Student Grade Prediction System

Shihabur Rahman Samrat

Abstract

This mini project uses supervised learning to forecast students' final grades in an online machine learning course. Key steps include identifying relevant features through correlation analysis, data splitting, and comparing the performance of Random Forest and Support Vector Machine (SVM) classifiers. Random Forest outperforms SVM with an accuracy of 88% compared to SVM's 73%, excelling in predicting students' grades. Feature importance analysis identifies Week 8 Total, Week 5 Mini Project 2, and Week 7 Mini Project 3 as the most influential features. This project highlights the potential of machine learning for grade prediction of online courses.

1. **Introduction**

This project uses supervised learning approaches to predict students' final grades in an online machine-learning course. The dataset contains anonymized information on 107 enrolled students, including their grades, course logs, and other relevant details. In this report, we will detail the steps involved in data processing, model training, performance evaluation, and the identification of essential features.

2. **Step 1: Data Processing**

In this phase, we checked for missing values and selected the most relevant features for predicting final grades. There were no missing values in the dataset. We also checked whether any of the columns have the same value for all students; in that case, we dropped that column since that value would be redundant. In our case, *Week1_Stat1* had the same value for all entries, so we dropped it.

After this, a correlation matrix was formed in relation to our target variable, "*Grade*". This showed all the strong and weak correlations that the other attributes have with our target variable. This matrix was used to cherry-pick the features that had a correlation greater than 0.3 or lesser than -0.3. Anything other than this range was discarded since they did not have any substantial impact on the grade of the student. So, we did not use all the available features. Our final feature set included the following features -

'Week2_Quiz1', 'Week3_MP1', 'Week3_PR1', 'Week5_MP2', 'Week5_PR2', 'Week7_MP3', 'Week7_PR3', 'Week4_Quiz2', 'Week6_Quiz3', 'Week8_Total', 'Week2_Stat1', 'Week3_Stat0', 'Week3_Stat1', 'Week4_Stat0', 'Week4_Stat1', 'Week5_Stat0', 'Week5_Stat1', 'Week6_Stat0', 'Week6_Stat1', 'Week7_Stat0', 'Week7_Stat1', 'Week7_Stat3', 'Week8_Stat0', 'Week8_Stat1', 'Week9_Stat0', 'Week9_Stat1'

3. **Step 2: Data Split - Training and Test Sets**

The dataset was divided into training and test sets with a 70/30 split using the *test_train_split* function. A 70/30 split has a balance between utilizing a significant portion of the data for training while still reserving a substantial amount for testing. It helps prevent issues like overfitting, where the model learns the training data too well but struggles to generalize.

4. **Step 3: Model Training**

Two different supervised learning approaches were selected for our classification task - Random Forest and Support Vector Machine (SVM). Random Forest is a versatile ensemble learning method that works well with a wide range of data types. It is robust to outliers and can handle both categorical and numerical features. This was my primary choice. For the sake of this task, since we had to do a comparison, I chose the SVM classifier just to understand how both would stack against each other.

Both these models were trained on the training dataset and evaluated on the test dataset.

- **Random Forest Classifier** achieved an accuracy of 88% in predicting students' final grades.
- **SVM Classifier** achieved an accuracy of 73% in predicting students' final grades.

Random Forest outperformed SVM, indicating that it was a better model for this prediction task.

5. **Step 4: Performance Evaluation**

Performance was evaluated by comparing predicted and actual grades. The scatter plots in figure 1 and 2 below demonstrate that both models had a reasonable fit to the data, with Random Forest showing a slightly better prediction.

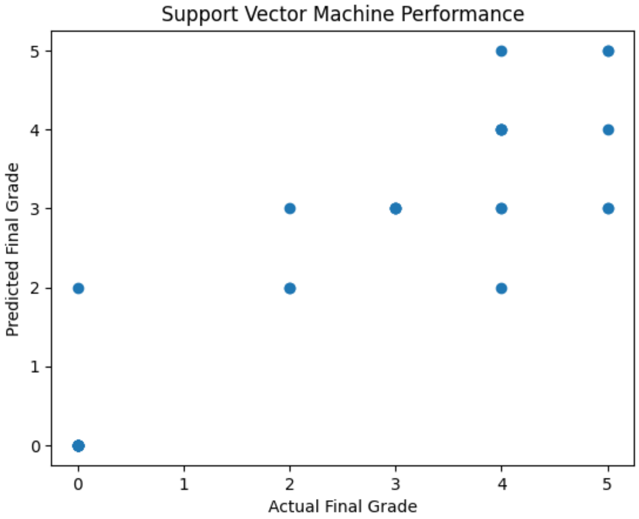


Figure 1: SVM Classifier scatter plot

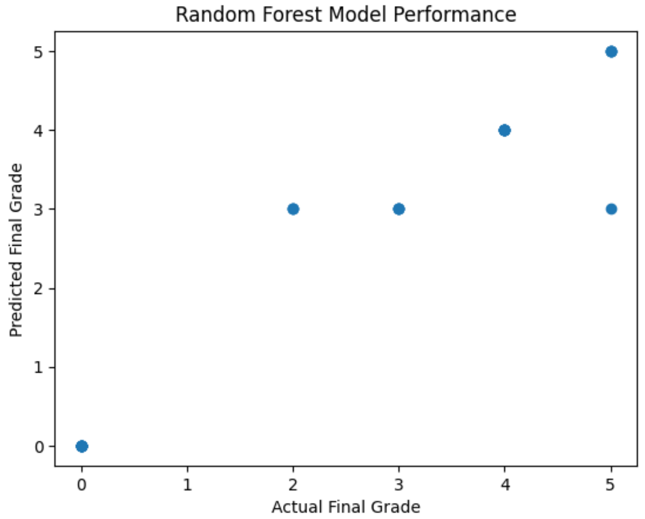
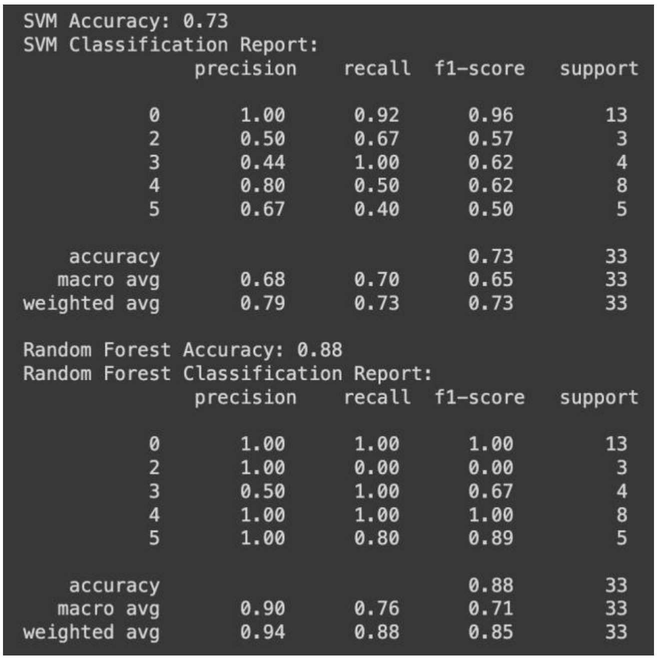


Figure 2: Random Forest Classifier scatter plot

The accuracy alongside a classification report (which contains precision, recall, f1-score, etc.) was generated for both models as shown in figure 3. The Random Forest model generally outperforms the SVM model in terms of accuracy (0.88 vs. 0.73). This suggests that, on average, the Random Forest model is making more accurate predictions.

However, both models seem to struggle with class 2, as indicated by the low recall and F1-score for that class. This suggests that there may be issues with class imbalance or the inherent difficulty of classifying this class. The Random Forest model tends to have higher precision for some classes, meaning that it is often correct when it predicts a class (e.g., class 0 and class 4). The SVM model has more balanced precision and recall for some classes (e.g., class 0 and class 3), indicating that it does not over-predict or under-predict as much as Random Forest.



```
SVM Accuracy: 0.73
SVM Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.92      0.96        13
           2       0.50      0.67      0.57         3
           3       0.44      1.00      0.62         4
           4       0.80      0.50      0.62         8
           5       0.67      0.40      0.50         5

    accuracy                           0.73        33
   macro avg       0.68      0.70      0.65        33
weighted avg       0.79      0.73      0.73        33

Random Forest Accuracy: 0.88
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        13
           2       1.00      0.00      0.00         3
           3       0.50      1.00      0.67         4
           4       1.00      1.00      1.00         8
           5       1.00      0.80      0.89         5

    accuracy                           0.88        33
   macro avg       0.90      0.76      0.71        33
weighted avg       0.94      0.88      0.85        33
```

Figure 3: Accuracy and classification report of both models

6. **Step 5: Important Features**

The three most important features in predicting students' final grades were found using feature importance analysis in the Random Forest model. These features were:

1. Week 8 Total
2. Week 5 Mini Project 2
3. Week 7 Mini Project 3

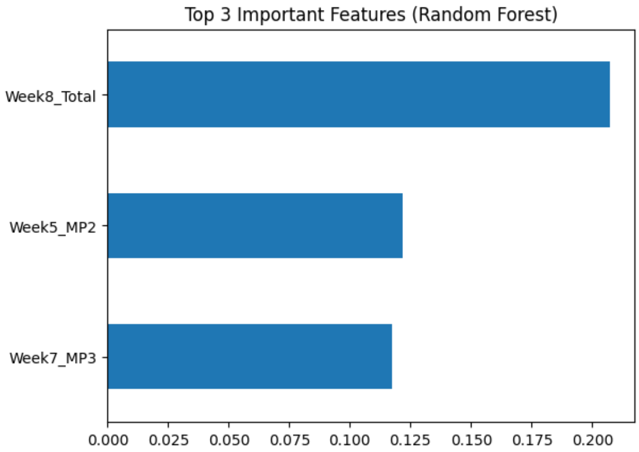


Figure 4: Top 3 important features for Random Forest

7. Conclusion

We successfully predicted students' final grades using supervised learning models in this project. Random Forest emerged as the better model compared to SVM, achieving an accuracy of 88%. We identified the three most important features contributing to this prediction: Week 8 Total, Week 5 Mini Project 2, and Week 7 Mini Project 3. The project encountered no significant bottlenecks, and the challenges were addressed through proper data preprocessing and model selection.

Overall, this project demonstrates the potential of using machine learning to predict students' performance in online courses. Further improvements can be made by fine-tuning the model hyperparameters and exploring other algorithm options.