

Student Performance Prediction Using Machine Learning and Explainability

Harshitha H B¹ Vinay T M² Madhurya H V³ Prerana R⁴ Moulya N⁵

^{1,2,3,4,5}Department of Computer Science and Design

^{1,2,3,4,5}ATME College of Engineering, India

Abstract — The increasing digitalization of academic activities has created large volumes of structured and unstructured student data. Traditional evaluation strategies rely heavily on retrospective assessments and fail to provide early intervention insights. This paper presents a machine-learning-based Student Performance Prediction System using Linear Regression and Explainable AI (SHAP) to forecast a student's final CGPA. The model integrates academic history, attendance, quiz performance, test averages, and assignment behavior using a sector-weighted feature engineering technique. A Streamlit-based interface enables individual and batch predictions while producing transparent explanation dashboards. Experimental results demonstrate a Mean Absolute Error (MAE) of 1.17 and RMSE of 1.37, with Linear Regression outperforming Random Forest and XGBoost. SHAP visualizations enhance interpretability and support data-driven academic decision-making. The system effectively identifies “At-Risk” students and generates intervention-ready reports for educational institutions.

Keywords: Student Performance Prediction, Machine Learning, Linear Regression, Explainable AI (SHAP), Academic Analytics, At-Risk Student Identification

I. INTRODUCTION

Academic institutions increasingly require predictive systems capable of identifying students likely to underperform. Manual evaluation methods provide limited insights and lack early warning capabilities. With the availability of digital academic records, machine learning offers promising solutions for forecasting performance based on multi-dimensional behavioral and academic data.

The proposed system predicts final CGPA using historical SGPA, CGPA, attendance, assessment patterns, and assignment behaviors. An Explainable AI layer (SHAP) provides transparency, enabling educators to understand the contribution of each feature to the prediction. A full-featured Streamlit interface provides individual predictions, manual entry predictions, batch processing, and automated PDF reports.

II. LITERATURE REVIEW

Machine learning in educational data mining (EDM) has evolved significantly. Prior works highlight the importance of interpretability in academic settings.

- 1) Alamri & Alharbi (2021) emphasize the need for explainable models to build trust among educators. Decision Trees, SVM, and Logistic Regression have been commonly used but lack interpretability.
- 2) Sahlaoui et al. (2021) show that ensemble models such as Random Forest and Gradient Boosting provide higher accuracy while SHAP values offer fine-grained explanations.

- 3) Zhao et al. (2021) demonstrate the importance of behavioral data including online learning logs, assignment delays, and attendance, improving prediction accuracy.
- 4) Feng et al. (2022) confirm the effectiveness of preprocessing and feature engineering using Decision Trees, Naïve Bayes, and k-NN.
- 5) Dake et al. (2021) highlight that attendance and assignment behavior became dominant predictors during online learning.

These works collectively underline three gaps:

- 1) Lack of systems combining prediction + interpretability
- 2) Weak integration of assignment timing features
- 3) Limited user-friendly dashboards for educators

III. METHODOLOGY

The system follows a seven-stage pipeline:

A. Data Collection

Dataset includes:

- SGPA, CGPA
- Attendance (%)
- Test Avg (0–15)
- Quiz Marks (0–10)
- Assignment Submission Dates
- Assignment Due Dates
- Student IDs & Demographics

B. Preprocessing

- Missing value imputation
- Date-to-datetime conversion
- Delay calculation: submission – due date
- StandardScaler normalization
- Categorical handling

C. Feature Engineering

A sector-based weighting model was developed.

- 1) *Academic Sector (40%)*
 - SGPA (30%) + CGPA (30%) + Test Avg (40%)
- 2) *Assignment Sector (25%)*
 - On-time submission rate (40%)
 - Consistency of submission delays (30%)
 - Quiz Marks (30%)
- 3) *Attendance Sector (25%)*
 - attendance (%) → normalized to [0,1]
- 4) *Weighted Score*
 - $WS=0.50A+0.25B+0.25C$

D. Model Training

Models evaluated:

- Linear Regression (selected)
- Random Forest
- XGBoost

E. Evaluation Metrics

- MAE
- RMSE
- R² Score

Linear Regression achieved the best generalization.

F. Explainability (SHAP)

- SHAP illustrates positive/negative feature impact per prediction.
- Supports local and global interpretability.

G. UI Development

Streamlit interface includes:

- Individual Prediction using Student ID
- Manual Entry Prediction
- Dataset Upload + PDF Generation
- Explainability Dashboard

IV. RESULTS AND DISCUSSION

A. Model Performance

Linear Regression proved most stable.

Model	MAE	RMSE	R ²
Linear Regression	1.17	1.37	0.07
Random Forest	1.28	1.50	0.11
XGBoost	1.38	1.64	0.33

B. Classification Threshold

- On Track → CGPA ≥ 7.0
- At Risk → CGPA < 7.0

C. SHAP Interpretation

Feature importance ranking:

Feature	Rank
Academic Sector	1
Attendance Sector	2
Assignment Sector	3
Weighted Score	4

D. System Outputs

- 1) Individual prediction card with sector contribution
- 2) Batch prediction → generates two PDFs:
 - on_track_students.pdf
 - at_risk_students.pdf
- 3) Explainability dashboard
- 4) Manual entry prediction with optimistic boost

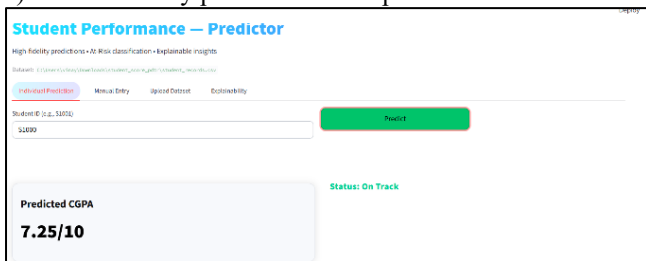


Fig. 1: Individual Prediction

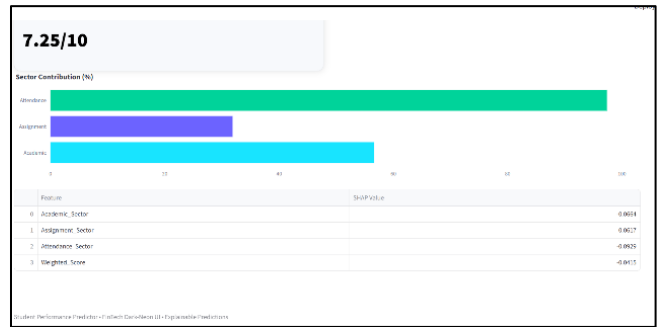


Fig. 2: Sector Contribution

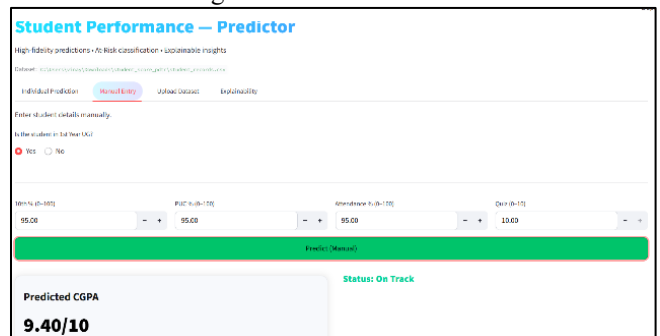


Fig. 3: Manual Entry Sector

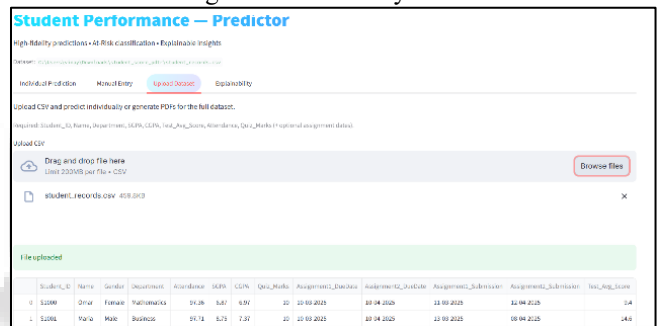


Fig. 4: Prediction based on uploaded dataset

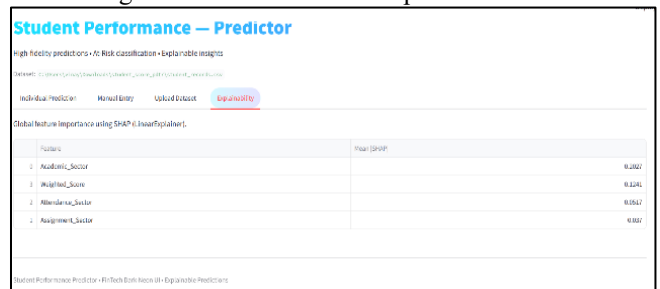


Fig. 5: Explainability

V. CONCLUSION

This paper presents a machine-learning-based system for predicting student academic performance using a sector-weighted feature engineering strategy. Linear Regression emerged as the most reliable model, and SHAP explainability significantly enhances interpretation. The interactive Streamlit interface enables real-time predictions, batch processing, and educator-friendly dashboards. The system supports early identification of academically weak students, enabling timely interventions and improved academic outcomes.

VI. FUTURE WORK

- Integration with LMS/ERP systems
- Deep learning (LSTM, GRU) for time-series prediction
- Behavioral and psychological factor inclusion
- Cloud deployment for large institutions
- Reinforcement learning for personalized study plans
- Automated alert system for parents and mentors
- Enhanced XAI using LIME or counterfactuals

REFERENCES

- [1] A. Alamri and A. Alharbi, "Explainable Student Performance Prediction Models," *IEEE Access*, 2021.
- [2] I. Sahlaoui, S. Sassi, and M. Jemni, "Predicting and Interpreting Student Performance Using Ensemble Models and SHAP," *Procedia Computer Science*, vol. 192, pp. 2357–2366, 2021.
- [3] Z. Zhao, J. Ye, and H. Cheng, "Academic Performance Prediction Using Multisource Behavioral Data," *Journal of Educational Technology*, 2021.
- [4] Y. Feng, H. Li, and W. Zhang, "Analysis and Prediction of Academic Performance Using Educational Data Mining," *Applied Sciences*, vol. 12, no. 1, p. 436, 2022.
- [5] A. Dake, M. S. Asare, and J. Opoku, "Predicting Academic Performance During COVID-19 Using Machine Learning," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 9, pp. 120–133, 2021.
- [6] W. Ahmed, "Machine Learning-Based Academic Performance Prediction," *PubMed Central (PMC)*, 2025.
- [7] E. Kalita and P. Sharma, "LSTM-SHAP Based Academic Performance Prediction," in *Springer Lecture Notes in Networks and Systems*, 2025.
- [8] M. Chen, "Predicting Student Performance by Optimizing Tree Components of Ensembles," *PubMed Central (PMC)*, 2024.
- [9] M. Skittou, "Development of an Early Warning System to Support Educational Planning by Identifying At-Risk Students," *ResearchGate*, 2024.
- [10] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," *IEEE Transactions on Learning Technologies*, 2021.
- [11] W. Liu, W. Xu, X. Zhan, and W. Cheng, "Student Performance Prediction by LMS Data and Classroom Videos," in *IEEE International Conference on Computer Science and Education (ICCSE)*, pp. 323–328, 2021.
- [12] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access*, vol. 10, pp. 21035–21048, 2022.