Mini Project 2: Predicting Student Performance Using Machine Learning Techniques

Mudita Shakya

1. Introduction

Grades and course activities are key indicators of how well a student performs in a course. It is important for instructors to identify students who are struggling with the course and those at the risk of dropping the course in order to provide extra support so that they can improve their performance.

To solve this issue, we can use the grades and course activities data collected from the students to train machine learning algorithms and predict the overall student performance. This would help instructors and student tutors focus and help the students with low predicted grades to improve their performance and help decrease course dropout rates.

In this project, two approaches: Random Forest Classifier and Support Vector Machine Classifier (SVC) models from scikit-learn were used for prediction. The tools used for the project were:

i. Scikit-learn
ii. NumPy
iii. Pandas
iv. Seaborn
v. Matplotlib

2. Data description

The dataset we are using contains the grade and the course logs collected from 107 students for a 9-week long course on Machine Learning, hosted on Moodle. The data includes the grades received from 3 mini projects, 3 quizzes, and 3 peer reviews and the different status columns signify scores related to the following:

Status0: course/ lectures/ content related

Status1: assignment related

Status2: grade related

Status3: forum related

3. Data Processing

Before using the data for model training and prediction purposes, we need to preprocess the data. The dataset was checked for null values and no null values were found. It was also important to check for and remove columns that did not have any varying values. After checking, the column “Week1_Stat1” was found to have only 0 as the values so this column was removed from the data frame. The values for the final “Grade” column were checked and it was found that the dataset was not balanced. We have 36 status columns (4 for each week) and 10 grade columns. We can reduce the number of columns for the status data by summing all the data for each status so that all status data can be reduced to 4 columns (‘stat0’, ‘stat1’, ‘stat2’, ‘stat4’).

4. Feature selection

For training a good model, it is necessary to identify features that will provide the most relevant information to the model. Identifying important features and reducing the features used helps to decrease the noise introduced by irrelevant features and improve the performance of the model. It is important to select features that have the strongest correlation with the value to be predicted. The columns such as “ID” were removed along with the final prediction value column: “Grade”. The column “Week8_Total” also was removed since it is just a different scaled version of the Grade column. The correlation of the remaining columns was checked with the grade column.

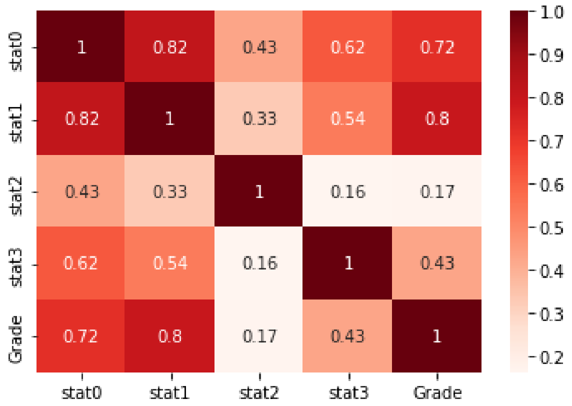


Fig 1.: Heat Map of the correlation matrix of the Status columns with Grade

From the correlation matrix of the Status columns, it was evident that the Status0 (course related) and Status1 (assignment related) columns had high correlation with grade whereas Status2 (grade related) and Status3 (forum related) columns had low correlation with grade.

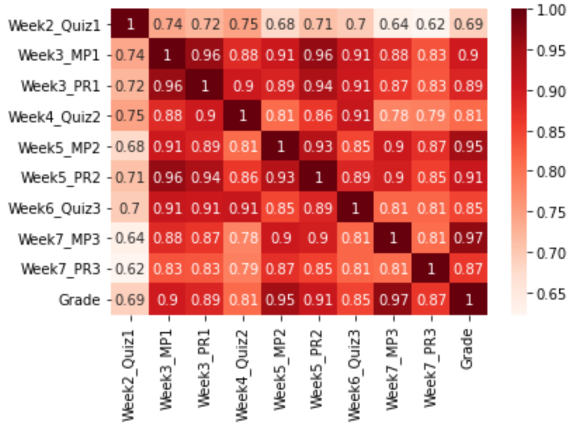


Fig 2.: Heat Map of the correlation matrix of the Grade columns and the final Grade

From the correlation matrix of the Grade column, it was clear that most of the grade columns had high correlation with the final grade except the Week2_Quiz1 column.

For selecting the best features, the feature importance property of Random Forest Classifier was used. The columns with feature importance values greater than a threshold of 0.07 (found through tests) were set to be the final features. The final features used were:

- Week7_MP3
- Week5_MP2
- Week3_MP1
- stat1
- Week4_Quiz2

All features other than these were not used for the model training and prediction.

5. Model training

The dataset was divided in 80:20 ratio for preparing the train and test datasets which resulted in a training dataset with 91 rows and a test dataset with 16 rows.

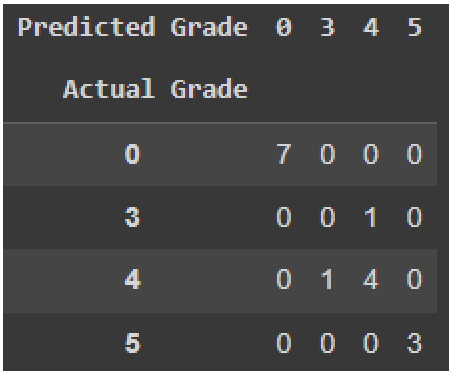
The selected final features were used for training both the Random Forest classifier as well as the Support Vector Machine Classifier. The Random Forest classifier was trained on the training dataset and the trained model was used for prediction. In the case of the SVC model, an extra step to scale the feature values was added. This step was not necessary for Random Forest since it uses decision trees which is why the transformation of a single variable is captured by a tree.

The trained model was used to predict the grade for the test set and the predictions were saved in the test data frame as **random_forest_predictions** and **svc_predictions** respectively.

6. Performance evaluation

The accuracy score metric was used for evaluating the performance of the two trained models. The accuracy score was computed by comparing the actual grade values with the predicted grades.

The Random Forest Classifier was able to get an accuracy of 87.5% and produced the following result.



Predicted Grade	0	3	4	5
Actual Grade				
0	7	0	0	0
3	0	0	1	0
4	0	1	4	0
5	0	0	0	3

Fig 3.: Predicted grades compared with actual grade for Random Forest Classification

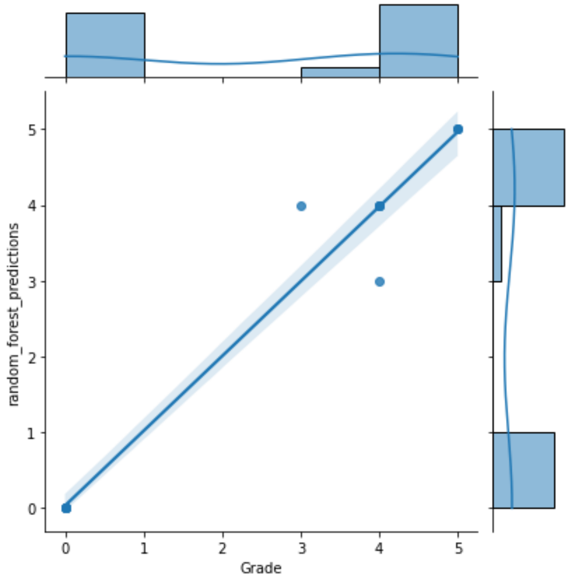
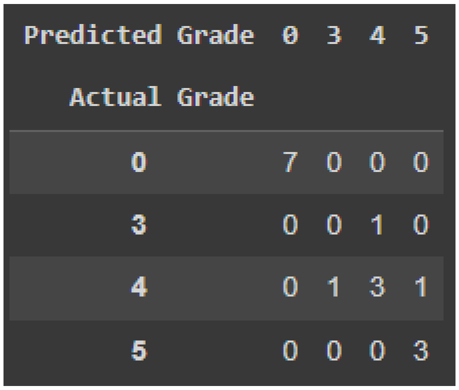


Fig 4.: Scatter plot of Grade and Predictions from Random Forest Classifier

The Support Vector Machine classifier was able to get an accuracy of 81.25% and produced the following result.



Predicted Grade	0	3	4	5
Actual Grade				
0	7	0	0	0
3	0	0	1	0
4	0	1	3	1
5	0	0	0	3

Fig 5.: Predicted grades compared with actual grades for SVC

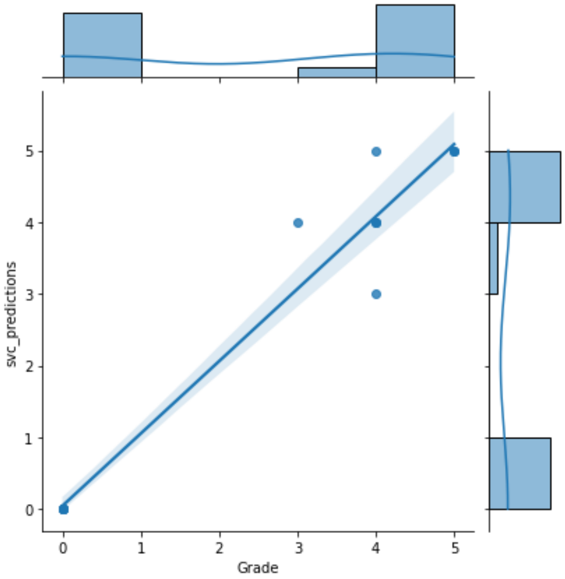


Fig 6.: Scatter plot of Grade and Predictions from SVC classifier

The top three most important features were identified to be:

i. Week7_MP3
ii. Week5_MP2
iii. Week3_MP1

7. Conclusion

There were some bottlenecks experienced when implementing this. Some of the issues found with this case are:

i. The dataset is not balanced. For a dataset to be balanced, it is necessary to have almost the same number of data points for each label.

ii. There is not enough data. The dataset contains the grade and course interaction data for 107 students. However, this is not enough to create a generalized model that works with all kinds of data.

iii. The data contains a lot of features. It is difficult to identify the important features that would provide the most information to the model to provide a decent prediction. Using the feature importance property of random forest classifier was very useful to identify the most important features for the model.

It was found that the Random Forest Classifier performed better than the Support Vector Machine Classifier. Random Forest Classifier uses many decision trees in an ensemble. The trees protect each other from their individual errors.

Throughout this project, I was able to learn more about feature selection and understand how important it is in the machine learning pipeline. I was able to train a Random Forest classifier and an SVC classifier on the data and get prediction results.

The most important features found were:

i. Week7_MP3
ii. Week5_MP2
iii. Week3_MP1

From this, it can be concluded that mini projects have a huge influence on the overall grade of the student and can be used as an indicator to identify struggling students and help improve their performance. Among the status variables, status0 and status1 were found to have the highest correlation with the grade value so keeping track of course interaction and if the student does his/ her assignments diligently and on time can be valuable indicators to the final performance of the student.