International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 11, November 2025)

Explainable AI for Student Performance Prediction in E-Learning Environments

Chitra Ganesh Desai

Professor and Head, Department of Computer Science, National Defence Academy, Pune, India

***Abstract*— This study investigates the use of machine learning to predict student academic outcomes using socio-demographic and academic performance features derived from a well-structured student performance dataset. A binary classification model was developed using the Random Forest algorithm to predict whether a student will pass or fail, based on attributes such as gender, lunch type, test preparation course completion, and scores in mathematics, reading, and writing. To address the challenge of interpretability in black-box models, this work employs two powerful explainable AI techniques—LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME provides local interpretability by approximating the model with an interpretable surrogate model around individual predictions, while SHAP offers a unified framework grounded in game theory to quantify the contribution of each feature to global and local predictions.**

The dataset used is particularly valuable for analyzing the impact of socio-demographic and preparatory factors on academic performance in core subjects. It enables predictive modelling of academic success, exploration of educational equity (e.g., the influence of parental education or lunch type), and the development of interpretable models for decision-making. Moreover, insights from these models support the design of adaptive e-learning systems that can proactively identify at-risk learners in K–12 environments and offer personalized interventions.

By combining predictive accuracy with model transparency, this research contributes toward actionable, data-driven strategies in education, supporting both inclusivity and performance improvement in line with modern educational priorities.

Keywords **— Explainable Artificial Intelligence (XAI), Random Forest Classifier, Educational Data Mining, Student Performance Prediction, LIME and SHAP Interpretability, E-learning Participation Analysis**

I. INTRODUCTION

The evolution of the internet has been instrumental in transforming the education sector, laying the foundation for digital learning systems that transcend traditional classroom boundaries.

With the rise of e-learning platforms, education has become more inclusive, flexible, and data-driven, encompassing diverse contexts ranging from K–12 classrooms [1] and higher education institutions to massive open online courses (MOOCs). This transformation has been further accelerated by the integration of artificial intelligence (AI) [2], which has introduced intelligent capabilities into educational environments—enabling personalized learning, adaptive assessments, and predictive analytics.

E-learning today is not confined to content delivery alone; it involves a complex ecosystem of interconnected components including learning management systems (LMS), automated evaluation, progress monitoring, real-time feedback mechanisms, and intelligent tutoring [3]. Among these, learner analytics powered by machine learning has gained prominence, offering the ability to predict academic outcomes, identify at-risk students, and support instructional planning [4]. Such capabilities are particularly valuable in large-scale digital learning environments where educators and administrators must monitor and respond to learner needs across diverse populations and geographies.

However, the effectiveness of machine learning in educational applications depends not only on predictive performance but also on the interpretability of its outcomes [5]. In many educational contexts—particularly K–12 and higher education—stakeholders such as teachers, students, and policymakers require explanations for model predictions to ensure trust, fairness, and actionable insights. This has led to the growing relevance of Explainable AI (XAI), a field that seeks to make machine learning models more transparent and understandable [6].

This study explores the use of machine learning to predict student academic outcomes using socio-demographic and academic performance features derived from a well-structured student performance dataset. A binary classification model based on the Random Forest algorithm is developed to predict whether a student is likely to pass or fail.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 11, November 2025)

To address the interpretability challenge of black-box models, the study employs LIME (Local Interpretable Model-agnostic Explanations) [7] and SHAP (SHapley Additive exPlanations)[8]. LIME offers local interpretability by approximating complex models with simpler, interpretable ones around specific predictions, while SHAP provides a game-theoretic framework to attribute the contribution of each feature to the model's output across the dataset. By incorporating these explainable AI techniques, the model not only delivers accurate predictions but also provides transparency and accountability—making it a valuable tool for enhancing student support in K–12 systems, improving decision-making in higher education, and optimizing learner engagement in MOOCs. It also contributes to enhancing its usability and reliability for targeted intervention strategies in alignment with the goals of NEP 2020.

II. Dataset

The dataset titled StudentsPerformance [9] is an enhanced version of the widely used student academic performance dataset, designed to study how various socio-demographic and educational factors influence students' outcomes in core subjects—mathematics, reading, and writing. It comprises both categorical variables, such as gender, race/ethnicity, parental level of education, lunch type, and test preparation course, and numerical variables including scores in math, reading, and writing—all on a scale of 0 to 100. These variables collectively offer a holistic view of the student's background and academic readiness.

A significant addition to this version of the dataset is the inclusion of a new column, e-learning, which is a hypothetical binary indicator (0 or 1) representing whether a student has access to or participates in e-learning platforms. The inclusion of this feature is particularly timely and relevant in the context of current educational research, especially in the aftermath of the COVID-19 pandemic, which led to a global surge in online learning adoption. The paper under consideration investigates the evolving role of digital interventions in education, and this e-learning variable provides a valuable opportunity to examine its potential correlation with academic performance. By analyzing patterns and model explanations involving this column, the aim is to explore whether access to e-learning environments contributes positively to student achievement, and if so, to what extent it can mitigate gaps caused by other socio-economic factors. This feature enriches the dataset, enabling more contemporary and policy-relevant analyses.

III. Methodology

The methodological framework for this study follows a structured, multi-stage approach. The process begins with the enhanced StudentsPerformance dataset, which includes an additional e-learning variable to capture students' engagement in digital learning environments. The first stage, data preprocessing, involves encoding categorical variables numerically and calculating the average of the three core subject scores—mathematics, reading, and writing. A binary target variable, Pass, is then generated by assigning a value of 1 for students with an average score greater than or equal to 50, and 0 otherwise. In the model development phase, a Random Forest classifier is trained on the processed data, using an 80:20 train–test split and hyperparameter tuning through grid search with cross-validation to ensure optimal performance. The model evaluation phase assesses predictive accuracy using metrics such as accuracy, precision, recall, and F1-score. To enhance the interpretability of model predictions, two complementary explainable AI techniques—LIME and SHAP—are employed. LIME provides local explanations for individual predictions by approximating the model's behavior around a specific instance, while SHAP offers both global and local insights by quantifying the contribution of each feature to the final output. These interpretability tools collectively enhance transparency and trust in the model's decision-making process. Finally, the insights generated are intended to support educators and policymakers in understanding performance drivers and designing data-driven interventions to improve student outcomes.

IV. Result And Analysis

A. Random Forest

The Random Forest classifier demonstrated highly robust predictive performance in classifying students as "Pass" or "Fail." The model achieved an overall accuracy of 99.0%, indicating its strong ability to generalize across unseen test data. Performance metrics including precision (0.9942), recall (0.9942), and F1-score (0.9942) further validate the model's reliability and balanced prediction capability across both classes.

The classification report shows that the model effectively distinguished between passing and failing students, achieving 96% accuracy for the 'Fail' class and 99% accuracy for the 'Pass' class, as reflected by their respective precision and recall values.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 11, November 2025)

The high recall for both classes suggests minimal false negatives, which is critical in educational applications where identifying at-risk students early is of high importance. Similarly, the strong precision values indicate a low rate of false positives, ensuring that the model does not incorrectly label successful students as failing.

While the Random Forest model exhibits exceptional predictive performance, the integration of Explainable Artificial Intelligence (XAI) techniques such as LIME and SHAP is crucial to enhance the transparency and interpretability of its decisions. In educational contexts, predictive accuracy alone is insufficient—educators, administrators, and policymakers must understand why a student is predicted to pass or fail. Explainable AI bridges this gap by revealing the underlying feature contributions that drive each prediction.

Through LIME (Local Interpretable Model-agnostic Explanations), the model's behavior can be locally approximated to highlight the most influential features for an individual student's outcome, such as reading and writing scores or test preparation status. Similarly, SHAP (SHapley Additive exPlanations) provides a global and consistent interpretation by quantifying the contribution of each feature across the entire dataset. This transparency fosters trust and accountability in AI-assisted educational tools, enabling educators to take informed, data-driven actions—such as providing targeted interventions to students identified as at-risk. Thus, incorporating XAI not only validates the model's fairness and reliability but also transforms it into a practical decision-support system in the e-learning ecosystem.

B. LIME Based Local Interpretability

For interpretability analysis, the LIME (Local Interpretable Model-agnostic Explanations) framework was applied to the instance corresponding to student index 10, whose feature vector is summarized as follows: gender = 1, race/ethnicity = 2, parental level of education = 0, lunch = 1, test preparation course = 1, math score = 58, reading score = 54, writing score = 52, and e-learning = 0.

Although the student's average score marginally exceeds the 50 % passing criterion, the Random Forest classifier predicted a "Fail" outcome with 0.90 probability, while assigning only 0.10 probability to the "Pass" class.

The LIME explanation as shown in Figure 1 revealed that low standardized scores in writing (−1.29), reading (−1.56), and mathematics (−1.43) were the primary contributors driving the prediction toward "Fail."

In contrast, contextual attributes such as parental education (−0.25), lunch type (−1.37), and lack of e-learning access (0 → −1.01 standardized) had smaller but still negative influences. These results indicate that the model's decision was predominantly shaped by the student's below-average academic performance across all core subjects, outweighing the marginally passing aggregate.

From an educational analytics perspective, the LIME output highlights how specific feature contributions can be localized to an individual learner, allowing educators to pinpoint weaknesses—in this case, deficiencies in literacy and numeracy skills—and plan remedial interventions. The ability to trace each feature's weight in the model's decision path enhances transparency and accountability, addressing one of the core challenges of deploying black-box algorithms in educational contexts. Thus, LIME serves as a vital interpretability bridge between predictive accuracy and actionable pedagogical insight.

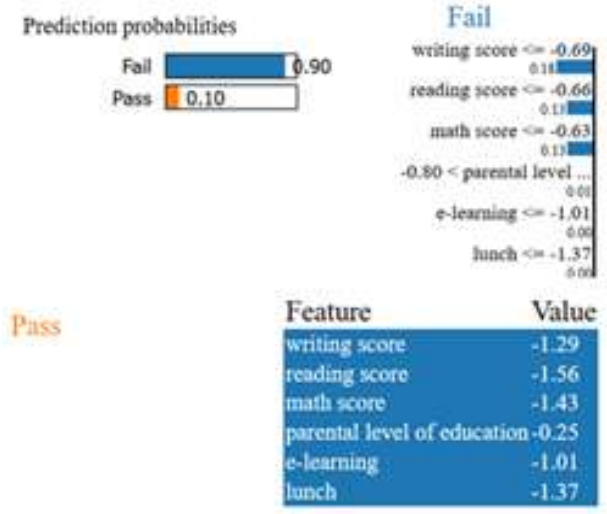

Figure 1 Lime

C. SHAP-Based Interpretability Analysis

The SHAP force plot (Figure 2) provides a local explanation for the classifier's decision regarding the selected student. The model output, f(x)=0.10f(x) = 0.10f(x)=0.10, represents the probability of the student being classified as "Pass." Given that this probability is substantially lower than the neutral baseline (~0.5), the model ultimately predicts the "Fail" class with high confidence.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 11, November 2025)

The visualization shows the base value (average model output across the dataset) near 0.5, and the bars indicate how individual feature contributions shift the prediction away from this baseline.

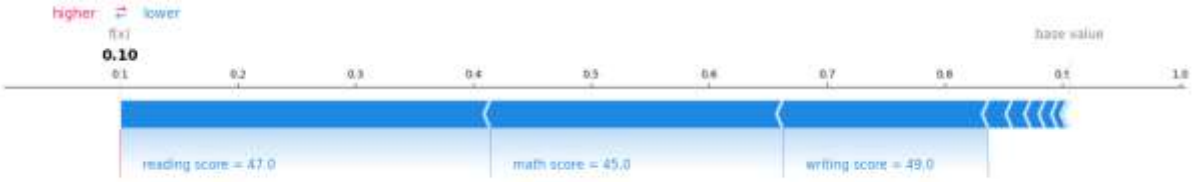

Figure 2 SHAP

Here, three academic performance features—reading score (47.0), math score (45.0), and writing score (49.0)—exert the most substantial negative influence on the model's output, pushing the prediction toward the "Fail" class. These are represented by blue bars, which denote features that decrease the likelihood of passing.

The absence of any red (positive) contributions indicates that none of the input features provided a strong counterbalance to offset the downward shift caused by low core subject scores. Essentially, the student's below-average performance in all three areas collectively outweighed any neutral or slightly positive socio-demographic factors.

From an interpretability standpoint, SHAP confirms the same insight derived from LIME: the classifier's decision for this student was dominated by subject-specific academic performance, while variables such as lunch type, e-learning access, or parental education had negligible impact on this particular outcome.

This type of feature attribution is critical in educational analytics—by isolating which variables most strongly influence predictions, educators can design targeted academic support interventions. In this case, the explanation clearly signals that remedial attention in reading, writing, and mathematics would be the most effective intervention strategy.

D. E-Learning Experiment and Explainability Analysis

To further extend the interpretability study, an additional experiment was conducted by introducing the e-learning variable to assess its influence on students' academic outcomes and its interaction with other socio-demographic features. The SHAP dependence plot in Figure 3 illustrates the marginal effect of e-learning on the prediction output, color-coded by math score. The SHAP values remain predominantly centered around zero, indicating that e-learning participation has a relatively subtle yet consistent influence on the model's decision boundary.

Notably, students engaged in e-learning (value = 1) tend to exhibit slightly higher SHAP values, suggesting a positive contribution toward the "Pass" prediction when accompanied by higher math scores. This pattern reflects that e-learning access alone may not drastically alter outcomes unless reinforced by strong foundational skills in core subjects.

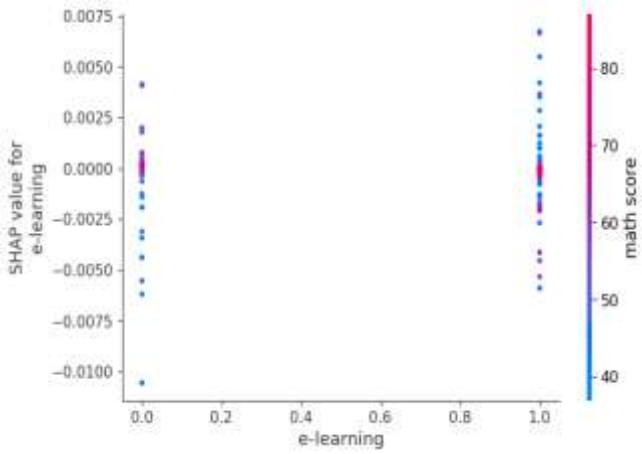

Figure 3 The marginal effect of e-learning on the prediction output, color-coded by math score

The interaction analysis in Figure 4 examines the relationship between e-learning and parental level of education. The visualization shows that students with both higher parental education levels and e-learning participation tend to have mildly positive SHAP values, implying a synergistic effect between home educational support and access to online learning environments. Conversely, students lacking e-learning access (value = 0) exhibit marginally lower SHAP values, particularly when parental education levels are limited, indicating a compounding disadvantage.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 11, November 2025)

These results highlight that while e-learning access independently exerts a modest impact, its combination with socio-demographic factors such as parental education and math competency can meaningfully influence academic success. The Random Forest classifier, supplemented by SHAP interpretability, effectively surfaces such nuanced relationships that might otherwise remain obscured in aggregate performance metrics. This integration of explainable AI enables educators and policymakers to identify the contextual value of digital learning interventions and supports targeted strategies for bridging equity gaps in line with NEP 2020 objectives.

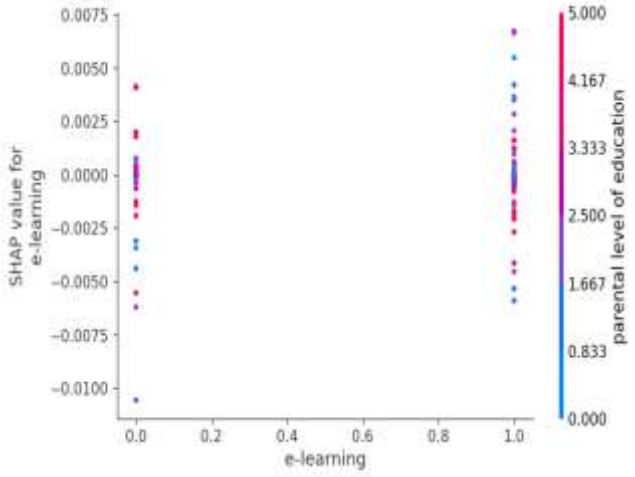

Figure 4 The relationship between e-learning and parental level of education

V. Conclusion

This study demonstrates that machine learning—specifically the Random Forest classifier—can reliably predict student academic outcomes using socio-demographic, academic, and e-learning–related features. While the model achieved exceptionally high predictive accuracy, the integration of LIME and SHAP was crucial in transforming this predictive capability into actionable educational insight. LIME provided transparent, instance-level explanations that pointed to specific learning deficiencies, while SHAP offered a consistent global understanding of feature importance, revealing that core subject scores overwhelmingly drive outcomes.

The extended e-learning experiment further highlighted that digital learning access exerts a modest yet positive influence, particularly when reinforced by strong academic performance and supportive socio-economic conditions such as higher parental education. Together, these findings underscore the value of Explainable AI in educational analytics—ensuring trust, fairness, and interpretability—while supporting data-driven interventions that align with NEP 2020 objectives and enhance equity and learner success in evolving digital education ecosystems.

REFERENCES

[1] L. Casal-Otero, M. Rosa, and E. Pérez, "AI literacy in K-12: A systematic literature review," International Journal of STEM Education, vol. 10, no. 1, pp. 1–22, 2023. [Online]. Available: https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-023-00418-7

[2] M. Giannakos, D. Chrysafiadi, and K. Papadopoulos, "The promise and challenges of generative AI in education," Interactive Learning Environments, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0144929X.2024.2394886

[3] S. Ghimire, "Explainable artificial intelligence-machine learning models," Education and Information Technologies, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24001346

[4] J. Yuan, C. Zhang, and M. Li, "Integrating behavior analysis with machine learning to predict online learning performance: A scientometric review and empirical study," arXiv preprint, arXiv:2406.11847, 2024. [Online]. Available: https://arxiv.org/abs/2406.11847

[5] J. Prentzas, K. Kapsalis, and P. Alexopoulos, "Explainable artificial intelligence approaches in primary education," Electronics, vol. 14, no. 11, p. 2279, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/11/2279

[6] U.S. Department of Education, "Artificial Intelligence and the Future of Teaching and Learning," Office of Educational Technology, Washington, DC, USA, 2023. [Online]. Available: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), pp. 1135–1144, 2016.

[8] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.

[9] S. P. Scientist, "Students Performance in Exams," Kaggle Dataset, 2018. [Online]. Available: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. [Accessed: 05-Jun-2025].