Predictive modeling of academic outcomes based on socioeconomic variables*

Rostyslav Zatserkovnyi[1,*,†], Roksoliana Zatserkovna[2,†]

[1] Lviv University of Trade and Economics, 10 Tuhan-Baranovskyi Str., Lviv, 79008, Ukraine
[2] Lviv Polytechnic National University, 12 Stepan Bandera Str., Lviv, 79013, Ukraine

Abstract

Students' academic performance, both in schools and universities, is influenced by a wide variety of socioeconomic, demographic, and behavioral factors. Identifying which of these factors most strongly correlate with poor academic outcomes can help educational institutions to more effectively allocate resources, as well as support students at risk of dropping out. In this study, we apply predictive machine learning models to a public dataset from Portuguese secondary schools in order to forecast student success. This forecasting is based on features such as parental education, employment status, access to educational support, and family relationships. Our results show that predictive modeling can effectively predict potential low academic performance, as well as highlight the socioeconomic indicators most critical in shaping a student's final grade. Information like this can be used as a basis for early intervention to help troubled students in the education system.

Keywords

machine learning, predictive modeling, classification, student performance, socioeconomic indicators [1]

1. Introduction

Academic performance plays a significant role in shaping the future opportunities of young individuals. Unfortunately, a significant number of students struggle to meet academic expectations due to a variety of factors outside of their personal control. Socioeconomic conditions – such as parental education, family support, work obligations, and access to educational resources – can have a significant influence on a student's ability to succeed in school. This means that recognizing these factors early is a pivotal task.

Existing literature suggests that there is a wide variety of factors which determine academic success. Aside from personal skill and motivation, student outcomes are often affected by:

- household income;
- the education level of parents and guardians;
- the presence of a stable and supportive home environment;
- quality of instruction in schools;
- availability of academic assistance;
- relationships with peers and classmates.

These non-academic factors can heavily influence a student's ability to learn, especially in countries with significant social inequality.

Machine learning, with its ability to model complex, nonlinear relationships, is well-suited to analyzing such nonlinear relationships. Aside from achieving high predictive accuracy, certain algorithms are "explainable", i.e., come with the ability to reveal the importance and contribution of individual features towards a decision. This makes it possible to determine how specific socioeconomic variables contribute to student success or failure.

*The 5th International Conference on Information Technologies: Theoretical and Applied Problems (ITTAP-2025), October 22-24, 2025, Ternopil, Ukraine

[1,*] Corresponding author.
[†] These authors contributed equally.
zatserkovnyi.rostyslav@gmail.com (R. Zatserkovnyi); roksoliana.s.zatserkovna@lpnu.ua (R. Zatserkovna)
0000–0001–6991–2866 (R. Zatserkovnyi); 0000–0003–1011–053X (R. Zatserkovna)

In this study, we focus on a publicly available dataset of Portuguese secondary school students. This dataset contains detailed information on academic performance (grades in two core subjects), and a variety of socioeconomic or lifestyle variables. We apply machine learning techniques to identify which indicators are most predictive of low academic performance. The core machine learning problem is framed as a classification task – determining if a student is a high or low performer based on known indicators. Therefore, we aim to create models that can help best predict this value, potentially helping educators target students which match the indicators for personalized intervention.

2. Literature review

Machine learning techniques have been used in a wide variety of domains and fields, and education is no exception. Several studies have examined the use of predictive models to identify students who are at risk of dropping out of school, with a particular emphasis on socioeconomic and behavioral variables [1, 2]. These approaches often frame the problem as a classification task, using algorithms such as Random Forests, Gradient Boosting, and Support Vector Machines to predict whether a student will pass or fail.

One well-cited dataset in this area is the UCI Student Performance Dataset, originally introduced by Cortez and Silva [3]. It has been used in numerous studies to explore the influence of parental education, school support, and daily routines on student outcomes. Kabakchieva [1], for example, used decision tree classifiers to predict final grades and found that family relationship quality, absences, and study time were among the most important predictors. Similarly, Kotsiantis et al. [4] used ensemble methods and logistic regression to analyze dropout risk. They reported that early performance and socioeconomic background were significant contributors to this risk.

Recent work has also focused on the interpretability of models in educational contexts. For example, Lundberg and Lee [5] introduced SHAP (SHapley Additive exPlanations), which has become a foundational tool for interpreting complex models. These approaches allow us to identify key socioeconomic indicators in a transparent way, which is critical for models used by educators or social workers. Similarly, Romero and Ventura [6] emphasized that in education, predictive accuracy is not enough. It's key for educators to understand why a model is assigning a high risk value to a particular student.

Another relevant factor in existing research is fairness and bias in such prediction systems. A study by Chen et al. [7] has shown that machine learning models trained on biased data can reinforce existing social inequalities. This is especially prominent when socioeconomic status or parental background is strongly correlated with educational achievement.

Our study contributes to this research area by using a dataset collected at the secondary school level, while applying a modern classification pipeline as well as an explainable classifier. The main goal of this project is to develop modern models that can help school administrators, teachers, and policy designers intervene early to help a student in need.

3. Dataset Overview

The primary data source for this study is the UCI Student Performance Dataset [3], which contains anonymized data on students from two Portuguese secondary schools.

The dataset includes detailed information on a student's academic performance in two subjects: mathematics, and Portuguese language. For the purposes of this research, we use the combined dataset of both subjects, which contains a total of 1044 rows across 33 features.

Each entry corresponds to an individual student, with some crossover between mathematics and Portuguese language datasets. Its variables and features can be grouped into the following categories:

- *Academic performance*: First-period (G1), second-period (G2), and final (G3) grades, on a scale from 0 to 20.
- *Family and parental background*: Includes parental education (Medu, Fedu), parental job types (Mjob, Fjob), family relationship quality (famrel), and whether the student lives in a two-parent home (Pstatus).
- *Support and study habits*: Includes whether the student receives educational support (schoolsup), family support (famsup), time spent studying per week (studytime), past class failures (failures), and access to extracurricular activities.
- *Demographics and daily life*: Includes student age, gender, travel time to school (traveltime), internet access (internet), free time after school (freetime), going out frequency (goout), and weekday/weekend alcohol consumption (Dalc, Walc).

In this study, the target variable for classification is the final grade (G3), which we binarize into two classes: pass (G3 ≥ 10) and fail (G3 ≤ 10). This is consistent with the passing standard in the Portuguese school system. A transformation like this allows us to formulate the task as a *binary classification problem*, which is aimed at predicting academic underperformance based on socioeconomic features.

Initial data exploration reveals some mild class imbalance: approximately 65% of students pass, while only 35% of them fail. To address this, we preserve class distribution in all train/test splits using stratified sampling. Several features also require preprocessing. Categorical features, such as parental job types and school names, are encoded numerically using one-hot encoding or ordinal mappings. Continuous features, such as study time and absences, are scaled using standard normalization techniques.

The dataset does not contain missing values, but there is a potential correlation issue between academic period grades (G1, G2 and G3). G1 and G2 are intermediate grades, which makes them good predictors of the final grade G3; but this obscures the socioeconomic factors we actually aim to analyze. Therefore, the intermediate grades are excluded from the input feature set: this decision ensures that the model primarily focuses on background and lifestyle indicators found in the dataset.

To better understand the data structure and feature relevance, we conduct exploratory visual analysis. Figure 1 illustrates the distribution of final grades. A significant concentration is present around the 10-14 range.

Figures 2 and 3 show correlations between parental education and student outcomes, as well as the impact of family relationship quality on passing rates. These initial plots confirm thehypothesis that socioeconomic and household factors can significantly contribute to academic success.

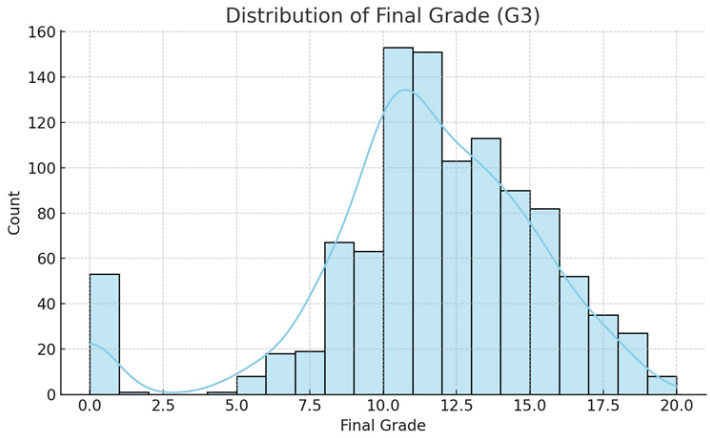

Figure 1: Distribution of final grade (G3) scores.

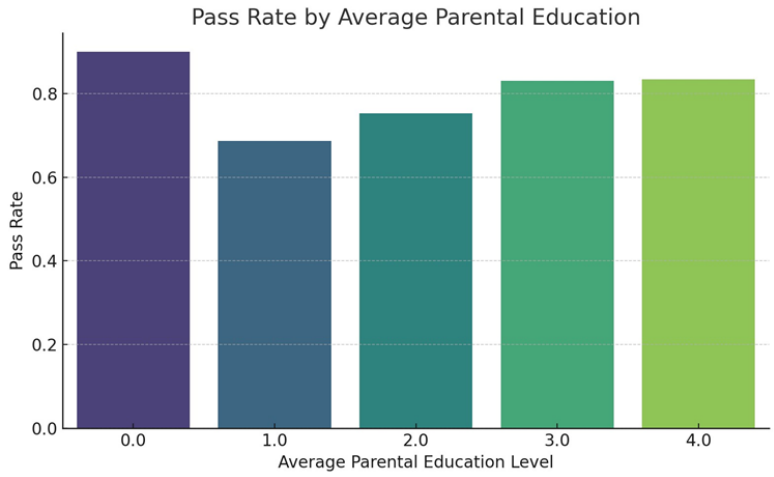

Figure 2: Pass rate grouped by parental education level (Medu, Fedu).

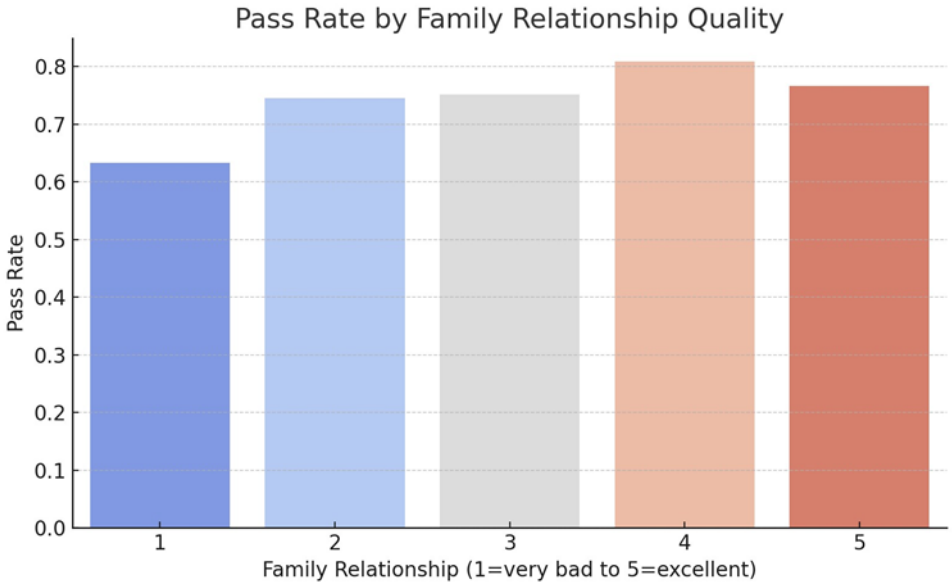

Figure 3: Effect of family relationship quality (famrel) on pass rate.

4. Implementation & Results

To model student academic outcomes, we frame the task as a binary classification problem: predicting whether a student will pass or fail based on non-academic variables. As noted in the previous section, we exclude early grade variables (G1 and G2). This makes sure that the model only focuses on socioeconomic, behavioral, and demographic inputs, instead of simply predicting future academic performance based on past performance.

We train and evaluate four supervised machine learning models, selected for their performance and compatibility with structured tabular data:

1. *XGBoost*: A scalable gradient boosting algorithm optimized for accuracy and training speed [8];
2. *LightGBM*: A leaf-wise boosting model that performs particularly well with categorical and imbalanced data [9];
3. *CatBoost*: An algorithm with built-in handling of categorical features, which reduces preprocessing complexity and makes it well-suited for our heavily categorical dataset;
4. *Explainable Boosting Machine (EBM)*: An inherently interpretable model that balances accuracy with the ability to be transparent about its output [10].

Tree-based ensemble models – a category to which all of the above models belong – are based on the principle of gradient boosting, where multiple decision trees are sequentially trained to minimize a loss function by correcting the errors of the previous trees. The prediction function in a boosting model is constructed in an additive manner:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

where F is the space of regression trees, $f_k$represents an individual decision tree, and K is the total number of trees. The objective function to be minimized usually includes both the empirical loss and a regularization term:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} Omega(f_k) \tag{2}$$

where l is a differentiable loss function (e.g., logistic or squared loss), and $Omega(f)$ is a complexity penalty term, defined as:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

Here, T is the number of leaves in the tree, w_j are the leaf weights, and γ, λ are the regularization parameters controlling model complexity. During training, each new tree f_t is fit to the negative gradients (residuals) of the loss with respect to the current predictions:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \qquad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{4}$$

To find the best split for a node, the model evaluates the gain in the objective:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{5}$$

where G_L, $G_R$and H_L, H_R are the sums of first and second derivatives in the left and right child nodes respectively.

Furthermore, XGBoost and LightGBM implement optimizations such as histogram-based splitting and leaf-wise growth. CatBoost incorporates ordered boosting and native categorical feature handling to reduce overfitting and improve generalization, and EBM focuses on interpretability of its results. This is done by making each term describe either a single feature, or a combination of or interaction between features.

The dataset is split into 80% training and 20% test data using stratified sampling to preserve the pass/fail class distribution. For each model, hyperparameters are optimized using 5-fold cross-validation. Evaluation metrics include:

- *Accuracy*: Overall proportion of correctly classified students;
- *Precision*: Proportion of predicted failures that were correct;
- *Recall*: Proportion of actual failures correctly identified;
- *F1 Score*: Harmonic mean of precision and recall;
- *ROC-AUC*: Area under the Receiver Operating Characteristic curve – this captures overall classification effectiveness.

Table 1
Performance comparison of classification models on predicting academic outcomes. Best-in-class performance is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 score	ROC-AUC
EBM	0,7847	0,8318	**0,9387**	0,8718	**0,6850**
CatBoost	**0,7895**	0,8182	**0,9387**	**0,8743**	0,6761
LightGBM	0,7847	0,8242	0,9202	0,8696	0,6692
XGBoost	0,7751	0,8187	0,9141	0,8638	0,6634

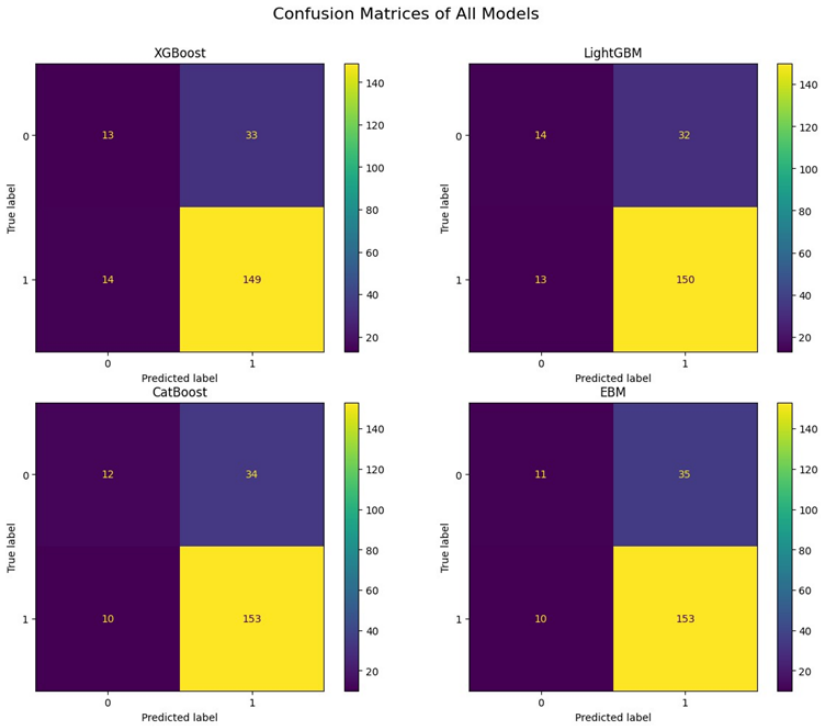

Figure 4: Confusion matrices of all models.

The results of the models are presented in Table 1, with confusion matrices and ROC curves visualized in Figures 4 and 5, respectively.

All models demonstrate reasonably high performance in identifying students at risk of academic failure, based entirely on socioeconomic and behavioral variables. Among them, EBM achieved the highest ROC-AUC and F1 score.

Aside from a slight boost in performance, its most major advantage is transparency. Figure 6 shows a global feature importance plot from the EBM model. This allows us to directly interpret which of the features most strongly influence model predictions. The feature importance ranking indicates that the most predictive factors include:

- *The number of past class failures*;
- *Parental education levels*, especially the mother's education;
- *Weekly study time*;
- *Frequency of social outings*.

Intuitively, support-related features such as access to educational help (schoolsup) and family support (famsup) also rank highly, suggesting actionable insights for educators.

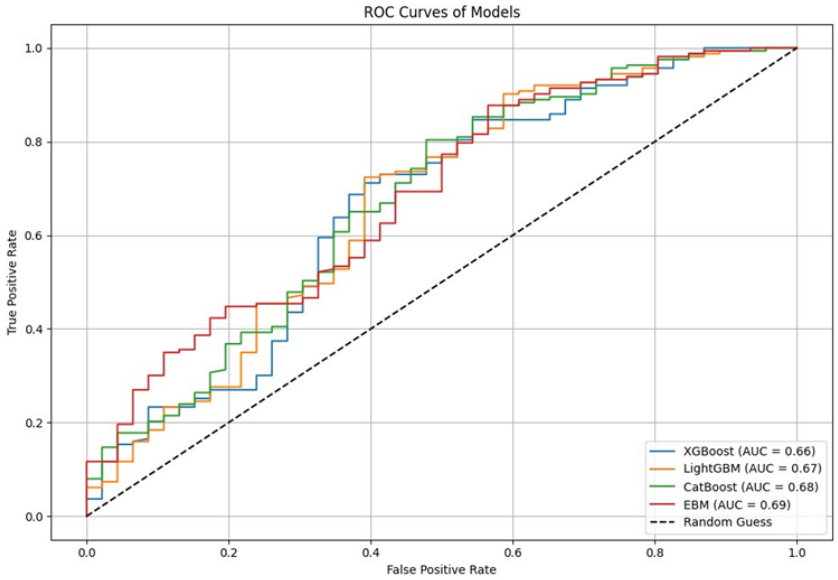

Figure 5: ROC curves of models.

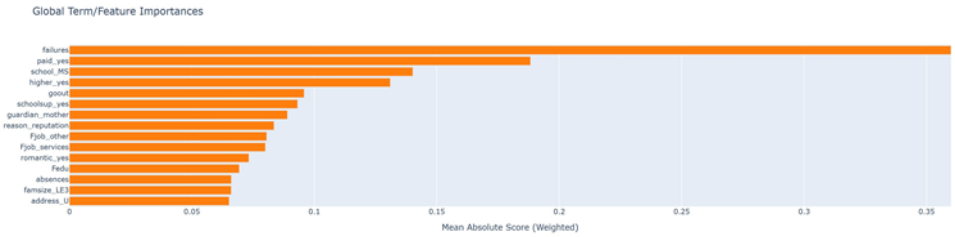

Figure 6: EBM global feature importance ranking.

Aside from summary rankings, EBM also allows us to visualize the contribution of individual variables in isolation. That is, we can plot the effect of a single variable - such as failures - on the predicted outcome across its entire value range. This provides interpretable insights into how specific socioeconomic indicators affect classification decisions globally. Figure 7 suggests that as the number of past class failures increases, the student's grade is far more likely to decrease.

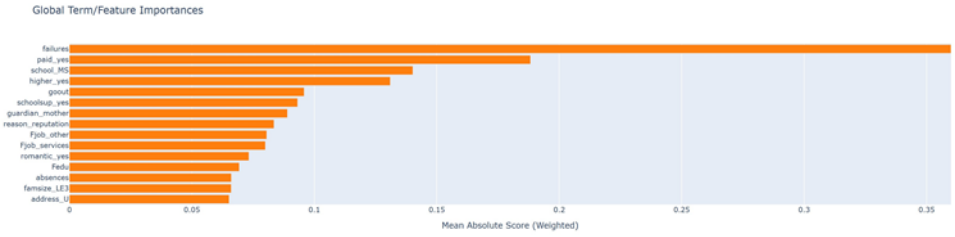

Figure 7: EBM shape function showing global contribution of the failures variable to model output.

These findings suggest that both strong predictive performance and transparency are achievable using modern machine learning approaches, as the EBM model offers fully interpretable diagnostics for each of its choices.

5. Discussion

This study explored the use of machine learning models to predict student academic outcomes based solely on socioeconomic and behavioral factors. Using a dataset of Portuguese secondary school students, we developed and evaluated several classification models that identified students at risk of academic failure without relying on prior academic grades.

Our results show that modern ensemble models - including XGBoost, LightGBM, and CatBoost - are generally effective at predicting academic risk using variables such as parental education, study habits, family support, and daily life indicators. The Explainable Boosting Machine (EBM) model achieved the best performance overall: it performed competitively while offering significant advantages in model interpretability. Its ability to produce global and local explanations makes it suitable for educational contexts where transparency and explainability are critical.

Among the most predictive features, we observed that students with large numbers of past class failures, low parental education levels, little time to study, and frequent social outings lower were substantially more likely to fail. Importantly, our approach avoids the use of prior grades (G1, G2) to ensure that interventions can be proposed even at the beginning of the academic year, when prior performance data from the current semesters may not yet be available.

That said, our study has certain limitations. The dataset is limited in size (1044 rows) and scope, restricted to two schools in a single country. The binary classification of student success into pass/fail categories also oversimplifies the nuance of the grading system; and cultural factors specific to Portugal may not generalize well to other education systems. Despite these limitations, our findings demonstrate the feasibility of predictive modelling to assess educational risk. Future extensions to our work may include:

- Incorporating multiple datasets from different countries, where different data columns with the same substance are matched together;
- Researching predictive modelling with a focus on Ukrainian schools and universities;
- Developing user interfaces for the model, such as a dashboard that can be used by school administrators and teachers.

Acknowledgements

This Word template was created by Tiago Prince Sales (University of Twente, NL) in collaboration with Manfred Jeusfeld (University of Skövde, SE). It is derived from the template designed by Aleksandr Ometov (Tampere University of Applied Sciences, FI). The template is made available under a Creative Commons License Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Kabakchieva, Predicting student performance by using data mining methods for classification, Cybernetics and Information Technologies 13 (2013) 61–72. doi:10.2478/cait-2013-0006.

[2] R. Cortez, A. Silva, Using data mining to predict secondary school student performance, in: Proceedings of 5th Future Business Technology Conference (FUBUTEC), Porto, Portugal, 2008.

[3] R. Cortez, A. Silva, Student performance data set, https://archive.ics.uci.edu/ml/datasets/Student+Performance, 2008. UCI Machine Learning Repository.

[4] S. Kotsiantis, C. Pierrakeas, P. Pintelas, Predicting students' performance in distance learning using machine learning techniques, Applied Artificial Intelligence 18 (2004) 411–426. doi:10.1080/08839510490442058.
[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
[6] C. Romero, S. Ventura, Data mining in education, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3 (2013) 12–27. doi:10.1002/widm.1075.
[7] I. Chen, F. Johansson, D. Sontag, Why is my classifier discriminatory?, in: Advances in Neural Information Processing Systems (NeurIPS), volume 31, 2018, pp. 3539–3550.
[8] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, volume 30, 2017, pp. 3149–3157.
[10] H. Nori, E. Jenkins, J. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019).