Grade Prediction of Student Performance On Machine Learning Course

Thinakone Louangdy

---

1. Introduction

Given collected data of students from a fully online nine-week machine learning course on the Moodle online management system. We aim to predict the final grade of 107 enrolled students by performing two selected supervised learning models of our choice.

To achieve the goal and tackle this issue, we will perform exploratory data analysis (EDA) from our dataset, which has 36 logs, including the complete status of learning material, quiz attempt and submission completion status, and grade and forum-related score. We will also look at the final total grade in a category and the logs of each category.

2. Data Processing and Analysis

We conducted a descriptive statistical analysis of the numeric data. The comprehensive summary included key statistical metrics such as the count, mean, standard deviation, minimum and maximum values, and quartile ranges. This descriptive analysis is essential for a deeper understanding of the distribution, variability, and spread of the numerical values in the dataset. We investigated the description of the numerical data. The mean grade of 107 students was 2.07, and the standard deviation was 1.99, with a minimum of 0 and a max of 5. To ensure the integrity of our analysis, we checked for duplicate entries in the dataset. This step is vital for maintaining data quality, as duplicates can lead to skewed results in further analysis. The count of duplicate rows provided insights into the uniqueness and redundancy of the data records. In our pursuit of focusing on quantitative analysis, non-numeric columns, specifically the 'ID' column, were removed from the dataset.

A heatmap correlation analysis was employed to identify key predictors in the feature selection process for a predictive modeling task targeting the variable “Grade,”. The initial step involved computing the correlation matrix of the dataset to quantify linear relationships between feature pairs. This matrix was then visualized using a heatmap, with the “BuPu” color scheme enhancing the interpretability of varying correlation strengths. The heatmap was rendered in a sizeable format to facilitate an easy understanding of the correlations.

A critical phase in the analysis was setting a correlation threshold of 0.5, which helped isolate features that exhibited a strong linear relationship with the target variable “Grade.” Features surpassing this threshold were deemed good predictors. The final selection list, named “features_to_consider,” excluded the target variable and included only those

features with significant correlation which are ['Week2_Quiz1', 'Week3_MP1', 'Week3_PR1', 'Week5_MP2', 'Week5_PR2', 'Week7_MP3', 'Week7_PR3', 'Week4_Quiz2', 'Week6_Quiz3', 'Week8_Total', 'Week3_Stat0', 'Week3_Stat1', 'Week4_Stat0', 'Week4_Stat1', 'Week5_Stat0', 'Week6_Stat0', 'Week6_Stat1', 'Week8_Stat1', 'Week9_Stat0']. This strategic approach streamlined the predictive modeling process, focusing on impactful features to enhance the accuracy and efficiency of the predictive model.

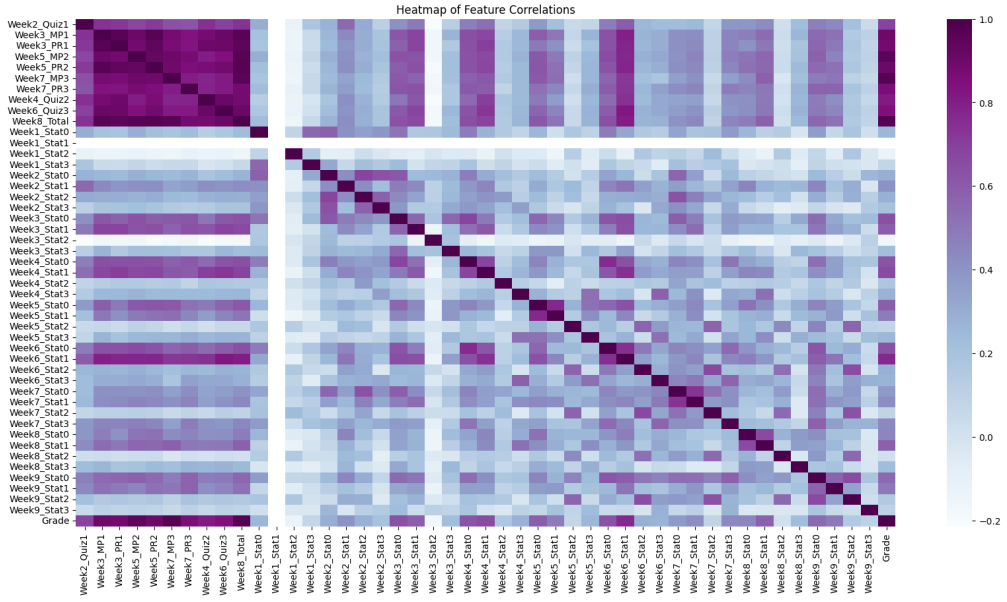

Figure 1. Heatmap of Feature Correlations

3. Model Training

In developing the predictive model, the dataset was meticulously partitioned into a set of features and a target variable, the latter being the “Grade” the model aims to predict. The feature set was carefully selected based on a correlation threshold, ensuring that only variables with a substantial relationship to the target variable were included. This selection was underpinned by the hypothesis that features with higher correlation coefficients are more likely to yield accurate predictions.

Following the feature selection, the dataset underwent a splitting process which allocated 80% of the data to training and the remaining 20% to testing. This stratagem is designed to train the model on a substantive portion of the data, facilitating the model's ability to generalize from the training data to unseen data. The Linear Regression model was then trained on this training set. Linear Regression was chosen for its simplicity and efficiency in establishing a linear relationship between the independent and dependent variables.

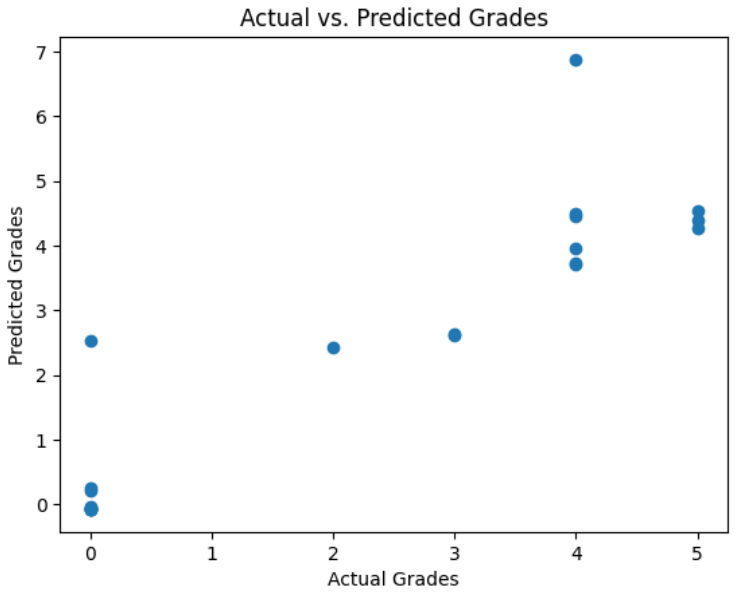

Figure 2. Actual Vs. Predicted Grade

The culmination of this process was the evaluation phase, where the model's predictions, derived from the test set, were juxtaposed with the actual grades. The performance metric, Mean Squared Error (MSE), was computed, yielding a value of 0.772. This metric, indicative of the model's accuracy, measures the average squared difference between the estimated values and what is estimated. A complementary scatter plot visually represented the model's predictions against the actual grades, clearly depicting the prediction accuracy. The distribution of data points around the line of perfect prediction highlighted the variances and served as a diagnostic tool to assess the model's predictive behavior.

To further enhance the model's predictive capabilities, a suite of different algorithms was employed, each with its distinct mechanics and potential for accuracy. These included Random Forest, Support Vector Machine (SVM), and Decision Tree classifiers. The Random Forest Classifier, known for its robustness and ability to manage overfitting, was set up with 100 estimators. The SVM, with a linear kernel, was chosen for its effectiveness in high-dimensional spaces, and the Decision Tree Classifier was selected for its interpretability and ease of use.

Each model was trained using the same data, ensuring consistency and comparability across results. The models' performances were rigorously evaluated against the testing set, with key metrics—accuracy, precision, recall, and F1-score—being compiled into a comprehensive classification report. This report served as the basis for a comparative analysis of the models, encapsulating the predictive success of each algorithm in numeric terms.

	Model	Accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)
0	Random Forest	90	0.885281	0.909091	0.888112
1	Support Vector Machine	81	0.886364	0.818182	0.830622
2	Decision Tree	100	1.000000	1.000000	1.000000

Figure 3. Model Scores Comparison

The Random Forest model achieved a commendable accuracy of 90%, with precision and recall metrics closely aligned, indicating a balanced prediction capability across different grade categories. While less accurate at 81%, the SVM model still maintained a high precision, suggesting that when it predicted a grade category, it did so with a high degree of confidence. The Decision Tree model stood out with a perfect accuracy of 100%. This exceptional result may warrant further investigation to validate robustness and ensure it is not a result of overfitting.

These results were then visualized in a bar plot, offering a clear and immediate graphical representation of each model's accuracy. Such a visualization is invaluable in communicating the outcomes to technical and non-technical stakeholders, allowing for informed decision-making regarding the model's future application and potential deployment. The analysis concluded with the Random Forest and Decision Tree models showing extreme performance, suggesting they could be promising candidates for the final predictive modeling task.

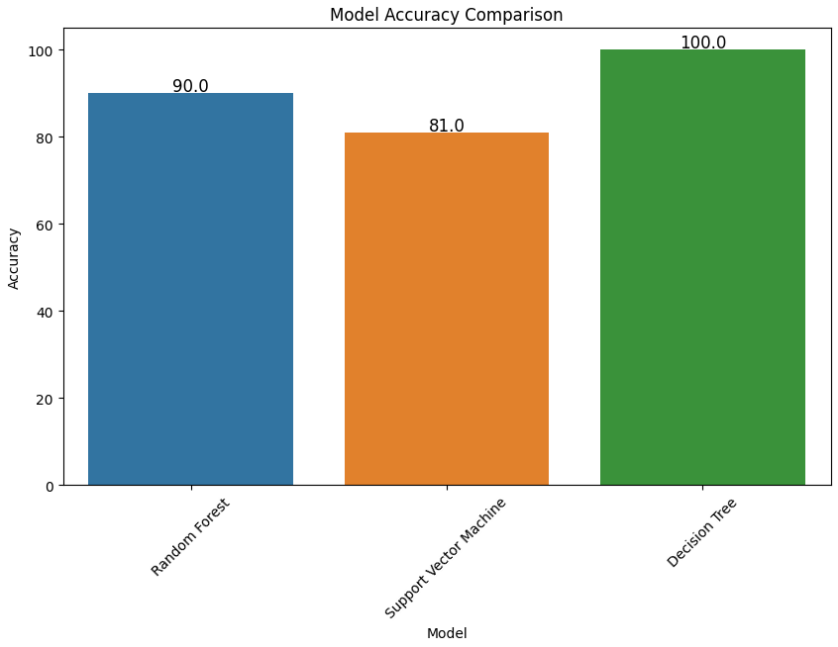

Figure 4. Model Accuracy Comparison

The scatter plot you have created serves as a comparative visualization that juxtaposes the True values of the target variable against the predicted values obtained from multiple predictive models. Each model's performance is quantified by its Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values, offering a gauge of the model's prediction accuracy.

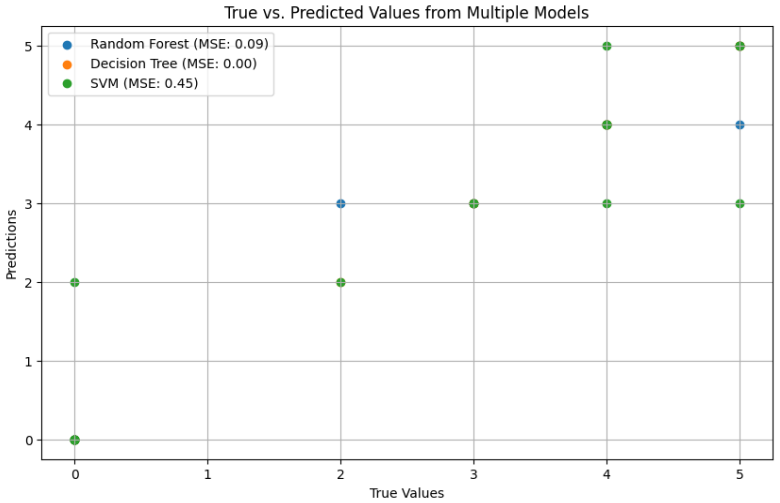

Figure 5. True Values Vs. Predicted Values

From the plot, it is discernible that the Decision Tree model achieves a perfect MSE score of 0.00, indicating that all its predictions match the true values exactly. While this may suggest an ideal model performance, it can also raise concerns about overfitting, where the model learns the training data too well, including the noise, which could lead to poor generalization of new data.

The Random Forest model, an ensemble method known for handling overfitting more effectively than individual decision trees, shows a low MSE, indicating high accuracy but not to the extent of the Decision Tree model. On the other hand, the Support Vector Machine (SVM) models exhibit higher MSE scores, implying less accuracy in their predictions compared to the ensemble methods.

4. Conclusion

In concluding this report, it's essential to acknowledge the initial bottlenecks faced due to my limited experience in predictive modeling. Despite this, the comprehensive journey through the project phases—from data processing and model training to evaluation—has been a substantial learning curve. The project began with the clear goal of predicting student grades from an online course, utilizing a dataset composed of various logs. Rigorous data processing was conducted to ensure quality, involving the careful removal of non-numeric entries and duplicates. The use of heatmap correlation was instrumental in selecting features that significantly correlated with the target variable, “Grade.”

The model training phase started with Linear Regression, favored for its simplicity in illustrating linear relationships. Model efficacy was assessed using Mean Squared Error (MSE) and visualized through a scatter plot, revealing a comparative analysis of actual versus predicted grades.

Exploring further, I trained additional models—Random Forest, SVM, and Decision Tree classifiers. The Random Forest model displayed a robust stance against overfitting, the Decision Tree presented perfect accuracy with an asterisk of potential overfitting, while the SVM maintained high precision despite a lower accuracy score. A visual comparison using scatter plots across all models elucidated their predictive performance. Coupled with an exhaustive classification report, this illustrated the strengths and limitations inherent in each model.

As I delved deeper into the course content, my understanding and knowledge expanded, enabling me to overcome the initial challenges. The progress made is clearly reflected in the results of this project, where informed decisions have now been made possible for future model optimizations and selections based on the insights garnered through this experience.