model. Contrary to previous works, our approach is able to provide both local and global explainability feedback about the predictions of the NGBoost model by exploting the LIME and SHAP importance methods, respectively.

The proposed framework consists of two main components: the prediction model and the explainability modules. The former is used for predicting the students' performance, while the latter for providing the reasoning behind the decisions and individual predictions of the model. Figure 1 provides a high-level overview of the proposed framework. Initially, the training data are preprocessed (invalid values removal, one-hot encoding of categorical variables and data imputation using 3NN) and transformed into a suitable form to be used as input to the ML algorithm for developing an efficient prediction model. Our goal is to develop a classifier with strong classification ability. For this reason, we selected the recently proposed NGBoost algorithm [4]. As soon as the prediction model is developed, it can be used for conducting predictions on new data; simultaneously, its predictions, along with the LIME and SHAP methods, are used for providing local and global explainability, respectively.

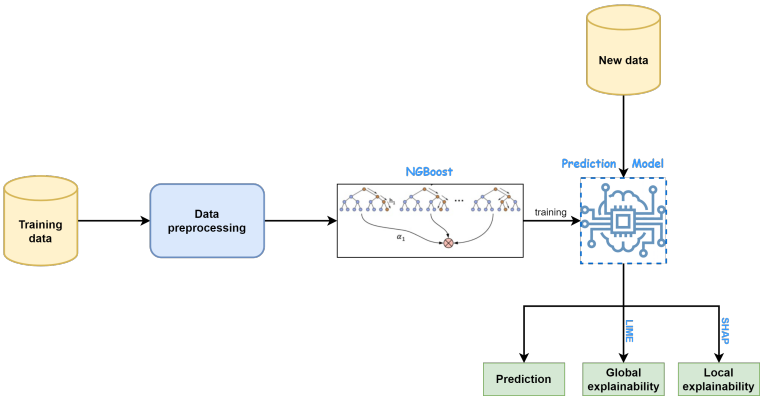


Fig. 1. The proposed framework

It is worth mentioning that the reason for selecting LIME for local explainability is that it probably constitutes the most flexible and widely utilized method for interpreting a single prediction [2]. Additionally, the reason for preferring the SHAP method for global explainability over NGBoost's feature importance is that feature importance can be biased or unstable and is heavily depending on the specific method for its calculation [1]. On the other hand, SHAP is based on game theory, ensuring that the importance scores are fair and unbiased [15]. In the following, we briefly present the main tech-

nologies on which the proposed framework is based, i.e. the NGBoost classifier, as well as the LIME and SHAP explainability methods.

3.1 NGBoost algorithm

NGBoost (Natural Gradient Boosting) [4] is a probabilistic boosting algorithm that can be used for classification and regression tasks. Its key innovation is the employment of natural gradient for performing gradient boosting by casting it as a problem of determining the parameters of a probability distribution. It is noted that ordinary gradients may be unsuitable for learning multi-parameter probability distributions (such as the normal distribution). On the other hand, the natural gradients with the use of training dynamics tends to be much more stable and robust, something that results in a better fitting process. The NGBoost algorithm consists of three components: (i) *base learners*, which uses weak ML models for making predictions of the input data and form the conditional probability, (ii) *parametric probability distribution*, which constitutes a conditional distribution and it is formed by an additive combination of base learners outputs, and (iii) *scoring rule*, which measures the quality of its probabilistic predictions and optimizes the ensemble-based model. In our experiments, NGBoost was implemented with decision trees as base learners for making the predictions of the input data and Negative Log Likelihood (NLL) as scoring rule.

3.2 Explainability methods

LIME (Local Interpretable Model-agnostic Explanations) [17] constitutes a model-agnostic and local interpretability method for explaining the predictions of any ML model. The rationale behind LIME is rooted in the need for transparency and interpretability in ML models, which is fundamental in many challenging real-world applications [2]. LIME has the advantage that it does not rely on any assumptions about the underlying model architecture or training process. In addition, it is highly transparent, which makes it easier to understand how the feature importances are calculated and how they contribute to the final prediction, flexible and computationally efficient in providing explanations. However, a limitations of LIME is that it does not provide a global view of the model's behavior or insights into how the model makes decisions across the entire dataset. Therefore, it should be used in conjunction with other methods for model interpretability and transparency.

SHAP (SHapley Additive exPlanations) [15] is a state-of-the-art explainability method [2], which measures the contribution of each data point to each feature value based on cooperative game theory (Shapley values). By doing so, it is able to deliver fine grain explanations and calculate the global feature contributions, including their direction. The key idea behind this method is that each feature's contribution is the Shapley value, which provides information about the model's performance if it was trained without that feature. However, a significant drawback of calculating the SHAP values is the computational cost, since the training time grows exponentially with the number of features.

4 Dataset

For the purpose of this work, we utilized two real-world educational datasets. Specifically, DATASET$_1$ includes information about 337 university students who attended an academic course using a Learning Content Management System (LMS) in a blended learning environment. This dataset was collected from the Department of Educational Sciences and Early Childhood Education, University of Patras, Greece, during the academic years 2007-2010. Each instance contains information related to the students' perceptions about the Moodle LMS and their opinions about its educational value and usefulness, as well as information related to the students' activity. Demographic values were not included due to the similar characteristics of all participants. In this dataset, students were classified upon whether they passed the lesson ("Pass") or not ("Fail"). More information about this dataset can be found in [5].

DATASET$_2$ contains data about 3716 students that attended courses of Mathematics of a secondary school (namely, the "Avgoulea-Linardatou" Microsoft Showcase School, Greece; data concern the time period 2007-2016). This dataset summarizes information about the students' performance from the first two out of three semesters such as tests grades, final examination grades oral grades, as well as semester grades. Note that an academic year of a secondary school in Greece consists of three semesters. The students were classified according to their performance upon a four-level classification scheme, i.e. "Fail", "Good", "Very Good" and "Excellent" (more information is provided in [12]). [3].

5 Experiments and Use Cases

In this section, we report on the evaluation of the performance of the NGBoost algorithm that was adopted in the proposed framework against state-of-the-art ML classification algorithms. In addition, we present the explanations produced by our framework for two representative use cases.

5.1 Experimental Results

We have evaluated the performance of the NGBoost algorithm on two educational datasets against that of three state-of-the-art ML algorithms, namely Random-Forest[11], XGBoost [3] and LGBM [9]. Admittedly, these algorithms constitute the most efficient ones for handling classification tasks with tabular data [6]. In our original experiments, we included several algorithms such as neural networks, support vector machines and k-nearest neighbors; nevertheless, their performance was inferior to that of ensemble tree-based algorithms.

The classification performance of all algorithms was evaluated using stratified 10-fold cross-validation and the following performance metrics: *Accuracy, Area under*

[3] It is noted that a detailed description of the features of both datasets, as well as their descriptive statistics and a complete exploratory data analysis, can be found in https://github.com/novelcore/A-new-explainable-and-interpretable-ML-based-framework-for-educational-data-mining

Curve (AUC), Geometric Mean (GM), Precision and *Recall*. It is worth mentioning that the performance metrics AUC and GM, as well as the balance between Precision and Recall, present the information provided by a confusion matrix in compact form; therefore, they constitute the proper metrics to evaluate the classification ability of a prediction model [13]. The implementation code was written in Python 3.8, while the hyperparameters of all algorithms were optimized in order to achieve the best possible performance [4].

Table 1 summarizes the performance of all classification algorithms on DATASET$_1$. As shown, NGBoost demonstrates the best overall classification performance for all performance metrics. Specifically, NGBoost reported the best Accuracy (71.2%), followed by Random-Forest (70.8%) and XGBoost (66.6%). Additionally, it reported the highest AUC (0.649) and GM (8.221) scores, while Random-Forest presented the second best performance. In contrast, LGBM was unable to distinguish between noise and rare cases, which resulted in poor performance.

Table 2 presents the performance of all classification algorithms on DATASET$_2$. NGBoost reported the highest GM metric (45.24) and the best balance between Precision and Recall, which suggests that it presents the best overall performance. Additionally, XGBoost and Random-Forest reported the second best performance, while LGBM reported the worst performance for all performance metrics. From the above results, we can conclude that the NGBoost algorithm is able to develop the most accurate prediction model for both datasets used in this work.

Algorithm	Accuracy	AUC	GM	Precision	Recall
LGBM	60.4%	0.633	8.034	0.581	0.633
Random-Forest	70.8%	0.641	8.094	0.605	0.641
XGBoost	66.6%	0.640	8.156	0.592	0.640
NGBoost	71.2%	0.649	8.221	0.608	0.648

Table 1. Performance evaluation of classification algorithms on DATASET$_1$

Algorithm	Accuracy	AUC	GM	Precision	Recall
LGBM	86.9%	0.911	44.01	0.862	0.867
Random-Forest	89.4%	0.923	44.99	0.906	0.884
XGBoost	89.7%	0.924	45	0.905	0.885
NGBoost	90.0%	0.927	45.24	0.897	0.927

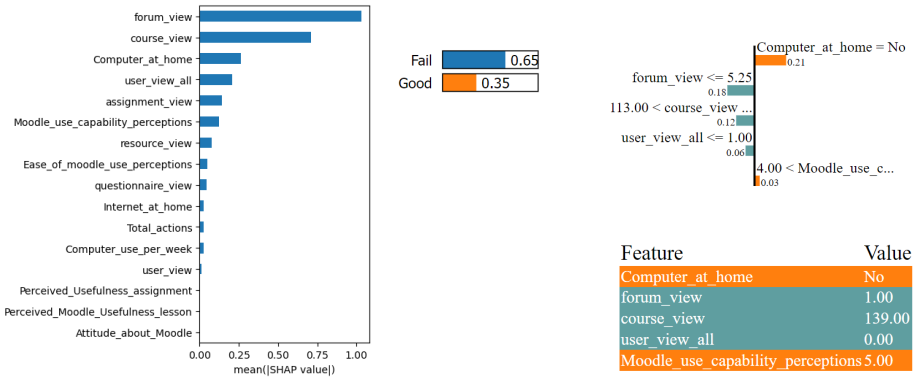
Table 2. Performance evaluation of classification algorithms on DATASET$_2$

5.2 Use cases

This subsection reports on the application of the proposed framework on two use cases corresponding to DATASET$_1$ and DATASET$_2$, aiming to demonstrate the usefulness of the feedback provided to educators and students. Figure 2 illustrates the output from the application of the proposed framework on the problem of identifying the university students who are at-risk of failing the examinations (DATASET$_1$). The interpretation of Fig-

[4] Additional information can be found in at https://github.com/novelcore/A-new-explainable-and-interpretable-ML-based-framework-for-educational-data-mining

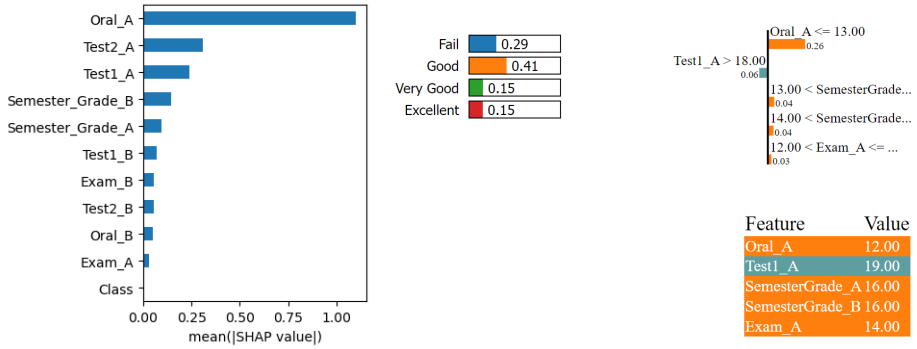
ure 2(a) suggests that the three most important features for predicting if a university student will fail in the examinations are *forum_view, course_view* and *Computer_at_home*, while the features *Perceived_Moodle_Usefulness, Perceived_Usefulness_assignment* and *Attitude_about_Moodle* seem to have no effect in the model's decisions. This implies that an educator should provide special attention to the number of times each student accesses the description and the basic material of each week's laboratory session, accesses the forum section and if he/she owes a computer. In addition, the feedback for a student could be that he/she pays more attention to the basic material of each week's laboratory. Local explainability focuses on understanding how the model made decisions for a single instance. For a specific student, the model predict that he/she will fail to the final course grade with probability 65% and the fact that this student has no computer at home has the highest impact on the model's decision.



(a) Global explanations (b) Local explanations

Fig. 2. Application of the proposed framework on DATASET$_1$

Figure 3 presents the output from the application of the proposed framework on the problem of predicting the performance of high-school students at the final examinations (DATASET$_2$). The global explainability component shows that the three features that influence most the model are *Oral_A, Test2_A* and *Test1_A*, while the three least important ones are *Oral_B, Exam_A* and *Class*. A possible recommendation to the educator could be that the students' performance on the first semesters, especially their performance on oral examinations and on 15-minutes tests, seem to considerably affect the students' performance on the final examination. The interpretation of Figure 3(b) suggests that for a specific student, the model predicts that he/she will have "good" performance on the final examinations, while the fact that his/her oral grade on the first semester is less than 13 (i.e. *Oral_A* = 12) has the highest impact on the model's decision. Therefore, a recommendation that could be given to the student (as well as to the educator) is to improve his/her oral argumentation skills.



(a) Global explanations (b) Local explanations

Fig. 3. Application of the proposed framework on DATASET$_2$

6 Conclusions

This work proposes an advanced explainable framework for predicting students' performance, which provides accurate, reliable and interpretable predictions. The proposed framework is based on the recently developed NGBoost algorithm, as well as on the LIME and SHAP importance methods for providing local and global explainability, respectively. Our numerical experiments showed that NGBoost is able to outperform traditional state-of-the-art ML algorithms, hence providing empirical evidence that its adoption could lead to the development of an accurate prediction model. The adoption of SHAP ensures that the importance scores are fair and unbiased, while the adoption of LIME provides a flexible, fast and reliable technique for interpreting a single prediction.

A limitation of this work is that the proposed framework was evaluated only on two real-world datasets, containing a limited number of attributes with the scope of predicting the students' performance on the examinations. A future work direction is to consider and elaborate additional educational datasets, and accordingly evaluate the proposed framework across diverse challenging issues in the educational domain. Another direction for future research is the automatic generation of recommendations in text form based on the feedback provided by the proposed framework, the main objective being to provide more human-interpretable recommendations and enhance the students' learning process.

Acknowledgements. This work received funding from the Horizon Europe research and innovation programme under Grant Agreement No. 101061509, project augMENTOR (Augmented Intelligence for Pedagogically Sustained Training and Education). We would like to thank the Department of Educational Sciences and Early Childhood Education, University of Patras, Greece, and the "Avgoulea-Linardatou" Microsoft Showcase School for providing us with the data used in this work.

References

1. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
2. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
3. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)
4. Duan, T., Anand, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A., Schuler, A.: Ngboost: Natural gradient boosting for probabilistic prediction. In: International Conference on Machine Learning. pp. 2690–2700. PMLR (2020)
5. Filippidi, A., Tselios, N., Komis, V.: Impact of moodle usage practices on students' performance in the context of a blended learning environment. Proceedings of Social Applications for Life Long Learning pp. 2–7 (2010)
6. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data? arXiv preprint arXiv:2207.08815 (2022)
7. Guleria, P., Sood, M.: Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. Education and Information Technologies **28**(1), 1081–1116 (2023)
8. Hur, P., Lee, H., Bhat, S., Bosch, N.: Using machine learning explainability methods to personalize interventions for students. International Educational Data Mining Society (2022)
9. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
10. Knapič, S., Malhi, A., Saluja, R., Främling, K.: Explainable artificial intelligence for human decision support system in the medical domain. Machine Learning and Knowledge Extraction **3**(3), 740–770 (2021)
11. Liaw, A., Wiener, M., et al.: Classification and regression by Random-Forest. R news **2**(3), 18–22 (2002)
12. Livieris, I.E., Drakopoulou, K., Tampakas, V.T., Mikropoulos, T.A., Pintelas, P.: Predicting secondary school students' performance utilizing a semi-supervised learning approach. Journal of educational computing research **57**(2), 448–470 (2019)
13. Livieris, I.E., Kiriakidou, N., Stavroyiannis, S., Pintelas, P.: An advanced CNN-LSTM model for cryptocurrency forecasting. Electronics **10**(3), 287 (2021)
14. Livieris, I.E., Kotsilieris, T., Tampakas, V., Pintelas, P.: Improving the evaluation process of students' performance utilizing a decision support software. Neural Computing and Applications **31**, 1683–1694 (2019)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
16. Ramaswami, G., Susnjak, T., Mathrani, A.: On developing generic models for predicting student outcomes in educational data mining. Big Data and Cognitive Computing **6**(1), 6 (2022)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
18. Tampakas, V., Livieris, I.E., Pintelas, E., Karacapilidis, N., Pintelas, P.: Prediction of students' graduation time using a two-level classification algorithm. In: Technology and Innovation in Learning, Teaching and Education. pp. 553–565. Springer (2019)