International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 10 Issue 04 December 2024
DOI: https://doi.org/10.69980/5nc7k059

A MULTIVARIATE STATISTICAL AND MACHINE LEARNING ANALYSIS OF FACTORS INFLUENCING UNIVERSITY STUDENT ACADEMIC PERFORMANCE

Dr. Elias M. Hartwell[1*], Prof. Sofia R. Lindström[2], Dr. Adrian K.Moretti[3]

[1*]*Department of Educational Data Science, Northbridge Institute of Applied Analytics, Toronto, Canada*
[2]*School of Computational Social Sciences, Westhaven University of Technology, Stockholm, Sweden*
[3]*Department of Statistics and Machine Learning, European Institute for Educational Research, Milan, Italy*

Abstract:
The present study aimed to explore the academic, socioeconomic, demographic, and behavioral determinants of the academic performance of university students through an integrated multivariate statistical and machine learning approach. Data from 493 university students and 16 variables were analysed with the continuous outcome being overall academic performance. Using seven supervised regression algorithms, R2, MAE, MSE, and RMSE were calculated to find significant relationships, which were then identified by descriptive statistics, correlation analysis, and group comparisons. Overall performance was most highly correlated with previous academic performance ($r = 0.925$, $p < .001$) and attendance ($\rho = 0.664$, $p < .001$), and moderately negatively associated with gaming ($\rho = -0.412$, $p < .001$). All of the variance in academic performance was accounted for by the multivariate regression model, and previous performance and preparation continued to be significant. For the machine-learning models, Support Vector Regression (SVR) gave the highest overall performance ($R^2 = 0.909$; RMSE = 0.184) followed by Random Forest (R2 = 0.908). Previous academic performance was also the most important predictor, as evidenced by the feature-importance analysis. From all these results, it is concluded that using hybrid of statistical inference and machine learning is an effective and interpretable method to predict student achievement and for early identification of academically at-risk students.

Keywords: *Academic performance; Machine learning; Multivariate analysis; Support Vector Regression; Educational data mining*

1. Introduction:

Student academic achievement is a key measure of success in a university or college since it demonstrates achievement of learning objectives, advancement, retention, and readiness for the next phase of employment. Reliable methods for understanding why some students perform better than others and methods for determining learners who might need to be provided with timely academic help are increasingly sought after by universities. Previous studies indicate that a student's academic performance is affected by academic history, the student's demographic characteristics, socioeconomic status, attendance, engagement, learning behavior, and extracurricular activities. Academic achievement, therefore, needs to be viewed as a multi-faceted phenomenon rather than the effect of a single determinant. Predictive models have also been shown to be useful for identifying students who are at risk of poor performance and for helping to inform more personalized educational interventions (Adejo & Connolly, 2018; Akçapınar et al., 2019).

Learning analytics, educational data mining and machine learning are increasingly being used in higher education, as more data becomes available for education. Machine-learning methods can uncover both nonlinear patterns and complex interactions that could be hard to recognize using only traditional statistical methods (Alsariera et al., 2022; Alshabandar et al., 2020). Academic-success prediction reviews acknowledged that applying correct data preprocessing, selecting the right predictors, evaluating the models, and providing transparency of the method are at the heart of achieving reliable results (Alyahyan & Düştegör, 2020; Cui et al., 2019). More recent applications also suggest that machine-learning models can be helpful in predicting student performance in a variety of educational contexts (Balcioğlu & Artar, 2025). The focus of data mining in education has thus shifted to exploring meaningful patterns in student data and guiding evidence-based academic decisions (Batool et al., 2023; Chaka, 2021).

Various analytical approaches are suggested to offer student outcome predictions. Associative classification has been effective for uncovering interpretable relationships between student characteristics and academic outcomes and traditional supervised algorithms have been found to be effective in predicting academic outcomes in educational datasets (Cagliero et al., 2021; Dabhade et al., 2021). Meta-analytic evidence also indicates that the use of machine learning can be a useful tool in predicting academic performance, student risk, and student attrition in higher education, albeit with performance varying based on the characteristics of the data, choice of machine learning algorithm and evaluation strategy (Fahd et al., 2022). Beyond the formal academic record, there are other factors that might also play a role in student achievement. For instance, extracurricular involvement can help foster students' overall growth and academic engagement and early indicators of involvement can offer useful insights into future academic performance (Ginosyan et al., 2020; Gray & Perkins, 2019). In a higher education context, too, learning analytics based on digital learning activities have been applied to predict students' results (R. Hasan et al., 2020).

Although there have been significant advances in predicting student performance, determining what factors are most important is still difficult because academic success is determined by multiple, interacting academic, socioeconomic, demographic, and behavioral factors. In many studies, emphasis is placed on prediction while failing to incorporate a proper statistical inference and machine-learning interpretation. This can hinder the ability to determine if a variable is statistically significant, practically significant or just for prediction purposes. Thus a combined multivariate statistical and machine-learning solution is required to analyse the correlations between student features, statistically determine the effects, compare prediction algorithms, and determine the most important student features affecting the academic performance of a university.

The purpose of this study is to explore some academic, socioeconomic, demographic and behavioral factors that are correlated with academic performance of university students. Specifically, it aims to use multivariate statistical methods to find statistically significant predictors, develop and compare machine-learning models that predict overall academic performance, use meaningful metrics to assess the performance of the models, and use model-based feature importance analysis to identify predictors that have the strongest influence. The study will combine statistical inference and predictive modeling to enable a more complete and interpretable understanding of the factors that affect student academic achievement.

2. Materials and Methods

2.1 Research Design

The study design used a quantitative data design using multivariate statistical analysis together with an assisted machine-learning approach. The integrated methodology was applied to determine factors that are linked to the academic performance of university students and also to assess predictive power of various algorithms. The overall academic performance was assumed to be a continuous outcome variable.

2.2 Dataset and Variables

There were 493 university students and 16 variables in the dataset (T. Hasan et al., 2024). The dependent variable was Overall which was the measure of overall academic performance. Gender, HSC, SSC, Income, Hometown, Computer Skills, Preparation, Gaming, Attendance, Job, English Skills, Extra Activities, Semester, Department, and previous academic performance were used as predictors (Last). These variables were demographic, socioeconomic, academic, skill-related and behavioral variables that were applicable to student achievement.

2.3 Data Preprocessing

This dataset was filtered for missing values, duplicates, inconsistent category labels and invalid values. No missing data or records were found. Irregularity in formatting in categorical variables, especially Income, was normalized. Binary and nominal variables were coded numerically whereas ordered categories were stored as per their logical order. Standardised numerical predictors were used where necessary where scale-sensitive algorithms including Support Vector Regression and K-Nearest Neighbors, were used. Descriptive statistics and graphical inspection were used to examine potential outliers.

2.4 Descriptive and Exploratory Analysis

The nature of the sample was summarized using descriptive statistics. The description of continuous variables was presented as a mean, standard deviation, minimum, and maximum, and the description of categorical variables was presented as frequencies and percentages. The graphical techniques that were applied to analyze the distribution of Overall performance were the relevant techniques. Correlation analysis was used to initially examine relationships between numerical predictors and academic performance with performance differences across the demographic and behavioral groups of interest also evaluated.

2.5 Multivariate Statistical Analysis

Pearson correlation was used when the variables were continuous and numerical, and Spearman correlation when the variables were ordinal and suitable. The comparisons of the academic performance of binary and multi-category groups were conducted through independent-samples t-tests and one-way ANOVA. This was followed by multiple linear regression to establish the independent contribution of the predictors to the overall academic performance. The measure of multicollinearity was done with the help of variance inflation factors, and the level of statistical significance was set at $p < 0.05$. The special focus was on Last due to the close connection with the Overall.

2.6 Machine-Learning Model Development

A number of supervised regression algorithms with supervision were designed to predict total academic performance. These were Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, XGBoost Regression, Support Vector Regression and K-Nearest Neighbors Regression. A baseline model was Linear Regression, and tree-based and nonlinear algorithms were also considered to be able to capture more complicated interactions between student characteristics.

2.7 Model Training and Evaluation

The dataset was divided into approximately 80% training data and 20% testing data. The training set was cross-validated to enhance stability of the model and minimize overfitting. The coefficient of determination (R2), mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) were used to assess predictive performance. The most successful model was selected with reference to the higher R and lower values of the prediction error.

2.8 Feature Importance and Model Interpretation

The feature-importance analysis was conducted to find out the most significant predictors of student academic performance. Random Forest and boosting-based model importance scores were checked and SHAP analysis of the best-performing model was taken into consideration to find the direction and the intensity of predictor impacts. Academic, socioeconomic and behavioral variables were then ranked based on their predictive contribution.

2.9 Robustness and Reproducibility

To enhance robustness, the development of all models was based on the same cleaned dataset and the same train-test partitions. Reproducibility was supported using cross-validation and fixed random seeds. Since the previous academic performance (Last) was significantly correlated with Overall, additional models were used that included this variable and the absence of it to determine whether it overshadowed the effect of other variables or not. The results of the departments were viewed with caution due to the significant imbalance in the dataset with regard to academic departments.

3. Results

3.1 Characteristics of the Student Sample

A total of 493 university students were analyzed. The sample was 66.5% male and 33.5% female. The majority of students were village residents (56.8%), 93.1% were not employed and 41.6% indicated their involvement in extracurricular activities. In terms of study behavior, 46.2% of them said that they prepared for between 2 and 3 hours, and 61.3% had more than three hours of gameplay. The attendance was also fairly high with 40.0% indicating 80-100% attendance and 44.0% indicating 60-79. The sample was also very concentrated in Computer Science and Engineering (89.9%), which means that comparisons based on departments should be viewed with care. The mean overall academic performance was 3.188/0.592 and the mean past academic performance (Last) was 3.164/0.641 shown in Table 1.

Table 1. Demographic, Academic, Socioeconomic, and Behavioral Characteristics of Students

Characteristic	Category/Statistic	n (%) or Mean ± SD
Sample size	Total	493
Gender	Male	328 (66.5%)
	Female	165 (33.5%)
Hometown	Village	280 (56.8%)
	City	213 (43.2%)
Income	Low	74 (15.0%)
	Lower middle	180 (36.5%)
	Upper middle	110 (22.3%)
	High	129 (26.2%)
Attendance	80–100%	197 (40.0%)
	60–79%	217 (44.0%)
Employment	Yes	34 (6.9%)

	No	459 (93.1%)
Extra activities	Yes	205 (41.6%)
Preparation	2–3 hours	228 (46.2%)
Gaming	>3 hours	302 (61.3%)
HSC	Mean ± SD	4.157 ± 0.547
SSC	Mean ± SD	4.768 ± 0.350
Previous performance (Last)	Mean ± SD	3.164 ± 0.641
Overall performance	Mean ± SD	3.188 ± 0.592

3.2 Distribution of Overall Academic Performance
The overall academic performance was between 1.00 and 4.00 with a mean of 3.188 and standard deviation of 0.592. As shown in Figure 1, the focus of observations was largely towards the top middle range of the performance scale, but also included lower-performing students. The distribution allowed for adequate variation to perform multivariate regression and continuous-outcome machine learning analysis.

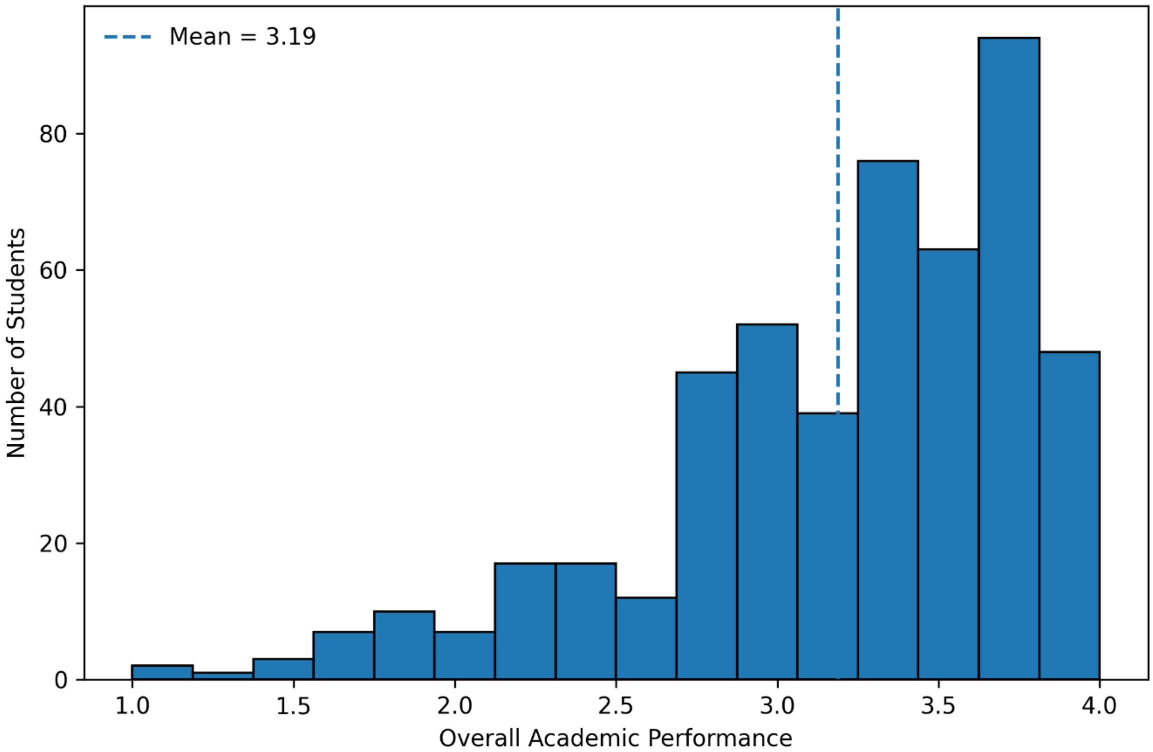

Figure 1. Distribution of Overall Academic Performance

3.3 Correlation and Group-Based Analysis
Several important relationships emerged with overall academic performance in bivariate analysis. The strongest positive correlation was with Overall ($r = 0.925$, $p < .001$), whereas previous academic performance had the highest positive correlation. Performance was also strongly positively associated with attendance ($\rho = 0.664$, $p < .001$) and preparation was moderately positively associated ($\rho = 0.377$, $p < .001$). The association between gaming and other measures was moderate and negative ($\rho = -0.412$, $p < .001$). A positive relation emerged between HSC and overall performance as well as between SSC and overall performance, but with considerably less strength than for previous academic performance.
There was a higher mean score for females than males (3.322 vs. 3.121; $p < .001$), and a higher mean score for students involved in extra-curricular activities than those who were not (3.408 vs. 3.032; $p < .001$). There was no significant association of overall performance with hometown or employment or income status. There were significant differences found between attendance, preparation and gaming categories shown in Table 2.

Table 2. Associations Between Student Factors and Overall Academic Performance

Predictor	Statistical Test	Statistic	p-value
HSC	Pearson r	0.252	<.001
SSC	Pearson r	0.099	.028
Previous performance (Last)	Pearson r	0.925	<.001
Computer skills	Pearson r	0.178	<.001
English skills	Pearson r	0.030	.506
Income	Spearman ρ	−0.021	.643
Preparation	Spearman ρ	0.377	<.001

Gaming	Spearman ρ	−0.412	<.001
Attendance	Spearman ρ	0.664	<.001
Semester	Spearman ρ	−0.024	.591
Gender	Welch t	−3.879	<.001
Hometown	Welch t	−1.164	.245
Employment	Welch t	0.503	.618
Extra activities	Welch t	7.385	<.001

3.4 Multivariate Correlation Structure

The multivariate correlation structure of the significant numerical and ordinal predictors is shown in Figure 2. Previous academic performance and overall academic performance had the strongest relationship (0.93). There were positive correlations between attendance and past and total performance but gaming had negative correlations with performance. Academic achievement and attendance had positive relationships with preparation. Conversely, the direct relationships between income, semester, and English skill and overall performance were also weak.

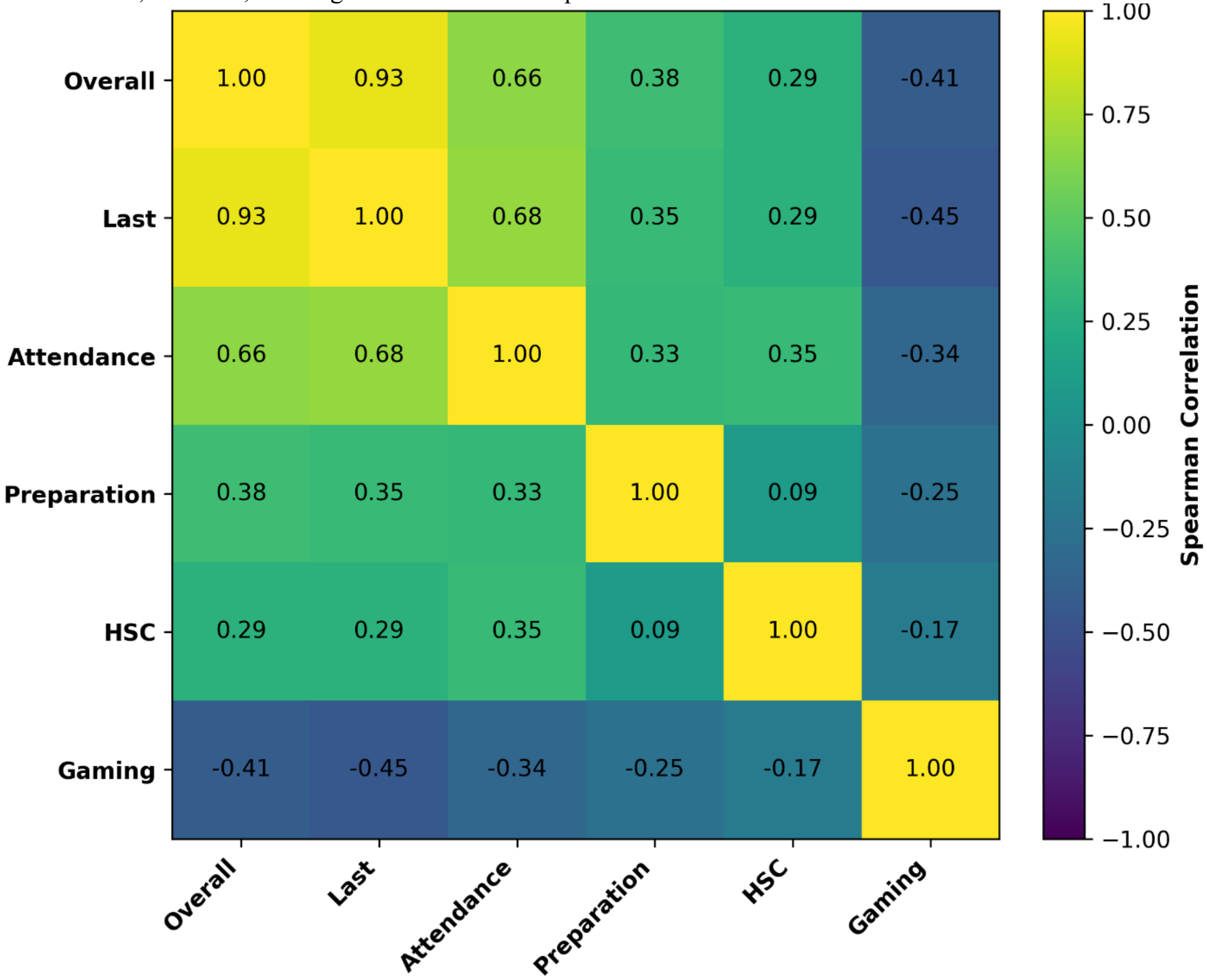

Figure 2. Correlation Matrix of Academic and Behavioral Variables

3.5 Multivariate Regression Results

The multiple linear regression model was able to explain 86.6 percent of the variance in overall academic performance (R2 = 0.866; adjusted R2 = 0.859). Previous academic performance turned out to be the most powerful independent predictor ($B = 0.812$, $p < .001$), meaning that a one-unit increase in previous academic performance was linked to an average improvement of about 0.81 units in overall performance after holding the rest of the variables constant. The preparation time was found to be statistically significant ($B = 0.040$, $p = .032$).

HSC, SSC, computer skills, English skills, income, gaming, attendance, semester, gender, hometown, employment, and extracurricular participation failed to maintain independent statistical significance after the previous performance and the other variables were adjusted. This result indicates that a number of the robust bivariate relationships, especially attendance and gaming, were overlapping each other significantly with students' previous academic success. The multicollinearity was tolerated among the key predictors, and VIF values were mostly lower than 2.5 shows in Table 3.

Table 3. Multiple Linear Regression Predicting Overall Academic Performance

Predictor	B	SE	95% CI	p-value

HSC	0.023	0.023	−0.022, 0.069	.318
SSC	−0.008	0.039	−0.084, 0.069	.843
Computer skills	0.012	0.009	−0.005, 0.029	.174
English skills	−0.008	0.015	−0.037, 0.022	.614
Previous performance (Last)	0.812	0.049	0.716, 0.908	<.001
Income	0.005	0.009	−0.012, 0.023	.561
Preparation	0.040	0.019	0.003, 0.077	.032
Gaming	0.005	0.020	−0.034, 0.043	.812
Attendance	0.017	0.028	−0.038, 0.072	.550
Semester	−0.001	0.004	−0.008, 0.006	.760
Female gender	0.030	0.023	−0.015, 0.075	.197
City hometown	−0.015	0.024	−0.061, 0.032	.533
Employment	0.046	0.068	−0.087, 0.179	.496
Extra activities	0.026	0.027	−0.026, 0.078	.335

Model $R^2 = 0.866$; adjusted $R^2 = 0.859$. Department indicators were included as control variables.

3.6 Predicted Versus Actual Academic Performance

A comparison of the predicted and actual results of the best machine learning model is shown in Figure 3. The predictions of Support Vector Regression (SVR) were very close to the line of perfect agreement over the range of observed performances. The fact that observations were grouped near the line of agreement suggests that the predictive agreement was strong; however, there were a few instances of students at the extremes of the academic-performance distribution.

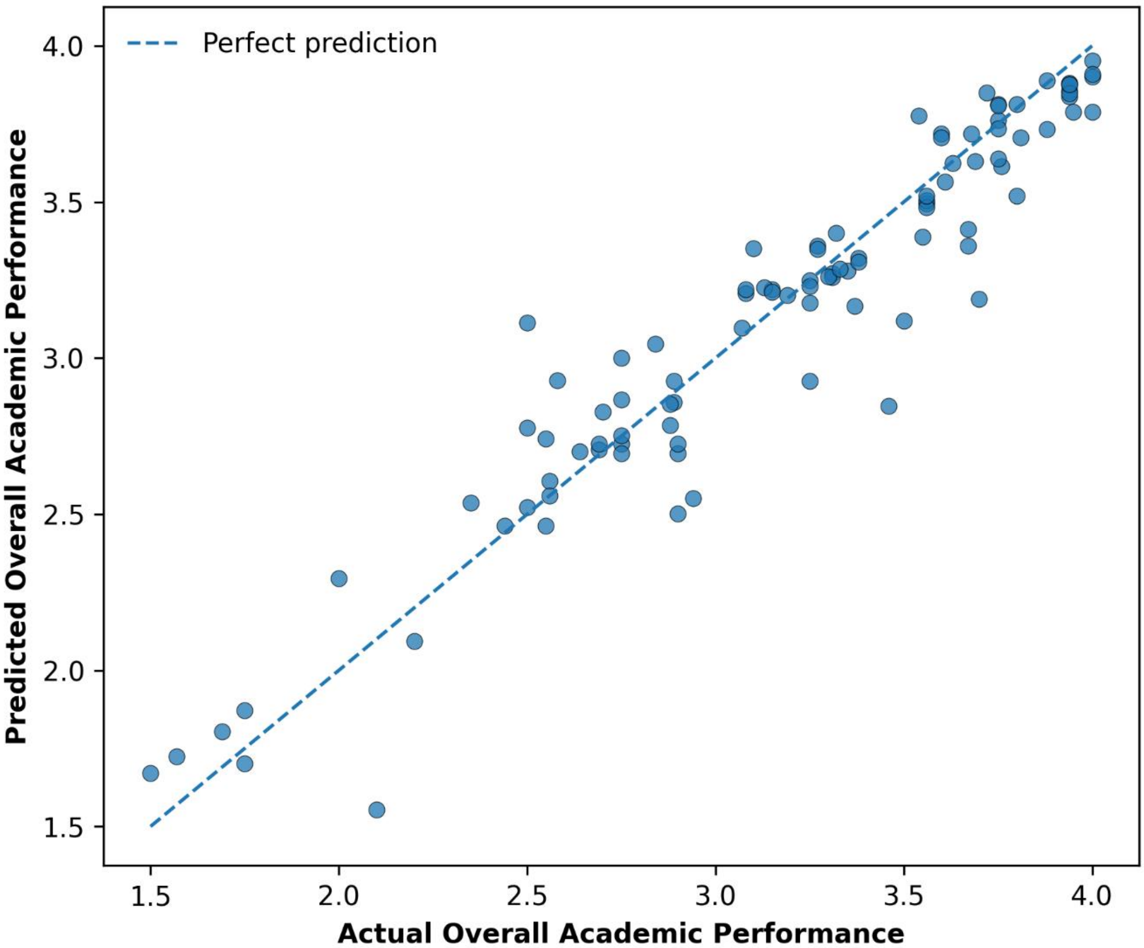

Figure 3. Predicted Versus Actual Overall Academic Performance Using Support Vector Regression

3.7 Machine-Learning Model Performance

All seven regression algorithms could be shown to have varying predictive accuracy. The model with the highest R2 value (0.909) and the lowest RMSE value (0.184) was the SVR model, which means that SVR was the best model overall. Random Forest has an almost similar R2 of 0.908 and the lowest MAE of 0.114. Gradient Boosting also did well with an R^2 of 0.901. The R^2 values for XGBoost and Decision Tree models were close at around 0.887, while Linear Regression and KNN models had relatively lower predictive power shown in Table 4.

Table 4. Comparative Performance of Machine-Learning Models

Model	R^2	MAE	MSE	RMSE
Support Vector Regression	0.909	0.130	0.034	0.184
Random Forest	0.908	0.114	0.035	0.186
Gradient Boosting	0.901	0.140	0.037	0.192
XGBoost	0.887	0.143	0.042	0.205
Decision Tree	0.887	0.145	0.042	0.205
Linear Regression	0.851	0.172	0.056	0.236
K-Nearest Neighbors	0.803	0.205	0.074	0.272

The very similar performance of SVR and Random Forest indicates that both nonlinear approaches effectively captured the predictive structure of the dataset, although SVR provided the best overall R^2 and RMSE.

3.8 Important Predictors and Model Interpretation

The variables that were the most important predictors from the best-performing SVR model were confirmed by the permutation-based interpretation of the model. Attendance was the second most important item, followed by semester, SSC, computer skills, HSC, and gaming, although they weren't as large as Last's. The regression results are thus confirmed in Figure 4, which also indicates that the variables that were not a significant factor in the linear model may add value to the prediction in a non-linear model.

Overall, the results of the statistical analysis combined with the results of the machine learning analysis suggest that academic history was the main factor in terms of performance. Other variables such as attendance, preparation, gaming behavior, prior school performance, and computer skills added more information, while income, employment, hometown, and English skills added relatively little once other variables were taken into account. The results highlight the importance of using multivariate inference along with predictive modeling in the study of academic achievement of university students.

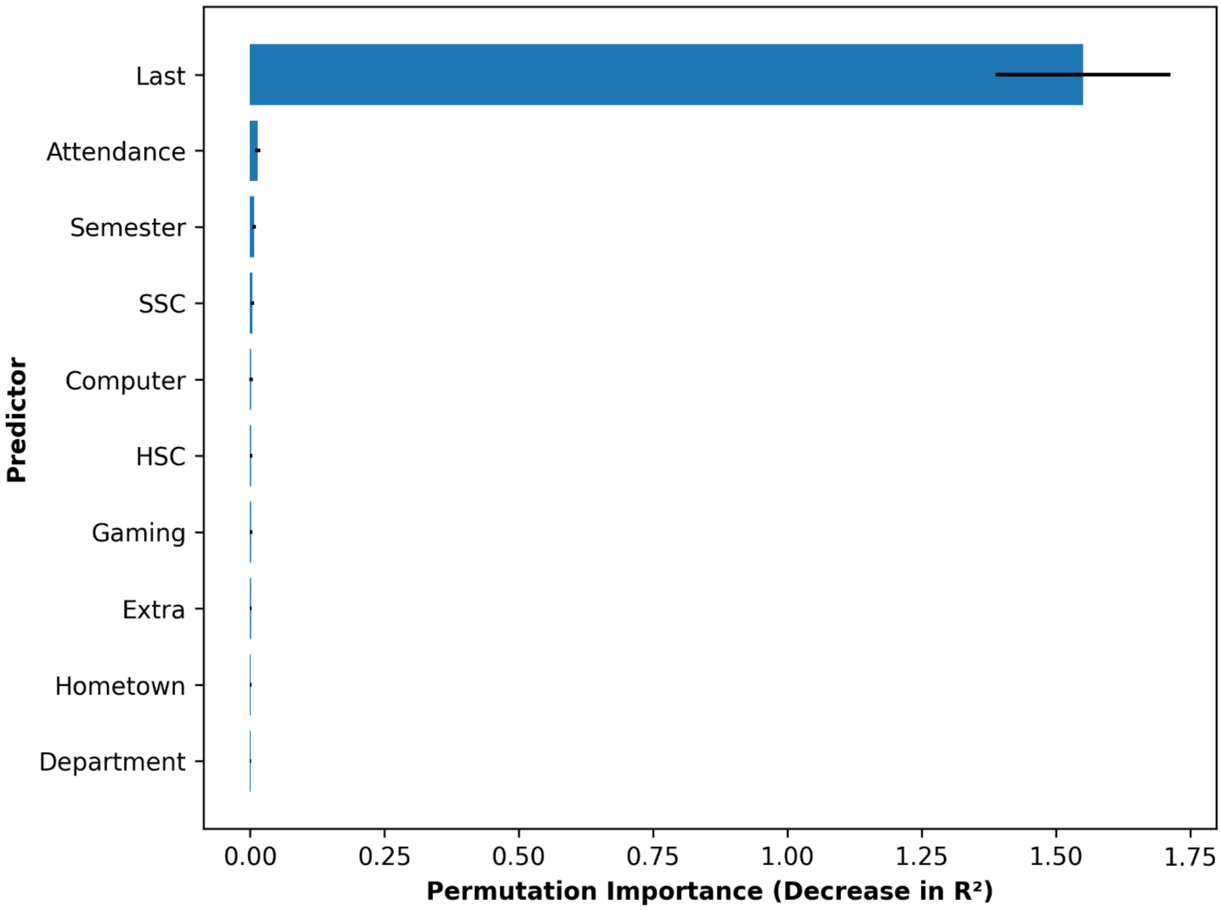

Figure 4. Feature Importance for Prediction of Overall Academic Performance

4. Discussion

The study revealed that a combined multivariate statistical and machine learning approach can be used to effectively analyze university student academic performance. Overall performance was most strongly predicted by previous academic performance (Last) and attendance, preparation, HSC, SSC, gaming, and selected behavioral characteristics were found to be significant. The regression model accounted for a good amount of the variance, and of the machine-learning models examined Support Vector Regression performed best in terms of predictive accuracy. The results align with the increasing adoption of data-driven methods in the identification of patterns of academic risk in higher education (Rastrollo-Guerrero et al., 2020; Tomasevic et al., 2020).

In general, there was a strong continuity of students' academic trajectories as evidenced in the positive correlation between previous academic performance and overall achievement. There was also a strong positive correlation between attendance and performance, and a positive correlation between preparation and performance. HSC score and achievement showed a positive relationship, but this relationship was less pronounced when other more current measures of achievement were included. Similarly, SSC scores were positively related to achievement, and this relationship was weaker when other more recent measures of achievement were introduced. The results of similar studies indicate that past academic performances and learning trajectories are among the most predictive predictors of achievement in the future (Musso et al., 2020; Trakunphutthirak & Lee, 2022). Artificial neural-network studies have also highlighted the need to comprehensively analyze previous achievement variables when predicting student outcomes (Rodríguez-Hernández et al., 2021).

At the bivariate level, gaming was negatively correlated with overall academic performance while extracurricular activities were positively associated with academic performance. Other variables controlled for included income, employment status and hometown, which had comparatively weaker effects. The findings suggest that the socioeconomic factors might have an indirect effect on achievement due to their relationship with learning opportunities, engagement, and academic resources of the student, not simply as independent variables. There is a significant, but context-specific, relationship between SES and higher education attainment as indicated in previous research (Rodríguez-Hernández et al., 2020; Selvitopu & Kaya, 2023). Behavioral and emotional factors can also impact academic outcomes, such as student engagement and learning interaction (Kukkar et al., 2023).

The overall predictive performance of the model was best for Support Vector Regression with R^2 of 0.909 and RMSE of 0.184, followed closely by the Random Forest model. The results showed that the nonlinear model had greater accuracy in depicting the complex relationships between academic and behavioral predictors than the linear model. The results of previous comparison studies also show that student performance prediction can significantly benefit from incorporating multiple interacting variables, with the help of supervised machine-learning methods (Tomasevic et al., 2020; Wang & Yu, 2025). Advanced neural and transfer-learning techniques have also demonstrated good potential in educational prediction tasks (Ljubičić & Hell, 2023; Tsiakmaki et al., 2020).

Feature-importance analysis confirmed that previous academic performance was the most important predictor followed by attendance and selected academic and behavioral factors. This regression/machine-learning interpretation agreement further bolsters confidence in the robustness of the findings. In the educational context, explainable machine-learning methods can help take prediction models beyond accuracy to provide meaningful interpretations of student-risk factors (Jang et al., 2022). Previous prediction research using LMS data has also shown that meaningful and understandable measures of student engagement can aid in timely academic intervention (Santos & Henriques, 2023).

The present results are consistent with previous research on the role of academic history, engagement patterns and behavioral information in predicting student performance. Machine-learning methods have shown their effectiveness in learning from complex relationships throughout learners' education and in early prediction of learners' outcomes (Musso et al., 2020; Rastrollo-Guerrero et al., 2020). This evidence is further developed in the present study by introducing multivariate statistical inference and predictive modeling within the same analytical framework.

The results are of practical significance for the university that wants to detect students at academic risk. Targeted monitoring can be done based on previous performance and attendance, and preparation and behavioral variables can provide for personalized academic counselling. Explainable predictive systems can be useful for facilitating proper allocation of academic-support resources and for creating interventions that can be implemented in time, before a significant decrease in performance occurs (Jang et al., 2022; Santos & Henriques, 2023).

The study has limitations due to the relatively small sample size, moderate departmental imbalance, Data design, and the lack of detailed psychological, institutional, and learning environment variables. Such limitations might limit generalizability. Larger multi-university datasets and a holistic integration of psychological and digital-learning indicators would benefit future research, as would model validation in other institutional environments. Higher-order machine learning techniques, such as explainable models, and longitudinal techniques could enhance the reliability and applicability of the predictions.

5. Conclusion

Multivariate statistical analysis together with machine-learning methods were found to be effective tools for the analysis of the academic performance of the university students in this study. The results revealed that previous academic performance was the most significant factor in overall achievement with attendance, preparation, HSC, SSC, gaming behaviour, and selected student characteristics also accounting for differences in achievement. Multivariate regression model accounted for a significant amount of variance in academic performance, indicating the significance of prior achievement and preparation given the other variables. The machine-learning algorithms evaluated were Support Vector Regression (SVR) and Random Forest (RF), with SVR performing best overall in predicting the machine's performance.Support Vector Regression (SVR) and Random Forest (RF) had comparable predictive performance among the machine-learning algorithms evaluated, and both algorithms performed the best overall predictions of the machine's performance. The results of the feature importance analysis also showed that previous academic performance was the most important feature with the remaining academic and behavioral features having a lesser impact. Other factors such as income, employment status, hometown, and some skill-related factors had relatively less independent effect. Combining statistical inference with predictive modelling gave an overall perspective on the significance and predictive role of student-level factors. The results of this research can also help universities to identify students who are at risk of poor performance and to design their academic-support measures accordingly. These results need to be confirmed in future studies with larger, more balanced, multi-institutional datasets, incorporating psychological and institutional and digital-learning variables.

References

1. Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61–75.
2. Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16(1), 40. https://doi.org/10.1186/s41239-019-0172-z
3. Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. Computational Intelligence and Neuroscience, 2022, 1–11. https://doi.org/10.1155/2022/4151487
4. Alshabandar, R., Hussain, A., Keight, R., & Khan, W. (2020). Students performance prediction in online courses using machine learning algorithms. 2020 International Joint Conference on Neural Networks (IJCNN), 1–7. https://ieeexplore.ieee.org/abstract/document/9207196/
5. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 3. https://doi.org/10.1186/s41239-020-0177-7
6. Balcioğlu, Y. S., & Artar, M. (2025). Predicting academic performance of students with machine learning. Information Development, 41(3), 896–915. https://doi.org/10.1177/02666669231213023
7. Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y
8. Cagliero, L., Canale, L., Farinetti, L., Baralis, E., & Venuto, E. (2021). Predicting student academic performance by means of associative classification. Applied Sciences, 11(4), 1420.
9. Chaka, C. (2021). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews. https://www.preprints.org/frontend/manuscript/2483a8ecb51b982fec76b4be2859ace7/download_pub
10. Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. Information and Learning Sciences, 120(3–4), 208–227.
11. Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. Materials Today: Proceedings, 47, 5260–5267.
12. Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies, 27(3), 3743–3775. https://doi.org/10.1007/s10639-021-10741-7
13. Ginosyan, H., Tuzlukova, V., & Ahmed, F. (2020). An investigation into the role of extracurricular activities in supporting and enhancing students' academic performance in tertiary foundation programs in Oman. Theory and Practice in Language Studies, 10(12), 1528–1534.
14. Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. Computers & Education, 131, 22–32.
15. Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Applied Sciences, 10(11), 3894.
16. Hasan, T., Hasan, M. M., & Manzoor, T. (2024). Student Performance Metrics Dataset. 1. https://doi.org/10.17632/5b82ytz489.1
17. Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. Education and Information Technologies, 27(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6
18. Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2023). Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms. Education and Information Technologies, 28(8), 9655–9684. https://doi.org/10.1007/s10639-022-11573-9
19. Ljubičić, T., & Hell, M. (2023). Predikcija uspješnosti studenata primjenom umjetnih neuronskih mreža. Ekonomska Misao i Praksa, 32(2), 361–374.
20. Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. Higher Education, 80(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7
21. Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. Applied Sciences, 10(3), 1042.
22. Rodríguez-Hernández, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. Educational Research Review, 29, 100305.

23. Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence, 2, 100018.
24. Santos, R. M., & Henriques, R. (2023). Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs. Computers and Education: Artificial Intelligence, 5, 100175.
25. Selvitopu, A., & Kaya, M. (2023). A Meta-Analytic Review of the Effect of Socioeconomic Status on Academic Performance. Journal of Education, 203(4), 768–780. https://doi.org/10.1177/00220574211031978
26. Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education, 143, 103676.
27. Trakunphutthirak, R., & Lee, V. C. S. (2022). Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data. Journal of Educational Computing Research, 60(3), 742–776. https://doi.org/10.1177/07356331211048777
28. Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer learning from deep neural networks for predicting student performance. Applied Sciences, 10(6), 2145.
29. Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. PloS One, 20(1), e0299018.