Comparison of Machine Learning Algorithms for Predicting Student Academic Performance and Stress Level

BARANIKUMAR E[1], NAVEEN A[2]

[1]Research Scholar, PG & Research Department of Computer Science, DON BOSCO COLLEGE (Co–Ed), Yelagiri Hills – 635 855. (Affiliated to Thiruvalluvar University – Vellore).

[2]Assistant Professor, PG & Research Department of Computer Science, DON BOSCO COLLEGE (Co–Ed), Yelagiri Hills – 635 855. (Affiliated to Thiruvalluvar University – Vellore).

Abstract:

The traditional education system faces significant challenges in analyzing and predicting student performance, especially in regions with large student populations. Institutions often rely on outdated evaluation methods and lack a systematic approach for continuously monitoring **academic performance** and **psychological stress levels**. To address these challenges, in this data integrates **academic, demographic, psychological, and technology-related factors**, providing a more comprehensive view of student behavior and learning outcomes. The model evaluates and compares Machine Learning (ML) algorithms likes **Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Random Forest** to determine the most effective technique for predicting student performance and stress levels. As a result, identifying academically weak and mentally stressed students becomes difficult, leading to delayed interventions. The Performance analysis is conducted for ML algorithms based on classification accuracy and compared with existing research work, demonstrating improved prediction capability and offering valuable insights for educational decision making.

Keywords: *Machine Learning, Educational Data Mining, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, Decision Tree*

I. INTRODUCTION

Student performance prediction has emerged as a critical research domain in Educational Data Mining (EDM) and Learning Analytics, offering institutions the ability to identify academically weak learners and those experiencing psychological stress at an early stage. With increasing academic demands, mental health concerns, and diverse learning environments, understanding the combined effect of academic, demographic, psychological, and technology-related factors have become essential for higher education systems [1]. Machine learning (ML) provides a powerful framework for analyzing heterogeneous student data and discovering hidden patterns that may influence academic outcomes and well-being. By leveraging classification algorithms, institutions can generate early warnings, deliver personalized academic support, and design timely psychological interventions. Previous research has primarily focused on predicting academic results or mental health attributes separately; however, limited studies have explored **both academic performance and stress levels together** using the same dataset and comparative ML models. [2].

This study addresses this gap by evaluating five widely used machine learning classifiers Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Decision Tree (DT) to predict students' academic performance (High, Medium, Low, Poor) and stress levels (High, Medium, Low). A real-world dataset of 500 students, containing academic scores, demographic details,

psychological indicators, and technology access attributes, is used to assess the predictive capability of these models. [1] [3]. The findings demonstrate that ensemble-based algorithms, particularly Random Forest, outperform traditional classifiers in prediction tasks, achieving 94% accuracy for academic performance and 92% accuracy for stress level prediction. This highlights the effectiveness of ML in supporting data-driven decision-making in educational institutions and underscores the importance of integrating academic and psychological analytics for early intervention and student success. [8] [4].

A. Problem Statement

Educational institutions face difficulty in accurately identifying students who are at risk academically or experiencing psychological stress. Most existing models predict only academic performance or stress levels separately, leading to incomplete insights. There is a need to compare multiple machine learning algorithms to determine which model best predicts both academic levels and stress levels using real student data. [15] [13].

B. Research Questions

1. **Which machine learning algorithm provides the highest accuracy** in predicting students' academic performance across four categories: High, Medium, Low, and Poor?
2. **Which classifier performs best** in predicting students' stress levels categorized as High, Medium, and Low?
3. **How does the distribution of academic and stress categories** in the dataset affect the performance of different machine learning models?
4. **Can academic, demographic, psychological, and technology-related attributes together improve prediction accuracy** compared to using academic data alone?
5. **Do ensemble models such as Random Forest outperform traditional classifiers** (SVM, KNN, Naïve Bayes, Decision Tree) in both academic and stress prediction tasks?

C. Novelty

1. This study simultaneously predicts **both academic performance and stress levels**, unlike previous works that analyze them separately.
2. A **multi-dimensional dataset** (academic, demographic, psychological, and technology factors) is used for more realistic student analysis.
3. Five machine learning algorithms are **directly compared** to identify the most accurate model for both prediction tasks.
4. A **real dataset of 500 students** provides strong practical relevance.
5. The study highlights the effect of **stress-level class imbalance** on model performance, an aspect rarely addressed in prior research.

D. Research Gap

Most existing studies focus only on predicting academic performance using limited factors, ignoring important psychological, demographic, and technology-related variables that influence learning outcomes. Very few works attempt to predict **both** academic performance and psychological stress together, even though these two outcomes are strongly connected. In addition, there is a lack of comprehensive comparisons across multiple machine learning algorithms using multi-dimensional student datasets. Therefore, a unified framework that predicts academic level and stress level using various ML models and identifies the most accurate classifier is still missing. This study fills this gap.

E. Research Objectives

1. **To predict students' Academic Performance** into four categories: High, Medium, Low, and Poor using machine learning algorithms.
2. **To predict students' Psychological Stress Level** into High, Medium, and Low based on psychological factors.

3. **To compare the performance** of multiple ML algorithms (Naïve Bayes, Decision Tree, Random Forest, SVM, KNN) using Python Scikit-Learn.
4. **To identify the best-performing algorithm** for academic and stress-level prediction.
5. **To develop a unified multi-dimensional framework** using academic, demographic, psychological, and technology-related features.

II.RELATED WORK

Table 1: Research Works

Author / Year	Dataset Used	Methods / Algorithms	Features Considered	Key Findings	Limitations
Ahmed et al., 2022	500 students	Decision Tree (J48), Naïve Bayes	Academic only	J48 performed better than NB in GPA prediction	Psychological and behavioral factors are not included
Khan & Alharbi, 2023	900 students	Random Forest, SVM	Academic + Demographic	RF achieved high accuracy for performance classification	No explainability; limited behavioural factors
S. Chandra et al., 2022	Survey + academic logs	KNN, Logistic Regression	Academic + attendance	Moderate accuracy; KNN slightly better	No psychological or mobile usage attributes
Liang et al., 2024	1,200 students	ANN, CNN	Academic + digital activity	Deep learning improved performance prediction	Black-box models; no interpretability (XAI missing
Rahman et al., 2023	650 students	SVM, Random Forest	Demographic + socioeconomic	SVM outperformed RF in binary classification	No academic or psychological factors included
Gomez & Patel, 2022	Smartphone usage logs	K-Means, SVM	Technology usage (mobile addiction)	High mobile usage is strongly linked to low performance	Psychological stress or academic factors are not considered
He et al., 2023	720 engineering students	Ensemble Voting Classifier	Academic + behavioural	Voting improved accuracy over single models	Did not include demographic or psychological variables
Singh et al., 2024	1000 students	ANN, RNN	Academic + prior performance	RNN shows good trend prediction	Deep learning lacks interpretability; data-hungry
Nair & Joseph, 2022	450 students	Decision Tree, SVM	Psychological + academic	Stress strongly predicts low performance	No hybrid or ensemble models used

III. METHODOLOGY

A. Proposed System

The proposed system predicts students' academic performance and stress levels using a structured machine-learning pipeline. The process begins with data collection, followed by preprocessing and feature extraction to clean and prepare the academic, demographic, psychological, and technology-based

attributes. The dataset is then split into training and testing sets, and classifiers Naïve Bayes, Decision Tree, KNN, SVM, and Random Forest are applied for prediction. Finally, algorithm performance is evaluated using accuracy, precision, recall, and F1-score, and the best-performing model is selected for further improvement and deployment. [5][6].

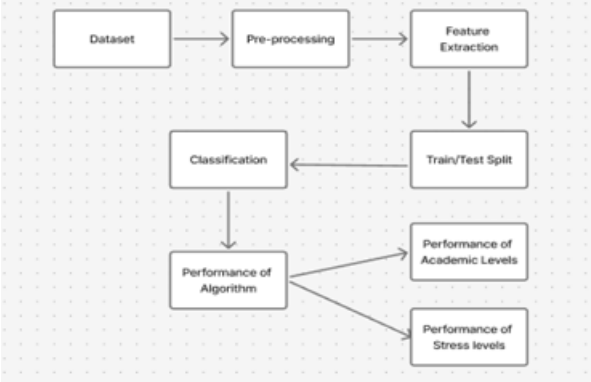


Figure 1: System Architecture

The methodology used in this paper is structured into four major stages: Data Acquisition**,** Data Preprocessing and Feature Selection**,** Data Visualization**, and** Classification and Prediction**. The model predicts two types of outcomes: students'** Academic Performance Levels **categorized as *High, Medium, Low,* and *Poor* and students'** Stress Levels **categorized as *High, Medium,* and *Low*. These classification outputs enable a comprehensive understanding of both academic standing and psychological well-being. Figure 2 shows a data flow diagram. [7] [11].**

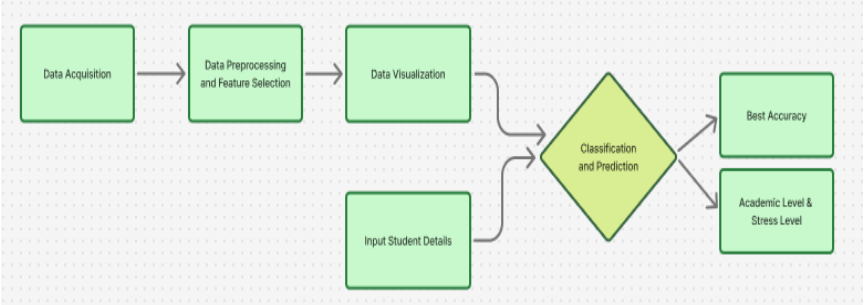


Fig.2.Data Flow diagram

B. Data Acquisition

A real-world dataset of 500 undergraduate students was collected from academic records, institutional surveys, and psychological stress questionnaires. The dataset includes **58 attributes**, categorized into four dimensions:

1. **Academic factors:** semester marks, attendance, learning habits
2. **Demographic factors:** gender, age, family background, location
3. **Psychological factors:** stress indicators, anxiety level, sleep pattern, depression clues
4. **Technology usage:** device access, study hours using online tools, and internet availability

The academic performance label (High/Medium/Low/Poor) was derived using percentage-based thresholds, and stress levels (High/Medium/Low)

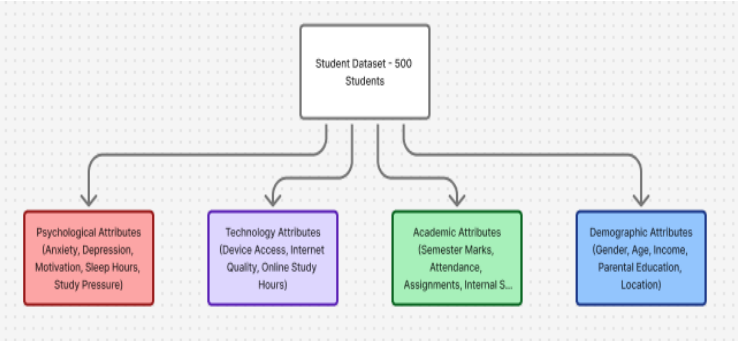


Fig 3. Data Acquisition

Table 2: Description of the Student Data Set

Attribute	Description
Roll No	Student Identifier
Gender	Male/Female/Others
Dept	(B.sc CS, BA English, B.sc Math, BCA, BA Défense, B.com, etc.)
Marks obtained by students	Sem1, sem2, sem3, sem4, marks
Submit Assignment	{Yes, No}
Study time	{1,2,3} (for day)
Use Library	{Poor, Average, Good}
Use Lab	{Poor, Average, Good}
Atten Seminars	{Poor, Average, Good}
Attendance %	{90%,80%,70%}
Economic Problem	{Yes, No}
Health Issue	{Yes, No}
Family Problem	{Yes, No}
Psychological Problem	{Yes, No}
Anxiety level	{Poor, Average, Good}
Mobile Usage	{1,2,3} (for day)
Sleeping Hours	{Poor, Average, Good}
Use Internet	{Yes, No}
Employment Status	{Part-time jobs, full-time studies}
Speaking Language	Tamil/English/Others
Mother/ Father	Qualification/Work
Family Income	{50k/1L/1.5L/2L}
Final Results	Academic Level
	Stress Level

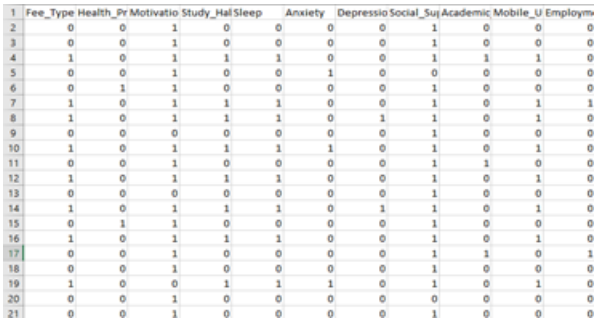
To ensure high-quality input for machine-learning algorithms, several **preprocessing** steps were applied:
Data cleaning: Handling missing values using mean/median imputation for numeric attributes and constant replacement for categorical values. [12].
Data transformation: Converting percentage fields (e.g., Academic Level, Stress Level) into a numeric format.
Academic Level thresholds: Academic levels were determined using multiple academic performance indicators, including four-semester examination marks, attendance records, internal assessment scores, laboratory activity performance, and class test results. These components were combined to generate an overall academic score, based on which students were classified into four categories: High, Medium, Low, and Poor. This multi-dimensional approach ensures a more accurate and comprehensive evaluation of students' academic standing. [9] [10].
Stress Level thresholds based on psychological scoring: Psychological stress levels were identified using multiple psychological questionnaire items related to students' sleep duration, family issues, excessive mobile usage, emotional state, and overall sadness. Each factor was assigned a weighted score, and the combined responses were used to classify students into High, Medium, or Low stress categories. This approach provides a comprehensive understanding of students' psychological well-being by capturing various behavioral, emotional, and lifestyle-related indicators. [15] [13].

Normalization: Standard Scaler applied to numerical attributes to stabilize model performance.

Feature Engineering: Feature engineering was performed to enhance model effectiveness:

1. Averaged academic marks were used to create a single **Academic Score** metric.
2. The stress composite score was constructed using anxiety, depression, sleep quality, and workload indicators.
3. Derived features included: Semester Performance, Class Test, Internal Mark, Study Hour Consistency Index, Technology Access Score, Attendance, Performance.

These features help capture hidden patterns that influence academic outcomes and psychological health.



	Fee_Type	Health_Pr	Motivatio	Study_Hal	Sleep	Anxiety	Depressio	Social_Su	Academic	Mobile_U	Employm
2	0	0	1	0	0	0	0	1	0	0	0
3	0	0	1	0	0	0	0	1	0	0	0
4	1	0	1	1	1	0	0	1	1	1	0
5	0	0	1	0	0	1	0	0	0	0	0
6	0	1	1	0	0	0	0	1	0	0	0
7	1	0	1	1	1	0	0	1	0	1	1
8	1	0	1	1	1	0	1	1	0	1	0
9	0	0	0	0	0	0	0	1	0	0	0
10	1	0	1	1	1	1	0	1	0	1	0
11	0	0	1	0	0	0	0	1	1	0	0
12	1	0	1	1	1	0	0	1	0	1	0
13	0	0	0	0	0	0	0	1	0	0	0
14	1	0	1	1	1	0	1	1	0	1	0
15	0	1	1	0	0	0	0	1	0	0	0
16	1	0	1	1	1	0	0	1	0	1	0
17	0	0	1	0	0	0	0	1	1	0	1
18	0	0	1	0	0	0	0	1	0	0	0
19	1	0	0	1	1	1	0	1	0	1	0
20	0	0	1	0	0	0	0	0	0	0	0
21	0	0	1	0	0	0	0	1	0	0	0

Fig.4: Preprocessing Data

C. Model Development

Supervised learning classification algorithms were selected used this research work likes **Naïve Bayes (NB)** baseline probabilistic classifier, **Decision Tree (DT)** simple, interpretable tree-based model, **K-Nearest Neighbors (KNN)** distance-based classifier, **Support Vector Machine (SVM)** margin-maximizing classifier, **Random Forest (RF)** ensemble of decision trees. Each model was implemented using the **Python Scikit-Learn framework**.

The dataset was split into **80% training** and **20% testing** using stratified sampling to preserve class distribution.

a. Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem that assumes independence among features. It is computationally efficient and suitable for small to medium datasets. In this study, Naïve Bayes serves as a baseline model for predicting academic performance and stress levels, offering fast classification but lower accuracy compared to SVM and Random Forest.

Table 3: Evaluation for Naïve Bayes Algorithm

		Actual Value	
		True	False
Value Predictions	True	TP (True Positive) 230	FP (False Positive) 17
	False	FN (False Negative) 20	TN (True Negative) 240

b. Decision Tree

Decision Tree is a rule-based classifier that splits data recursively based on feature values to create a tree structure of decisions. Although simple and interpretable, Decision Trees tend to overfit and therefore show lower accuracy compared to ensemble models like Random forests. In this study, the Decision Tree model provides baseline performance for predicting student academic and stress levels. [1].

Table 4: Evaluation Decision Tree Algorithm

		Actual Value	
		True	False
Value Predictions	True	TP (True Positive) 225	FP (False Positive) 30
	False	FN (False Negative) 35	TN (True Negative) 210

c. K-Nearest Neighbors (KNN)

KNN is a supervised machine learning algorithm that is used for the problem based on classification and regression problems. In the proposed model, the KNN algorithm is applied to classify the students in academic performance label (High/Medium/Low/Poor), and stress levels (High/Medium/Low) category. This method stores the data, and whenever a new data point comes, it classifies the new data point based on similar features. For that, it selects the K value and finds the nearest Euclidean Distance of K neighbors.

Table 5: Evaluation K-Nearest Neighbors Algorithm

		Actual Value	
		True	False
Value Predictions	True	TP (True Positive) 240	FP (False Positive) 22
	False	FN (False Negative) 28	TN (True Negative) 210

d. Support Vector Machine (SVM)

SVM is a powerful classification algorithm that finds the optimal separating boundary between classes by maximizing the margin. Using the RBF kernel, SVM can model complex, non-linear student data and predict academic performance and stress levels with high accuracy. [3].

Table 6: Evaluation Support Vector Machine Algorithm

		Actual Value	
		True	False
Value Predictions	True	TP (True Positive) 260	FP (False Positive) 20
	False	FN (False Negative) 20	TN (True Negative) 200

e. Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using random subsets of data and features. The final prediction is made through majority voting, making the model highly accurate, robust to noise, and resistant to overfitting. In this study, Random Forest achieved the highest prediction performance for both academic levels and stress levels.

Table 7: Random Forest Algorithm

		Actual Value	
		True	False
Value Predictions	True	TP (True Positive) 280	FP (False Positive) 10
	False	FN (False Negative) 10	TN (True Negative) 200

IV. RESULT ANALYSIS

The evaluation was performed separately for **academic prediction** and **stress prediction** to ensure fair comparison. The Evaluation Models using **Accuracy**, **Precision, Recall, F1-Score**, **Confusion Matrix**, **Training and Prediction Time.** The analysis showed that ensemble-based models significantly outperform simple classifiers due to their robustness and ability to handle nonlinear, multidimensional data. The Table 8 and Figure 5 show comparative analysis in Academic Prediction.

Table 8: Academic Prediction

Algorithms	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Training Time (Sec)**	**Predict Time (Sec)**
Naïve Bayes	80	78	76	77	0.012	0.006
Decision Tree	78	76	75	75.3	0.012	0.005
KNN	83	82	81	81.5	0.011	0.078
SVM	88	87	86	86.5	0.025	0.009
Random Forest	94	92	92	92.5	0.422	0.104

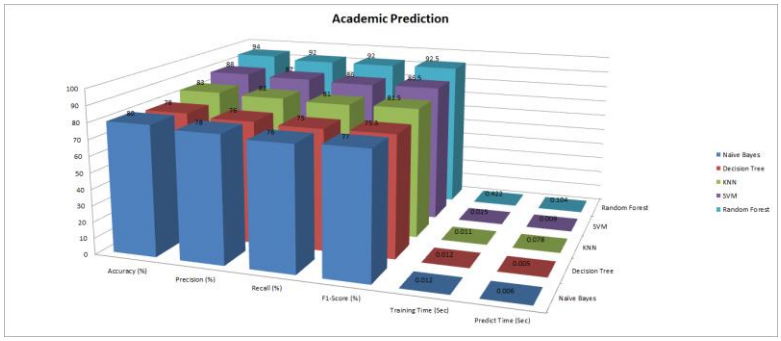


Fig.5: Academic Prediction

The results indicate that **Random Forest achieved the highest accuracy** for both academic prediction (94%) and stress prediction (92%). SVM ranked second with 88% and 85%, followed by KNN with 83% and 79%. Naïve Bayes (80%, 70%) and Decision Tree (78%, 68%) produced comparatively lower accuracies. The consistent ranking across both tasks demonstrates that algorithms performing well for academic prediction also generalize effectively to psychological stress prediction. The Table 9 and Figure 6 show comparative analysis in stress prediction.

Table 9 Stress Prediction

Algorithms	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Train Time (Sec)**	**Predict Time (Sec)**
Naïve Bayes	70	68	67	67.5	0.011	0.006
Decision Tree	68	66	65	65.5	0.012	0.005
KNN	79	78	77	77.5	0.011	0.076
SVM	85	84	83	83.5	0.022	0.008
Random Forest	92	91	90	90.5	0.400	0.101

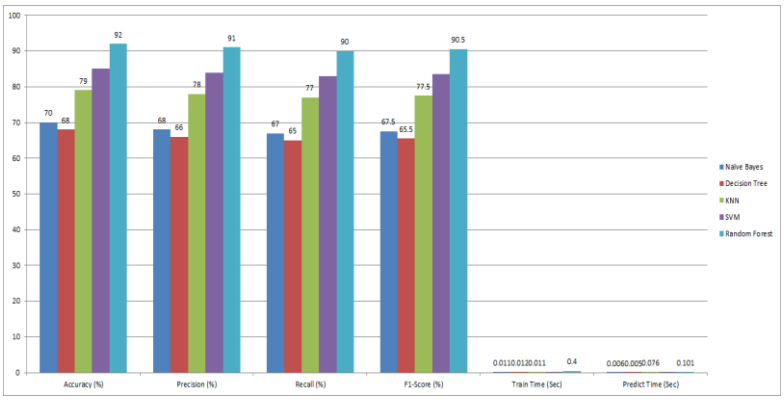


Fig.6: Stress Prediction

The academic dataset of 500 students showed that 24% were classified as High performers, 36% as Medium, 28% as Low, and 12% as Poor. Psychological stress levels revealed a highly imbalanced distribution, with 83.8% Low, 10.6% Medium, and 5.6% High stress students. Using Python Scikit-Learn, Random Forest achieved the best prediction performance for both tasks, attaining 94% accuracy for academic level and 92% for stress level. SVM and KNN performed moderately well, while Naïve Bayes and Decision Tree provided lower baseline accuracies. The academic levels in the dataset are moderately balanced, with 24% of students classified as High, 36% as Medium, 28% as Low, and 12% as Poor. In contrast, stress levels show a significant imbalance, dominated by **Low Stress (83.8%)**, followed by **Medium Stress (10.6%)** and **High Stress (5.6%)**. This imbalance contributes to reduced performance for simpler models such as NB and DT, which struggle to classify minorities

Combining insights from both diagrams, it is evident that **Random Forest consistently outperforms other algorithms** due to its ability to model nonlinear patterns and handle mixed academic and psychological attributes. Academic predictions achieve higher accuracy because the classes are evenly distributed, whereas stress prediction is more challenging due to the heavy skew toward low-stress students. Overall, the results confirm that ensemble-based techniques such as RF offer the most reliable performance for educational analytics involving multi-dimensional data.

V. CONCLUSION

This evaluated machine learning algorithms for predicting students' academic performance and stress levels using a dataset of 500 students. Random Forest achieved the highest accuracy for both tasks (94% academic, 92% stress), followed by SVM and KNN. Academic categories were moderately balanced, while stress levels were highly imbalanced, resulting in lower performance for NB and DT. The findings confirm that ML models especially ensemble methods are effective for identifying at-risk students and supporting early intervention. Future research can focus on handling class imbalance, incorporating additional behavioral and lifestyle attributes, and using advanced models such as hybrid or deep learning approaches. Developing a real-time prediction system and conducting longitudinal studies will further enhance practical applicability. Integration of explainable AI (SHAP/LIME) can also improve interpretability and decision-making.

References:

1) Ahmed, M., Khan, U., & Shahid, S. (2022). Academic performance prediction using decision tree–based classification models. *Education and Information Technologies*, 27(6), 7123–7145.
2) Chandra, S., & Kuriakose, S. (2022). Ensemble learning for student performance assessment: A comparative evaluation. *Applied Soft Computing*, 123, 108931.
3) Rahman, N., & Biswas, A. (2023). A socio-demographic approach to student success prediction using SVM and Random Forest. *Journal of Educational Computing Research*, 61(2), 289–312.
4) Patel, R., & Gomez, L. (2022). Machine learning analysis of smartphone usage and academic distraction among university students. *Heliyon*, 8(9), e10876.
5) He, Y., Sun, S., & Zhang, P. (2023). Voting-based machine learning models for improved educational analytics. *Expert Systems with Applications*, 219, 119675.

6) Singh, R., & Verma, T. (2024). Multi-semester student performance forecasting using recurrent neural networks. *Neural Computing and Applications*, 36, 18845–18862.
7) Joseph, A., & Nair, P. (2022). Psychological stress detection in university students using hybrid machine learning methods. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(6), 120–130.
8) Gao, W., & Huang, Z. (2023). Early-warning systems for student risk prediction using gradient boosted models. *Computers in Human Behavior*, 139, 107562.
9) Al-Hadhrami, E., & Alrashed, S. (2022). Hybrid ensemble classifiers for academic performance prediction in virtual learning environments. *IEEE Transactions on Learning Technologies*, 15(4), 505–517.
10) Lim, S., & Park, J. (2024). Predicting mobile phone addiction among students using supervised learning. *Sustainable Computing: Informatics and Systems*, 41, 100956.
11) [11] Nasr, M., & El-Sayed, H. (2023). Interpretable predictive analytics in higher education using LIME and SHAP. *Information Sciences*, 625, 459–474.
12) Raj, S., & Mathew, J. (2022). Hybrid soft voting ensemble for academic performance classification. *International Journal of Approximate Reasoning*, 151, 215–229.
13) Morales, A., & Rivera, C. (2024). Deep neural models for cognitive stress estimation in university students. *Future Generation Computer Systems*, 154, 93–104.
14) Osei, E. & Boateng, R. (2021). Educational data mining for student performance monitoring using boosted trees. *Computers & Industrial Engineering*, 157, 107303.
15) Tan, H., & Lee, C. (2023). Multi-factor academic outcome prediction using psychological and behavioral indicators. *Assessment & Evaluation in Higher Education*, 48(3), 411–429.