Student Performance Predictor

1st Rwittik Sarker
Department of CSE
ID: 22101634
BRAC University
Dhaka, Bangladesh
rwittik.sarker@g.bracu.ac.bd

2nd Samiha Tahiat
Department of CSE
ID: 22101737
BRAC University
Dhaka, Bangladesh
samiha.tahiat@g.bracu.ac.bd

3rd Hasin Saleh
Department of CSE
ID: 24141161
BRAC University
Dhaka, Bangladesh
hasin.saleh.alvi@g.bracu.ac.bd

***Abstract*—Early prediction of student performance is crucial for enabling timely support through tutoring, counseling, and adaptive resources. This research develops machine learning models using demographic data, prior grades, attendance, LMS activity, and assessment records to generate accurate and interpretable risk assessments. Models such as linear regression, decision trees, random forests, SVMs, neural networks, and ensemble methods are evaluated using mean squared error, and R^2 score. Results show that demographic and academic features are strong predictors, with ensemble methods offering superior performance. Interpretability tools like SHAP and LIME reveal key risk factors, making predictions actionable for educators. The study identifies the most effective models and provides insights to support early-warning systems and improve student retention through data-driven interventions.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

A. Research Problem

Many educational institutions need early, reliable signals about which students are at risk of poor performance so they can provide targeted interventions (tutoring, counselling, adaptive resources). Existing manual monitoring is labor-intensive and often too late. This project aims to build and evaluate machine-learning models that predict student performance early in a course using available student background, academic, and in-course activity data. The goal is to get accurate prediction about the most important risk factors so that educators can act on them.

B. Research Objective

The objectives of this study are as follows:

1) Collect and preprocess a representative student dataset (e.g., UCI Student Performance dataset or institutional records [1]), and engineer features from demographics, prior grades, attendance, LMS activity, and assessments.
2) Build multiple predictive models, including linear regression, decision trees/Random Forest, support vector machines (SVM), and at least one neural network–based model.
3) Evaluate model performance using appropriate metrics (e.g., Mean Squared Error (MSE), R^2 score) and select the best-performing model(s).
4) Analyze feature importance and generate interpretable explanations (e.g., SHAP, LIME) to surface actionable factors for educators.
5) Compare individual models with ensemble approaches and evaluate benefits of ensembling.

II. Literature Review

Machine learning (ML) and data mining have shown strong potential in predicting student performance and enabling early interventions. Studies highlight that integrating diverse data sources—such as academic records, LMS activity, and surveys—enhances prediction accuracy and robustness. For instance, [2] identifies internal (e.g., CGPA, attendance, quiz scores) and external (e.g., demographics, family background) factors, showing improved outcomes using deep neural networks over logistic regression or SVMs. Feature selection is also emphasized in [3] to eliminate irrelevant attributes, boosting accuracy and reducing processing time. Decision trees remain the most widely used and interpretable algorithm, followed by SVMs, neural networks, and ensemble models. A systematic review [4] confirms that classification techniques, especially Decision Trees, Bayesian Networks, and Random Forests dominate performance prediction studies, often relying on demographic and academic features. Ensemble methods generally outperform single classifiers, as shown in [5], where a hybrid model combining Decision Tree, ANN, and SVM achieved 81.67% accuracy. Similarly, [6] found that hybrid Random Forest models reached up to 99.72% accuracy, significantly outperforming standard models. Most research targets higher education and supports data-driven intervention strategies, though challenges remain. These include inconsistent metrics, lack of generalizability, and limited use of explanatory models. Namoun and Alshanqiti, after reviewing 62 studies, advocate for broader outcome measures beyond grades and call for stronger ensemble and interpretability-focused approaches [7]. Overall, ML is a valuable tool for education, but further research must address methodological gaps and improve validation across diverse learning contexts.

III. Dataset

The dataset used in this study is derived from the UCI Student Performance dataset, specifically from two educational datasets: `student-mat.csv` (Mathematics course) and `student-por.csv` (Portuguese language course). Both datasets describe student demographics, social, and school-related features, as well as their academic performance. The cleaned dataset combines both sources, resulting in a total of:

- **Number of instances:** 682 students
- **Number of attributes:** 68 attributes
- **Missing values:** None
- **Duplicate rows:** None

A. Attribute Information

The dataset includes the following types of features:

- **Demographic information:** school, sex, age, address, family size, parental status.
- **Parental background:** mother's education (Medu), father's education (Fedu), mother's job (Mjob), father's job (Fjob).
- **Academic support:** school support classes, family support, extra paid classes, higher education aspiration.
- **Social and lifestyle factors:** extracurricular activities, nursery attendance, internet access, romantic relationships, family relationships, free time, social outings, alcohol consumption (workdays and weekends), health condition.
- **Academic performance:** absences, grades (G1, G2, G3).

B. Target Variable

Although three grade variables are available (`G1`, `G2`, and `G3`), the final grade (`G3`, ranging from 0 to 20) is used as the primary output variable. It represents the students' ultimate performance and therefore serves as the target label for predictive modeling.

C. Notes

- The dataset contains 382 students who appear in both the Math and Portuguese datasets. They are identified by matching demographic and family attributes.
- Attributes are a mix of binary, nominal, and numeric values, making the dataset suitable for classification and regression tasks.

D. Exploratory Data Analysis

Several visualizations were generated to better understand the dataset:

- **Distribution of Final Grades (G3)**

Figure 1 presents histograms of final grades(G3) students in two subjects: Math(left,blue) and Portuguese (right,orange)

Math(G3_math): The grades are spread with a massive peak at grade 0. Further, most of the other students clustered between 10-13. This reflects more failure and more uneven performance in math.

Portuguese(G3_por): Grades are more normally distributed, peaking around 10-13, with fewer extreme failures. This suggests that performance is generally stronger,more consistent and balanced around average.

- **Coorelation between Grades**

Figure 2 displays the correlation between student grades in math and Portuguese over three grading periods (G1, G2, G3). It reveals that:

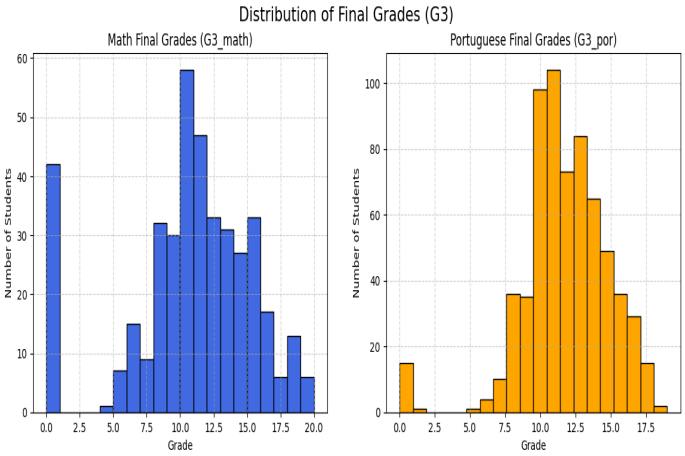

Fig. 1. Distribution of Final grades (G3).

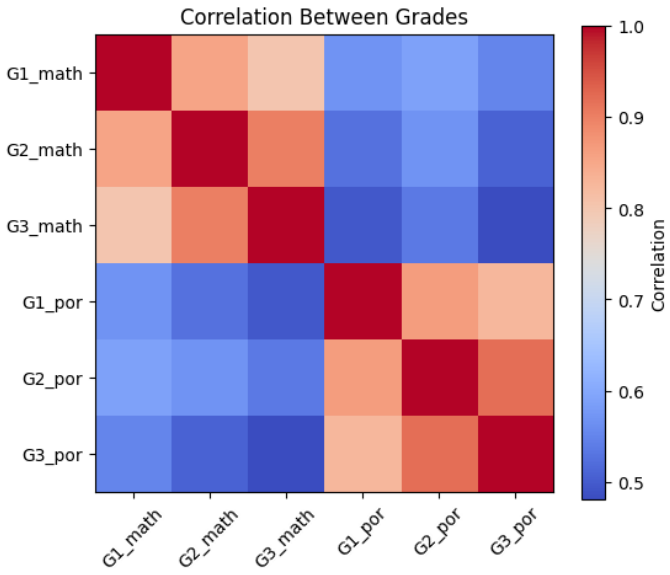

Fig. 2. Correlation Between Grades

Grades within each subject (math or Portuguese) are highly correlated over time, showing consistent performance across terms.

Math and Portuguese grades have only moderate correlations, suggesting that students who perform well in one subject don't necessarily perform equally well in the other. The color gradient (red = strong correlation, blue = weak) visually emphasizes these patterns.

- **Average Final Math Grade vs. Study Time**

Figure 3 compares the average final grades (G3) in Math (left) and Portuguese (right) in different categories of weekly study time.

- **Math(left,blue):** Average grades increase slightly with more study time, but the effect is modest and the error bars (variability) are wide. Even students studying $>$ 10h don't score much higher than those studying less.

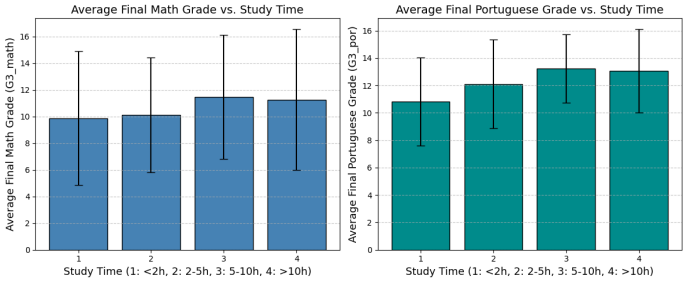

Fig. 3. Final Grade vs Study Time

- **Portuguese(right,green):** Average grades show a clearer positive trend with study time: students who study more achieve higher grades, with less overlap in variability compared to math.

IV. Methodology

A. Preprocessing

Before modeling, several preprocessing steps were applied to clean and prepare the dataset:

- **Dropping Irrelevant Columns:** Columns such as `reason`, `guardian_math`, and `guardian_por` were removed based on correlation analysis, as they showed little predictive power for final grades. This reduced dimensionality and noise.
- **Handling Missing Values:** The dataset had no missing values, confirmed during initial exploration, so no imputation was needed.
- **Feature Scaling:** Due to differing numeric ranges (e.g., `absences` up to 93 vs. `study time` scaled from 1 to 4), standardization was applied to ensure balanced feature contribution.
- **Duplicate Removal:** No duplicate rows were found, so no records were removed.
- **Encoding Categorical Variables:** Nominal and binary categorical features (e.g., `sex`, `address`, `schoolsup`) were transformed using one-hot encoding or binary mapping to make them compatible with machine learning algorithms.
- **Merging Math and Portuguese Datasets:** Data from Mathematics and Portuguese subjects were merged by matching students using demographic and family attributes, ensuring consistency and avoiding duplicates.

These steps improved efficiency, reduced redundancy, and ensured the dataset was ready for training.

B. Models

In this study, multiple machine learning techniques were applied to predict students' final academic performance. Each method was chosen for its predictive power and interpretability in educational data mining.

- **Decision Tree For Multi output Regression:** A Decision Tree for Multi-Output Regression is an extension of regression trees where the model predicts multiple continuous target variables simultaneously instead of just one. At each split, the tree chooses the feature and threshold that minimizes the combined error (e.g., sum of squared errors) across all target outputs.
- **Random Forest Regressor for Multi-Output Regression:** A Random Forest Regressor for Multi-Output Regression extends the standard random forest to predict multiple continuous target variables simultaneously. It builds an ensemble of decision trees, where each tree outputs a vector of predictions instead of a single value. The final prediction is obtained by averaging across trees for all outputs.
- **Linear Regression / Multiple Linear Regression** For predicting continuous values such as final grades, linear regression models the relationship between input features and the target variable using a linear function. Multiple linear regression incorporates several predictors simultaneously, allowing the analysis of how various social, demographic, and academic features influence performance.
- **Support Vector Regression (SVR) for Multi-Output Regression:** Support Vector Regression (SVR) for Multi-Output Regression adapts the conventional SVR, which is designed for a single continuous target, to predict multiple dependent variables simultaneously. By extending its margin-based learning framework, multi-output SVR can generate predictions for several continuous outcomes, making it particularly effective when the outputs are interrelated (e.g., predicting student grades across different subjects).
- **Neural Network (MLPRegressor) for Multi-Output Reression:** It employs a multilayer perceptron to predict multiple continuous target variables simultaneously. The network learns shared hidden representations from the input features and outputs a vector of predictions, one for each target.In our neural network two hidden layers are used. The first hidden layer has 100 neurons and second hidden layer has 50 neurons.
- **Manual Voting Ensemble:** Instead of relying on a built-in ensemble method, we can manually combine the models by averaging their predictions. In this ensemble approach, SVR, Linear Regression and Random Forest is used which made the prediction step of a trained VotingRegressor, mathematically equivalent. By blending outputs from these diverse models, the ensemble often achieves a more accurate or at least more stable performance compared to any single model alone.

Each of these models was trained, validated, and compared to evaluate their predictive accuracy and interpretability. The ensemble methods (Random Forest) typically offered the best balance between accuracy and robustness, while linear regression provided interpretable baselines.

C. Evaluation Metrics

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Here, equation 1 is Mean Squared Error (MSE) which measures the average squared difference between the actual values (y_i) and the predicted values ($\hat{y}_i$). A lower MSE indicates better model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Here, equation 2 is the coefficient of determination (R^2) which indicates the proportion of variance in the dependent variable (y_i) that is explained by the model. Values closer to 1 imply stronger explanatory power.

D. Model Interpretability

Knowing which features are important is not enough; educators need interpretable explanations to act on them.

- **SHAP (SHapley Additive exPlanations):** Uses game theory to assign each feature a contribution to an individual prediction. Example: For a struggling student, SHAP might show that low attendance contributed 40% and poor quiz scores 30% to the predicted risk, helping teachers provide targeted support.
- **LIME (Local Interpretable Model-agnostic Explanations):** Explains one prediction at a time by locally approximating the model with a simple, interpretable model. Example: For an "at-risk" student, LIME might highlight LMS activity and class participation as main drivers, clarifying why the system flagged the student.

V. Result and Analysis

The dataset was evaluated using regression models, since the target variable (G3) is continuous. The performance was measured using Mean Squared Error (MSE) and the coefficient of determination (R^2). Multiple experiments were conducted on different train-test splits and feature configurations, hence multiple result rows per model.

A. Comparative Insights

Table I compares six machine learning models—Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), Neural Network, and the proposed Voting Ensemble—for predicting student grades in Mathematics and Portuguese across three periods. Performance is evaluated using Mean Squared Error (MSE) and R^2, where lower MSE and higher R^2 indicate better predictions. The proposed Voting Ensemble—combining SVR (20%), Random Forest (60%), and Linear Regression (20%)—achieves the best results with the lowest average MSE (5.862) and highest R^2 (0.266), showing the benefit of leveraging complementary model strengths. Random Forest alone is the top individual model (MSE = 5.961, R^2 = 0.250), reinforcing its performance. Linear Regression also performs well, often rivaling SVR and Neural Networks, suggesting the dataset has moderate linearity. In contrast, Decision Tree performs poorly (MSE = 12.484, $R^2 = -0.602$), likely due to overfitting. Overall, the results highlight the robustness of the Voting Ensemble in balancing accuracy, complexity, and interpretability for grade prediction.

- **Comparison of Model Performance (Mean Squared Error)**

Fig. 4. Comparison of Model Performance (Mean Squared Error)

Figure 4 compares model performance based on Mean Squared Error (MSE). Voting Ensemble and Random Forest consistently achieve the lowest errors, while Decision Tree performs worst with much higher errors, highlighting the stability of ensemble methods and the overfitting tendency of single trees. and the tendency of single trees to overfit.

- **Comparison of Model Performance (R^2 Score)**

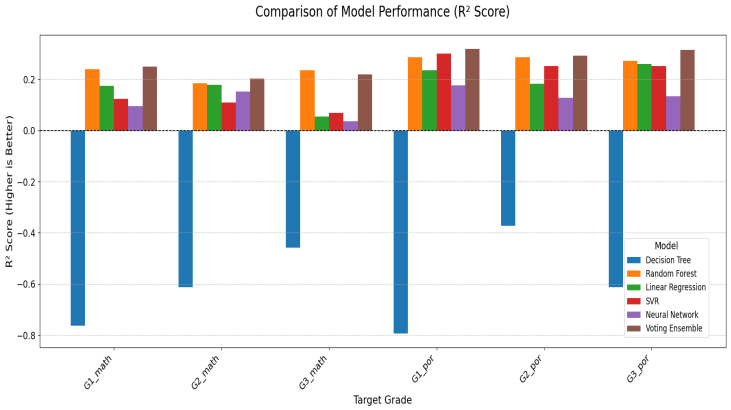

Fig. 5. Comparison of Model Performance (R^2 Score)

Figure 5 compares model performance based on R^2 scores. The Voting Ensemble shows the highest predictive power with consistently positive scores, while the Decision Tree performs worse than a mean predictor, yielding negative R^2 values.

SHAP Explanation for Random Forest Model

The following section explains three SHAP (SHapley Additive exPlanations) plots used to interpret the Random Forest

TABLE I
COMPREHENSIVE MODEL PERFORMANCE COMPARISON INCLUDING AVERAGE SCORES

Model	G_1 (Math)		G_2 (Math)		G_3 (Math)		G_1 (Port.)		G_2 (Port.)		G_3 (Port.)		Average	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
Linear Regression	5.208	0.175	7.577	0.178	11.669	0.054	4.535	0.235	5.042	0.183	5.431	0.259	6.577	0.181
Decision Tree	11.128	-0.763	14.855	-0.612	17.996	-0.459	10.632	-0.793	8.475	-0.373	11.818	-0.612	12.484	-0.602
Random Forest	4.807	0.238	7.521	0.184	9.438	0.235	4.238	0.285	4.413	0.285	5.347	0.271	5.961	0.250
SVR	5.539	0.122	8.214	0.109	11.488	0.069	4.154	0.299	4.624	0.251	5.493	0.251	6.585	0.183
Neural Network	5.713	0.095	7.811	0.152	11.906	0.035	4.880	0.177	5.388	0.127	6.348	0.134	7.008	0.120
Voting Ensemble	**4.742**	**0.249**	**7.348**	**0.203**	**9.640**	**0.219**	**4.043**	**0.318**	**4.368**	**0.292**	**5.028**	**0.314**	**5.862**	**0.266**

model's predictions for students' final grades in Mathematics (G3 math) and Portuguese (G3 por).

- **SHAP force plot for a single prediction of the final Math grade (G3_math).**

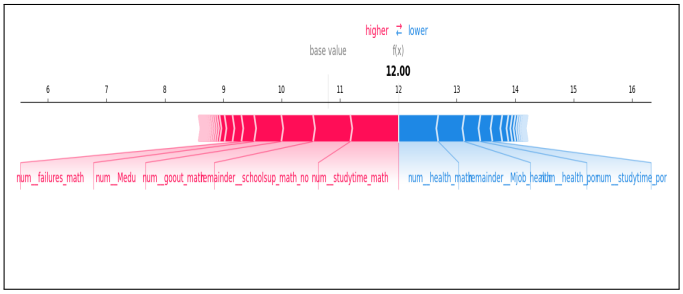

Fig. 6. SHAP force plot for a single prediction of the final Math grade (G3_math).

Figure 6 shows the factors influencing a single student's predicted Math grade. The *base value* represents the dataset's average grade, while the model predicts **12.00** for this student. Features in red (`num_failures_math`, `num_Medu`, `num_goout_math`, `schoolsup_no`, `num_studytime_math`) increase the prediction, whereas features in blue (`num_health`, `Mjob_health`, `health_por`, `num_studytime_por`) decrease it. Overall, positive effects dominate, yielding a predicted grade above average.

- **SHAP force plot for a single prediction of the final Portuguese grade (G3_por).**

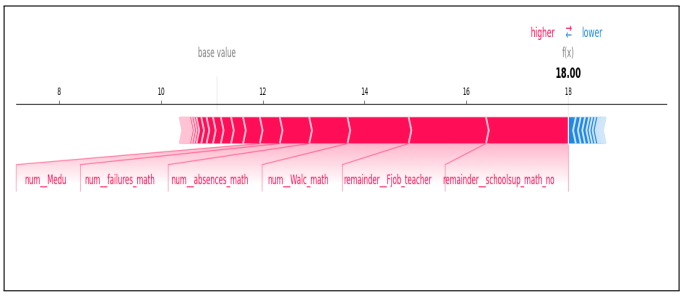

Fig. 7. SHAP force plot for a single prediction of the final Portuguese grade (G3_por).

Figure 7 shows the factors influencing a single student's predicted Portuguese grade. The *base value* is the dataset's average grade, while the model predicts **18.00** for this student. Features in red (`num_Medu`, `num_failures_math`, `num_absences_math`, `num_Walc_math`, `Fjob_teacher`, `schoolsup_no`) increase the prediction, whereas blue features have a minor negative effect. Overall, strong positive contributions yield a grade well above average.

- **Summary of feature impacts on all test set predictions:**

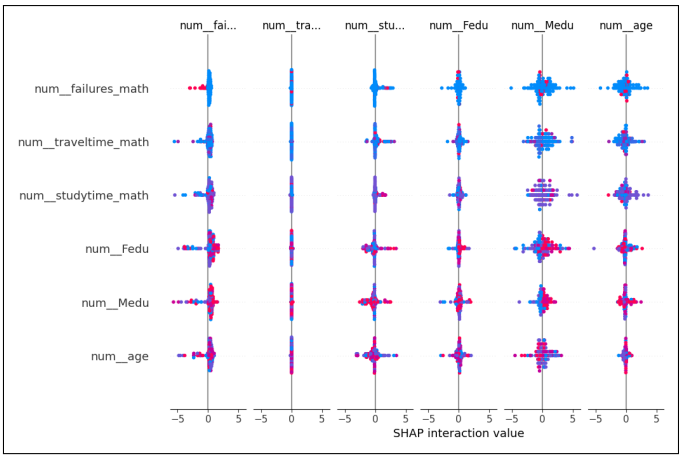

Fig. 8. SHAP summary plot showing the impact of all features on the model's predictions.

Figure 8 provides a global view of feature importance in the Random Forest model.

- **Feature Importance:** Features are ranked by influence, with `num_failures_math` as the most important, followed by `num_traveltime_math` and `num_studytime_math`.
- **Impact:** The x-axis shows SHAP values: positive values increase predicted grades, negative values decrease them.
- **Top Features:**
  - `num_failures_math`: More failures reduce grades.
  - `num_studytime_math`: More study time improves grades.
  - `num_traveltime_math`: Longer travel time often lowers grades.

TABLE II
LIME EXPLANATION FOR A SAMPLE STUDENT'S FINAL MATH GRADE (G_3 MATH)

Feature	Condition	Contribution
Predicted Value: 13.13 — Actual Value: 15.0		
Features Pushing Prediction Higher (Positive Contribution)		
Failures (Math)	≤ -0.33	+1.750
Study Time (Math)	> 0.01	+0.478
Absences (Math)	> 0.00	+0.316
Higher Education (Math)	is no	+0.280
Father's Education	> 0.65	+0.242
School Support (Math)	is no	+0.232
Father's Job (Teacher)	is yes	+0.218
Sex (Male)	is yes	+0.217
Higher Education (Math)	is yes	+0.191
Features Pushing Prediction Lower (Negative Contribution)		
Failures (Port.)	≤ -0.37	-0.338

TABLE III
LIME EXPLANATION FOR A SAMPLE STUDENT'S FINAL PORTUGUESE GRADE (G_3 PORT.)

Feature	Condition	Contribution
Predicted Value: 14.67 — Actual Value: 13.0		
Features Pushing Prediction Higher (Positive Contribution)		
Failures (Port.)	≤ -0.37	+2.375
Failures (Math)	≤ -0.33	+0.963
School (is GP)	≤ 0.00	+0.607
School Support (Math)	is no	+0.501
Study Time (Math)	> 0.01	+0.383
Study Time (Port.)	> 0.12	+0.375
Higher Education (Port.)	is yes	+0.372
Mother's Education	> 0.48	+0.327
Father's Job (Teacher)	is yes	+0.324
Features Pushing Prediction Lower (Negative Contribution)		
Sex (Male)	is yes	-0.250

LIME Analysis (Local Explanations): While SHAP provides a global perspective, LIME offers local explanations for individual predictions. Tables II and III illustrate LIME explanations for two sample students: one in Mathematics (G3 Math) and one in Portuguese (G3 Port.).

Table II shows that for the Math student, low failures, higher study time, and fewer absences increased the predicted grade, while failures in Portuguese slightly decreased it.

Table III shows that for the Portuguese student, fewer failures, higher study time in both subjects, and family background factors (parents' education and job) raised the prediction, whereas being male had a negative effect.

These local explanations demonstrate how LIME identifies the key factors driving a single student's predicted performance, complementing SHAP's global insights.

Comparative Insights: SHAP vs. LIME: LIME explanations (Tables III and IV) provide clear, case-specific insights. For Math, fewer failures, more study time, and lower absences increased the predicted grade, while Portuguese failures had a minor negative effect. In the Portuguese case, study habits, fewer failures, and supportive family background raised the prediction, with male gender slightly lowering it. These results align with educational expectations and are easily interpretable for non-technical stakeholders.

SHAP explanations (Figures 6 and 7) add a quantitative perspective by assigning contributions relative to a base value. In Math, study time, mother's education, and fewer failures raised the score, while health and parental job reduced it, yielding a prediction of 12. In Portuguese, family background, study behavior, and cross-subject performance dominated, producing a prediction of 18. SHAP therefore complements LIME by offering both direction and magnitude of influence, enabling more fine-grained interpretation.

VI. CONCLUSION

This study explores machine learning models to predict student performance in Mathematics and Portuguese, focusing on accuracy and interpretability. Various models, including Random Forest, Linear Regression, Support Vector Regression (SVR), and Decision Trees, were tested, with a particular emphasis on ensemble methods. The Voting Ensemble combining SVR, Random Forest, and Linear Regression, with weighted averaging, consistently outperformed individual models, achieving the lowest mean squared error (5.862) and the highest R^2 (0.266).

Interpretability was ensured through LIME and SHAP, which provided both case-based insights and quantitative attributions of feature influence. These methods ensured the predictions were transparent and actionable. The study confirms that ensemble-based approaches are effective for educational grade prediction, simple models remain valuable benchmarks, and incorporating interpretability techniques makes the predictions more trustworthy. Future work could expand the feature set and optimize model weighting to enhance generalizability.

REFERENCES

[1] P. Cortez and A. M. G. Silva, "Student Performance" [Dataset], UCI Machine Learning Repository, 2008. doi: 10.24432/C5TG7T.

[2] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, Dec. 2015, doi: 10.1016/j.procs.2015.12.157.

[3] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies*, vol. 28, no. 1, pp. 905–971, Jan. 2023, doi: 10.1007/s10639-022-11152-y.

[4] M. H. bin Roslan and C. J. Chen, "Educational data mining for student performance prediction: A systematic literature review (2015–2021)," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 05, pp. 147–179, Mar. 2022, doi: 10.3991/ijet.v17i05.27685.

[5] O. W. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," *Journal of Applied Research in Higher Education*, vol. 10, no. 1, pp. 61–75, 2018, doi: 10.1108/JARHE-09-2017-0113.

[6] S. O. Oppong, "Predicting students' performance using machine learning algorithms: A review," *Asian Journal of Research in Computer Science*, vol. 16, no. 3, pp. 128–148, Jul. 2023, doi: 10.9734/AJRCOS/2023/v16i3351.

[7] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, art. 237, 2021, doi: 10.3390/app11010237.