Predicting Student Exam Performance and Providing Actionable Study Recommendations

Jack Zammit, Odin Fouchier, Rafael Costa

Abstract

This project presents a machine learning approach to predicting student exam performance and translating these predictions into actionable advice for improving outcomes. We apply several models—Ridge Regression, Random Forest, Gradient Boosting, XGBoost, and a Multi-Layer Perceptron (MLP)—to a dataset of student performance factors. Ridge Regression performed best in our initial tests and after hyperparameter tuning. We then identify key features driving exam scores and develop interactive tools: one predicts the student's percentile given their features, and another suggests the required study hours and attendance needed to achieve a target percentile. Our results highlight the importance of attendance, study hours, and resource access and demonstrate how predictive modeling can inform student success strategies.

1 Introduction

Predicting student performance is critical for educators and learners. By understanding the factors influencing exam scores, students can adjust their study habits, and educators can provide targeted support. Our task is to predict exam scores from various academic and behavioral factors (e.g., hours studied, attendance, and motivation) and use this model to give actionable feedback.

1.1 Main Contributions

1. **Prediction Model Selection:** We compare multiple regression models, including linear (Ridge), tree-based (Random Forest, Gradient Boosting, XGBoost), and neural network (MLP) methods.

2. **Feature Importance Analysis:** After identifying the best model, we determine the most influential factors driving exam scores.

3. **User Interaction Tools:** We develop interactive code that predicts a student's percentile rank from their input features. We also provide a mechanism for students to set a target percentile and receive recommended attendance and study hours to achieve it.

2 Related Work

Various studies have explored machine learning models for predicting student academic performance, employing diverse algorithms and datasets to identify key influencing factors.

One study investigated the use of machine learning to predict student performance based solely on midterm grades. The dataset consisted of 1,854 students in a Turkish language class, and models such as Random Forests, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Logistic Regression, and Naive Bayes were tested. The accuracy ranged between 70% and 75%, likely due to the limited use of only three predictor features. In contrast, our approach incorporates additional variables, such as attendance and extracurricular activities, to enhance predictive accuracy. This research highlights the potential for machine learning to support timely interventions for students at risk of underperformance (Yagci, 2022).

Another study employed models like Random Forests, Decision Trees, SVM, and XGBoost to analyze data from two high schools. Attendance and study habits were identified as crucial predictors, aligning closely with our findings. Notably, the use of XGBoost achieved a high accuracy of 97.12%, demonstrating the model's effectiveness in regression tasks. This underscores the importance of identifying at-risk students early to facilitate interventions that can improve learning outcomes (Ojajuni, 2021). This study is the most similar to ours in the way that it is structured.

Additional research applied machine learning to a dataset of 649 students, incorporating 30 variables related to socioeconomic background, social activities, past grades, and study habits. Unlike our approach, this study relied heavily on categorical variables, and the target outcome was a 5 class classification model. Despite these differences, the study supports the broader conclusion that machine learning models can provide valuable insights to help educators ensure student success (Dervenis, 2022)

Prior research has demonstrated the effectiveness of linear models, ensemble methods, and deep learning techniques for predicting academic performance. Common factors influencing performance include parental involvement, socioeconomic status, and consistent study habits. While neural networks have shown promise, simpler linear models can still perform competitively under certain conditions due to their interpretability and efficiency. Our approach aligns with these findings, showing that well-regularized models can offer reliable predictions while remaining easy to interpret. .

3 Dataset and Evaluation

3.1 Dataset

We got this dataset from Kaggle, as listed in the references. We used a dataset of student performance factors (e.g., attendance percentage, study hours, sleep hours, parental involvement, teacher quality, and access to resources). The dataset contains 6607 rows and 20 features. Missing values in columns such as teacher quality, parental education level, and distance from home were imputed. We removed the *Previous_Scores* column to focus on current performance predictors. The target variable is the *Exam_Score*.

The dataset was split into training (80%) and test (20%) sets, ensuring a fair evaluation of generalization capability. A StandardScaler was applied to standardize the target variable.

3.2 Evaluation Metrics

We measured performance using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2. R^2 indicates how much variance is explained, while MSE and MAE provide prediction accuracy in the original score scale.

4 Methods

We worked with five different models at once. We used a Ridge model, which applies L2 regularization to prevent overfitting, three different tree-based models (Random Forest, Gradient Boosting, XGBoost) and an MLP to capture potential nonlinearities.

- **Ridge Regression:** A linear model with L2 regularization. Hyperparameter tuning was applied to *alpha*.
- **Tree-Based Models:** Random Forest, Gradient Boosting, and XGBoost were tested with different hyperparameters (e.g., number of estimators, max depth, learning rate).
- **MLP:** A feed-forward neural network with hidden layers, ReLU activations, and the Adam optimizer.

Feature engineering involved imputing numeric features, ordinal encoding for ordered categorical variables, and one-hot encoding for nominal categorical variables.

Why Ridge Overcomes Limitations:

The Ridge model emerged as the best performer. Although we tried more complex methods, the linear relationship in our data and careful tuning made Ridge both accurate and interpretable. While MLP offers flexibility, it did not outperform Ridge in this scenario, likely due to the dataset size, linear relationships, and limited complexity.

5 Experiments

We first trained all models with default parameters and found that Ridge achieved the highest R^2 (0.737) and lowest MSE and MAE. The results were as follows:

5.1 Results

Initial model performance:

Model	**MSE**	**MAE**	R^2
Ridge	3.71	0.79	0.737
Random Forest	4.77	1.18	0.663
Gradient Boosting	4.13	0.95	0.708
XGBoost	5.14	1.19	0.636
MLP	5.04	1.17	0.643

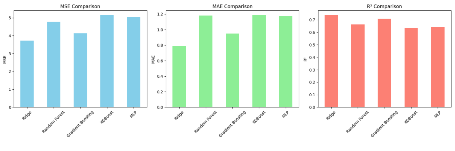


Figure 1: Visualisation of initial model performance

After hyperparameter tuning, Ridge regression retained its superiority. The top six features driving performance were:

1. Attendance
2. Hours Studied
3. Access to Resources
4. Parental Involvement
5. Internet Access
6. Learning Disabilities

6 Discussion

6.1 Error Analysis

Surprisingly, more complex models like MLP or Gradient Boosting did not outperform Ridge, indicating a relatively linear relationship or that the dataset size and complexity did not benefit more flexible models. That is why we decided to stick with ridge to follow our feature importance and user interaction research.

6.2 Graphs and Visualizations

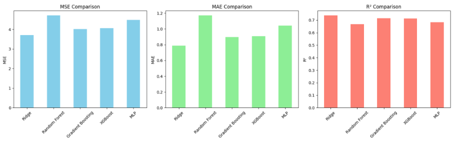


Figure 2: Model Performance Comparison after hyperparameter tuning

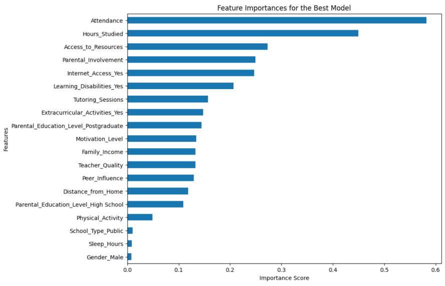
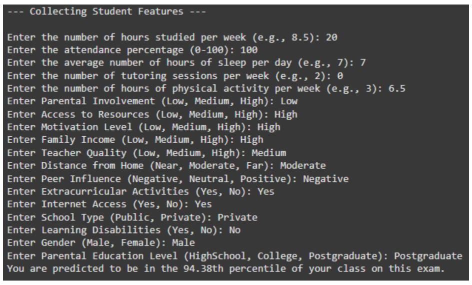


Figure 3: Feature Importance Analysis

6.3 User Interaction

Interactive functionalities include:

- **Percentile Prediction:** Given all features including attendance and hours studied, the model predicts the student's percentile ranking.

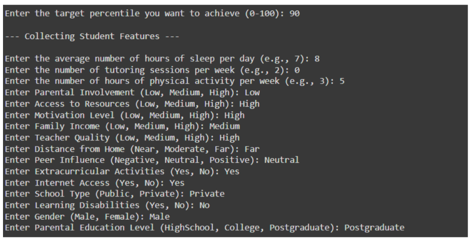


```
--- Collecting Student Features ---

Enter the number of hours studied per week (e.g., 8.5): 20
Enter the attendance percentage (0-100): 100
Enter the average number of hours of sleep per day (e.g., 7): 7
Enter the number of tutoring sessions per week (e.g., 2): 0
Enter the number of hours of physical activity per week (e.g., 3): 6.5
Enter Parental Involvement (Low, Medium, High): Low
Enter Access to Resources (Low, Medium, High): High
Enter Motivation Level (Low, Medium, High): High
Enter Family Income (Low, Medium, High): High
Enter Teacher Quality (Low, Medium, High): Medium
Enter Distance from Home (Near, Moderate, Far): Moderate
Enter Peer Influence (Negative, Neutral, Positive): Negative
Enter Extracurricular Activities (Yes, No): Yes
Enter Internet Access (Yes, No): Yes
Enter School Type (Public, Private): Private
Enter Learning Disabilities (Yes, No): No
Enter Gender (Male, Female): Male
Enter Parental Education Level (HighSchool, College, Postgraduate): Postgraduate
You are predicted to be in the 94.38th percentile of your class on this exam.
```

Figure 4: Collecting student features to calculate percentile

- **Target Percentile Guidance:** When a student inputs their uncontrollable features and desired percentile, the system iterates over attendance and hours studied levels to provide a range of feasible combinations that achieve the target percentile. This user-friendly interface demonstrates how students can achieve their desired percentile by adjusting controllable factors.

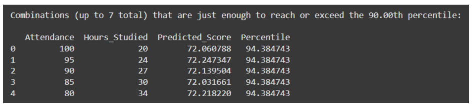


```
Enter the target percentile you want to achieve (0-100): 90

--- Collecting Student Features ---

Enter the average number of hours of sleep per day (e.g., 7): 8
Enter the number of tutoring sessions per week (e.g., 2): 0
Enter the number of hours of physical activity per week (e.g., 3): 5
Enter Parental Involvement (Low, Medium, High): Low
Enter Access to Resources (Low, Medium, High): High
Enter Motivation Level (Low, Medium, High): High
Enter Family Income (Low, Medium, High): Medium
Enter Teacher Quality (Low, Medium, High): High
Enter Distance from Home (Near, Moderate, Far): Far
Enter Peer Influence (Negative, Neutral, Positive): Neutral
Enter Extracurricular Activities (Yes, No): Yes
Enter Internet Access (Yes, No): Yes
Enter School Type (Public, Private): Private
Enter Learning Disabilities (Yes, No): No
Enter Gender (Male, Female): Male
Enter Parental Education Level (HighSchool, College, Postgraduate): Postgraduate
```

Figure 5: Collecting student features



```
Combinations (up to 7 total) that are just enough to reach or exceed the 90.00th percentile:

   Attendance  Hours_Studied  Predicted_Score  Percentile
0         100             20        72.060788   94.384743
1          95             24        72.247347   94.384743
2          90             27        72.139504   94.384743
3          85             30        72.031661   94.384743
4          80             34        72.218220   94.384743
```

Figure 6: The result that gives you the combinations

7 Conclusion

We set out to predict student exam performance and give actionable study advice. After exploring multiple ML models, Ridge regression proved most effective. By identifying key features, we gained insights into what matters most for exam success and built interactive tools to help students strategize their study efforts and attendance.

7.1 Future Directions

- Making it major specific or school level specific (Highschool, College, Postgraduate, etc).
- Explore deeper neural architectures or incorporate more advanced feature engineering to

improve the MLP's competitiveness. This dataset might have more of a linear relationship, but in the future, we can use more datasets and try more complex models.

- Enhance the user interface for broader accessibility.

References

References

[1] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Minaei-Bidgoli, Behrouz, and William F. Punch. "Predicting Student Academic Performance Using Machine Learning." *Springer Professional*, 2021. Available at: https://www.springerprofessional.de/en/predicting-student-academic-performance-using-machine-learning/19648996.

[3] Dervenis, Charalampos . "Exploring the Use of Machine Learning Models in Predicting Student Success." *ACM Digital Library*, 2022. Available at: https://dl.acm.org/doi/10.1145/3564982.3564990.

[4] Almulla, Mutlaq, et al. "The Effectiveness of Machine Learning in Education Research." *Smart Learning Environments*, 2022. Available at: https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z.

[5] mjmagno . "Student Performance Factors EDA." *Kaggle*, 2024. Available at: https://www.kaggle.com/code/mjmagno/student-performance-factors-eda/