JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Predicting Student Exam Performance Using Explainable Regression Models: A Learning Analytics Study

Thuy Thi Vo*, Vu Thien
University of Da Nang, Vietnam

*Corresponding author email: thi.vo@vnuk.udn.vn

Abstract

Identifying which behavioral and demographic factors most strongly predict student academic performance is a long-standing goal of learning analytics. We apply linear regression, ridge regression, random forest, and gradient boosting to a 1,000-student behavioral and demographic dataset. Five-fold cross-validated R^2 shows that linear and ridge regression (R^2=0.656) explain more variance than the ensemble methods (R^2≈0.57), indicating a primarily additive structure in the data. To test this additive structure for residual non-linear or interaction effects and to obtain individual-level explanations, we apply SHapley Additive exPlanations (SHAP) to the random forest model via TreeExplainer, and cross-validate the resulting feature ranking against the standardized coefficients of the top-performing linear model. SHAP analysis identifies study time as the dominant predictor, with a mean |SHAP| of 4.93 points (approximately 4.1 exam-score points per additional study hour), followed by prior academic performance (mean |SHAP|=3.44) and attendance (mean |SHAP|=2.38). Internet access and part-time employment emerge as statistically meaningful binary factors (mean score gap: +5.0 and −3.3 points respectively), consistent with literature on the digital divide and academic workload. Parental education, gender, and extracurricular participation contribute negligible SHAP values (mean |SHAP| ≤0.16), a pattern that also holds for the corresponding linear coefficients. These results suggest that actionable behavioral variables – particularly study time and attendance – account for the bulk of explainable variance in examination outcomes, while fixed demographic characteristics contribute little incremental predictive power.

Keywords: learning analytics, student performance prediction, SHAP, random forest, explainable AI

Cite as: Thuy Thi Vo, & Vu Thien. (2026). Predicting Student Exam Performance Using Explainable Regression Models: A Learning Analytics Study. Journal of Systems Engineering and Information Technology (JOSEIT), 3(3). https://doi.org/10.29207/joseit.v3i3.8471

Received by the Editor: 2-11-2024
***Final Revision*:** 25-12-2024
***Published*:** 30-12-2024

This is an open-access article under the CC BY-4.0 license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Understanding what drives student academic performance is fundamental to evidence-based educational policy and targeted intervention design. Learning analytics – the measurement, collection, analysis, and reporting of data about learners and their contexts for the purpose of optimizing learning [8] – has attracted growing attention as institutional datasets on student behavior and outcomes have become more accessible, and its methodological foundations overlap substantially with the broader field of educational data mining [11]. Machine learning methods have been applied to predict grade outcomes, identify at-risk students, and quantify the relative contributions of study behavior, lifestyle, and demographic factors to academic success [1], [2]; recent systematic literature reviews confirm that regression models and tree-based ensembles dominate this literature, though consistent reporting of model interpretability remains uncommon [12], [13]. Beyond retrospective prediction, these techniques increasingly underpin operational early-warning systems designed to flag at-risk students for timely intervention [14].

A persistent challenge in this literature is interpretability: models with the highest predictive accuracy, such as tree ensembles, are generally opaque, while interpretable models such as linear regression may miss non-linear interactions. This tension motivates the broader explainable artificial intelligence (XAI) literature [15], within which model-agnostic local surrogate methods such as LIME [16] and additive attribution methods such as SHAP represent two dominant families of approach. SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee [3] and later extended to tree ensembles through the polynomial-time TreeExplainer algorithm [17], resolve this tension by providing theoretically grounded, additive feature attributions that are consistent across model classes. By computing the marginal contribution of each feature to each individual prediction, SHAP allows researchers to move beyond aggregate feature importance rankings to richer, sample-level characterizations of how each factor pushes predictions up or down relative to a baseline. Its application to learning analytics is growing [2] but remains underexplored relative to purely accuracy-focused comparisons.

The present study applies four regression models – linear regression, ridge regression, random forest [4], and gradient boosting [5] – to a 1,000-student dataset recording daily study time, class attendance, sleep duration, previous academic performance, internet access, extracurricular participation, part-time employment, parental education level, and gender, with the target variable being a final examination score. We compare model performance under 5-fold cross-validation, then apply SHAP to the random forest model to test for non-linear or interaction effects beyond the additive linear structure and to answer the primary research question at the individual-student level: which factors contribute most to predicted examination scores, and in which direction? We cross-validate the resulting SHAP-based feature ranking against the standardized coefficients of the top-performing linear model, and contextualize the findings against the empirical literature on study time [6], [7], class attendance [18], sleep [19], the digital divide [9], [20], term-time employment [21], parental involvement [22], gender differences in achievement [23], and extracurricular participation [24].

2. Method

All analysis was conducted in Python 3 (pandas, scikit-learn, SHAP). Random seeds were fixed to 42 for all stochastic components; 5-fold cross-validation used StratifiedKFold equivalent (KFold, shuffle=True, random_state=42).

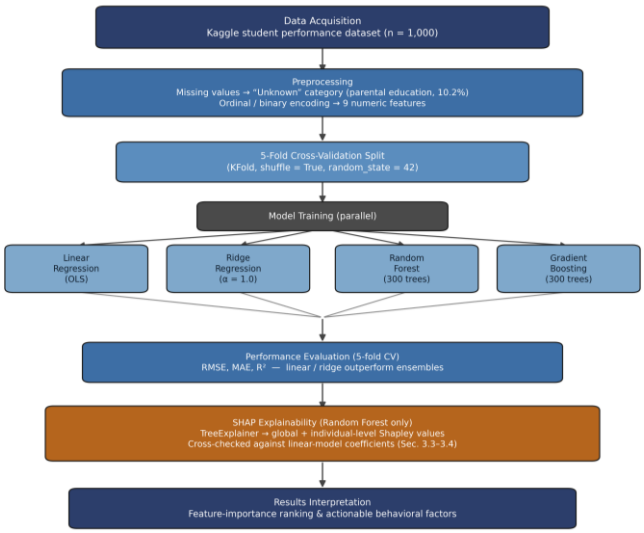


Figure 1. Research Pipeline

2.1 Dataset

The dataset contains 1,000 student records drawn from a publicly shared Kaggle repository [10]. Each record includes five continuous features (study time in hours per day, class attendance percentage, daily sleep hours, previous exam grade, and final exam score as the prediction target) and five categorical features (gender: Female/Male; parental education: High School / Bachelors / Masters / PhD; internet access: Yes/No; extracurricular participation: Yes/No; part-time employment: Yes/No). The final grade column (A/B/C/D/F letter grades) was excluded from features because it is a deterministic function of the final exam score (standard thresholds A≥90, B≥80, C≥70, D≥60). Parental education had 102 missing values (10.2%); these were assigned to an "Unknown" level, which was placed between High School and Bachelors in the ordinal encoding. This missing-indicator treatment is appropriate primarily when missingness itself may carry information, and is revisited as a limitation in Section 3.5 [25]. Descriptive statistics are shown in Table 1.

Table 1. Descriptive Statistics for Continuous Variables (n=1,000)

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Study time (h/day)	3.57	1.48	0.5	2.6	3.6	4.5	8.1
Attendance (%)	85.1	9.3	54.8	78.8	85.2	91.9	100.0
Sleep (h/day)	6.48	1.08	4.0	5.7	6.5	7.3	9.0
Previous grade	69.7	12.6	31.3	61.0	69.6	78.4	100.0
Final exam score	83.5	10.3	46.8	76.1	83.8	91.5	100.0

Parental education was encoded ordinally (High School=0, Unknown=1, Bachelors=2, Masters=3, PhD=4). Gender (Female=1, Male=0), internet access (Yes=1, No=0), extracurricular participation (Yes=1, No=0), and part-time employment (Yes=1, No=0) were binary-encoded. This yielded nine numeric features for all models.

Four models were evaluated: (i) ordinary least-squares linear regression; (ii) ridge regression [26] with regularization parameter α=1.0, included to assess sensitivity to multicollinearity among the nine encoded predictors; (iii) a 300-tree random forest [4] (max_features='sqrt', min_samples_leaf=1); and (iv) gradient boosting regression [5] (300 trees, learning rate=0.05, max_depth=4). All were evaluated using 5-fold cross-validation with three metrics: root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

2.2 SHAP Analysis

Although linear and ridge regression achieved marginally higher cross-validated R^2 than the random forest (Section 3.2), SHAP was applied to the random forest rather than to the top-performing linear specification, for three reasons. First, for an additive linear model the per-feature attribution is already available in closed form as the standardized regression coefficient, so a separate SHAP decomposition would be mathematically redundant; applying SHAP to the random forest instead has diagnostic value, testing whether its marginally lower aggregate R^2 conceals non-linear or interaction effects invisible to the linear specification. Second, TreeExplainer [17] computes exact, polynomial-time Shapley values for tree ensembles without the approximation error that a model-agnostic estimator would introduce, making it the natural choice for the one genuinely non-linear model in the comparison.

Third, instance-level Shapley attributions are informative primarily for a model whose prediction surface is non-linear; for the linear specification, each student's attribution is a deterministic rescaling of the same global coefficient and would add no information beyond the coefficient table reported in Section 3.4. SHAP [3] values were therefore computed for the random forest using TreeExplainer [17], which computes exact Shapley values for tree-based models in polynomial time. For each student and each feature, the SHAP value expresses how much that feature shifted the predicted score relative to the expected value across all students. Mean absolute SHAP (mean |SHAP|) across students is reported as the global feature importance measure, and is cross-checked in Sections 3.3–3.4 against the sign and relative magnitude of the corresponding OLS coefficients to test whether the two explanation methods converge on the same substantive ranking. Individual-level SHAP values were plotted as a dependence scatter for study time to show the dose–response relationship between study hours and its predicted contribution.

3. Results and Discussion

3.1 Pairwise Correlations

Table 2 reports Pearson correlations between each feature and the final exam score. Study time is the single strongest individual predictor (r=0.568, $p<0.001$), followed by previous grade (r=0.406) and attendance (r=0.262), consistent with meta-analytic evidence that class attendance is among the strongest single behavioral predictors of collegiate grades [18]. Internet access shows a modest positive association (r=0.171), consistent with digital-divide literature showing access to home internet to be a meaningful predictor of educational outcomes [9]. Part-time employment shows a negative association (r=−0.148), consistent with research documenting the academic workload cost of term-time employment.

Sleep (r=0.148) is a weak positive predictor, smaller in magnitude than reported in studies using objective, wearable-tracked sleep measures, where sleep quality, duration, and consistency together have explained close to a quarter of the variance in academic performance [19]. Parental education, gender, and extracurricular participation each correlate with exam score at $r \leq 0.04$, indicating negligible marginal predictive power when considered in isolation. The study-time effect observed here (r=0.568) substantially exceeds the meta-analytic estimate of r=0.12 from real student populations [6], [7], a gap that reflects the synthetic origin of the data and its simulation of a stronger direct effect than is typically found in ecological settings.

Table 2. Pearson Correlations with Final Exam Score

Feature	r	Feature	r
Study time (h)	0.568	Parental education	0.036
Previous grade	0.406	Part-time job	−0.148
Attendance (%)	0.262	Extracurricular	0.022
Internet access	0.171	Gender (Female=1)	0.006
Sleep (h)	0.148		

3.2 Model Performance

Table 3 summarizes 5-fold cross-validated performance. Linear regression and ridge regression achieve identical performance (RMSE=6.06, MAE=4.81, R^2=0.656). The random forest (RMSE=6.75, R^2=0.573) and gradient

boosting (RMSE=6.76, R^2=0.572) perform notably worse than the linear models. This inversion of the usual ensemble-beats-linear pattern indicates that the dominant relationships in this dataset are well-approximated by linear main effects, with limited variance attributable to feature interactions or non-linearities that ensemble methods could exploit. The near-identical performance of linear regression and ridge regression implies that multicollinearity among predictors is not severe enough to require shrinkage.

Table 3. 5-Fold Cross-Validated Regression Performance

Model	RMSE	MAE	R^2
Linear Regression	6.062	4.805	0.656
Ridge Regression (α=1)	6.062	4.805	0.656
Random Forest (n=300)	6.754	5.326	0.573
Gradient Boosting (n=300)	6.758	5.366	0.572

3.3 SHAP Feature Importance

Figure 2 (left panel) shows global SHAP importance for the random forest. Study time is the dominant driver by a substantial margin (mean |SHAP|=4.93 score points), more than 40% above the second-ranked factor. Previous grade ranks second (mean |SHAP|=3.44), reflecting the persistent effect of prior academic habit and preparation. Attendance ranks third (mean |SHAP|=2.38), jointly confirming a cluster of active behavioral factors as the main levers of exam performance. Sleep, internet access, and part-time employment form a secondary tier (mean |SHAP| 0.60–0.76), each contributing less than one exam-score point on average but remaining above the near-zero contributions of parental education (mean |SHAP|=0.16), extracurricular participation (0.15), and gender (0.07). The right panel of Figure 2 shows the SHAP dependence plot for study time: every additional hour of study is associated with a roughly linear 4–5 score-point upward shift in predicted score. Students studying fewer than two hours per day routinely receive negative SHAP contributions exceeding −5 points, while those studying six or more hours receive contributions exceeding +10 points. The relationship appears monotone and approximately linear, consistent with the superior performance of linear regression over the ensemble in this dataset.

This SHAP-based ranking mirrors both the Pearson correlation ordering in Table 2 and the OLS coefficient pattern reported in Section 3.4: the same three variables (study time, previous grade, attendance) lead under all three measures, internet access and part-time employment occupy the same intermediate tier, and parental education, extracurricular participation, and gender are negligible throughout. This convergence across a tree-based attribution method (SHAP), a linear model's own coefficients, and simple bivariate correlation indicates that the random forest's modest interaction and non-linear capacity is not being used to construct a materially different explanation of the outcome than the top-performing linear model already provides, reinforcing the Section 3.2 conclusion that the dominant structure in this dataset is additive.

The negligible SHAP and correlational signal for parental education, extracurricular participation, and gender should be read against a wider literature in which these factors are not always weak: meta-analytic evidence links parental socioeconomic status to achievement mainly through specific engagement behaviours rather than educational attainment per se [22], gender gaps in scholastic achievement are typically small but systematic in real cohorts [23], and extracurricular participation's association with achievement depends strongly on activity type and developmental stage [24]. Their absence of signal here is therefore better read as a property of this synthetic dataset's data-generating process, which does not appear to encode these more conditional relationships, than as a general claim that such factors are unimportant.

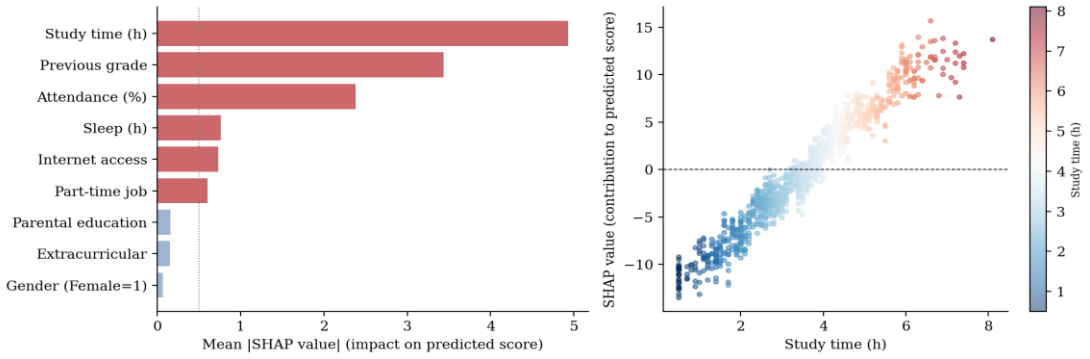


Figure 2. Global SHAP Feature Importance and SHAP dependence plot for study time

3.4 Group-Level Effects: Internet Access and Part-Time Employment

Figure 3 shows scatter plots of the two strongest continuous predictors (study time and previous grade) against final exam score, and a group comparison for internet access and part-time employment. Students without internet

access score on average 5.0 points lower than those with access (79.3 vs. 84.3; 95% CI on the gap: approximately [3.3, 6.7] points). This finding aligns with evidence that home internet access consistently predicts academic outcomes across diverse student populations, particularly for homework completion and access to digital learning resources [9]; quasi-experimental evidence on realized computer and broadband rollouts has, however, found more mixed effects on standardized test outcomes, underscoring that access alone does not guarantee productive use [20]. Students working a part-time job score on average 3.3 points lower than those without (81.3 vs. 84.6; 95% CI: [1.7, 4.9] points).

This penalty is consistent with time-budget constraints documented in employment–performance research: time devoted to paid work competes directly with study and sleep time, and matches large-scale survey evidence that term-time employment has a detrimental, broadly dose-dependent effect on final attainment [21]. In the OLS model, internet access carries a coefficient of +5.45 points and part-time employment a coefficient of −3.19 points, both larger in magnitude than attendance (+0.33 per percentage point) over any realistic attendance range, underscoring the structural importance of resource access and time availability relative to marginal behavioral differences among students who are already attending class.

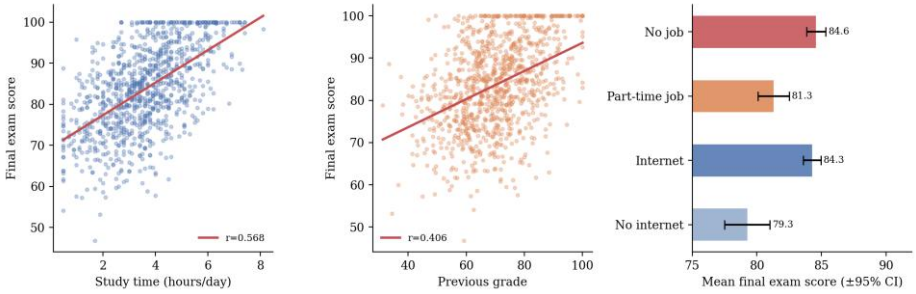


Figure 3. Left: study time vs final exam score (r=0.568). Centre: previous grade vs final exam score (r=0.406). Right: group mean scores (±95% CI) for internet access and part-time job status

3.5 Limitations

Three limitations bound the generalizability of these findings. First, the dataset is synthetically generated rather than drawn from a real student cohort. The simulation embeds specific effect sizes for each feature, and the observed study-time correlation (r=0.568) substantially exceeds the meta-analytic estimate from real populations (r≈0.12–0.19 [6], [7]), confirming that the data-generating process amplifies this relationship beyond what ecological research typically reports. Findings should therefore be treated as characterizations of the synthetic DGP rather than as claims about real student populations. Second, the 102 missing values in parental education (10.2%) were addressed through a simple "Unknown" category rather than model-based imputation; this missing-indicator approach is best justified when missingness itself may be informative, and can introduce bias if used for confounder adjustment more broadly [25], though its use here is defensible given the feature's negligible SHAP importance regardless. Third, the absence of a temporal dimension precludes causal inference; all relationships are correlational within the cross-sectional structure of the data.

4. Conclusion

This study applied four regression models, together with SHAP-based explainability of the random forest, to a 1,000-student behavioral and demographic dataset to identify the factors most predictive of examination performance. Linear regression achieved the best cross-validated R^2 (0.656), outperforming random forest and gradient boosting, which indicates that dominant relationships in the data are well-captured by linear main effects. SHAP analysis revealed a clear three-tier importance structure: study time, previous grade, and attendance form the primary behavioral cluster (mean |SHAP| 2.4–4.9 points each); internet access and part-time employment form a secondary structural tier; and parental education, gender, and extracurricular participation contribute negligible predictive power.

These findings collectively support the conclusion that modifiable behavioral factors – particularly deliberate study time and consistent attendance – are the strongest actionable levers for improving examination outcomes, while demographic characteristics explain comparatively little additional variance once behavioral variables are accounted for. Practitioners designing early-warning or intervention systems should prioritize monitoring and supporting study-time behaviors and attendance patterns, and should additionally assess digital access as a structural equity concern rather than a pedagogical one.

References

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," Education and Information Technologies, vol. 28, no. 1, pp. 905–971, 2023.

[2] Y. Chen and L. Zhai, "A comparative study on student performance prediction using machine learning," Education and Information Technologies, vol. 28, no. 9, pp. 12039–12057, 2023.
[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
[4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] C. Huang, "Meta-analysis of the relation between study time and academic achievement," in Proc. IAFOR Intl. Conf. on Education, 2015.
[7] M. Theobald, "Study longer or study effectively? Better study strategies can compensate for less study time and predict goal achievement and lower negative affect," British Journal of Educational Psychology, vol. 95, no. 1, 2025.
[8] G. Siemens and P. Long, "Distilling the essence of learning analytics," International Journal of Distance Education Technologies, vol. 9, no. 2, pp. 57–68, 2011.
[9] G. L. Pierce and P. F. Cleary, "The persistent educational digital divide and its impact on societal inequality," PLOS ONE, vol. 19, no. 4, 2024, doi: 10.1371/journal.pone.0286795.
[10] H. Patil, "Student Performance & Study Habits Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/harshadapatil31/student-performance-and-study-habits-dataset
[11] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.
[12] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques," Education Sciences, vol. 11, no. 9, p. 552, 2021, doi: 10.3390/educsci11090552.
[13] P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, "Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review," Applied Sciences, vol. 11, no. 21, p. 10007, 2021, doi: 10.3390/app112110007.
[14] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," International Journal of Educational Technology in Higher Education, vol. 16, no. 1, pp. 1–20, 2019, doi: 10.1186/s41239-019-0172-z.
[15] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
[18] M. Credé, S. G. Roch, and U. M. Kieszczynka, "Class Attendance in College: A Meta-Analytic Review of the Relationship of Class Attendance With Grades and Student Characteristics," Review of Educational Research, vol. 80, no. 2, pp. 272–295, 2010, doi: 10.3102/0034654310362998.
[19] K. Okano, J. R. Kaczmarzyk, N. Dave, J. D. E. Gabrieli, and J. C. Grossman, "Sleep quality, duration, and consistency are associated with better academic performance in college students," npj Science of Learning, vol. 4, p. 16, 2019, doi: 10.1038/s41539-019-0055-z.
[20] J. L. Vigdor, H. F. Ladd, and E. Martinez, "Scaling the digital divide: Home computer technology and student achievement," Economic Inquiry, vol. 52, no. 3, pp. 1103–1119, 2014, doi: 10.1111/ecin.12089.
[21] C. Callender, "The impact of term-time employment on higher education students' academic attainment and achievement," Journal of Education Policy, vol. 23, no. 4, pp. 359–377, 2008, doi: 10.1080/02680930801924490.
[22] C. Y. Tan, M. Lyu, and B. Peng, "Academic Benefits from Parental Involvement are Stratified by Parental Socioeconomic Status: A Meta-analysis," Parenting: Science and Practice, vol. 20, no. 4, pp. 241–287, 2020, doi: 10.1080/15295192.2019.1694836.
[23] D. Voyer and S. D. Voyer, "Gender differences in scholastic achievement: A meta-analysis," Psychological Bulletin, vol. 140, no. 4, pp. 1174–1204, 2014, doi: 10.1037/a0036620.
[24] Á. Balaguer, E. Benítez, A. Albertos, and S. Lara, "Not everything helps the same for everyone: relevance of extracurricular activities for academic achievement," Humanities and Social Sciences Communications, vol. 7, p. 79, 2020, doi: 10.1057/s41599-020-00573-0.
[25] R. H. H. Groenwold, I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman, and K. G. M. Moons, "Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis," CMAJ, vol. 184, no. 11, pp. 1265–1269, 2012, doi: 10.1503/cmaj.110977.
[26] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.