Explainable Machine Learning for Student Level Prediction: A Case Study in UAE Education Systems

Shatha Ghareeb[a], Krishna Sharma[a], Hagar Elbatanouny[b], Rawaa Al Jumeily[c], Bilal M. Khan[d,*], Wasiq khan[e], Abir Hussain[b,*]

[a] *Department of Computing & Games, Teesside University, Middlesbrough, U.K*
[b] *Department of Electrical Engineering , University of Sharjah, Sharjah, United Arab Emirates*
[c] *Founding Director and General Manager, Belvedere British School, Abu Dhabi, United Arab Emirates*
[d] *Institute of the Environment and Sustainability (IoES), University of California Los Angeles, California, USA*
[e] *School of Computer Science and Mathematics, Liverpool John Moores university, Liverpool, U.K*

Abstract

In recent years, the use of machine learning methods for predicting and evaluating student performance has become increasingly popular in the educational sector. This study examines how well machine learning models predict student academic performance, especially when it comes to curriculum transitions, which are a frequent problem in multicultural and multinational educational systems. This study investigates the predictive power of several machine learning algorithms, including Decision Trees, Support Vector Machines, Artificial Neural Networks, and ensemble techniques like Random Forest and eXtreme Gradient Boosting, using a real-world dataset gathered from UAE schools. With an accuracy and F1 score of 98.9%, support vector machines outperformed all other models in terms of prediction performance.

Beyond obtaining high predicted accuracy, it is critical to guarantee that the decisions made by these models are visible and interpretable, especially

*Corresponding author.

Email addresses: s.ghareeb@tees.ac.uk (Shatha Ghareeb), s3454618@tees.ac.uk (Krishna Sharma), U22102889@sharjah.ac.ae (Hagar Elbatanouny), r.aljumeily@belvederebritishschool.com (Rawaa Al Jumeily), m.bilal@ucla.edu (Bilal M. Khan), w.khan@ ljmu.ac.uk (Wasiq khan), abir.hussain@sharjah.ac.ae (Abir Hussain)

in sensitive fields such as education. By using explainable AI approaches like SHAP, LIME, and partial dependence graphs, this study emphasizes model interpretability beyond prediction, providing both local and global insights into model behavior. These methods demonstrated recurring trends in how academic term scores, especially in science, math, and English, affect student outcomes. This research advances data-driven educational practices that can guide curriculum preparation, promote early intervention, and customize student support by incorporating transparent and accurate machine learning models. In addition to being useful for UAE-based educational systems, the framework can be used to a variety of other educational situations.

Keywords: Student Failure Prediction, Explainable AI, Machine Learning, Educational Data Mining, Academic Risk Assessment.

---

1. Introduction

Over the last decade, the rapid improvements in technology have influenced a number of industries, including education [1]. Innovative strategies are becoming more and more necessary to assist and improve student learning as educational institutions develop and succeed [2]. Traditionally, grades, test results, and teacher evaluations have been used to measure academic achievement [3]. The entire range of elements impacting a student's academic path, including behavioral patterns, individual circumstances, past and present curriculum, and use of learning tools, is not fully captured by these approaches [4][5]. As a result, there is a growing interest in leveraging data and machine learning (ML) algorithms to predict student performance and develop more personalized, data-driven educational strategies [6][7].

In recent years, ML approaches have emerged as useful tools for identifying at-risk students and prediction academic outcomes with high precision [8]. Large amounts of organized and unstructured data, including as attendance, demographics, past academic records, and behavioral patterns, can be analyzed by ML models, which may reveal hidden tendencies and offer early warning signs of academic failure or underperformance [9]. For educators, these predicted insights are priceless since they allow for prompt and individualized interventions meant to avoid academic deterioration [6][7].

Though they are capable of making correct predictions, many ML models are viewed as "black boxes" since they don't disclose the reasoning behind their choices [10]. Since trust and transparency are essential in the educa-

tional setting, this lack of interpretability presents a serious obstacle to the adoption of ML [11]. Parents, teachers, administrators, and students must all be able to comprehend and respond to model outputs. Consequently, the incorporation of Explainable Artificial Intelligence (XAI), which aims to explain and interpret machine learning predictions, has become crucial in educational applications [12].

This study aims to address the challenge of predicting student failure and ensuring that model decisions are interpretable. This study explores a variety of machine learning models, including traditional classifiers (e.g., Decision Trees (DT), Support Vector Machines (SVM), and advanced ensemble learners (e.g., eXtreme Gradient Boosting (XGBoost), Voting Classifier), using a real-world dataset gathered from UAE schools that offer British and American curricula. The study also uses several explainability approaches, including interaction plots, Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) lines, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME), to analyze how the top-performing models make decisions.

This study proposes an interpretable machine learning framework for forecasting student academic failure, focusing on the heterogeneous educational setting of the United Arab Emirates (UAE). Unlike many prior efforts, this study provides cutting-edge predicted accuracy and incorporates a diverse set of explainability methodologies to improve transparency and practical understanding. The main contributions are as follows:

- **Explainable machine learning framework:** The framework provides both global and instance-level insights by incorporating multiple interpretability methods, such as SHAP, LIME, ICE plots, and 2D PDP interaction visualizations. This allows teachers and administrators to understand the underlying reasons for predictions and tailor interventions accordingly.

- **Feature interaction analysis for educational insight:** The model demonstrates how individual academic term scores and their interactions influence student outcomes, filling a major gap in existing research that frequently ignores how subject combinations effect risk.

- **UAE-specific academic risk prediction:** The study uses real-world data from UAE schools with varied curricula (British and American)

to build a strong model capable of identifying at-risk students, an issue specific to the UAE's multicultural education system.

The remaining sections of this work are organized as follows. In Section 2, the literature on machine learning applications in predicting student performance is reviewed, with an emphasis on the strengths and weaknesses of the present methods. The research methodology, including the datasets used, the machine learning models, and the evaluation criteria used, is covered in Section 3. The experiments' results are shown in Section 4, along with a discussion of the interpretability findings. Section 5 provides a final summary of the work and ends with recommendations for future research.

2. Literature Review

The use of ML in predicting student performance has grown significantly in recent years, driven by the vast amounts of educational data generated across institutions. As educational systems evolve, it becomes increasingly important to develop methods that not only predict academic success but also identify at-risk students early and provide appropriate interventions. The literature summarized in Table 1 illustrates a growing interest in leveraging machine learning for student performance prediction across diverse educational contexts.

In multicultural settings, such as the UAE, where there are multiple educational curriculums, machine learning has been applied to address challenges in student leveling. Ghareeb et al. [13] proposed a framework that uses ML algorithms to help place students in the right academic year group. This is crucial as curriculums vary significantly in terms of assessment techniques, exam boards, and academic year schedules. By implementing machine learning classifiers such as Random Forest (RF) and Artificial Neural Networks (ANN), their approach helped ease the transition for students switching between curriculums, ensuring smoother integration and better tracking of their academic progress.

Similar concerns regarding student performance prediction are tackled by Qureshi and Lokhande [14], who focus on the use of Educational Data Mining (EDM) to predict academic outcomes. The paper evaluates six classification algorithms (including RF, Decision Trees (DT), and SVMs) and identifies how factors such as academic performance, personal traits, and

family background influence a student's academic results. Their work underscores the importance of using diverse datasets to understand the various factors impacting student success.

Ahmed [15] built on this by exploring the predictive power of ML models in higher education, specifically in e-learning environments. Through techniques such as K-means clustering and Support Vector Machines, this study highlighted how machine learning can be used to predict student success based on their interaction with learning platforms. By identifying the features that contribute to success, the study also emphasized how these models can improve student outcomes and institutional rankings.

In a similar vein, Chandra and Kumar [16] investigated the role of ML in student placement prediction, looking at factors like technical and communication skills, as well as academic scores. They used various preprocessing and visualization techniques to better understand how these factors contribute to students' chances of securing job placements after graduation. This research highlighted how academic performance influences career outcomes and how ML can aid in predicting these opportunities, ultimately enhancing job placement strategies.

Lagrazon et al. [17] explored the application of ML in predicting success in licensure examinations for Electronics Engineering graduates. They demonstrated how ensemble models—by combining predictions from multiple algorithms can enhance accuracy in forecasting licensure exam results. Their study highlighted the potential of machine learning to improve educational strategies and curriculum design by providing actionable insights into students' academic progress.

Rimpy et al. [18] provided a comprehensive review of EDM techniques used for performance prediction, noting how machine learning helps identify patterns in educational data. They stressed the importance of early identification of weak students, enabling schools to intervene before poor performance impacts their academic journey. Their research serves as a reminder of the power of data mining in transforming educational systems by providing insights that can inform decision-making at various levels.

In higher education, machine learning has also been employed to predict performance based on both academic and non-academic factors. Bird [19] discussed the promises and challenges of predictive analytics in improving student success. While not providing specific data, they highlighted how ML can inform decisions about student support and academic strategies. Similarly, Issah et al. [20] conducted a systematic review of various ML methods

used to determine the factors influencing student performance. Their review confirmed that academic attributes, including grades and demographic factors, are the most influential in predicting outcomes. They also pointed out the gap in research on prescriptive intervention strategies, urging further exploration into how predictive models can be used to prevent poor academic performance.

The application of regression-based machine learning models in predicting student performance is also highlighted by Asthana et al. [21]. Their study proposed the concept of 'Learning Coefficients', a measure of students' learning potential, which can guide targeted interventions. By employing models like RF, SVM, and ANN, the study revealed that linear regression (LR) achieved the highest accuracy in predicting academic success. This demonstrated how regression models can provide valuable metrics not only for prediction but also for helping students improve their performance.

Graph-based approaches to student performance prediction are explored by Mubarak et al. [22], who used Graph Convolutional Networks (GCN) to classify students based on their engagement with course materials. Their study used a semi-supervised approach to classify students into behavioural categories such as "high engagement" and "at-risk", showcasing the power of GCNs in handling complex student interaction data. This model can help educators identify students who might need additional support based on their engagement patterns, thus providing more personalised and targeted interventions.

The review by Oppong [23] brings attention to the broader use of machine learning in predicting student performance, with an emphasis on neural networks. The study revealed that supervised learning techniques dominate the field, with neural networks yielding the best prediction accuracy. This finding highlighted the importance of using appropriate algorithms for various prediction tasks, particularly when dealing with complex datasets that include multiple variables influencing student success.

The novel deep learning model introduced by Fazil et al. [24] offered an innovative approach to performance prediction by incorporating student behaviour data, such as interaction with virtual learning environments (VLEs). Their system, ASIST, combined attention mechanisms with convolutional and bidirectional Long Short-Term Memory (BiLSTM) networks to predict student performance. The model's ability to process various behavioural and assessment data enables it to classify students into different performance categories, providing early predictions of academic success.

Table 1: Categorical features, content values and description.

Ref	Population	Region	ML model	Performance	Explainability
[13]	Students in UAE with multicultural curriculum transitions	UAE	RFC, ANN, Combined Classifiers	Accuracy (up to 0.816), AUC (up to 0.824), F1 (up to 0.602)	Not applied
[14]	Undergraduate Computer Engineering students (2nd–final year)	India	RF, DT, SVM, KNN, NB, LR, XGBoost	RF best avg. accuracy: 77.5%, SVM: 69.5%, XGBoost: 72.7%	Feature importance (embedded techniques), no SHAP/LIME
[15]	32,582 students from 2017–2022, Wollo University	Ethiopia (multiple regions incl. Oromia, Amhara, Tigrai, etc.)	SVM, DT, NB, KNN	SVM: 96.0%, DT: 93.4%, NB: 83.3%, KNN: 87.3%	Not addressed explicitly
[16]	Undergraduate students (2018–2021 batches)	India	Not explicitly stated (focus on preprocessing and visualization)	Not reported	Not applied (only feature correlation via visualization)
[17]	500 Electronics Engineering graduates (2014–2019)	Philippines (Southern Luzon State University)	Ensemble (Bagged Trees, RUSBoost), SVM, DT, NN, NB	Accuracy up to 98% (Ensemble, 80:20 split)	Not applied
[24]	6,068 students across 3 modules (business, science, humanities)	Ireland (University of Limerick)	ASIST (Attention-aware CNN + Stacked BiLSTM)	AUC: 0.86–0.90 across datasets; F1: up to 0.83; Acc: up to 0.82	Attention mechanism used to highlight important features; ablation study for behavior analysis
[21]	91 undergraduate CSE students	India (Amity University, Lucknow)	LR, RF, DT, SVR, ANN	LR: 97% accuracy; ANN: 79%	Correlation analysis of proposed learning coefficients with academic features
[22]	Students from Stanford MOOCs (CS courses)	USA (MOOC data, Stanford)	Graph Convolutional Network (GCN)	Accuracy: 84%, F1-score: 78%	Not explicitly applied, though interpretability via t-SNE visualization discussed
[25]	1268 university students (Engineering programs)	Türkiye	Hybrid PSO-DNN (Particle Swarm Optimization + Deep Neural Network)	Accuracy: 0.633 (custom data), 0.806 (xAPI-Edu-Data); F1-score: 0.561	SHAP (global & local), LIME (local)
[26]	872 undergraduate students	Bangladesh	Stacking ensemble (RF + GB base, SVC meta)	Accuracy: 86.38%, F1-score: 86.49%	SHAP, LIME
[27]	University students (preparatory year, course-level)	Saudi Arabia	RF, SVM, ANN	Accuracy: 75.4%–80.8%	SHAP (global), LIME (local), global surrogate model

The existing body of research demonstrates considerable progress in applying machine learning to student performance prediction; however, a closer examination reveals a persistent gap when it comes to combining high model performance with actionable explainability—particularly in the context of the Middle East, and the UAE in particular. While some models achieve commendable predictive accuracy, they often lack transparency in their decision-making processes. Conversely, studies that apply explainable AI methods tend to overlook more nuanced techniques such as feature interaction analysis, or fail to fully leverage these tools to support meaningful educational

interventions.

This disconnect becomes especially critical in multicultural and rapidly evolving educational systems like those in the UAE, where curriculum transitions and diverse student backgrounds add complexity to academic assessment. Despite the region's increasing focus on AI integration in education, there remains a shortage of frameworks that predict student outcomes reliably and also provide educators, parents, and policymakers with interpretable insights into why those predictions were made. This study addresses that gap by introducing a methodology that blends strong predictive performance with comprehensive explainability, enabling targeted support strategies and informed decision-making within the UAE's dynamic educational landscape.

3. Methodology

This work uses a machine learning pipeline to predict student failure and generate transparent, interpretable results to help data-driven educational interventions. Figure 1 shows the whole methodological flow, which includes five essential phases: data collection and cleaning, preprocessing, model training, model evaluation, and XAI. The methodology was developed to ensure that all aspects of student performance are captured effectively and that the best model is selected based on performance metrics.

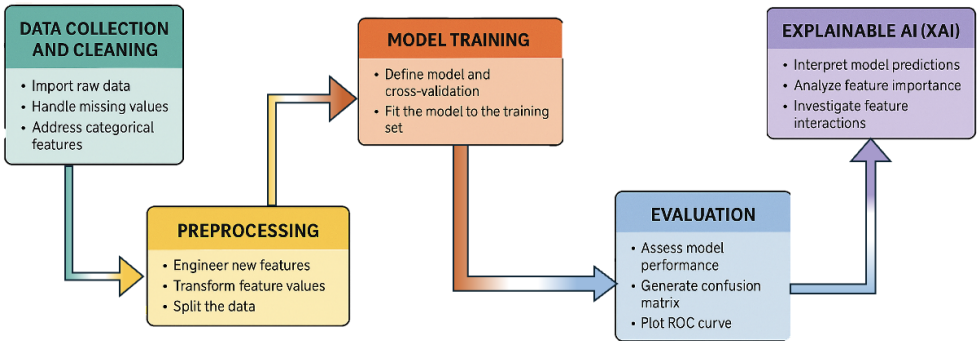


Figure 1: Machine learning pipeline for student level prediction.

3.1. Dataset

This study used a publicly available educational dataset created by Ghareeb et al. [28] from two international schools in Abu Dhabi, United Arab

Emirates—one following the British curriculum and the other the American curriculum. Ethical approval was obtained through Liverpool John Moores University (Ref: 19/CMS/001), and formal approval was granted by the Ministry of Education and Knowledge in Abu Dhabi. The data were obtained through institutional gatekeepers and completely anonymized to ensure compliance with data protection and AI ethical requirements, erasing any sensitive or personally identifiable information.

The dataset contains 1500 rows (unique student information) and 30 columns that capture a variety of variables. These features include both categorical and numerical data types, providing a comprehensive perspective of students' academic and demographic profiles throughout two academic years (2017-2018 and 2018-19).

The categorical attributes are personal information (e.g., gender and age), academic advancement indicators (e.g., current year, intended year, and year of admission), and school system identifiers (e.g., previous curriculum, current curriculum, previous school, and current school). These characteristics provide vital context for understanding each student's educational path and are required for predicting the effects of curriculum transfer and academic placement. Table 2 provides a full overview of the categorical features, including values and descriptions.

In addition to categorical data, the dataset contains a collection of numerical features that serve as the primary academic performance indicators. These include entrance exam scores in Mathematics, Science, and English, all scored out of 100. For each of the two academic years, three term-wise performance scores per subject are recorded as percentages. These numerical features are critical in feature engineering and help assess student performance trends over time. Table 3 contains a complete summary of the numerical features, including values and descriptions. These features are used as input variables for model training, while the target variable is a binary class label indicating whether the student is expected to succeed (class 0) or fail (class 1) academically.

3.2. Preprocessing

The preprocessing phase involved comprehensive data cleaning to ensure consistency and reliability. This included correcting column names, fixing spacing and symbol issues, and standardizing text formatting using custom Python scripts. Addressing these inconsistencies was essential to prevent errors during downstream modeling.

Table 2: Categorical features, content values and description [28].

Attribute name	**Value**	**Description**
Gender	Male/Female	Gender of the student
Student age (as of 2017/18)	6, 7, 8, 9,... etc	Age of the student calculated from 2017/18 academic year
Year of admission	Old BBS Student, Old GEMS Student, New Admission 18/19	The data collected is for two or more academic years; 2017/18 and before academic years + 2018/19 academic year
Current year (17/18)	Foundation Stage 1, Foundation Stage 2, Year 1–12/Grade 1–11	This is the year or grade group assigned to the student by the school
Proposed year/grade (18/19)	Foundation Stage 1, Foundation Stage 2, Year 1–13/Grade 1–12	This is the year or grade group assigned to the student by the school
Previous school (17/18)	Many schools in UAE	Previous schools that the student was in before this study
Previous curriculum	British/ American/ MOE/ Canadian/ Indian/ Australian/ CBSE/ German.	The curriculum the student transferred from
Current school	GEMS, Belvedere	Name of the school that the data has been collected
Current curriculum	Binary value	The curriculum the student transferred to

Missing values were primarily found in numerical exam-related fields and were handled using mean imputation. A new target variable was derived by averaging students' scores across all terms and entrance exams. This continuous metric was then binarized into two classes: students with averages above 80% were labeled as academically strong (Class 0), and those below as at-risk (Class 1).

Categorical variables were transformed using one-hot encoding to enable proper model interpretation. Min-Max Normalization was applied to scale all numerical features to the [0,1] range, ensuring balanced feature contributions and improving model convergence. These preprocessing steps established a

Table 3: Numerical features, content values and description [28].

Attribute name	**Value**	**Description**
Math-exam	Mark out of 100	Exam marks for school entry exam in math
Science-exam	Mark out of 100	Exam marks for school entry exam in science
English-exam	Mark out of 100	Exam marks for school entry exam in English
Math19-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2018/19
Science19-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2018/19
English19-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2018/19
Math19-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2018/19
Science19-2	Percentage out of 100%	Term 2 student science Exam marks during academic year 2018/19
English19-2	Percentage out of 100%	Term 2 student English Exam marks during academic year 2018/19
Math19-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2018/19
Science19-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2018/19
English19-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2018/19
Math20-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2019/20
Science20-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2019/20
English20-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2019/20
Math20-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2019/20
Science20-2	Percentage out of 100%	Term 2 student science Exam marks during academic year 2019/20
English20-2	Percentage out of 100%	Term 2 student English Exam marks during academic year 2019/20
Math20-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2019/20
Science20-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2019/20
English20-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2019/20

clean, normalized, and well-structured dataset for training robust machine learning models.

3.3. Model Training and Explainability

This work used a variety of supervised machine learning algorithms to create a predictive method for identifying students who are at risk of academic failure. To provide consistent comparison, the chosen models are trained and assessed independently using the same dataset, reflecting both traditional and ensemble-based approaches. The models employed in this work are DT, SVM, RF, XGBoost, AdaBoost, stacking, K-Nearest Neighbors (KNN), Extra Trees, ANN, Bernoulli Naive Bayes and ensemble voting classifiers.

Table 4: Parameters of Machine Learning Models Used in the Study

Model	**Key Parameters**
Random Forest	`n_estimators=100`, `random_state=42`
ANN (MLP)	`hidden_layer_sizes=(50,)`, `max_iter=1000`, `random_state=42`
SVM	`kernel='linear'`, `random_state=42`, `probability=True`
KNN	`n_neighbors=5`
Bernoulli Naive Bayes	Default parameters
AdaBoost	`random_state=42`
XGBoost	`random_state=42`
Extra Trees	`random_state=42`
Stacking	`estimators=[`RF, ANN, KNN`]`, `final_estimator=LogisticRegression()`
Voting Classifier	`estimators=[`RF, ANN, SVM, KNN`]`, `voting='soft'`

All models were trained using an 80:20 train-validation split, where 80% of the data was used for model training and the remaining 20% was reserved for validation. To ensure optimal performance, each model underwent hyperparameter tuning using Optuna, allowing the exploration of multiple parameter configurations to identify the most effective setup for each classifier.

The performance of the machine learning models that were put into practice was evaluated using a number of evaluation measures that gave a thorough picture of the quality of classification and the ability to generalize to new data. Accuracy, precision, recall, F1 score, and the confusion matrix are among the metrics that were chosen.

This study used a number of XAI strategies that offer both global and local interpretability in order to increase transparency and foster confidence in the model's predictions. These methods were used to interpret individual

predictions, investigate interaction effects, and assess overall feature relevance in the top-performing model.

Globally, Permutation Feature Importance was used to quantify how much each feature affected model accuracy, while PDP and ICE lines illustrated how individual and average changes in features influenced predictions [29]. Two-dimensional PDPs further revealed how feature combinations, such as exam scores and past academic performance, interacted to impact student risk classification.

For local explanations, SHAP force plots and LIME were applied to visualize how specific features influenced predictions for individual students. SHAP offered consistent, game-theory-based explanations of feature contributions, while LIME created interpretable surrogate models around specific predictions [29].

The study guaranteed interpretability at the dataset and individual levels by combining these several explainability techniques. The ethical application of AI in education is supported by this combination of accuracy and transparency, which gives educators and stakeholders accurate, data-driven insights into student outcomes.

4. Results and Discussion

This section highlights the results of using various machine learning models to predict how well students perform in school. Accuracy, precision, recall, F1-score, and ROC AUC are some of the frequently used classification metrics that are part of the evaluation framework. Graphical interpretations are provided by ROC curves and confusion matrices. This section also provides a thorough review of explainability techniques, including SHAP, LIME, permutation importance, partial dependence plots using ICE, and 2D plots, in order to better understand model decision-making.

4.1. Performance Evaluation

The first step was to ensure a balanced class distribution throughout both the training and testing datasets. This was accomplished by thorough preprocessing, which included removing bias-introducing variables such as "Year_of_Admission" and excluding records linked with underrepresented groups (e.g., new students with only 103 instances). As a result, the final training set had 594 and 524 instances of Class 0 and Class 1, respectively,

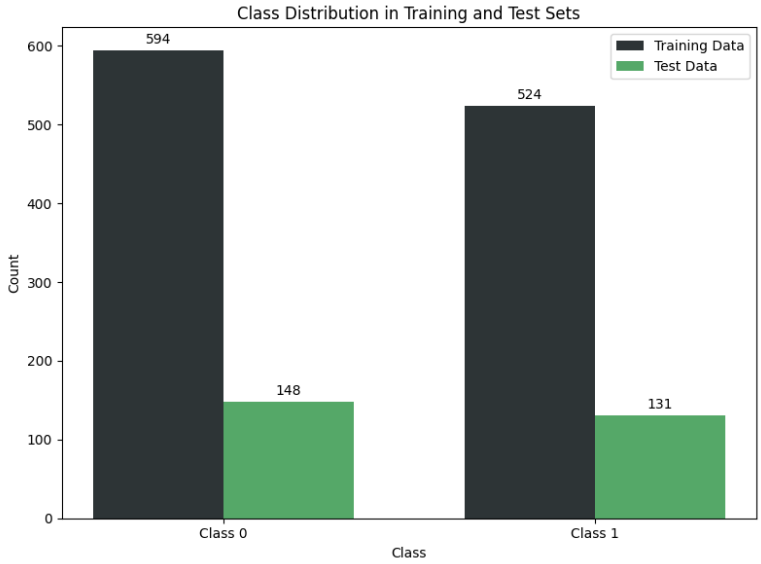


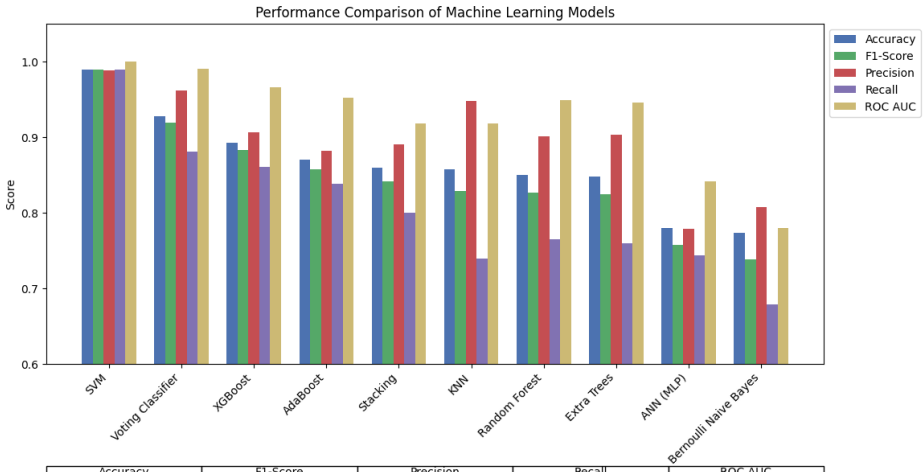
Figure 2: Class Distribution in Training and Test Sets.

while the test set had 148 and 131. Figure 2 visually verifies that the class distribution is balanced across both datasets.

All models in this study were tuned using Optuna to ensure optimal configurations, and the best-performing hyperparameters are presented in Table 5. This tuning stage was critical for ensuring a fair and accurate comparison of various model types.

Table 5: Best Hyperparameters for Each Model after Tuning

Model	**Best Parameters**
Random Forest	`n_estimators=52, max_depth=2, min_samples_split=6, min_samples_leaf=4`
MLP (ANN)	`hidden_layer_sizes=170, alpha=0.0005476, learning_rate_init=0.0001435`
SVM	`C=0.4554`
KNN	`n_neighbors=13`
AdaBoost	`n_estimators=199, learning_rate=0.9908`
XGBoost	`n_estimators=194, max_depth=2, learning_rate=0.2993`
Extra Trees	`n_estimators=74, max_depth=3`



	Accuracy	F1-Score	Precision	Recall	ROC AUC
SVM	0.989	0.989	0.988	0.989	1.000
Voting Classifier	0.928	0.919	0.962	0.881	0.991
XGBoost	0.893	0.883	0.906	0.861	0.966
AdaBoost	0.870	0.858	0.882	0.838	0.952
Stacking	0.860	0.842	0.890	0.800	0.918
KNN	0.858	0.829	0.948	0.739	0.918
Random Forest	0.850	0.827	0.901	0.765	0.949
Extra Trees	0.848	0.825	0.903	0.760	0.946
ANN (MLP)	0.780	0.758	0.779	0.744	0.842
Bernoulli Naive Bayes	0.773	0.738	0.808	0.679	0.780

Figure 3: Performance Comparison of Machine Learning Models.

Figure 3 compares tuned models across several performance indicators. The SVM outperformed all other classifiers, delivering high results with accuracy, precision, recall, and F1-scores of 0.989 and a ROC AUC of 1.000. These findings demonstrate the SVM's capacity to generalize effectively on unseen data, making it the most reliable and accurate model in this predictive framework.

Other models, including ensemble approaches such as Voting Classifier, XGBoost, and AdaBoost, performed well. While they did not outperform SVM, they did show stability and effectiveness at capturing complex decision boundaries. Tree-based approaches such as Random Forest and Extra Trees produced reasonable results, although with lower recall scores, indicating some limits in detecting at-risk pupils. Simpler models, such as KNN and ANN, showed higher variations and scored poorly in contrast, indicating difficulties in adjusting to the dataset's structure despite tuning. Overall, while some models produced acceptable results, SVM consistently outperformed in terms of accuracy and reliability.

The ROC AUC analysis, shown in Figure 4 and Table 6, provides a more intuitive view of class separability among models. As expected, SVM had a perfect AUC of 1.000, followed by the Voting Classifier (0.991) and

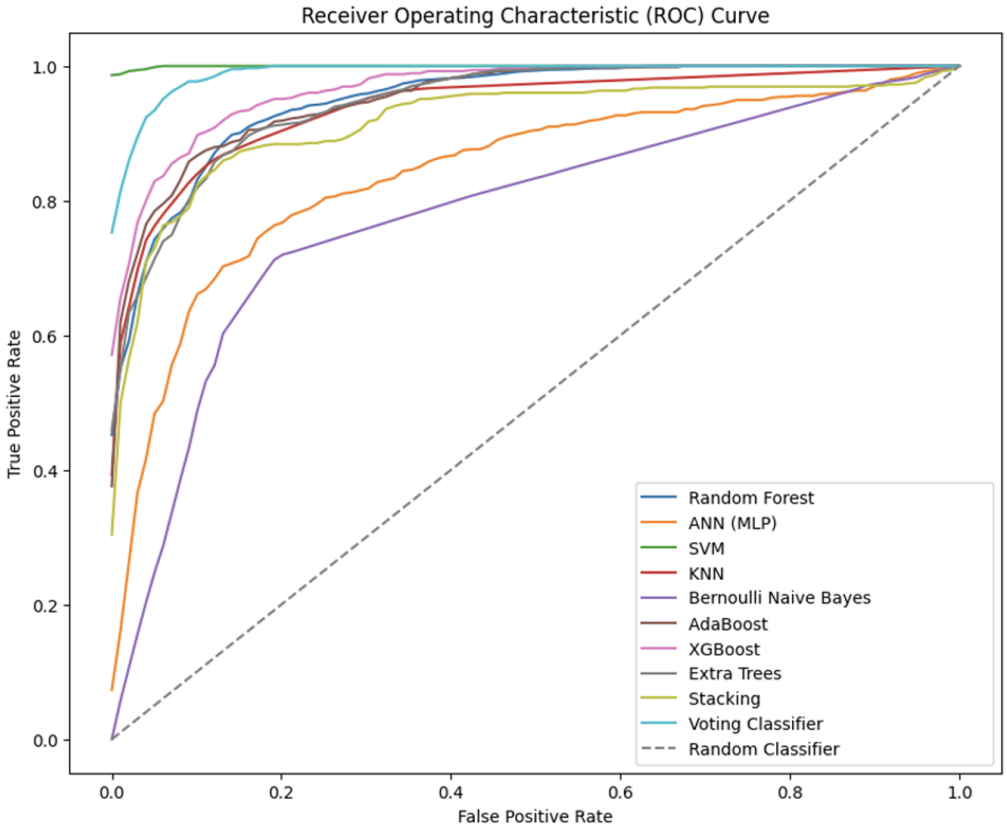


Figure 4: ROC curve for each model.

XGBoost (0.966). AdaBoost, Random Forest, and Extra Trees all displayed strong discriminatory strength, with AUCs greater than 0.94. In contrast, the ANN (0.842) and Naive Bayes (0.780) models performed poorly, supporting previous conclusions reached from the key metrics.

Table 6: Tabular representation of ROC curve for each model

Model	SVM	Voting Classifier	XGBoost	AdaBoost	Stacking	KNN	Random Forest	Extra Trees	ANN	Bernoulli Naive Bayes
ROC AUC	1.000	0.991	0.966	0.952	0.918	0.940	0.949	0.946	0.842	0.780

The confusion matrix components—false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN)—are broken out in depth in Figure 5 to offer a better understanding of model behavior. The SVM model showed only one false positive and zero false negatives, demonstrating good control over both types of errors. Similarly, the Voting Classifier and

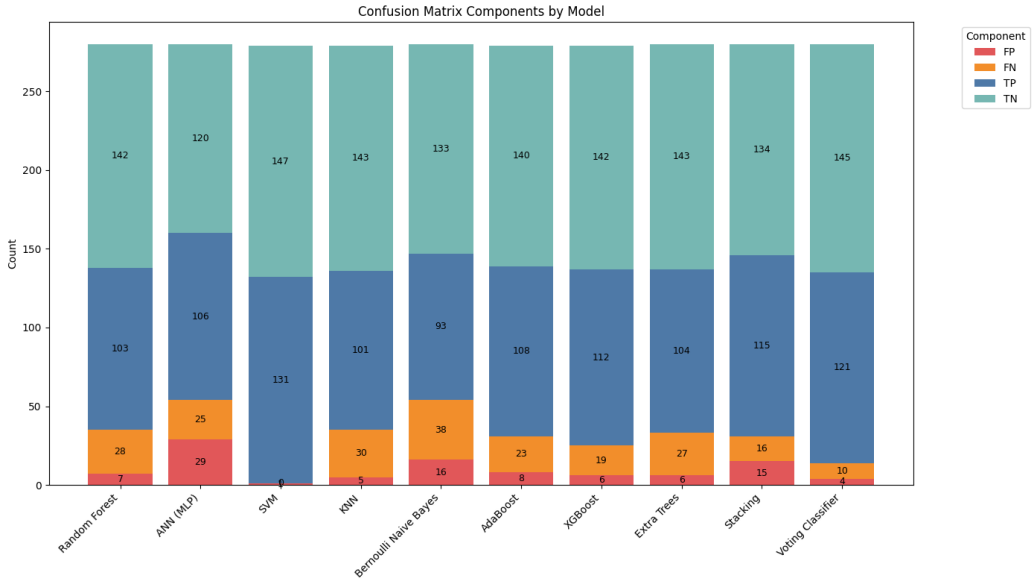


Figure 5: Confusion Matrix Components by Model.

XGBoost produce balanced and dependable results, with low misclassification rates and high true positive counts.

In contrast, models such as Bernoulli Naive Bayes and ANN have greater false negative rates, indicating a proclivity to overlook students at risk of poor performance. Despite their high precision, KNN and Extra Trees have moderate false negative rates, indicating a relative conservatism in predicting positive cases.

4.2. Explainability

To gain a better understanding of the SVM model's decision-making process, which demonstrated the best overall performance across all evaluation metrics, several explainability techniques were used, such as permutation importance, PDP with ICE lines, and 2D interaction plots. These methods contribute in determining which features most strongly influenced predictions, particularly for class 1, which corresponds to students expected to fail.

The Permutation Importance plot in Figure 6 shows that numerical academic performance features like 'Mathexam', 'Math191_', and 'English191_' have the largest influence on the model's output. These variables were ranked

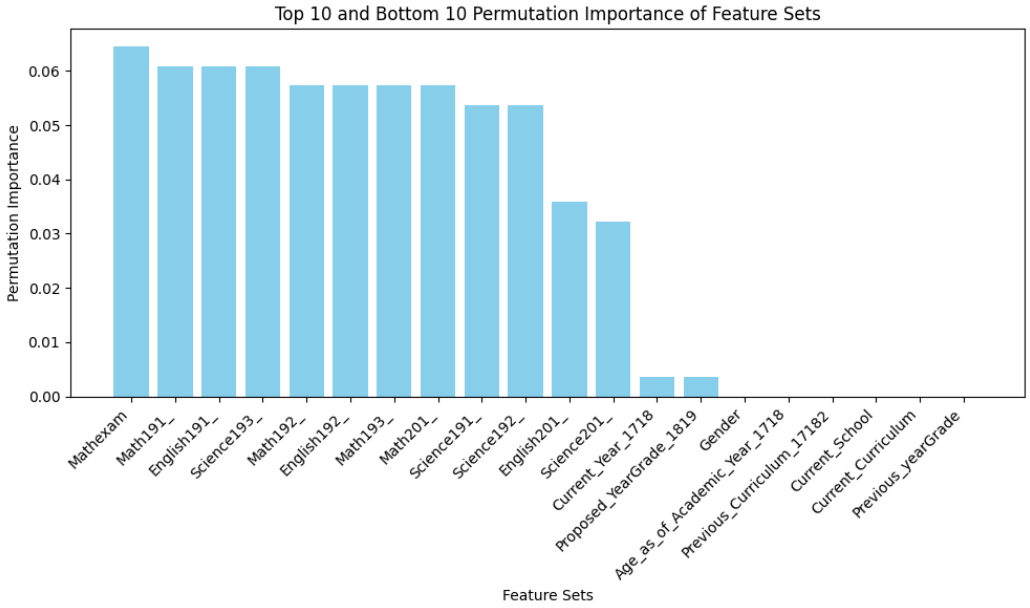


Figure 6: Top 10 and Bottom 10 Permutation Importance.

first in relevance, indicating that the model depends largely on subject-specific scores to evaluate student success. In contrast, categorical characteristics such as 'Gender', 'Previous_YearGrade', and curricular codes contributed minimally, implying that they had little influence on the model's assessment of pupils as at risk of failing.

PDP plots with ICE lines were created in Figures 7-9 for features 'Math192_', 'Science192_', and 'English192_' in order to investigate how feature values impact the likelihood of predicting class 1 (fail). Each feature shows a distinct downward trend, indicating that higher scores in these subjects are linked to a lower probability of failure. This confirms that the model's projections match actual academic risk and is in line with educational intuition. These variables represent student performance in the second term, when students tend to show more consistency and adaptability to the academic environment, which is why the 192-term subject scores were chosen. The first term may be a transitional period with varied performance due to unfamiliarity with subjects or learning conditions. While maintaining the overall pattern seen in the global trend, the addition of ICE lines highlights individual diversity across student records.

2D PDP interaction plots were created in Figures 10-13 to further examine

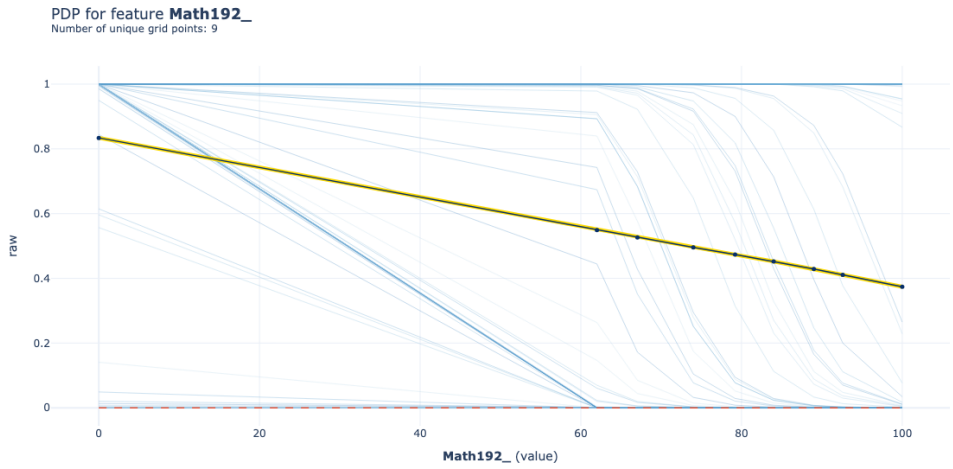


Figure 7: PDP plot for term 2 student Maths Exam marks during academic year 2018/19.

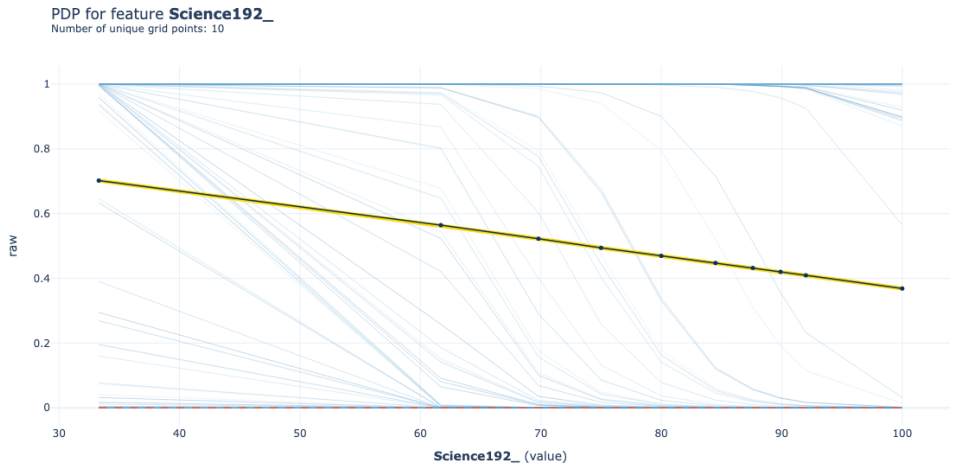


Figure 8: PDP plot for term 2 student science Exam marks during academic year 2018/19.

how feature combinations affect the model's predictions for the failure class. These graphs show how two variables work together to affect the model's output, highlighting the significance of each feature separately as well as how features interact. The interaction plot between 'Math191_' and 'Math192_' in Figure 10 shows that the failure probability is highest when both scores

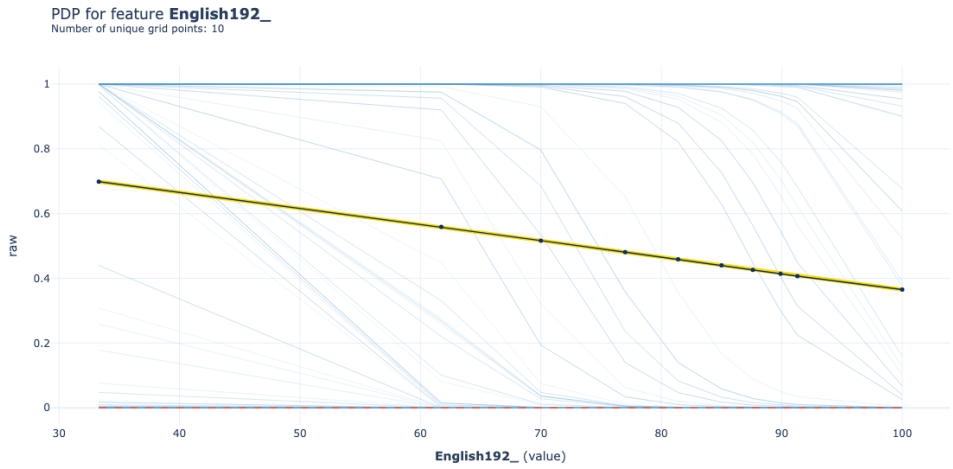


Figure 9: PDP plot for term 2 student English Exam marks during academic year 2018/19.

are low and drastically decreases as either or both scores improve. Students who perform poorly in both terms are more likely to be categorized as at risk, which emphasizes the significance of consistent performance throughout terms. Similar to this, Figure 11 shows an interaction between 'Science192_' and 'Math192_'. Students who perform poorly in both courses have a high chance of failing, whereas those who score well in one or both disciplines experience a significant decrease in predicted risk.

Figure 12 illustrates another interaction between 'Mathexam' and 'Englishexam' that shows how entrance-level test results may be combined together to support the model's risk prediction. The plot demonstrates the model's sensitivity to imbalance in basic knowledge by showing that, even in cases when one entrance subject score is strong, a very poor score in another subject may still increase the failure probability. In contrast, the response surfaces of interactions between categorical features, like 'Previous_Curriculum_17182' and 'Current_Curriculum' in Figure 13, are flat or uniform, suggesting that they have little effect on the prediction. This demonstrates that while calculating the likelihood of failure, the model gives priority to academic performance indicators above background category characteristics.

A global interpretation of feature contributions for both success and failure predictions was provided via SHAP summary graphs. High SHAP values

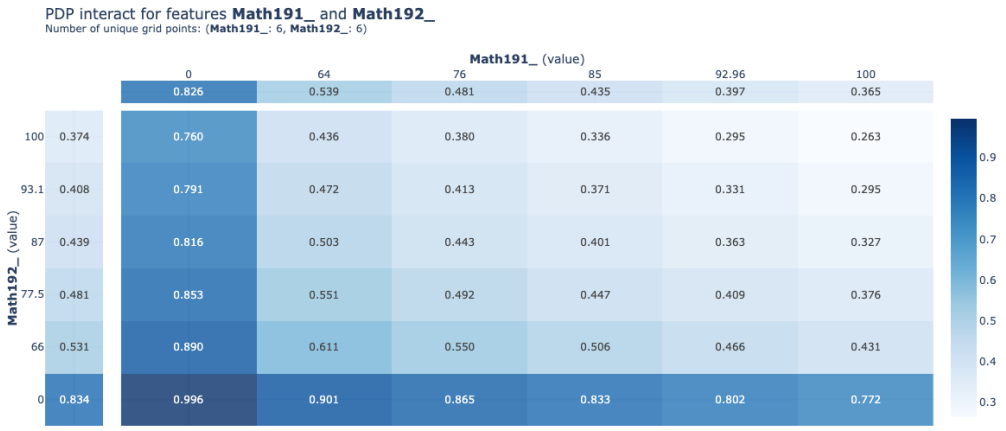


Figure 10: 2D interacting plots (failure) for the features 'Math191_' and 'Math192_'

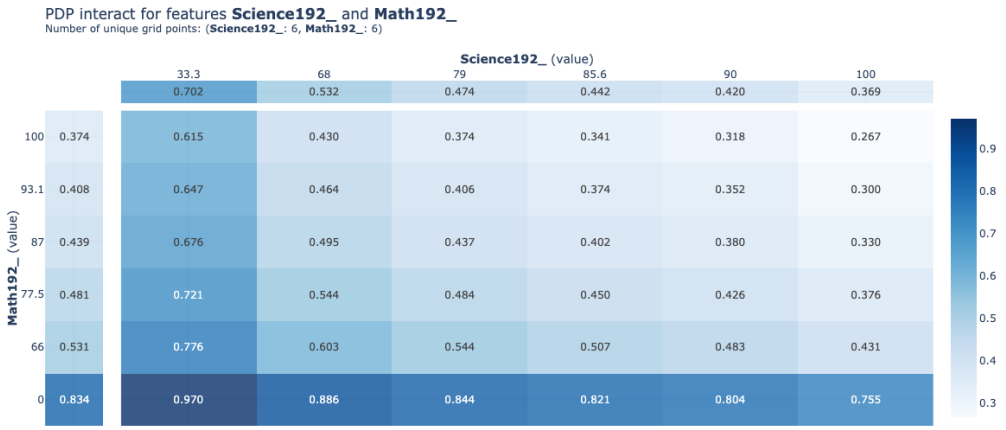


Figure 11: 2D interacting plots (failure) for the features 'Science192_' and 'Math192_'

are closely linked to higher input values across important academic courses like 'Math191', 'Math192', and 'Science193' as shown in Figure 14a, which represents class 0 (success). This suggests that students are more likely to be categorized as successful if they have higher grades in these subjects, and the model heavily depends on these characteristics for assigning a favorable

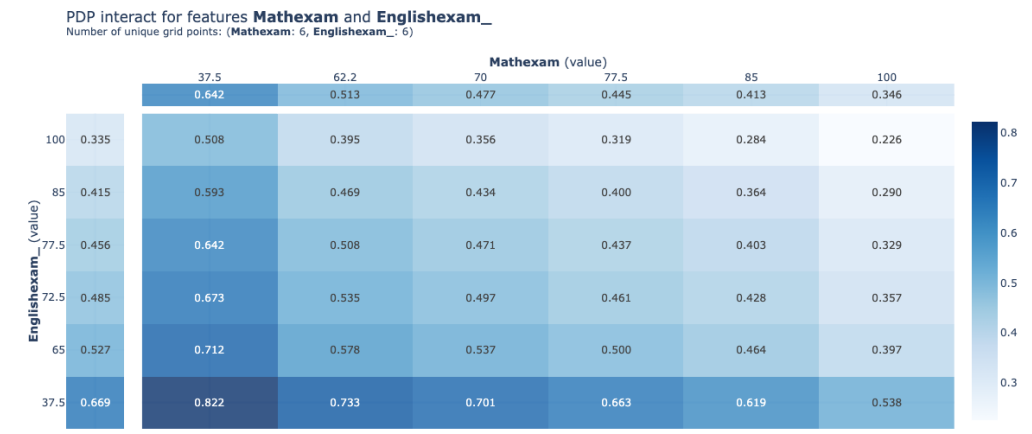


Figure 12: 2D interacting plots (failure) for the features 'Mathexam' and 'Englishexam'

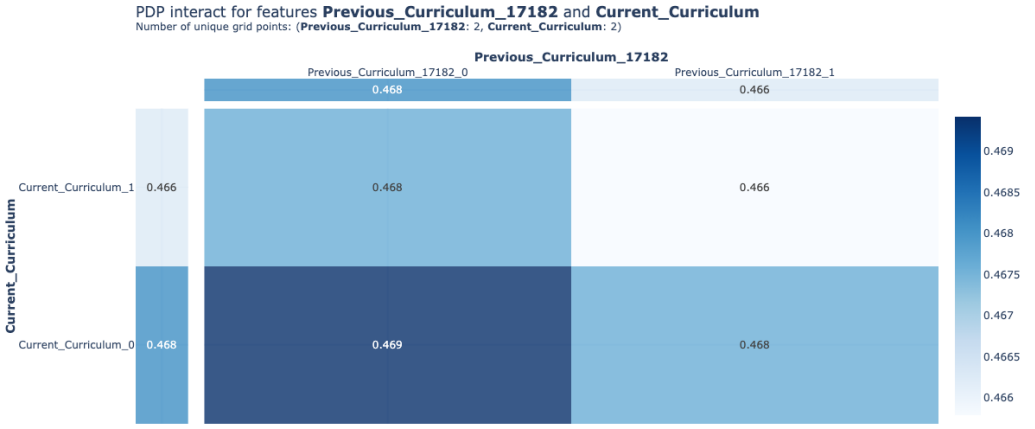
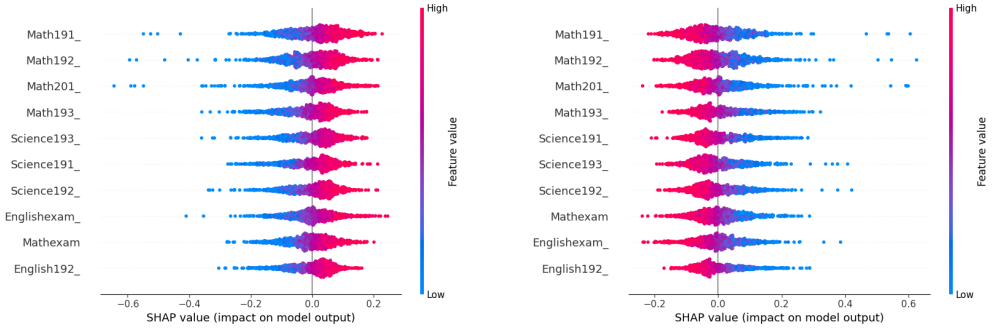


Figure 13: 2D interacting plots (failure) for the features 'Previous_Curriculum_17182' and 'Current_Curriculum'

outcome.

On the other hand, the SHAP summary for class 1 (fail) is shown in Figure 14b, where stronger SHAP contributions to failure predictions are linked to lower values of the same academic courses. The model's susceptibility to score distributions is demonstrated by the inversion of impact between classes. It

is noteworthy that 'Math192' frequently ranks among the top contributors in both classes, highlighting its significance. The validity of the model's interpretation and its reliance on recent academic achievement to distinguish between successful and struggling students are confirmed by this alignment with permutation important results.



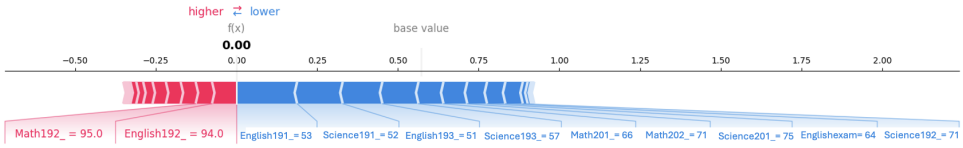
(a) Class 0 – Success (b) Class 1 – Failure

Figure 14: SHAP summary plots showing global feature impact for both success and failure classes.

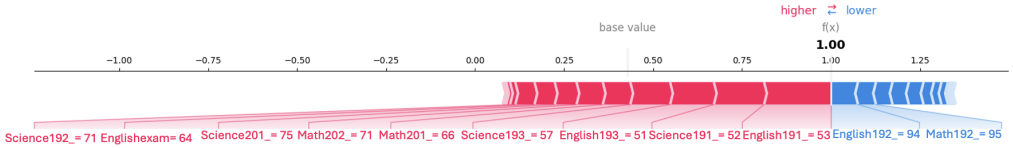
The local explainability tools SHAP and LIME in Figures 15 and 16 offer additional insight into the underlying causes of this prediction for Student 55, who was categorized as "Fail." The dataset indicates that this student is male and ages between 17 and 18. What distinguishes this student, though, is the trend in his academic performance over terms and topics, which finally caused the model to give the fail class a high probability.

For class 1, the SHAP force diagram in Figure 15b shows that a combination of moderate to low scores in 'Science192' = 71, 'Science201' = 75, 'Math201' = 66, and 'Englishexam' = 64.5 substantially pushes the prediction toward failure. The model consistently found underperformance in foundational topics, as seen by the early-term scores such as 'Science193' = 57, 'English193' = 51, and 'English191' = 53. The high scores in 'Math192' = 95 and 'English192' = 94 demonstrate improvements, however they are displayed in blue, suggesting that they act against the failure prediction but were not enough to offset the cumulative negative contributions.

This balance is evident in the class 0 SHAP force plot in Figure 15a, where the many weaker features, like 'Science193', 'English191', and 'Math201', which appear in blue and push the prediction to the left, overpowered the high-performing features that tried to push the model toward a "pass" classi-



(a) Local Shap plot for Class 0 – Success



(b) Local Shap plot for Class 1 – Failure

Figure 15: Local SHAP force plots for student # 55.

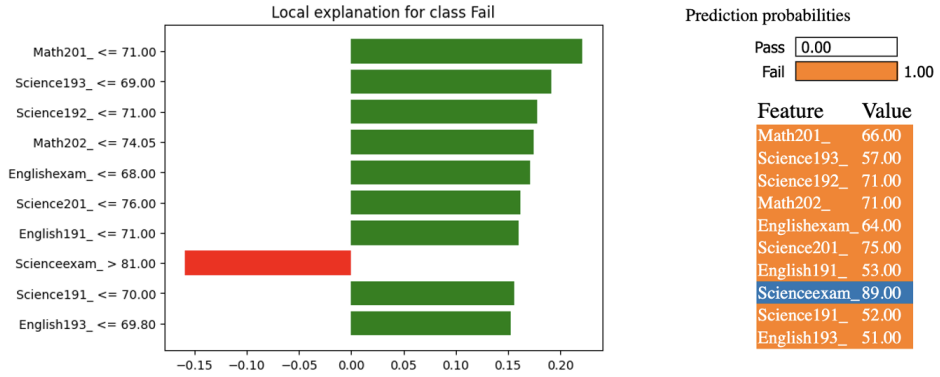


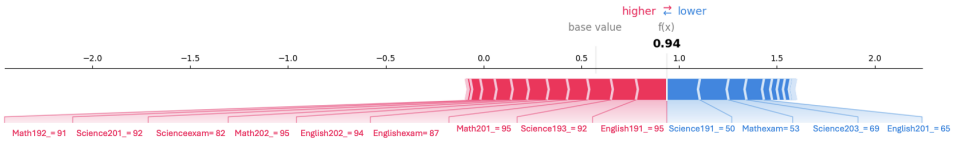
Figure 16: LIME explanation plot for student # 55.

fication. The importance of early academic patterns in the model's decision-making is highlighted by this difference between factors that are supportive and those that are antagonistic.

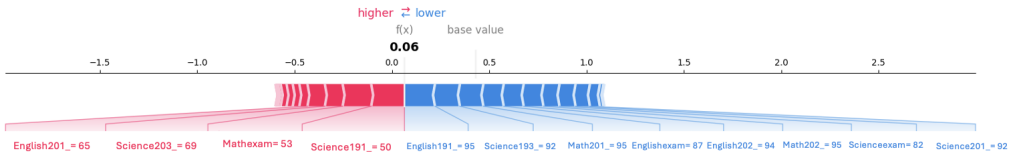
The SHAP is further supported by the LIME explanation for Student 55 in Figure 16, which highlights certain academic results that helped the model classify this student as likely to fail. The science score in term 3 of 2018/2019 (=57), the mathematics score in term 1 of 2019/2020 (=66), and the science score in term 2 of 2018/2019 (=71) were the most significant

negative contributions. The prediction model places a strong emphasis on fundamental courses, and their comparatively moderate to low performances point to a lack of persistent academic proficiency in these areas.

Student 267 is a 7-year-old female at Belvedere school under the British curriculum. According to the SVM expectations, she was appropriately identified as a student who is likely to succeed (class 0). The local SHAP force plot in Figure 17a clearly shows that recent academic performance plays the main role in driving the forecast to success. Despite a few low marks in earlier terms (e.g., Term 1 English and Term 3 Math), the Class 1 SHAP plot in Figure 17b shows that these had little opposing influence, with pink bars representing features that pushed the prediction marginally toward failure.



(a) Local Shap plot for Class 0 – Success



(b) Local Shap plot for Class 1 – Failure

Figure 17: Local SHAP force plots for student # 267.

This interpretation is further reinforced by the LIME explanation shown in Figure 18. LIME recognizes the same collection of high-value features — particularly Term 2 Math and English marks from 2019/2020 — as key contributors to the "Success" prediction. In contrast, earlier performance (e.g., Term 1 English and Term 1 science) was among the negatively contributing factors, but it was offset by greater recent academic signals. The clear rising trend in grades across terms was significant, indicating both academic improvement and topic knowledge. Thus, the model correctly predicted a positive academic outcome for this student. The consistency between SHAP and LIME explanations demonstrates the interpretability and reliability of the model's conclusions and confirms the prediction's robustness.

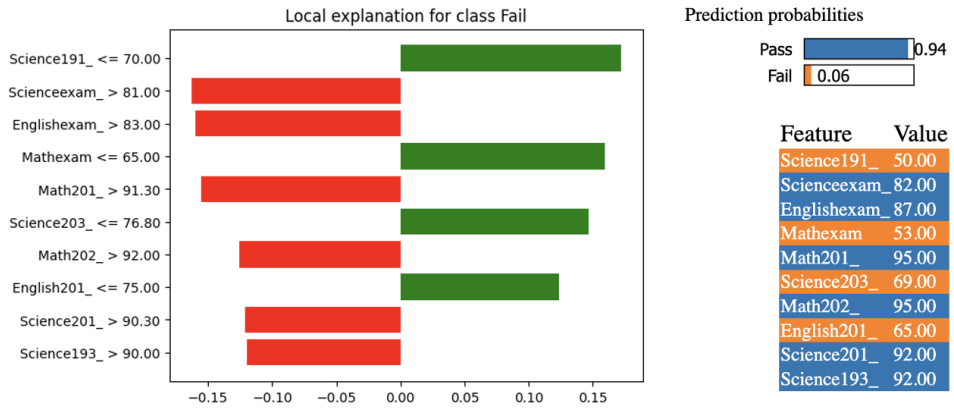


Figure 18: LIME explanation plot for student # 267.

4.3. Discussion

The present study advances the field of student performance prediction by combining high predictive accuracy with rich, multi-level interpretability. Unlike many previous works, which emphasize prediction or interpretability, but rarely both, this research offers a framework applicable to multicultural education systems, such as the UAE.

In terms of data region, our study is distinctly localized, leveraging real-world data from UAE schools following both British and American curricula. This specificity is rare; while some prior works like Ghareeb et al. [13]] also used UAE data, their scope was narrower and primarily focused on academic placement rather than performance prediction. Others, such as [24] (Ireland), [25] (Turkey), and [26] (Bangladesh), employed region-specific datasets but lacked the multicultural complexity seen in UAE systems.

From a performance standpoint, the proposed framework delivers state-of-the-art results. The SVM model achieved an outstanding accuracy and F1-score of 0.989 and a perfect AUC of 1.000, surpassing the benchmarks reported in studies such as [24], where the ASIST model achieved an F1-score of up to 0.83, and [25], which reported a maximum accuracy of 0.806 on the xAPI-Edu-Data dataset. These improvements are attributable to both algorithmic tuning via Optuna and thoughtful feature engineering based on term-wise academic scores.

Regarding explainability and interpretability, our work surpasses many of the recent contributions. While [24] used an attention mechanism to ex-

plain BiLSTM-based predictions, and [[25]–[27]] employed SHAP and LIME, our study integrates a more diverse suite of interpretability tools: global methods (Permutation Importance, SHAP summary plots, PDP, ICE), local methods (SHAP force plots, LIME), and interaction-based explanations (2D PDP). This multi-angle strategy supports model transparency and provides actionable insights, which many reviewed studies either overlook or treat superficially.

Notably, our framework stands apart by emphasizing feature interaction analysis (e.g., 'Math191' vs. 'Math192'), which is absent in most reviewed works. For example, [26] and [27] provide both global and local explainability but do not delve into interaction effects, which are critical for understanding academic trajectories across multiple terms.

5. Conclusion and Future Work

Academic underperformance is still a significant issue for teachers, and the ability to effectively support at danger students is frequently limited by delayed interventions. This study addressed this issue using a dataset collected from students in the UAE, where students were drawn from multiple international curricula. We investigated the application of machine learning models to predict student performance in order to provide prompt and individualized interventions in response to this challenge. SVM outperformed the other models in precision, recall, and F1-score, all of which had a score of 0.989, confirming its appropriateness for binary classification tasks in the educational field. The use of explainable AI strategies to improve usability and transparency is a key component of this research. Permutation importance and SHAP summary plots, two global interpretability techniques, have repeatedly shown that Term 2 scores in science, math, and English are among the most important factors affecting model predictions. Lower academic achievement in these disciplines significantly raised the probability of failure predictions.

Local interpretability techniques, such as SHAP force plots and LIME, offered tailored explanations that aided in figuring out the reasoning behind the prediction.

Although this study was conducted using data from UAE-based students, the methodology and insights are broadly applicable to other international and multicultural educational contexts. The flexible design of the machine

learning pipeline and the emphasis on explainability make the approach suitable for generalization across schools with diverse student populations. In order to improve the model's interpretability and usability in actual decision-making situations, future research should broaden the feature set to incorporate behavioral, economic, and longitudinal data.

References

[1] R. Raja, P. Nagasubramani, Impact of modern technology in education, Journal of applied and advanced research 3 (1) (2018) 33–35.

[2] D. Sun, Y. Zhan, Z. H. Wan, Y. Yang, C.-K. Looi, Identifying the roles of technology: A systematic review of stem education in primary and secondary schools from 2015 to 2023, Research in Science & Technological Education 43 (1) (2025) 145–169.

[3] H. A. Gallagher, Vaughn elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement?, Peabody Journal of Education 79 (4) (2004) 79–107.

[4] M. Thomran, A. E. Alshammari, A. Al-Subari, H. Ahmed, Investigating the role of psychological elements in advancing it skills among accounting students: insights from saudi arabia, Humanities and Social Sciences Communications 12 (1) (2025) 1–12.

[5] C. L. Huang, L. Fu, S.-C. Hung, S. C. Yang, Effect of visual programming instruction on students' flow experience, programming self-efficacy, and sustained willingness to learn, Journal of Computer Assisted Learning 41 (1) (2025) e13110.

[6] N. F. Ab Rahman, S. L. Wang, T. F. Ng, A. S. Ghoneim, Artificial intelligence in education: A systematic review of machine learning for predicting student performance.

[7] J. Wang, Y. Yu, Machine learning approach to student performance prediction of online learning, PloS one 20 (1) (2025) e0299018.

[8] S. Alturki, I. Hulpuș, H. Stuckenschmidt, Predicting academic outcomes: A survey from 2007 till 2018, Technology, Knowledge and Learning 27 (1) (2022) 275–307.

[9] A. D. Smith, Event detection in educational records: an application of big data approaches, International Journal of Business and Systems Research 15 (3) (2021) 271–291.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[11] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Computers and education: artificial intelligence 3 (2022) 100074.

[12] K. Fiok, F. V. Farahani, W. Karwowski, T. Ahram, Explainable artificial intelligence for education and training, The Journal of Defense Modeling and Simulation 19 (2) (2022) 133–144.

[13] S. Ghareeb, A. J. Hussain, D. Al-Jumeily, W. Khan, R. Al-Jumeily, T. Baker, A. Al Shammaa, M. Khalaf, Evaluating student levelling based on machine learning model's performance, Discover Internet of Things 2 (1) (2022) 3.

[14] R. Qureshi, P. S. Lokhande, A comprehensive review of machine learning techniques used for designing an academic result predictor and identifying the multi-dimensional factors affecting student's academic results, in: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), IEEE, 2024, pp. 1–6.

[15] E. Ahmed, Student performance prediction using machine learning algorithms, Applied Computational Intelligence and Soft Computing 2024 (1) (2024) 4067721.

[16] K. S. Kumar, et al., Data preprocessing and visualizations using machine learning for student placement prediction, in: 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), IEEE, 2022, pp. 386–391.

[17] P. G. G. Lagrazon, J. E. E. Japor, M. R. D. De Veluz, R. R. Maaliw, F. T. Villa, M. C. B. Abejo, L. P. Arroyo, J. F. S. Marqueses, Ensemble-based prediction model for enhanced electronics engineering licensure

examination results using student performance analysis, in: 2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), IEEE, 2023, pp. 1–5.

[18] A. Dhankhar, K. Solanki, et al., Educational data mining tools and techniques used for prediction of student's performance: A study, in: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), IEEE, 2022, pp. 1–5.

[19] K. Bird, Predictive analytics in higher education: The promises and challenges of using machine learning to improve student success, AIR Professional File (12 2023). doi:10.34315/apf1612023.

[20] I. Issah, O. Appiah, P. Appiahene, F. Inusah, A systematic review of the literature on machine learning application of determining the attributes influencing academic performance, Decision analytics journal 7 (2023) 100204.

[21] P. Asthana, S. Mishra, N. Gupta, M. Derawi, A. Kumar, Prediction of student's performance with learning coefficients using regression based machine learning models, IEEE Access 11 (2023) 72732–72742.

[22] A. A. Mubarak, H. Cao, I. M. Hezam, F. Hao, Modeling students' performance using graph convolutional networks, Complex & Intelligent Systems 8 (3) (2022) 2183–2201.

[23] S. O. Oppong, Predicting students' performance using machine learning algorithms: a review, Asian Journal of Research in Computer Science 16 (3) (2023) 128–148.

[24] M. Fazil, A. Rísquez, C. Halpin, A novel deep learning model for student performance prediction using engagement data., Journal of Learning Analytics 11 (2) (2024) 23–41.

[25] A. Kala, O. Torkul, T. T. Yildiz, I. H. Selvi, Early prediction of student performance in face-to-face education environments: A hybrid deep learning approach with xai techniques, IEEE Access (2024).

[26] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, F. Akter, An explainable ai-based approach for predicting undergraduate students academic performance, Array 26 (2025) 100384.

[27] S. Alwarthan, N. Aslam, I. U. Khan, An explainable model for identifying at-risk student at higher education, IEEE Access 10 (2022) 107649–107668.

[28] S. Ghareeb, A. Hussain, W. Khan, D. Al-Jumeily, T. Baker, R. Al-Jumeily, Dataset of student level prediction in uae, Data in Brief 35 (2021) 106908.

[29] L. Gianfagna, A. Cecco, Explainable AI with Python, 2021. doi:10.1007/978-3-030-68640-6.