JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
JSLHU

ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, 2026, 27, 029-038

ỨNG DỤNG MACHINE LEARNING TRONG DỰ BÁO KẾT QUẢ HỌC TẬP CỦA HỌC SINH TRUNG HỌC CƠ SỞ

Nguyễn Thị Ngọc Diễm, Phạm Nguyễn Khánh An, Nguyễn Thị Phương Anh, Lê Phương Long*
Trường Đại học Lạc Hồng, số 10, Đường Huỳnh Văn Nghệ, Phường Trấn Biên, Thành phố Đồng Nai, Việt Nam
*Tác giả liên hệ: phuonglong@lhu.edu.vn

THÔNG TIN BÀI BÁO

Ngày nhận:	05/04/2026
Ngày hoàn thiện:	03/06/2026
Ngày chấp nhận:	03/08/2026
Ngày đăng:	25/08/2026

TỪ KHÓA

Học máy;
Kết quả học tập;
Học sinh cấp THCS;
Dự báo kết quả học tập;
Yếu tố ảnh hưởng.

TÓM TẮT

Nghiên cứu này phân tích các yếu tố ảnh hưởng đến kết quả học tập của học sinh trung học cơ sở và phát triển ứng dụng dự báo dựa trên Machine Learning. Dữ liệu gồm 957 quan sát hợp lệ được thu thập từ học sinh tại tỉnh Đồng Nai, Việt Nam. Kết quả kiểm định cho thấy các thang đo đạt độ tin cậy chấp nhận được, với Cronbach's Alpha từ 0,602 đến 0,851. Phân tích nhân tố khám phá cho thấy dữ liệu phù hợp với KMO = 0,915 và $p < 0,001$, trích xuất được bảy nhân tố đại diện. Kết quả hồi quy cho thấy mô hình giải thích 53,4% sự biến thiên của kết quả học tập, trong đó nhân tố nhận thức - thái độ - cảm xúc có ảnh hưởng mạnh nhất. Trong số chín mô hình dự báo, các mô hình phi tuyến cho hiệu quả tốt hơn mô hình tuyến tính. Random Forest và XGBoost đạt $R^2 = 0,951$, trong khi SVR đạt $R^2 = 0,899$ và RMSE = 0,456. SVR được chọn để triển khai nhờ cân bằng giữa độ chính xác, tính ổn định và khả năng ứng dụng thực tiễn. Hệ thống web demo cho phép học sinh hoàn thành khảo sát trực tuyến, nhận kết quả dự báo và khuyến nghị học tập phù hợp. Kết quả cho thấy tính khả thi của việc ứng dụng Machine Learning trong hỗ trợ học tập ở cấp trung học cơ sở.

MACHINE LEARNING FOR PREDICTING THE ACADEMIC PERFORMANCE OF LOWER SECONDARY SCHOOL STUDENTS

Nguyen Thi Ngoc Diem, Pham Nguyen Khanh An, Nguyen Thi Phuong Anh, Le Phuong Long*
Lac Hong University, No.10, Huynh Van Nghe St., Tran Bien ward, Dong Nai City, Vietnam
*Corresponding Author: phuonglong@lhu.edu.vn

ARTICLE INFO

Received:	Apr 5th, 2026
Revised:	Jun 3rd, 2026
Accepted:	Aug 3rd, 2026
Published:	Aug 25th, 2026

KEYWORDS

Machine Learning;
Academic performance;
Lower secondary students;
Prediction;
Influencing factors.

ABSTRACT

This study analyzes factors influencing academic performance among lower secondary school students and develops a Machine Learning-based predictive application. The dataset included 957 valid observations collected from students in Dong Nai Province, Vietnam. Reliability testing showed acceptable internal consistency, with Cronbach's Alpha values ranging from 0.602 to 0.851. Exploratory Factor Analysis confirmed data suitability, with KMO = 0.915 and $p < 0.001$, and extracted seven representative factors. The regression results showed that the model explained 53.4% of the variance in academic performance, with the cognition-attitude-emotion factor having the strongest influence. Among nine predictive models, nonlinear models outperformed linear models. Random Forest and XGBoost achieved the highest R^2 value of 0.951, while SVR achieved $R^2 = 0.899$ and RMSE = 0.456. SVR was selected for deployment because of its balance between predictive accuracy, stability, and practical applicability. A web-based demo system was then developed to allow students to complete an online survey, receive predicted academic performance, and obtain learning recommendations. The findings demonstrate the feasibility of applying Machine Learning to academic support systems at the lower secondary level.

Doi:
Available online at: https://lhj.vn

1. Introduction

Analyzing students' academic performance plays a central role in teaching practice and educational management, particularly at the lower secondary level, where students begin to establish relatively stable study habits, learning orientations, and motivational patterns. At this stage, academic outcomes not only reflect students' attainment of curriculum standards but also indicate their ability to adapt to instructional approaches, learning environments, and support conditions provided by families and schools.

In practice, students' academic performance is shaped by multiple interacting factors, including individual characteristics, family socio-economic conditions, levels of engagement, and the organization of teaching activities within schools. Such relationships often involve nonlinear dynamics and complex dependencies that cannot be sufficiently addressed through classical statistical modeling techniques. This situation calls for more advanced analytical approaches capable of modeling intricate patterns and uncovering the relative contribution of different factors.

Machine Learning (ML) has emerged as a promising approach in educational data analysis due to its ability to handle large datasets and model complex, nonlinear relationships among variables. Prior studies have shown that ML models can achieve strong predictive performance in forecasting academic outcomes and identifying students at risk of underperformance. However, much of the existing research focuses on university or upper secondary students, while lower secondary students remain comparatively underexplored, despite the critical role this stage plays in long-term academic development.

In response to this practical need and research gap, the present study applies ML techniques to analyze factors influencing academic performance among lower secondary school students using educational survey data. The study develops and compares multiple ML models to identify significant predictors of academic outcomes and to assess the feasibility of applying ML in supporting teaching and educational management at the lower secondary level. The findings are expected to provide empirical evidence for designing targeted educational interventions aligned with students' developmental characteristics and learning needs.

2. Research Content

2.1 Theoretical Background on Academic Performance of Lower Secondary School Students

Academic performance at the lower secondary level refers to the extent to which students meet the required knowledge and skill standards defined in the general education curriculum. According to the 2019 Education Law of Vietnam, academic performance serves as an essential basis for evaluating students' learning and development within schools [5]. At this level, academic outcomes are typically reflected through grade point averages, academic classifications, and the degree to which students complete subject-specific learning tasks.

Circular No. 22/2021/TT-BGDDT stipulates that student assessment at the lower secondary level should combine formative and summative evaluation in order to provide a comprehensive view of students' progress throughout the learning process. This framework indicates that academic performance is not merely an end result but is closely linked to ongoing learning engagement and development.

Recent studies suggest that academic performance is influenced by multiple interacting factors, and the relationships among these factors are often nonlinear. An analysis of PISA 2022 data revealed that Vietnamese students' academic competence is associated with individual characteristics, family socio-economic conditions, and learning environments [3]. Several domestic studies have also demonstrated that ML models are capable of capturing complex relationships between input variables and academic outcomes, thereby achieving high predictive accuracy [7], [13].

2.2. Factors Influencing Academic Performance of Lower Secondary School Students

At the lower secondary stage, students' academic outcomes are influenced by both internal characteristics and external contextual conditions. This classification aligns with the cognitive and emotional developmental characteristics of students at this stage.

2.2.1. Internal Factors (Student-Related)

This group reflects students' psychological characteristics and learning behaviors, comprising four key dimensions:

- Perceived learning competence: students' level of self-confidence, self-evaluation ability, and belief in their capacity to accomplish academic tasks.
- Learning attitude: the degree of seriousness, proactiveness, and responsibility demonstrated in the learning process.
- Learning-related emotions: feelings of interest, anxiety, or pressure experienced during academic activities.
- Learning motivation: personal goals, intrinsic motivation, and persistence in striving for achievement.

These factors play a significant role in shaping students' learning engagement and their overall effectiveness in knowledge acquisition.

2.2.2. External Factors (Family and School Context)

External factors reflect the broader learning environment and support conditions, including:

- Socio-economic conditions: access to learning devices, educational resources, and a supportive study environment.
- Family academic support: the extent of parental attention, encouragement, and assistance when students encounter academic difficulties.
- School instructional organization: teaching methods, feedback from teachers, and the presence of a positive learning climate.

These contextual factors interact with students' personal characteristics and collectively influence academic outcomes to varying degrees.

2.3. Application of Machine Learning in Educational Data Analysis

The development of learning management and assessment systems in schools has generated large volumes of data related to students' behaviors, personal characteristics, and academic outcomes. Within this context, Machine Learning (ML) has emerged as an effective tool for mining educational data and modeling complex relationships between input variables and academic performance.

International research has highlighted the role of Educational Data Mining and Learning Analytics in predicting academic outcomes and analyzing influencing [2], [11]. Papamitsiou and Economides (2014) further demonstrated that ML algorithms are capable of identifying nonlinear relationships and providing empirical evidence to support decision-making in educational settings [10].

In Vietnam, several studies have begun applying ML techniques to predict academic performance and identify key determinants. ML models can improve predictive accuracy compared to traditional statistical approaches [7]. The interpretability of ML models in analyzing PISA data through explanation techniques such as SHAP [3].

Overall, ML not only enhances predictive capability but also helps clarify the relative contribution of individual and contextual factors, thereby supporting data-driven educational decision-making.

Although several studies have applied machine learning to predict students' academic performance, most of them have been conducted in international contexts. In Vietnam, related studies have mainly focused on university students, while research on lower secondary school students remains limited. This shows a clear research gap, especially because students at the lower secondary level are still developing their learning habits, attitudes, motivation, and academic behavior.

In addition, many previous studies mainly stop at analyzing influencing factors or comparing prediction models. They have not fully demonstrated how the results can be developed into an automatic prediction system that allows students to complete a survey, receive predicted academic performance, and obtain suitable learning recommendations.

To address these gaps, this study uses survey data collected from lower secondary school students currently studying in Dong Nai Province, Vietnam. The collected data are first analyzed through reliability testing, exploratory factor analysis, and regression analysis to identify key factors influencing academic performance. After that, machine learning models are trained and compared to select a suitable predictive model. Finally, the selected model is deployed in a web-based demo system that automatically predicts students' academic performance after they complete the survey and provides corresponding learning recommendations.

The novelty of this study does not lie in proposing a new machine learning algorithm. Instead, it lies in the use of a survey dataset collected from lower secondary school students in Dong Nai Province and in the integration of statistical analysis, machine learning prediction, and a web-based demo application for practical academic support.

3. Research Model

The research model was developed to examine the relationship between influencing factors and academic performance among lower secondary school students using Machine Learning techniques. Academic performance was defined as the dependent variable, while the independent variables comprised two main groups: internal factors (perceived competence, learning attitude, learning-related emotions, and learning motivation) and external factors (family support, socio-economic conditions, and school instructional organization).

The survey dataset was employed to train and assess various models, with the objective of identifying the most robust predictive framework. In addition to evaluating predictive performance, the analysis also aimed to assess the relative contribution of each factor to students' academic outcomes.

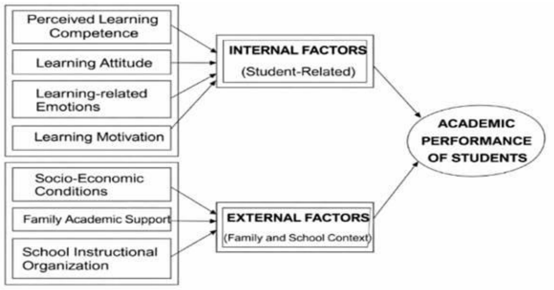

Figure 1: Research Model

4. Research Results

4.1. Data Description

After data cleaning and coding, the final dataset consisted of 957 valid observations collected from lower secondary school students in Dong Nai Province, Vietnam. The dataset included 29 analytical variables, comprising one dependent variable, Academic Performance, and independent variables grouped into seven theoretical factors: perceived learning competence, learning attitude, learning-related emotions, learning motivation, socio-economic conditions, family academic support, and school instructional organization.

The demographic characteristics of the respondents are presented in Table 1. In terms of gender, the sample was relatively balanced, with male students accounting for 51.20% and female students accounting for 48.80%. Regarding grade level, Grade 8 students represented the largest proportion of the sample at 39.81%, followed by Grade 7 students at 24.35%, Grade 6 students at 18.81%, and Grade 9 students at 17.03%. For the most recent semester GPA, the largest group of students reported scores from 8.0 to 8.9, accounting for 35.01%, followed by the 6.5 to 7.9 group at 25.81% and the 9.0 to 10.0 group at 20.48%.

Table 1: Demographic Characteristics of Respondents

Variable	**Category**	**Frequency (N)**	**Percentage (%)**
Gender	Male	490	51.20
	Female	467	48.80
Grade level	Grade 8	381	39.81
	Grade 7	233	24.35
	Grade 6	180	18.81
	Grade 9	163	17.03
Most recent semester GPA	9.0-10.0	196	20.48
	8.0-8.9	335	35.01
	6.5-7.9	247	25.81
	5.0-6.4	63	6.58
	Below 5.0	116	12.12
Total		**957**	**100.00**

4.2. Reliability Analysis of the Scales

Reliability was assessed using Cronbach's Alpha. The results indicate that all scales reached an acceptable level or higher.

Table 2: Reliability Analysis Results of Measurement Scales

Factor Group	**Code**	**Number of Items**	**Cronbach's Alpha**
Perceived Learning Competence	NT	4	0.851
Learning Attitude	TD	4	0.768
Learning-related Emotions	CX	4	0.715
Learning Motivation	DC	4	0.710
Socio-Economic Conditions	DK	3	0.602
Family Academic Support	SHT	3	0.702
School Instructional Organization	TCDH	4	0.796

Table 2 presents the reliability analysis results of the measurement scales. The Cronbach's Alpha coefficients ranged from 0.602 to 0.851, indicating that all constructs achieved an acceptable level of internal consistency. Perceived Learning Competence recorded the highest coefficient (α = 0.851), followed by School Instructional Organization (α = 0.796) and Learning Attitude (α = 0.768). Although Socio-Economic Conditions had the lowest coefficient (α = 0.602), it still met the acceptable threshold for exploratory research. Therefore, all constructs were retained for subsequent exploratory factor analysis.

4.3. Exploratory Factor Analysis

EFA was conducted to determine the underlying structure of the measurement scales and to examine the convergence of observed variables before being used in subsequent analytical models.

Data suitability was assessed using the KMO and Bartlett's tests.

Table 3: KMO and Bartlett's Test

	KMO	**Bartlett's Test (Chi-square)**	**Sig. (p-value)**
Value	0.915	Significant	0.000

The KMO statistic of 0.915 exceeds the recommended threshold of 0.90, indicating excellent sampling adequacy and strong intercorrelations among variables. In addition, Bartlett's Test of Sphericity was statistically significant ($p < 0.001$), demonstrating that the correlation matrix is not an identity matrix and that the data are suitable for factor extraction.

The EFA procedure resulted in the extraction of seven distinct factors. To ensure clarity and interpretability, Table 4 reports only those observed variables with factor loadings equal to or greater than 0.50 for each retained factor.

Table 4: Extracted Factors and Representative Factor Loading

Factor	**Converged Variables / Construct Codes**	**Interpretation**	**Factor Loading Range**
Factor 1	NT-TD-CX+	Positive cognition-attitude-emotion	0.56 - 0.77
Factor 2	DC, TCDH	Intrinsic motivation and pedagogical support	0.51 - 0.67
Factor 3	SHT	Family academic support	0.39 - 0.72
Factor 4	CX-	Negative emotions (anxiety, pressure)	0.58 - 0.77
Factor 5	TCDH	School instructional organization	0.55 - 0.72
Factor 6	DK	Learning material conditions	0.92
Factor 7	TD (discipline)	Discipline and task completion	0.65

The results indicate that:

· Factor 1 strongly converges NT-TD-CX+ variables, reflecting proactive learning tendencies, confidence in personal competence, and positive attitudes toward learning.

· Factor 4 separates CX- variables (e.g., anxiety and academic pressure), suggesting that negative emotions form a distinct psychological construct.

· Factor 6 shows a very high loading (0.92) on the learning equipment variable within DK, highlighting the role of material conditions in the learning context.

· Variables in SHT and TCDH form relatively independent factors, reflecting a clear distinction between family environment and school environment.

Overall, the extracted factor structure aligns with the theoretical framework comprising two major groups: internal factors (NT, TD, CX, DC) and external factors (SHT, DK, TCDH). The extracted factor scores were calculated and subsequently incorporated into both the regression analysis and the Machine Learning models in the subsequent stages of the study.

4.4. Multiple Linear Regression (OLS)

After extracting seven factors from EFA, a multiple linear regression model was constructed to examine the impact of these factors on Academic Performance (Score).

Table 5: Results of the Ordinary Least Squares (OLS) Regression Model

	Value
N	957
R^2	0.534
Adjusted R^2	0.531
F-statistic	155.5

Prob (F-statistic)	1.06e-152
Durbin-Watson	1.728

The coefficient of determination ($R^2 = 0.534$) indicates that the model accounts for approximately 53.4% of the total variance in academic performance. The F-statistic is statistically significant ($p < 0.001$), demonstrating that the overall regression model provides a meaningful explanation of the dependent variable.

Moreover, the Durbin-Watson statistic (1.728) is close to the reference value of 2, suggesting that the residuals do not exhibit serious autocorrelation.

Table 6: Estimated Coefficients of the Extracted Factors

Factor		β	t-value	p-value
Constant		7.776	251.306	0.000
1	NT-TD-CX+	0.880	26.040	0.000
2	DC	0.408	11.093	0.000
3	SHT	0.092	2.488	0.013
4	CX-	0.325	9.045	0.000
5	TCDH	0.133	3.580	0.000
6	DK	-0.072	-2.299	0.022
7	TDdis	0.197	5.192	0.000

Note: β = regression coefficient; $p < 0.05$ indicates statistical significance.

The regression equation is expressed as:

KQHT = 7.776 + 0.880(NT-TD-CX+) + 0.408(DC) + 0.092(SHT) + 0.325(CX-) + 0.133(TCDH) - 0.072(DK) + 0.197(TDdiscipline) (1)

The results of the regression analysis reveal that all included factors remain significant at the 0.05 threshold. Among them, the NT-TD-CX+ factor shows the strongest positive effect on academic performance, suggesting that positive cognition, attitude, and emotions play the most influential role in students' academic outcomes.

4.5. Model Diagnostics and Optimization Analysis

Multicollinearity assessment indicates that the Variance Inflation Factor (VIF) values fall within the range of 1.00 to 1.02, confirming the absence of substantial multicollinearity among the independent variables. In addition, the Durbin-Watson statistic (1.728) is close to the benchmark value of 2, suggesting that serial correlation in the residuals is not a major concern.

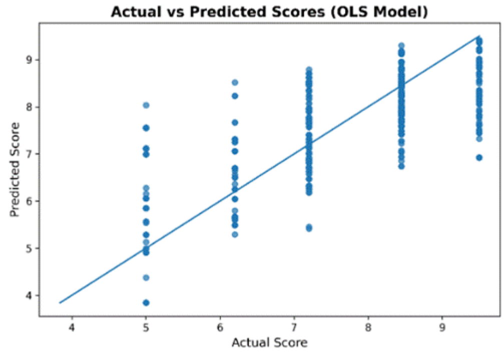

Figure 1: Actual vs. Predicted Plot

The plotted observations are distributed closely along the 45-degree reference line, reflecting a satisfactory level of agreement between actual and estimated values.

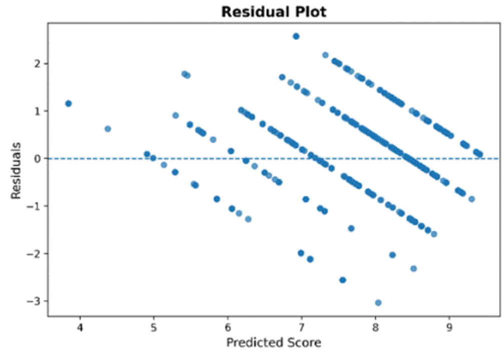

Figure 2: Residual Plot

Although the residuals are generally centered around zero, their variance appears uneven across the fitted values, indicating possible heteroscedasticity.

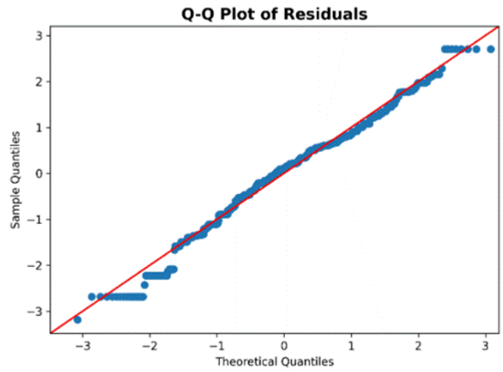

Figure 3: Q-Q Plot

The points align reasonably well with the diagonal reference line, suggesting that the residuals approximate a normal distribution.

The Breusch-Pagan test produces a p-value of 0.000, providing statistical evidence of heteroscedasticity. This finding implies that the underlying relationships in the dataset may involve nonlinear patterns.

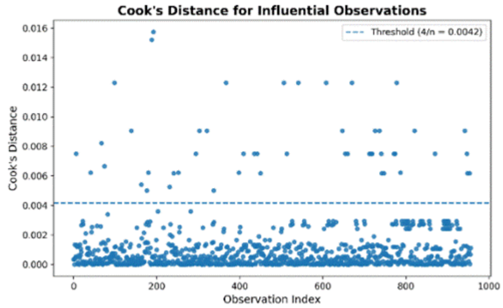

Figure 4: Cook's Distance Plot

Applying a Cook's Distance cutoff of 0.00418, 50 observations exceed the influence threshold. However, given the total sample size (N = 957), this proportion remains relatively small, and the overall model stability is preserved.

Overall, while the regression model satisfies key classical assumptions, the moderate R^2 value and the presence of heteroscedasticity indicate that nonlinear Machine Learning models may provide improved predictive performance in subsequent analyses.

4.6. Model Training and Comparison

To ensure the robustness of the predictive results, the dataset was divided into training and testing subsets using an 80-20 split. In addition, a 10-fold cross-validation procedure was applied during model training to evaluate the stability and generalization capability of the algorithms. This approach reduces the risk of overfitting and provides a more reliable assessment of model performance across different subsets of the data.

To evaluate the predictive capability of academic performance based on the factors extracted from EFA, nine regression models were trained and compared, including both traditional linear models and nonlinear Machine Learning models.

To evaluate predictive accuracy, three statistical metrics were employed: MAE, RMSE, and R^2, with detailed results provided in Table 7:

Table 7: Comparison of Predictive Model Performanc

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	0.796	1.021	0.493
Gradient Boosting	0.227	0.400	0.922
Random Forest	0.121	0.317	0.951
XGBoost	0.100	0.317	0.951
Decision Tree	0.596	0.831	0.664
Lasso Regression	0.790	1.011	0.503
Ridge Regression	0.792	1.011	0.503
Elastic Net	0.791	1.011	0.503
SVR	0.250	0.456	0.899

Performance Analysis

Table 7 reveals a clear performance gap between linear models and nonlinear Machine Learning models.

The linear models (Linear, Lasso, Ridge, Elastic Net) achieved R^2 values of approximately 0.50, indicating limited explanatory power. The evidence suggests that the relationship between the explanatory factors and academic performance extends beyond a strictly linear framework.

Tree-based and boosting models (Random Forest, XGBoost, Gradient Boosting) achieved substantially higher R^2 values (above 0.92), demonstrating strong capability in modeling nonlinear relationships and interactions among predictors.

The SVR model achieved an R^2 of 0.899, which, although slightly lower than boosting models, is significantly higher than linear models. Its MAE and RMSE values remain relatively low and stable, indicating consistent predictive performance.

Model Selection

The study selected SVR (Support Vector Regression) as the representative predictive model based on the following considerations:

· SVR achieves high predictive performance (R^2 = 0.899) with relatively low error metrics, ensuring reliable forecasting accuracy.

· Through kernel functions, SVR effectively captures nonlinear relationships while maintaining control over model complexity via hyperparameters (C & ε).

· Compared to boosting models, SVR tends to exhibit greater stability and a lower risk of overfitting, particularly with a moderate sample size (N = 957).

· SVR is well-suited for continuous prediction tasks and demonstrates strong generalization capability when deployed on new datasets.

Therefore, although it does not produce the highest R^2 value, SVR represents a balanced choice between predictive accuracy and model stability, making it appropriate for practical application in educational data analysis.

4.7. Application

After selecting the SVR model as the predictive model, the study developed a web-based demo to illustrate the practical applicability of the model. The system was designed based on a survey questionnaire corresponding to the identified influencing factors. Students respond to the online questionnaire by indicating their level of agreement using a Likert scale.

Upon submission, the responses are processed by a previously trained and stored SVR model. The system automatically predicts the student's academic performance and generates corresponding learning recommendations. These recommendations focus on improving learning attitudes, strengthening intrinsic motivation, and developing effective study habits.

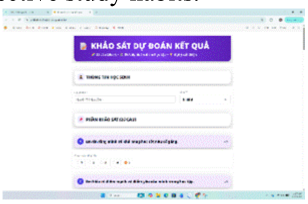

Figure 5: Survey Interface of the Demo System

The development of this web demo not only demonstrates the feasibility of applying ML in educational settings but also highlights its potential integration into learning management systems to support early identification of students at risk of low academic performance.

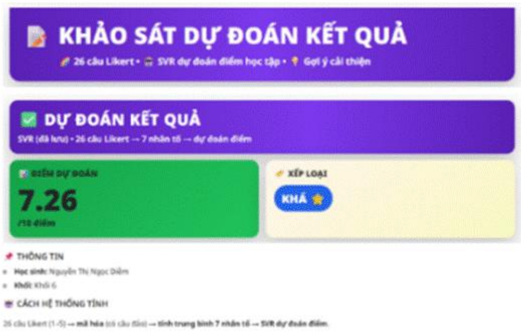

Figure 6: Predicted Academic Performance Based on

The online survey interface includes questions grouped into seven factor categories. Students select their responses before submitting the data to the system.

SCAN ME

Figure 7: QR Code for Accessing the Demo Video

After completing the survey and clicking the "Predict Result" button, the system processes the data and displays the predicted academic performance, as illustrated below.

Survey Responses

The results page presents the predicted score along with tailored learning recommendations.

To further demonstrate system feasibility, a video recording of the demo operation was produced.

4.8. Research Contributions

This study provides several important contributions to the literature on educational data analytics.

First, it constructs a context-specific dataset designed for lower secondary school students within the Vietnamese education system, thereby extending empirical research beyond the university-focused datasets commonly used in prior studies.

Second, the study conducts a comprehensive comparison of nine predictive models, demonstrating that nonlinear machine learning algorithms outperform traditional regression models in predicting academic performance.

Third, the study integrates the predictive model into a web-based demonstration system that enables students to receive automated predictions and learning recommendations. This practical implementation illustrates how machine learning models can be embedded into educational support systems to assist both students and educators.

4.9. Research Implications

4.9.1. Implications for Teachers

The findings indicate that factors related to cognition, learning attitudes, and positive academic emotions exert a substantial influence on students' academic performance. This suggests that teachers should adopt instructional strategies that actively engage students in the learning process, such as problem-based learning, collaborative learning, and continuous formative feedback.

Additionally, teachers may utilize online survey tools or learning analytics platforms to monitor students' motivation and learning attitudes. Integrating data analysis tools and Machine Learning-based predictive systems can assist teachers in identifying early changes in students' learning behavior and adjusting instructional approaches accordingly.

4.9.2. Implications for School Management

The results suggest that the application of data analytics and Machine Learning models can assist schools in identifying factors that influence students' academic outcomes at an early stage. Schools may therefore consider developing learning data management systems that integrate analytical and predictive functions in order to monitor students' academic progress over time.

Predictive models can be employed to detect students who may be at risk of low academic performance. Based on these insights, schools can implement targeted interventions such as academic counseling, supplementary classes, or programs that support the development of effective learning strategies.

4.9.3. Implications for Educational Policy

The findings demonstrate that Machine Learning techniques have considerable potential for supporting the analysis and prediction of students' academic performance. Consequently, educational authorities may consider promoting the adoption of data-driven analytical systems in educational management.

The development of learning analytics platforms at the school or regional level could provide timely insights for teachers and administrators, thereby enabling evidence-based educational decision-making and more effective monitoring of student learning outcomes.

4.9.4. Implications for Data Analytics and Artificial Intelligence in Education

The results of this study indicate that combining educational survey data with Machine Learning models provides an effective approach for analyzing and predicting students' academic performance. This suggests that researchers and practitioners in the field of Data Analytics can leverage educational datasets derived from learning management systems or educational surveys to develop predictive models and decision-support tools for instructional improvement.

Furthermore, the findings show that nonlinear algorithms such as Random Forest, XGBoost, and Support Vector Regression are capable of capturing complex interactions among psychological, behavioral, and environmental variables. This opens opportunities for future research integrating Machine Learning techniques with Explainable AI and Learning Analytics in order to enhance the interpretability and practical applicability of predictive models in educational contexts.

Finally, the implementation of the predictive model in a web-based application demonstrates the feasibility of integrating Machine Learning systems into digital learning environments or Learning Management Systems (LMS). Such integration may support the development of early-warning systems and personalized learning support tools within smart education ecosystems.

5. Conclusion and Future Directions

5.1. Conclusion

This study analyzed factors influencing academic performance among lower secondary school students and applied Machine Learning models to predict academic outcomes. Based on 957 valid observations collected from students in Dong Nai Province, the study conducted reliability testing, exploratory factor analysis, regression analysis, and predictive modeling. The reliability analysis showed acceptable internal consistency, with Cronbach's Alpha values ranging from 0.602 to 0.851. The EFA results also confirmed the suitability of the data for factor analysis, with KMO = 0.915 and $p < 0.001$, resulting in seven extracted factors.

The multiple linear regression model explained 53.4% of the variance in academic performance. Among the extracted factors, the positive cognition-attitude-emotion factor had the strongest influence, indicating that students' confidence, learning attitude, and positive learning emotions play an important role in academic outcomes.

The comparison of nine predictive models showed that nonlinear models performed better than linear models. Random Forest and XGBoost achieved the highest predictive performance, with $R^2 = 0.951$, while SVR achieved $R^2 = 0.899$ and RMSE = 0.456. Although SVR did not produce the highest R^2 value, it was selected for deployment due to its balance between predictive accuracy, model stability, and suitability for practical implementation.

Based on the trained SVR model, a web-based demo system was developed. The system allows students to complete an online survey, automatically predicts their academic performance, and provides corresponding learning recommendations. Overall, the findings clarify key factors affecting lower secondary students' academic performance and demonstrate the feasibility of integrating Machine Learning into educational support tools.

5.2. Future Directions

Despite its contributions, this study has several limitations. First, the dataset was collected from lower secondary school students in Dong Nai Province; therefore, the findings may not fully represent students in other regions. Second, the study mainly used survey data, which may be affected by self-report bias. Third, the data were collected at one point in time, so the study could not examine changes in students' academic performance over different semesters or school years.

Future research should expand the sample to include students from different provinces, school types, and grade levels to improve the generalizability of the findings. Longitudinal data should also be collected to analyze changes in academic performance over time and improve the stability of prediction results.

In terms of modeling, future studies may apply more advanced algorithms, such as deep learning or ensemble learning, to improve predictive accuracy. In addition, model interpretation techniques such as SHAP or LIME should be integrated to explain how each factor contributes to the predicted outcome. This would make the prediction results more transparent and useful for teachers, students, and school administrators.

From an application perspective, the current web demo can be further developed into a practical decision-support system for schools. Future versions may include prediction history, student progress tracking, early warning functions, and analytical reports for teachers and parents.

Integration with Learning Management Systems could also support more effective monitoring and personalized academic support.

6. REFERENCES

[1] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," International Journal of Educational

[3] T. T. H. Duong, T. T. Nguyen, T. P. L. Tran, and T. H. G. Tran, "Exploring factors affecting Vietnamese students' mathematical competence using PISA 2022 data: A machine learning approach with SHAP-based interpretation," Journal of Education, vol. 25, no. 16, pp. 24 - 29, 2025.

[4] D. Ifenthaler and J. Y. K. Yau, "Utilising learning analytics for study success: Reflections on current empirical findings," Research and Practice in Technology Enhanced Learning, vol. 15, no. 1, pp. 1 - 17, 2020
DOI: 10.1186/s41039-020-00129-y.

[5] Ministry of Education and Training, Law on Education No. 43/2019/QH14. Hanoi, Vietnam, 2019.

[6] Ministry of Education and Training, Circular No. 22/2021/TT-BGDĐT on the Assessment of Lower and Upper Secondary School Students. Hanoi, Vietnam, 2021.

[7] N. L. Nguyen, "Predicting students' academic performance using supervised machine learning models," Equipment with New General Education Program, vol. 2, no. 93, pp. 5 - 7, 2023.

[8] T. H. Nguyen, H. L. Tran, and T. K. Do, "Applying machine learning to identify influencing factors and predict students' academic performance in blended learning models," HNUE Journal of Science, vol. 68, no. 1, pp. 63 - 76, 2023.

Technology in Higher Education, vol. 17, no. 1, pp. 1 - 21, 2020
DOI: 10.1186/s41239-020-0177-7.

[2] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3 - 17, 2009.

[9] T. K. S. Nguyen, H. H. Nguyen, T. T. T. Hoang, and Q. N. Tran, "Building a dataset of factors influencing academic performance of students at Hanoi Metropolitan University," Journal of Science, vol. 82, pp. 29 - 43, 2024.

[10] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," Computers & Education, vol. 67, pp. 49 - 64, 2014
DOI: 10.1016/j.compedu.2013.02.004.

[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] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 252-254
DOI: 10.1145/2330601.2330661.

[13] B. T. Tran, "Applying machine learning to predict students under academic warning at the University of Economics, Hue University," Journal of Management and Economics, vol. 21, pp. 1 - 16, 2022.