DOI: https://doi.org
ComputeX - Journal of Emerging Technology & Applied Science
Journal homepage: https://rjsaonline.org/index.php/ComputeX

Lightweight Explainable Machine Learning for Early Detection of Academic Performance Risk among Students

Furqan Ali[1]

[1] Fast University, Islamabad, Pakistan
Email: furqanalii80@gmail.com

ARTICLE INFO

Received:
July 27, 2026
Revised:
August 26, 2026
Accepted:
September 22, 2026
Available Online:
October 01, 2026

Keywords:
Educational Data Mining, Learning Analytics, Academic Risk Prediction, Explainable Artificial Intelligence, SHAP, XGBoost, Class Imbalance

Corresponding Author:
furqanalii80@gmail.com

ABSTRACT

Predictive models are rarely used unless educators can understand them and if they can, they can intervene before students complete their education and get the signs of academic failure. Three classifiers were designed and evaluated in this study: logistic regression, random forest, and extreme gradient boosting (XGBoost) with publicly available UCI Student Performance dataset (math course, n = 395) to predict academic performance risk. The final grade of less than 10 out of 20 was defined as risk, and 32.9% of students were at risk. The models were fine-tuned with stratified five-fold cross-validation on 80% of the data, and then tested once on a 20% stratified hold-out set, using cost-sensitive learning techniques to address the class imbalance problem. All models discriminated well (hold-out ROC-AUC 0.933–0.941; recall 0.846–0.962) and differences among them were small when compared with sampling uncertainty for the first (only) period for which the grade was available. The discrimination of the first period grade alone was 0.698–0.713 ROC-AUC. Results from the SHAP analysis of the XGBoost model revealed that most important variable for making predictions was prior assessment performance, with a mean absolute SHAP value of about 27 times higher than the next important variable. The results indicate that the predictive signal of the early term grades is relatively high, and that the signal of behavioural and demographic variables is relatively low. They also show that any methodology used to make the model explainable should be used in tandem with careful consideration of ethics and fairness when implementing it into real world use.

Introduction

Sentiment Failure and student attrition can be costly to students, institutions and societies, and the ability to identify poorly performing students at an early point in time is a major goal of educational data mining and learning analytics (EDM/LA) (Baker & Yacef, 2009; Romero & Ventura, 2010). The development of data collection systems and machine learning has led to the generation of an extensive body of research predicting grade, pass–fail, and dropout (Albreiki et al., 2021; Romero & Ventura, 2020). Systematic reviews have found that previous academic performance and internal assessment grades are the

most commonly used and most informative predictors, along with demographic and behavioural variables, which are used less consistently (Hellas et al., 2018; Shahiri et al., 2015).

In spite of this development, there are three issues that remain. The first is that many studies focus on predictive accuracy and little or no attention is devoted to the timing of the prediction. The model may be accurate, but be of no use for early intervention, if assessments are done at the end of a course. Second, there are ensemble models that are often more powerful than simpler ones, like random forests (Breiman, 2001) and gradient boosting (Chen & Guestrin, 2016), which the teachers and advisers may not be able to interpret. Thirdly, the use of predictive models in the education context involves ethical considerations relating to privacy, labelling and algorithmic bias (Baker & Hawn, 2022; Slade & Prinsloo, 2013).

The second problem is another area that Explanable AI can help solve, albeit partially. A theoretically sound framework for interpreting a prediction into the features that were input to the model is called SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017), and efficient algorithms have been developed for tree-based models (Lundberg et al., 2020). However, in high stakes settings, post hoc explanations of complex models should not be viewed as faithfully replicating inherently interpretable models, cautioned Rudin (2019), which is why it makes sense to compare a transparent baseline with ensemble methods.

The basic public data for this problem was presented by Cortez and Silva (2008) and were used to model the maths and Portuguese performance of secondary-school students. That data set was used as the basis for this study to answer the following questions. (1) What is the performance of logistic regression, random forest and XGBoost in determining the risk of poor final performance of students? (2) What is the effect of having an early-term grade available on predictive performance? (3) What are the most important predictors for the predictions, based on SHAP? The study provides a reproducible example of comparison with leakage reporting and explicit reporting of uncertainty and does not make claims based on a small sample size.

Literature Review

Since early surveys of educational data mining tasks, methods and data sources (Baker & Yacef, 2009; Romero & Ventura, 2010), research to provide a prediction of student performance has increased in a steady manner. An updated survey indicated that the field has now grown from basic classification to encompass learning analytics, deep learning and issues of interpretability (Romero & Ventura, 2020). According to systematic reviews conducted, the variables that are most commonly used as the predictor variables are prior grades and assessment results, while the most popular algorithms used are decision trees, neural networks and naive Bayes, support vector machines (Shahiri et al., 2015). A review of the literature of programming education also found a similar mix of studies, with different predictors, different outcome definitions, and different evaluation protocols, making it challenging to make direct comparisons across the literature (Hellas et al., 2018). A more recent systematic review found that academic and demographic factors remain the most significant variables, and ensemble methods are increasingly being used, with a continued lack of model transparency and external validation (Albreiki et al., 2021).

One of the most used open datasets was provided by Cortez and Silva (2008) with records from two Portuguese secondary schools. They used binary, five-level and regression formulations of the four models and found that the models with period grades predicted significantly better than models without them. This discovery set up a recurring conflict. The variables that are the best predictors are those that are taken the closest to the outcome, but they also have the least time available to intervene. Later research with the same data has confirmed this correlation, and the relationship to previous grades has been the main warning against using early-warning systems.

In terms of method, tree ensembles are now the norm for tabular education data. Random forests are based on averaging many decorrelated trees (Breiman, 2001), and XGBoost incorporates regularised and sequentially fitted trees, with efficient handling of sparse inputs (Chen & Guestrin, 2016). This is not ensured to be their advantage, though, especially with small sample sizes and near-linear relationships. Evaluation practice is also crucial. Error estimates if hyperparameters are tuned on the same folds as used for performance estimation are optimistically biased (Varma & Simon, 2006) and information not used at prediction time can creep into the feature set, and boost results (Kaufman et al., 2012). To calculate precision–recall results, the ROC-AUC is used in conjunction with it, since the latter can be more favorable in imbalanced problems (Saito & Rehmsmeier, 2015).

Interpretability is also an emerging field of research. SHAP uses Shapley values from cooperative game theory to assign a contribution to each feature for a specific prediction, which can then be aggregated to obtain global importance rankings

(Lundberg & Lee, 2017). The TreeSHAP algorithm was shown to enable an exact computation of the Shapley values for tree ensembles and to yield both local and global explanations (Lundberg et al., 2020). Though these tools make complex models more inspectable, Rudin (2019) pointed out that explanation is not same as interpretability, and simpler models should be preferred when they do just as well. In addition, explanations also refer to model behaviour and do not necessarily refer to causal mechanisms, meaning that if there is a feature that helps to predict the model, that does not mean there is a feature to intervene on.

Lastly, the ethical literature emphasizes learning analytics and a power imbalance, consent, data stewardship, and the potential for stereotyping students with labels (Slade & Prinsloo, 2013). Predictive models in education are known to be unevenly applied to different groups and proxy variables and biased training data are known to cause harm (Baker & Hawn, 2022). The literature justifies this setup: a clear baseline, two ensemble comparators, careful validation, and a clear discussion about fairness and limits.

Methodology

Research Design

A quantitative, predictive, supervised-learning design was used. Students were classified in a binary fashion as at-risk class or not at-risk class from variables pre-collected prior to the student assessment. The study utilized only secondary data and there was no interaction with the respondents. The workflow involves data acquisition, data preprocessing, stratified train–test split, hyperparameter tuning by cross-validation, one hold-out test set, and explanation using SHAP.

Data and Selection Criteria

This study employed the Student Performance dataset from the UCI Machine Learning Repository (Cortez and Silva, 2008). It includes data from the school reports and questionnaires of students from two secondary schools in Portugal, separated into Portuguese and Portuguese language and mathematics files. The mathematics file was downloaded from a public GitHub mirror of the original CSV, and not from UCI, therefore users should verify the dimensions and checksums of the file against that in the repository. The criteria used to select the data were: (i) the data must be publicly available and anonymous, (ii) have a continuous end-of-course grade, and (iii) have prior grades, absences, study time, and demographic and social covariates. The mathematics file was selected as it contained the most comprehensive set of measures for the analyses planned. Importantly, the data do not refer to university students, but to secondary-school students, and this should be taken into account when generalising to higher education.

The file had 395 students and 33 variables, no row was duplicated and no variables were missing. Students were aged 15 to 22 years (M = 16.70, SD = 1.28), and 208 (52.7%) were female. The mean final grade (G3) was 10.42 (SD = 4.58) on a 0–20 scale.

The result was a low grade (G3) lower than 10, the standard Portuguese pass mark. According to this definition, 32.9% (or 130) of the students were identified as being at risk. The four groups of predictors were: academic history and behaviour (e.g., academic failures, absences, study time, studying at school, studying at home, extra paid classes); social and lifestyle factors (e.g., extra circular activities, going out, free time, internet access, alcohol consumption, romantic relationship); demographic and household characteristics (e.g., address type, parents' occupation, parents' education level); and the first period grade (G1). The second period grade (G2) was not included because it was taken near the end of the outcome, and adding it to the model would have resulted in a model that was more an end-of course predictor than an early-detector, which would have undermined the goal of early detection (Kaufman et al., 2012). G3 was only used in the construction of the label. This resulted in 31 candidate predictors. A sensitivity analysis was also conducted by omitting G1 to see how performance would change if an assessment score was not available.

Data Preprocessing

No data was missing, which means no imputation was necessary. If missing data were found, the pipeline would have replaced the missing data for numeric values with the median and for categorical values with the mode within the cross-validation loop. Categorical variables were encoded by one-hot encoding; unseen categories were not used in the prediction. The numeric variables were standardised for logistic regression but not for the tree-based models because the two methods are said to be monotonic scaling invariant. Training data only were used to fit all preprocessing steps within a pipeline (no influence by validation or test observations).

Feature Selection

No univariate filter applied prior to modelling as filtering the whole data set could be a problem of information leakage and the sample was small. Rather, the selection was performed in the models. L2 regularisation reduced small coefficients in logistic regression and tree-based models did implicit selection via the choice of splitting and depth. Conceptual subsets were defined with only inclusion or exclusion of G1 and G2 subsets; these are defined in Section 3.3.

Class Imbalance

There was a moderate imbalance in the sample with 32.9% of students being at risk. Cost-sensitive learning was chosen over artificial oversampling due to a small dataset. The positive-class weight for logistic regression and random forest is balanced class weights, and XGBoost is the ratio of negative to positive training cases (2.04 in the training partition). To report alongside the accuracy and ROC-AUC (Saito & Rehmsmeier, 2015), recall and precision for the at-risk class and F1-score were also reported, as was precision–recall AUC.

Models

The comparisons were made among three classifiers. Logistic regression with L2 penalty was used as a meaningful baseline model. The bagged and boosted tree ensembles were represented by random forest (Breiman 2001) and XGBoost (Chen & Guestrin 2016), respectively. For logistic regression, the inverse regularisation strength was tuned by exhaustive grid search, the number of trees for random forest was tuned by exhaustive grid search, and the maximum depth and minimum leaf size for random forest were tuned by exhaustive grid search, as well as the number of boosting rounds, maximum depth, and learning rate for XGBoost. The tuning criterion used was ROC-AUC.

Validation Strategy

The data was split once in a 80/20 ratio (316/104 at risk, 79/26 at risk hold-out test set) with a fixed random seed of 42, stratified by the outcome. Stratified five-fold cross validation was used to tune hyper-parameters on training set exclusively. A set of cross-validated metrics were then calculated for the same folds for the selected configuration. These cross-validation estimates are therefore somewhat optimistic due to the tuning and estimation shared folds (Varma & Simon, 2006) and the hold-out test set gave the unbiased estimate. For final evaluation the test set was used once and a classification threshold of 0.5 was used.

Evaluation Metrics

Besides the precision-recall AUC and confusion matrix, accuracy, precision, recall, F1-score and ROC-AUC were calculated for at-risk class. Results were presented as mean and standard deviation between folds as a cross validation. A hold-out set of 79 students is not sufficient to draw conclusions between the models, so no significance tests were performed.

Explainability

For the XGBoost model, SHAP values were calculated using TreeSHAP algorithm (Lundberg et al., 2020) on the holdout set. The best cross-validated ROC-AUC was obtained with XGBoost, hence this was chosen for explanation. Global importance was summarised as the mean absolute SHAP value per feature, in log-odds units. The direction of the association was given by the correlation between the value of the feature and the value of the corresponding SHAP. Logistic regression coefficients were explored.

Software and Reproducibility

Analyses were performed in Python 3.12 with scikit-learn 1.8.0 (Pedregosa et al., 2011), XGBoost 3.4.1, SHAP 0.53.0, and pandas 3.0.2. The data split, cross validation folds and model initialisation was controlled by a fixed seed (42). The entire script will be made available on request from the author and deposited in a public repository upon acceptance.

Results and Analysis

Table 1

Characteristics of the Analytic Sample

Characteristic	Value
Students (N)	395
Candidate predictors	31
Missing values	0
Age, M (SD)	16.70 (1.28)
Female, n (%)	208 (52.7)
Final grade (G3), M (SD)	10.42 (4.58)
At risk (G3 < 10), n (%)	130 (32.9)
Training set, n (at risk)	316 (104)
Hold-out test set, n (at risk)	79 (26)

Note. Data from the UCI Student Performance dataset (mathematics course; Cortez & Silva, 2008). G2 and G3 were not used as predictors.

Table 1 displays a fully unbalanced samples. All three models had high levels of cross-validated discrimination (ROC-AUC .903–.922) with standard deviations of .03 to .05 that were not appreciably larger than the differences between models. The operating characteristics of models varied more. The most accurate model was XGBoost, which had the highest recall score (.933) and the lowest precision score (.646), and the most precise model was random forest, with the lowest recall score (.790) and the highest precision score (.701).

Table 2

Five-Fold Cross-Validated Performance on the Training Set (M ± SD), Including G1

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic regression	.823 ± .034	.687 ± .047	.857 ± .128	.758 ± .057	.916 ± .032

Random forest	.817 ± .042	.701 ± .071	.790 ± .110	.738 ± .065	.903 ± .048
XGBoost	.807 ± .043	.646 ± .056	.933 ± .071	.761 ± .051	.922 ± .050

Note. Selected hyperparameters: logistic regression C = 0.1; random forest 500 trees, unrestricted depth, minimum leaf size 3; XGBoost 100 rounds, depth 2, learning rate 0.03. Estimates are slightly optimistic because tuning and estimation used the same folds.

On unseen data, this is the same pattern as displayed in Table 3. ROC-AUC ranged from .933 to .941 in all three models and logistic regression had the nominally highest, suggesting that the discrimination did not increase with model complexity. XGBoost failed to catch one at-risk student but triggered 13 false alarms, whereas the random forest triggered the least false alarms (7), and missed 4 at-risk students. In the test set, with 26 positive cases, any differences are not considered rank order.

Table 3

Hold-Out Test Performance and Confusion Matrices (n = 79; 26 At Risk)

Model	Acc.	Prec.	Rec.	F1	ROC-AUC	PR-AUC	TN	FP	FN	TP
Logistic regression	.848	.719	.885	.793	.941	.896	44	9	3	23
Random forest	.861	.759	.846	.800	.933	.887	46	7	4	22
XGBoost	.823	.658	.962	.781	.935	.889	40	13	1	25

Note. Threshold = 0.5. Acc. = accuracy; Prec. = precision; Rec. = recall; TN = true negatives; FP = false positives; FN = false negatives; TP = true positives. Positive class = at risk.

The sensitivity analysis that eliminated G1 resulted in a significant drop in performance. The cross-validated ROC-AUC was .665 (logistic regression), .672 (random forest), and .666 (XGBoost) and the hold-out ROC-AUC was .698, .702, and .713, respectively. On the hold out set, XGBoost achieved recall .654 and F1 .618 while random forest achieved recall .500 and F1 .565 (a third to a half of students classified as at-risk were not picked up). The values suggest that there was some (but not all) predictive information in the non-grade variables.

Table 4

Ten Most Influential Predictors in the XGBoost Model (SHAP, Hold-Out Set)

Rank	Feature	Mean |SHAP| (log-odds)	Association with risk
1	First-period grade (G1)	1.819	Lower grade, higher risk
2	Going out with friends	0.067	Higher value, higher risk

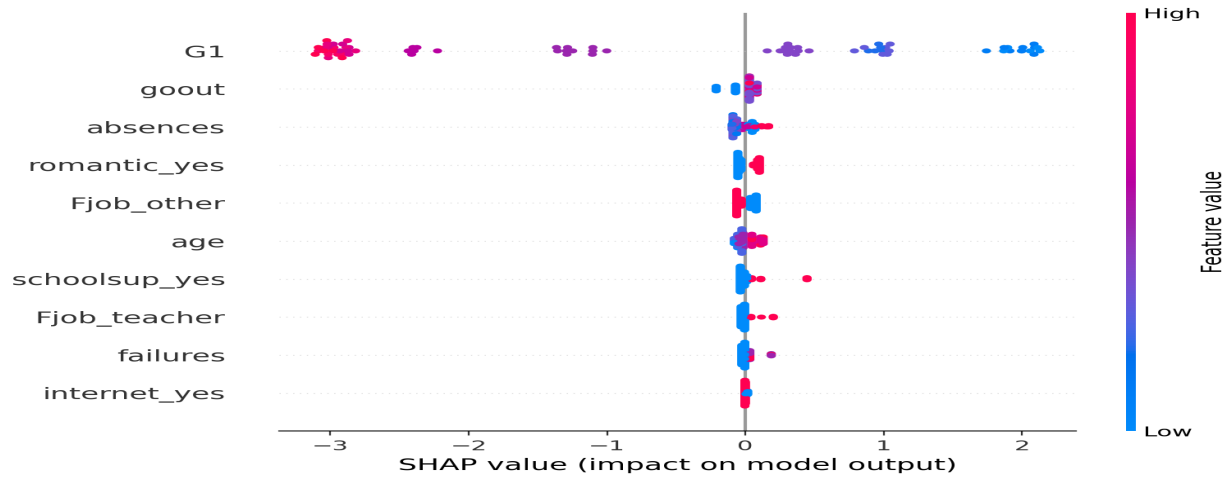
3	Absences	0.066	More absences, higher risk
4	Romantic relationship (yes)	0.063	Yes, higher risk
5	Father's job: other	0.058	Yes, lower risk
6	Age	0.052	Older, higher risk
7	School support (yes)	0.049	Yes, higher risk
8	Father's job: teacher	0.026	Yes, higher risk
9	Past failures	0.025	More failures, higher risk
10	Internet access (yes)	0.003	Yes, lower risk

Note. Direction was inferred from the correlation between feature values and SHAP values and is descriptive, not causal.

The sensitivity analysis that eliminated G1 resulted in a significant drop in performance. The cross-validated ROC-AUC was .665 (logistic regression), .672 (random forest), and .666 (XGBoost) and the hold-out ROC-AUC was .698, .702, and .713, respectively. On the hold out set, XGBoost achieved recall .654 and F1 .618 while random forest achieved recall .500 and F1 .565 (a third to a half of students classified as at-risk were not picked up). The values suggest that there was some (but not all) predictive information in the non-grade variables.

Figure 1

SHAP Summary Plot for the XGBoost Model (Hold-Out Set)



Note. Each point is one student. Horizontal position shows the SHAP value (log-odds contribution toward the at-risk class); colour shows the feature value (red = high, blue = low).

Discussion

The conclusions drawn from the results are that in this dataset an early-term grade was the key contributor to the predictive signal. Since G1 was available, all three classifiers had a ROC-AUC score high enough for detecting at-risk students, as expected, since period grades were found to be the most useful in the original work (Cortez & Silva, 2008). The sensitivity analysis, to the contrary, revealed that, with no grade whatsoever, discrimination was modest, corroborating the general finding that behavioural and demographic variables alone are poor bases for early-warning systems (Hellas et al., 2018; Shahiri et al., 2015).

What is interesting about the comparable performance of the three models is noteworthy. Envelopes around the logistic regression baseline were comparable to or slightly larger than the ensembles on the hold-out set, and differences from the baseline, as cross-validated, were small and well within one standard deviation. This seems possible if one dominant predictor is near-monotonically related to the outcome, and is consistent with Rudin's (2019) claim that complex models are not always the better choice. In practice, this might then be a matter of working characteristics and transparency among models. For an organization with a strong focus on recapturing nearly all at-risk youth, recall is more important and may warrant sacrificing precision rate and the increased number of false alarms, while for an organization with limited counselling resources, reducing the threshold of a more precise algorithm such as random forest may be more appropriate.

The baseline interpretation was corroborated by the SHAP analysis. Secondary predictors (going out, absences, age) were in the expected direction, but their contributions were small, and some were not reasonable targets for intervention (paternal occupation categories, romantic relationship status), and may be proxies for other circumstances. These associations are not causal in nature and the attributions of this order may be unstable in a sample of 79 test students. The explanation also does not explain why students fail (Lundberg & Lee, 2017; Rudin, 2019) nor does it describe the fitted model.

There are a number of implications of this. If there is already an early assessment, that is most useful, and institutions may want to focus on gathering low-stakes, regular assessments instead of increasing the collection of demographic data. Risk scores should not be used for automatic decisions but for discussion and support. In addition, it is important that alerts review error rates by subgroups as proxy variables like parental occupation may capture inequity (Baker & Hawn, 2022).

Limitations

The results are subject to a number of limitations. The number of samples was small (n = 395, 79 of which constituted the test set), and uncertainty is introduced into the estimates when the performance, and model differences, could not be tested formally. The data is from secondary schools in a single subject and single country, and has been obtained over 10 years ago, and so the generalisation to universities and the current cohorts is untested. The at risk threshold (G3<10) was a convention, not a school policy, and students whose final grades reflect events that are not recorded may be labeled at risk. The cross validation estimates were skewed in their favor and the estimates had been more conservative had they been done in a nested and/or repeated fashion (Varma & Simon, 2006). The models did not take the form of calibration, nor were they tested with a sample separate from the original in time or institution and they were not evaluated for subgroup fairness. SHAP attributions are not causal, but rather model-dependent.

Ethics and privacy

The data were made publicly available anonymised, there was no attempt to restake persons. Nevertheless, making a prediction about academic risk entails sensitive inferences about minors or young adults. Informed consent or clear legal basis is required for deployment, data minimisation and transparency regarding the use of scores, as well as human oversight (Slade & Prinsloo, 2013). Some attributes are sensitive, e.g., sex, family background, relationship status, etc., and may introduce or exacerbate bias; an institution should check for error rates across attributes and may choose to exclude sensitive attributes (Baker & Hawn, 2022). Risk labels should help students, not hurt them, and there should be awareness of the potential for self-fulfilling prophecies. The data, seed information, library versions, split proportions, search grids and definitions of the metrics were presented to ensure data reproducibility. It should be stored in a public script repository, and replication should check the data file against the original from UCI.

Conclusion

This study employed logistic regression, random forest and XGBoost models to compare their performance in the early identification of academic performance risk, based on a publicly available student-performance dataset. In the first period, all

three models discriminated well, and did not differ greatly from each other, with hold-out ROC-AUCs of 0.933–0.941, and the simplest model was competitive with the ensembles. If you don't have that grade, then the discrimination is about 0.70 and the SHAP analysis revealed that the first-period grade was responsible for almost every bit of model's contribution. Results showed that the assessment data from the beginning of the school term are the most influential inputs into the early-warning model, and that transparent baselines should also be provided along with more complex methods. The results are found to be secondary, small and dated; should be reproduced on modern university data using fair audits, calibration, and nested and repeated validations before operational use.

References

1. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), Article 552. https://doi.org/10.3390/educsci11090552
2. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9
3. Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.
4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
6. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of the 5th Future Business Technology Conference* (pp. 5–12). EUROSIS.
7. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education: Companion* (pp. 175–199). ACM. https://doi.org/10.1145/3293881.3295783
8. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. https://doi.org/10.1145/2382577.2382579
9. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
12. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532
13. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), Article e1355. https://doi.org/10.1002/widm.1355
14. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
15. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432
16. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157
17. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366
18. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, Article 91. https://doi.org/10.1186/1471-2105-7-91