Predicting Student Performance Using Supervised Machine Learning Approaches

Mete Harun Akcay

I. Introduction

Accurately predicting the final grade of a student beforehand is a critical challenge in education. Such predictions can provide valuable insights into student behaviour, help educators intervene when necessary and apply personalized strategies for students not having promising final grades. In this study, the goal is to predict students' final grades based on their assessments and engagement patterns.

The dataset contains detailed information about students' academic performance, including scores from quizzes, mini-projects, and peer reviews, as well as their interactions with course materials. Four machine learning models were chosen for this analysis: Linear Regression, Random Forest Regression, Random Forest Classifier, and XGBoost Classifier. Predictions made by the regression models were rounded to the closest integer since the target class contained discrete values. All the models were trained on a selection of features representing student performance and activity, and their results were compared to determine which model is more effective at predicting grades.

This study does not only aim to make grade predictions, but it also to seeks to reveal the most significant factors that influence student performance, providing useful insights that can help improve teaching methods.

II. Data Processing

The dataset used in this study, which consists of 107 rows and 48 columns, includes grades from various assessments such as quizzes, mini projects and peer reviews, as well as logs of student activity during the semester. Upon examining the dataset, there were no missing values; therefore, no imputation was necessary. However, all the values for the column Week1_Stat1 were 0, which makes sense since there were probably no assignments for students to review in the first week. Therefore, this column was dropped with the first column which has the ID of the students.

A correlation analysis, which is given in the heat map given in Figure 1, was conducted to evaluate the relationship between each feature and the target variable. Features that were highly correlated with the target variable were the total grade of the week 8, mini project grades and peer review grades. It was also discovered that course and assignment related logs were more correlated with the final grade than grade and forum related logs.

Features having correlation coefficient lower than 0.5 with the target were removed from the dataset to enhance model performance and reduce complexity. Additionally, the column Week8_Total, which had a correlation coefficient of 0.97 with the target variable, was removed from the dataset. This decision was made not only to prevent overfitting but also because the column represented the total points accumulated by a student up to the 8^{th} week. Given its strong correlation with the final grade, retaining this feature would undermine the predictive nature of the model, as it would essentially serve as a direct indicator of the target rather than contributing to a meaningful prediction. As a result, the data frame ready for the modelling phase had 107 rows and 19 columns.

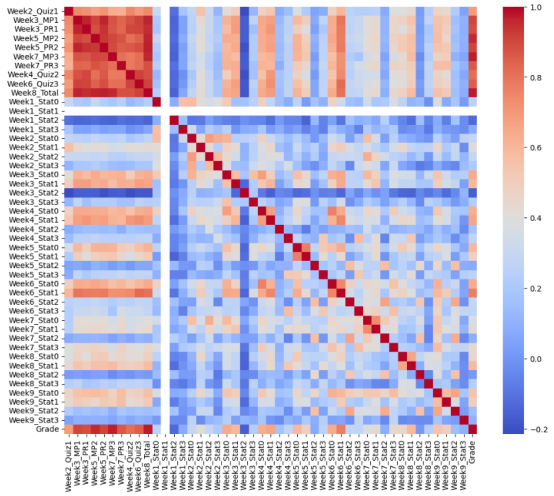

Fig. 1: Correlation Matrix

III. Data Analysis

To simplify the dataset and enhance interpretability, a process of feature aggregation was applied. Columns representing similar metrics across different weeks, such as Week1_Stat1, Week2_Stat1, and so forth, were consolidated into a single feature, labeled as Stat1. This aggregation was done for all similar statistical features across multiple weeks, resulting in a unified representation of each stat.

Additionally, columns representing different metrics within the same week were combined to form week-specific features, providing a broader view of student activity and performance during each week. This transformation enabled a clearer understanding of the dataset by reducing redundancy and allowing for more intuitive analysis of trends across both individual metrics and time periods.

Another correlation matrix, shown in Figure 2, was created to provide a broader view of the correlations between key

performance metrics such as Quiz, MP, and other statistical features. This matrix not only highlights the relationships between these individual metrics but also provides a comprehensive understanding of how they collectively contribute to the final grade. Based on the correlation matrix, the following observations about the features were made:

- **MP (Mini Project)** shows the highest correlation with the final grade (98%).
- **PR (Peer Review)** and **Quiz** follow as other highly correlated features with coefficients 0.92 and 0.85, respectively.
- **Stat1 (assignment-related logs)** demonstrates a significant correlation as well (80%), followed by **Stat0 (course-related logs)** (72%)
- **Stat3 (forum-related logs)** has a relatively low correlation coefficient of 0.43 while **Stat2 (grade-related logs)** has the weakest correlation among the statistical features with a coefficient 0.16.
- Except for the first and last weeks, which do not correlate with the target at all, the remaining weeks were found to be moderately correlated with the final grade.

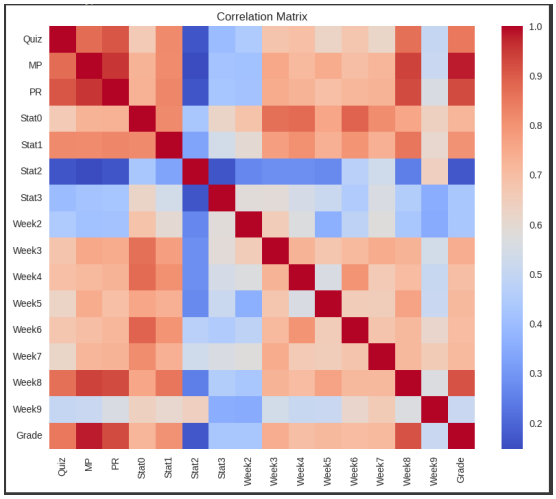

Fig. 2: Generalized Correlation Matrix

This correlation pattern can be interpreted based on student behaviors throughout the course. Students who performed well in the course typically displayed consistent engagement with assignment and course-related content, as indicated by the strong correlations for assignment-related logs and course-related logs.

On the other hand, grade-related logs and forum-related logs exhibited weaker correlations. It is likely that students who were concerned about their grades, particularly those who were at risk of failing, frequently checked their grades. This behavior, commonly observed among underperforming students, explains why no significant correlation was found between grade-related logs and the final grade. Similarly, forum participation may not have been a consistent indicator of performance, as students used the forum less systematically, regardless of their success in the course.

Furthermore, the lack of correlation between the first and last weeks and the final grade can be explained by the fact that all students, regardless of performance, tend to check information during these weeks. In contrast, the moderate correlation observed for the weeks in between suggests that sustained engagement during these periods had a stronger impact on the students' overall performance.

The assignment grades (MP, PR, Quiz) on the other hand are unsurprisingly highly correlated with the final grade. This is due to the fact that these grades directly contribute to the final grade itself, rather than serving as independent predictors. As a result, they are not just indicators of student performance but rather integral components of the final grade calculation, which explains their strong correlation with the target class.

An interesting finding about this study is the relationship between mini project and peer review grades. The scatter plot given in Figure 3 illustrates the relationship between Week 5 Mini Project (MP2) grades and Week 5 Peer Review (PR2) grades. As seen in the plot, there is little correlation between the two sets of grades. This suggests that the peer review grades and mini project grades are not aligned.

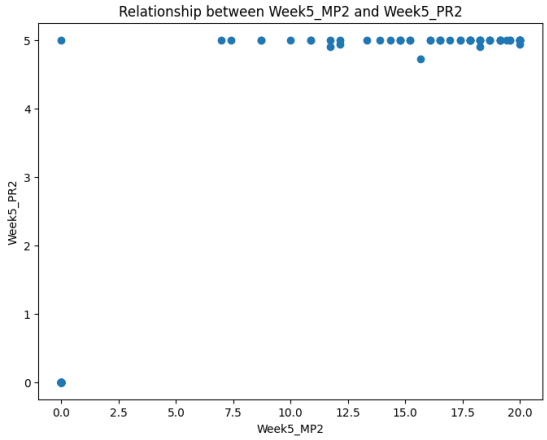

Fig. 3: Relationship Between MP2 and PR2

One possible explanation for this lack of correlation is the differing grading perspectives between the professor and the students. While the professor may assign lower grades based on stricter evaluation criteria, students tend to give higher peer review grades to their classmates, possibly due to personal biases or the desire to be lenient.

Another finding that can be seen in the scatter plot below (Figure 4) shows the relationship between course-related logs and Quiz scores. Although the correlation is not particularly strong, there is a noticeable trend indicating that students who access course-related content more frequently tend to achieve higher quiz scores.

This suggests that students who engage with the course materials by watching lecture videos are generally better prepared for quizzes. As seen in the plot, students with higher values for Stat0 tend to score higher on quizzes. This relationship indicates that consistent interaction with course content

contributes positively to academic performance, particularly in quizzes.

However, the scatter also highlights that some students may perform well on quizzes despite relatively low course-related log activity, indicating that while engagement is beneficial, it is not the sole factor influencing quiz performance.

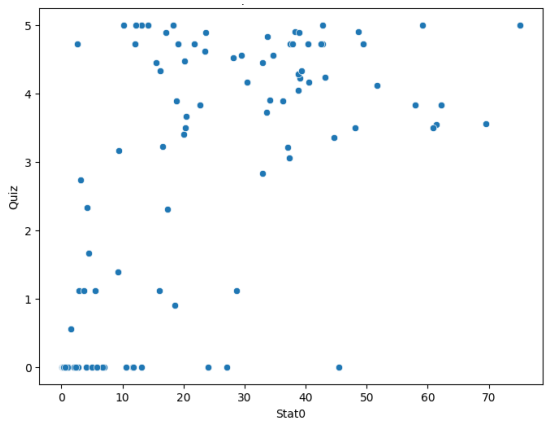

Fig. 4: Relationship Between Course-related Logs and Quizzes

One of the other interesting findings is that while students' scores on mini projects show a declining trend across the three assignments, the scores on Quizzes improve over time. Since the mini projects have different maximum scores, they were firstly normalized to have a valid comparison. As observed in Figure 5, MP1 has the highest average score, followed by a noticeable drop in MP2, and the scores further decline for MP3. This trend may suggest that students either found the major projects progressively more challenging, or their level of engagement with these assignments decreased over time.

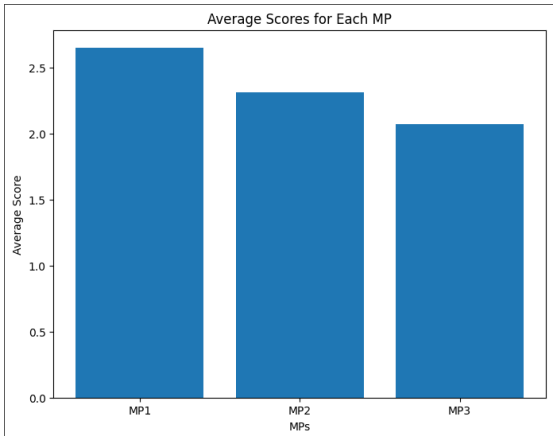

Fig. 5: Average Score of Mini Projects

In contrast, quiz scores show an upward trend, as it can be seen in Figure 6 below. Quiz1 has the lowest average score, while Quiz2 and Quiz3 show gradual improvements, with Quiz3 having the highest average score. This suggests that students may have become more accustomed to the format and content of the quizzes, possibly improving their understanding of the material as the course progressed. The upward trend in quiz performance could also indicate that quizzes, being shorter and perhaps less demanding than mini projects, allowed students to refine their test-taking strategies and perform better over time.

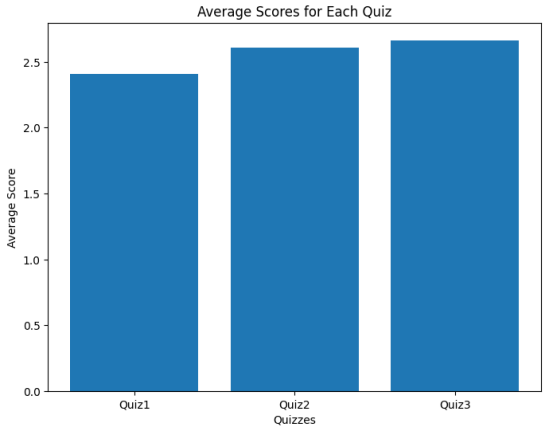

Fig. 6: Average Score of Quizzes

This contrast between improving quiz performance and declining MP performance highlights differing student engagement levels or capabilities with different types of assessments. While students seem to adapt and improve with quizzes, the increasing difficulty or complexity of the mini projects might have posed a challenge as the course content got harder.

IV. Modelling

For this project, multiple machine learning models were employed to predict students' final grades. This task is fundamentally a classification problem, as the final grades are discrete values. However, given that the final grade is determined by rounding the sum of grades to the nearest integer (for instance, if a student's calculated grade is 4.6, it is rounded up to 5), regression models were also included. The rationale behind this approach is that regression models can predict continuous values, allowing for more precise predictions that reflect the underlying patterns in the data.

After obtaining the continuous predictions from the regression models, the values were rounded to the closest integer, effectively transforming the task back into a classification problem. This hybrid approach allowed for greater flexibility in capturing subtle variations in student performance while maintaining the discrete nature of the final grade for evaluation. In this way, the regression models provided a robust foundation for predicting final grades, which were then handled as classification outcomes in the final evaluation.

The dataset was firstly decided to split into training and test sets. However, due to the fact that the dataset is quite small, splitting it once could lead to unstable or biased results depending on how the data is divided. To address this issue, 5-fold cross-validation was employed to ensure more robust and reliable model evaluation. The dataset is divided into 5 equal folds. The model is trained 5 times, each time using a

different fold as the test set while the remaining 4 folds are used for training. This process is repeated until each fold has been used as a test set exactly once. The final performance is then averaged over the 5 iterations, providing a more stable estimate of the model's generalization capability.

The dataset was trained with two regression and two classification models. Performances were compared based on multiple metrics such as mean squared error (MSE), coefficient of determination (R^2), accuracy, recall, precision and f1 scores. Moreover, confusion matrices were analyzed to determine which classes were hard to be accurately predicted.

A. Regression Models

Linear Regression and Random Forest Regression models were trained and performance of each model was measured based on MSE, R^2, and the accuracy score based on the rounded values. Figure 7 below presents a comparative analysis of Random Forest Regression and Linear Regression models.

Both models yield similar values for MSE, with Linear Regression slightly outperforming Random Forest. A lower MSE indicates that, on average, the predictions from Linear Regression are marginally closer to the actual values. Similarly, R^2 values for both models are nearly identical, while again Linear Regression performs slightly better than Random Forest. These metrics suggest that both models explain approximately the same proportion of the variance in the target variable.

A significant difference between the models is observed when rounding the predictions and calculating the accuracy. Linear Regression demonstrates a notably higher accuracy compared to Random Forest, indicating that its predictions are more aligned with the actual integer values of the target variable.

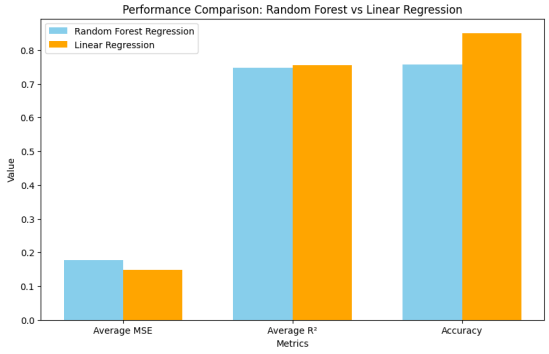

Fig. 7: Performance of Regression Models

Confusion matrices for both models, given in Figure 8, were analyzed to see where exactly the difference in accuracy stemmed from. Both models perform almost perfect when predicting grade 0. However, for higher grades, differences in performance start to emerge. For grades 3, 4 and 5, Linear Regression seems to perform significantly better than Random Forest, by accurately classifying 43 over 54 instances, compared to Random Forest's performance of correctly predicting 35 out of 54 instances. Especially, for grade 3, Random Forest misclassified almost half of the instances as grade 4.

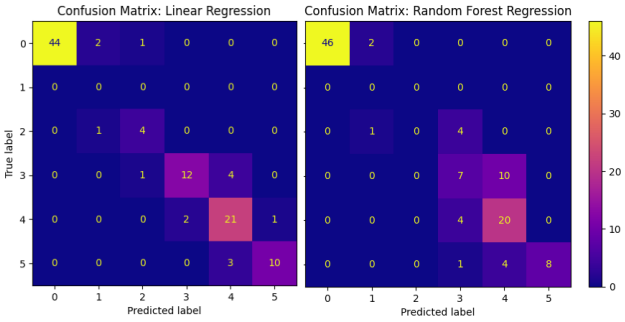

Fig. 8: Confusion Matrices of Regression Models

B. Classification Models

Random Forest Classifier and XGBoost Classifier were trained. Since XGBoost requires encoding, label encoding was applied before the training. The accuracy scores in overall were similar to the accuracy score achieved by Random Forest Regressor. While Random Forest achieved an accuracy score of 0.79, XGBoost achieved 0.75. More detailed performance analysis can be done by referring to the following figures, Figure 9 and 10, where the classification reports of both models are given.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	48
2	0.00	0.00	0.00	5
3	0.50	0.53	0.51	17
4	0.60	0.75	0.67	24
5	0.82	0.69	0.75	13
accuracy			0.79	107
macro avg	0.58	0.59	0.59	107
weighted avg	0.76	0.79	0.77	107

Fig. 9: Classification Report of Random Forest Classifier

	precision	recall	f1-score	support
0	0.98	1.00	0.99	48
1	0.00	0.00	0.00	5
2	0.47	0.53	0.50	17
3	0.58	0.62	0.60	24
4	0.62	0.62	0.62	13
accuracy			0.75	107
macro avg	0.53	0.55	0.54	107
weighted avg	0.72	0.75	0.73	107

Fig. 10: Classification Report of XGBoost Classifier

Both models handle class 0 remarkably well, with Random Forest achieving perfect precision and recall, and XGBoost coming very close. This indicates that both models are highly effective at identifying students with the lowest grades, just like the two previous models.

Where the two models begin to diverge is in their handling of the higher grade categories, particularly grade 4 and 5. Random Forest shows a clear advantage in this range, delivering

stronger precision and recall scores compared to XGBoost. This suggests that Random Forest is better equipped to manage the complexity and variation within the higher-performing students. However, both models struggle significantly with class 1, where no predictions were made correctly. This could be attributed to the small number of instances for this class, which limits the models' ability to generalize.

When considering the overall accuracy, Random Forest outperforms XGBoost slightly, achieving 79% compared to XGBoost's 75%. While this may not seem like a dramatic difference, the higher weighted averages in precision and recall for Random Forest suggest that it handles class imbalances better, especially in classes with more instances.

In short, while XGBoost and Random Forest are comparable in certain areas, Random Forest demonstrates a more consistent ability to differentiate between adjacent grades, particularly in higher and mid-range classes, making it the stronger performer overall in this task; however, when compared to linear regression, its performance is still noticeably weaker.

C. Feature Importance and Final Comparison

Based on the correlation matrices given before, most correlated variables were the following, in descending order: Week8_Total, Week7_MP3, Week5_MP2, Week5_PR2, Week3_MP1, Week3_PR1, Week7_PR3, and the quizzes. However, this does not necessarily mean that these variables are the most important for every model. Feature importance may vary depending on the underlying algorithm, as different models evaluate and utilize features in distinct ways. For instance, Linear Regression emphasizes the features that contribute most directly to reducing the prediction error, while models like Random Forest and XGBoost focus on selecting features that improve overall decision boundaries or reduce uncertainty in classification. Table 1 below shows the top 3 most important features along with their importance coefficients for each model. Since the importance order for Random Forest Regressor and Classifer was naturally the same, they were combined under the name Random Forest. Mini Project 3 is by far the most important feature for all of the models. Mini Project 2 is the second most important feature for Linear Regression and Random Forest models, while it is not in the top 3 for XGBoost. Third most important feature for each model is different: Peer Review 2 for Linear Regression, Mini Project 1 for Random Forest, and interestingly, Quiz 1 for XGBoost.

Model	First Feature	Second Feature	Third Feature
Linear Regression	MP3 (1.3)	MP2 (0.8)	PR2 (0.5)
Random Forest	MP3 (2.2)	MP2 (1.4)	MP1 (0.8)
XGBoost	MP3 (2.7)	PR3 (1.6)	Quiz1 (1.1)

TABLE I: Top 3 Most Important Features for Each Model

When the importance coefficients are analyzed, it can be said that the success of Linear Regression may be due to the fact that its feature importance coefficients are more balanced. Random Forest and XGBoost, on the other hand, seem to heavily rely on Mini Project 3 value, which might indicate that they are capturing patterns strongly tied to this specific variable, potentially at the expense of generalizing well across other aspects of the data.

Figure 11 below shows the comparison between the performances of all four of the models used in this study in terms of four different performance metrics. Linear Regression consistently outperforms the other models across all metrics, including accuracy, precision, recall, and F1-score. The strong performance of Linear Regression may be attributed to its more balanced feature importance, which enables it to capture the relationships between the input features and the target variable more evenly.

In contrast, both Random Forest models (Regressor and Classifier) demonstrate similar performance, with slight variations across the metrics. They perform relatively well, but their reliance on certain key features, particularly Mini Project 3, may have led to their more uneven performance when compared to Linear Regression. XGBoost, while competitive, tends to lag behind the other models in most metrics, indicating that its feature importance distribution and ability to generalize may not be as strong in this specific dataset.

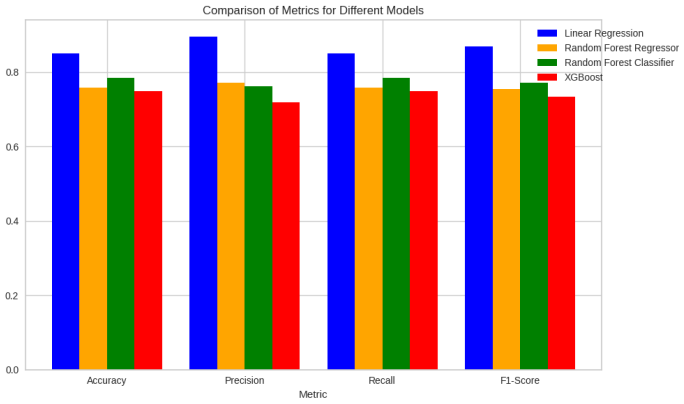

Fig. 11: Performance of the Models

V. Conclusion

This study set out to predict student performance using a variety of machine learning approaches, while also identifying the most influential factors that contribute to final grades. Several key insights emerged, both in terms of the models' effectiveness and the challenges encountered during the process.

One of the primary challenges was the small size of the dataset, which made it difficult to ensure robust generalization. To mitigate this issue, 5-fold cross-validation was employed, ensuring that model evaluations were based on multiple iterations rather than a single train-test split.

Another bottleneck was the high correlation between some features and the target variable, such as Week8 Total. While these features appeared to be strong predictors, their direct contribution to the final grade undermined the predictive nature of the models. This issue was addressed by carefully selecting

and eliminating overly correlated features to prevent overfitting, ensuring the models could generalize more effectively.

While each model demonstrated strengths and weaknesses, the use of both regression and classification approaches, along with feature engineering and cross-validation, allowed for a comprehensive evaluation of student performance. Moving forward, larger datasets with more diverse features could further improve the generalization and predictive capabilities of the models. Additionally, the inclusion of domain-specific knowledge in feature engineering could help tailor models even more closely to the context of the data.

In summary, this study demonstrates that machine learning models can be utilized in predicting student grades and identifying key performance factors. However, the effectiveness of these models depends heavily on the quality of feature selection, proper handling of data correlations, size of the dataset, and ensuring that the models generalize well beyond the training data. With careful preprocessing and validation techniques, models can provide meaningful insights into student performance, offering potential pathways for educational interventions and personalized learning strategies.